UC San Diego
UC San Diego Electronic Theses and Dissertations

Title

Understanding How Context Influences Function Across Biological Scales in Multicellular Mammalian Systems

Permalink

https://escholarship.org/uc/item/91m6s7f8

Author

Baghdassarian, Hratch Matthew

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org
Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Understanding How Context Influences Function Across Biological Scales in Multicellular Mammalian Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Hratch Matthew Baghdassarian

Committee in charge:

Professor Nathan E. Lewis, Chair
Professor Terence Hwa, Co-Chair
Professor Eric Bennett
Professor Bernhard Palsson
Professor Gene Yeo

2023

Copyright

Hratch Matthew Baghdassarian, 2023

All rights reserved.

The dissertation of Hratch Matthew Baghdassarian is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

TO MY PARENTS, JOYCE AND ALEX
TO MY SISTER, KARINE, and
TO ARJANA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people in my personal and professional life without whom I could not have arrived at this stage. I would like to thank Professor Nathan E. Lewis for his role as my advisor. He has provided key insights for all content in my dissertation, and beyond this, he has been an invaluable mentor who has supported me in all aspects of my graduate career. I could not have asked for a kinder and more supportive mentor. I first met Nate when he interviewed me for the program, and from that very first moment throughout the five years we worked together, he has made me feel comfortable to come to him with any challenges I am facing. We've had countless discussions ranging from specific details dealing with technical challenges in projects to being proactively creative in generating scientific ideas, to managing work-life balance. This openness helped me to advance my technical knowledge, my ability as a scientist, and find a group of colleagues who were constantly encouraging me and became close friends.

There's also many other people in the graduate program who I would like to acknowledge. I would like to acknowledge Erick Armingol not only for his contributions as co-author in Chapter 3-5, but for becoming a lifelong friend whose scientific brilliance is only surpassed by his kindness. Erick, we had an amazing run here, and we got to work on some cool projects while also having a lot of fun. I'm excited to see the great heights your career will take you. I also want to thank some friends from my cohort who helped me a lot. Thanks Azhar Khandekar for putting up with my late night rants and for always showing up on the soccer field. Thanks Cameron Martino for imparting some of your wisdom-- which is way beyond your years—on me; I really think you've lived at least two lifetimes already. I also would like to thank Benjamin Kellman, Isaac Shamie, Curtis Kuo, and Helen Masson from the Lewis Lab, who helped me adjust when I was starting out and were always willing to go out of their way to help me. Beyond being amazing colleagues, you all became great friends. Finally, to Juan Tibocha-Bonilla and Gaby Canto, two of the sweetest people I've met and who Arjana and I both count as great friends. To all of you, I can't tell you

how much I appreciate the moments we shared together and the friendships we formed. I would also like to acknowledge my friends outside the program, most of whom have known me since I was quite young and have been a great source of support for me in these years: to Matthew Shintaku, Alex and Tori Sierra, Cristian Garcia, Andrew Mahoney, James Bradley, Kirill Slobodyanyuk, Charles Ayers, and Hemingway He, who give me an excellent reason to visit the rest of California.

Finally, I would like to acknowledge my family for their advice, support, and love during my time in graduate school. To my parents, Alex and Joyce Baghdassarian, without whom I would would not have had the courage to start a Ph.D. Thank you for always pushing me to be my best, supporting me to pursue my passions, and for your endless advice and unconditional love. To my twin sister, Karine Baghdassarian, who inspired me to pursue this: you have the sweetest heart, and live life to the fullest. To Arash Zadeh and Nelli Petikyan, who sometimes seem to know me better than I know myself. You've been there for the toughest times and for the best ones. Also, you've had a substantial impact on how I approach life, and I think the ski getaways were some of the most fun times I had during graduate school. Finally, to Arjana Begzati, who has loved and supported me day-in and day-out, celebrating my successes and facing my challenges alongside me. I can't imagine my life without you, and I certainly would not have gotten through graduate school without you. It's amazing to have found someone who makes everything more enjoyable and who I can trust with anything: work-related matters, travel, to simply go out and enjoy life, or to relax at home, content in each others' company.

Chapter 1 has been submitted for publication as a review article. The dissertation author was the primary author. Hratch Baghdassarian and Nathan Lewis conceptualized the work, carefully reviewed, discussed and edited the paper. Hratch Baghdassarian wrote the paper and generated the figures.

Chapter 2, in part, is a reprint of the material as it appears in The New England Journal of Medicine: "Variant STAT4 and Response to Ruxolitinib in an Autoinflammatory Syndrome."

NEJM. May 2023. https://doi.org/10.1056/NEJMoa2202318. Reproduced with permission from The New England Journal of Medicine, Copyright Massachusetts Medical Society. Lori Broderick, Hratch Baghdassarian, Sarah A. Blackstone, Rachael Philips, Hirotsugu Oda, John J. O'Shea, Joshua D. Milner, Daniel L. Kastner and Christopher D. Putnam conceived the overall experimental designs. Lori Broderick, Hratch Baghdassarian, Nathan Lewis, Vivian K. Hua, David R. Murdock, Nobuyuki Horita, Shimul Chowdhury, David Dimmock, Sofia Rosenzweig, Brynja Matthiasardottir, Elaine Remmers, Adam Mark, Roman Sasik, Kathleen M. Fisch, and Kristen Jepsen performed sequencing and analysis for genomic and single cell RNA sequencing studies. Brynja Matthiasardottir and Rachael Philips performed bulk RNA sequencing and analyses. Christopher D. Putnam, Elif Erin, Norman R. Watts and Hirotsugu Oda performed structural modeling and analyses. Lori Broderick, Owen Clay, Rachael McVicar and Yang Liu initiated and maintained primary fibroblast lines. Lori Broderick, Vivian K. Hua, and Owen Clay performed and analyzed wound healing assays. Lori Broderick, Sarah A. Blackstone, and Pallavi Pimpale Chavan performed and analyzed studies involving in vitro expression, plasmid construction, flow cytometry, and ELISA. Lori Broderick, Sarah A. Blackstone, and Davide Randazzo performed immunofluorescence staining and microscopy. Suzanne Tucker performed clinical histology and immunohistochemistry. Chi Ma, Massimo Gadina, Daniella M. Schwartz and Joshua Milner performed and analyzed ex vivo flow cytometry. Lori Broderick, Suzanne M. Tucker, Natalie Deuitch, Michele Nehrebecky, Anwesha Sanyal, Giffin Werner, Raphaela Goldbach-Mansky, William A. Gahl, Johanna Chang, and Kathryn Torok were responsible for clinical samples. Lori Broderick, Hratch Baghdassarian, Johanna Chang, Daniel L. Kastner and Christopher D. Putnam wrote the initial draft of the manuscript with contributions from Sarah A. Blackstone, Brynja Matthiasardottir, David Dimmock and Kathryn Torok. All other authors reviewed the final draft and agreed to publish the manuscript.

Chapter 3, in part, is a reprint of the materials as they appear in Nature Communications: "Context-aware deconvolution of cell-cell communication with Tensor-cell2cell". Nat. Commun.

June 2022. https://doi.org/10.1038/s41467-022-31369-2. Erick Armingol, Hratch Baghdassarian, and Nathan Lewis conceived the work. Cameron Martino contributed important insights for creating Tensor-cell2cell. Erick Armingol implemented Tensor-cell2cell and performed the analyses on the datasets of COVID-19 and ASD. Hratch Baghdassarian designed and created the simulated 4D-communication tensor and performed the analyses on the simulated data. Erick Armingol, Hratch Baghdassarian, and Cameron Martino performed benchmarking and statistical analyses. Cameron Martino trained classifiers and compared Tensor-cell2cell to CellChat. Hratch Baghdassarian performed benchmarking analyses using different external CCC tools. Erick Armingol performed benchmarking analyses using different preprocessing and batch-correction methods. Erick Armingol and Hratch Baghdassarian developed downstream analyses. Araceli Perez-Lopez helped to interpret the COVID-19 results and researched literature. Caitlin Aamodt helped to interpret the ASD study case and researched literature. Rob Knight contributed to the benchmarking analyses. Erick Armingol and Hratch Baghdassarian wrote the paper and all authors carefully reviewed, discussed and edited the paper.

Chapter 4 materials, in part, have been submitted for publication as they appear in: "Combining LIANA and Tensor-cell2cell to decipher cell-cell communication across multiple samples." https://doi.org/10.1101/2023.04.28.53873. 1. Hratch Baghdassarian, Daniel Dimitrov, and Erick Armingol conceived the project, adapted the computational tools, developed the protocol, and wrote the initial version of the manuscript. Julio Saez-Rodriguez and Nathan Lewis revised the manuscript and supervised the project. Hratch Baghdassarian, Daniel Dimitrov, and Erick Armingol contributed equally. Julio Saez-Rodriguez and Nathan Lewis are both corresponding authors and have contributed equally.

Chapter 5, in part, is currently being prepared for submission for publication of the material. Hratch Baghdassarian and Nathan Lewis conceived the project. Hratch Baghdassarian implemented the software for constructing and analyzing the model. Juan Tibocha-Bonilla and Erick Armingol provided conceptual input on achieving model feasibility. Laurence Yang, Juan

Tibocha-Bonilla, and Nathan Lewis provided conceptual input on formulating the coupling constraints. Erick Armingol implemented code to parse model parameters. Nathan Lewis supervised the project. All authors provided conceptual input on downstream analyses.

VITA

2015-2016	Research Assistant, Lawrence Berkeley National Laboratory
2017	Bachelor of Arts, University of California Berkeley
2017	Bachelor of Science, University of California Berkeley
2017-2018	Research Assistant, University of California San Francisco
2023	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

1. H. Baghdassarian*, S. Blackstone*, O. Clay, et al . "Variant STAT4 and Response to Ruxolitinib in an Autoinflammatory Syndrome." NEJM. (May 2023). https://doi.org/10.1056/NEJMoa2202318.
2. H. Baghdassarian, N. Lewis. "Resource Allocation In Mammalian Systems". In Submission, Journal of Biomedical Science.
3. H. Baghdassarian*, D. Dimitrov*, E. Armingol*, J. Saez-Rodriguez, N. Lewis. "Combining LIANA and Tensor-cell2cell to decipher cell-cell communication across multiple samples." Under Review, Nature Protocols. https://doi.org/10.1101/2023.04.28.538731.
4. E. Armingol*, H. Baghdassarian*, C. Martino, A. Perez-Lopez, C. Aamodt, R. Knight, N. Lewis. "Context-aware deconvolution of cell-cell communication with Tensor-cell2cell". Nat. Commun. (June 2022). https://doi.org/10.1038/s41467-022-31369-2.
5. A. Chiang*, H. Baghdassarian*, B. Kellman, B. Bao, J. Sorrentino, C. Liang, C. Kuo, H. Masson, N. Lewis. "Systems glycobiology for discovering drug targets, biomarkers, and rational designs for glyco-immunotherapy". Journal of Biomedical Science. (June 2021). https://doi.org/10.1186/s12929-021-00746-2.
6. B. Kellman*, H. Baghdassarian*, T. Pramparo, I. Shamie, V. Gazestani, A. Begzati, S. Li, S. Nalabolu, S. Murray, L. Lopez, K. Pierce, E. Courchesne, N. Lewis. “Multiple Freeze-Thaw Cycles Lead to a Loss of Consistency in poly(A)-Enriched RNA Sequencing.” BMC Genomics. (Jan. 2021). https://doi.org/10.1186/s12864-021-07381-z.
7. E. Armingol. H. Baghdassarian, N. Lewis. “Next-generation tools for studying cell–cell interactions and communication.” In Submission, Nat. Rev. Genet.
8. E. Armingol, R. Larsen, M. Cequeira, H. Baghdassarian, N. Lewis. “Unraveling the coordinated dynamics of protein- and metabolite-mediated cell-cell communication.” In Preparation. https://doi.org/10.1101/2022.11.02.514917.
9. E. Armingol, A. Ghaddar, C. Joshi, H. Baghdassarian, I. Shamie, J. Chan, H. Her, S. Berhanu, A. Dar, F. Rodriguez-Armstrong, O. Yang, E. O’Rourke, N. Lewis. “Inferring a spatial code of cell-cell interactions and communication across a whole animal body.” PLOS Comp. Bio. (Nov. 2022) . https://doi.org/10.1371/journal.pcbi.1010715.
10. C. Kuo, A. Chiang, H. Baghdassarian, N. Lewis. "Dysregulation of the secretory pathway connects Alzheimer's disease genetics to aggregate formation". Cell Systems. (June 2021). https://doi.org/10.1016/j.cels.2021.06.001.
11. H. Pinkard, H. Baghdassarian, A. Mujal, E. Roberts, K. Hu, D. Friedman, I. Malenica, T. Shagan, A. Fries, K. Corbin, M. Krummel*, L. Waller*. “Learned adaptive

multiphoton illumination microscopy for large-scale immune response imaging”. Nat Commun. (March 2021). https://doi.org/10.1038/s41467-021-22246-5.

12. S. Palluk*, D. Arlow*, T. Rond, R. Bector, J. Kang, H. Baghdassarian, A. Truong, P. Kim, A.Singh, N.Hillson, J.Keasling. "De novo DNA synthesis using polymerase-tethered nucleotides." Nat Biotechnol. (June 2018). https://doi.org/10.1038/nbt.4173.

*contributed equally to work

ABSTRACT OF THE DISSERTATION

Understanding How Context Influences Function Across Biological Scales in Multicellular Mammalian Systems

by

Hratch Matthew Baghdassarian

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Nathan E. Lewis, Chair
Professor Terence Hwa, Co-Chair

In mammalian systems, no cell acts in isolation, but rather coordinates to achieve higher-order function. Such cell behaviors are complex and influenced by context. To comprehensively understand them, we must understand how molecular interactions affect cell phenotypes, and analogously, how cell interactions affect higher-order phenotypes.

I begin by examining the role of resource allocation in cellular decision-making processes. I underscore the significance of resource constraints and context in shaping cellular phenotypes and enabling population-level behaviors. My research then pivots to a detailed investigation into a rare systemic inflammatory disorder, Disabling Pansclerotic Morphea (DPM). I report on the discovery of novel variants in the STAT4 gene that are linked to DPM. Leveraging these insights, we propose a successful therapeutic approach using the JAK inhibitor, ruxolitinib, thereby demonstrating the importance of a context-informed genetic understanding in disease management.

The JAK-STAT pathway is a key signaling pathway mediating immune cell communication. Thus, I shift the focus of my research to intercellular communication. My novel unsupervised method, Tensor-cell2cell, deciphers complex cell-cell communication patterns across multiple contexts (e.g., time points, disease severities, or spatial contexts). Given the generalizability of this approach to use other communication methods' outputs as its input, I then introduce a protocol integrating two computational tools, LIANA and Tensor-cell2cell. LIANA is similarly generalizable in that it provides a centralized resource to run many methods, thus providing a natural preceding step to Tensor-cell2cell. The protocol enhances robustness and flexibility in identifying cell-cell communication programs across multiple samples. Finally, I present humanME, a computational tool for generating and analyzing human ME-Models from input metabolic models. This approach refines the prediction accuracy of growth rate and offers unique solutions, highlighting the importance of machinery resources in constraining intracellular activities.

Collectively, this body of work leverages omics to provide mechanistic insights to how cellular context impacts interactions and functions in mammalian systems across molecular-, cell-, and tissue-scales. The new methods and tools proposed herein pave the way for more nuanced, context-driven research, underpinning future advancements in human health and disease.

Chapter 1: Resource Allocation in Mammalian Systems

Each function of a cell has resource costs and fitness benefits. Cell decisions manage resource allocation to optimize these functions. In fact, the management of resource constraints (e.g., nutrient availability, bioenergetic capacity, and macromolecular machinery production) shape activity and ultimately impact phenotype. In mammalian systems, the quantification of resource allocation provides important insights into higher-order multicellular functions; it shapes intercellular interactions and relays environmental cues for tissues to coordinate individual cells to overcome resource constraints and achieve population-level behavior. Furthermore, these constraints, objectives, and phenotypes are context-dependent, with cells adapting their behavior according to their microenvironment, resulting in distinct steady-states. This review will highlight the biological insights gained from probing resource allocation in mammalian cells and tissues.

1.1 Introduction

Resource allocation governs economies and biology alike. Each biological function has an associated **resource cost** while also conferring a **fitness** benefit by contributing to a cellular **objective**, such as growth (Appendix A). Cells **optimize** for these objectives under the constraints of their **resource budget**. The accumulation of resource allocation decisions to fulfill cell objectives results in observed cell phenotypes. As such, resource constraints limit cells' activity and, consequently, their range of possible phenotypes[1]. In this sense, resource allocation can be viewed as a cost-benefit[2] or supply-demand[3] analysis (Fig. 1.1). Despite the complexity of mammalian cells, resource allocation is a fundamental principle underlying decision-making. From an evolutionary perspective, organisms that best apply resource allocation strategies will have higher fitness. As such, the consideration of resource allocation illuminates *how* and *why* cells respond to their environment. Specifically, resource costs limit *how* a cell can achieve its objective by constraining the possible mechanisms the cell can use. Fitnessilluminates *why* the cell chooses one specific mechanism over other possibilities. Understanding these choices is highly informative considering the degeneracy encoded within biological networks[4].

Decision-making depends strongly on **cellular context**[5]. To make decisions, a cell perceives extracellular cues[6] such as nutrient availability and communication signals[7], and processes this information based on its intracellular state (e.g., cell type, genomic variants, epigenetic state). Consequently, the extracellular cues provide the cell with its resource budget and shape its objectives. Intracellularly, resource availability and objectives in a given context determine pathway activity[3]. Finally, context can change with time[8–10], space[11,12], and disease[13,14], introducing new objectives that cause trade-offs and transitional costs that further constrain the cell.

The cellular context includes other cells. Mammalian cells do not act in isolation, but rather in multicellular systems to achieve higher-order functions[15–17]. Constraints, contexts, and phenotypes are ubiquitous across biological scales. Thus, the insights into resource allocation may be generalized to tissues and even the whole-organism (Appendix C). With multicellularity, cells become specialized to limit the burden of trade-offs. Additionally, decision-making accounts for coordination and competition from other cells, leading to synergistic effects[18].

In this Review, we discuss how resource allocation impacts mammalian cell decisions and multicellularity. We begin with two questions at the cellular scale (Fig. 1.2a):

(1) How do metabolic resources (nutrients, machinery, and bioenergetics) constrain the cell?

(2) How do cells allocate resources to coordinate activity across molecular processes?

Building on these concepts to understand multicellularity (Fig. 1.2b), we ask:

(3) How do trade-offs imposed by resource constraints affect cellular decision-making, leading to cell specialization?

(4) How do specialized cells with distinct tasks coordinate within multicellular systems to achieve higher-order functions?

Here we highlight the role of resource allocation, which is one valuable concept among many to understand biological mechanisms. Our aim is to demonstrate the broad utility and unique insights provided by resource allocation across various areas of biology. Resource allocation provides a unique perspective to uncover fundamental principles that can be applicable across diverse systems. We structure our discussion by first introducing an overarching principle and subsequently illustrating them with wide-ranging examples from various systems in both health and disease. While bioenergetic and metabolic optimality are not always the driving forces or not yet fleshed out in mammalian systems, such a perspective has been invaluable to the study of prokaryotes and lower eukaryotes in support of mammalian systems, which we also highlight here.

Throughout these discussions, we explore the plasticity of resource allocation as it changes across contexts. We also briefly highlight powerful systems biology methods that now help address such questions (Appendix B, Table 1.1). Quantifying and modeling resource allocation provides insights into how cells regulate gene expression, intracellular pathway activity, cell-cell interactions, and ultimately phenotypes. Associated technological and computational innovations are providing high-throughput measurements and analysis tools to decipher how resource allocation, as a governing design principle, shapes the complex processes underlying mammalian phenotypes.

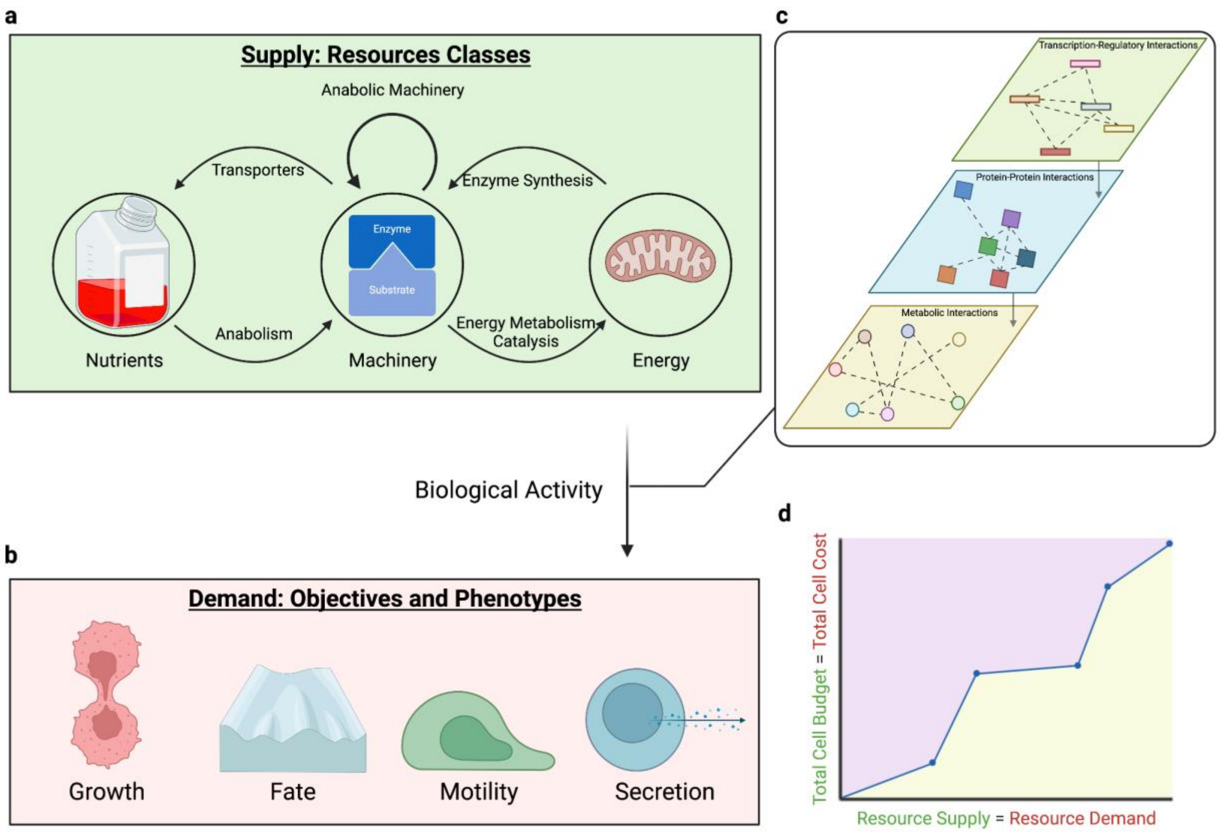


Figure 1.1: Generation and use of a balanced cellular resource budget.
(a) The total cellular budget is determined by the supply of nutrients, machinery, and energy resources. There is extensive crosstalk between these three resource classes, each depending on the others to generate its supply. **(b)** Once the budget is generated, it is used to conduct diverse biological activities that contribute to the cell objectives and result in observed phenotypes. The amount of resources used to meet these functional demands represents the cost of executing biological function. **(c)** Intracellular processes (e.g., transcriptional activity, protein activity, and metabolism) interact to drive the activities which connect resources to objectives. **(d)** In an efficiently allocated system, the cellular budget generated will equal the resource demanded (blue line). Generating an excess budget (purple shaded region) results in higher resource costs than necessary, whereas not generating enough budget (yellow shaded region) means that not all the demands can be met. Overall, the total cellular budget constrains biological activity, and in turn, observable phenotypes.

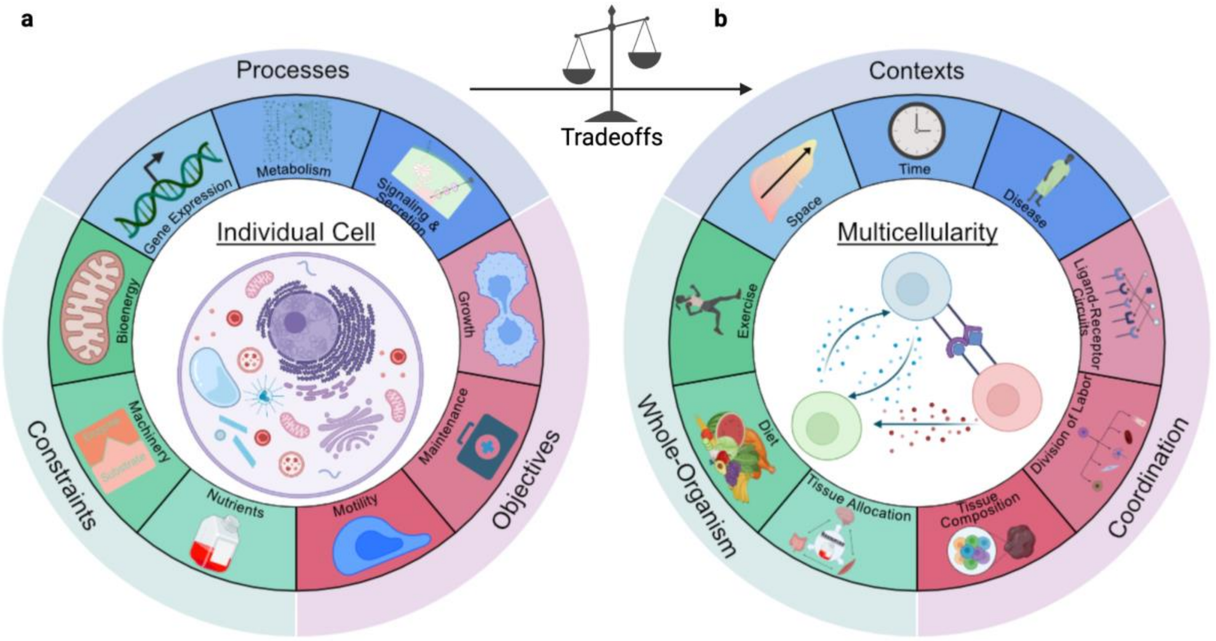


Figure 1.2: Key factors determining mammalian resource allocation strategies.
(a) The individual cell must account for its resource constraints when trying to optimize various objectives. Mechanistically, constraints are tied to objectives by intracellular processes that determine the global activities of the cell. **(b)** Constraints lead to trade-offs between multiple objectives, resulting in the evolution of multicellular organisms for more efficient allocation strategies. Coordination between cells enables specialization, communication, and tissue-scale steady-states that function robustly across multiple contexts. These concepts extend to the whole-organism scale, demonstrating that resource allocation can provide broad insights into biological function.

1.2 Cellular Resources Constrain Phenotype

The availability of nutrients, **machinery**, and bioenergy define the cellular resource budget, and consumption of these resources defines the resource costs of biological activity (Fig. 1.1). How does the cell manage its resource budget and how do these constraints affect phenotype?

1.2.1 Nutrients: resources informing allocation

Supplies and signals. Extracellular nutrients contribute to the total resource budget as substrates for machinery synthesis and bioenergetic pathways (Appendix A, Appendix D). For example, amino acids can be incorporated as building blocks for proteins, whereas glucose and glutamine can be catabolized for energy or nucleotide synthesis (e.g., for oligonucleotides)[19,20]. However, in managing resource allocation, nutrients play a particularly important role as extracellular cues. Their presence informs the cells of which metabolic pathways may be utilized and ultimately which objectives may be achieved, guiding allocation across the metabolic network. In this sense, through regulatory programs[21], nutrients will induce activation of specific metabolic pathways and express the associated machinery that catalyze those pathways.

Signals of scarcity. Nutrient allocation has evolved to cope with nutrient scarcity[21]. Mammalian cells have many strategies to handle nutrient scarcity; the global metabolic network is robust to nutrient inputs, capable of utilizing distinct nutrient-pathway combinations to meet cellular demands[4,22,23]. For example, cells will shift glucose usage from energy metabolism to *de novo* serine synthesis when serine is scarce[24] and use fatty acid oxidation to generate energy when glucose is scarce[25]. Additionally, cells may degrade intracellular components through autophagy and divert the resultant substrates to the most necessary pathways[21]. Finally, cells may scavenge for more complex extracellular resources, such as proteins and lysophospholipids, to catabolize through mechanisms such as macropinocytosis[26]. These resources are often produced by other cells in a multicellular system.

Multicellularity decreases the likelihood of individual cells facing nutrient scarcity. This is because mammalian cells have multiple subcompartments and specialized cell types that improve nutrient storage and delivery systems to create nutrient-rich microenvironments [27,28]. Storage is a multicellular example of the **hedging** strategies discussed later, diverting resources away from current objectives in anticipation of future context-dependent fluctuations in nutrient levels[29].

Allocation in abundance. Resource allocation in nutrient-rich environments becomes a decision-making problem. Cell activity still comes at a resource cost. Thus, the cell must choose which nutrients to shuttle to which metabolic pathways to most efficiently achieve its objective

(Appendix D). Sometimes, it will even disregard available extracellular nutrients, such as non-essential amino acids, choosing instead to produce them intracellularly[28].

Since cell activity evolved under nutrient scarcity, mammalian cells must tightly control nutrient uptake to prevent irregular phenotypes such as uncontrolled growth (see **Appendix C** for consequences of overabundance). Unlike prokaryotes, which typically uptake nutrients upon sensing them, mammalian cells tend to couple nutrient sensors with signaling proteins such as growth factors for an additional layer of regulation (Fig. 1.3c)[28,30]. This additional regulatory layer serves a dual purpose of allowing cells and tissues across the organism to coordinate via combined nutrient- and signaling protein- circuits, maintaining steady-state circulating nutrient levels and mobilizing nutrient stores when necessary[31,32]. Thus, nutrient scarcity tends to be a local and context-specific constraint in multicellular systems, such as in wound-healing and poorly vascularized regions[26]. The cell's microenvironment (e.g., nutrient concentration, interactions with other cell types, and presence of other extracellular nutrients) affect metabolic activity. Thus, it is important to appropriately account for a cell's microenvironment when studying resource allocation, which may not always be accurately represented *in vitro*. For example, *in vivo* early-activated CD8+ T-cells utilize glucose differently than those under the super-physiological conditions of cell culture[33], shifting flux from aerobic glycolysis to oxidative phosphorylation to create a larger bioenergy budget.

Distribution drives function. Ultimately, nutrient availability and allocation affect cell phenotypes, such as growth[26,34,35] and secretion[36], as well as organismal phenotypes, such as development[37] and immunity. In immunity, nutrient availability affects cell fate, function, and composition. For example, under both amino acid and glucose deprivation, mTORC1 activation and CD4+ T-regulatory cell proliferation decreases[38]. Additionally, effector T-cells rely on both glucose[39,40] and glutamine[41] for cytokine secretion whereas invariant Natural Killer T cell cytotoxicity is independent of these nutrients[42]. In early development, glucose is shuttled to anabolic pathways that trigger synthesis of key machinery controlling blastocyst formation, with cells using other carbon sources such as pyruvate and lactate to generate bioenergy. In the absence of glucose prior to compaction, zygotes do not invest biosynthetic resources into expressing the glucose transporter SLC2A3, resulting in reduced cell growth and degenerated morulae[43]. Mechanistically, SLC2A3 expression is attributable to glucose being metabolized via the hexosamine biosynthetic pathway (HBP). Coordinated utilization of glucose by the HBP and the pentose phosphate pathway is also required for synthesis of key machinery involved in trophectoderm fate-specification[44].

We note that there is an interplay between the three resource classes: nutrients, machinery, and bioenergetics. So far, this is largely presented as nutrients used to synthesize the cell's machinery, and machinery and metabolic substrates together dictating the bioenergetic pathways that the cell utilizes. However, this crosstalk is not unidirectional, with each resource class depending on the others (Fig. 1.1a). Building machinery requires energy[45], and without the right machinery, nutrients can't be used effectively. Transporters, for example, are machinery that deliver extracellular nutrients to the cell. In CD8+ T-cells, knocking-out amino acid transporters limits nutrient intake, altering the ratio of terminal effector to memory precursor subpopulations[46].

1.2.2 Machinery: resources actuating allocation

As actuators of biological function, the machinery budget contributes to pathway activity and cell phenotype. The machinery component of the resource budget depends on two factors: machinery activity and machinery abundance (Appendix F). Higher activity increases the amount of substrate a single unit of machinery is able to convert into a product. There are innumerable examples in which altered catalytic efficiency, e.g. via point mutations, has disrupted homeostasis and led to disease states (reviewed here[47]). On the other hand, higher abundance increases the proportion of machinery to its substrate. This is evident for individual machinery. For example, CARKL expression levels change depending on activation signals to alter energy metabolism and ultimately dictate macrophage fate.[48] Similarly, pyruvate kinase isoforms interact based on expression levels to affect nucleotide metabolism and dictate proliferation[49]. Abundance is also coordinated across multiple components. CRISPR screens have tested thousands of genes and found that machinery group by shared effects across multiple phenotypes;[50] meanwhile, the abundance of hundreds of secretory pathway machinery together coordinate a cell's capacity for secreting specific proteins[51]. Unlike the cell's nutrient budget, which is largely limited by extracellular availability, machinery abundance is constrained by the resource costs of synthesis. Beyond diverting nutrients to anabolism, cells must invest bioenergy and biosynthetic machinery for gene expression.

Machinery Pose Non-negligible Auto-catalytic Costs. The machinery facilitating anabolism and gene expression are auto-catalytic, meaning they contribute to their own synthesis. As a result, machinery synthesis costs include the use of biosynthetic machinery[52]. For example, in eukaryotes, individual pre-mRNAs compete for the shared pool of splicing machinery to be properly processed[53]. Since ribosomal genes represent a substantial mass fraction of the proteome, global splicing efficiency is associated with ribosomal expression. Given that ribosome expression increases with growth rate to accommodate increasing biomass production demand, and that nutrient signaling via TOR influences ribosome gene expression, these observations couple cell growth with nutrient and machinery constraints of macromolecular synthesis (Appendix A). In fact, under nutrient scarcity, reducing splicing efficiency through intron-mediated regulation improves cell survival by decreasing ribosomal expression[54].

Jones *et al.* generalize this concept of **resource loading** by modeling gene expression to study the limiting biosynthetic machinery[55]. Typically, costs of translational and ribosome biogenesis are seen as the main constraints on gene synthesis. For instance, rough estimates for HeLa cells show that all ribosomes must be constantly active to maintain global protein levels[56]. However, Jones *et al.* find that, in engineered circuits in human cell lines, transcriptional resources limit protein expression levels of target genes. Similarly, one study concluded that the relative impact of transcriptional and translational costs change across nutrient-limiting conditions, pointing to the crosstalk between machinery and nutrients[57].

Machinery Biosynthesis has Substantial Energetic Costs. The energetic costs of machinery synthesis can be quantified across genomic, transcriptional, and translational processes as a function of gene features. Protein synthesis costs are highly sensitive to the energy budget[58] and are a large resource burden: ribosomal translation alone represents 30% (BioNumbers[59] ID 110441) of a mammalian cell's energy budget. Also, net protein synthesis costs represent more than 70% (BioNumbers[59] ID 111918) of a generic cell's energy budget (due to the higher contribution of amino acid synthesis relative to polymerization[45]). Lane and Martin argue that protein synthesis costs represent such a substantial portion of the total cellular energy budget that they prevent the evolution of eukaryotic genome complexity withoutmitochondria[60]. In contrast, Lynch and Marinov contend that, considering the cell's total lifetime, the energy budget scales with cell volume to more than compensate for increased costs of genome complexity, regardless of the presence of mitochondria[45].

Strategies to Minimize Machinery Costs. Due to these costs, cells employ various strategies to efficiently generate the machinery budget. For example, protein degradation is costly (Appendix A, Appendix D) because proteasomal degradation consumes ATP[61] and sequesters proteases. From a resource allocation perspective, this aligns with the fact that short-lived proteins have low abundance[62,63,64], representing just 5%[63] of the human proteome. By only dedicating degradation resources to lowly expressed proteins, the cell reduces the cost of tuning protein abundance in modes beyond translational control. In contrast to proteasomal degradation, protein turnover via autophagy recovers more energy and nutrient resources than it consumes and represents a large fraction of total protein degradation in mammalian cells[65,66].

Short-lived proteins also constitute only a small portion of complexes[63], supporting the notion that cells conserve protein degradation resources by minimizing their use in more abundant proteins. An alternative strategy proposed to reduce costs in complexes is proportional synthesis. Observed in both prokaryotes[67] and eukaryotes[68], cells express complex subunits in proportion to their stoichiometries by tuning synthesis rates, avoiding excess resource expenditure on synthesis. Additionally, degradation would be required to achieve proper stoichiometries and clear misfolded proteins resulting from excess expression. Furthermore, there are spatio-temporal variations in complexes' stoichiometries, many of which are regulated beyond transcription[69]. Together with proportional synthesis, this suggests context-specific tuning of translation rates to minimize synthesis costs.

Cells also employ strategies beyond the direct tuning of machinery abundance to minimize machinery costs. For example, a metabolic modeling approach[70] showed that coenzyme redundancy (e.g., distinct pools of NAD and NADP) is not necessary for growth but reduces the minimal abundance of protein required to catalyze coenzyme coupled reactions[71]. These various strategies demonstrate that cells work to optimize machinery expression levels by minimizing synthesis costs while maximizing the associated fitness benefits[2]. This cost-benefit trade-off in gene expression was characterized in a high-throughput manner in eukaryotes, testing the impact of 81 genes' expression levels on growth rate. Crucially, 83% of genes demonstrated distinct fitness curves correlating with differential gene expression[72]. Thus, minimizing machinery costs has proven to be an accurate optimality principle to predict and understand metabolic activity[73,74] (Appendix B), such as in relationships between energy metabolism and growth (Appendix A).

1.2.3 Bioenergy: resources fueling allocation

Energy metabolism (e.g., glycolysis and oxidative phosphorylation) uses machinery to transfer the energy stored in extracellular nutrients to its main intermediate currency, ATP, and long-term nutrient stores. This energy budget is spent to fuel a multitude of tasks[22,75,76] that prompts the cell to organize its activities accordingly. Several methods identify the energetic cost of executing function at varying resolutions, broadly by measuring or estimating the metabolic flux associated with bioenergy generation or consumption (reviewed here[10,77]); these include estimates from measures such as the oxygen consumption rate and ATP equivalents, i.e. mechanistic delineation of the number of ATP molecules consumed, each of which yield ~50 kJ/mol (BioNumbers[59] ID 100775, 100776) of energy from hydrolysis.

As discussed previously, a large fraction of the energy budget is spent on gene expression and protein translation[45]. The remainder of the energy budget is available for other tasks. For example, migrating cells consume energy to displace the extracellular matrix[78,79]. The energetic cost of motility changes with physical features of the matrix such as stiffness and spatial confinement. As such, cells minimize these migratory costs by choosing context-specific migratory mechanisms[80] and migrating through paths that require lower energy expenditure[81]. Neurons also demonstrate energy allocation. They consume energy for functions distributed across maintenance and activity, including neurotransmitter uptake and release by synaptic vesicles and action potential generation[82–84]. Consequently, the brain has a high rate of energetic expenditure[85], accounting for 20% of the total energy costs of the human body (BioNumbers[59] ID: 103264, 110878), and must use its energy budget efficiently. At the cell-scale, neurons employ combinatorial strategies to express sets of ion channels with specific kinetics that minimize energetic costs while meeting functional requirements (e.g., spiking rate)[86,87]. At the tissue-scale, energetic constraints limit the total number[88] and active fraction[89] of neurons in the brain. As such, the brain analogously employs combinatorial strategies to efficiently encode information into sets of neurons in such a manner that limits the number of active neurons, and thus energy demands[90,82].

While whole-cell modeling has demonstrated that the total energy budget is nearly equivalent to the total energy cost of a synthetic minimal cell[91], it has also demonstrated excess production of energetic intermediates in *Mycoplasma genitalium*[92]. Whole-cell modeling approaches (Appendix B, Table 1.1) are yet to be translated to mammalian cells, but such insights could prove invaluable to understanding energy resource allocation. An excess budget indicates either a deprioritization of evolutionary optimality for the energy budget over other objectives, cell hedging for future energy demands, or incomplete accounting of energy consumption.

1.3 Molecular Processes Coordinate Resources

Mechanistically, resource constraints are tied to the activity of molecular processes such as gene synthesis, metabolism, and molecular communication. These processes can be coupled to each other using systems approaches (Appendix B, Table 1.1), providing quantitative details regarding how resources are used. Resource constraints that affect the activity of one process propagate to others and eventually affect phenotype (Fig. 1.1c).

1.3.1 Gene Expression: Coordinating mRNA with Protein

To understand the global relationship between mRNA and protein expression levels, proteomics is often compared with transcriptomics using correlative metrics[93,94]. Comparing per cell absolute protein copy numbers to transcripts-per-million mRNA levels in multiple human tissue cell lines identified an average Pearson correlation of 0.6[95]. Importantly, a gene-specific, tissue-independent protein-to-RNA (PTR)[95,96] ratio increases the correlation. This indicates a gene-specific effect on resource loading, wherein intrinsic features[64,97] yield a competitive advantage for shared translational resources. For example, sequence length is a positive indicator of the variation in the PTR ratio, consistent with longer genes having a larger resource cost. Sequence length is one of many intrinsic features relating mRNA to protein abundance[64,97]. Furthermore, the PTR ratio increases with transcriptional abundance[98,99], signifying cooperation between transcription and translation in generating highly abundant proteins. Gene-specific competitive advantages also coordinate across processes within transcription. For example, mammalian cells use “economies of scale”, wherein splicing efficiency increases with transcription rate to prioritize production of strongly transcribed genes[100].

The role of gene-intrinsic features makes it apparent that abundance alone does not sufficiently explain mRNA to protein coupling. It follows that such features inform gene expression rates; broadly, there are four rates to consider: transcription, translation, mRNA decay, and protein degradation. The gene-specific combinations of these rates give rise to global steady-state levels of mRNA and protein in a context-dependent manner, especially with dynamic responses requiring rapid adaptation[101,102]. Resource allocation limits the existing rate combinations, with certain combinations resulting in synthesis costs that outweigh the gene's contribution to the cell objective. Specifically, there is an evolutionary lack of genes with high transcription rates and low translation rates, despite these rates being mechanistically feasible. This is driven by a "precision-economy trade-off" between stochastic protein abundance (precision) and resource costs of synthesis (economy), with high transcription and low translation not providing an advantage in either[103].

The four synthesis rates are commonly modeled using a simplified set of ordinary differential equations (Table 1.1). Derivation of these rates requires absolute quantification–the number of molecules per cell[104]. Notably, these rates have a larger contribution to changes in the absolute level of a protein than to its relative level[102]. This makes sense from a resource allocation perspective since there is high inequality in the distribution of absolute protein abundances, with a small number of proteins constituting the majority of the total proteome by mass and copy number[63,98,105,106]. Thus, highly abundant proteins will sequester a disproportionate fraction of cellular resources, independent of the fact that they tend to have an intrinsic competitive advantage in using cellular resources.

Translational efficiency also impacts gene-specific PTRs[107]. Gene-specific sequences regulate translational efficiency to prevent ribosomal jamming and excess translational machinery costs[108]. Genes with high translational efficiency also have high mRNA abundance, enabling them to outcompete their lower efficiency counterparts for ribosomes by both concentration and affinity. Furthermore, these genes are functionally enriched for macromolecular synthesis and energy metabolism[109], indicating the cell is optimizing for the production of its machinery and bioenergy resource budgets. From a resource loading perspective, ribosome machinery saturation imposes an upper bound on the dynamic range of protein abundance (e.g., translation rates plateau at ~1000 protein molecules per mRNA molecule per hour in NIH3T3 mouse fibroblasts).

1.3.2 Actuation: Coordinating Protein with Metabolism

Metabolism links gene expression to cell phenotype[3,110]. Metabolic inputs, outputs, and intermediates are incorporated into all the resource classes and molecular processes discussed, forming the basis upon which cell activity occurs[92]. There are innumerable examples of how metabolic activity affects cell phenotype. For example, shifts from purine to serine synthesis promote cell motility[111,112], mitochondrial metabolism modulates stem cell fate[113–115], and many pathways affect growth[26].

Gene expression produces the machinery that catalyzes metabolism. As mentioned previously, machinery activity and abundance together determine the flux that an enzyme-catalyzed biochemical reaction can carry (Appendix F)[116,117]. Assuming Michaelis-Menten kinetics, the maximum flux that a reaction can carry is the product of the catalytic rate constant and the enzyme concentration. There are several considerations *in vivo* that may prevent a cell from realizing these maximum rates, which tend to be reported under *in vitro*, nonphysiological conditions. First, the maximum rate assumes that the enzyme is fully saturated. However, cells try to minimize intermediate metabolite concentrations for homeostatic maintenance and quick adaptation to new contexts[118]. Yet, enzymes are more efficient at high (saturating) substrate concentrations. Thus, efficient enzyme usage must be balanced against minimizing metabolic intermediates[119] and rapid substrate consumption. More generally, there is a trade-off between control over reaction fluxes and metabolic intermediate concentrations[3]. Second, reaction thermodynamics affects reaction kinetics via the mass-action ratio due to such variables as intracellular pH and metabolite concentrations[120]. The extent of backwards flux due to thermodynamics requires a higher machinery investment to maintain the same reaction rate (Appendix B)[74]. Finally, regulatory effects such as allostery and post-translational modifications can alter kinetics to alter reaction rates.

Thus, the observed reaction rate will change with variables such metabolite concentration in a context-dependent manner. Combining proteomic measurements of enzyme abundance with metabolic modeling estimates of fluxomics to estimate *in vivo* observed catalytic rate constants demonstrated that the maximum value identified across multiple growth conditions agrees with the *in vitro* catalytic rate constants. Such studies decompose discrepancies between experimental and theoretical values into the underlying saturating, thermodynamic, and regulatory factors (Appendix F)[73,121]. The extent to which a reaction is active, the "capacity utilization", can be defined as the ratio between this context-specific, observed reaction flux and the maximum reaction flux[117]. If this ratio is one, enzymes are being utilized at full capacity and activity is **machinery-limited**. However, if this ratio is less than one, one of the aforementioned factors is decreasing the reaction rate. If this is due to saturation effects, not all of the expressed enzyme is used[122] and activity is instead **nutrient-limited**. Unused, free enzymes may point to hedging for rapid adaptation to future conditions which require increased flux through those reactions (Appendix A).

1.3.3 Communication: Coordinating Signaling and Secretion with Gene Expression

Signaling and secretion link a cell's extracellular environment with its intracellular activity, regulating the higher-order functions of multicellular systems. While signaling pathways sense and respond to extracellular signals, the secretory pathway produces communicatory molecules to send such information. These two molecular processes are not only complementary conceptually, but also biologically, coordinating each others' activity[123,124]. Secreted proteins are the product of gene expression. Unlike machinery, these proteins do not directly contribute to biomass production or intracellular activity. Yet, human cells allocate a massive amount of resources to protein secretion: secreted proteins represent >25% of the proteome by mass[51], despite the fact that these proteins do not contribute to intracellular tasks such as biomass production. This speaks to the importance of secretory tasks such as cell-cell communication[125–127], cytotoxicity[128,129], and remodeling of the extracellular matrix[130] to multicellular systems. Signal transduction pathways and their downstream transcription regulatory networks use extracellular cues, i.e. nutrients and communicatory molecules, to induce gene expression. Specifically, a receptor will sense the extracellular cue, downstream machinery processes the information encoded by that signal, and transcription factors induce the synthesis of target genes. As such, signaling pathways enable context-specific cellular decision-making[123,131–134].

Signaling pathways have multiple possible objectives, including **signal amplification**, **sensing precision**[135], **information transfer**, **parameter robustness**, and **response time**[136]. However, the activities underlying these objectives are constrained by energetic and machinery resource costs that the cell minimizes[137–139]. For example, Goldbeter–Koshland push–pull network sensing systems prioritize sensing precision, with receptors, downstream signaling machinery, and energy, independently constraining the objective. For optimality, these three constraints evolved to be equally limiting[140]. The efficiency by which signaling pathways and downstream transcription regulatory networks achieve their objective depends on their network topology[141]. Consequently, evolution has converged on a small number of prevalent topologies, termed network motifs, that determine the dynamics and robustness of network input-output relationships. Since there are multiple possible signaling objectives, it is interesting to consider how the interplay between these various objectives may affect resource allocation. For example, only two experimentally observed network motifs demonstrate fold-change detection across a wide range of organisms because they are uniquely **Pareto optimal** for response time, sensing precision, and signal amplification[142].

Network motifs are building blocks for global network structures. Convex analysis[143] of signaling networks identifies a minimal set of pathways representative of the network state from its global topology. These "extreme pathways" reveals how cells divert metabolic and machinery resources across the network, encoding for **signaling crosstalk** and redundancy between pathway reactions to achieve robust input-output relationships[144,145]. This concept has been extended to integrate signaling and metabolic modules together in a dynamic manner that accounts for differences in reaction time-scales[146]. Reducing crosstalk trades-off with increasing sensing precision[147] due to the energetic costs of **signaling modularity**[148]. Thus, cells employ combinatorial strategies leveraging crosstalk to reduce resource costs. For example, by sharing the tumor necrosis factor (TNF) ligand, the nuclear factor kappa B (NF-κB) and c-Jun N-terminal protein kinase (JNK) pathways can increase information transfer relative to either pathway acting in isolation. In contrast to the signaling pathways of an individual cell, sharing a ligand across cells does not have an upper bound on information transfer[149], demonstrating the utility of multicellularity. In fact, when coordinating in multicellular systems (Fig. 1.3b), a number of strategies reduce the resource costs associated with synthesis and secretion of communicatory molecules. For example, a single shared ligand can achieve diverse population-level behaviors such as bistability, e.g. all-or-nothing responses, and bimodality, e.g. cell fate decisions, by combining autocrine and paracrine communication[150]. Also, promiscuous ligand-receptor combinations enable more robust activation of multiple cell types as compared to a one-to-one binding[151]. Overall, mammalian cells have developed a number of resource optimization strategies to efficiently communicate, lowering the barriers to multicellularity.

1.3 Trade-offs Occur Due to Multiple Objectives

Cells balance multiple objectives[152]. Allocating resources towards one objective induces a trade-off because the shared and limited pool of resources must be diverted away from the pathways that enable a competing objective. A simple case is in the expression of two different genes within a cell: under a fixed resource budget, due to resource loading of biosynthetic machinery, increased expression of one gene is achieved by decreasing expression of the other[153]. Additionally, at the tissue-scale, different systems in the brain are each specialized for one or more tasks and trade-off against each other[154]. These tissue-scale tasks are coordinated within the multicellular system[18] to appropriately distribute resources and enable context-specific task prioritization.

Under resource optimality assumptions, such trade-offs are mathematically represented by a Pareto front (Fig. 1.3a), or the set of all optima for which performance of one objective cannot be improved without decreasing performance of another objective. Pareto analysis reduces the resource-constrained phenotype space[155], enabling more accurate identification of the biological mechanisms evolution converged upon. Using genome-scale models (GEMs) (Appendix B, Table 1.1) in Pareto analysis has demonstrated trade-offs between five objectives in hepatocytes[156] and between growth and protein secretion in CHO cells[157]. Pareto analysis has also been adopted to understand omics data, enabling the identification of the number and type of objectives present in the dataset, as well as the features that support each objective[158].

1.3.1 Context-specific Trade-offs Underlie Cellular Decision-Making

Trade-offs between multiple objectives require a cell to choose how to allocate its resources and prioritize certain activities. Context-dependent changes in phenotype and the underlying cell state can be understood by cellular decision-making: different contexts introduce different objectives and resource budgets to the cell, imposing trade-offs along the context dimension (Fig. 1.3a). Context-dependent changes to the biological objective are reflected in the underlying molecular processes that drive activity, such as global changes in translational efficiency during differentiation[159] and spatial changes in glycolytic fluxes during development[160].

Multiple objectives are not always present simultaneously, but may arise in sequence according to some context variable (e.g., as a cell migrates, ages, or encounters stress). For example, trade-offs between growth and non-growth associated maintenance (NGAM) over time can explain age-related declines in biological function (Appendix D). Additionally, the extent to which eukaryotic gene expression programs optimize for growth depends on nutrient availability, indicative of context-specific tuning of machinery for non-growth objectives[72]. It is worth noting that mammalian cells, particularly in homeostatic tissue, are likely often prioritizing for non-growth objectives; however, research in this area remains limited, partially because studies using physiological readouts beyond growth are uncommon (see Conclusion for details).

Proteome allocation demonstrates how global strategies of machinery expression are selected according to context-specific trade-offs. Proteome allocation assumes a context-independent, constant upper bound allocated to total protein abundance due to synthesis costs and spatial constraints[116,122,161–164]. Due to this upper bound, increases in expression of one or a group of proteins that support a given function requires compensatory decreased expression of others. An insightful consequence of this proteome allocation perspective is in the use of respiratory or glycolytic pathways in energy metabolism (Appendix A). Proteome allocation trade-offs explain why cells employ the machinery cost minimization strategies previously discussed, as it enables them to not only efficiently perform a single task, but also to have a larger capacity to perform multiple tasks simultaneously and express some machinery in excess to hedge for future conditions. Highly abundant proteins perform “core” functions, such as gene expression and energy metabolism, that tend to be conserved across contexts. In contrast, context-specific proteins have lower abundance, suggesting proteome re-allocation across contexts minimizes machinery biosynthetic costs by unevenly distributing their abundance according to function[165].

1.3.2 Cells Hedge for Future Contexts

Cells adjust their pathway activity accordingly to adapt to the resource demands of each context. To rapidly adapt, cells employ hedging: rather than being fully optimized for one context, cells divert some of their resources in anticipation of new tasks.

For example, the *E. coli* proteome is not fully optimized for growth, re-allocating up to 95% of its proteome by mass fraction from growth functions such as energy production to those such as cell signaling and membrane transport in anticipation of stresses[163]. Indeed, *E. coli* balances multiple objectives, including growth, minimizing total metabolic fluxes (Appendix B), and ATP yield. However, they do not lie exactly on the Pareto front to decrease the cost of adjustment between objectives[166]. In *S. cerevisiae*, cells allocate some of their ribosomes for future growth, and under hyperosmotic stress undergo cell cycle arrest, sacrificing the speed of their adaptive response to maintain glycogen reserves in preparation for subsequent stresses[167]. Similarly, mammalian cells may allocate some of their proteome towards aerobic glycolysis to hedge for hypoxia (Appendix A).

The extent to which cells utilize hedging varies, as some cells specialize for specific contexts. For example, different microbial strains tend to be optimized for glycolytic or gluconeogenic growth, resulting in long lag times upon nutrient shifts in one direction of central carbon metabolism. Probing this further, trade-offs between lag time, growth rate, and futile cycling prevent optimality in both directions, driving these cells towards **specialization**[168]. While these trade-offs shape the decision-making of unicellular organisms, multicellular, mammalian organisms have evolved strategies to limit trade-offs and the need for hedging.

1.4 Multicellularity: From Cells to Organisms

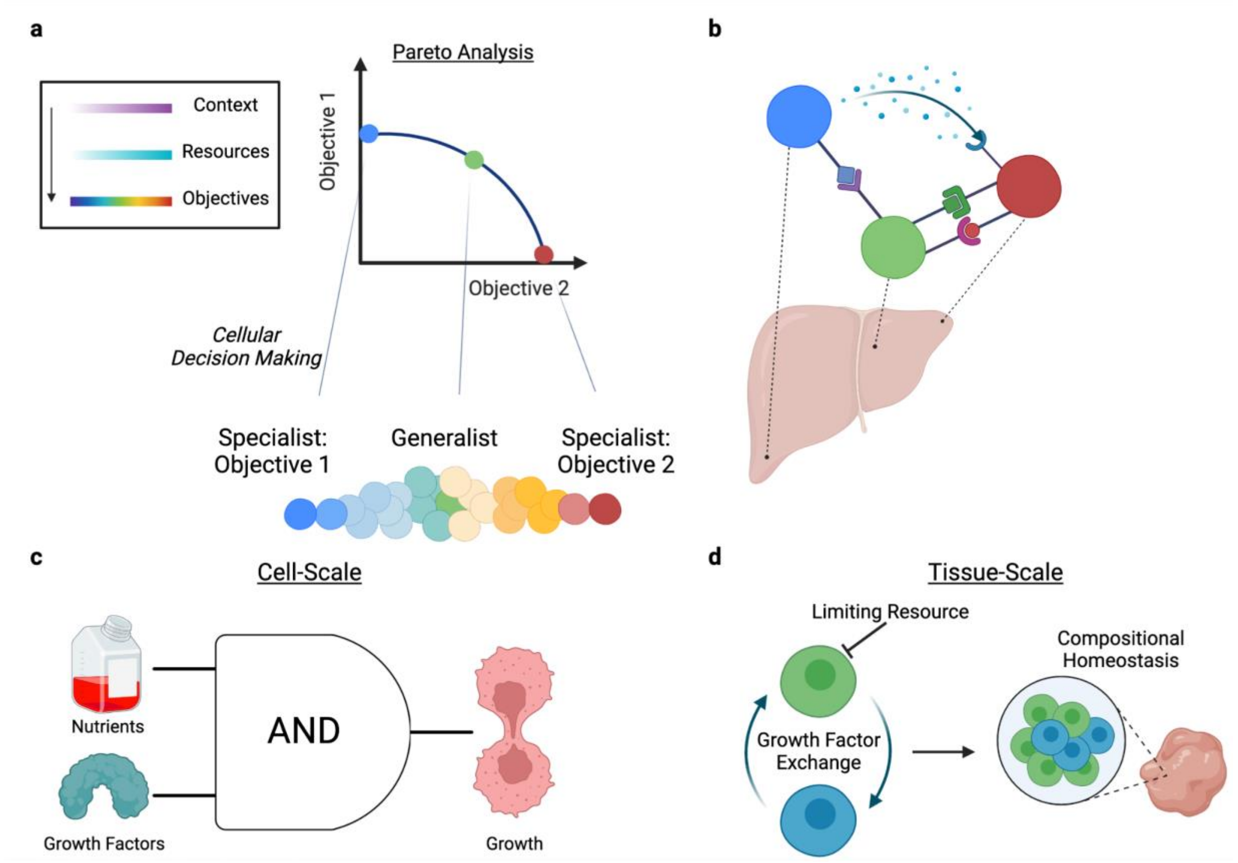


Figure 1.3: Extending resource allocation to multicellularity.
(a) The cellular context determines the available resources and cell objectives. Resource constraints limit the number of objectives a cell can achieve and the efficiency by which it performs each objective; thus, individual cells that face multiple objectives must manage trade-offs. Pareto analysis provides a quantitative view of trade-offs wherein competing objectives are plotted. Along the Pareto front, a cell cannot improve its performance in one objective without worsening its performance in another. Within multicellular systems, this leads to cell specialization for specific objectives along a context gradient. **(b)** In a tissue, cell specialization can occur across, for example, spatial context gradients (see ref[169]). With division of labor, each cell performs a particular set of tasks, but the higher-order biological function results from the synergistic effects of the population. Tissue-level behavior is achieved by coordination between cells via mechanisms such as cell-cell communication, in which distinct signals are sent in a cell-type- and context-dependent manner. **(c)** Individual cells receive communicatory signals such as growth-factors that, in combination with local nutrient availability, regulate growth phenotypes. **(d)** In a tissue, multi-cell circuits involving growth factor exchange, in combination with resource constraints, maintain steady-state proportions of cell types (see ref[170]).

Multicellularity emerged to efficiently manage resource trade-offs (Fig. 1.2). In a multicellular system, a tissue receives myriad cues for distinct objectives, which impose demands on distinct cellular resources and pathways[6]. Cell-scale tradeoffs induce coordination, resulting in higher-order functions[15,16] that ultimately impact the whole-organism (Appendix C). Pareto analysis of mammalian tissue demonstrates that, subject to trade-offs, single-cells can become specialists optimized for a particular task or generalists that can perform multiple tasks (Fig. 1.3a)[158,171].

1.4.1 Division of Labor Distributes Resource Burdens

Cell specialization is a **division of labor** resource allocation strategy employed by multicellular organisms. Due to resource constraints, individual cells specialize to perform a subset of the tasks a tissue must execute. However, division of labor distributes the tissue tasks across multiple cell types in a manner yielding synergistic effects: the cells' combined functions amplify performance across a range of tissue-level tasks. For example, division of labor mitigates both the cost of switching between multiple objectives in a tissue[172] and the loss of tissue function when cells proliferate (insead of performing the objective they are specialized for) for homeostatic maintenance and turnover[173]. Furthermore, diversity in the extent of specialization enables robustness against perturbations[18].

In tissues, resource constraints and cell objectives are often associated with spatial context gradients that alter the extracellular cues each individual cell faces (Fig. 1.3b). For example, terminally differentiated cells have varying degrees of specialization to optimize tissue function (i.e., maximize net performance of all objectives) when external spatial gradients are present, hence explaining observed gene-expression continua within a cell type. Enterocytes are distributed continuously across three objectives–cell adhesion and lipid transport, carbohydrate and amino-acid uptake, and anti-bacterial defense–in a manner correlated to their location in the intestinal villus[169]. Hepatocytes also alter expression patterns of ~50% of their genes in accordance with their spatial location. These patterns reflect coordination between all three resource classes–liver zones with higher oxygen availability also demonstrated machinery shifts towards oxidative phosphorylation, likely generating energy to support higher protein secretion[174].

1.4.2 Coordination Enables Higher-Order Functions

As individual cells allocate their resources to specialize, emergent tissue-level functions arise from coordination, wherein resources are distributed across systems[154] and yields synergistic effects that enable phenotypes no individual cell type can[18]. For example, division of labor enables tissue to perform multiple tasks more optimally than any individual cell. Here, the type and range of cell-cell communication influences tissue spatial patterning to coordinate specialization, enabling cells to self-organize as efficient multicellular systems[175]. In the rove beetle tergan gland, cell specialization resulted in the co-evolution of two specialized cell types that each produce small molecules which are innocuous in isolation, but form a defensive toxin when combined [176].

Due to communication costs, resource allocation also impacts how the information to organize multicellular systems is distributed (Fig. 1.3b). For example, signaling pathways optimize for specific objectives such as sensing precision[140]. To manage spatial-gradient-induced noise during wound healing, cells maximize sensing precision by coordinating within optimal local distances via paracrine growth factor communication[177]. Indeed, Pareto optimal task-distribution in tissues can be separately influenced by spatial gradients and local communication[175]. We saw previously that the three individual resource classes affect cell fate, changing the composition of cell populations. Communication also maintains steady-state cell type proportions in tissue by tuning cell proliferation and removal rates (Fig. 1.3c-d). Zhou *et al.* identified a two-cell macrophage-fibroblast circuit that uses growth factor exchange for compositional homeostasis. In this circuit, fibroblast proliferation is limited by resource constraints whereas macrophages are limited by negative feedback of the growth factor CSF1[170]. This is particularly interesting given that cells are relatively more sensitive to autocrine signaling at low densities and that feedback loops can generate bistability within isogenic cells[150]. These cells even optimize the specific type of negative feedback they use, using auto-regulation via endocytosis to enhance the response time and robustness of the circuit in achieving steady-state[178]. The specific growth factors used are context-dependent across nutrient-limiting conditions, and the distinct cell types are competing for different limiting resources[179]. However, while context affects individual parameters, such as cell growth rates, the circuit's stability is generalizable. Furthermore, the topology of this stable circuit can be extrapolated to circuits with more cell types[178].

GEMs can be used to model intercellular resource allocation and metabolic crosstalk between cells[180]. For example, multicellular GEMs modeling brain astrocytes and neurons delineate metabolic interactions, demonstrating that energy pathways are distributed across cell types to minimize protein costs[181] and revealing metabolic phenotypes underlying Alzheimer's Disease[182]. GEMs also elucidated how fibroblasts reprogram colorectal cancer cell metabolism by stimulating pathways such as glycolysis and glutaminolysis without altering growth[183].

Coordination between cell types can also be advantageous for disease states. In cancer, for example, groups of circulating tumor cells demonstrate more than an order of magnitude increased metastatic potential compared to individual cells [184]. Extending the previous discussion of motility resource allocation[81] to the multicellular dynamics of cancer metastasis, cells invade cooperatively to minimize energetic costs of migration. A "leader" cell disproportionately expends its resources to displace the matrix, making migration easier for "follower" cells. After it has depleted its energy budget, the leader cell is replaced by a follower cell that still has a high energy budget[185]. In a simplified model, one study demonstrated that metastatic invasion occurs in the presence of a nutrient gradient, indicating the metastatic population is willing to pay the energetic costs of migration in search of a location with more nutrient resources[186].

1.4.3 Resource Competition Maintains Homeostasis

While division of labor improves the efficiency by which cells allocate resources and allows for coordination, much like the intracellular pathways of an individual cell, cells in the same microenvironment compete for a shared pool of extracellular resources. In some cases, cells in the same microenvironment–particularly if they share the same identity–are not only using the same pool of resources but are also using those resources in the same manner because they have a shared objective. We saw this in the macrophage-fibroblast circuits, in which macrophages competed for growth factors whereas fibroblasts competed for space[179]. During development, competition actually serves as a coordinating mechanism to improve morphogenesis (i.e., growth, differentiation, and structure), with resource scarcity being the coordinating signal[187]. Thus, competition is an important homeostatic mechanism constraining any individual or population of cells' objective (e.g., growth) from dominating.

Canonically, cell competition is defined as the "active elimination of intrinsically viable cells that differ in some way from their neighbors"[188]. Within this definition, resource competition is one such mechanism for elimination, with less fit "loser" cells unable to use resources as efficiently as "winner" cells to achieve their objective, thus optimizing overall tissue fitness[189]. For example, under conditions of nutrient abundance, lower Myc expression causes cells to have a lower anabolic capacity, ribosomal abundance, protein synthesis rates, and proliferative capacity[190,191]. These less-fit cells are eliminated via apoptotic signaling and asymmetric cell division. In epidermal expansion, during mouse embryogenesis, for example, this elimination ensures appropriate tissue architecture.

Similarly, oncogenic epithelial cells are eliminated by wild-type cells via apical extrusion. Metabolically, the less-fit cells are inefficient, shifting towards aerobic glycolysis and exhausting glucose without substantially increasing ATP[192]. In contrast to homeostatic maintenance, metabolic competition in tumors, which have high resource demands, can lead to disease progression. For example, tumor cells reduce tumor infiltrating lymphocyte (TIL) effector function by competing for glucose. Lower glucose availability decreases TILs' oxidative phosphorylation flux and their biosynthetic capacity to secrete interferon gamma[193]. The high lactic acid levels produced by tumors via aerobic glycolysis in low glucose conditions also decreases immune cell ATP levels and biosynthetic capacity, ultimately suppressing TIL infiltration and survival[194].

However, the presence of other nutrient sources and cell types may mitigate the role of competition[195]. Recent work has leveraged the high resource demands of cancer cells by engineering adipocyte glucose and lipid metabolism to outcompete tumors[196,197].

1.5 Conclusions

Resource allocation provides a unique perspective to the mechanisms mediating mammalian biology across cell-, tissue-, and whole-organism- scales. The tradeoffs induced by resource constraints have favored the evolution of multicellular systems with specialized cell types that synergistically contribute to higher-order functions. As outlined in this Review, resource allocation impacts homeostasis, and context-dependent changes in allocation are accompanied by physiological changes, as seen in diseases such as obesity, Alzheimer's, aging, infection, and cancer. However, to comprehensively understand these changes and better define general principles that are predictive across contexts, the discussed systems biology approaches must be extended.

Multi-omics measurements alongside systems biology analyses (Appendix B, Table 1.1) enable studies that account for the interconnectivity of molecular processes and provide a mechanistic connection between resource allocation and phenotype[198]. Models that incorporate multiple molecular processes are important to fully and explicitly account for the resource costs of biological activity. Many models are limited to prokaryotes due to the complexity of mammalian systems (e.g., multiple subcompartments with distinct localization[199], protein secretion, etc.) and limitations in joint molecular (e.g., omics) and physiological (e.g., cell morphology[200], size[201], density[202] and growth rate[203]) measurements. Additionally, many modeling approaches have focused on protein machinery. However, other macromolecules, such as noncoding RNA (e.g. lncRNAs, tRNA[204], etc.), also constrain cell activity; without incorporating these molecular details, models may miss key regulatory components that affect phenotype. Finally, while resource allocation is apparent at the whole-organism scale (Appendix C), the underlying molecular mechanisms regulating this have not been extensively studied. Recently, whole-body models have extended the GEM framework to address this gap[205], but these can benefit from more detailed measurements, especially at the resolution of specific cell-types. Improving modeling approaches, when complemented by more comprehensive quantitation, will provide excellent opportunities to uncover principles of resource allocation in mammalian cells and will also be important to link extra- and intra-cellular activity to understand multicellularity.

1.6 Appendix

1.6.1 Appendix A: Resource Constraints and Growth

While some cell types endure dynamic phenotypic changes (e.g., immune responses, development, and hepatocyte communication), many cells in tissue exist in a steady-state, with resources allocated to homeostatic maintenance. Thus, resource allocation is often simplified into a stratification between growth- and non-growth associated maintenance (NGAM) costs (Fig. 1.4a-b). NGAM represents biological processes that do not directly contribute to growth, such as error-checking[206], maintenance of membrane potentials, and macromolecular maintenance given turnover. While some of these processes may change with growth, they have baseline NGAM activity quantitatively defined as resource costs at zero growth[207,208]. Underlying the balance of growth and NGAM are alterations in gene expression of cellular machinery and energy management that ultimately affect organismal phenotypes (Appendix D, Fig. 1.4c).

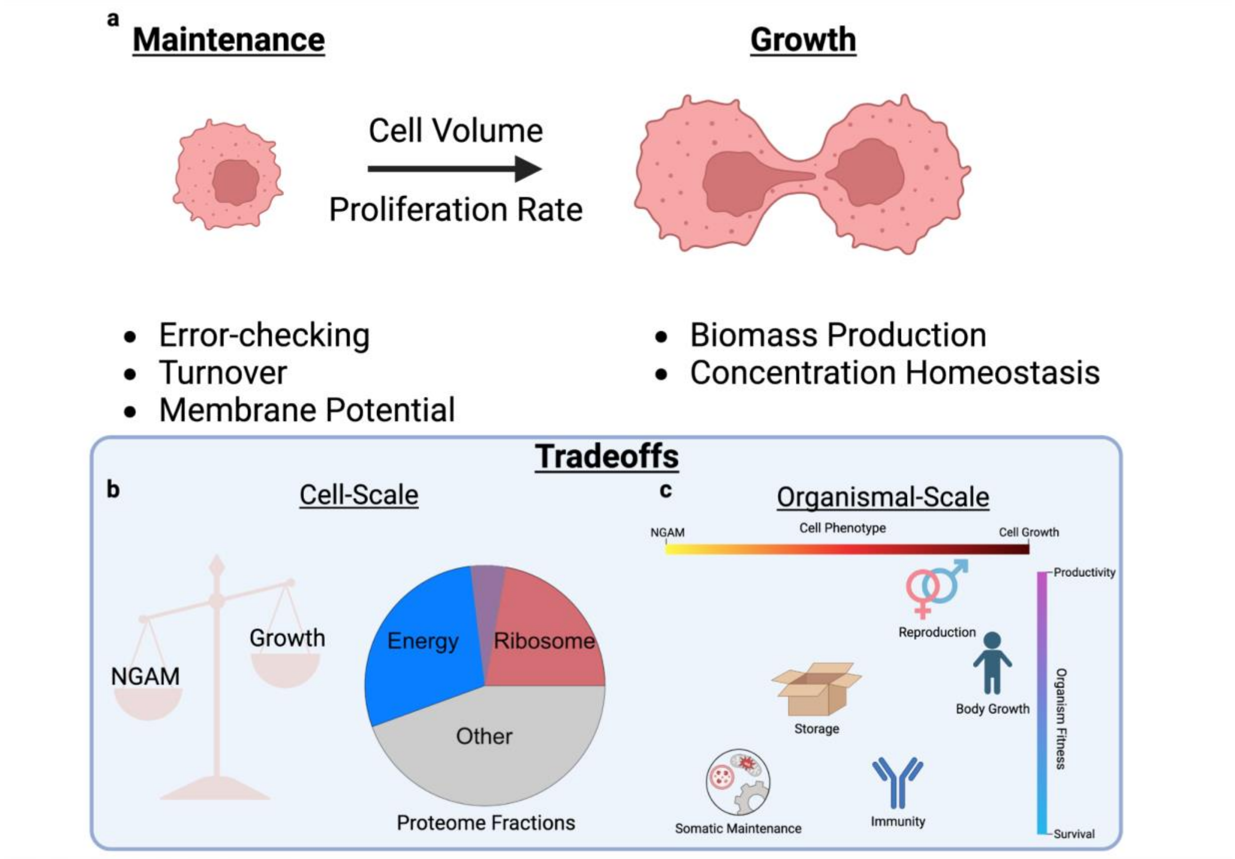


Figure 1.4: Maintenance and growth objectives in cells and organisms.
(a) Cell maintenance and growth both require resource investments and incorporate distinct biological activities. **(b)** Consequently, they trade-off with each other. Trade-offs are observed by the proteome fraction allocated to translational machinery for biomass production (red) and machinery for catalysis of bioenergy metabolism (blue). The purple shaded region represents varying cell proteome allocation towards translation or bioenergy with changing context. **(c)** Life-history traits optimally trade-off against each other to maximize whole-organism fitness via productivity and survival. While these trade-offs mirror those at the cell-scale, rather than being binary, they use a combination of cell growth and NGAM.

1.6.1.1 Growth Phenotypes Depend On Gene Expression

Gene expression indirectly (as machinery facilitating biosynthesis) and directly (as biosynthetic products) contributes to biomass production. The relationship between gene expression and growth maintains concentration homeostasis–which is necessary for appropriate cell function[201]–with cell volume expansion by dynamically adjusting synthesis rates through resource loading of transcriptional and translational machinery[209]. Thus, in prokaryotes[210] and eukaryotes[122,211,212], the proteome fraction allocated to ribosomes increases linearly with increasing growth rate to support increased biosynthetic demand. Such linear relationships delineate the extent to which cells are nutrient- or machinery- limited in growth, with cells allocating a baseline unused portion of ribosomes to hedge for future increases in growth rate.

In rapidly proliferating prokaryotes, protein degradation is often assumed to be negligible because dilution rates due to cell division far outweigh degradation rates[103]. Yet, in slower-growing mammalian cells, degradation rates play a substantially more important role in dictating overall protein abundance[102,213]. Macromolecular synthesis must not only contribute to biomass production and counteract biomass loss due to dilution, but also biomass loss due to degradation. Consequently, the linear relationship between ribosomal mass fraction and growth rate is altered at slow growth, shifting a larger than predicted fraction of proteome mass to active ribosomes to maintain biomass when faced with non-negligible protein degradation[214].

1.6.1.2 Energy Budgets are Balanced Between Growth and NGAM

In mammalian cells, linear relationships exist between individual genes' abundance with cell size[215] and between other phenotypes such as migration with growth rate[216]. Omics measurements of individual genes enable further exploration of proteome re-allocation between translational machinery and other functional protein sectors, such as energy metabolism (Fig. 1.4b). Consistently, there is a shift in energy metabolism from respiratory to glycolytic activity with larger cell size and higher growth rate[122,211,215].

By knocking out key enzymes facilitating aerobic glycolysis, rapidly proliferating mammalian cells shift back to OxPhos without reducing their growth rate[217]. Additionally, under both respiratory and glycolytic conditions, the effect of ribosomal proteins on growth rate is independent of the energy budget[218]. This indicates that, unlike in microbes (Appendix D), proteome constraints alone cannot explain the shift to anaerobic glycolysis at high growth rate in mammalian cells. The metabolic shift may not provide enough of a fitness benefit, with mammalian OxPhos having higher protein efficiency (Appendix D) as compared to aerobic glycolysis[219]. An alternative explanation may lie in proteome hedging, wherein cells express excess anaerobic machinery for future contexts, such as hypoxia at high growth[220,20]. However, in this case, resource allocation might not be as applicable to mammals, whose coordinated tissue systems mitigate trade-offs and typically do not anticipate high growth.

1.6.2 Appendix B Systems Biology Approaches to Understand Resource Allocation

The coupling of activity between different intracellular molecular processes, constrained by the cellular resource budget, dictates phenotype. Many systems biology approaches combine high-throughput measurements with computational algorithms to model the interplay between intracellular processes and understand phenotype.

Phenomenological models, for example, are coarse-grained representations that identify quantitative relationships between resource constraints, cellular activity, and phenotype. Such models have described relationships between gene expression and growth[210], proteome allocation under resource limitation[162], and macromolecular concentration homeostasis[209]. Describing such relationships in a few meaningful parameters that are robust across contexts, these models shed light on principles of resource allocation. However, they lack molecular details.

Thus, models that integrate resource availability with mechanisms are necessary. Kinetic ordinary differential equations (ODEs) are commonly used to model gene expression (Fig. 1.5a). However, these models require extensive measurements of kinetic parameters and cannot account for resource availability. In contrast, genome-scale metabolic metabolic models (M-Models) map out all metabolic pathways of the cell, represent them mathematically using their stoichiometric coefficients[221], and use flux balance analysis (FBA) to simulate the reaction fluxes[221,222]. These fluxes optimize a cellular objective that represents a phenotype of interest (often growth rate[208]) while directly accounting for nutrient and bioenergy resource constraints. In mammalian cells, FBA has explored numerous questions, including polyamine metabolism in T-helper 17 cell pathogenicity[223] and enzymes that affect migration, but not proliferation, of cancer cells[224].

M-Models do not comprehensively account for machinery resources. They can partially do so using methods to overlay gene expression measurements that prune the metabolic network[225], but this only incorporates machinery in a binary manner. Other methods have constrained M-Models by more explicitly minimizing machinery costs. Parsimonious FBA, for example, minimizes the sum of fluxes throughout the network as a secondary objective. Here, the total flux is a proxy for the global "effort" or resource cost of biological activity[226,227]. Another method estimates the extent of total reaction flux that is backwards as a proxy for machinery costs, with higher backwards flux requiring a larger enzyme budget to drive the reaction forward[74]. Finally, a number of strategies constrain reaction fluxes by machinery abundance at varying resolution[73,161,228,229]. Machinery coupling is not limited to canonical catalytic enzymes, but can be extended to transport[230], gene expression[64], and protein secretion[157].

Such methods that incorporate individual enzymes into genome-scale models differentiate between nutrient- and machinery-limiting conditions. However, to explicitly account for the entire machinery budget and costs, genome-scale models of metabolism and expression (ME-Models) add gene synthesis reactions for each enzyme[231] and couple them to metabolism (Fig. 1.5b)[232,233]. In prokaryotes, ME-Models[234] have shed light on proteome allocation strategies under oxidative[235], acidic[236], and thermal[237] stress. In eukaryotes, the ME-Modeling framework accounts for compartment-specific proteome constraints, enabling analysis of metabolic shifts, machinery protein costs, and proteome reallocation[164]. Resource Balance Analysis (RBA) models employ similar concepts to encode gene expression and metabolism while also accounting for spatial constraints[238]. While RBA is limited to prokaryotes, a theoretical roadmap for eukaryotes exists[239]. Finally, whole-cell models (WCMs) additionally characterize kinetics and encompass more cellular processes[240]. By modeling each intracellular system separately then connecting them via common inputs and outputs, WCMs can account for resources shared between multiple intracellular subsystems[92] and interconnectivity of molecular processes, e.g. the effect of glycolytic flux for producing dNTPs used in chromosomal replication[91].

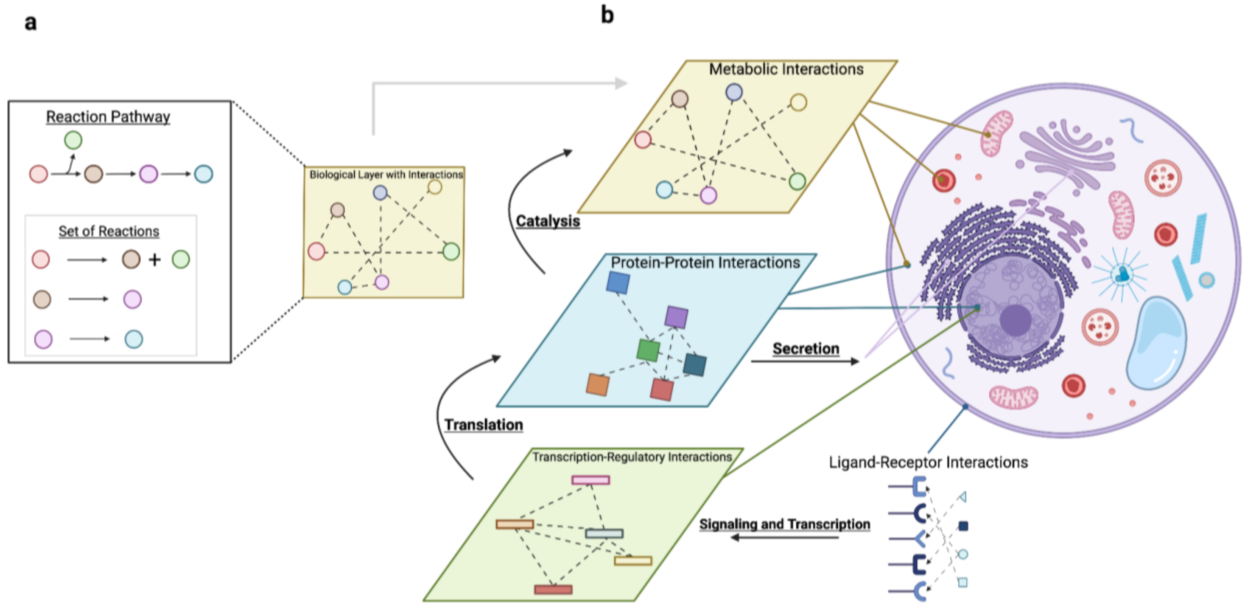

a
b
Reaction Pathway
Set of Reactions
Biological Layer with Interactions
Metabolic Interactions
Catalysis
Protein-Protein Interactions
Secretion
Translation
Transcription-Regulatory Interactions
Ligand-Receptor Interactions
Signaling and Transcription

Figure 1.5: Modeling cellular activity from individual reactions to global networks.
(a) A pathway with a set of connected molecules can be decomposed into its individual reactions. These reactions can be used for high-resolution modeling of a particular process, incorporating detailed kinetic parameters. Globally, all the reaction pathways form a network defining some molecular process such as metabolism. **(b)** Processes can be coupled (black arrows) using first-principles to create more comprehensive genome-scale models such as ME-models. Signaling and secretion incorporate extracellular interactions, providing a basis for modeling multicellular systems.

1.6.3 Appendix C: Resource Allocation Affects the Whole-Organism

Much like coordination between cells in tissue, tissues interact to efficiently allocate resources across the whole-organism. Such strategies reveal how whole-organism physiology is affected by factors such as diet and exercise.

Mammals use complex endocrine signaling mechanisms to sense their internal nutrient state[241–244], which affects their food choices and, ultimately, the composition[245] and elemental stoichiometry[246] of nutrients available to cells. Regulatory mechanisms[247–249] ensure that diets optimally balance nutrient ratios to maximize organismal performance. If individual foods do not meet these ratios, mammals consume mixed diets with complementary foods to target this optimum[250–253]. Furthermore, when food availability is constrained and the ideal balance cannot be met, mammals may over- or under-consume some nutrients to prioritize target amounts of others[254,255]. Humans, for example, prioritize protein intake over carbohydrates[256]. Multicellular organisms may also employ post-ingestion strategies to balance nutrient ratios, particularly under scarcity[257]. Insects, for example, modulate secretion of digestive enzymes to re-balance nutrient ratios. Excess carbohydrates decrease carbohydrase secretion and excess proteins decrease protease secretion, balancing nutrient ratios while also minimizing machinery costs[258].

Diets are tightly regulated because nutrient imbalances have associated costs that constrain performance[259]. Mammals must handle excesses by voiding or storing[260] them and deficits by conserving them or mobilizing nutrient stores. Protein excesses lead to increased excretion[261] and macromolecular degradation[262] costs, while protein deficits causeincreased digestive passage times to increase nitrogen intake, resulting in decreased digestive enzyme efficiency and loss of non-diet proteins[263]. Nutrient imbalances can also force cells to utilize less efficient metabolic pathways. For example, gluconeogenesis is used in high-protein, low-carbohydrate diets to form glucose from amino acid precursors. This has substantially higher energetic costs than using carbohydrates directly[264]. Metabolic changes from imbalanced diets are tied to systemic changes involving multiple tissues[265], such as starvation from deficits[266] and obesity from excesses[267]. Human prioritization of protein intake[256] in the presence of modern high fat and carbohydrate foods, in combination with dysregulated insulin secretion in overweight individuals, has been suggested to contribute to the modern obesity epidemic[267]. For example, amino acid variability in human diets in association with obesity is higher than that of carbohydrates and fats[268].

After ingestion, the body optimally allocates resources across tissues with different metabolic rates and resource expenditures[85,269]. This allocation is coordinated through the aforementioned sensing and signaling mechanisms that connect resource distribution with demand. Competing demands between tissues inevitably lead to trade-offs that impact whole-organism performance[270]. The high-energy demands of the brain, for example, impose evolutionary decreases in net resource expenditure of other organs such as the gut[271] and trade-off with human juvenile body growth[272]. Since the brain processes the regulatory signals, one theory proposes that the brain prioritizes its own glucose supply to maintain intracellular ATP homeostasis[273–275]. These allocation decisions have also explained trade-offs between life-history traits[276] such as survival and reproduction (Appendix D).

Diet influences these trade-offs because certain nutrient balances are more optimal for certain traits[268,277]. For example, specific diets can affect performance of specific types of exercise[278] (e.g. aerobic endurance training[279–281] vs. anaerobic high-intensity training[282,283]). This makes sense since distinct nutrient stores and bioenergetic pathways are used for different exercise[284]. Furthermore, diet interacts with exercise to yield synergistic effects on performance[285–288].

Skeletal muscle accounts for ~25% of the whole-body basal metabolic rate[289]. At the cell-scale, this is due to the ATP-consuming processes that enable muscles to perform mechanical work: activation via Ca^{2+} cycling and contraction via myosin cross-bridge cycling (i.e., ion pumping and force generation)[290]. The energetic costs depend on factors such as the type[291,292], frequency[293], and duration[294] of contraction, with more rigorous exercises consuming more resources. These high resource demands cause physical activity to trade-off with other energy-demanding functions, such as reproduction[295] and immunity[296]. For example, upon antigen exposure, mice must divert energy resources from tasks such as locomotor activity to immunity[297]. Furthermore, during nutrient scarcity, insulin-dependent GLUT4 glucose uptake by skeletal muscle and adipose tissue is suppressed to preserve resources for the brain[275]. To manage these trade-offs, mammals have evolved various strategies for efficient muscle use. This includes tight regulation of concentration homeostasis for maximal bioenergy metabolism[298,299], cell-type specialization[300] and tissue heterogeneity[301,302] optimized for distinct mechanical tasks, and use of energy-conserving motions[303–306].

1.6.4 Appendix D: Further Insights into Growth and Maintenance

1.6.4.1 Microbial Shifts in Energy Metabolism at High Growth Rate

In microbes, resource allocation has provided a possible explanation for the shift from respiratory to glycolytic activity with higher growth rate. First, machinery facilitating protein synthesis impose a large resource burden[307]; in human cells, for example, translational machinery comprise more than 15% of total proteome mass[308]. Given an upper limit on the total proteome mass, increases in the ribosome mass fraction with growth rate induces trade-offs with other protein sectors. Second, as growth rate increases, the energy budget must also increase[22218], largely to accommodate increased ATP consumption for biosynthesis[45]. Thus, the cell needs a strategy to accommodate increased proteome allocation towards translational machinery while still maintaining or increasing its energy budget. As such, cells will shift from high yield (ATP generated per unit glucose) but low protein efficiency (ATP generated per unit machinery mass) oxidative phosphorylation to low yield but high protein efficiency aerobic glycolysis as growth increases[309].

1.6..4.2 Energetic Trade-offs Beyond Proteome Allocation

Energetic trade-offs extend beyond proteome allocation, with nutrients used for anabolism incurring indirect costs–defined by their potential to be fully oxidized to yield energy–in addition to the direct energetic costs of biosynthesis[310]. As such, cells must account for the opportunity cost of the fitness gain from energy production relative to biosynthesis. In line with this, mammalian cells demonstrate sensitivity[35] and specificity[19] in the nutrient type used for biomass generation. These opportunity costs also contribute to trade-offs between NGAM and growth[311]. For example, maintaining protein levels at slow-growth consumes 20% of the mammalian energy budget (BioNumbers[59] ID 113244).

1.6.4.3 Growth and NGAM Scale to Whole-Organisms

Life-history traits maximize organismal-scale fitness by contributing to survival and productivity (i.e., energy invested in activities other than survival)[312]. Traits classically include body growth prior to maturity, reproduction after maturity, and somatic maintenance; these traits optimally trade-off against each in a manner analogous to cell growth and NGAM. The connection between these scales lies in nutrient sensing and signaling (e.g., insulin-signaling[313] and target of rapamycin pathways) that stimulate either cell growth or NGAM depending on resource availability, thus prioritizing certain traits over others[314]. While cell growth largely contributes to body growth and reproduction, NGAM contributes strongly to somatic maintenance[312,314]. However, these traits tend to lie on a gradient requiring both cell growth and NGAM (Fig. 1.4c). Somatic maintenance, for example, uses cell growth to balance apoptosis of damaged cells during tissue turnover.

Additional functions such as immunity[315] and nutrient storage[29] also contribute to organismal fitness, but with additional costs that induce further trade-offs. These trade-offs are dynamic over changing resource availability in an attempt to optimize lifetime fitness[316]; whereas immunity contributes to immediate survival, storage is a hedging strategy for future scarcity. For example, children that exhibit higher levels of immune activity demonstrate decreased body growth, but this trade-off is tempered if those individuals have more body fat[315]. Over time, such trade-offs introduce various stresses to the body that must be dealt with[317], leading to a decrease in cell- and organismal-fitness that possibly explain aging[318]. Finally, maintenance tasks can also trade-off between each other; for example dairy cow calcium homeostasis is Pareto optimal between effectiveness–or a rapid return to physiological blood concentrations after milk production–and economy–or avoiding excess release of calcium from body stores[155].

1.6.5 Appendix F: Modeling Machinery Activity and Abundance to Couple Protein with Metabolism

As an example of systems biology approaches , we model here enzyme-catalyzed reactions. The assumptions regarding their reaction kinetics are used to constrain models with machinery costs and couple gene expression to metabolism.

For conversion of substrate (S) to product (P) by a machinery enzyme (E), Michaelis-Menten kinetics is often assumed:

(1)

$$v = k_{cat}[E]\frac{[S]}{[S] + K_M}$$

Under *in vitro*, fully saturated conditions ([S] >> K_M), the maximal reaction rate is achieved:

(2)

$$v_{max} = k_{cat}[E]$$

In equation (2), reaction rates depend on machinery abundance and activity. However, *in vivo* observed reaction rates often deviate from this maximum value due to saturating, thermodynamic, and regulatory effects:

(3)

$$v_{obs} = k_{cat}[E]\eta(C) = v_{max}\eta(C)$$

η(C) is a context-dependent scaling parameter that accounts for the discrepancy between v_{max} and v_{obs} due to *in vivo* variables such as metabolite concentration, pH, and molecular crowding. η(C) can be decomposed into saturating, thermodynamic, and regulatory components. Saturating and thermodynamic components can only decrease v_{obs} relative to v_{max}. However, regulatory effects such as allostery and post-translational modifications can cause $v_{obs} > v_{max}$.

Focusing on saturating effects, if substrate concentration is no longer sufficiently high such that K_M is not negligible, then enzymes are not fully saturated. In this case, we can represent total enzyme concentration as:

(4)

$$[E] = [E]_{used} + [E]_{free}$$

From Michaelis-Menten derivations:

(5)

$$[E]_{used} \equiv [ES] = \frac{[E][S]}{[S] + K_M}$$

We assume the scaling factor as proportional to the fraction of used enzyme :

(6)

$$\eta_{saturation}(C) = \frac{[E]_{used}}{[E]} = \frac{[S]}{[S] + K_M}$$

Thus, $\frac{[S]}{[S] + K_M}$is the enzyme saturation scaling factor. From equations (3) and (6), and also shown in the Michaelis-Menten derivation, Equation (1) can also be written as:

(6)

$$v_{obs} = k_{cat}[E]_{used}$$

Defining the capacity utilization of a reaction as the ratio between the observed and maximal reaction rate yields the following equality from equations (2) and (6):

(7)

$$\frac{v_{obs}}{v_{max}} = \frac{[E]_{used}}{[E]}$$

Note that the capacity utilization and saturation scaling factor are equivalent and analogous concepts, and many of these equations are different conceptual representations of the same mathematical equivalencies. Broadly, equation (7) indicates that with saturating effects, there is some fraction of unused enzyme that is decreasing the overall reaction rate.

Next, we will focus on thermodynamic effects. For simplicity, we will consider an isothermal, isobaric reaction of a substrate converted into product, disregarding Michaelis-Menten kinetics (conclusions are the same):

$$S \underset{k_r}{\overset{k_f}{\rightleftharpoons}} P$$

The reaction kinetics can be written as:

(8)

$$v_{obs} = v_f - v_r = k_f[S] - k_r[P]$$

The reaction thermodynamics can be written as:

(9)

$$\Delta G = -RT\ln\left(\frac{[S]K_{eq}}{[P]}\right)$$

The kinetics and thermodynamics are related to each other via the equilibrium constant K_{eq} from the following:

(10)

$$K_{eq} = \frac{k_f}{k_r}$$

Thus, equations (9) can be rewritten using equations (8) and (10) to link thermodynamics with kinetics as:

(11)

$$\Delta G = -RT\ln\left(\frac{v_f}{v_r}\right)$$

Equation (11) holds for more complex reactions, including Michaelis-Menten kinetics and the Hill equation. From (11), reactions far from equilibrium ($\Delta G << 0$) tend not to have a high reverse flux. However, enzymes catalyzing reactions near equilibrium ($\Delta G \sim 0$) face a higher reverse flux. In this sense, thermodynamic forces (e.g., intermediate metabolite concentrations) can affect machinery costs, with reactions that have a higher reverse flux requiring a larger enzyme investment per unit of forward flux.

Relevant Sources: ref.[73,74,120,121,319]

1.6.6 Appendix G: Glossary

- **Cellular context:** A combination of the current intracellular state (e.g., genomic variants, cell type, and machinery concentrations and localization) and extracellular cues from the microenvironment (e.g., nutrients and communicatory molecules) that together inform cellular decision-making and change as a function of factors such as time, space, and disease.

- **Cell specialization**: The extent to which a cell is optimized for the performance of a specific, single objective.
- **Division of labor**: A resource allocation strategy in which multicellular systems distribute multiple objectives across cells with varying degrees of specialization.
- **Fitness:** The efficiency by which a system uses its resource budget to achieve its objective; mathematically, this is the extent to which the system minimizes resource costs while simultaneously maximizing its objectives.
- **Hedging**: Resource allocated in preparation for future objectives, particularly at the cost of a current objective.
- **Information Transfer**: The extent to which the output depends on or is informed by the input (e.g., mutual information).
- **Machinery**: The macromolecular products of anabolism and gene expression, often enzymes, that catalyze and enable cell functions.
- **Machinery-limiting:** The flux through a reaction is limited by saturation of the machinery.
- **Nutrient-limiting:** The flux through a reaction is limited by the availability of a metabolic substrate (nutrient or downstream intermediate).
- **Objective:** The biological goal that a cell or system is trying to achieve (e.g., motility, proliferation, and differentiation) through the integration of its various biological activities.
- **Optimality**: The maximization of an objective that is constrained by the resource budget.
- **Resource budget**: The total quantity of a resource (i.e., nutrient, machinery, and bioenergy) that is available to the cell for use.
- **Resource cost**: The total quantity of resources that are consumed or sequestered for biological activities contributing to the cell objective.
- **Resource loading:** Competition in the cell for a shared and often limiting resource.
- **Response Time**: The time between when the pathway senses the input and when it generates the output.
- **Signal Amplification**: The magnitude change in the output relative to the magnitude change in the input.
- **Signaling Crosstalk:** the interaction of shared components between signaling pathways, particularly in the presence of multiple inputs and/or outputs, which requires resource sharing across the signaling network
- **Signaling Modularity:** A set of signaling components that can convert inputs to outputs while limiting retroactivity (i.e., instances in which the inputs and the outputs are not unidirectional).
- **Sensing Precision**: The ability of a signaling pathway to accurately convert a given input to the desired output with limited variance.
- **Parameter Robustness**: The change in the output response given a change to one of the system's parameters (e.g. binding affinities).
- **Pareto optimality:** A state in which increasing performance of one objective can only occur by decreasing the performance of another objective due to resource constraints, leading to trade-offs.

1.6.7 Appendix H: Authors, Contributions, and Acknowledgements

Authors: Hratch M. Baghdassarian, Nathan E. Lewis

H.M.B and N.E.L. conceptualized the work. H.M.B. wrote the paper and generated the figures. All authors carefully reviewed, discussed and edited the paper.

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	Approach	Method	Biological Finding	Organism	Reference
Signal Transduction	Mass Action ODEs	Tests each of five possible objectives as a function of signaling resources estimated by mass action models that incorporate machinery abundance and activity.	Phosphorelay pathways tend to prioritize information transfer as their primary objective.	*B. subtilis, S. cerevisiae*	136
Signal Transduction	Mass Action ODEs	Estimates signal sensing error as a function of 1) receptor abundance, binding, and integration time, 2) downstream signaling machinery abundance and binding, and 3) ATP turnover. Further summarizes this function as a set of parameters that conceptually represent the resource classes of interest.	Goldbeter–Koshland push–pull network sensing systems maximize sensing precision. Each of receptors, signaling machinery, and energy resources are equally limiting on sensing precision to achieve efficient resource allocation.	*E. coli* (network type studied is ubiquitous across prokaryotes and eukaryotes)	140
Signal Transduction	Genome-Scale Modeling	Extreme Pathway Analysis uses convex analysis to identify a minimal set of "extreme pathways". Conceptually, these are pathways whose activities, according to the network stoichiometry and model constraints, each represent a functional state of the network. Mathematically, these pathways are the edges of the steady-state convex solution space, with distributions that are independent and whose non-negative linear combinations represent all possible solutions to the objective.	Proof-of-concept application of the genome-scale metabolic modeling framework to assess signaling pathway activity, crosstalk, and redundancy in the context of a network with distinct combinations of ligand inputs and transcription factor activity outputs.	Prototypic example	144
Signal Transduction	Genome-Scale Modeling		Demonstrated certain properties of the human B-cell JAK-STAT signaling network, such as minimal crosstalk between pathways and high redundancy in STAT1-STAT3 heterodimerization	*H. Sapiens*	145

Table 1.1: **Methods to Probe Resource Allocation**

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Signal Transduction	Genome-Scale Modeling	Combines three molecular processes (signaling, transcriptional regulatory networks, and metabolism) which are connected by their respective inputs/outputs as well as metabolic intermediates. Developes integrated dynamic FBA (idFBA) to simulate reaction fluxes and resolve issues that arise due to mixed time-scales between reactions. idFBA treats reactions on fast time-scales as quasi steady-state, applying FBA at discrete time points; it treats slow reactions as instantaneous (steady-state) after some time-delay, incorporating it into the stoichiometric matrix.	Proof-of-concept integration of metabolism with signaling and transcription regulatory networks within the genome-scale modeling framework.	*S. cerevisiae*	146
Gene Expression	Mass Action ODEs	Estimate gene synthesis rates using ordinary differential equations modeling production (transcription/translation) and degradation of mRNA and protein and inputting experimentally measured turnover rates (using metabolic pulse labeling) and abundance.	Translation rates have a large overall contribution to protein abundance. Combinations of mRNA and protein turnover rates (high vs low) are enriched for specific cell processes largely associated with metabolism and cell maintenance.	*M. musculus*	64
Gene Expression	Mass Action ODEs		The extent each of mRNA abundance, translation rates, and protein degradation rates contributes to protein abundance changes with context (e.g., steady-state, LPS stimulation).	*M. musculus*	102

Table 1.1: **Methods to Probe Resource Allocation.**

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Gene Expression	Mass Action ODEs	Use gene synthesis ODEs to identify a simplified, two-dimensional Crick space for the investigation of the observed range of gene synthesis rates. Used these synthesis rates to derive meaningful quantitative relationships between transcription and translation rates at the boundary of the Crick space, expressions for the fitness cost of transcription, and the relationship between fitness costs and stochasticity within the Crick space.	There is a lack of genes with high transcription rates and low translation rates due to a precision (stochasticity) - economy (resource costs) tradeoff in gene expression.	*E. coli, S. cerevisiae, M. musculus, H. sapiens*	103
Gene Expression	Coarse-Graining/Phenomenological Modeling	Builds on gene synthesis ODEs by further accounting for resource loading of RNA polymerases in transcription and ribosomes in translation rather than assuming constant synthesis rates. Resource loading is accounted for by modeling the total available machinery as well as fraction of said machinery that should be allocated to a particular gene. The model can be adjusted to have the machinery or substrates be limiting. Coarse-graining is in relation to modeling synthesis as a function of the mass fraction allocated to a gene rather than gene-specific features.	By dynamically adjusting synthesis rates through accounting of resource loading, the model reproduces observations that under machinery-limiting conditions, protein and mRNA numbers grow proportionally to cell volume, maintaining concentration homeostasis under exponential cell growth	generalizable	209

Table 1.1: **Methods to Probe Resource Allocation.**

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Gene Expression	Coarse-Graining /Phenomenological Modeling	Measures growth rate and biomass composition across nutrient- and translational machinery-limiting conditions, fit parameters to the linear relationships between these variables, and used these variables to quantitatively reveal context-dependent resource allocation across coarse-grained proteome fractions (e.g., ribosomal fraction).	Under translation-limiting conditions, cells will re-allocate resources away from other proteome sectors and to ribosomal proteins in order to better support bioamss production.	*E. coli*	210
Gene Expression	Coarse-Graining /Phenomenological Modeling	Grouped proteins into coarse-grained sectors by clustering proteomics measured across differing resource-limiting conditions (anabolic, catabolic, and translational). Treated these sectors as "coarse-grained enzymes" and modeled how fluxes move through these enzymes to understand how they change with respect to each other (proteome re-allocation) as a function of growth rate in a context-dependent manner.	In general, the fraction of proteome sectors associated with a specific limitation increases with increasing limitation, indicating proteome re-allocation to the machinery enabling the processes effected by the limiting condition.	*E. coli*	162
Metabolism	Genome-Scale Modeling	Flux balance analysis (FBA) estimates optimal steady-state metabolomic fluxes from stoichiometric representations of metabolic networks and an objective function (one of the encoded reactions) using linear programming. The referenced reviews provides an overview of metabolic models (M-Models), FBA, and other methods to analyze M-models.	*Review Paper: Genome-scale modeling is a mathematical formulation of resource allocation that can also reveal mechanistic insights. Encoded pathway activity is constrainted by nutrient, energy, and machinery resource availability and fluxomics is predicted on the basis of cells diverting those resources to optimize for a specified objective or multiple objectives.	prokaryotes, eukaryotes	221,222

Table 1.1: Methods to Probe Resource Allocation

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Protein-Constrainted Metabolism	Genome-Scale Modeling	Parsimonious FBA (pFBA) is a two-step linear programming approach. The first step uses FBA to optimize the primary objective function. The second step sets a secondary objective of minimizing the sum of reaction fluxes, holding the optimal level of the primary objective constant. Conceptually, this secondary objective represents minimizing the total resource costs of metabolic activity while still optimizing the first objective.	GEM simulations are consistent with experimental data when predicting which machinery the cell uses to achieve growth.	*E. coli*	226
Protein-Constrained Metabolism	Genome-Scale Modeling	Constrained allocation flux balance analysis (CAFBA) accounts for machinery costs by constraining coarse-grained proteome sectors each metabolic reaction is associated with.	Proof-of-concept recapitulation of various findings, including the linear relationships between growth rate and ribosomal abundance as well as demonstration of overflow metabolism at high growth rates.	*E. coli*	161

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Protein-Constrained Metabolism	Genome-Scale Modeling	Enzyme cost minimization (ECM) implements an approach to estimate the enzyme abundance required to sustain a given reaction flux based on the optimality principle that machinery costs are minimized. The enzyme cost is proportional to the net flux generated per unit enzyme. To consider *in vivo* condition-specific effects, the cost function is not only dependent on the catalytic rate constant, but also on factors that may prevent the enzyme from operating at maximal kinetic efficiency (which may be estimated from metabolite levels). This includes substrate concentrations that do not saturate the enzyme and reaction thermodynamics (i.e., reverse fluxes) which constrain the feasible space of metabolite concentrations from the equilibrium binding constants. Finally, the function is scaled by a "burden" parameter that accounts for maintenance costs (e.g., molecular mass, misfolding, post-translational modifications, etc.).	Proof-of-concept demonstration that machinery cost minimization can accurately predict enzyme abundance associated with central carbon metabolism. Decomposing the factors that prevent enzymes from operating at maximal capacity *in vivo* into saturation, reversibility, and regulatory-based parameters enables determining which factors contribute to the accuracy of the prediction.	*E. coli*	73

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Protein-Constrained Metabolism	Genome-Scale Modeling	Max-min Driving Force (MDF) calculates the thermodynamic efficacy, or the ratio between the net flux and the total flux, of a given reaction as a function of pH, metabolite concentration, and the Gibbs free energy change. This metric is related to the kinetics of a reaction via the mass-action ratio and conceptually serves as a proxy for protein cost (i.e., a lower driving force requires a higher abundance of enzyme or higher catalytic rate per unit of forward flux). Using the optimality principle of enzyme cost minimization, MDF then uses linear programming to solve for the objective of maximizing the driving force throughout the metabolic network.	Analyzes central carbon metabolism to identify MDF bottlenecks and how pathways have evolved to resolve these bottlenecks via enzyme abundances, enzyme activity, and intermediate metabolite concentrations. Further uses MDF as an explanation for organisms choosing yield-inefficient pathways.	*E. coli*	74
Protein-Constrained Metabolism	Genome-Scale Modeling	GECKO directly constrains reaction fluxes with machinery resources by incorporating a simple enzyme production reaction (bounded by proteomics abundance) for catalysis of the respective metabolic reaction (coupled by a coefficient proportional to catalytic rate of enzyme), limiting reaction fluxes to a maximum according to first-principles	Ability to distinguish between nutrient- and machinery-limiting conditions and a reduced solution space (lower flux variability). Proof-of-concept demonstration of various findings with improved performance/accuracy over tradiation M-models, including the Crabtree effect and physiologically accurate growth rates.	*S. cerevisiae, H. Sapiens, multiple other organisms*	228,229

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Gene Expression	Genome-Scale Modeling	Reconstructed an expression network in an analgous manner to M-Model networks. Synthesis reactions are the analogue of metabolic reactions, macromolecules that are being synthesized are the analogue of metabolites, and gene expression machinery (e.g., RNA polymerase and ribosomes) are the analogue of metabolic enzymes, resulting in a genome-scale stoichiometric matrix of gene expression (E-matrix).	Proof-of-concept recapitulation of various findings including the relationship between ribosome abundance and growth rate and rRNA operon redundancy based on nutrient uptake rates.	*E. coli*	231
Gene Expression & Metabolism	Genome-Scale Modeling	ME-Models combine gene expression and metabolism into a coherent genome-scale model by explicitly encoding the gene expression reactions for each enzyme, then deriving "coupling constraints", or first principle quantitative relationships that link the enzyme products of expression reactions with the metabolic reactions they catalyze. Because coupling constraints are a function of growth, non-linear programs must be applied to find an optimal solution. Coupling gene expression with machinery allows for an explicit accounting of the machinery costs of metabolic activity and allows for variable RNA and protein content within the biomass composition.	*In silico* differentially expressed genes across different carbon sources yields hypothesis generation for the de-orphaning of L-arabinose transport.	*T. maritima*	232
Gene Expression & Metabolism	Genome-Scale Modeling		E. coli utilize low-yield carbon metabolism to divert machinery resources to the catalysis of enzymes not operating at their maximal capacity.	*E. coli*	233

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Gene Expression & Metabolism	Genome-Scale Modeling	Review of extensions and further applications of ME-Models.	*Review Paper	prokaryotes, human red blood cells	234
Gene Expression & Metabolism	Genome-Scale Modeling	ME-Model framework extended to eukaryotes with additional considerations of subcompartments (e.g., transport and compartment-specific proteome constraints), proteome sector constraints, and molecular crowding.	Proof-of-concept results recapitulating various findings or experimental data; a result specific to ME-Models is the ability to assess the effect of excess or unecessary protein production on growth rate.	*S. cerevisiae*	164
Gene Expression & Metabolism	Genome-Scale Modeling	Resource Balance Analysis (RBA) connects gene expression to metabolism at the genome-scale using similar approaches as the ME-Model framework while also accounting for spatial constraints.	*Review Paper	prokaryote, theory in eukaryotes	320
Metabolism & Secretory Pathway	Genome-Scale Modeling	Combines protein secretion and metabolism into a coherent genome-scale model by representing the secretory pathway in an analogous manner to metabolic networks inl M-Models. Secreted proteins are the analogue of metabolites and secretory machinery are the analogue of metabolic enzymes.	There is a negative correlation between secreted protein abundance and the energetic cost of producing that protein, and there exists tradeoffs between cell growth and protein secretion.	*C. griseus, M. musculus, H. sapiens*	157

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Secretory Pathway	Network Propagation	Applies a random-walk network propagation on a protein-protein interaction network between secretory pathway machinery and secreted proteins. Converts resultant transition probabilities to a "machinery support score" that indicates the extent to which machinery abundance aligns with production of a given secreted protein according to the network topology.	Changes to amyloid precursor protein (APP) abundance in Alzheimer's Disease, which cannot be explained at the transcriptional level, can be explained by how well APP is supported by secretory machinery abundance.	*H. sapiens*	51
All Intracellular Systems	Whole-Cell Modeling	Creates a separate module for each of 28 intracellular systems and applies a relevant modeling approach to each module. Next, integrates these modules into a single, coherent model by linking their common inputs and outputs, treating each module as independent at short time-scales, and dividing the shared resources across the modules at each time point.	The duration of replication initiation has high variability due to stochastic DnaA expression, but this variability is mitigated across the entire cell cycle due to decreased dNTP availability for polymerization over the course of the cell cycle.	*M. genitalium*	92
All Intracellular Systems	Whole-Cell Modeling	Accounts for 3D cell structure and comprehensively incorporates kinetic parameters across cell subsystems.	The total energy budget in the minimal cell at each time point is estimated to be similar to the total energy cost at that time point, indicating efficient energy resource allocation.	Synthetic minimal cell (JCVI-syn3A)	91
Multicellularity: Division of Labor	Tradeoffs/Pareto Analysis	Identified three relevant objective functions by aligning M-model predicted fluxomics with experimentally measured fluxomics. Computes the Pareto surface generated by the three objectives.	Experimentally measured fluxomics are near but not exactly on the Pareto front in order to minimizing the cost of switching between objectives.	*E. coli*	166

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Multicellularity: Division of Labor	Tradeoffs/Pareto Analysis	Pareto task inference (ParTI) estimates a low-dimensional polytope (n vertices fit using principal convex hull analysis) that encompasses omics datasets and represents Pareto optimal fronts in gene expression space. Such polytopes can be found in systems that perform multiple objectives and are Pareto optimal because, for any point outside the polytope, there exists a point inside it that can perform equally well or better at all tasks. The polytopes' vertices represent "archetypes" or gene expression profiles optimized for a particular objective. As samples move in gene expression space from one vertex to another, they undergo tradeoffs in their ability to perform each objective.	Generally, identified the biological objective of each archetype using enrichment analysis of the gene expression profiles of cells that clustered at the polytope vertices.	*M. musculus, H. sapiens*	158
Multicellularity: Division of Labor	Tradeoffs/Pareto Analysis		Generally, the continuum of gene expression profiles (in contrast to distinct clusters) found in some samples from single-cell sequencing technologies may be explained by varying extents of cell specialization, or the distance of a cell from the polytope vertex in gene expression space, within a Pareto optimal front. Less specialized cells perform multiple tasks whereas more specialized cells are optimized for a specific task.	*M. musculus, H. sapiens*	171

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	Approach	Method	Biological Finding	Organism	Reference
Multicellularity: Division of Labor	Tradeoffs/Pareto Analysis	Uses ParTI to further estimate tissue-level performance of each objective as well as overall tissue performance as a function of single-cell gene expression profiles. To quantify tissue-level function, individual cells' objective performance, based on their distance from polytope vertices, is aggregated across all cells and objectives.	By characterizing tissue-level performance as a mathematical function in gene expression space, this study is able to 1) calculate the optimal gene expression profile of individual cells, 2) characterize the expected behavior of cells to skew towards specialization and 3) incorporate spatial context gradients to accurately predict a continuum of expression profiles. This continuum is representative of a tissue-scale "division of labor" strategy in which individual cells tune their gene expression to perform one or more objectives based on their context, with the population-level outcome being optimized performance across all objectives.	*M. musculus*	169
Multicellularity: Cell Coordination	Genome-Scale Modeling	First, creates a context-specific M-model of brain tissue. Next, cell type specific M-models are built based on gene expression profiles. The models are also curated for consistency with known cell type specific metabolic functions. Finally, based on the expression of metabolite transporters, cell type specific models are joined by metabolites that can be transferred between models.	Model simulations focus on metabolic crosstalk between astrocytes and different neuron cell types and were used for three analyses to identify key genes and reactions, enriched metabolic pathways, and mechanisms of perturbation (e.g., drug treatment) associated with cell-type and context-specific responses of metabolism.	*H. sapiens*	182

Table 1.1: methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Multicellularity: Cell Coordination	Genome-Scale Modeling	Review of existing multi-cellular and multi-tissue M-models. Generally, a context-specific M-model of each cell type or tissue type is built and these models are connected to each other via a shared subcompartment and transport of shared metabolites. Just as standard M-models of a cell have compartments (e.g., nucleus, cytoplasm, mitochondria, etc.), individual cells or tissue types form the compartments of a single, large multi-cellular model.	*Review Paper	generalizable	180
Multicellularity: Cell Coordination	Genome-Scale Modeling	Rather than explicitly creating a multicellular M-model, this study experimentally measures metabolomics data of one cell type in the presence or absence of another cell type. It then uses this metabolomics data to constrain the M-models and compares the simulated fluxomics.	Used the M-models to investigate the metabolic effects of cancer-associated fiborblasts on colorectal cancer, observing activation or inhibition of a number of central carbon pathways such as glycolysis, glutaminolysis, and the TCA cycle.	*H. sapiens*	183
Multicellularity: Cell Coordination	Cell Circuits and Communication	Models two cell types (macrophages and fibroblasts) as a network that can exchange growth factors. The amount of each cell type was dependent on growth factor exchange, cell proliferation and death rates, and a limiting environmental resource upon which cell growth depends (space). Assesses the stability, or ratio of cell type quantities over time, of different network topologies.	Macrophage-fibroblast circuits achieve stability (compositional homeostasis) by exchange of growth factors PDGF and CSF1 and, necessarily, spatial constraints on the carrying capacity of fibroblasts. Generally, two-cell circuits can ahieve stability if one cell is environmentally limited in its carrying capacity and growth factor exchange occurs to regulate proliferation.	*M. musculus*	170

Table 1.1: Methods to Probe Resource Allocation.

Molecular Process	**Approach**	**Method**	**Biological Finding**	**Organism**	**Reference**
Multicellularity: Cell Coordination	Cell Circuits and Communication	Models the binding of a ligand to receptors, each associated with signaling pathways. Repeats this for multiple ligands and receptors and quantifies the cumulative signaling pathways' activity from the binding strength. Assesses the capability of these networks to activate cell types across a range of binding parameters and possible ligand-receptor combinations.	Relative to non-promiscuous interactions, promiscuous ligand-receptor binding enables few ligands to communicate specific signals to a wider range of cell types. This is achieved by utilizing specific combinations of multiple ligand-receptor pairs and is indicative of a strategy to minimize the resources required to synthesize communicatory macromolecules.	*M. musculus*	151

1.7 References

1. Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K. & Alon, U. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336,** 1157–1160 (2012).

2. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436,** 588–592 (2005).

3. Hofmeyr, J. S. & Cornish-Bowden, A. Regulating the cellular economy of supply and demand. *FEBS Lett.* **476,** 47–51 (2000).

4. Edelman, G. M. & Gally, J. A. Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 13763–13768 (2001).

5. Shakiba, N., Jones, R. D., Weiss, R. & Del Vecchio, D. Context-aware synthetic biology by controller design: Engineering the mammalian cell. *Cell Syst* **12,** 561–592 (2021).

6. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nat. Biotechnol.* (2022). doi:10.1038/s41587-022-01288-0

7. Armingol, E., Baghdassarian, H. M., Martino, C., Perez-Lopez, A., Aamodt, C., Knight, R. & Lewis, N. E. Context-aware deconvolution of cell-cell communication with Tensor-cell2cell. *Nat. Commun.* **13,** 3665 (2022).

8. Rooyackers, O. E., Adey, D. B., Ades, P. A. & Nair, K. S. Effect of age on in vivo rates of mitochondrial protein synthesis in human skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 15364–15369 (1996).

9. Gerashchenko, M. V., Peterfi, Z., Yim, S. H. & Gladyshev, V. N. Translation elongation rate varies among organs and decreases with age. *Nucleic Acids Res.* **49,** e9 (2021).

10. Ghosh, S., Körte, A., Serafini, G., Yadav, V. & Rodenfels, J. Developmental energetics: Energy expenditure, budgets and metabolism during animal embryogenesis. *Semin. Cell Dev. Biol.* (2022). doi:10.1016/j.semcdb.2022.03.009

11. Kleinridders, A., Ferris, H. A., Reyzer, M. L., Rath, M., Soto, M., Manier, M. L., Spraggins, J., Yang, Z., Stanton, R. C., Caprioli, R. M. & Kahn, C. R. Regional differences in brain glucose metabolism determined by imaging mass spectrometry. *Mol Metab* **12,** 113–121 (2018).

12. Ben-Moshe, S. & Itzkovitz, S. Spatial heterogeneity in the mammalian liver. *Nat. Rev. Gastroenterol. Hepatol.* **16,** 395–410 (2019).

13. Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., Herbst, R. H., Rogel, N., Slyper, M., Waldman, J., Sud, M., Andrews, E., Velonias, G., Haber, A. L., Jagadeesh, K., Vickovic, S., Yao, J., Stevens, C., Dionne, D., Nguyen, L. T., Villani, A.-C., Hofree, M., Creasey, E. A., Huang, H., Rozenblatt-Rosen, O., Garber, J. J., Khalili, H., Desch, A. N., Daly, M. J., Ananthakrishnan, A. N., Shalek, A. K., Xavier, R. J. & Regev, A. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178,** 714–730.e22 (2019).

14. Gazestani, V. H., Pramparo, T., Nalabolu, S., Kellman, B. P., Murray, S., Lopez, L., Pierce, K., Courchesne, E. & Lewis, N. E. A perturbed gene network containing PI3K-AKT, RAS-ERK and WNT-β-catenin pathways in leukocytes is linked to ASD genetics and symptom severity. *Nat. Neurosci.* **22,** 1624–1634 (2019).

15. Toda, S., Frankel, N. W. & Lim, W. A. Engineering cell-cell communication networks: programming multicellular behaviors. *Curr. Opin. Chem. Biol.* **52,** 31–38 (2019).

16. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell-cell communication through single-cell transcriptomics. *Curr Opin Syst Biol* **26,** 12–23 (2021).

17. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22,** 71–88 (2021).

18. Rueffler, C., Hermisson, J. & Wagner, G. P. Evolution of functional specialization and division of labor. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E326–35 (2012).

19. Hosios, A. M., Hecht, V. C., Danai, L. V., Johnson, M. O., Rathmell, J. C., Steinhauser, M. L., Manalis, S. R. & Vander Heiden, M. G. Amino Acids Rather than Glucose Account for the Majority of Cell Mass in Proliferating Mammalian Cells. *Dev. Cell* **36,** 540–549 (2016).

20. Fan, J., Kamphorst, J. J., Mathew, R., Chung, M. K., White, E., Shlomi, T. & Rabinowitz, J. D. Glutamine-driven oxidative phosphorylation is a major ATP source in transformed mammalian cells in both normoxia and hypoxia. *Mol. Syst. Biol.* **9,** 712 (2013).

21. Efeyan, A., Comb, W. C. & Sabatini, D. M. Nutrient-sensing mechanisms and pathways. *Nature* **517,** 302–310 (2015).

22. Bennett, N. K., Nguyen, M. K., Darch, M. A., Nakaoka, H. J., Cousineau, D., Ten Hoeve, J., Graeber, T. G., Schuelke, M., Maltepe, E., Kampmann, M., Mendelsohn, B. A., Nakamura, J. L. & Nakamura, K. Defining the ATPome reveals cross-optimization of metabolic pathways. *Nat. Commun.* **11,** 4319 (2020).

23. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).

24. Chaneton, B., Hillmann, P., Zheng, L., Martin, A. C. L., Maddocks, O. D. K., Chokkathukalam, A., Coyle, J. E., Jankevics, A., Holding, F. P., Vousden, K. H., Frezza, C., O'Reilly, M. & Gottlieb, E. Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature* **491,** 458–462 (2012).

25. Cantó, C., Jiang, L. Q., Deshmukh, A. S., Mataki, C., Coste, A., Lagouge, M., Zierath, J. R. & Auwerx, J. Interdependence of AMPK and SIRT1 for metabolic adaptation to fasting and exercise in skeletal muscle. *Cell Metab.* **11,** 213–219 (2010).

26. Zhu, J. & Thompson, C. B. Metabolic regulation of cell growth and proliferation. *Nat. Rev. Mol. Cell Biol.* **20,** 436–450 (2019).

27. Chantranupong, L., Wolfson, R. L. & Sabatini, D. M. Nutrient-sensing mechanisms across evolution. *Cell* **161,** 67–83 (2015).

28. Palm, W. & Thompson, C. B. Nutrient acquisition strategies of mammalian cells. *Nature* **546,** 234–242 (2017).

29. Fischer, B., Dieckmann, U. & Taborsky, B. When to store energy in a stochastic environment. *Evolution* **65,** 1221–1232 (2011).

30. Rathmell, J. C., Vander Heiden, M. G., Harris, M. H., Frauwirth, K. A. & Thompson, C. B. In the absence of extrinsic signals, nutrient utilization by lymphocytes is insufficient to maintain either cell size or viability. *Mol. Cell* **6,** 683–692 (2000).

31. Itoh, Y., Kawamata, Y., Harada, M., Kobayashi, M., Fujii, R., Fukusumi, S., Ogi, K., Hosoya, M., Tanaka, Y., Uejima, H., Tanaka, H., Maruyama, M., Satoh, R., Okubo, S., Kizawa, H., Komatsu, H., Matsumura, F., Noguchi, Y., Shinohara, T., Hinuma, S., Fujisawa, Y. & Fujino, M. Free fatty acids regulate insulin secretion from pancreatic beta cells through GPR40. *Nature* **422,** 173–176 (2003).

32. Zisman, A., Peroni, O. D., Abel, E. D., Michael, M. D., Mauvais-Jarvis, F., Lowell, B. B., Wojtaszewski, J. F., Hirshman, M. F., Virkamaki, A., Goodyear, L. J., Kahn, C. R. & Kahn, B. B. Targeted disruption of the glucose transporter 4 selectively in muscle causes insulin resistance and glucose intolerance. *Nat. Med.* **6,** 924–928 (2000).

33. Ma, E. H., Verway, M. J., Johnson, R. M., Roy, D. G., Steadman, M., Hayes, S., Williams, K. S., Sheldon, R. D., Samborska, B., Kosinski, P. A., Kim, H., Griss, T., Faubert, B., Condotta, S. A., Krawczyk, C. M., DeBerardinis, R. J., Stewart, K. M., Richer, M. J., Chubukov, V., Roddy, T. P. & Jones, R. G. Metabolic Profiling Using Stable Isotope Tracing Reveals Distinct Patterns of Glucose Utilization by Physiologically Activated CD8 T Cells. *Immunity* **51,** 856–870.e5 (2019).

34. Chen, Y., McConnell, B. O., Gayatri Dhara, V., Mukesh Naik, H., Li, C.-T., Antoniewicz, M. R. & Betenbaugh, M. J. An unconventional uptake rate objective function approach enhances applicability of genome-scale models for mammalian cells. *NPJ Syst. Biol. Appl.* **5,** 25 (2019).

35. Son, S., Stevens, M. M., Chao, H. X., Thoreen, C., Hosios, A. M., Schweitzer, L. D., Weng, Y., Wood, K., Sabatini, D., Vander Heiden, M. G. & Manalis, S. Cooperative nutrient accumulation sustains growth of mammalian cells. *Sci. Rep.* **5,** 17401 (2015).

36. da Silva Novaes, A., Borges, F. T., Maquigussa, E., Varela, V. A., Dias, M. V. S. & Boim, M. A. Influence of high glucose on mesangial cell-derived exosome composition, secretion and cell communication. *Sci. Rep.* **9,** 6270 (2019).

37. Hu, K. & Yu, Y. Metabolite availability as a window to view the early embryo microenvironment in vivo. *Mol. Reprod. Dev.* **84,** 1027–1038 (2017).

38. Long, L., Wei, J., Lim, S. A., Raynor, J. L., Shi, H., Connelly, J. P., Wang, H., Guy, C., Xie, B., Chapman, N. M., Fu, G., Wang, Y., Huang, H., Su, W., Saravia, J., Risch, I., Wang, Y.-D., Li, Y., Niu, M., Dhungana, Y., Kc, A., Zhou, P., Vogel, P., Yu, J., Pruett-Miller, S. M., Peng, J. & Chi, H. CRISPR screens unveil signal hubs for nutrient licensing of T cell immunity. *Nature* **600,** 308–313 (2021).

39. Cham, C. M. & Gajewski, T. F. Glucose availability regulates IFN-gamma production and p70S6 kinase activation in CD8+ effector T cells. *J. Immunol.* **174,** 4670–4677 (2005).

40. Cham, C. M., Driessens, G., O'Keefe, J. P. & Gajewski, T. F. Glucose deprivation inhibits multiple key gene expression events and effector functions in CD8+ T cells. *Eur. J. Immunol.* **38,** 2438–2450 (2008).

41. Carr, E. L., Kelman, A., Wu, G. S., Gopaul, R., Senkevitch, E., Aghvanyan, A., Turay, A. M. & Frauwirth, K. A. Glutamine uptake and metabolism are coordinately regulated by ERK/MAPK during T lymphocyte activation. *J. Immunol.* **185,** 1037–1044 (2010).

42. Khurana, P., Burudpakdee, C., Grupp, S. A., Beier, U. H., Barrett, D. M. & Bassiri, H. Distinct Bioenergetic Features of Human Invariant Natural Killer T Cells Enable Retained Functions in Nutrient-Deprived States. *Front. Immunol.* **12,** 700374 (2021).

43. Pantaleon, M., Scott, J. & Kaye, P. L. Nutrient sensing by the early mouse embryo: hexosamine biosynthesis and glucose signaling during preimplantation development. *Biol. Reprod.* **78,** 595–600 (2008).

44. Chi, F., Sharpley, M. S., Nagaraj, R., Roy, S. S. & Banerjee, U. Glycolysis-Independent Glucose Metabolism Distinguishes TE from ICM Fate during Mammalian Embryogenesis. *Dev. Cell* **53,** 9–26.e4 (2020).

45. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15690–15695 (2015).

46. Huang, H., Zhou, P., Wei, J., Long, L., Shi, H., Dhungana, Y., Chapman, N. M., Fu, G., Saravia, J., Raynor, J. L., Liu, S., Palacios, G., Wang, Y.-D., Qian, C., Yu, J. & Chi, H. In vivo CRISPR screening reveals nutrient signaling processes underpinning CD8+ T cell fate decisions. *Cell* **184,** 1245–1261.e21 (2021).

47. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* **425,** 3919–3936 (2013).

48. Haschemi, A., Kosma, P., Gille, L., Evans, C. R., Burant, C. F., Starkl, P., Knapp, B., Haas, R., Schmid, J. A., Jandl, C., Amir, S., Lubec, G., Park, J., Esterbauer, H., Bilban, M., Brizuela, L., Pospisilik, J. A., Otterbein, L. E. & Wagner, O. The sedoheptulose kinase CARKL directs macrophage polarization through control of glucose metabolism. *Cell Metab.* **15,** 813–826 (2012).

49. Lunt, S. Y., Muralidhar, V., Hosios, A. M., Israelsen, W. J., Gui, D. Y., Newhouse, L., Ogrodzinski, M., Hecht, V., Xu, K., Acevedo, P. N. M., Hollern, D. P., Bellinger, G., Dayton, T. L., Christen, S., Elia, I., Dinh, A. T., Stephanopoulos, G., Manalis, S. R., Yaffe, M. B., Andrechek, E. R., Fendt, S.-M. & Vander Heiden, M. G. Pyruvate kinase isoform expression alters nucleotide synthesis to impact cell proliferation. *Mol. Cell* **57,** 95–107 (2015).

50. Funk, L., Su, K.-C., Ly, J., Feldman, D., Singh, A., Moodie, B., Blainey, P. C. & Cheeseman, I. M. The phenotypic landscape of essential human genes. *Cell* (2022). doi:10.1016/j.cell.2022.10.017

51. Kuo, C.-C., Chiang, A. W. T., Baghdassarian, H. M. & Lewis, N. E. Dysregulation of the secretory pathway connects Alzheimer’s disease genetics to aggregate formation. *Cell Syst* **12,** 873–884.e4 (2021).

52. Reuveni, S., Ehrenberg, M. & Paulsson, J. Ribosomes are optimized for autocatalytic production. *Nature* **547,** 293–297 (2017).

53. Munding, E. M., Shiue, L., Katzman, S., Donohue, J. P. & Ares, M., Jr. Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. *Mol. Cell* **51,** 338–348 (2013).

54. Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V. & Abou Elela, S. Introns are mediators of cell response to starvation. *Nature* **565,** 612–617 (2019).

55. Jones, R. D., Qian, Y., Siciliano, V., DiAndreth, B., Huh, J., Weiss, R. & Del Vecchio, D. An endoribonuclease-based feedforward controller for decoupling resource-limited genetic modules in mammalian cells. *Nat. Commun.* **11,** 5690 (2020).

56. Yewdell, J. W. Not such a dismal science: the economics of protein synthesis, folding, degradation and antigen processing. *Trends Cell Biol.* **11,** 294–297 (2001).

57. Kafri, M., Metzl-Raz, E., Jona, G. & Barkai, N. The Cost of Protein Production. *Cell Rep.* **14,** 22–31 (2016).

58. Buttgereit, F. & Brand, M. D. A hierarchy of ATP-consuming processes in mammalian cells. *Biochem. J* **312 (Pt 1),** 163–167 (1995).

59. Milo, R. & Phillips, R. *Cell biology by the numbers*. (Garland Science, 2015). doi:10.1201/9780429258770

60. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–934 (2010).

61. Peth, A., Nathan, J. A. & Goldberg, A. L. The ATP costs and time required to degrade ubiquitinated proteins by the 26 S proteasome. *J. Biol. Chem.* **288,** 29215–29222 (2013).

62. Cambridge, S. B., Gnad, F., Nguyen, C., Bermejo, J. L., Krüger, M. & Mann, M. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J. Proteome Res.* **10,** 5275–5284 (2011).

63. Li, J., Cai, Z., Vaites, L. P., Shen, N., Mitchell, D. C., Huttlin, E. L., Paulo, J. A., Harry, B. L. & Gygi, S. P. Proteome-wide mapping of short-lived proteins in human cells. *Mol. Cell* **81,** 4722–4735.e5 (2021).

64. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

65. Marchingo, J. M. & Cantrell, D. A. Protein synthesis, degradation, and energy metabolism in T cell immunity. *Cell. Mol. Immunol.* **19,** 303–315 (2022).

66. Singh, R. & Cuervo, A. M. Autophagy in the cellular energetic balance. *Cell Metab.* **13,** 495–504 (2011).

67. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635 (2014).

68. Taggart, J. C. & Li, G.-W. Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes. *Cell Syst* **7,** 580–589.e4 (2018).

69. Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P. & Beck, M. Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biol.* **17,** 47 (2016).

70. Armingol, E., Tobar, E. & Cabrera, R. Understanding the impact of the cofactor swapping of isocitrate dehydrogenase over the growth phenotype of Escherichia coli on acetate by using constraint-based modeling. *PLoS One* **13,** e0196182 (2018).

71. Goldford, J. E., George, A. B., Flamholz, A. I. & Segrè, D. Protein cost minimization promotes the emergence of coenzyme redundancy. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2110787119 (2022).

72. Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., Weinberger, A., Alon, U., Milo, R. & Segal, E. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell* **166,** 1282–1294.e18 (2016).

73. Noor, E., Flamholz, A., Bar-Even, A., Davidi, D., Milo, R. & Liebermeister, W. The Protein Cost of Metabolic Fluxes: Prediction from Enzymatic Rate Laws and Cost Minimization. *PLoS Comput. Biol.* **12,** e1005167 (2016).

74. Noor, E., Bar-Even, A., Flamholz, A., Reznik, E., Liebermeister, W. & Milo, R. Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS Comput. Biol.* **10,** e1003483 (2014).

75. Yang, X., Heinemann, M., Howard, J., Huber, G., Iyer-Biswas, S., Le Treut, G., Lynch, M., Montooth, K. L., Needleman, D. J., Pigolotti, S., Rodenfels, J., Ronceray, P., Shankar, S., Tavassoly, I., Thutupalli, S., Titov, D. V., Wang, J. & Foster, P. J. Physical bioenergetics: Energy fluxes, budgets, and constraints in cells. *Proc. Natl. Acad. Sci. U. S. A.* **118,** (2021).

76. Rolfe, D. F. & Brown, G. C. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiol. Rev.* **77,** 731–758 (1997).

77. Schmidt, C. A., Fisher-Wellman, K. H. & Neufer, P. D. From OCR and ECAR to energy: Perspectives on the design and interpretation of bioenergetics studies. *J. Biol. Chem.* **297,** 101140 (2021).

78. van Helvert, S., Storm, C. & Friedl, P. Mechanoreciprocity in cell migration. *Nat. Cell Biol.* **20,** 8–20 (2018).

79. Mosier, J. A., Wu, Y. & Reinhart-King, C. A. Recent advances in understanding the role of metabolic heterogeneities in cell migration. *Fac Rev* **10,** 8 (2021).

80. Li, Y., Yao, L., Mori, Y. & Sun, S. X. On the energy efficiency of cell migration in diverse physical environments. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 23894–23900 (2019).

81. Zanotelli, M. R., Rahman-Zaman, A., VanderBurgh, J. A., Taufalele, P. V., Jain, A., Erickson, D., Bordeleau, F. & Reinhart-King, C. A. Energetic costs regulated by cell mechanics and confinement are predictive of migration path during decision-making. *Nat. Commun.* **10,** 4185 (2019).

82. Attwell, D. & Laughlin, S. B. An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab.* **21,** 1133–1145 (2001).

83. Du, F., Zhu, X.-H., Zhang, Y., Friedman, M., Zhang, N., Ugurbil, K. & Chen, W. Tightly coupled brain activity and cerebral ATP metabolic rate. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 6409–6414 (2008).

84. Pulido, C. & Ryan, T. A. Synaptic vesicle pools are a major hidden resting metabolic burden of nerve terminals. *Sci Adv* **7,** eabi9027 (2021).

85. Wang, Z., Ying, Z., Bosy-Westphal, A., Zhang, J., Schautz, B., Later, W., Heymsfield, S. B. & Müller, M. J. Specific metabolic rates of major organs and tissues across adulthood: evaluation by mechanistic model of resting energy expenditure. *Am. J. Clin. Nutr.* **92,** 1369–1377 (2010).

86. Hasenstaub, A., Otte, S., Callaway, E. & Sejnowski, T. J. Metabolic cost as a unifying principle governing neuronal biophysics. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 12329–12334 (2010).

87. Alle, H., Roth, A. & Geiger, J. R. P. Energy-efficient action potentials in hippocampal mossy fibers. *Science* **325,** 1405–1408 (2009).

88. Herculano-Houzel, S. Scaling of brain metabolism with a fixed energy budget per neuron: implications for neuronal activity, plasticity and evolution. *PLoS One* **6,** e17514 (2011).

89. Lennie, P. The cost of cortical computation. *Curr. Biol.* **13,** 493–497 (2003).

90. Levy, W. B. & Baxter, R. A. Energy efficient neural codes. *Neural Comput.* **8,** 531–543 (1996).

91. Thornburg, Z. R., Bianchi, D. M., Brier, T. A., Gilbert, B. R., Earnest, T. M., Melo, M. C. R., Safronova, N., Sáenz, J. P., Cook, A. T., Wise, K. S., Hutchison, C. A., 3rd, Smith, H. O., Glass, J. I. & Luthey-Schulten, Z. Fundamental behaviors emerge from simulations of a living minimal cell. *Cell* **185,** 345–360.e28 (2022).

92. Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr, Assad-Garcia, N., Glass, J. I. & Covert, M. W. A whole-cell computational model predicts phenotype from genotype. *Cell* **150,** 389–401 (2012).

93. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21,** 630–644 (2020).

94. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13,** 227–232 (2012).

95. Edfors, F., Danielsson, F., Hallström, B. M., Käll, L., Lundberg, E., Pontén, F., Forsström, B. & Uhlén, M. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **12,** 883 (2016).

96. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

97. Vogel, C., Abreu, R. de S., Ko, D., Le, S.-Y., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R., Marcotte, E. M. & Penalva, L. O. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6,** 400 (2010).

98. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L.-H., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F.,

Uhlen, M., Gagneur, J., Hahne, H. & Kuster, B. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15,** e8503 (2019).

99. Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.* **11,** e1005206 (2015).

100. Ding, F. & Elowitz, M. B. Constitutive splicing and economies of scale in gene expression. *Nat. Struct. Mol. Biol.* **26,** 424–432 (2019).

101. Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. & Regev, A. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29,** 436–442 (2011).

102. Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E. H., Fields, A. P., Schwartz, S., Raychowdhury, R., Mumbach, M. R., Eisenhaure, T., Rabani, M., Gennert, D., Lu, D., Delorey, T., Weissman, J. S., Carr, S. A., Hacohen, N. & Regev, A. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347,** 1259038 (2015).

103. Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat. Commun.* **10,** 68 (2019).

104. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165,** 535–550 (2016).

105. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537,** 328–338 (2016).

106. Hukelmann, J. L., Anderson, K. E., Sinclair, L. V., Grzes, K. M., Murillo, A. B., Hawkins, P. T., Stephens, L. R., Lamond, A. I. & Cantrell, D. A. The cytotoxic T cell proteome and its shaping by the kinase mTOR. *Nat. Immunol.* **17,** 104–112 (2016).

107. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* **7,** 481 (2011).

108. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I. & Pilpel, Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141,** 344–354 (2010).

109. Al-Bassam, M. M., Kim, J.-N., Zaramela, L. S., Kellman, B. P., Zuniga, C., Wozniak, J. M., Gonzalez, D. J. & Zengler, K. Optimization of carbon and energy utilization through differential translational efficiency. *Nat. Commun.* **9,** 4474 (2018).

110. Basan, M. Resource allocation and metabolism: the search for governing principles. *Curr. Opin. Microbiol.* **45,** 77–83 (2018).

111. Soflaee, M. H., Kesavan, R., Sahu, U., Tasdogan, A., Villa, E., Djabari, Z., Cai, F., Tran, D. H., Vu, H. S., Ali, E. S., Rion, H., O'Hara, B. P., Kelekar, S., Hallett, J. H., Martin, M., Mathews, T. P., Gao, P., Asara, J. M., Manning, B. D., Ben-Sahra, I. & Hoxhaj, G. Purine nucleotide depletion prompts cell migration by stimulating the serine synthesis pathway. *Nat. Commun.* **13,** 2698 (2022).

112. Kiweler, N., Delbrouck, C., Pozdeev, V. I., Neises, L., Soriano-Baguet, L., Eiden, K., Xian, F., Benzarti, M., Haase, L., Koncina, E., Schmoetten, M., Jaeger, C., Noman, M. Z., Vazquez, A., Janji, B., Dittmar, G., Brenner, D., Letellier, E. & Meiser, J. Mitochondria preserve an autarkic one-carbon cycle to confer growth-independent cancer cell migration and metastasis. *Nat. Commun.* **13,** 2699 (2022).

113. Carey, B. W., Finley, L. W. S., Cross, J. R., Allis, C. D. & Thompson, C. B. Intracellular α-ketoglutarate maintains the pluripotency of embryonic stem cells. *Nature* **518,** 413–416 (2015).

114. Vannini, N., Girotra, M., Naveiras, O., Nikitin, G., Campos, V., Giger, S., Roch, A., Auwerx, J. & Lutolf, M. P. Specification of haematopoietic stem cell fate via modulation of mitochondrial activity. *Nat. Commun.* **7,** 13125 (2016).

115. Schell, J. C., Wisidagama, D. R., Bensard, C., Zhao, H., Wei, P., Tanner, J., Flores, A., Mohlman, J., Sorensen, L. K., Earl, C. S., Olson, K. A., Miao, R., Waller, T. C., Delker, D., Kanth, P., Jiang, L., DeBerardinis, R. J., Bronner, M. P., Li, D. Y., Cox, J. E., Christofk, H. R., Lowry, W. E., Thummel, C. S. & Rutter, J. Control of intestinal stem cell function and proliferation by mitochondrial pyruvate metabolism. *Nat. Cell Biol.* **19,** 1027–1036 (2017).

116. Nilsson, A., Nielsen, J. & Palsson, B. O. Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst* **5,** 538–541 (2017).

117. Davidi, D. & Milo, R. Lessons on enzyme kinetics from quantitative proteomics. *Curr. Opin. Biotechnol.* **46,** 81–89 (2017).

118. Schuster, S., Schuster, R. & Heinrich, R. Minimization of intermediate concentrations as a suggested optimality principle for biochemical networks. II. Time hierarchy, enzymatic rate laws, and erythrocyte metabolism. *J. Math. Biol.* **29,** 443–455 (1991).

119. Tepper, N., Noor, E., Amador-Noguez, D., Haraldsdóttir, H. S., Milo, R., Rabinowitz, J., Liebermeister, W. & Shlomi, T. Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. *PLoS One* **8,** e75370 (2013).

120. Beard, D. A. & Qian, H. Relationship between thermodynamic driving force and one-way fluxes in reversible processes. *PLoS One* **2,** e144 (2007).

121. Davidi, D., Noor, E., Liebermeister, W., Bar-Even, A., Flamholz, A., Tummler, K., Barenholz, U., Goldenfeld, M., Shlomi, T. & Milo, R. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3401–3406 (2016).

122. Xia, J., Sánchez, B. J., Chen, Y., Campbell, K., Kasvandik, S. & Nielsen, J. Proteome allocations change linearly with the specific growth rate of Saccharomyces cerevisiae under glucose limitation. *Nat. Commun.* **13,** 2819 (2022).

123. Shvartsman, S. Y., Hagan, M. P., Yacoub, A., Dent, P., Wiley, H. S. & Lauffenburger, D. A. Autocrine loops with positive feedback enable context-dependent cell signaling. *Am. J. Physiol. Cell Physiol.* **282,** C545–59 (2002).

124. Farhan, H. & Rabouille, C. Signalling to and from the secretory pathway. *J. Cell Sci.* **124,** 171–180 (2011).

125. Le Bihan, M.-C., Bigot, A., Jensen, S. S., Dennis, J. L., Rogowska-Wrzesinska, A., Lainé, J., Gache, V., Furling, D., Jensen, O. N., Voit, T., Mouly, V., Coulton, G. R. & Butler-Browne, G. In-depth analysis of the secretome identifies three major independent secretory pathways in differentiating human myoblasts. *J. Proteomics* **77,** 344–356 (2012).

126. Wegrzyn, J. L., Bark, S. J., Funkelstein, L., Mosier, C., Yap, A., Kazemi-Esfarjani, P., La Spada, A. R., Sigurdson, C., O'Connor, D. T. & Hook, V. Proteomics of dense core secretory vesicles reveal distinct protein categories for secretion of neuroeffectors for cell-cell communication. *J. Proteome Res.* **9,** 5002–5024 (2010).

127. Francis, K. & Palsson, B. O. Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 12258–12262 (1997).

128. Reefman, E., Kay, J. G., Wood, S. M., Offenhäuser, C., Brown, D. L., Roy, S., Stanley, A. C., Low, P. C., Manderson, A. P. & Stow, J. L. Cytokine secretion is distinct from secretion of cytotoxic granules in NK cells. *J. Immunol.* **184,** 4852–4862 (2010).

129. Lopez, J. A., Brennan, A. J., Whisstock, J. C., Voskoboinik, I. & Trapani, J. A. Protecting a serial killer: pathways for perforin trafficking and self-defence ensure sequential target cell death. *Trends Immunol.* **33,** 406–412 (2012).

130. Bonnans, C., Chou, J. & Werb, Z. Remodelling the extracellular matrix in development and disease. *Nat. Rev. Mol. Cell Biol.* **15,** 786–801 (2014).

131. Shen, Y., Zhou, M., Cai, D., Filho, D. A., Fernandes, G., Cai, Y., de Sousa, A. F., Tian, M., Kim, N., Lee, J., Necula, D., Zhou, C., Li, S., Salinas, S., Liu, A., Kang, X., Kamata, M., Lavi, A., Huang, S., Silva, T., Do Heo, W. & Silva, A. J. CCR5 closes the temporal window for memory linking. *Nature* **606,** 146–152 (2022).

132. Hill, S. M., Nesser, N. K., Johnson-Camacho, K., Jeffress, M., Johnson, A., Boniface, C., Spencer, S. E. F., Lu, Y., Heiser, L. M., Lawrence, Y., Pande, N. T., Korkola, J. E., Gray, J. W., Mills, G. B., Mukherjee, S. & Spellman, P. T. Context Specificity in Causal Signaling Networks Revealed by Phosphoprotein Profiling. *Cell Syst* **4,** 73–83.e10 (2017).

133. Larson, R. C., Kann, M. C., Bailey, S. R., Haradhvala, N. J., Llopis, P. M., Bouffard, A. A., Scarfó, I., Leick, M. B., Grauwet, K., Berger, T. R., Stewart, K., Anekal, P. V., Jan, M., Joung, J., Schmidts, A., Ouspenskaia, T., Law, T., Regev, A., Getz, G. & Maus, M. V. CAR T cell killing requires the IFNγR pathway in solid but not liquid tumours. *Nature* **604,** 563–570 (2022).

134. Klumpe, H. E., Langley, M. A., Linton, J. M., Su, C. J., Antebi, Y. E. & Elowitz, M. B. The context-dependent, combinatorial logic of BMP signaling. *Cell Syst* **13,** 388–407.e10 (2022).

135. Lestas, I., Vinnicombe, G. & Paulsson, J. Fundamental limits on the suppression of molecular fluctuations. *Nature* **467,** 174–178 (2010).

136. Alves, R., Salvadó, B., Milo, R., Vilaprinyo, E. & Sorribas, A. Maximization of information transmission influences selection of native phosphorelay architectures. *PeerJ* **9,** e11558 (2021).

137. Wang, T.-L., Kuznets-Speck, B., Broderick, J. & Hinczewski, M. The price of a bit: energetic costs and the evolution of cellular signaling. *bioRxiv* 2020.10.06.327700 (2022). doi:10.1101/2020.10.06.327700

138. Mehta, P. & Schwab, D. J. Energetic costs of cellular computation. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 17978–17982 (2012).

139. Lan, G., Sartori, P., Neumann, S., Sourjik, V. & Tu, Y. The energy-speed-accuracy tradeoff in sensory adaptation. *Nat. Phys.* **8,** 422–428 (2012).

140. Govern, C. C. & Ten Wolde, P. R. Optimal resource allocation in cellular sensing systems. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 17486–17491 (2014).

141. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8,** 450–461 (2007).

142. Adler, M., Szekely, P., Mayo, A. & Alon, U. Optimal Regulatory Circuit Topologies for Fold-Change Detection. *Cell Syst* **4,** 171–181.e8 (2017).

143. Schilling, C. H., Letscher, D. & Palsson, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203,** 229–248 (2000).

144. Papin, J. A. & Palsson, B. O. Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.* **227,** 283–297 (2004).

145. Papin, J. A. & Palsson, B. O. The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* **87,** 37–46 (2004).

146. Lee, J. M., Gianchandani, E. P., Eddy, J. A. & Papin, J. A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput. Biol.* **4,** e1000086 (2008).

147. Barton, J. P. & Sontag, E. D. The energy costs of insulators in biochemical networks. *Biophys. J.* **104,** 1380–1390 (2013).

148. Saez-Rodriguez, J., Kremling, A. & Gilles, E. D. Dissecting the puzzle of life: modularization of signal transduction networks. *Comput. Chem. Eng.* **29,** 619–629 (2005).

149. Cheong, R., Rhee, A., Wang, C. J., Nemenman, I. & Levchenko, A. Information transduction capacity of noisy biochemical signaling networks. *Science* **334,** 354–358 (2011).

150. Youk, H. & Lim, W. A. Secreting and sensing the same molecule allows cells to achieve versatile social behaviors. *Science* **343,** 1242782 (2014).

151. Su, C. J., Murugan, A., Linton, J. M., Yeluri, A., Bois, J., Klumpe, H., Langley, M. A., Antebi, Y. E. & Elowitz, M. B. Ligand-receptor promiscuity enables cellular addressing. *Cell Syst* **13,** 408–425.e12 (2022).

152. Alon, U. in *An Introduction to Systems Biology* 249–272 (Chapman and Hall/CRC, 2019).

153. Gyorgy, A., Jiménez, J. I., Yazbek, J., Huang, H.-H., Chung, H., Weiss, R. & Del Vecchio, D. Isocost Lines Describe the Cellular Economy of Genetic Circuits. *Biophys. J.* **109,** 639–646 (2015).

154. Alonso, R., Brocas, I. & Carrillo, J. D. Resource Allocation in the Brain. *Rev. Econ. Stud.* **81,** 501–534 (2013).

155. Szekely, P., Sheftel, H., Mayo, A. & Alon, U. Evolutionary tradeoffs between economy and effectiveness in biological homeostasis systems. *PLoS Comput. Biol.* **9,** e1003163 (2013).

156. Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A. & Yarmush, M. L. Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. *Ann. Biomed. Eng.* **35,** 863–885 (2007).

157. Gutierrez, J. M., Feizi, A., Li, S., Kallehauge, T. B., Hefzi, H., Grav, L. M., Ley, D., Baycin Hizal, D., Betenbaugh, M. J., Voldborg, B., Faustrup Kildegaard, H., Min Lee, G., Palsson, B. O., Nielsen, J. & Lewis, N. E. Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nat. Commun.* **11,** 68 (2020).

158. Hart, Y., Sheftel, H., Hausser, J., Szekely, P., Ben-Moshe, N. B., Korem, Y., Tendler, A., Mayo, A. E. & Alon, U. Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods* **12,** 233–5, 3 p following 235 (2015).

159. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).

160. Bulusu, V., Prior, N., Snaebjornsson, M. T., Kuehne, A., Sonnen, K. F., Kress, J., Stein, F., Schultz, C., Sauer, U. & Aulehla, A. Spatiotemporal Analysis of a Glycolytic Activity Gradient Linked to Mouse Embryo Mesoderm Development. *Dev. Cell* **40,** 331–341.e4 (2017).

161. Mori, M., Hwa, T., Martin, O. C., De Martino, A. & Marinari, E. Constrained Allocation Flux Balance Analysis. *PLoS Comput. Biol.* **12,** e1004913 (2016).

162. Hui, S., Silverman, J. M., Chen, S. S., Erickson, D. W., Basan, M., Wang, J., Hwa, T. & Williamson, J. R. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Mol. Syst. Biol.* **11,** 784 (2015).

163. Yang, L., Yurkovich, J. T., Lloyd, C. J., Ebrahim, A., Saunders, M. A. & Palsson, B. O. Principles of proteome allocation are revealed using proteomic data and genome-scale models. *Sci. Rep.* **6,** 36734 (2016).

164. Elsemman, I. E., Rodriguez Prado, A., Grigaitis, P., Garcia Albornoz, M., Harman, V., Holman, S. W., van Heerden, J., Bruggeman, F. J., Bisschops, M. M. M., Sonnenschein, N., Hubbard, S., Beynon, R., Daran-Lapujade, P., Nielsen, J. & Teusink, B. Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies. *Nat. Commun.* **13,** 801 (2022).

165. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J. & Aebersold, R. The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549 (2011).

166. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336,** 601–604 (2012).

167. Bonny, A. R., Kochanowski, K., Diether, M. & El-Samad, H. Stress-Induced Transient Cell Cycle Arrest Coordinates Metabolic Resource Allocation to Balance Adaptive Tradeoffs. *bioRxiv* 2020.04.08.033035 (2020). doi:10.1101/2020.04.08.033035

168. Schink, S. J., Christodoulou, D., Mukherjee, A., Athaide, E., Brunner, V., Fuhrer, T., Bradshaw, G. A., Sauer, U. & Basan, M. Glycolysis/gluconeogenesis specialization in microbes is driven by biochemical constraints of flux sensing. *Mol. Syst. Biol.* **18,** e10704 (2022).

169. Adler, M., Korem Kohanim, Y., Tendler, A., Mayo, A. & Alon, U. Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. *Cell Syst* **8,** 43–52.e5 (2019).

170. Zhou, X., Franklin, R. A., Adler, M., Jacox, J. B., Bailis, W., Shyer, J. A., Flavell, R. A., Mayo, A., Alon, U. & Medzhitov, R. Circuit Design Features of a Stable Two-Cell System. *Cell* **172,** 744–757.e17 (2018).

171. Korem, Y., Szekely, P., Hart, Y., Sheftel, H., Hausser, J., Mayo, A., Rothenberg, M. E., Kalisky, T. & Alon, U. Geometry of the Gene Expression Space of Individual Cells. *PLoS Comput. Biol.* **11,** e1004224 (2015).

172. Goldsby, H. J., Dornhaus, A., Kerr, B. & Ofria, C. Task-switching costs promote the evolution of division of labor and shifts in individuality. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 13686–13691 (2012).

173. Rodríguez-Caso, C. Can cell mortality determine division of labor in tissue organization? *J. Theor. Biol.* **332,** 161–170 (2013).

174. Halpern, K. B., Shenhav, R., Matcovitch-Natan, O., Toth, B., Lemze, D., Golan, M., Massasa, E. E., Baydatch, S., Landen, S., Moor, A. E., Brandis, A., Giladi, A., Avihail, A. S., David, E., Amit, I. & Itzkovitz, S. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542,** 352–356 (2017).

175. Adler, M., Moriel, N., Goeva, A., Avraham-Davidi, I., Mages, S., Adams, T. S., Kaminski, N., Macosko, E. Z., Regev, A., Medzhitov, R. & Nitzan, M. Emergence of division of labor in tissues through cell interactions and spatial cues. *bioRxiv* 2022.11.16.516540 (2022). doi:10.1101/2022.11.16.516540

176. Brückner, A., Badroos, J. M., Learsch, R. W., Yousefelahiyeh, M., Kitchen, S. A. & Parker, J. Evolutionary assembly of cooperating cell types in an animal chemical defense system. *Cell* **184,** 6138–6156.e28 (2021).

177. Handly, L. N., Pilko, A. & Wollman, R. Paracrine communication maximizes cellular response fidelity in wound signaling. *Elife* **4,** e09652 (2015).

178. Adler, M., Mayo, A., Zhou, X., Franklin, R. A., Jacox, J. B., Medzhitov, R. & Alon, U. Endocytosis as a stabilizing mechanism for tissue homeostasis. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E1926–E1935 (2018).

179. Zhou, X., Franklin, R. A., Adler, M., Carter, T. S., Condiff, E., Adams, T. S., Pope, S. D., Philip, N. H., Meizlish, M. L., Kaminski, N. & Medzhitov, R. Microenvironmental Sensing by Fibroblasts Controls Macrophage Population Size. *bioRxiv* 2022.01.18.476683 (2022). doi:10.1101/2022.01.18.476683

180. Martins Conde, P. do R., Sauter, T. & Pfau, T. Constraint Based Modeling Going Multicellular. *Front Mol Biosci* **3,** 3 (2016).

181. Gustafsson, J., Robinson, J. L., Zetterberg, H. & Nielsen, J. Brain energy metabolism is optimized to minimize the cost of enzyme synthesis and transport. *bioRxiv* 2022.11.14.516523 (2022). doi:10.1101/2022.11.14.516523

182. Lewis, N. E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M. P., Cheng, J. K., Patel, N., Yee, A., Lewis, R. A., Eils, R., König, R. & Palsson, B. Ø. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.* **28,** 1279–1285 (2010).

183. Wang, J., Delfarah, A., Gelbach, P. E., Fong, E., Macklin, P., Mumenthaler, S. M., Graham, N. A. & Finley, S. D. Elucidating tumor-stromal metabolic crosstalk in colorectal cancer through integration of constraint-based models and LC-MS metabolomics. *Metab. Eng.* **69,** 175–187 (2022).

184. Aceto, N., Bardia, A., Miyamoto, D. T., Donaldson, M. C., Wittner, B. S., Spencer, J. A., Yu, M., Pely, A., Engstrom, A., Zhu, H., Brannigan, B. W., Kapur, R., Stott, S. L., Shioda, T., Ramaswamy, S., Ting, D. T., Lin, C. P., Toner, M., Haber, D. A. & Maheswaran, S. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158,** 1110–1122 (2014).

185. Zhang, J., Goliwas, K. F., Wang, W., Taufalele, P. V., Bordeleau, F. & Reinhart-King, C. A. Energetic regulation of coordinated leader-follower dynamics during collective invasion of breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 7867–7872 (2019).

186. Liu, L., Duclos, G., Sun, B., Lee, J., Wu, A., Kam, Y., Sontag, E. D., Stone, H. A., Sturm, J. C., Gatenby, R. A. & Austin, R. H. Minimization of thermodynamic costs in cancer cell invasion. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 1686–1691 (2013).

187. Smiley, P. & Levin, M. Competition for finite resources as coordination mechanism for morphogenesis: An evolutionary algorithm study of digital embryogeny. *Biosystems.* **221,** 104762 (2022).

188. Baker, N. E. Emerging mechanisms of cell competition. *Nat. Rev. Genet.* **21,** 683–697 (2020).

189. Matamoro-Vidal, A. & Levayer, R. Multiple Influences of Mechanical Forces on Cell Competition. *Curr. Biol.* **29,** R762–R774 (2019).

190. Clavería, C., Giovinazzo, G., Sierra, R. & Torres, M. Myc-driven endogenous cell competition in the early mammalian embryo. *Nature* **500,** 39–44 (2013).

191. Ellis, S. J., Gomez, N. C., Levorse, J., Mertz, A. F., Ge, Y. & Fuchs, E. Distinct modes of cell competition shape mammalian tissue morphogenesis. *Nature* **569,** 497–502 (2019).

192. Kon, S., Ishibashi, K., Katoh, H., Kitamoto, S., Shirai, T., Tanaka, S., Kajita, M., Ishikawa, S., Yamauchi, H., Yako, Y., Kamasaki, T., Matsumoto, T., Watanabe, H., Egami, R., Sasaki, A., Nishikawa, A., Kameda, I., Maruyama, T., Narumi, R., Morita, T., Sasaki, Y., Enoki, R., Honma, S., Imamura, H., Oshima, M., Soga, T., Miyazaki, J.-I., Duchen, M. R., Nam, J.-M., Onodera, Y., Yoshioka, S., Kikuta, J., Ishii, M., Imajo, M., Nishida, E., Fujioka, Y., Ohba, Y.,

Sato, T. & Fujita, Y. Cell competition with normal epithelial cells promotes apical extrusion of transformed cells through metabolic changes. *Nat. Cell Biol.* **19,** 530–541 (2017).

193. Chang, C.-H., Qiu, J., O'Sullivan, D., Buck, M. D., Noguchi, T., Curtis, J. D., Chen, Q., Gindin, M., Gubin, M. M., van der Windt, G. J. W., Tonc, E., Schreiber, R. D., Pearce, E. J. & Pearce, E. L. Metabolic Competition in the Tumor Microenvironment Is a Driver of Cancer Progression. *Cell* **162,** 1229–1241 (2015).

194. Brand, A., Singer, K., Koehl, G. E., Kolitzus, M., Schoenhammer, G., Thiel, A., Matos, C., Bruss, C., Klobuch, S., Peter, K., Kastenberger, M., Bogdan, C., Schleicher, U., Mackensen, A., Ullrich, E., Fichtner-Feigl, S., Kesselring, R., Mack, M., Ritter, U., Schmid, M., Blank, C., Dettmer, K., Oefner, P. J., Hoffmann, P., Walenta, S., Geissler, E. K., Pouyssegur, J., Villunger, A., Steven, A., Seliger, B., Schreml, S., Haferkamp, S., Kohl, E., Karrer, S., Berneburg, M., Herr, W., Mueller-Klieser, W., Renner, K. & Kreutz, M. LDHA-Associated Lactic Acid Production Blunts Tumor Immunosurveillance by T and NK Cells. *Cell Metab.* **24,** 657–671 (2016).

195. Reinfeld, B. I., Madden, M. Z., Wolf, M. M., Chytil, A., Bader, J. E., Patterson, A. R., Sugiura, A., Cohen, A. S., Ali, A., Do, B. T., Muir, A., Lewis, C. A., Hongo, R. A., Young, K. L., Brown, R. E., Todd, V. M., Huffstater, T., Abraham, A., O'Neil, R. T., Wilson, M. H., Xin, F., Tantawy, M. N., Merryman, W. D., Johnson, R. W., Williams, C. S., Mason, E. F., Mason, F. M., Beckermann, K. E., Vander Heiden, M. G., Manning, H. C., Rathmell, J. C. & Rathmell, W. K. Cell-programmed nutrient partitioning in the tumour microenvironment. *Nature* **593,** 282–288 (2021).

196. Nguyen, H. P., Sheng, R., Murray, E., Ito, Y., Bruck, M., Biellak, C., An, K., Lynce, F., Dillon, D. A., Magbanua, M. J. M., Huppert, L. A., Hammerlindl, H., Esserman, L., Rosenbluth, J. M. & Ahituv, N. Implantation of engineered adipocytes that outcompete tumors for resources suppresses cancer progression. *bioRxiv* 2023.03.28.534564 (2023). doi:10.1101/2023.03.28.534564

197. Seki, T., Yang, Y., Sun, X., Lim, S., Xie, S., Guo, Z., Xiong, W., Kuroda, M., Sakaue, H., Hosaka, K., Jing, X., Yoshihara, M., Qu, L., Li, X., Chen, Y. & Cao, Y. Brown-fat-mediated tumour suppression by cold-altered global metabolism. *Nature* **608,** 421–428 (2022).

198. Nagle, M. P., Tam, G. S., Maltz, E., Hemminger, Z. & Wollman, R. Bridging scales: From cell biology to physiology using in situ single-cell technologies. *Cell Syst* **12,** 388–400 (2021).

199. Mah, C. K., Ahmed, N., Lam, D., Monell, A., Kern, C., Han, Y., Cesnik, A. J., Lundberg, E., Zhu, Q., Carter, H. & Yeo, G. W. Bento: A toolkit for subcellular analysis of spatial transcriptomics data. *bioRxiv* 2022.06.10.495510 (2022). doi:10.1101/2022.06.10.495510

200. Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E. & Singh, S. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nat. Methods* **19,** 1550–1557 (2022).

201. Neurohr, G. E., Terry, R. L., Lengefeld, J., Bonney, M., Brittingham, G. P., Moretto, F., Miettinen, T. P., Vaites, L. P., Soares, L. M., Paulo, J. A., Harper, J. W., Buratowski, S., Manalis, S., van Werven, F. J., Holt, L. J. & Amon, A. Excessive Cell Growth Causes Cytoplasm Dilution And Contributes to Senescence. *Cell* **176,** 1083–1097.e18 (2019).

202. Bryan, A. K., Hecht, V. C., Shen, W., Payer, K., Grover, W. H. & Manalis, S. R. Measuring single cell mass, volume, and density with dual suspended microchannel resonators. *Lab Chip* **14,** 569–576 (2014).

203. Cermak, N., Olcum, S., Delgado, F. F., Wasserman, S. C., Payer, K. R., A Murakami, M., Knudsen, S. M., Kimmerling, R. J., Stevens, M. M., Kikuchi, Y., Sandikci, A., Ogawa, M., Agache, V., Baléras, F., Weinstock, D. M. & Manalis, S. R. High-throughput measurement of single-cell growth rates using serial microfluidic mass sensor arrays. *Nat. Biotechnol.* **34,** 1052–1059 (2016).

204. Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A. & Madan Babu, M. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Sci. Signal.* **11,** (2018).

205. Thiele, I., Sahoo, S., Heinken, A., Hertel, J., Heirendt, L., Aurich, M. K. & Fleming, R. M. Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol. Syst. Biol.* **16,** e8982 (2020).

206. Sartori, P. & Pigolotti, S. Thermodynamics of error correction. *Phys. Rev. X.* **5,** (2015).

207. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5,** 93–121 (2010).

208. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13,** 344–349 (2010).

209. Lin, J. & Amir, A. Homeostasis of protein and mRNA concentrations in growing cells. *Nat. Commun.* **9,** 4496 (2018).

210. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330,** 1099–1102 (2010).

211. Metzl-Raz, E., Kafri, M., Yaakov, G., Soifer, I., Gurvich, Y. & Barkai, N. Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *Elife* **6,** (2017).

212. Björkeroth, J., Campbell, K., Malina, C., Yu, R., Di Bartolomeo, F. & Nielsen, J. Proteome reallocation from amino acid biosynthesis to ribosomes enables yeast to grow faster in rich media. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 21804–21812 (2020).

213. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2,** e270 (2014).

214. Calabrese, L., Grilli, J., Osella, M., Kempes, C. P., Lagomarsino, M. C. & Ciandrini, L. Protein degradation sets the fraction of active ribosomes at vanishing growth. *PLoS Comput. Biol.* **18,** e1010059 (2022).

215. Miettinen, T. P., Pessa, H. K. J., Caldez, M. J., Fuhrer, T., Diril, M. K., Sauer, U., Kaldis, P. & Björklund, M. Identification of transcriptional and metabolic programs related to mammalian cell size. *Curr. Biol.* **24,** 598–608 (2014).

216. Kochanowski, K., Sander, T., Link, H., Chang, J., Altschuler, S. J. & Wu, L. F. Systematic alteration of in vitro metabolic environments reveals empirical growth relationships in cancer cell phenotypes. *Cell Rep.* **34,** 108647 (2021).

217. Hefzi, H. & Lewis, N. Mammalian cells devoid of lactate dehydrogenase activity. *World Patent* (2017). at <https://patentimages.storage.googleapis.com/1c/42/f9/2251487b0ba775/WO2017192437A1.pdf>

218. Mendelsohn, B. A., Bennett, N. K., Darch, M. A., Yu, K., Nguyen, M. K., Pucciarelli, D., Nelson, M., Horlbeck, M. A., Gilbert, L. A., Hyun, W., Kampmann, M., Nakamura, J. L. & Nakamura, K. A high-throughput screen of real-time ATP levels in individual cells reveals mechanisms of energy failure. *PLoS Biol.* **16,** e2004624 (2018).

219. Shen, Y., Dinh, H. V., Cruz, E., Call, C. M., Baron, H., Ryseck, R.-P., Pratas, J., Subramanian, A., Fatma, Z., Weilandt, D., Dwaraknath, S., Xiao, T., Hendry, J. I., Tran, V., Yang, L., Yoshikuni, Y., Zhao, H., Maranas, C. D., Wühr, M. & Rabinowitz, J. D. Proteome capacity constraints favor respiratory ATP generation. *bioRxiv* 2022.08.10.503479 (2022). doi:10.1101/2022.08.10.503479

220. Argüello, R. J., Combes, A. J., Char, R., Gigan, J.-P., Baaziz, A. I., Bousiquot, E., Camosseto, V., Samad, B., Tsui, J., Yan, P., Boissonneau, S., Figarella-Branger, D., Gatti, E., Tabouret, E., Krummel, M. F. & Pierre, P. SCENITH: A Flow Cytometry-Based Method to Functionally Profile Energy Metabolism with Single-Cell Resolution. *Cell Metab.* **32,** 1063–1075.e7 (2020).

221. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10,** 291–305 (2012).

222. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–248 (2010).

223. Wagner, A., Wang, C., Fessler, J., DeTomaso, D., Avila-Pacheco, J., Kaminski, J., Zaghouani, S., Christian, E., Thakore, P., Schellhaass, B., Akama-Garren, E., Pierce, K., Singh, V., Ron-Harel, N., Douglas, V. P., Bod, L., Schnell, A., Puleston, D., Sobel, R. A., Haigis, M., Pearce, E. L., Soleimani, M., Clish, C., Regev, A., Kuchroo, V. K. & Yosef, N. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity. *Cell* **184,** 4168–4185.e21 (2021).

224. Yizhak, K., Le Dévédec, S. E., Rogkoti, V. M., Baenke, F., de Boer, V. C., Frezza, C., Schulze, A., van de Water, B. & Ruppin, E. A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Mol. Syst. Biol.* **10,** 744 (2014).

225. Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C. & Lewis, N. E. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst* **4,** 318–329.e6 (2017).

226. Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D. & Palsson, B. Ø. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6,** 390 (2010).

227. Holzhütter, H.-G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271,** 2905–2922 (2004).

228. Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E. J. & Nielsen, J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13,** 935 (2017).

229. Domenzain, I., Sánchez, B., Anton, M., Kerkhoven, E. J., Millán-Oropeza, A., Henry, C., Siewers, V., Morrissey, J. P., Sonnenschein, N. & Nielsen, J. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nat. Commun.* **13,** 3766 (2022).

230. Sahoo, S., Aurich, M. K., Jonsson, J. J. & Thiele, I. Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease. *Front. Physiol.* **5,** 91 (2014).

231. Thiele, I., Jamshidi, N., Fleming, R. M. T. & Palsson, B. Ø. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5,** e1000312 (2009).

232. Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K. & Palsson, B. O. In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3,** 929 (2012).

233. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9,** 693 (2013).

234. Dahal, S., Zhao, J. & Yang, L. Genome-scale Modeling of Metabolism and Macromolecular Expression and Their Applications. *Biotechnol. Bioprocess Eng.* **25,** 931–943 (2021).

235. Yang, L., Mih, N., Anand, A., Park, J. H., Tan, J., Yurkovich, J. T., Monk, J. M., Lloyd, C. J., Sandberg, T. E., Seo, S. W., Kim, D., Sastry, A. V., Phaneuf, P., Gao, Y., Broddrick, J. T., Chen, K., Heckmann, D., Szubin, R., Hefner, Y., Feist, A. M. & Palsson, B. O. Cellular responses to reactive oxygen species are predicted from molecular mechanisms. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 14368–14373 (2019).

236. Du, B., Yang, L., Lloyd, C. J., Fang, X. & Palsson, B. O. Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in Escherichia coli. *PLoS Comput. Biol.* **15,** e1007525 (2019).

237. Chen, K., Gao, Y., Mih, N., O'Brien, E. J., Yang, L. & Palsson, B. O. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 11548–11553 (2017).

238. Cell design in bacteria as a convex optimization problem. *Automatica* **47,** 1210–1218 (2011).

239. Goelzer, A. & Fromion, V. RBA for eukaryotic cells: foundations and theoretical developments. *bioRxiv* 750182 (2019). doi:10.1101/750182

240. Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., Sun, G., Agmon, E., DeFelice, M. M., Maayan, I., Lane, K., Spangler, R. K., Gillies, T. E., Paull, M. L., Akhter, S., Bray, S. R., Weaver, D. S., Keseler, I. M., Karp, P. D., Morrison, J. H. & Covert,

M. W. Simultaneous cross-evaluation of heterogeneous datasets via mechanistic simulation. *Science* **369,** (2020).

241. Lutter, M. & Nestler, E. J. Homeostatic and hedonic signals interact in the regulation of food intake. *J. Nutr.* **139,** 629–632 (2009).

242. Hardie, D. G., Ross, F. A. & Hawley, S. A. AMPK: a nutrient and energy sensor that maintains energy homeostasis. *Nat. Rev. Mol. Cell Biol.* **13,** 251–262 (2012).

243. French, S. S., Dearing, M. D. & Demas, G. E. Leptin as a physiological mediator of energetic trade-offs in ecoimmunology: implications for disease. *Integr. Comp. Biol.* **51,** 505–513 (2011).

244. Wang, J., Liu, R., Hawkins, M., Barzilai, N. & Rossetti, L. A nutrient-sensing pathway regulates leptin gene expression in muscle and fat. *Nature* **393,** 684–688 (1998).

245. Simpson, S. J. & Raubenheimer, D. A multi-level analysis of feeding behaviour: the geometry of nutritional decisions. *Phil. Trans. R. Soc. Lond. B* **342,** 381–402 (1993).

246. Sterner, R. W. & Elser, J. J. *Ecological Stoichiometry*. (Princeton University Press, 2017).

247. Khan, M. S., Spann, R. A., Münzberg, H., Yu, S., Albaugh, V. L., He, Y., Berthoud, H.-R. & Morrison, C. D. Protein Appetite at the Interface between Nutrient Sensing and Physiological Homeostasis. *Nutrients* **13,** (2021).

248. Solon-Biet, S. M., Cogger, V. C., Pulpitel, T., Heblinski, M., Wahl, D., McMahon, A. C., Warren, A., Durrant-Whyte, J., Walters, K. A., Krycer, J. R., Ponton, F., Gokarn, R., Wali, J. A., Ruohonen, K., Conigrave, A. D., James, D. E., Raubenheimer, D., Morrison, C. D., Le Couteur, D. G. & Simpson, S. J. Defining the Nutritional and Metabolic Context of FGF21 Using the Geometric Framework. *Cell Metab.* **24,** 555–565 (2016).

249. Yan, Y., Zhang, L., Zhu, T., Deng, S., Ma, B., Lv, H., Shan, X., Cheng, H., Jiang, K., Zhang, T., Meng, B., Mei, B., Li, W.-G. & Li, F. Reconsolidation of a post-ingestive nutrient memory requires mTOR in the central amygdala. *Mol. Psychiatry* **26,** 2820–2836 (2021).

250. Hewson-Hughes, A. K., Colyer, A., Simpson, S. J. & Raubenheimer, D. Balancing macronutrient intake in a mammalian carnivore: disentangling the influences of flavour and nutrition. *R Soc Open Sci* **3,** 160081 (2016).

251. Johnson, C. A., Raubenheimer, D., Rothman, J. M., Clarke, D. & Swedell, L. 30 days in the life: daily nutrient balancing in a wild chacma baboon. *PLoS One* **8,** e70383 (2013).

252. Felton, A. M., Felton, A., Wood, J. T., Foley, W. J., Raubenheimer, D., Wallis, I. R. & Lindenmayer, D. B. Nutritional Ecology of Ateles chamek in lowland Bolivia: How Macronutrient Balancing Influences Food Choices. *Int. J. Primatol.* **30,** 675–696 (2009).

253. Rode, K. D. & Robbins, C. T. Why bears consume mixed diets during fruit abundance. *Canadian Journal of Zoology* **78,** 1640–1645 (2000).

254. Rothman, J. M., Raubenheimer, D. & Chapman, C. A. Nutritional geometry: gorillas prioritize non-protein energy while consuming surplus protein. *Biol. Lett.* **7,** 847–849 (2011).

255. Jensen, K., Simpson, S. J., Nielsen, V. H., Hunt, J., Raubenheimer, D. & Mayntz, D. Nutrient-specific compensatory feeding in a mammalian carnivore, the mink, Neovison vison. *Br. J. Nutr.* **112,** 1226–1233 (2014).

256. Simpson, S. J., Batley, R. & Raubenheimer, D. Geometric analysis of macronutrient intake in humans: the power of protein? *Appetite* **41,** 123–140 (2003).

257. Cavigliasso, F., Dupuis, C., Savary, L., Spangenberg, J. E. & Kawecki, T. J. Experimental evolution of post-ingestive nutritional compensation in response to a nutrient-poor diet. *Proc. Biol. Sci.* **287,** 20202684 (2020).

258. Clissold, F. J., Tedder, B. J., Conigrave, A. D. & Simpson, S. J. The gastrointestinal tract as a nutrient-balancing organ. *Proc. Biol. Sci.* **277,** 1751–1759 (2010).

259. Anderson, T. R., Raubenheimer, D., Hessen, D. O., Jensen, K., Gentleman, W. C. & Mayor, D. J. Geometric Stoichiometry: Unifying Concepts of Animal Nutrition to Understand How Protein-Rich Diets Can Be 'Too Much of a Good Thing'. *Front. Ecol. Evol.* **8,** (2020).

260. Azzout-Marniche, D., Gaudichon, C., Blouet, C., Bos, C., Mathé, V., Huneau, J.-F. & Tomé, D. Liver glyconeogenesis: a pathway to cope with postprandial amino acid excess in high-protein fed rats? *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **292,** R1400–7 (2007).

261. Bender, D. A. The metabolism of 'surplus' amino acids. *Br. J. Nutr.* **108 Suppl 2,** S113–21 (2012).

262. Harber, M. P., Schenk, S., Barkan, A. L. & Horowitz, J. F. Effects of dietary carbohydrate restriction with high protein intake on protein metabolism and the somatotropic axis. *J. Clin. Endocrinol. Metab.* **90,** 5175–5181 (2005).

263. Lambert, J. E., Fellner, V., McKenney, E. & Hartstone-Rose, A. Binturong (Arctictis binturong) and Kinkajou (Potos flavus) digestive strategy: implications for interpreting frugivory in Carnivora and primates. *PLoS One* **9,** e105415 (2014).

264. Veldhorst, M. A. B., Westerterp-Plantenga, M. S. & Westerterp, K. R. Gluconeogenesis and energy expenditure after a high-protein, carbohydrate-free diet. *Am. J. Clin. Nutr.* **90,** 519–526 (2009).

265. Bisschop, P. H., De Sain-Van Der Velden, M. G. M., Stellaard, F., Kuipers, F., Meijer, A. J., Sauerwein, H. P. & Romijn, J. A. Dietary carbohydrate deprivation increases 24-hour nitrogen excretion without affecting postabsorptive hepatic or whole body protein metabolism in healthy men. *J. Clin. Endocrinol. Metab.* **88,** 3801–3805 (2003).

266. Soeters, M. R., Soeters, P. B., Schooneman, M. G., Houten, S. M. & Romijn, J. A. Adaptive reciprocity of lipid and glucose metabolism in human short-term starvation. *Am. J. Physiol. Endocrinol. Metab.* **303,** E1397–407 (2012).

267. Simpson, S. J. & Raubenheimer, D. Obesity: the protein leverage hypothesis. *Obes. Rev.* **6,** 133–142 (2005).

268. Dai, Z., Zheng, W. & Locasale, J. W. Amino acid variability, tradeoffs and optimality in human diet. *Nat. Commun.* **13,** 6683 (2022).

269. Heymsfield, S. B., Peterson, C. M., Bourgeois, B., Thomas, D. M., Gallagher, D., Strauss, B., Müller, M. J. & Bosy-Westphal, A. Human energy expenditure: advances in organ-tissue prediction models. *Obes. Rev.* **19,** 1177–1188 (2018).

270. Lailvaux, S. P. & Husak, J. F. The life history of whole-organism performance. *Q. Rev. Biol.* **89,** 285–318 (2014).

271. Aiello, L. C. & Wheeler, P. The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution. *Current Anthropology* **36,** 199–221 (1995).

272. Kuzawa, C. W., Chugani, H. T., Grossman, L. I., Lipovich, L., Muzik, O., Hof, P. R., Wildman, D. E., Sherwood, C. C., Leonard, W. R. & Lange, N. Metabolic costs and evolutionary implications of human brain development. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 13010–13015 (2014).

273. Peters, A., Schweiger, U., Pellerin, L., Hubold, C., Oltmanns, K. M., Conrad, M., Schultes, B., Born, J. & Fehm, H. L. The selfish brain: competition for energy resources. *Neurosci. Biobehav. Rev.* **28,** 143–180 (2004).

274. Peters, A. The selfish brain: Competition for energy resources. *Am. J. Hum. Biol.* **23,** 29–34 (2011).

275. Hitze, B., Hubold, C., van Dyken, R., Schlichting, K., Lehnert, H., Entringer, S. & Peters, A. How the selfish brain organizes its supply and demand. *Front. Neuroenergetics* **2,** 7 (2010).

276. van Noordwijk, A. J. & de Jong, G. Acquisition and Allocation of Resources: Their Influence on Variation in Life History Tactics. *The American Naturalist* **128,** 137–142 (1986).

277. Cotter, S. C., Simpson, S. J., Raubenheimer, D. & Wilson, K. Macronutrient balance mediates trade-offs between immune function and life history traits. *Functional Ecology* **25,** 186–198 Preprint at https://doi.org/10.1111/j.1365-2435.2010.01766.x (2011)

278. Lee, E. C., Fragala, M. S., Kavouras, S. A., Queen, R. M., Pryor, J. L. & Casa, D. J. Biomarkers in Sports and Exercise: Tracking Health, Performance, and Recovery in Athletes. *J. Strength Cond. Res.* **31,** 2920–2937 (2017).

279. Vandenbogaerde, T. J. & Hopkins, W. G. Effects of acute carbohydrate supplementation on endurance performance: a meta-analysis. *Sports Med.* **41,** 773–792 (2011).

280. Harber, M. P., Konopka, A. R., Jemiolo, B., Trappe, S. W., Trappe, T. A. & Reidy, P. T. Muscle protein synthesis and gene expression during recovery from aerobic exercise in the fasted and fed states. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **299,** R1254–62 (2010).

281. Ferguson-Stegall, L., McCleave, E., Ding, Z., Doerner, P. G., Iii, Liu, Y., Wang, B., Healy, M., Kleinert, M., Dessard, B., Lassiter, D. G., Kammer, L. & Ivy, J. L. Aerobic exercise training adaptations are increased by postexercise carbohydrate-protein supplementation. *J. Nutr. Metab.* **2011,** 623182 (2011).

282. Wroble, K. A., Trott, M. N., Schweitzer, G. G., Rahman, R. S., Kelly, P. V. & Weiss, E. P. Low-carbohydrate, ketogenic diet impairs anaerobic exercise performance in exercise-trained women and men: a randomized-sequence crossover trial. *J. Sports Med. Phys. Fitness* **59,** 600–607 (2019).

283. Sale, C., Saunders, B., Hudson, S., Wise, J. A., Harris, R. C. & Sunderland, C. D. Effect of β-alanine plus sodium bicarbonate on high-intensity cycling capacity. *Med. Sci. Sports Exerc.* **43,** 1972–1978 (2011).

284. Hargreaves, M. & Spriet, L. L. Skeletal muscle energy metabolism during exercise. *Nat Metab* **2,** 817–828 (2020).

285. Magkos, F., Hjorth, M. F. & Astrup, A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. *Nat. Rev. Endocrinol.* **16,** 545–555 (2020).

286. Gomez-Pinilla, F. The combined effects of exercise and foods in preventing neurological and cognitive disorders. *Prev. Med.* **52 Suppl 1,** S75–80 (2011).

287. Wu, A., Ying, Z. & Gomez-Pinilla, F. Docosahexaenoic acid dietary supplementation enhances the effects of exercise on synaptic plasticity and cognition. *Neuroscience* **155,** 751–759 (2008).

288. Webster, C. C., Noakes, T. D., Chacko, S. K., Swart, J., Kohn, T. A. & Smith, J. A. H. Gluconeogenesis during endurance exercise in cyclists habituated to a long-term low carbohydrate high-fat diet. *J. Physiol.* **594,** 4389–4405 (2016).

289. Gallagher, D., Belmonte, D., Deurenberg, P., Wang, Z., Krasnow, N., Pi-Sunyer, F. X. & Heymsfield, S. B. Organ-tissue mass measurement allows modeling of REE and metabolically active tissue mass. *Am. J. Physiol.* **275,** E249–58 (1998).

290. Barclay, C. J. Energetics of contraction. *Compr. Physiol.* **5,** 961–995 (2015).

291. Beltman, J. G. M., van der Vliet, M. R., Sargeant, A. J. & de Haan, A. Metabolic cost of lengthening, isometric and shortening contractions in maximally stimulated rat skeletal muscle. *Acta Physiol. Scand.* **182,** 179–187 (2004).

292. Tillin, N. A., Pain, M. T. G. & Folland, J. P. Contraction type influences the human ability to use the available torque capacity of skeletal muscle during explosive efforts. *Proc. Biol. Sci.* **279,** 2106–2115 (2012).

293. Bergström, M. & Hultman, E. Energy cost and fatigue during intermittent electrical stimulation of human skeletal muscle. *J. Appl. Physiol.* **65,** 1500–1505 (1988).

294. Hogan, M. C., Ingham, E. & Kurdak, S. S. Contraction duration affects metabolic energy cost and fatigue in skeletal muscle. *Am. J. Physiol.* **274,** E397–402 (1998).

295. Lailvaux, S. P. & Husak, J. F. Predicting Life-History Trade-Offs with Whole-Organism Performance. *Integr. Comp. Biol.* **57,** 325–332 (2017).

296. Kistner, T. M., Pedersen, B. K. & Lieberman, D. E. Interleukin 6 as an energy allocator in muscle tissue. *Nat Metab* **4,** 170–179 (2022).

297. Ganeshan, K., Nikkanen, J., Man, K., Leong, Y. A., Sogawa, Y., Maschek, J. A., Van Ry, T., Chagwedera, D. N., Cox, J. E. & Chawla, A. Energetic Trade-Offs and Hypometabolic States Promote Disease Tolerance. *Cell* **177,** 399–413.e12 (2019).

298. Glancy, B., Willis, W. T., Chess, D. J. & Balaban, R. S. Effect of calcium on the oxidative phosphorylation cascade in skeletal muscle mitochondria. *Biochemistry* **52,** 2793–2809 (2013).

299. Glancy, B. & Balaban, R. S. Energy metabolism design of the striated muscle cell. *Physiol. Rev.* **101,** 1561–1607 (2021).

300. Schiaffino, S. & Reggiani, C. Fiber types in mammalian skeletal muscles. *Physiol. Rev.* **91,** 1447–1531 (2011).

301. Liu, G., Mac Gabhann, F. & Popel, A. S. Effects of fiber type and size on the heterogeneity of oxygen distribution in exercising skeletal muscle. *PLoS One* **7,** e44375 (2012).

302. Lai, A. K. M., Dick, T. J. M., Biewener, A. A. & Wakeling, J. M. Task-dependent recruitment across ankle extensor muscles and between mechanical demands is driven by the metabolic cost of muscle contraction. *J. R. Soc. Interface* **18,** 20200765 (2021).

303. Schindler, H. J., Rues, S., Türp, J. C., Schweizerhof, K. & Lenz, J. Jaw clenching: muscle and joint forces, optimization strategies. *J. Dent. Res.* **86,** 843–847 (2007).

304. Rues, S., Lenz, J., Türp, J. C., Schweizerhof, K. & Schindler, H. J. Forces and motor control mechanisms during biting in a realistically balanced experimental occlusion. *Arch. Oral Biol.* **53,** 1119–1128 (2008).

305. Miller, R. H. & Hamill, J. Optimal footfall patterns for cost minimization in running. *J. Biomech.* **48,** 2858–2864 (2015).

306. Umberger, B. R., Gerritsen, K. G. M. & Martin, P. E. A model of human muscle energy expenditure. *Comput. Methods Biomech. Biomed. Engin.* **6,** 99–111 (2003).

307. Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J. & Milo, R. Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 8488–8493 (2014).

308. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S. & Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548 (2011).

309. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R. & Hwa, T. Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528,** 99–104 (2015).

310. Mahmoudabadi, G., Phillips, R., Lynch, M. & Milo, R. Defining the Energetic Costs of Cellular Structures. *bioRxiv* 666040 (2019). doi:10.1101/666040

311. Chen, Y. & Nielsen, J. Energy metabolism controls phenotypes by protein efficiency and allocation. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 17592–17597 (2019).

312. Kuzawa, C. W. Developmental origins of life history: growth, productivity, and reproduction. *Am. J. Hum. Biol.* **19,** 654–661 (2007).

313. Lewin, N., Swanson, E. M., Williams, B. L. & Holekamp, K. E. Juvenile concentrations of IGF -1 predict life-history trade-offs in a wild mammal. *Functional Ecology* **31,** 894–902 Preprint at https://doi.org/10.1111/1365-2435.12808 (2017)

314. Adler, M. I. & Bonduriansky, R. Why do the well-fed appear to die young? A new evolutionary hypothesis for the effect of dietary restriction on lifespan. *Bioessays* **36,** 439–450 (2014).

315. Urlacher, S. S., Ellison, P. T., Sugiyama, L. S., Pontzer, H., Eick, G., Liebert, M. A., Cepon-Robins, T. J., Gildner, T. E. & Snodgrass, J. J. Tradeoffs between immune function and childhood growth among Amazonian forager-horticulturalists. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E3914–E3921 (2018).

316. Fischer, B., Taborsky, B. & Dieckmann, U. Unexpected patterns of plastic energy allocation in stochastic environments. *Am. Nat.* **173,** E108–20 (2009).

317. Poganik, J. R., Zhang, B., Baht, G. S., Tyshkovskiy, A., Deik, A., Kerepesi, C., Yim, S. H., Lu, A. T., Haghani, A., Gong, T., Hedman, A. M., Andolf, E., Pershagen, G., Almqvist, C., Clish, C. B., Horvath, S., White, J. P. & Gladyshev, V. N. Biological age is increased by stress and restored upon recovery. *Cell Metab.* (2023). doi:10.1016/j.cmet.2023.03.015

318. Kirkwood, T. B. L. in *The Evolution of Senescence in the Tree of Life* 23–39 (Cambridge University Press, 2017).

319. Dourado, H., Mori, M., Hwa, T. & Lercher, M. J. On the optimality of the enzyme-substrate relationship in bacteria. *PLoS Biol.* **19,** e3001416 (2021).

320. Goelzer, A. & Fromion, V. Resource allocation in living organisms. *Biochem. Soc. Trans.* **45,** 945–952 (2017).

Chapter 2: Gain-of-function STAT4 mutations lead to an autoinflammatory syndrom responsive to ruxolitinib

Disabling pansclerotic morphea (DPM) is a rare systemic inflammatory disorder, characterized by poor wound healing, fibrosis, cytopenias, hypogammaglobulinemia and malignancy, of unknown etiology and high mortality. We describe 3 families with autosomal dominant DPM, and a novel therapeutic approach. We evaluated 4 patients from 3 unrelated families with autosomal dominant DPM. Genomic sequencing independently identified 3 heterozygous variants in the SH2 domain of *STAT4*. Primary skin fibroblast and cell line assays were used to define the functional nature of the genetic defect. Single cell transcriptomics of peripheral blood mononuclear cells identified the inflammatory pathways affected in DPM and targeted by therapy. Genome sequencing revealed 3 novel heterozygous missense variants in the transcription factor *STAT4.* In transfected cell lines, these variants exhibited gain-of-function. Primary skin fibroblasts demonstrated enhanced inflammation driven by IL-6, with impaired wound healing, contraction and matrix secretion *in vitro*. JAK-STAT signaling inhibition with ruxolitinib led to improvement in the hyperinflammatory fibroblast phenotype *in vitro*, and resolution of inflammatory markers and clinical symptoms in treated patients. Single cell transcriptomics revealed expression patterns consistent with an immunodysregulatory phenotype that were targeted by JAK inhibition. Gain-of-function variants in the *STAT4* gene underlie cases of DPM. The JAK inhibitor, ruxolitinib, attenuated the dermatologic and inflammatory phenotype of DPM both *in vitro* and clinically.

2.1 Introduction

Disabling pansclerotic morphea (DPM) is a severe systemic inflammatory disorder in the scleroderma continuum, characterized by poor wound healing with rapidly progressive deep fibrosis involving the mucous membranes, dermis, subcutaneous fat, fascia, muscles, and bone leading to contractures, musculoskeletal atrophy and articular ankylosis.[1] Scleroderma-associated autoantibodies are usually not present. DPM is refractory to therapy, including systemic glucocorticoids, immunosuppression, and autologous stem cell transplant.[2] Disease pathogenesis has been attributed to abnormal collagen synthesis and deposition, vascular damage, and altered immunoregulation similar to other forms of scleroderma.[3] Current treatment integrates multiple, broad-spectrum pharmaceutical and ancillary therapies (e.g. methotrexate, mycophenolate mofetil, ultraviolet-A light therapy)[3,4], directed at halting disease progression, but has limited success and unacceptable side effects.[5,6] DPM currently has high rates of morbidity and mortality due to squamous cell carcinoma, restrictive pulmonary disease, sepsis and gangrene, resulting in a post-diagnosis survival time of less than 10 years.[1] A genetic etiology has not previously been identified.

We describe 4 individuals in 3 independent families with DPM and monoallelic gain-of-function variants in *STAT4*. STAT4 belongs to the signal transducers and activators of transcription (STATs) family of transcription factors. STAT proteins are recruited to activated receptors at the plasma membrane through interaction of an SH2 domain with a receptor phosphotyrosine residue generated by JAK activity. STAT proteins are also JAK substrates, and phosphorylation leads to dimerization, nuclear import of STAT proteins, and transcriptional activation of downstream genes. STAT proteins mediate cytokine responses and act in immune responses, cell growth and differentiation, cell survival, apoptosis, and oncogenesis.[7–9] STAT4 is also essential for transcriptional activation downstream of IL-6 receptor signaling and for transcription of IL-6.[10] IL-6 family cytokines have been proposed to coordinate immune-stroma crosstalk, and have been implicated in other forms of morphea.[11–13] Understanding the pathophysiology of DPM may shed light on a much larger range of disorders characterized by poor wound healing and severe, unchecked fibrosis.

2.2 Results

2.2.1 Patient Characteristics

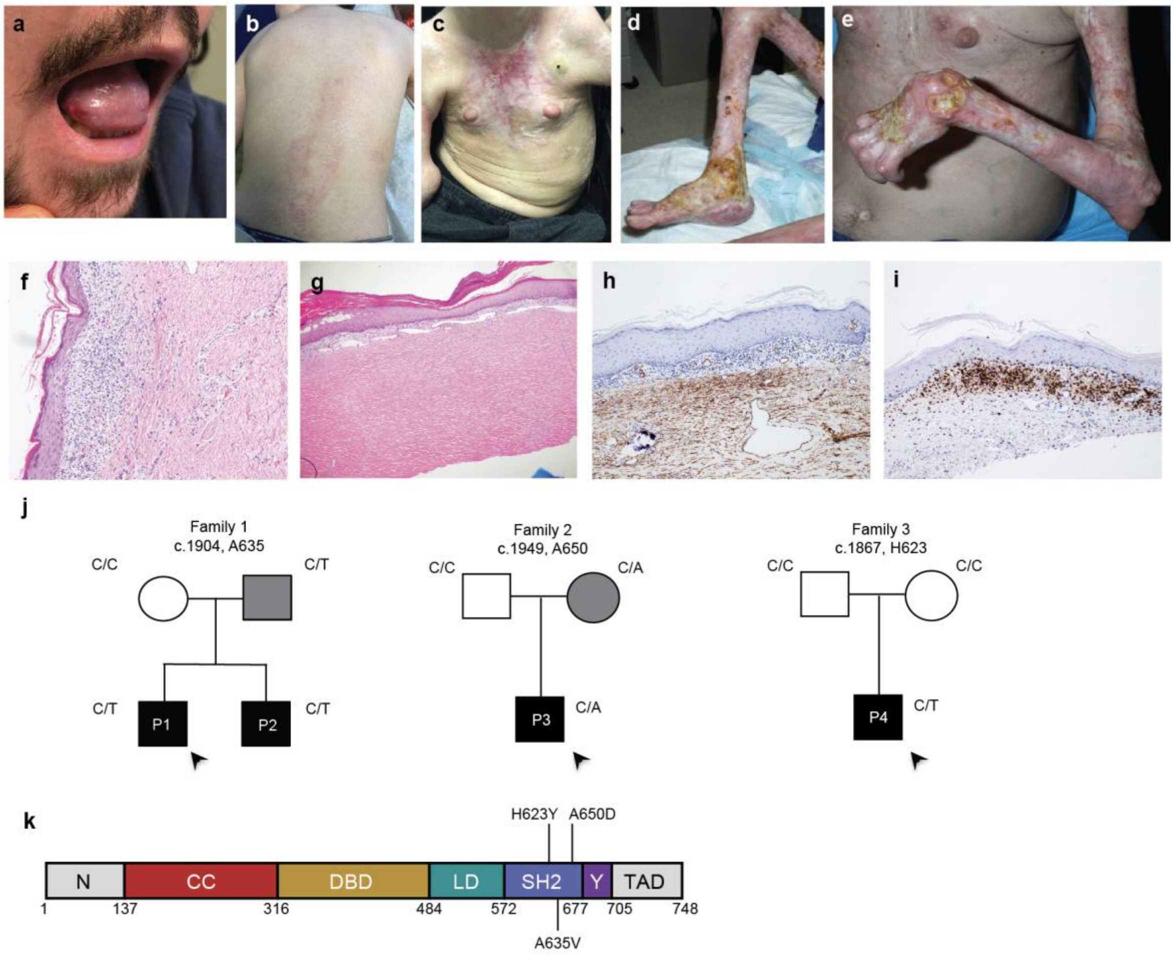


Figure 2.1: Clinical manifestations.
Clinical images showing oral ulceration and limitation in tongue protrusion **(a)**, spreading waxy, hypopigmented lesions on the back and waxy hypopigmented “tank top” sign on the chest **(b,c)**, ulcerations with articular ankylosis of the extremities **(d,e).** Histologic sections of skin biopsy show prominent inflammation **(f)** and dermal thickening and hyalinization of morphea **(g). (h,i)** Immunohistochemical staining for SMA **(h)** and CD3 **(i)** in skin biopsy samples prior to use of ruxolitinib. **(j)** Family pedigrees with probands indicated by arrowhead. Circles represent female subjects squares represent male subjects, solid form represents individuals diagnosed with DPM. Grey shading represents individuals with the STAT4 variants, but with milder symptoms. **(k)** Locations of the identified variants in a linear protein model showing protein domains :N, N-terminal domain; CC, coil-coil domain; DBD, DNA binding domain; LD, linker domain; SH2, src homology 2 domain; Y, phosphotyrosyl-tail segment, and TAD, transactivation domain. The approximate location of patient mutations are shown in the SH2 domain.

Table 2.1: Clinical phenotype of patients with disabling pansclerotic morphea.

Characteristic	**Subjects (n = 4)**	**Reference Ranges**
Demographics		
Male sex – no. (%)	4 (100)	
Median age at onset (range, yr)	3 (0.75 - 5 yrs)	
Key Features		
Skin sclerosis – no. (%)	4 (100)	
Skin/mucosal ulceration – no. (%)	4 (100)	
Muscular atrophy – no. (%)	3 (75)	
Joint contractures – no. (%)	3 (75)	
Squamous cell carcinoma – no. (%)	1 (25)	
Recurrent infections – no. (%)	2 (50)	
Laboratory findings*		
Median ANC (range) – x 10^3/ul	1.75 (0.76 - 2.79)	1000 – 6500
Median ALC (range) – x 10^3/ul	1.26 (0.5 - 2.82)	1200 – 3000
Median Platelets (range) – x 10^3/ul	278 (94 - 351)	135 – 380
Median IgG (range) – mg/dL	304 (272 - 721)	345 – 1236
Median IgA (range) – mg/dL	6.5 (<5 – 8)	14 – 154
Median IgM (range) – mg/dL	65 (13 – 84)	41 – 200
Median CRP (range) – mg/L	11 (<0.5 – 46.2)	< 0.5
Median ESR (range) – mm/h	10 (1 – 24)	<15
Histologic findings*		
Dermal thickening – no. (%)	4 (100)	
Hyalinization – no. (%)	4 (100)	
Inflammatory infiltrate – no. (%)	4 (100)	

**at initial evaluation. Range indicates observed patient values.*

We evaluated 3 unrelated kindreds with features of disabling pansclerotic morphea (DPM, **Table 2.1**). All four patients presented before the age of 5 years, with signs of mucosal ulcerations and skin sclerosis, with skin biopsy consistent with morphea by H&E staining (**Fig. 2.1a-g, S1, Table 2.1**). Musculoskeletal imaging showed deep tissue inflammation and sclerosis of several levels from subcutis to muscle, consistent with DPM (**Fig. 2.6**).

Laboratory evaluations demonstrated mild neutropenia and lymphopenia, especially of CD4+ T cells, with elevated serum inflammatory markers in two patients. Similarly, immunoglobulin classes IgG and IgA were reduced in the majority of cases, and no patients had detectable autoantibodies (**Fig. 2.7, Table 2.1**). Immunohistochemical staining of the skin (**Fig. 2.1h,i**) showed extensive alpha smooth muscle actin (SMA) staining consistent with fibrosis and CD3 positive staining in most of the subepidermal lymphocytes, confirming an inflammatory infiltrate comprised of mostly T cells.

Despite multiple systemic immunosuppressive therapies (**Fig. 2.8,4**), the majority of patients had spreading of their rash, worsening ulcerations, with rapid development of contractures, muscular atrophy and reduced mobility. Two patients also experienced recurrent infections, and one developed squamous cell carcinoma. Pulmonary nodules and infiltrates, pulmonary and portal hypertension, rapid-onset blindness (glaucoma and cataracts requiring surgery) and sensorineural hearing loss, were less frequent findings.

Two of the families had first degree relatives with oral ulcerations, mild skin disease, or early onset progressive hand swan-neck deformities, suggesting a possible inherited etiology (**Fig. 2.1j**).

2.2.2 STAT4 variants segregate with disease

Independent genomic analyses identified heterozygous variants in *STAT4* in all three kindreds (A635V, A650D, and H623Y, respectively), which were not reported in the population databases gnomAD, TopMed or deCAF, and were predicted to be damaging (**Fig. 2.1k**, **S5,6**). All of the identified *STAT4* variants (H623Y, A635V and A650D) affected the SH2 domain (**Fig. 2.1k, S6**); SH2 domain variants in STAT proteins have been implicated in other immune diseases, including those in STAT1 (H629Y) associated with chronic mucocutaneous candidiasis and in STAT3 (Q635L and N647D) associated with hyper IgE syndrome.[14–17]

2.2.3 STAT4 variants generate a gain-of-function phenotype

The clinical presentation and inheritance pattern are consistent with a gain-of-function etiology. Multiple species sequence alignment indicates that the three STAT4 variants are at conserved positions in the SH2 domain (**Fig. 2.11**). *In silico* modeling[18] predicts that they are directly involved with the SH2 phosphotyrosine peptide-binding pocket (**Fig. 2.1k, S6).** In the case of H623Y and A635V, the variants likely stabilize the dimer through hydrophobic interactions and thereby activate it, whereas the A650D mutation probably leads to intra-monomer interactions with R705, suggesting that it may contribute to the condition through a different mechanism.

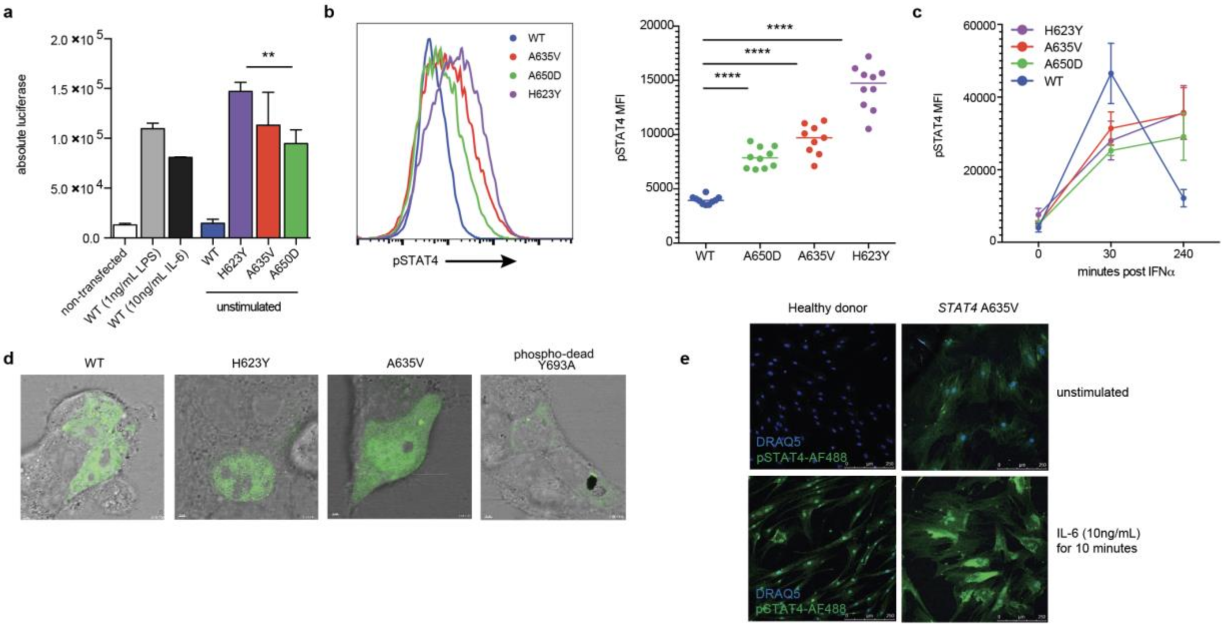


Figure 2.2: STAT4 variants generate a gain-of-function phenotype.
(a) Absolute luciferase emission from the IL-6 Leeporter cell line transfected with vector carrying wildtype (WT) or *STAT4* variants show enhanced transcriptional activity in the presence of STAT4 A635V, H623Y and A650D, compared to wildtype and the non-transfected cell line, with or without stimulation with LPS. **, p<0.01 (n=3). **(b,c)** STAT4 phosphorylation in U3A cells stably transfected with wildtype or variant STAT4. Flow cytometry used to measure mean fluorescent intensity of phosphorylated STAT4 (pSTAT4) shows increased pSTAT4 in unstimulated cells **(b)** transfected with A635V (red), A650D (green), and H623Y (purple) variants STAT4 compared to wild type STAT4 (blue). ****,p <0.0001. **(c)** In response to IFNα, pSTAT4 phosphorylation persists in variant cells at 240 minutes compared to WT cells. **(d)** HEK 293T cells transiently transfected with plasmids containing wildtype, H623Y, A635V or phospho-dead Y693A STAT4 tagged with GFP. Unstimulated cells transfected with H623Y or A635V variants show a greater accumulation of STAT4 in the nucleus compared to wildtype and Y693A cells. **(e)** Primary skin fibroblasts from patient have prominent pSTAT4 (green) compared to healthy donor fibroblasts; staining persists in a peri-nuclear location with IL-6 stimulation. Nuclear staining with DRAQ5 (blue).

To test whether the variants have a gain-of-function phenotype, we stably transfected a cell line containing a luciferase gene driven by the human IL-6 promoter, a STAT4 target,[10] with a plasmid containing WT or mutant *STAT4*. In the absence of stimulation, cells carrying the mutant plasmids exhibited enhanced luciferase activity compared to WT (**Fig. 2.2a, S7**). To further analyze phosphorylated STAT4 (pSTAT4) levels, U3A cells (deficient in STAT1 and STAT4) were stably transfected with plasmids containing wildtype STAT4 or patient variants. In unstimulated cells, pSTAT4 levels were increased with patient variants relative to wildtype STAT4 (**Fig. 2.2b**). After IFNα stimulation, cells carrying wildtype or patient STAT4 variants had similarly increased pSTAT4 levels at 30 minutes post-exposure. However, the increased pSTAT4 levels persisted in cells containing the patient STAT4 variants at 240 minutes, but decreased in cells containing wildtype STAT4, suggesting that the patient variants are resistant to mechanisms that inhibit/reduce STAT4 signaling (**Fig. 2.2c**).

RNAseq of transfected U3A cells was used to monitor global transcriptional changes caused by the variants. Variant cells stimulated with IFNα displayed a greater number and variety of differentially expressed genes compared to wildtype IFNα stimulated cells (**Fig. 2.13a,b**). Genes related to STAT4-pathways including IL-6, and the counterregulatory SOCS proteins demonstrated a greater increase in cells transfected with patient variants compared to wildtype transfected cells, even though baseline presence of total STAT4 was similar (**Fig. 2.13c,d**).

This enhanced transcriptional activity was associated with a greater accumulation of STAT4 in the nucleus of STAT4 H623Y or A635V transfected HEK 293T cells compared to wildtype and the phospho-dead control variant Y693A (**Fig. 2.2d,S9a**). Similarly, pSTAT4 was evident at baseline in primary patient skin fibroblasts but not healthy donor fibroblasts despite similar expression of *STAT4* (**Fig. 2.2e, S9b,c**). IL-6 stimulation enhanced nuclear and peri-nuclear pSTAT4 in fibroblasts (**Fig. 2.2e**), while IL-12 induced phosphorylation of STAT4 in patient peripheral blood T cells, typical of Th1 skewing, was not affected, though T cell subsets exhibited evidence for exhaustion (**Fig. 2.15a-c**). Taken together, these data make a strong case that the patient-identified *STAT4* variants cause a gain-of-function phenotype.

2.2.4 STAT4 variants lead to impaired wound healing, inflammation and fibrosis

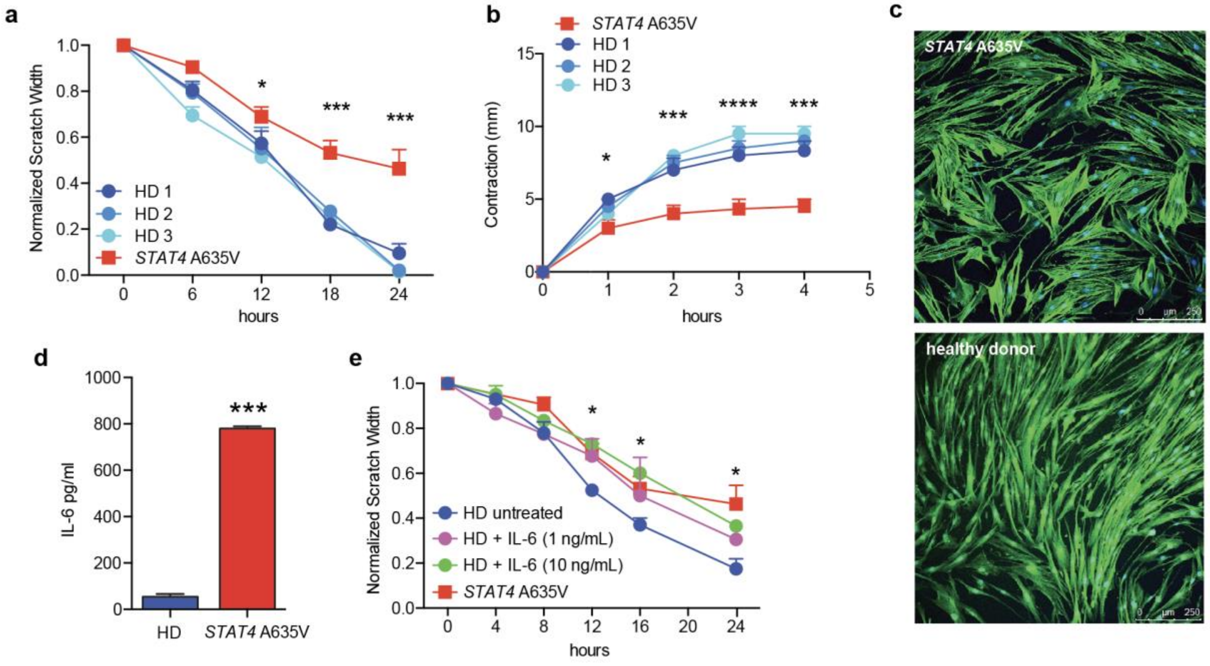


Figure 2.3: STAT4 A635V fibroblasts have a hyperinflammatory phenotype that impairs multiple aspects of cellular function and is driven by IL-6.

(a) Wound healing as measured by scratch assay is reduced in fibroblasts from STAT4 patients (red) compared to fibroblasts from healthy donors (HD, blue). (n = 3 experiments, 6 scratches each; *, p<0.05; ***, p<0.005 by 2-way ANOVA. **(b)** TGF-β-induced contraction of collagen matrix by patient-derived fibroblasts (red) is reduced relative to fibroblasts from healthy donors **(**n=3**,** ***, p<0.005, ****, p<0.0001 by 2-way ANOVA). **(c)** F-actin immunocytochemistry shows increased cell size in patient primary skin fibroblasts compared to healthy donor fibroblasts. **(d)** Patient fibroblasts show enhanced IL-6 secretion in the absence of stimulation (n=3, ***, p<0.005). **(e)** Wound healing is reduced in healthy donor primary skin fibroblasts when treated with varying concentrations of IL-6, approaching rates similar to those of cells from STAT4 patients (n=3 experiments, 6 scratches each, *, p<0.05).

STAT4 is expressed in leukocytes and skin[19–21] and may perpetuate inflammation in rheumatoid arthritis patient-derived fibroblasts.[10] Given the dramatic impairment in wound healing in DPM (**Fig. 2.1**), we used primary skin fibroblasts from patients P1 and P2 and unrelated healthy donors to study an *in vitro* model of wound healing. We first tested a scratch assay that requires all three wound healing processes, i.e., epithelialization, contraction, and connective tissue deposition. In the scratch assay, a fibroblast monolayer is scratched to induce a uniform gap, and the rate of closure is monitored (**Fig. 2.16a**). Compared to healthy donor skin fibroblasts, cells with STAT4 A635V failed to migrate as rapidly and ultimately did not close the induced gap (**Fig. 2.3a**). We also found defects in individual wound healing processes. Contraction was tested by embedding primary fibroblasts in a collagen matrix followed with TGF-β stimulation (**Fig. 2.16b,c**). Patient-derived fibroblasts had reduced contractility compared to healthy donor fibroblasts (**Fig. 2.3b**), consistent with the continued presence of inflammatory mediators.[22] Immunofluorescence staining of F-actin filaments in patient fibroblasts demonstrated greater disorganization and larger cell size (**Fig. 2.3c, S11d**). Similarly, matrix secretion of pro-collagen α1 by patient fibroblasts was diminished relative to healthy donor fibroblasts, while fibronectin secretion was unchanged consistent with poor wound healing (**Fig. 2.16e,f**).

2.2.5 IL-6 secretion by patient fibroblasts drives an autoinflammatory loop

Fibroblasts secrete proinflammatory mediators, including IL-1, IL-6, TGF-β, and VEGF.[23] We found that unstimulated patient skin fibroblasts secreted 12-fold more IL-6 than healthy donors (**Fig. 2.3d**). In contrast, the secretion of IL-1 and interferon were at the limits of detection for healthy donor and patient cultures. To explore the potential role of IL-6 in the pathology of the patient fibroblasts, healthy donor cells were pulsed with recombinant IL-6 every 8 hours for the duration of a scratch assay. Remarkably, IL-6 reduced the migration of healthy donor fibroblasts, prevented scratch closure, reduced TGF-β-induced contraction, and increased the size of cells (**Fig. 2.3e, S12a,b**). Together these data suggest that the gain-of-function STAT4 A635V variant causes an autoinflammatory loop, largely mediated by IL-6 which drives the fibroblast phenotype. *In vitro* treatment with anti-IL-6 led to a modest improvement in fibroblast function, but suggested that upstream targeting of this molecular pathway may be required to inhibit autoinflammation (**Fig. 2.17c**).

2.2.6 Therapeutic targeting with JAK inhibitors

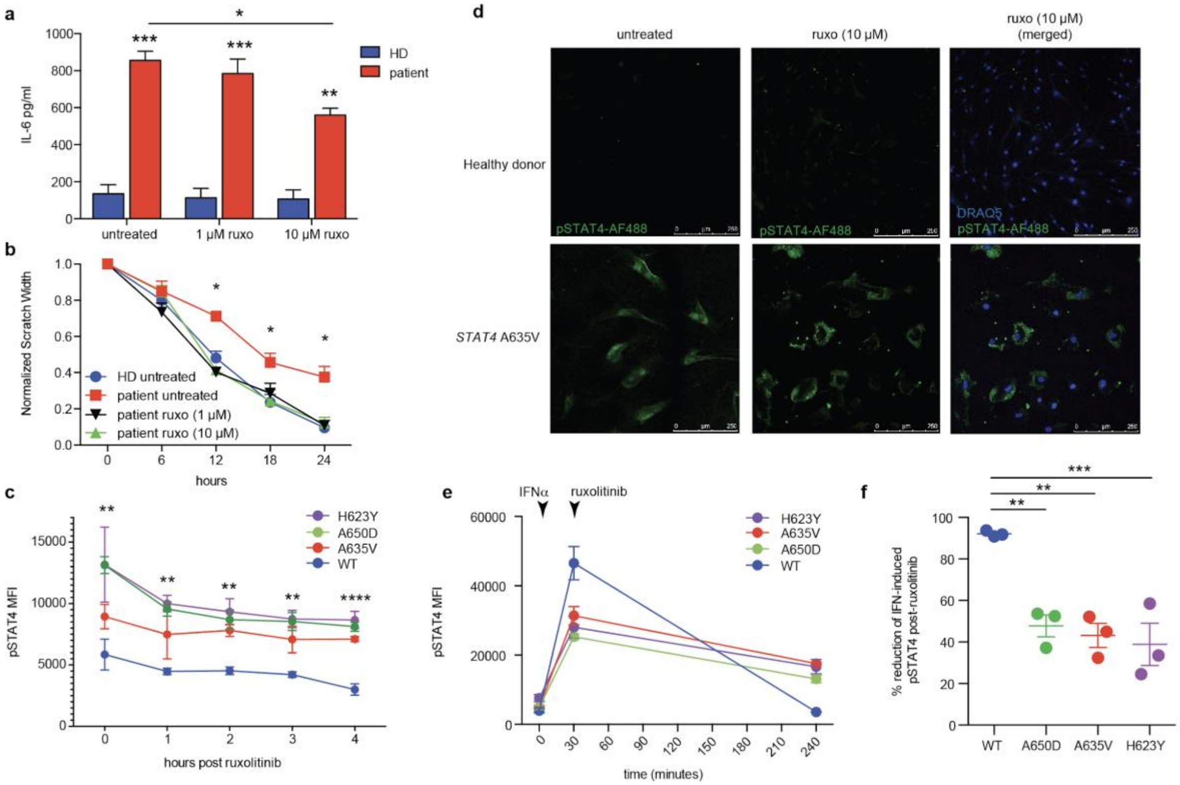


Figure 2.4: JAK inhibition reduces STAT4 phosphorylation and enhances wound healing in vitro. **(a)** Patient fibroblasts show enhanced IL-6 secretion which is responsive to ruxolitinib (n=3, *, p<0.05; **, p<0.01; ***,p<0.005; HD summary of 3 healthy donors). **(b)** Pre-treatment with ruxolitinib leads to enhanced fibroblast migration in wound healing assays, with closure near 24 hours similar to unaffected fibroblasts, n=3 experiments, 6 scratches each, *, p<0.05; by 2-way ANOVA. **(c)** STAT4 phosphorylation in unstimulated U3A cells is reduced after treatment with ruxolitinib (n=4), **, p<0.01; ****, p<0.0001, by 2-way ANOVA. **(d)** Nuclear pSTAT4 is reduced in response to ruxolitinib treatment of patient fibroblasts. (**e,f**) IFNα-stimulated U3A cells expressing variant STAT4 have higher levels of pSTAT4 210 minutes after ruxolitinib treatment compared to cells expressing wildtype (WT) STAT4 (n=3, **, p<0.01; ***, p<0.005, by ANOVA).

To determine if the hyperinflammatory fibroblast phenotype could be ameliorated *in vitro* by disrupting STAT4 signaling, we treated primary patient fibroblast cultures with the JAK inhibitor ruxolitinib. Ruxolitinib significantly reduced IL-6 secretion at concentrations achievable in serum (**Fig. 2.4a**). Similarly, ruxolitinib enhanced the rate of scratch closure of patient fibroblasts to one nearly identical to that of healthy donor fibroblasts (**Fig. 2.4b**) but did not affect the rate of scratch closure of healthy donor fibroblasts (data not shown). Other assays testing individual steps in wound healing were less responsive to ruxolitinib treatment (**Fig. 2.18**). Ruxolitinib treatment also reduced the total amount of pSTAT4 and its nuclear localization in unstimulated U3A cells and patient fibroblasts (**Fig. 2.4c,d**). However, the persistent levels of pSTAT4, even in the absence of new phosphorylation (**Fig. 2.4e,f**), suggest that the mutant STAT4 dimers are more stable than wild-type, consistent with *in silico* predictions.

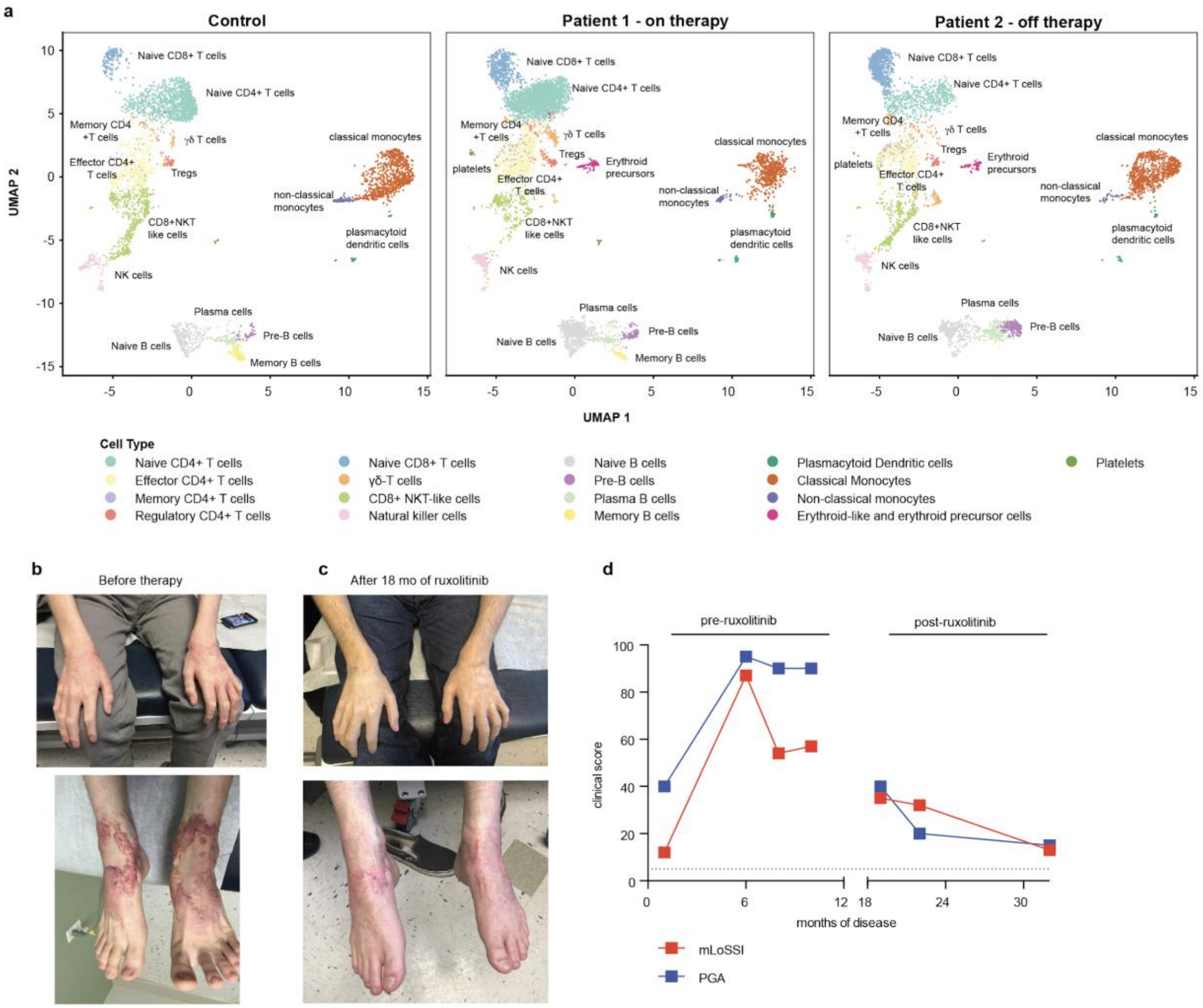


Figure 2.5: Ruxolitinib treatment leads to resolution of inflammatory phenotype.
(a) PBMCs analyzed by scRNA-seq plotted in UMAP space and clustered for each of the control and patient samples. **(b,c)** Waxy, erythematous nodular lesions on bilateral hands and feet prior to **(b)** and following **(c)** initiation of ruxolitinib therapy. (**d**) Clinical scoring by the modified LS Skin Severity Index (mLoSSI, red) and Physician Global Assessment (PGA, blue) of Disease Activity. Dotted line represents absence of disease.

The identification of gain-of-function STAT4 variants led us to initiate oral ruxolitinib treatment in patients P1 and P2. Peripheral blood samples, collected from P1 while on treatment and from P2 while off treatment, were analyzed by single-cell RNA sequencing to identify cell types (**Fig S14**) and to analyze differential gene expression in specific cell types such as NK cells and T-cells, which are known to highly express *STAT4.* In these cell types, pathway analysis based on differentially expressed genes revealed that ruxolitinib treatment results in decreased activity of genes such as *IFNG, IFNA, TNF, IL6* and *STAT1*, which control multiple inflammatory pathways. These cell types were also associated with increased expression of genes in pathways controlled by upstream regulators (**Fig. 2.5a, S15**).

Initiation of oral ruxolitinib therapy yielded clinical success. P1 had notable improvement in his nodular rash, without development of new lesions. After 11 months of therapy, the rash and oral ulcers had largely resolved. The most recent laboratory evaluation was notable for stable leukocyte counts, normal inflammatory markers, and normal IgG and IgM, with persistent IgA deficiency. No adverse events have been reported on ruxolitinib. Eighteen months after initiation of ruxolitinib, he had discontinued all other medications, with complete resolution of his chest rash, significant clearing of bilateral extremities and global clinical improvement (**Fig. 2.5b-d**). P2 has since begun therapy with ruxolitinib with improvement in his pulmonary hypertension. He continues on IVIG 2 g/kg for IgG levels <1000 mg/dL, but has been able to avoid infusions for up to 2 months at a time. His neutropenia resolved, inflammatory markers normalized, anemia improved, and thrombocytopenia stabilized.

2.3 Discussion

DPM, the most severe subtype of deep morphea within the juvenile localized scleroderma spectrum, is characterized by rapid sclerosis in all layers of the skin, fascia, muscle, and bone.[24] The etiology of DPM has been elusive since the first description in 1923.[24] Despite the general belief that a genetic component underlies disease development, a genetic basis for DPM has not been previously identified. Our identification of novel, autosomal dominantly inherited or *de novo* variants in *STAT4* is the first description of a gain-of-function variant in this gene and the first genetic link for DPM.

STAT4 roles that could drive the multiple clinical facets of DPM[10] include its involvement in T helper type-1 cell development and function[25] and regulation of IL-6 by stromal cells.[13,23] Genome-wide association studies have also implicated SNPs in *STAT4* with multiple autoimmune diseases (**Fig. 2.21**).[26,27] Given the number of pathways STAT4 acts in, additional genetic modifiers of disease severity could exist, but the presence of a parent with milder disease in 2 of the 3 families is more reflective of a clinical disease spectrum with variable expression or incomplete penetrance.

Increasing evidence points to fibroblasts as inflammatory mediators in sites of inflammation.[13,23] IL-6 family cytokines specifically have been proposed to coordinate immune-stroma crosstalk.[11–13] IL-6 has also been implicated in negatively regulating Th1 differentiation,[28] which may explain the lack of Th1 cell skewing and reduced phosphorylation to IL-12, as well as generating a chronic inflammatory disease state, consistent with the T cell exhaustion[29,30] we observed prior to treatment. The prominent IL-6 signature observed in our fibroblast cultures also suggests that anti-IL-6 monoclonal antibodies, such as tocilizumab, currently approved for interstitial lung disease in systemic sclerosis,[31] may be an alternative therapy or may be useful in combination with JAK inhibitors in patients with DPM.

We speculated that the *STAT4* gain-of-function mutations were dependent on JAK activity, and explored the use of the clinically available JAK inhibitor ruxolitinib for patients in the most severely affected family. For the patient on consistent therapy, we observed normalization of most immunologic parameters and resolution of systemic symptoms, without adverse effects. Given the multiple systems and body surface area affected, we expect that oral systemic therapy, rather than topical JAK inhibitor therapy would be required in patients with DPM. We propose that this immunomodulatory approach may be promising for patients with refractory disease.

2.4 Methods

2.4.1 Human Subjects

Blood samples were obtained from the affected patients and immediate family members. Skin biopsies were obtained from the Family 1 proband and healthy controls. All participants (legal guardians if the patient was a minor) provided written informed consent under their respective Institutional Review Boards. All procedures performed in studies involving human participants were in accordance with the ethical standards of the respective institutions.

2.4.2 Statistical analyses

For *in vitro* experiments, statistical analyses and graphing were performed in Microsoft Excel and Graphpad Prism (version 5.03; Graph Pad, Graph Pad Software Inc., CA) programs with the two-tailed, unpaired Student's t test, or the Kruskal–Wallis test. Flow cytometry data were analyzed with FlowJo software. Data are expressed as mean +/- SEM. No samples were excluded from the analyses, and no randomizations performed. A *p*-value <0.05 was considered statistically significant.

2.4.3 DNA isolation, library construction, and sequencing

Family 1. Rapid whole genome sequencing largely followed methods as previously outlined. Blood was drawn following consent for rapid whole-genome sequencing[32] of the family into Rady Children's Institute Genomic Biorepository NCT02917460. Return of results on this protocol was limited to pathogenic and likely pathogenic variants in response to the FDA oversight.[33] DNA was subsequently extracted using EZ1 DSP DNA Blood Kit and sequenced using 2 × 101 base pair run on a HiSeq 2500 System (Illumina) in rapid-run mode to a ~45-fold coverage. Rapid alignment and nucleotide variant calling were performed using the Dragen (Edico

Genome) hardware and software. Single nucleotide variants were annotated and analyzed in Opal (Omicia). Initially, variants were filtered to retain those with allele frequencies of <1% in the Exome Variant Server, 1000 Genomes Samples, and Exome Aggregation Consortium database[34] (http://evs.gs.washington.edu/EVS/2016). A gene panel was built in Phenolyzer[35] using Human Phenotype Ontology (HPO).[36] Structural variants were identified with Manta[37] and CNVnator,[38] a combination that provided the highest sensitivity and precision on 21 samples with known structural variants. Structural variants were filtered to retain those affecting coding regions of known disease genes and with allele frequencies <2% in the RCIGM database. Variants were classified based on the guidelines established by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.[39] No variants reportable per protocol were identified. Analysis in QIAGEN Ingenuity Pathway Analysis revealed a rare variant in *STAT4* shared by both brothers and not inherited from their unaffected mother. These results were shared with the research team.

The variant was confirmed by Sanger sequencing in peripheral blood mononuclear cells and fibroblasts. Subsequent to completion of functional studies the variant could be reclassified and reported clinically to the patients.

Family 2. Whole genome sequencing from genomic DNA was performed by HudsonAlpha Clinical Services Lab using the Illumina HiSeq X platform. Mapping of FASTQ files to the GRCh37 human reference was done using the Burrows-Wheeler Aligner[40] (v. 0.7.15) followed by processing with Picard tools (v.2.1.1, http://broadinstitute.github.io/picard) on the NIH HPC Biowulf cluster (http://hpc.nih.gov). A variant call file (VCF) was generated with GATK (v. 3.6)[41] and subsequently annotated using ANNOVAR.[42] Strict filtering was done by evaluating for variants absent from the ExAC database[34] with *in-silico* predictions of pathogenicity, ultimately identifying the *STAT4* c.1949C>A (p.Ala650Asp) variant in both the proband and mother. Sanger sequencing confirmed the *STAT4* variant was present in the proband and his mildly affected mother, but absent in his father.

Family 3. Exome sequencing of the proband and his unaffected parents was performed on DNA derived from peripheral blood. DNA extraction was performed using the Maxwell 16 Blood DNA Purification Kit. Illumina TruSeq DNA Sample Preparation Kit version 2 was used to prepare sequence libraries. Paired end sequencing was performed on the Illumina HiSeq2000 instrument at the National Institutes of Health (NIH) Intramural Sequencing Center (NISC). Sequence reads were aligned to the reference build hg19 with BWA. PCR duplicates were marked using Picard MarkDuplicates and the results were coordinate sorted using SAMtools. Base recalibration and realignment around microindels were performed using GATK which allowed for more accurate base quality scores. GATK's genotype workflow was used to identify SNVs and indels. Identified variants were annotated with the Variant Effect Predictor (VEP). Given the phenotypic severity, variants were filtered for ultrarare or novel (MAF < 0.0001) variants. GEMINI allowed for filtering based on de novo or inheritance patterns. Multiple variant impact prediction tools were then used for further filtering for variants predicted to be pathogenic. Subsequent analysis revealed the *STAT4* c.1867C>T, p.His623Tyr variant. The variant was confirmed by Sanger Sequencing. The BigDye Terminator v1.1 Cycle Sequencing kit (Applied Biosystems) was used for sequencing coding exons of *STAT4*. The sequencing was performed on a Seq Studio Genetic Analyzer (Applied Biosystems). The sequencing data were reviewed using Sequencher (Gene Codes) and the chromatograms shown for P4 and Family 3 are derived from Sequencher.

2.4.4 Molecular dynamics methods

A full-length human STAT4 model (amino acids 1-748) predicted by AlphaFold[43,44] was obtained from UniProt (Uniprot ID: Q14765). For molecular dynamics (MD) simulations a dimeric model containing the SH2 domain (amino acids 572-679), the phosphorylated tail segment (P-tail segment, amino acids 680-706) and the transcriptional activation domain (TAD, amino acids 708-748) was built in Chimera[45] based on the STAT1 crystal structure (PDB ID: 1BF5,[46]) The obtained model was in silico phosphorylated at Y693 using Coot.[47] The final phosphorylated model was then minimized sequentially using Yasara [48] and Chimera to remove clashes. Individual MD simulations of wild-type STAT4 and STAT4 containing the single mutations were carried out using the Desmond simulation package (Schrödinger Release 2017-3). Proteins were prepared by Protein Preparation Wizard, and the optimized potentials for liquid simulation force field (OPLS_2005) parameters were used in restraint minimization and system building [49]. The system was set up for simulation using a predefined water model (TIP3P) as a solvent. The electrically neutral system for simulation was built with 0.15 M NaCl in 10 Å buffer. The NPT ensemble with 300°K, and a pressure of 1 bar was applied in the run. The simulation was performed for 50 ns, and the trajectory sampling was done at an interval of 5 ps. The short-range coulombic interactions were analyzed using a cutoff value of 9.0 Å using the short-range method. The smooth particle mesh Ewald method was used for handling long-range coulombic interactions. MD simulations supported a gain-of-function model, and are available upon request.

2.4.5 Cell lines

Human primary skin fibroblasts were cultured from skin biopsy samples as previously described.[50] Briefly, 1mm skin biopsy pieces were placed in a 6 well plated coated with 0.1% gelatin in DMEM with 20% FBS and 1% Antibiotic-Antimycotic (Gibco). Media was replaced every 2-3 days, and cell cultures were confluent after 2-3 weeks. After passage 2, cells were maintained in Gibco Medium 106 (ThermoFisher) supplemented with Gibco Low Serum Growth Supplement (ThermoFisher) and 1% Antibiotic-Antimycotic (Gibco). Cells were used between passages 4 and 9.

The IL-6 Leeporter™ Luciferase Reporter, an NIH3T3 derivative stably expressing a luciferase construct driven by human *IL6* promoter, was purchased from Abeomics, Inc. (San Diego, CA) and maintained in DMEM medium (w/ L-glutamine, 4.5g/L glucose, sodium pyruvate) supplemented with 10% heat-inactivated FBS and 1% pen/strep, plus 3 μg/ml of puromycin. The cell line was validated with dose-response curves by seeding 5 x 10^4 cells/well on a white solid-bottom 96-well plate and stimulating for 16 hours with LPS (0.1-10,000 ng/mL) or IL-6 (1-1000 ng/mL). Abeomics luciferase assay reagent (#17-1101) was added to each well, and luminescence measured within 1-5 minutes on an EnSpire Plate Reader (PerkinElmer Inc.). For all assays, cells were used between passages 4-9.

U3A cells[51,52] were acquired from Sigma Aldrich. Cells were maintained in DMEM medium (w/ L-glutamine, 4.5g/L glucose, sodium pyruvate) supplemented with 10% heat-inactivated FBS and 1% pen/strep.

2.4.6 Mammalian transfection and expression of recombinant proteins

The human *STAT4* ORF was purchased from Dharmacon (Lafayette, CO) and cloned into the Gateway pcDNA DEST40 (ThermoFisher Scientific), which contains a C-terminal V5-6x His tag, per the manufacturer's instructions. The STAT4 variants c.1949C>A mutation encoding STAT4 A650D, c.1867C>T mutation encoding STAT4 H623Y, and c.1904 C>T mutation, encoding A635V were introduced using a QuikChange protocol, per the manufacturer's instructions. The *STAT4* coding sequence in each of the mutated plasmids was verified by sequencing by Eton Bioscience, Inc. (San Diego, CA). Plasmid DNA (1.5μg) containing either wild-type or mutant *STAT4* were transfected into the Leeporter cell line using LipoFectamine 2000 per manufacturer's instructions (ThermoFisher Scientific). Cells were co-transfected with pCXLE-EGFP (Addgene) to verify similar transfection efficiencies prior to proceeding with downstream assays.

For flow cytometry studies, human STAT4 ORF was cloned into pDONR223. The c.1949C>A mutation encoding STAT4 A650D, c.1867C>T mutation encoding STAT4 H623Y, and c.1904 C>T mutation, encoding A635V were induced using QuikChange Lightning protocol, following manufacturer instructions. Mutations were verified by Sanger sequencing performed by Psomagen (Rockville, MD). U3A cells were stably transfected with *PiggyBac*[53] vector (Addgene, cat# 80479) containing STAT4 wild-type or patient variants. Transfected cells were then selected using puromycin (0.5 µg/ml). After puromycin selection, cells were stimulated with doxycycline (0.5 µg/ml) for a period of 24 hours to induce STAT4 expression before use in experiments.

2.4.7 Flow cytometric assessment of STAT4 phosphorylation

Intracellular STAT4 phosphorylation was measured by flow cytometry. For experiments in which cell lines were stimulated or inhibited, IFNα (10,000 IU/mL) and/or ruxolitinib (2.5 µM) were added at indicated time points. Cells were fixed in 4% paraformaldehyde (PFA) at room temperature for 10 minutes, then permeabilized in 100% methanol overnight at -20 degrees. Cells were then washed 3 times in PBS containing 0.5% BSA (FACS buffer), and then resuspended in FACS buffer. Staining was performed at room temperature, using Phospho-STAT4 (Tyr693) Monoclonal Antibody in PE (ThermoFisher, 12-9044-42) at a 1:200 dilution. For experiments in which a live dead stain was used, the stain was added before the fixation step at 37 for 10 mins. Analysis was performed using FlowJo and GraphPad software.

For primary human peripheral blood cells, ficolled PBMC were cultured at 1-2 x 10^6 cells/mL complete media (RPMI 1640, 2 mM glutamine, 10% FBS, penicillin and streptomycin) with phytohemagglutinin (PHA, 3 µg/mL, Sigma) for 40-72 hours, then washed and rested for 2 hours in a 37 degree C incubator. Cells were filtered and aliquoted to 0.3-0.5 x 10^6/tube per condition, resuspending in RPMI 1640 without FBS. After resting for 1 hour, cells were stimulated with human recombinant IL-12 (25 ng/mL, R&D Systems) at 37 degrees C for 20 minutes. Live/dead dye (ThermoFisher) was added per manufacturer's protocol and cells were fixed with 1.6% PFA followed by permeabilization with cold 100% methanol overnight. Cells were stained with antibodies to cell surface markers (CD4 L200, CD45RO UCHL1) and pSTAT4 (38/p-Stat4, BD Biosciences) for 30 minutes at room temperature in the dark, washed, and then evaluated by flow cytometry. Cells were analyzed on the FSC/SSC-gated blasted cells.

2.4.8 Reverse transcription and quantitative PCR

RNA was isolated from patient and healthy donor primary skin fibroblasts using Trizol reagent (Life Technologies) per manufacturer's instructions. cDNA was synthesized using Taqman Reverse Transcription reagents (Applied Biosystems). Relative gene expression for STAT4 was determined using the following primers: STAT4: 5'-CAGTGAAAGCCATCTCGGAGGA-3' and 5'-TGTAGTCTCGCAGGATGTCAGC-3' with GAPDH 5'-ACATCGCTCAGACACCATG-3' and 5'-TGTAGTTGAGGTCAATGAAGGG-3' as reference gene. Quantitative PCR was performed using 100nM each forward and reverse primer and iQ SYBR Green supermix (Bio-Rad) with a Bio-Rad CFX96 Real-Time System, and reaction parameters as per Bio-Rad instructions Data was visualized with CFX Manager v3.0 software, and relative gene expression ws determined using the 2ΔΔct method.

2.4.9 Immunocytochemistry

Clinical immunohistochemistry was performed on skin biopsy samples with mouse anti-CD3 (clone# LN10) and mouse anti-SMA (clone# ASM-1) using standard protocols.

For immunofluorescence analysis of fibroblasts, primary skin fibroblasts were plated on glass coverslips (Corning) 24 hours prior to staining. Cells were washed with dPBS, and fixed in 4% PFA for 20 minutes. After washing, cells were permeabilized with 0.1% Triton X-100 for 20 minutes. After a second wash, cells were incubated in Alexa-Fluor 488-phalloidin (1:40 in 1% BSA for 30 minutes or phospho-STAT4 (PA5-105861, 1:200, for 1 hour) protected from light. After washing, cells were stained with 10 µM DRAQ5 (ThermoFisher) for 5 minutes, coverslips mounted to glass slides and sealed with ProLong Gold anti-fade mountant (Life Technologies). Slides were imaged using a Leica TCS-SPE confocal microscope with 10X objective.

For immunofluorescence analysis of transfected HEK 293T cells, 2.5 x 10^4 cells were plated on 35mm glass bottom dishes (MatTek, #P35G-1.5-14-C). The next day, cells were transiently transfected with plasmid DNA (20 ng) containing wild-type or variant *STAT4* tagged with GFP using LipoFectamine 2000 under manufacturer's instructions (ThermoFisher Scientific). Micrographs depicted in Fig. 2.2d represent single frames after live imaging performed on a Zeiss LSM780 confocal system driven by the ZEN Black software (Zeiss). During imaging, cells were maintained at 37°C and 5% CO2 throughout the experiment using dedicated CO2 and temperature controllers (Zeiss) connected to a heated stage insert (Pecon). Time lapse was captured employing a Plan Apochromatic 40X/1.4NA oil immersion lens (Zeiss) maintained at 37°C by an objective heater (Bioptechs). Expression and localization of the STAT4-GFP protein was visualized using a 488nm argon ion laser with pinhole size set at 1 Airy Units (AU) and digital zoom set at 1.0 whereas laser power and detector gain were adjusted to avoid pixel saturation. Phase contrast images were acquired simultaneously. Cells were imaged continuously for 30 min and stimulated with INFα, which was added directly to the cells while on the microscope stage. Recordings (1,024 pixels wide) were exported as .czi files and then converted to .ims files to generate movies and single frame images in Imaris 9.9 (Bitplane). For each image intensity levels were linearly adjusted (when needed) in Imaris 9.9.

2.4.10 mRNA sequencing

U3A cells stably transfected with A650D, A635V, H623Y, wildtype, or phospho-dead Y693A variants were plated at 6 x 10^5 cells per well, and doxycycline was added to induce STAT4 protein expression. After 24 hours, cells were either left unstimulated, or stimulated with IFNα (10,000 IU/mL) for 4 hours. Total RNA was prepared using RNeasy Micro Kit (Qiagen, #74004). A fraction of total RNA (400 ng) was processed into mRNA-seq library using Quant-seq 3' mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen, #015.96) with the PCR Add-on Kit (Lexogen, #020.96) following manufacturer's protocol. The libraries were sequenced as 100 cycles (read 1) and index (8 bases) on V1.5 kit on NovaSeq 6000 (Illumina). Raw sequencing data were processed with bcl2fastq (v2.20.0.422) to generate FastQ files.

2.4.11 RNA-seq analysis

Sequence reads were trimmed with Cutadapt (v2.10), aligned to human genome (build hg38) using STAR (v2.5.4a, options: --outFilterType BySJout --outFilterMultimapNmax 200 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.6 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000), bam files were created with SAMtools (v1.10) and transcript abundance quantified using subread (v2.0.2). Differentially expressed genes were identified using the edgeR R package with the following criteria: log2 fold change > 0.5, FDR < 0.1, and logCPM > 0.3.

2.4.12 Enzyme-linked immunosorbent assay (ELISA) and antibody arrays

Secreted IL-6 was assessed using the human IL-6 Duo-Set kit (R and D Systems, Inc.) per manufacturer's protocol. Matrix production of pro-collagen I alpha-1 and fibronectin was evaluated using human Duo-Set ELISA kits (R and D Systems, Inc.) per manufacturer's instructions. ELISA absorbances were measured on an EnSpire Plate Reader.

2.4.13 Wound healing assay

Primary skin fibroblasts were plated at 2 x 10^5 cells per well on 0.1% gelatin coated 6-well plates in Gibco Medium 106 (ThermoFisher, #M106500) supplemented with Gibco Low Serum Growth Supplement (ThermoFisher, #S00310) and 1% Antibiotic-Antimycotic (Gibco). Cells were allowed to adhere and become confluent for 48 hours, and monolayers scratched manually with a p200 pipette tip. Cells were washed once to remove cellular debris. Prior to experiments, culture medium was replaced with fresh medium. In some experiments, cells were pre-treated for 24h with ruxolitinib (Cayman Chemical, #11609) or neutralizing anti-IL-6 monoclonal antibody (MQ2-13A5, ThermoFisher). Images were collected every 4-6 hours with a 4X objective using ToupView software (ToupTek Photonics, version x64). A minimum of 5 images per timepoint were collected for each assay, until wound closure. All cultures were used between passages 4 and 9, and maintained in a 37°C incubator with 5% CO_2.

2.4.14 Imaging analysis

Images, standardized at 1024 x 768 pixels, were imported into Image J. For each scratch, the analysis grid tool was overlaid and a line drawn horizontally on the grid, and measured, with 8 measurements per scratch. Measurements were exported to Microsoft Excel and normalized to the initial scratch width. A minimum of 3 independent scratches were performed.

2.4.15 Contraction assay

A collagen-based contraction assay was modified from a protocol by Cell BioLabs, Inc. Briefly, type I bovine collagen (3 mg/mL, Advanced Biomatrix, 5005), 5X DMEM (Gibco, 12100-061), and 0.5 M NaOH were mixed on ice. Primary skin fibroblasts were harvested by trypsin-EDTA treatment, and resuspended in Gibco Medium 106 (ThermoFisher, M106500) with low serum growth supplement (ThermoFisher, S00310) at 2 x 10^6 cells per mL. The collagen lattice was prepared by mixing 2 parts of cell suspension with 8 parts of ice-cold collagen solution. The cell-collagen lattice was added to a 24 well plate, at 500 µl per well and incubated for 1 hour in a 37°C incubator with 5% CO_2 to support collagen polymerization. After 1 hour, 1 mL of Gibco Medium 106 (ThermoFisher, M106500) with low serum growth supplement (ThermoFisher, S00310), was added to each well. In some experiments, cells were treated with TGF-β at 10 ng/mL (R&D Systems, 240-B-010). The cultures were incubated for 2 days, and then collagen gels released with a sterile syringe. The collagen gel size was measured hourly.

2.4.16 10X Genomics single cell RNA-seq

2.4.16.1 Sample processing and data sets

Peripheral blood samples from both patients underwent red blood cell lysis using the ACK buffer protocol (ThermoScientific), and PBMCs were resuspended in 1X PBS with 0.04% BSA. Libraries were prepared using Chromium Single Cell 3' Reagent Kits v3 (User Guide CG00183 RevA). 10X Genomics RNA sequencing was conducted at the IGM Genomics Center, University of California, San Diego, La Jolla, CA. Less than 1 hour passed from time of collection, to loading on the 10X Genomics Chip. Cells were at 97% viability using trypan blue exclusion. The healthy donor control dataset ("5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (v3 chemistry)") was obtained from the publicly available FASTQ files at https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_v3.

2.4.16.2 Single-cell RNA-seq gene expression quantitation

Alignment of exonic reads to transcripts, identification of single-cell-specific barcodes, extraction of unique molecular identifiers (UMI), and filtering of low RNA-content barcodes were used to quantitate UMI counts in each single cell. UMI counts were determined for scRNA-seq data from the PBMCs from both patients and the control (3 samples). Quantitation was performed with CellRanger's (v3.1.0) *count* function.[54] Default parameters were used, with the exception of increasing the expected number of cells for the healthy donor sample to 5,000. The GRCh38 GTF and reference FASTA files required for the analysis were built from the Ensembl release 93 files using CellRanger.

2.4.16.3 Single-cell normalization and dataset integration

Raw UMI counts for the three samples were normalized and integrated using Seurat (v4.1.1).[55] Unless otherwise stated, all analyses were performed using default parameters. Cells with fewer than 200 non-zero genes and genes found in fewer than 3 cells were excluded. To further limit the impact of low-quality cells,[56] each sample was manually assessed on standard quality control metrics (the per cell total number of counts, total number of unique genes, and fraction of counts mapping to mitochondrial genes)[57] and filtered using heuristic thresholds.[58,59] Cells from the control sample with total counts greater than or equal to 20,000, total counts less than or equal to 3,750, or percent of counts mapping to mitochondrial genes greater than or equal to 20% were excluded. Cells from Patient 1 with total counts greater than or equal to 10,000, total counts less than or equal to 1250, total unique genes detected less than or equal to 350, or percent of counts mapping to mitochondrial genes greater than or equal to 20% were excluded. Cells from Patient 2 with total counts greater than or equal to 15,000, total counts less than or equal to 1500, total unique genes detected less than or equal to 500, or percent of counts mapping to mitochondrial genes greater than or equal to 20% were excluded. Next, each sample was normalized for count depth, scaled by a factor of 1e6 to arrive at counts per million (CPM), and the CPM values with a pseudo-count of 1 added were natural log-transformed.

Finally, we accounted for batch effects using Seurat's *IntegrateData*. First, for each sample, the top 2000 variable features for each normalized sample were identified using Seurat's *FindVariableFeatures* function to help identify integration features. The samples were then scaled (normalized by the standard deviation for each gene) and centered (subtracted the average expression for each gene) prior to running principal components analysis (PCA) to 50 principal components. Finally, the reciprocal PCA method was used to identify the integration anchors.

2.4.16.4 Dimensionality reduction and clustering

The integrated dataset was analyzed by PCA to reduce the dimensionality to 47 principal components (PCs), at which point the additional variance explained by considering more principal components was less than 0.1%. We ran the uniform manifold approximation and projection (UMAP) algorithm on the PCs to further reduce the dimensionality to two for visualization. We constructed a shared nearest neighbors (SNN) graph based on Euclidean distance in the PCA space with Seurat's *FindNeighbors* function. Next, we clustered cells using Louvain modularity optimization; the resolution parameter of Seurat's *FindClusters* function was set to 0.5. We confirmed that the integrated patient dataset did not possess batch-specific clusters; each cluster contained cells arising from multiple samples (**Fig. 2.16a**).

2.4.16.5 Cell Type Identification

We identified cell types from the scaled, integrated expression matrix and the cluster annotations using ScType[60] (**Fig. 2.16b**). Input markers for each cell type were selected from the "Immune system" tissue of ScType's database. Clusters were annotated as the cell type that received the highest ScType score.

To verify cell type annotations, a list of markers for each cluster was identified using differential expression (DE) analysis. DE tested for significantly up- or down-regulated genes in a given cluster relative to all other cells in the dataset using a Wilcoxon rank-sum test. Only genes present in at least 25% of cells in the cluster and that showed at least a log2-fold-change (LFC) of 0.5 were tested. Genes with a Bonferroni corrected p-value less than or equal to 0.05 and a LFC greater than 1.5 were retained for assessment. Uncertainty in the marker expression of three clusters (clusters 2, 6, and 9 labelled as Naïve CD4+ T-cells, Naïve B cells, and gamma-delta T-cells respectively) led us to further sub-cluster these data and manually annotate them based on DE markers. For each cluster, we subsetted the batch-corrected expression matrix to cells only in that cluster and re-ran the dimensionality reduction and clustering as previously described. The PCA dimensionality in these instances was 30 PCs. The resolution parameter of Seurat's *FindClusters* function was adjusted according to the number of mixed cell type populations we expected to observe in each cluster based on DE results. DE genes for each subcluster were identified using this same testing approach, and cell types were annotated based on significantly differentially expressed genes within these sub-clusters. Sub-clusters derived from cluster 2 were re-annotated as Effector or Memory CD4+ T-cells, those from cluster 6 were re-annotated as Plasma, Pre-, and Memory B cells, and those from cluster 9 were re-annotated as gamma-delta T-cells or Regulatory CD4+ T-cells.

2.4.16.6 Differential expression across samples and Ingenuity Pathway Analysis

Differential expression (DE) analysis between cell types across different samples was conducted using MAST (v1.20.0).[61] DE was performed on the log-normalized expression matrix. To account for technical variability and other nuisance factors, we introduced the cellular detection rate (CDR)--the fraction of genes expressed in a cell--as a covariate in the design matrix.

We tested for significant differential expression using a likelihood ratio test; the full model regressed expression on both the test condition and the CDR, and the reduced model regressed expression only on the CDR. We test for differential expression of genes in Patient 1 relative to Patient 2 (Condition 1) and genes in Patient 2 relative to the healthy donor (Condition 2). In Condition 1, we test each of NK cells, naïve CD4+ T-cells, effector CD4+ T-cells, naïve CD8+T-cells, and CD8+ NKT-like cells. In Condition 2, we test for CD8+ NKT-like cells to examine an exhaustive phenotype. For Condition 1, only genes present in at least 5% of both conditions and that showed at least a LFC of 0.5 were tested. Genes with a Benjamini-Hochberg false discovery rate (FDR) less than or equal to 0.1 were considered significantly differentially expressed. For Condition 2, only genes present in at least 10% of both conditions and that showed at least a LFC of 0.9 were tested. Genes with a Benjamini-Hochberg FDR less than or equal to 0.01 were considered significantly differentially expressed.[60]

Finally, we ran Ingenuity Pathway Analysis (IPA) on the significantly differentially expressed genes for each cell type in Condition 1. Because CD8+ NKT cells had so many significant DE genes (1,416), within IPA we further filtered for genes with an FDR less than or equal to 0.01, yielding 641 “analysis ready” genes. Upstream regulators were identified by conducting a “Core Expression Analysis” on the LFC values. Default filters were used except for: 1) the Confidence filter included both “Experimentally Observed” and “High (predicted)” relationships and 2) the Species filter only included “Human” and “Uncategorized” molecules and relationships.

For cell-type specific analyses, IPA readouts on upstream regulators were sorted by "predicted activation state." Values flagged as “bias,” and molecule types “chemical,” “drug,” “complex” or “group” were removed. A total of 67 upstream regulators were enriched across the selected cell types.

2.5 Appendix

2.5.1 Appendix A: Detailed case presentations

Family 1: The elder of the two siblings (STAT4 p.Ala635Val) presented with severe oral ulcerations in early childhood leading to inability to protrude his tongue by age 3 years. At age 16, he presented to Rheumatology clinic for a newly developed waxy, pale flat lesion on his chest, with biopsy consistent with morphea. Over a 5-week period, the skin lesions increased in size, to involve the forearms and posterior ears, and subsequently progressed to involve the anterior tibia, as well as the face. His initial laboratory evaluation was notable for mild elevation of CRP with normal IgG and IgM but absent IgA. White blood cell and absolute lymphocyte counts were mildly decreased and subset analysis demonstrated decreased CD4 and CD8 T cells. He was treated with intermittent pulse methylprednisolone 1 g daily x 3 days, methotrexate 25 mg SQ weekly and prednisone 20 mg daily. Despite therapy, he continued to rapidly develop new lesions on his back with ulceration and spreading waxy hypopigmentation, leading to addition of mycophenolate mofetil and PUVA therapy for 6 weeks. The rash continued to progress with ulcerating lesions on the skin over his buttocks, and new erythematous, raised, pruritic rash on his hands and feet, with biopsy suggestive of nodular keloids, with minimal response to laser therapy. Six months later, he developed erythematous nodular lesions on his chest, with persistence of lesions on his hands, wrists, ankles, feet and upper legs, while the waxy hypopigmented lesions on his face, upper arms and lower legs largely subsided. The patient continued to note intermittent aphthous ulcers. In an attempt to taper his daily prednisone and methotrexate, monthly intravenous immunoglobulin was initiated at 2g/kg, with no further spread of the lesions. Four years into his disease course, identification of a mutation in *STAT4* led to introduction of the JAK inhibitor ruxolitinib 5 mg BID, and discontinuation of mycophenolate mofetil. Ultimately, ruxolitinib was increased to 10 mg BID and the monthly IVIG dose decreased to 1 g/kg monthly. During this time, he had notable improvement in weight and in his nodular rash, without development of new lesions. After 11 months of therapy, the rash and oral ulcers had largely resolved, and IVIG was administered only for replacement dosing. His most recent laboratory evaluation is notable for stable white blood cell, neutrophil, and lymphocyte counts, normal inflammatory markers and normal IgG and IgM. He continues to have IgA deficiency. No adverse events have been reported on ruxolitinib. By eighteen months after initiation of ruxolitinib, he had discontinued all other medications, with complete resolution of rash on his chest and significant clearing of both hands and feet.

The younger sibling presented at age 5 with neck limitation, bilateral wrist arthritis, bilateral knee arthritis, and arthritis of the metatarsophalangeal joints. He had significant diffuse muscular weakness, with normal clinical laboratory studies including serum inflammatory markers. Radiographs were notable for diffuse osteopenia without joint space narrowing or erosive disease. He was diagnosed with polyarticular JIA, with initial treatment with prednisone and naproxen. However, due to lack of improvement, he was started on methotrexate. He subsequently developed methotrexate intolerance with complaints of shaking, fever, and abdominal pain and was switched to etanercept. Persistent elevations in ESR prompted a switch to infliximab with initial dosing at 6 mg/kg and increased up to 10 mg/kg. His exam continued to be notable for persistently active bilateral arthritis of elbows, wrists, knees and ankles with an unsteady gait, and flexion contractures of the hands. He was also noted to have erythema of the upper eyelids, as well as a rash on the cheeks that was thought to be due to pseudoporphyria from naproxen use. Despite apparent improved range of motion on infliximab, he subsequently developed a right leg contracture with shiny firmness of the skin over his hands and feet as well as a scaly rash over his upper chest and hyperpigmentation of the neck. He continued to have diffuse arthritis as well as subcutaneous tissue loss without calcinosis. He was started on daily leflunomide, but switched to cyclosporine given worsening of his skin findings including tautness of the skin over the dorsum of hands and forearms as well as over the dorsum of his feet and legs below the knees. He was also noted to have tautness of the skin around the lateral portion of his neck near his jaw as well as his upper chest and lower neck. Additional evaluations included an ophthalmologic exam negative for uveitis, negative chest CT, and normal cardiac echocardiogram. Light therapy was started with some improvement in skin lesions, but he continued to have severe limitation of wrists, no motion in the ankles or subtalar joints, severe flexion contractures of the fingers, and decreased range of motion of the wrists. Skin biopsy at an outside center was reported to be consistent with scleroderma.

He was started on treatment for localized scleroderma with methylprednisolone pulsing (3 consecutive days for 3 months) and combination therapy including leflunomide, cyclosporine and prednisone. Despite this regimen, he continued to develop large linear scleroderma lesions on the side of his face, along his neck and the back of his neck and extending from the distal arms into the fingers. He had severe contractures of the feet with the lesions extending up the thighs from the feet, sparing the knees. Leflunomide was discontinued due to elevated LFTs. Exam at that time demonstrated significant weight loss from the 40th percentile to the 10-25th percentile and height to be in the 3rd percentile of age appropriate normals. He continued to have worsening scleroderma of the face without ability to note any normal skin on the face and worsening knee contractions, and the diagnosis of pansclerotic morphea was made. The patient was continued on methylprednisolone pulses and cyclosporine, and minocycline was added. Labs were notable for worsening LFTs and aldolase values, preceding a presumed scleroderma renal crisis and right ankle cellulitis/bullous impetigo requiring IV antibiotics, and new diagnosis of restrictive lung disease. Given the refractory course and progressive systemic complications, autologous bone marrow transplant was proposed. For stem cell mobilization, he received 1 dose of 2 g/m^2 cyclophosphamide followed by G-CSF and stem cell collection. Shortly afterwards, he began to develop right lower leg ulcers with Rodnan score 3 diffusely (scleroderma skin scoring). He underwent a non-myeloablative preparatory regimen followed by stem cell rescue procedure, consisting of 50 mg/kg per day times 4 days of cyclophosphamide followed by 1.5 mg/kg of rapid anti-thymocyte globulin (ATG) with methylprednisolone followed by autologous stem cell infusion. After 3 months, his right leg ulcers persisted, however, the rest of his skin lesions appeared to have improved and he was noted to have new hair growth on the skin. Over the next several months, his Rodnan score improved from 3 to 1 or 2 in multiple quadrants of the body including the back, upper chest, lower abdomen, and upper arms, with improved laboratory values including normal CBCD, and LFTs. His IgG remained low at 473 (582-1441 mg/dL) and monthly replacement IVIG to keep IgG > 500 mg/dL was initiated. Unfortunately, 1 year post-transplant, he had relapse of his skin disease with worsening joint contractures, skin tightening, hair loss and loss of function. Immunomodulatory therapy was re-initiated with intravenous cyclophosphamide every 2 weeks and daily oral imatinib mesylate. The imatinib mesylate was stopped shortly afterwards due to persistent neutropenia and myelosuppression. He continued to have sclerodermatous skin with new breakdown in the chest region leading to initiation of bosentan. However, disease progression with resorption of the distal bones of his hands, worsening thrombocytopenia and bleeding from the skin led to discontinuation of bosentan. His case was further complicated by recurrent cellulitis requiring intravenous antibiotics and subsequent squamous cell carcinoma. Despite aggressive radiotherapy, he had poor healing after the biopsy and also developed several similar lesions on the dorsum of his right foot. Five years post-transplant, he underwent right above the knee amputation and bone marrow biopsy which showed mixed cellular marrow (75% cellularity) without malignancy, but his surgery was complicated by the development of an ulcer that required surgical revision, and his replacement immunoglobulin was increased to immunomodulatory dosing at 2 g/kg monthly. He continued to have persistent skin lesions with frequent cellulitis and bleeding episodes, ultimately electing to have bilateral through-humerus amputations, complicated by post-operative liver failure, and right ventricular hypertension with subsequent development of splenomegaly, and portal hypertension. Nearly 9 years post-transplant, prompted by the genetic results, he was started on ruxolitinib. However, the patient then developed anxiety attacks and ruxolitinib was held. He subsequently underwent elective left lower extremity through joint amputation as well as chest wall biopsy. Ruxolitinib was held during this time but then restarted 2 weeks after surgery, with stabilization of his pulmonary hypertension. He continues on IVIG 2 g/kg for IgG levels < 1000 but has been able to avoid infusions for up to 2 months at a time over the last year. Most recent labs are notable for resolution of neutropenia with normal inflammatory markers, with persistent IgM and IgA deficiency.

The father of P1 and P2 has a history of oral ulcerations and less severe skin disease without a formal diagnosis, whereas the mother has no history of similar disease.

Family 2: The proband of Family 2 (STAT4 p.Ala650Asp) presented at 3 years of age with joint swelling of the ankles, knees, and elbows, and an inability to keep up with his peers. At the age of 7, he developed painful bilateral hand contractures and shortly after, developed a white patch on his left leg. He was given the diagnosis of juvenile rheumatoid arthritis, was hospitalized and treated with naproxen and methotrexate. He subsequently underwent biopsy consistent with scleroderma and he was treated with steroids for 6 months. By the following year, the fibrotic skin lesions became confluent and a diagnosis of generalized morphea was made. He was subsequently evaluated for possible eosinophilic fasciitis after a small ulcer on the right foot spread to the whole foot (except the sole) and progression of other skin ulcerations occurred. He had depigmented areas on the extremities and face with pruritus. In addition, he had increased contractions in the hands and feet and decreased use of the extremities. Lab studies showed an elevated ESR and profound peripheral eosinophilia. ANA and rheumatoid factor were negative. At that time diagnoses considered were scleroderma with possible overlap features of eosinophilic fasciitis. He was treated with infusions of methylprednisolone and then put on methotrexate. He responded to methotrexate without any major extension of the morphea. Six months after the drug was stopped, there was almost an "explosion" of skin sclerosis, preceded by inflammatory pruritus. He was treated with IM methotrexate and penicillamine for 8 months. Penicillamine was discontinued due to neutropenia. After 5 years of disease, he had generalized morphea in all stages of activity, involving the legs, buttocks, groin, shoulders and arms plus some deformity of the hands, ankles and feet with fixed flexion in the hands and wrist joints. He was treated with three pulses of methylprednisolone 500 mg for three consecutive weeks, as well as 10 mg of prednisone on a daily basis. Azathioprine and cyclosporine were added but he developed a severe infection, and he was treated with IV prostacyclin and hyperbaric oxygen. By report, the lesions did not get any worse after treatment with hyperbaric oxygen however when it was discontinued the ulcers worsened, and methotrexate was restarted. He underwent surgery to the hand to improve flexibility; upon surgical debridement of the foot ulcer, cultures grew *Serratia marcescens* and *Enterococcus faecalis*. After 8 years of disease, he had developed contractures of the knees, with maximum straightening of the right leg of about 150 degrees and maximum straightening of the left knee of less than 90 degrees, and contractures/subluxation of the ankles and elbows, and he was unable to close his hands. He had chronic foot ulcers on the right side, covering the back and distal lower legs with small ulcers on the upper legs and arms. Biopsy of the foot ulcer showed only granulation tissue, without evidence of neoplasia. His skin thickened and scarred causing extensive contractures, both axial and peripheral, limiting movement and leading to the patient being wheel chair bound. At initial evaluation, laboratory examinations were significant for ESR of 32, aldolase 11, CPK 32. Imaging included a chest x-ray with mild interstitial changes. Pathology included a lymph node biopsy suggestive of CMV infection, with bronchoalveolar lavages demonstrating increased eosinophils in the vessels. His bone marrow biopsy was unremarkable. Over the course of disease, he had persistently elevated inflammatory markers including ESR and CRP with negative autoantibodies. Over the 15 years of disease, his immunosuppressive therapy regimen included oral and intravenous steroids with a maximum dose of 32 mg prednisolone, intermittent methotrexate, maximum dose of 25 mg subcutaneously, a trial of D-penicillamine, withdrawn because of severe leukopenia that resolved after using granulocyte and growth factors, azathioprine, cyclosporine and 17 pulses of intravenous gammaglobulin (1 g/kg/day for two days). Therapies were ultimately discontinued in the setting of open wounds and risks for infection. Ulcers were primarily treated by surgical debridement and local wound care. By his twenties, he was diagnosed with 50% hearing loss. He also lost vision suddenly in both eyes from cataracts. He had corrective cataract surgery with successful return of 20/20 vision in both eyes. At age 31, the patient died following an infection.

The proband’s mother in Family 2 also has a presentation of joint deformities that began at age 20 years. She reported “electric pain” in her upper extremities with progressive hand joint swan-neck deformities over a 6-month period. The progression subsequently stabilized and was no longer painful, but she was left with marked residual deformities. She later developed bilateral cataracts requiring surgical correction at age 40, and bilateral hearing loss requiring hearing aids by the age of 50.

Family 3: The proband in Family 3 (STAT4 p.His623Tyr) developed a non-healing wound on his lower extremity at 9 months of age, followed by poor weight gain, loose stools, ‘wasted appearance’ of his lower extremities with firmness of the subcutis throughout the upper and lower extremities, with decreased range of motion and mild swelling of several small to large joints by age 12 months. Further history revealed recurrent tonsillitis, two pneumonias, with normal CBC and IgG but low IgA. Evaluation was extensive including cardiopulmonary, gastroenterology, genetics, and musculoskeletal studies. Significant findings in the musculoskeletal imaging pathology were found consistent with deep tissue inflammation and sclerosis of several levels from subcutis to muscle, consistent clinically with disabling pansclerotic morphea of childhood. An ultrasound of the lower extremities demonstrated obvious irregular nodular thickening of the subcutaneous soft tissues throughout bilaterally and arthritis of the knees. Further imaging (before treatment) included MRI hip/girdle and lower extremities (**Fig. 2.8d**) demonstrated subcutis edema with extensive fasciitis with associated adjacent muscle edema, which was directly supported histologically via full thickness skin biopsy of the left thigh (**Fig. 2.8e,f**) that showed marked sclerotic changes of the deep fascia and fibrous trabeculae of the subcutaneous fat with extension into the dermis with an associated moderate lymphoplasmacytic infiltrate of the periadnexal, perivascular and septal areas. Noted was extensive expanded fascial tissue with inflammation and sclerosis and associated neighboring skeletal muscle lymphocytic infiltrate (**Fig. 2.8e,f**).

He was treated (**Fig. 2.9**) immediately with oral glucocorticoids (2mg/kg divided BID with taper over 9 months), in conjunction with methotrexate (1 mg/kg/week), UVA ~10 J/cm2 three times/week (4 months), and aggressive physical therapy and occupational therapy. Improvement in skin induration and joint range of motion (ROM) was noted within 6 months. Oral steroids were increased to potentially gain more benefit and tapered completely off by 24 months into therapy. Methotrexate continued. Repeat MRI of hip girdle/ lower extremities 18 months into treatment demonstrated full resolution of subcutis, fascia and muscle edema (and resolved inguinal and popliteal lymph node adenopathy). He had slow improvement with full resolution of remaining skin thickness and improvement of ROM, but with remaining deficits mild in his fingers, elbows, hips and knees and more moderate in ankles and subtalar joints, over the next several years. He was weaned off methotrexate completely after 6 years of therapy, and within 6 months had a flare of tenosynovitis of the upper extremities, wrists and fingers, mostly PIPs, documented clinically and radiographically with MRI. There was no flare of the skin, subcutis or fascia. Methotrexate with oral prednisone taper was restarted and led to improvement within 6 months. Due to tolerance issues 2 years later, he was switched from methotrexate to adalimumab to treat polyarthritis, with his only remaining symptoms being morning stiffness and synovial hypertrophy on examination. He has experienced full arthritis response to adalimumab and he was been in clinical remission in skin and joint symptoms for the past 6 years. He is currently 17 years old and fully participates in track and football sports at high school and has no complaints. He does have remaining moderate joint contractures of the subtalar and ankle joints with associated radiographic narrowing of the anterior subtalar joint space and osteophyte formation at the talonavicular joint supporting early degenerative changes, but other joints have little or no remaining joint contractures. There have been no additional immune lab abnormalities, serious or chronic infections or other autoimmune diagnoses over the 15 years of follow up.

2.5.2 Appendix B: Additional Figures

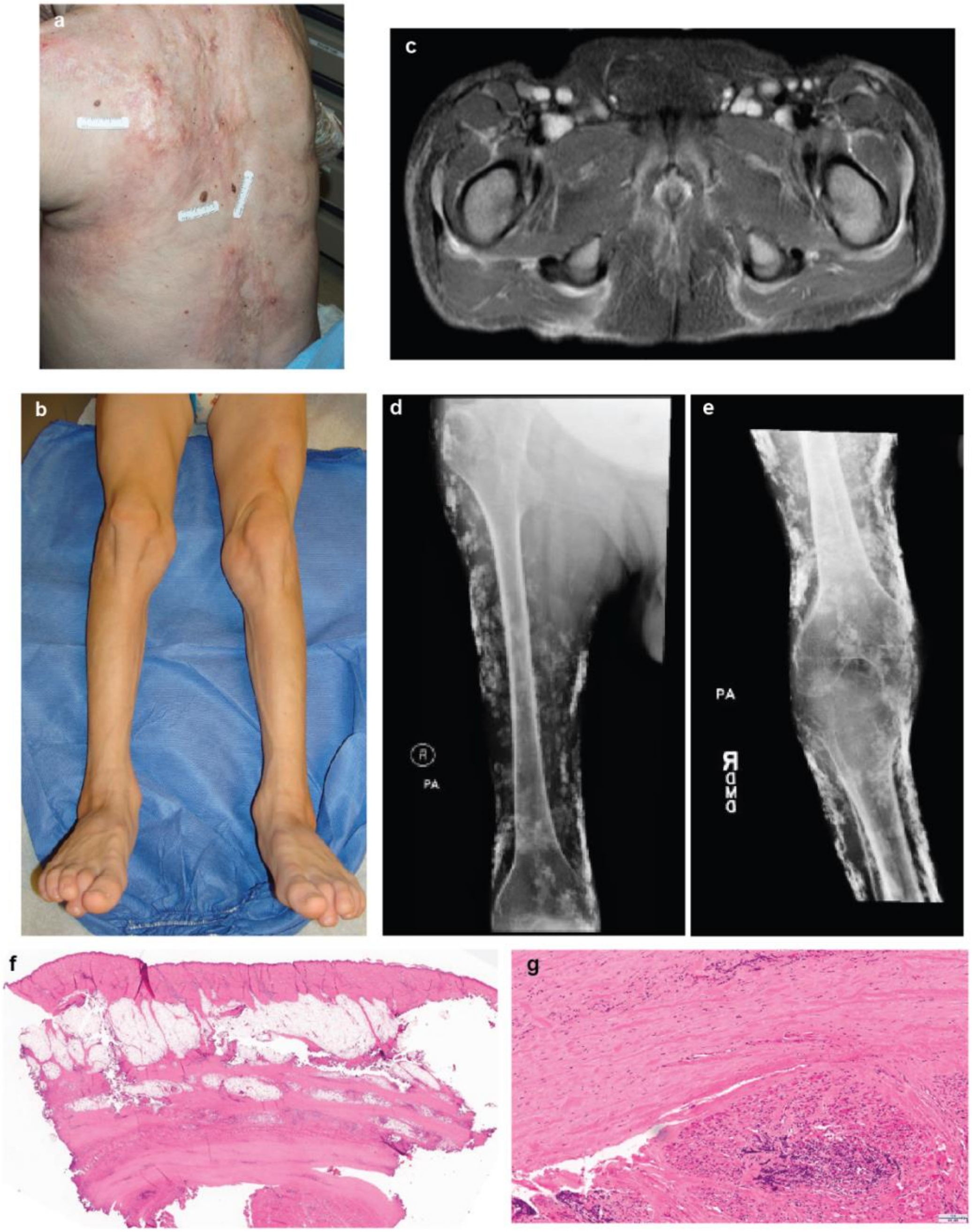


Figure 2.6: Additional clinical images and evaluations of patients.
Hypopigmented “tank top” sign on the back **(a, P3)**, and loss of entire subcutis of legs with joint contractures of ankle, subtalar and toe **(b, P4).** MRI of pelvic girdle **(c, P4)** and extremities (femur **(d)**, knee **(e), P3)** reveals diffuse fasciitis, adjacent myositis, and subcutaneous calcifications. Full thickness skin biopsy **(f,g, P4)** shows thickened fascia with sclerosis and pockets of lymphoplasmacytic inflammation throughout **(f)** with adjacent muscle demonstrating lymphocytic infiltrate throughout the muscle bundles **(g).**

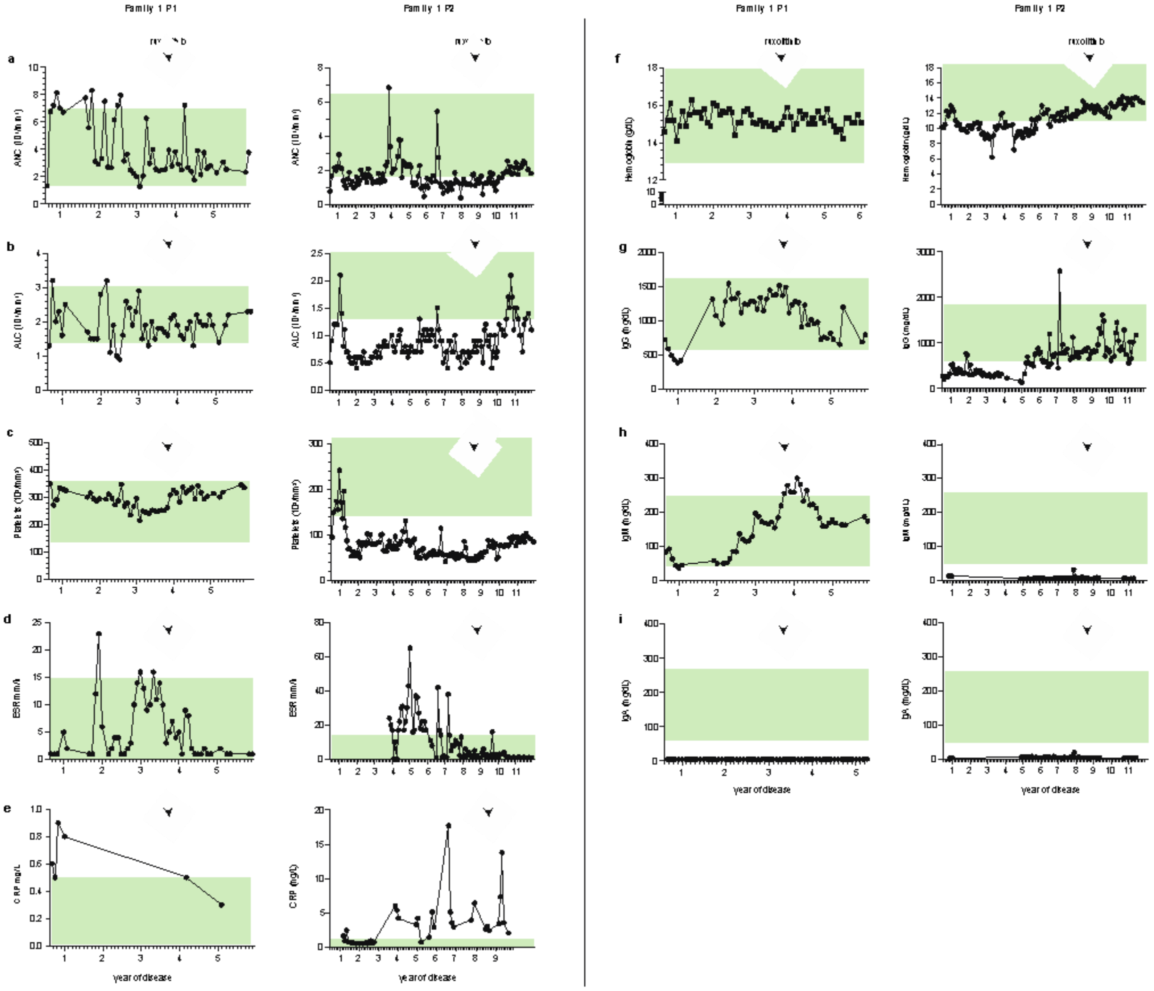


Figure 2.7: Summary of clinical labs for patients of Family 1.

(a) Absolute neutrophil count (ANC, normal range green), **(b)** absolute lymphocyte count (ALC, normal range green), **(c)** platelet count (normal range (green): 150-450 $10^3/mm^3$), **(d)** erythrocyte sedimentation rate (ESR, normal range green), **(e)** C-reactive protein levels (CRP, normal range (green): < 0.5 mg/dL), **(f)** hemoglobin (normal range (green): 11.5 – 18.0 g/dL), **(g)** IgG (normal range (green): 650-1600 mg/dL, **(h)** IgM (normal range (green): 50-300 mg/dL), and **(i)** IgA (normal range (green): 80-280 mg/dL levels over the course of disease. Arrowheads indicate initiation of ruxolitinib therapy.

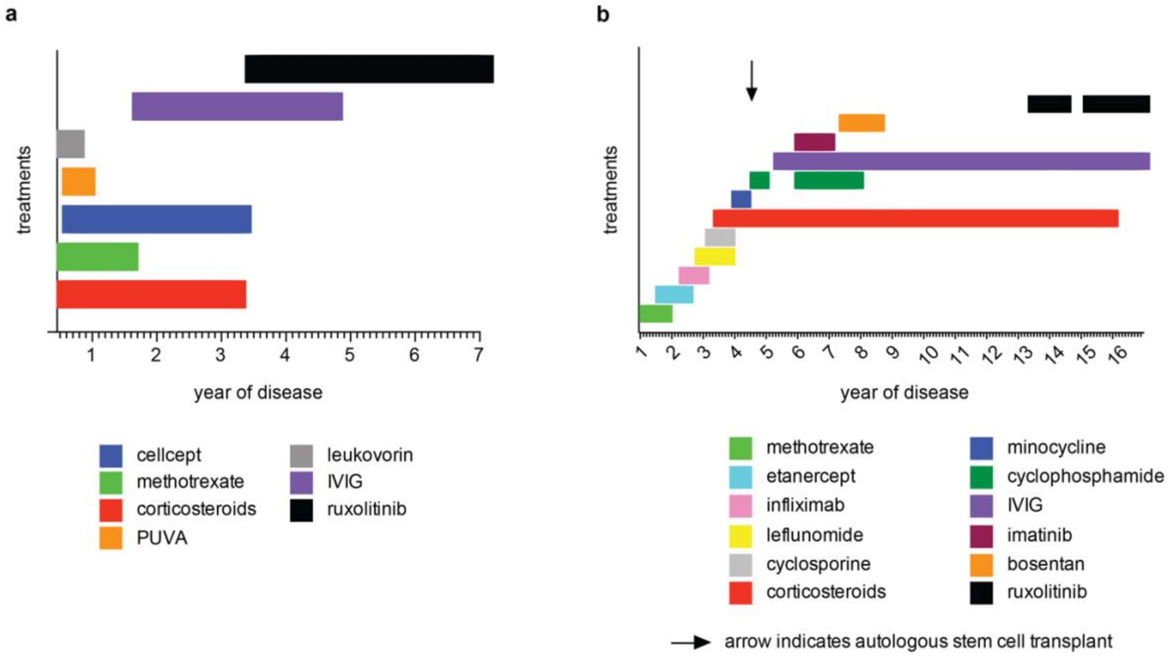


Figure 2.8: Treatment summary for Family 1.
Timeline for immunomodulatory treatments for Patient 1 **(a)** and Patient 2 **(b)**.

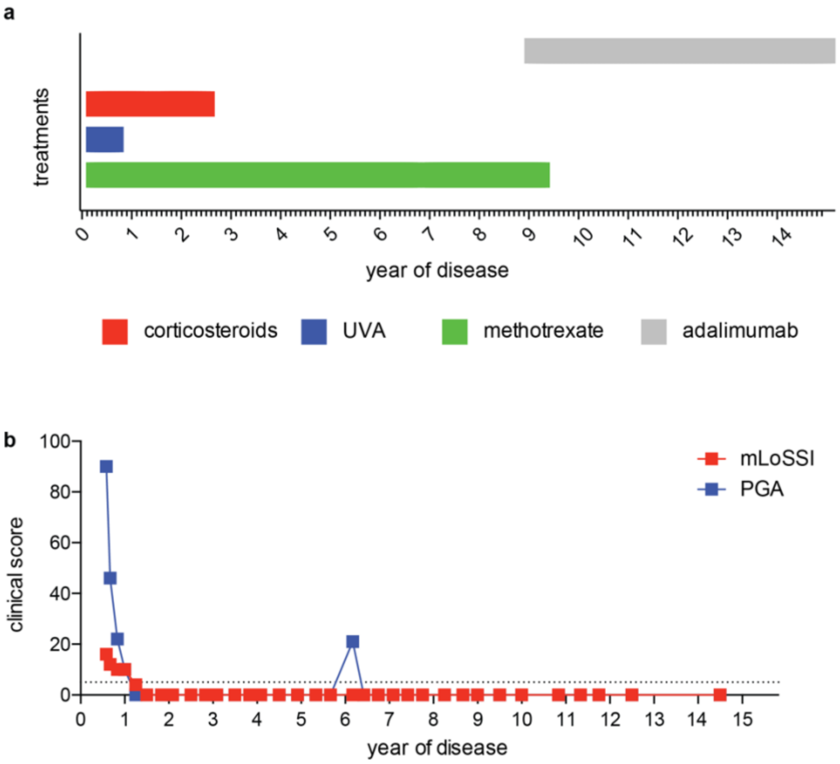


Figure 2.9: Treatment summary for Patient 4, Family 3.
Timeline for immunomodulatory treatments for the family 3 proband **(a;** *UVA, ultraviolet A***)** and clinical scoring by the modified LS Skin Severity Index (mLoSSI, red) and Physician Global Assessment (PGA, blue) of Disease Activity. Dotted line represents absence of disease **(b)**.

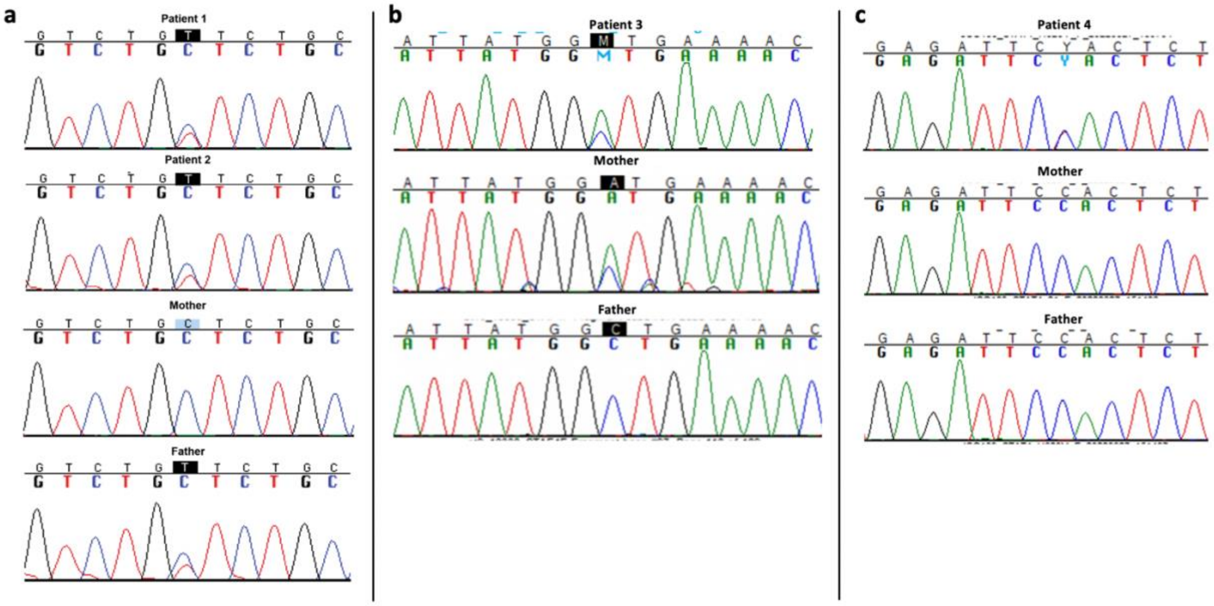
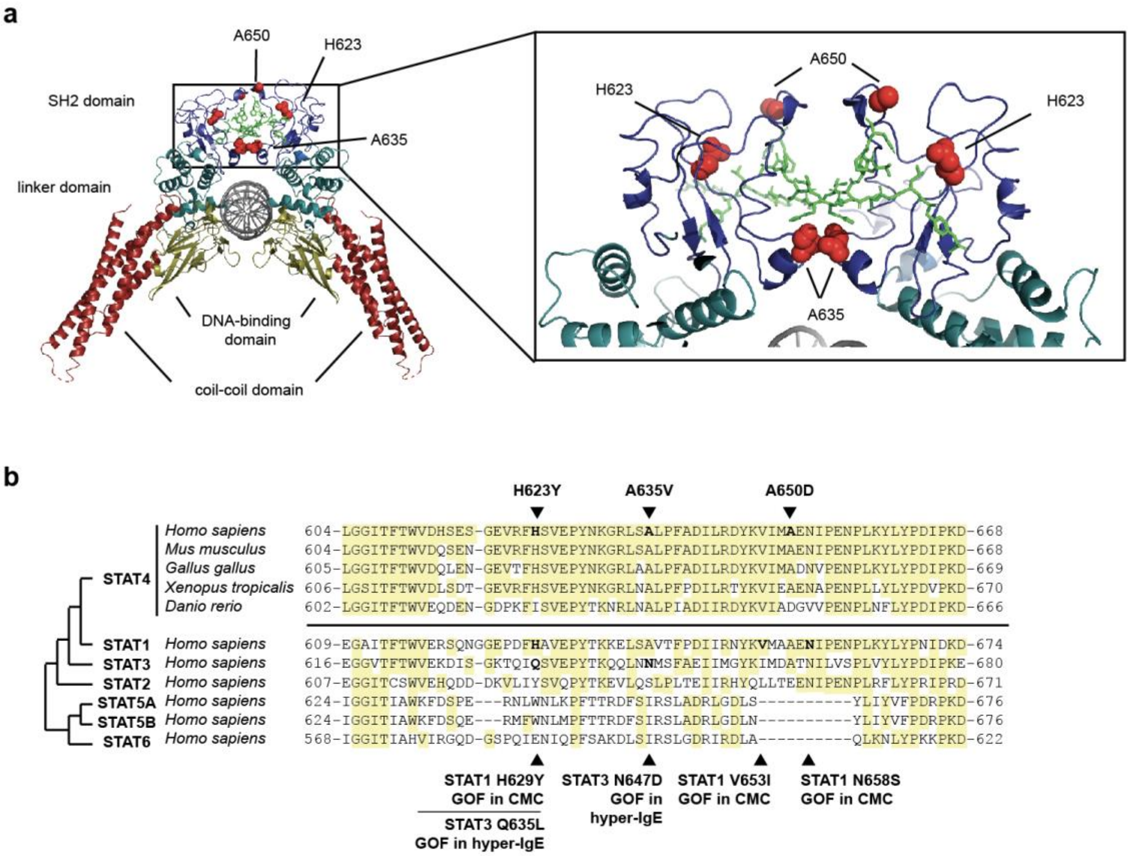


Figure 2.10: Sanger sequencing
(a) Chromatograms show confirmation of heterozygous *STAT4* mutation, c.1904 C>T, encoding A635V. **(b)** Confirmation of heterozygous *STAT4* mutation, c.1949 C>A, encoding A650D in Family 2. **(c)** Confirmation of *de novo* heterozygous *STAT4* mutation, c.1867 C>T, encoding H623Y in Family 3.

Figure 2.11: Structure, conservation and predicted impact of amino acid substitutions caused by STAT4 patient mutations.

(a) The structure of STAT4 modeled on STAT1 (PDB 1BF5; linker domain green, DNA binding domain yellow, coil-coil domains red, and the A635, H623 and A650 residues (red spheres) proximal to the SH2 domain (cyan)). The box shows a magnification of the phosphotyrosine binding pocket. **(b)** Sequences for *Homo sapiens* STAT1 (NP_009330.1), STAT2 (NP_005410.1), STAT3 (NP_644805.1), STAT4 (NP_003142.1), STAT5A (NP_003143.2), STAT5B (NP_036580.2), and STAT6 (NP_003144.3) were aligned with STAT4 sequences from *Mus musculus* (mouse, NP_035617.1), *Gallus gallus* (chicken, NP_001254484.2), *Xenopus tropicalis* (western clawed frog, XP_031749081.1), and *Danio rerio* (zebrafish, XP_005167937). Triangles at top indicate the positions of amino acids changed by patient mutations. Highlighted residues are identical to the *H. sapiens* STAT4 sequence. Numbers indicate amino acid positions. Alignments were performed with Clustal Omega.[62] Cladogram on left was derived from previous phylogenetic analyses.[63] Amino acid substitutions associated with a STAT1 gain-of-function and chronic mucocutaneous candidiasis (CMC) and STAT3 gain-of-function and hyper-IgE syndrome are indicated by triangles at bottom.[64–66] The amino acids at STAT1 H629 and STAT3 Q635 are homologous to STAT4 H623 and the STAT3 N647 amino acid is homologous to STAT4 A635. **(c)** The allele frequencies of the patient mutations in databases of single nucleotide variants reveal that these mutations are not present in the general population. **(d)** Eight programs were used to make *in silico* predictions for the effect of patient amino acid substitutions on STAT4 function (BayesDel,[67] CADD_phred,[68] FATHMM[69] GERP++,[70] PolyPhen2,[71] REVEL,[72] SIFT,[73] VEST[74]). Predictions indicating a potential or probable deleterious effect are highlighted in red. Cutoffs used for predicting deleterious effects for each program are reported at bottom.


a
A650
H623
SH2 domain
A635
linker domain
DNA-binding domain
coil-coil domain
A650
H623
H623
A635
b
H623Y
A635V
A650D
STAT4
Homo sapiens 604-LGGITFTWVDHSES-GEVRFHSVEPYNKGRLSALPFADILRDYKVIMAENIPENPLKYLYPDIPKD-668
Mus musculus 604-LGGITFTWVDQSEN-GEVRFHSVEPYNKGRLSALPFADILRDYKVIMAENIPENPLKYLYPDIPKD-668
Gallus gallus 605-LGGITFTWVDQLEN-GEVTFHSVEPYNKGRLAALPFADILRDYKVIMADNVPENPLKYLYPDIPKD-669
Xenopus tropicalis 606-LGSITFTWVDLSDT-GEVRFHSVEPYNKGRLNALPFPDILRTYKVIEAENAPENPLLYLYPDVPKD-670
Danio rerio 602-LGGITFTWVEQDEN-GDPKFISVEPYTKNRLNALPIADIIRDYKVIADGVVPENPLNFLYPDIPKD-666
STAT1 Homo sapiens 609-EGAITFTWVERSQNGGEPDFHAVEPYTKKELSAVTFPDIIRNYKVMAAENIPENPLKYLYPNIDKD-674
STAT3 Homo sapiens 616-EGGVTFTWVEKDIS-GKTQIQSVEPYTKQQLNNMSFAEIIMGYKIMDATNILVSPLVYLYPDIPKE-680
STAT2 Homo sapiens 607-EGGITCSWVEHQDD-DKVLIYSVQPYTKEVLQSLPLTEIIRHYQLLTEENIPENPLRFLYPRIPRD-671
STAT5A Homo sapiens 624-IGGITIAWKFDSPE---RNLWNLKPFTTRDFSIRSLADRLGDLS----------YLIYVFPDRPKD-676
STAT5B Homo sapiens 624-IGGITIAWKFDSQE---RMFWNLMPFTTRDFSIRSLADRLGDLN----------YLIYVFPDRPKD-676
STAT6 Homo sapiens 568-IGGITIAHVIRGQD-GSPQIENIQPFSAKDLSIRSLGDRIRDLA----------QLKNLYPKKPKD-622
STAT1 H629Y GOF in CMC
STAT3 Q635L GOF in hyper-IgE
STAT3 N647D GOF in hyper-IgE
STAT1 V653I GOF in CMC
STAT1 N658S GOF in CMC

c

Variant	Allele Frequencies gnomADg	gnomAD_exomes	ExAC	1000Gp3	TopMed	AllofUs	deCAF
STAT4:p.A650D	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STAT4:p.A635V	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STAT4:p.H623Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0

d

Variant	CADD _PHRED	BayesDel _addAF	VEST4	REVEL	FATHMM pred (score)	GERP++ _NR	SIFT pred (score)	PolyPhen2 pred (score)
STAT4:p.A650D	22.2	0.04026	0.169	0.367	D (-3.87)	5.38	tolerated (0.41)	benign (0.001)
STAT4:p.A635V	31	0.335585	0.49	0.718	D (-4.03)	5.38	deleterious (0.03)	probably damaging (0.995)
STAT4:p.H623Y	26.6	0.176164	0.451	0.623	D (-3.99)	5.38	deleterious (0.01)	possibly damaging (0.752)
Deleterious cutoff	≥20	≥0.0692655	≥0.5	≥0.65	from output	≥0.047	from output	from output

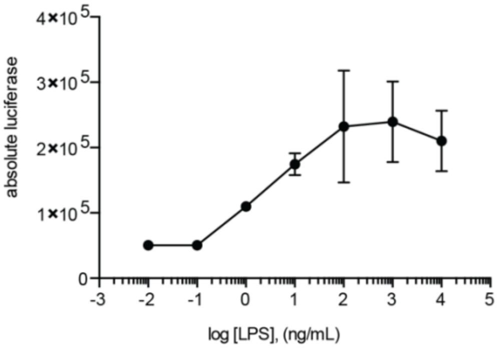


Figure 2.12: Induction of IL-6 promoter activity by LPS in IL-6 Leeporter™ cells.
IL-6 Leeporter™ cells were seeded at 5 x 10^4 cells/well into a white solid-bottom 96-well microplate. Cells were stimulated with various concentrations of LPS for 16 hours, and luciferase activity measured. Shown as average ± SEM (n=2).

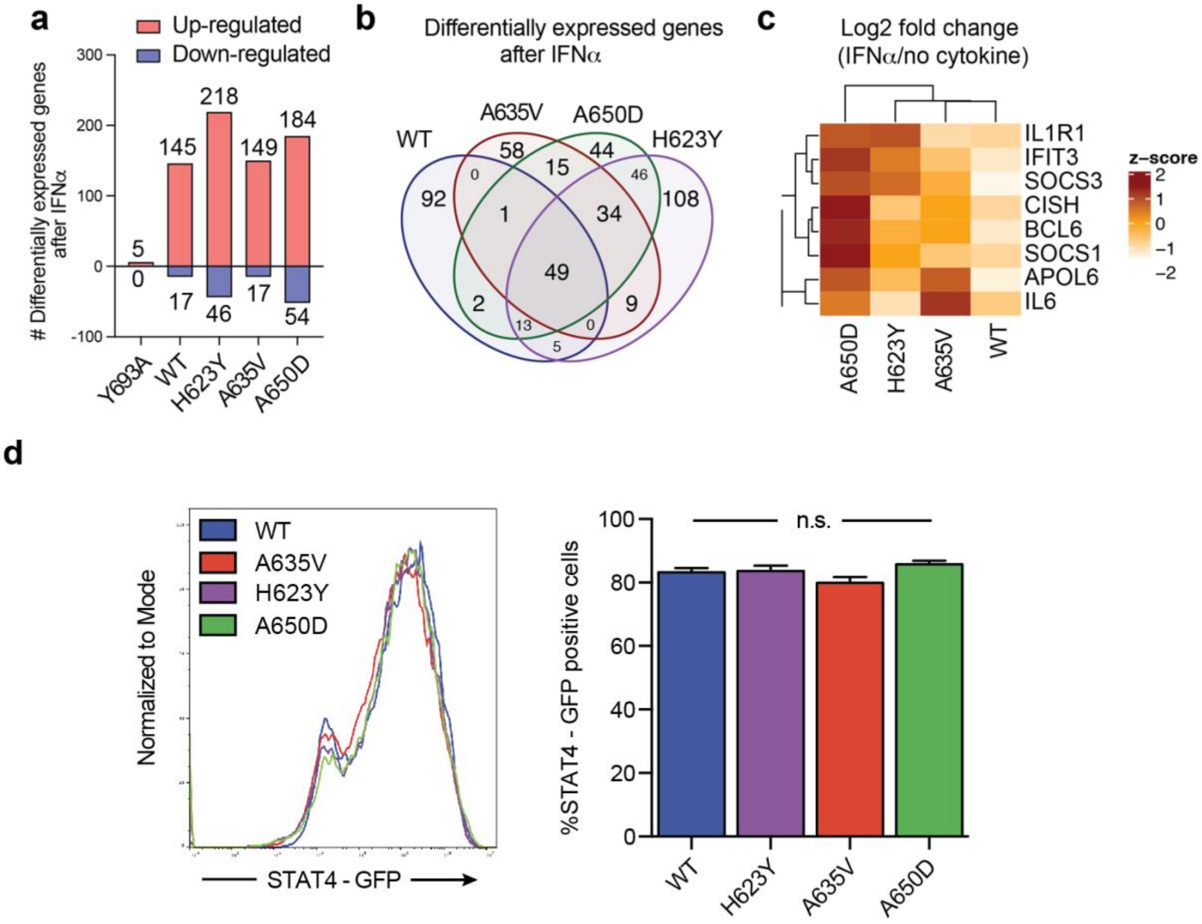


Figure 2.13: RNAseq demonstrates pro-inflammatory state.
(a) Number of differentially expressed genes between unstimulated and IFNα-stimulated U3A cells stably transfected with the indicated STAT4 variants (WT, H623Y, A635V, A650D, or the phospho-dead variant Y693A). Genes differentially expressed after IFNα stimulation were identified using the following criteria: log2 fold change > 0.5, FDR < 0.1, and logCPM > 0.3. (**b)** Venn Diagram showing the number of differentially expressed genes similar and different between each condition. **(c)** Log2 fold change expression of top induced genes with IFNα stimulation. **(d)** Baseline expression of total STAT4 does not vary among unstimulated U3A cells stably transfected with the indicated STAT4 variants (H623Y, A635V, A650D) compared to wildtype (WT) (n=3 per cell line; n.s., not significant).

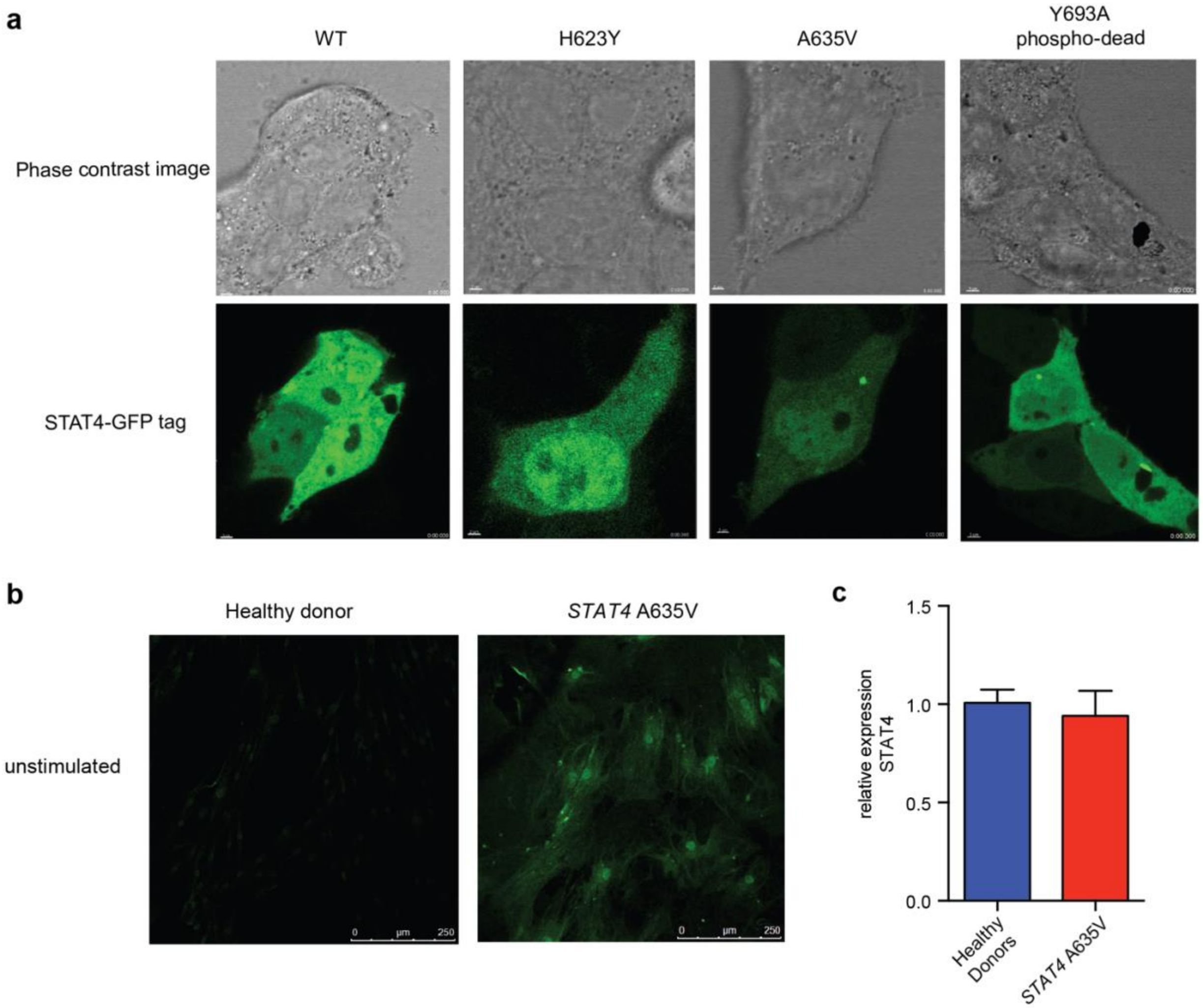


Figure 2.14: Nuclear localization of variant STAT4.
(a) HEK 293T cells transiently transfected with plasmids containing wildtype H623Y, A635V or phospho-dead Y693A STAT4 tagged with GFP. Unstimulated cells transfected with H623Y or A635V variants show a greater accumulation of STAT4 in the nucleus compared to wildtype and Y693A cells in single fluorescence channel images. **(b)** Single channel immunofluorescence imaging of primary skin fibroblasts from patient have prominent pSTAT4 (green) which is not observed in healthy donor fibroblasts at baseline. **(c)** Relative expression of STAT4 mRNA by qPCR in healthy donor and patient derived skin fibroblasts.

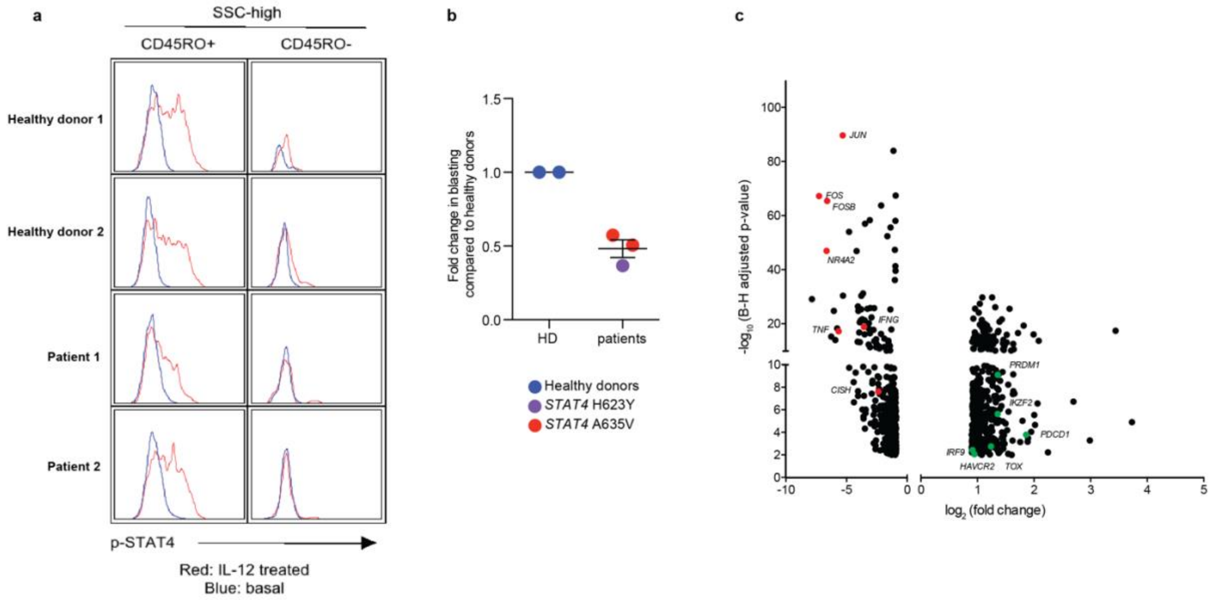


Figure 2.15: pSTAT4 activation and effects on T cells in peripheral blood.

(a) Flow cytometry of CD3+CD4+ CD45RO+ or – cells from 2 healthy donors compared to Family 1 P1 and P2. pSTAT4 measured in response to PHA and IL-12 stimulation. **(b)** PHA-induced blasting was reduced in peripheral blood T cells isolated from patients carrying either the H623Y or A635V variants, compared to healthy donor T cells. **(c)** Differential expression analysis of T cells from untreated patient compared to control, demonstrating upregulation of genes associated with T cell exhaustion (*PDCD1, HAVCR4, PRDM1, IKZF2, IRF9* and *TOX*, shown in green) and downregulation of *JUN, FOS, FOSB, NR4A2, CISH, TNF* and *IFNG*, shown in red. Non-significant genes with low fold changes are not shown. *STAT4* was not differentially expressed among T cell subsets between patient and control samples.

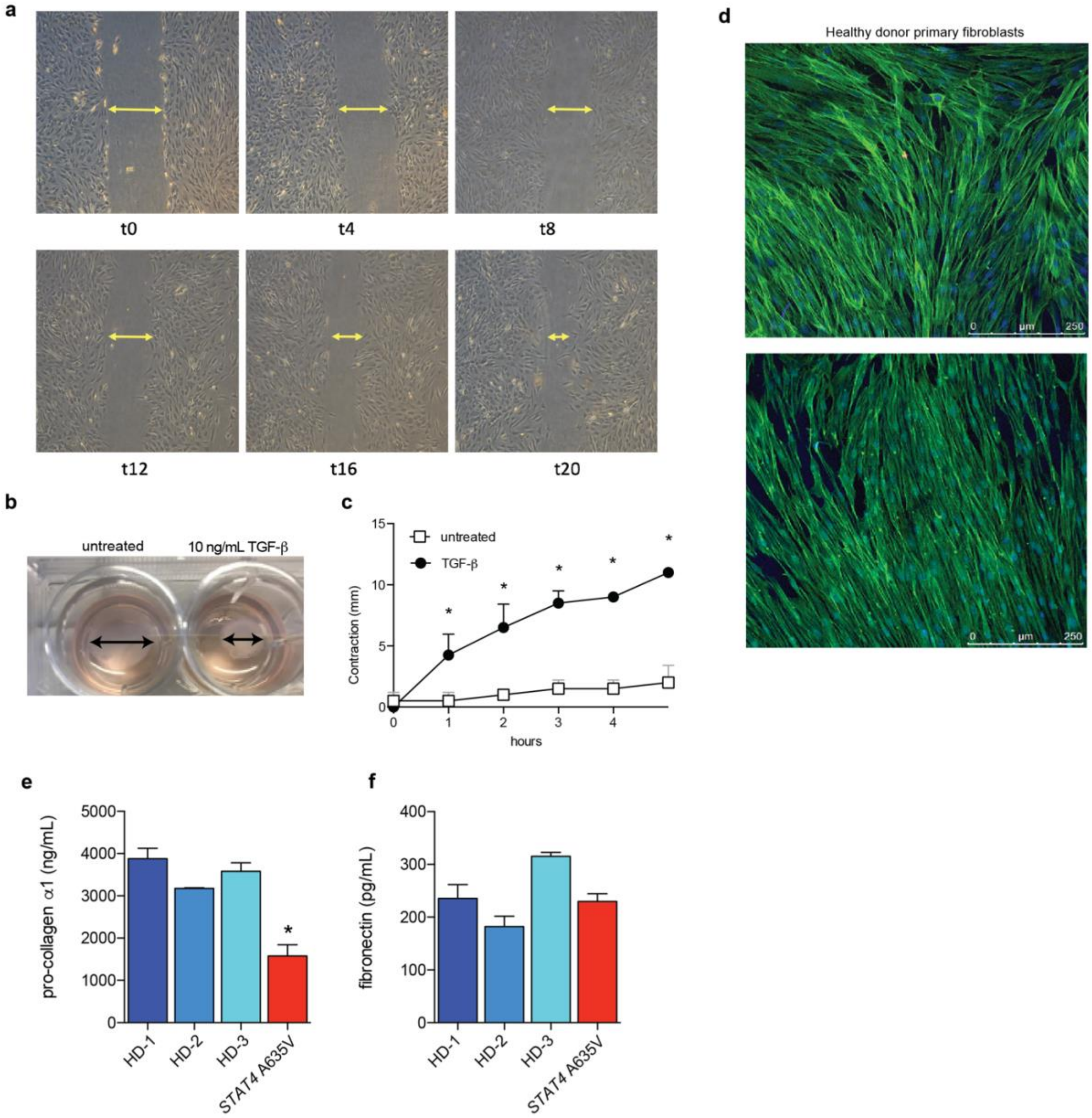


Figure 2.16: Evaluation of fibroblast function in vitro.

(a) Representative wound closure (“scratch”) assay. Confluent monolayers are scratched with a 200-μl pipette tip and diameter of scratch measured every 4-6 hours until closure. **(b,c)** TGF-β induces contraction in healthy donor control skin fibroblasts in a collagen matrix. Primary skin fibroblasts were embedded in a type I bovine collagen matrix, and incubated for 2 days with or without 10ng/mL TGF-β. After release, collagen gel diameter (arrow) was measured hourly. Representative collagen gel disks with contraction in the presence of TGF-β, compared to the untreated disk is shown in **(b)**. (*, p<0.05; n=4). **(d)** F-actin immunocytochemistry of 2 additional healthy donors. **(e,f)** Secretion of pro-collagen α1 **(e)** and fibronectin **(f)** at baseline in cell cultures from patient or healthy donors (n=3).

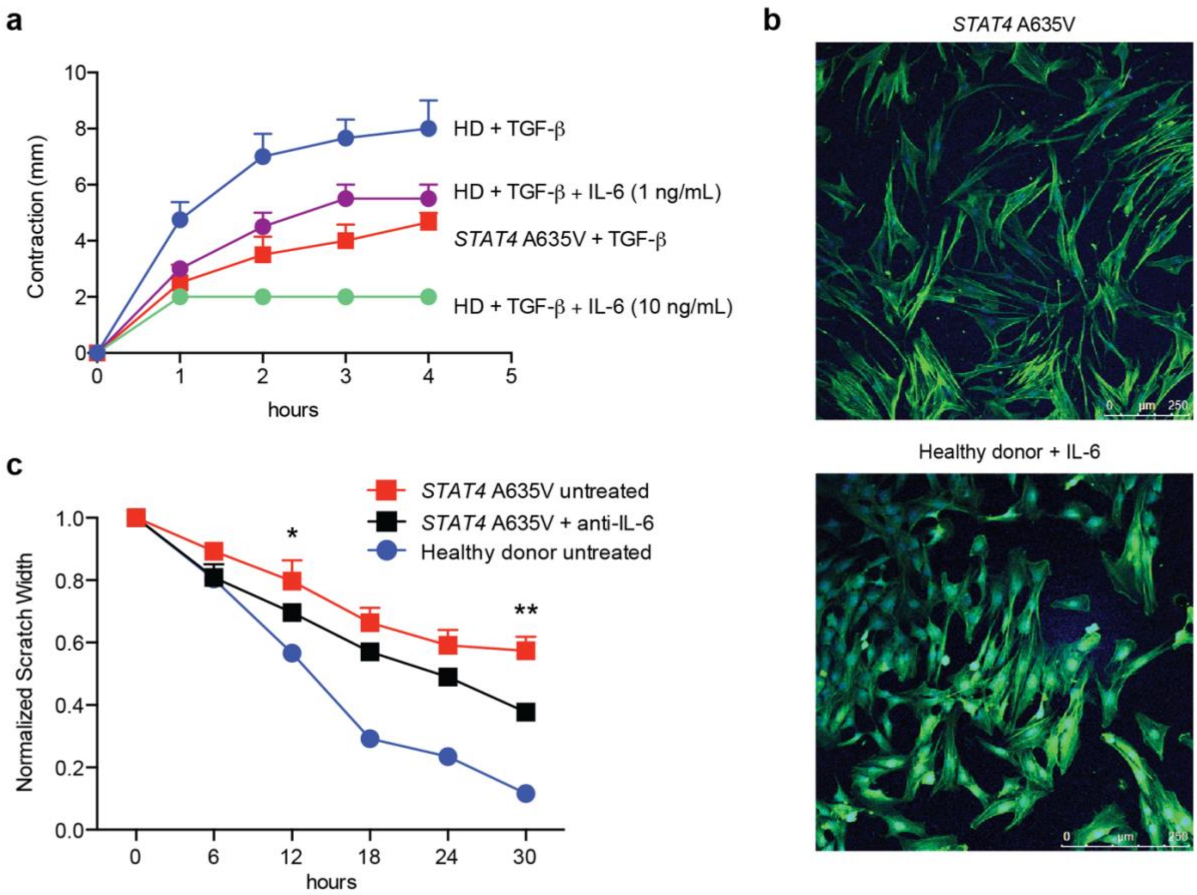


Figure 2.17: Role of IL-6 in fibroblast inflammation.

(a) TGF-β -induced contraction of collagen matrix by healthy donor derived fibroblasts is reduced by IL-6 in a dose-dependent fashion (n=4). **(b)** Treatment of healthy donor fibroblasts with IL-6 (10 ng/mL) leads to enlarged cells (10X, images representative of 2 independent experiments). **(c)** Pre-treatment with anti-IL-6 leads to improved fibroblast migration in wound healing assays. (n=3 experiments, 6 scratches each *, p<0.05; **, p<0.01; by 2-way ANOVA).

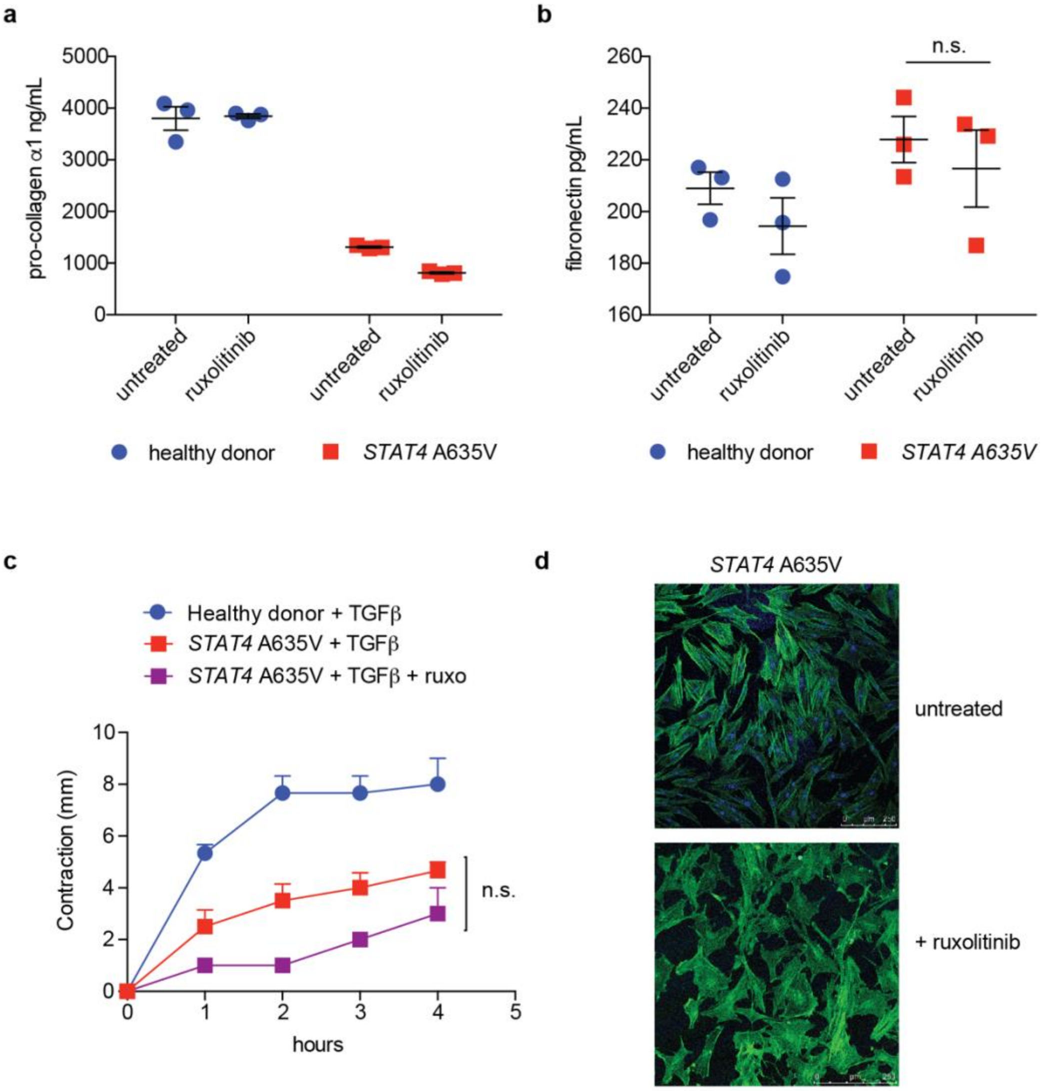


Figure 2.18: Ruxolitinib does not improve other aspects of wound healing in vitro.

(a,b) ELISA of primary skin fibroblast supernatants shows similar secretion of pro-collagen α1 **(a)** and fibronectin **(b)** at baseline, and when treated with ruxolitinib in both patient and normal donor samples (n = 3). **(c)** Collagen contraction induced by TGF-β remains impaired despite treatment with ruxolitinib, compared to healthy donor fibroblasts (n=3). **(d)** F-actin immunocytochemistry shows disorganized distribution and enhanced stress fibers in patient primary skin fibroblasts compared to healthy donor that is not dramatically improved with ruxolitinib treatment. (10X, images representative of 2 independent experiments)

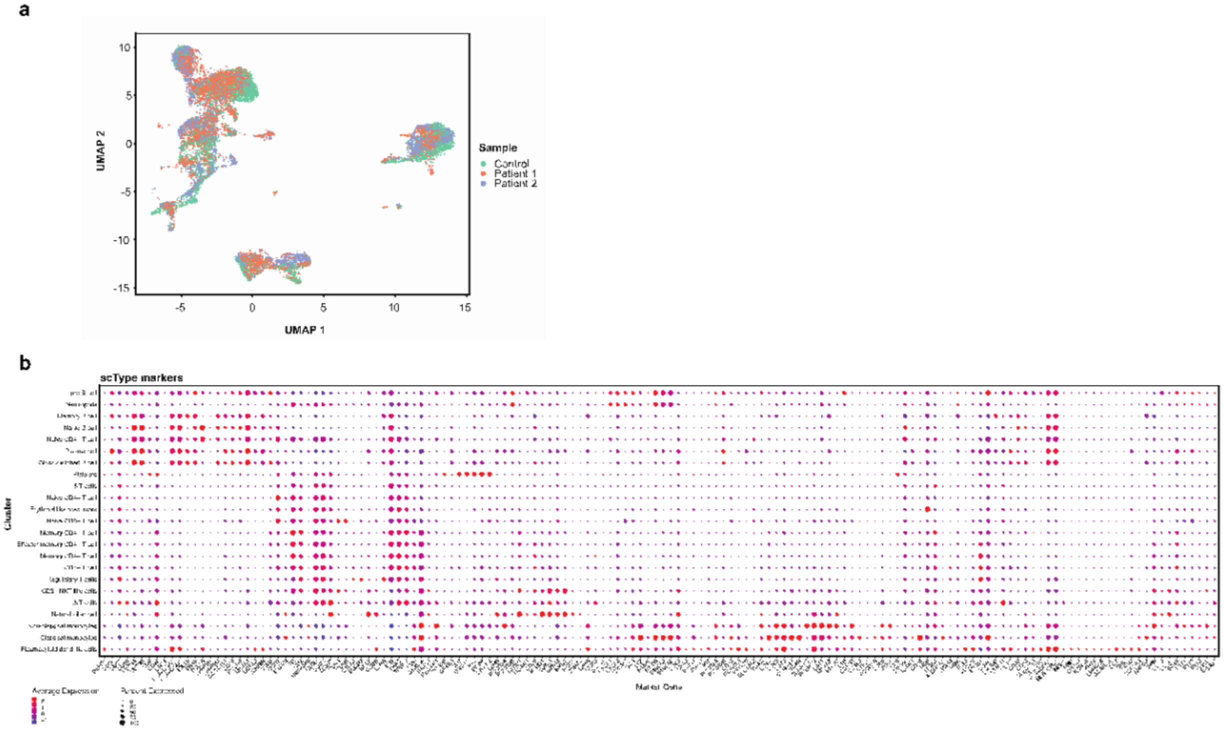


Figure 2.19: Cell type identification and integration of the scRNA-seq datasets.
(a) Cells plotted in UMAP space after running the dimensionality reduction on 47 PCs for the integrated patient dataset, colored by sample. **(b)** Dotplot of canonical PBMC markers (x-axis). Clusters (y-axis, left-hand side) were labeled as cell types based on expression. Markers visualized are taken from those expected to be upregulated in immune tissue according to ScType's database. Dot size reflects the fraction of cells in the cluster expression the marker, and color reflects the average expression of the marker. Expression values are taken from the integrated, scaled expression matrix.

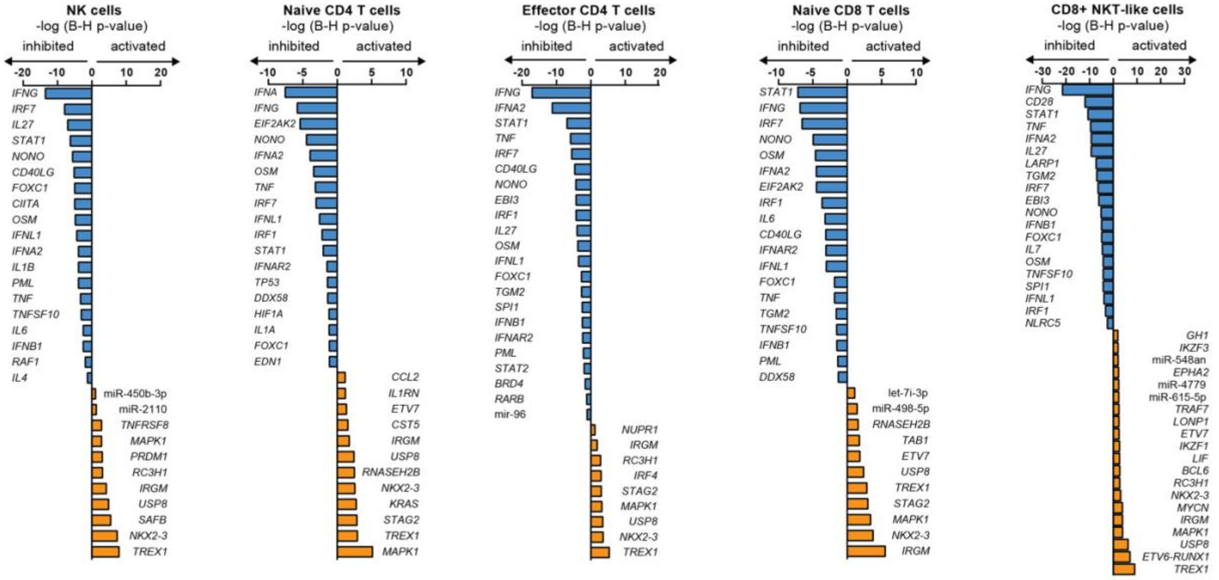


Figure 2.20: Upstream regulators identified by Ingenuity Pathway Analysis (IPA).
IPA analysis of NK cell, CD4+ T cell and CD8+ T cell clusters. Differentially expressed genes used as input to IPA were identified by comparing expression levels in Patient 1 relative to Patient 2 in each cell type. Upstream regulators that are activated are shown in orange, and those that are inhibited are shown in blue.

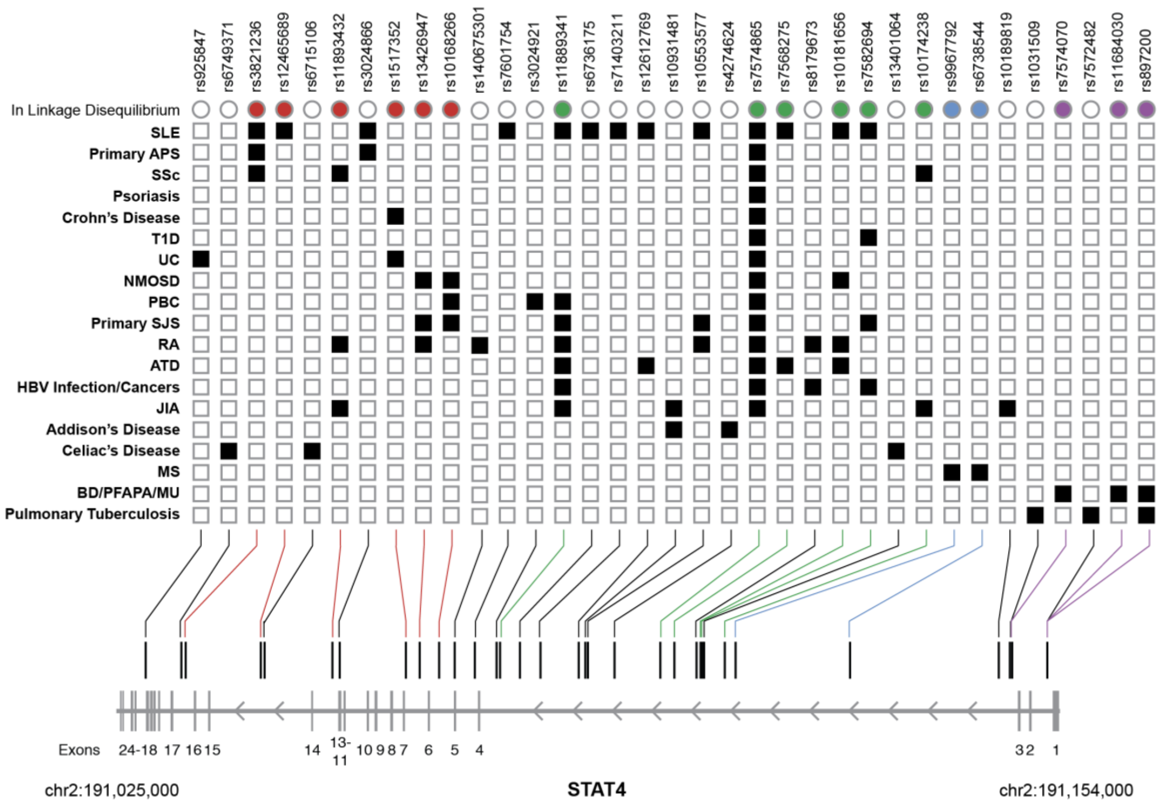


Figure 2.21: SNPs in STAT4 have been associated with immune disease by GWAS studies. **Top.** Filled in squares indicate an association between the SNP (columns) and the disease (rows). Colored circles (red, green, blue, purple) indicate groups of SNPs that are in linkage disequilibrium with each other (r^2>0.8). **Bottom.** Positions of the SNPs within the *STAT4* gene are shown. Coordinates are from the GRCh38 build of the human genome. APS: Antiphospholipid Syndrome; ATD: Autoimmune Thyroid Disease; BD: Behçet disease; JIA: Juvenile Idiopathic Arthritis; HBV: Hepatitis B Virus; MS: Multiple Sclerosis; MU: Mouth Ulcers; NMOSD: Neuromyelitis Optica Spectrum Disorder; PBC: Primary Biliary Cirrhosis or Cholangitis; PFAPA: Periodic fever, aphthous stomatitis, pharyngitis, and cervical adenitis syndrome; RA: Rheumatoid Arthritis; SjS: Sjögren's Syndrome; SLE: Systemic Lupus Erythematous; SSc: Systemic Sclerosis; T1D: Type I Diabetes; UC: Ulcerative Colitis.[26,27,75–117]

2.5.3 Appendix C: Data Availability

The authors declare that the data supporting the findings of this study are available within the paper. Single-cell data can be accessed at GEO (in process). Genome data for individual patients cannot be made publicly available for reasons of patient confidentiality. Researchers may apply for access to these data, pending approval of the individual Institutional Review Boards.

2.5.4 Appendix D: Authors, Contributions, and Acknowledgements

Authors: Hratch Baghdassarian, Sarah A. Blackstone, Owen S. Clay, Rachael Philips, Brynja Matthiasardottir, Michele Nehrebecky, Vivian K. Hua, Rachael McVicar, Yang Liu, Suzanne M. Tucker, Davide Randazzo, Natalie Deuitch, Sofia Rosenzweig, Adam Mark, Roman Sasik, Kathleen M. Fisch, Pallavi Pimpale Chavan, Elif Eren, Norman R. Watts, Chi A. Ma, Massimo Gadina, Daniella M. Schwartz, Anwesha Sanyal, Giffin Werner, David R. Murdock, Nobuyuki Horita, Shimul Chowdhury, David Dimmock, Kristen Jepsen, Elaine F. Remmers, Raphaela Goldbach-Mansky, William A. Gahl, John J. O’Shea, Joshua D. Milner, Nathan E. Lewis, Johanna Chang, Daniel L. Kastner, Kathryn Torok, Hirotsugu Oda, Christopher D. Putnam, Lori Broderick

Lori Broderick, Hratch Baghdassarian, Sarah A. Blackstone, Rachael Philips, Hirotsugu Oda, John J. O’Shea, Joshua D. Milner, Daniel L. Kastner and Christopher D. Putnam conceived the overall experimental designs. Lori Broderick, Hratch Baghdassarian, Nathan Lewis, Vivian K. Hua, David R. Murdock, Nobuyuki Horita, Shimul Chowdhury, David Dimmock, Sofia Rosenzweig, Brynja Matthiasardottir, Elaine Remmers, Adam Mark, Roman Sasik, Kathleen M. Fisch, and Kristen Jepsen performed sequencing and analysis for genomic and single cell RNA sequencing studies. Brynja Matthiasardottir and Rachael Philips performed bulk RNA sequencing and analyses. Christopher D. Putnam, Elif Erin, Norman R. Watts and Hirotsugu Oda performed structural modeling and analyses. Lori Broderick, Owen Clay, Rachael McVicar and Yang Liu initiated and maintained primary fibroblast lines. Lori Broderick, Vivian K. Hua, and Owen Clay performed and analyzed wound healing assays. Lori Broderick, Sarah A. Blackstone, and Pallavi Pimpale Chavan performed and analyzed studies involving *in vitro* expression, plasmid construction, flow cytometry, and ELISA. Lori Broderick, Sarah A. Blackstone, and Davide Randazzo performed immunofluorescence staining and microscopy. Suzanne Tucker performed clinical histology and immunohistochemistry. Chi Ma, Massimo Gadina, Daniella M. Schwartz and Joshua Milner performed and analyzed *ex vivo* flow cytometry. Lori Broderick, Suzanne M. Tucker, Natalie Deuitch, Michele Nehrebecky, Anwesha Sanyal, Giffin Werner, Raphaela Goldbach-Mansky, William A. Gahl, Johanna Chang, and Kathryn Torok were responsible for clinical samples. Lori Broderick, Hratch Baghdassarian, Johanna Chang, Daniel L. Kastner and Christopher D. Putnam wrote the initial draft of the manuscript with contributions from Sarah A. Blackstone, Brynja Matthiasardottir, David Dimmock and Kathryn Torok. All other authors reviewed the final draft and agreed to publish the manuscript.

2.6 References

1. Wollina, U., Buslau, M., Heinig, B., Petrov, I., Unger, E., Kyriopoulou, E., Koch, A., Kostler, E., Schonlebe, J., Haroske, G., Doede, T. & Pramatarov, K. Disabling pansclerotic morphea of childhood poses a high risk of chronic ulceration of the skin and squamous cell carcinoma. *Int. J. Low. Extrem. Wounds* **6,** 291–298 (2007).

2. Moll, M., Holzer, U., Zimmer, C., Rieber, N. & Kuemmerle-Deschner, J. Autologous stem cell transplantation in two children with disabling pansclerotic morphea. *Pediatr. Rheumatol. Online J.* **9,** 77 (2011).

3. Gruss, C., Stucker, M., Kobyletzki, G., Schreiber, D., Altmeyer, P. & Kerscher, M. Low dose UVA1 phototherapy in disabling pansclerotic morphoea of childhood. *Br. J. Dermatol.* **136,** 293–294 (1997).

4. Kowal-Bielecka, O. & Distler, O. Use of methotrexate in patients with scleroderma and mixed connective tissue disease. *Clin. Exp. Rheumatol.* **28,** S160–3 (2010).

5. Forsea, A. M., Cretu, A. N., Ionescu, R. & Giurcaneanu, C. Disabling pansclerotic morphea of childhood--unusual case and management challenges. *J. Med. Life* **1,** 348–354 (2008).

6. Soh, H. J., Samuel, C., Heaton, V., Renton, W. D., Cox, A. & Munro, J. Challenges in the diagnosis and treatment of disabling pansclerotic morphea of childhood: case-based review. *Rheumatol. Int.* **39,** 933–941 (2019).

7. Philips, R. L., Wang, Y., Cheon, H., Kanno, Y., Gadina, M., Sartorelli, V., Horvath, C. M., Darnell, J. E., Jr., Stark, G. R. & O'Shea, J. J. The JAK-STAT pathway at 30: Much learned, much more to do. *Cell* **185,** 3857–3876 (2022).

8. Shuai, K. Modulation of STAT signaling by STAT-interacting proteins. *Oncogene* **19,** 2638–2644 (2000).

9. Xin, P., Xu, X., Deng, C., Liu, S., Wang, Y., Zhou, X., Ma, H., Wei, D. & Sun, S. The role of JAK/STAT signaling pathway and its inhibitors in diseases. *Int. Immunopharmacol.* **80,** 106210 (2020).

10. Nguyen, H. N., Noss, E. H., Mizoguchi, F., Huppertz, C., Wei, K. S., Watts, G. F. M. & Brenner, M. B. Autocrine Loop Involving IL-6 Family Member LIF, LIF Receptor, and STAT4 Drives Sustained Fibroblast Production of Inflammatory Mediators. *Immunity* **46,** 220–232 (2017).

11. Kurzinski, K. & Torok, K. S. Cytokine profiles in localized scleroderma and relationship to clinical features. *Cytokine* **55,** 157–164 (2011).

12. Torok, K. S., Kurzinski, K., Kelsey, C., Yabes, J., Magee, K., Vallejo, A. N., Medsger, T., Jr. & Feghali-Bostwick, C. A. Peripheral blood cytokine and chemokine profiles in juvenile localized scleroderma: T-helper cell-associated cytokine profiles. *Semin. Arthritis Rheum.* **45,** 284–293 (2015).

13. West, N. R. Coordination of Immune-Stroma Crosstalk by IL-6 Family Cytokines. *Front. Immunol.* **10,** 1093 (2019).

14. Chakraborty, A., Dyer, K. F., Cascio, M., Mietzner, T. A. & Tweardy, D. J. Identification of a novel Stat3 recruitment and activation motif within the granulocyte colony-stimulating factor receptor. *Blood* **93,** 15–24 (1999).

15. Holland, S. M., DeLeo, F. R., Elloumi, H. Z., Hsu, A. P., Uzel, G., Brodsky, N., Freeman, A. F., Demidowich, A., Davis, J., Turner, M. L., Anderson, V. L., Darnell, D. N., Welch, P. A., Kuhns, D. B., Frucht, D. M., Malech, H. L., Gallin, J. I., Kobayashi, S. D., Whitney, A. R., Voyich, J. M., Musser, J. M., Woellner, C., Schaffer, A. A., Puck, J. M. & Grimbacher, B. STAT3 mutations in the hyper-IgE syndrome. *N. Engl. J. Med.* **357,** 1608–1619 (2007).

16. Jagle, S., Heeg, M., Grun, S., Rensing-Ehl, A., Maccari, M. E., Klemann, C., Jones, N., Lehmberg, K., Bettoni, C., Warnatz, K., Grimbacher, B., Biebl, A., Schauer, U., Hague, R., Neth, O., Mauracher, A., Pachlopnik Schmid, J., Fabre, A., Kostyuchenko, L., Fuhrer, M., Lorenz, M. R., Schwarz, K., Rohr, J. & Ehl, S. Distinct molecular response patterns of activating STAT3 mutations associate with penetrance of lymphoproliferation and autoimmunity. *Clin. Immunol.* **210,** 108316 (2020).

17. Okada, S., Asano, T., Moriya, K., Boisson-Dupuis, S., Kobayashi, M., Casanova, J. L. & Puel, A. Human STAT1 Gain-of-Function Heterozygous Mutations: Chronic Mucocutaneous Candidiasis and Type I Interferonopathy. *J. Clin. Immunol.* **40,** 1065–1081 (2020).

18. Mao, X., Ren, Z., Parker, G. N., Sondermann, H., Pastorello, M. A., Wang, W., McMurray, J. S., Demeler, B., Darnell, J. E., Jr. & Chen, X. Structural bases of unphosphorylated STAT1 association and receptor binding. *Mol. Cell* **17,** 761–771 (2005).

19. Nishio, H., Matsui, K., Tsuji, H., Tamura, A. & Suzuki, K. Immunolocalisation of the janus kinases (JAK)--signal transducers and activators of transcription (STAT) pathway in human epidermis. *J. Anat.* **198,** 581–589 (2001).

20. Yamamoto, K., Kobayashi, H., Arai, A., Miura, O., Hirosawa, S. & Miyasaka, N. cDNA cloning, expression and chromosome mapping of the human STAT4 gene: both STAT4 and STAT1 genes are mapped to 2q32.2-->q32.3. *Cytogenet. Cell Genet.* **77,** 207–210 (1997).

21. Zhong, Z., Wen, Z. & Darnell, J. E., Jr. Stat3 and Stat4: members of the family of signal transducers and activators of transcription. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 4806–4810 (1994).

22. Ehrlich, H. P. & Wyler, D. J. Fibroblast contraction of collagen lattices in vitro: inhibition by chronic inflammatory cell mediators. *J. Cell. Physiol.* **116,** 345–351 (1983).

23. Kendall, R. T. & Feghali-Bostwick, C. A. Fibroblasts in fibrosis: novel roles and mediators. *Front. Pharmacol.* **5,** 123 (2014).

24. Diaz-Perez, J. L., Connolly, S. M. & Winkelmann, R. K. Disabling pansclerotic morphea of children. *Arch. Dermatol.* **116,** 169–173 (1980).

25. Wurster, A. L., Tanaka, T. & Grusby, M. J. The biology of Stat4 and Stat6. *Oncogene* **19,** 2577–2584 (2000).

26. Lessard, C. J., Li, H., Adrianto, I., Ice, J. A., Rasmussen, A., Grundahl, K. M., Kelly, J. A., Dozmorov, M. G., Miceli-Richard, C., Bowman, S., Lester, S., Eriksson, P., Eloranta, M. L., Brun, J. G., Goransson, L. G., Harboe, E., Guthridge, J. M., Kaufman, K. M., Kvarnstrom, M., Jazebi, H., Cunninghame Graham, D. S., Grandits, M. E., Nazmul-Hossain, A. N., Patel, K.,

Adler, A. J., Maier-Moore, J. S., Farris, A. D., Brennan, M. T., Lessard, J. A., Chodosh, J., Gopalakrishnan, R., Hefner, K. S., Houston, G. D., Huang, A. J., Hughes, P. J., Lewis, D. M., Radfar, L., Rohrer, M. D., Stone, D. U., Wren, J. D., Vyse, T. J., Gaffney, P. M., James, J. A., Omdal, R., Wahren-Herlenius, M., Illei, G. G., Witte, T., Jonsson, R., Rischmueller, M., Ronnblom, L., Nordmark, G., Ng, W. F., Registry, U. K. P. S. S., Mariette, X., Anaya, J. M., Rhodus, N. L., Segal, B. M., Scofield, R. H., Montgomery, C. G., Harley, J. B. & Sivils, K. L. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjogren's syndrome. *Nat. Genet.* **45,** 1284–1292 (2013).

27. Remmers, E. F., Plenge, R. M., Lee, A. T., Graham, R. R., Hom, G., Behrens, T. W., de Bakker, P. I., Le, J. M., Lee, H. S., Batliwalla, F., Li, W., Masters, S. L., Booty, M. G., Carulli, J. P., Padyukov, L., Alfredsson, L., Klareskog, L., Chen, W. V., Amos, C. I., Criswell, L. A., Seldin, M. F., Kastner, D. L. & Gregersen, P. K. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.* **357,** 977–986 (2007).

28. Diehl, S., Anguita, J., Hoffmeyer, A., Zapton, T., Ihle, J. N., Fikrig, E. & Rincon, M. Inhibition of Th1 differentiation by IL-6 is mediated by SOCS1. *Immunity* **13,** 805–815 (2000).

29. Mognol, G. P., Spreafico, R., Wong, V., Scott-Browne, J. P., Togher, S., Hoffmann, A., Hogan, P. G., Rao, A. & Trifari, S. Exhaustion-associated regulatory regions in CD8(+) tumor-infiltrating T cells. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E2776–E2785 (2017).

30. Kurachi, M. CD8(+) T cell exhaustion. *Semin. Immunopathol.* **41,** 327–337 (2019).

31. Khanna, D., Lin, C. J. F., Furst, D. E., Wagner, B., Zucchetto, M., Raghu, G., Martinez, F. J., Goldin, J., Siegel, J. & Denton, C. P. Long-Term Safety and Efficacy of Tocilizumab in Early Systemic Sclerosis-Interstitial Lung Disease: Open-Label Extension of a Phase 3 Randomized Controlled Trial. *Am. J. Respir. Crit. Care Med.* **205,** 674–684 (2022).

32. Farnaes, L., Hildreth, A., Sweeney, N. M., Clark, M. M., Chowdhury, S., Nahas, S., Cakici, J. A., Benson, W., Kaplan, R. H., Kronick, R., Bainbridge, M. N., Friedman, J., Gold, J. J., Ding, Y., Veeraraghavan, N., Dimmock, D. & Kingsmore, S. F. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom Med* **3,** 10 (2018).

33. Milko, L. V., Chen, F., Chan, K., Brower, A. M., Agrawal, P. B., Beggs, A. H., Berg, J. S., Brenner, S. E., Holm, I. A., Koenig, B. A., Parad, R. B., Powell, C. M. & Kingsmore, S. F. FDA oversight of NSIGHT genomic research: the need for an integrated systems approach to regulation. *NPJ Genom Med* **4,** 32 (2019).

34. Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., The Exome Aggregation, Consortium, Daly, M. J. & MacArthur, D. G. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45,** D840–D845 (2017).

35. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12,** 841–843 (2015).

36. Kohler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Ayme, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., Brudno, M., Buske, O. J., Chinnery, P. F., Cipriani, V., Connell, L. E., Dawkins, H. J., DeMare, L. E., Devereau, A. D., de Vries, B. B., Firth, H. V., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jahn, J. A., James, R., Krause, R., Sj, F. L., Lochmuller, H., Lyon, G. J., Ogishima, S., Olry, A., Ouwehand, W. H.,

Pontikos, N., Rath, A., Schaefer, F., Scott, R. H., Segal, M., Sergouniotis, P. I., Sever, R., Smith, C. L., Straub, V., Thompson, R., Turner, C., Turro, E., Veltman, M. W., Vulliamy, T., Yu, J., von Ziegenweidt, J., Zankl, A., Zuchner, S., Zemojtel, T., Jacobsen, J. O., Groza, T., Smedley, D., Mungall, C. J., Haendel, M. & Robinson, P. N. The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45,** D865–D876 (2017).

37. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A. J., Kruglyak, S. & Saunders, C. T. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222 (2016).

38. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21,** 974–984 (2011).

39. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L. & Committee, Acmg Laboratory Quality Assurance. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–424 (2015).

40. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

41. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

42. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

43. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

44. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50,** D439–D444 (2022).

45. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

46. Chen, X., Vinkemeier, U., Zhao, Y., Jeruzalmi, D., Darnell, J. E., Jr. & Kuriyan, J. Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell* **93,** 827–839 (1998).

47. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501 (2010).

48. Land, H. & Humble, M. S. YASARA: A Tool to Obtain Structural Guidance in Biocatalytic Investigations. *Methods Mol. Biol.* **1685,** 43–67 (2018).

49. Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., Felts, A. K., Halgren, T. A., Mainz, D. T., Maple, J. R., Murphy, R., Philipp, D. M., Repasky, M. P., Zhang, L. Y., Berne, B. J., Friesner, R. A., Gallicchio, E. & Levy, R. M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **26,** 1752–1780 (2005).

50. Vangipuram, M., Ting, D., Kim, S., Diaz, R. & Schule, B. Skin punch biopsy explant culture for derivation of primary human fibroblasts. *J. Vis. Exp.* e3779 (2013).

51. Muller, M., Laxton, C., Briscoe, J., Schindler, C., Improta, T., Darnell, J. E., Jr., Stark, G. R. & Kerr, I. M. Complementation of a mutant cell line: central role of the 91 kDa polypeptide of ISGF3 in the interferon-alpha and -gamma signal transduction pathways. *EMBO J.* **12,** 4221–4228 (1993).

52. Torpey, N., Maher, S. E., Bothwell, A. L. & Pober, J. S. Interferon alpha but not interleukin 12 activates STAT4 signaling in human vascular endothelial cells. *J. Biol. Chem.* **279,** 26789–26796 (2004).

53. Kim, S. I., Oceguera-Yanez, F., Sakurai, C., Nakagawa, M., Yamanaka, S. & Woltjen, K. Inducible Transgene Expression in Human iPS Cells Using Versatile All-in-One piggyBac Transposons. *Methods Mol. Biol.* **1357,** 111–131 (2016).

54. Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** 14049 (2017).

55. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. & Satija, R. Integrated analysis of multimodal single-cell data. *Cell* **184,** 3573–3587 e29 (2021).

56. Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C. & Teichmann, S. A. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17,** 29 (2016).

57. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15,** e8746 (2019).

58. Buttner, M., Ostner, J., Muller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12,** 6876 (2021).

59. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett,

D. A., Kellis, M. & Tsai, L. H. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570,** 332–337 (2019).

60. Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13,** 1246 (2022).

61. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S. & Gottardo, R. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16,** 278 (2015).

62. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D. & Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).

63. Gorissen, M., de Vrieze, E., Flik, G. & Huising, M. O. STAT genes display differential evolutionary rates that correlate with their roles in the endocrine and immune system. *J. Endocrinol.* **209,** 175–184 (2011).

64. Meesilpavikkai, K., Dik, W. A., Schrijver, B., Nagtzaam, N. M., van Rijswijk, A., Driessen, G. J., van der Spek, P. J., van Hagen, P. M. & Dalm, V. A. A Novel Heterozygous Mutation in the STAT1 SH2 Domain Causes Chronic Mucocutaneous Candidiasis, Atypically Diverse Infections, Autoimmunity, and Impaired Cytokine Regulation. *Front. Immunol.* **8,** 274 (2017).

65. Ovadia, A., Sharfe, N., Hawkins, C., Laughlin, S. & Roifman, C. M. Two different STAT1 gain-of-function mutations lead to diverse IFN-gamma-mediated gene expression. *NPJ Genom Med* **3,** 23 (2018).

66. Sobh, A., Chou, J., Schneider, L., Geha, R. S. & Massaad, M. J. Chronic mucocutaneous candidiasis associated with an SH2 domain gain-of-function mutation that enhances STAT1 phosphorylation. *J. Allergy Clin. Immunol.* **138,** 297–299 (2016).

67. Feng, B. J. PERCH: A Unified Framework for Disease Gene Prioritization. *Hum. Mutat.* **38,** 243–251 (2017).

68. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47,** D886–D894 (2019).

69. Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N. & Gaunt, T. R. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34,** 57–65 (2013).

70. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A. & Batzoglou, S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).

71. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

72. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., Powell, I. J., Cussenot, O., Cancel-Tassin, G., Giles, G. G., MacInnis, R. J., Maier, C., Hsieh, C. L., Wiklund, F., Catalona, W. J., Foulkes, W. D., Mandal, D., Eeles, R. A., Kote-Jarai, Z., Bustamante, C. D., Schaid, D. J., Hastie, T., Ostrander, E. A., Bailey-Wilson, J. E., Radivojac, P., Thibodeau, S. N., Whittemore, A. S. & Sieh, W. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99,** 877–885 (2016).

73. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11,** 863–874 (2001).

74. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3,** S3 (2013).

75. Abelson, A. K., Delgado-Vega, A. M., Kozyrev, S. V., Sanchez, E., Velazquez-Cruz, R., Eriksson, N., Wojcik, J., Linga Reddy, M. V., Lima, G., D'Alfonso, S., Migliaresi, S., Baca, V., Orozco, L., Witte, T., Ortego-Centeno, N., Aadea group, Abderrahim, H., Pons-Estel, B. A., Gutierrez, C., Suarez, A., Gonzalez-Escribano, M. F., Martin, J. & Alarcon-Riquelme, M. E. STAT4 associates with systemic lupus erythematosus through two independent effects that correlate with gene expression and act additively with IRF5 to increase risk. *Ann. Rheum. Dis.* **68,** 1746–1753 (2009).

76. Aiba, Y., Yamazaki, K., Nishida, N., Kawashima, M., Hitomi, Y., Nakamura, H., Komori, A., Fuyuno, Y., Takahashi, A., Kawaguchi, T., Takazoe, M., Suzuki, Y., Motoya, S., Matsui, T., Esaki, M., Matsumoto, T., Kubo, M., Tokunaga, K. & Nakamura, M. Disease susceptibility genes shared by primary biliary cirrhosis and Crohn's disease in the Japanese population. *J. Hum. Genet.* **60,** 525–531 (2015).

77. Bi, C., Li, B., Cheng, Z., Hu, Y., Fang, Z. & Zhai, A. Association study of STAT4 polymorphisms and type 1 diabetes in Northeastern Chinese Han population. *Tissue Antigens* **81,** 137–140 (2013).

78. Clark, A., Gerlach, F., Tong, H., Hoan, N. X., Song le, H., Toan, N. L., Bock, C. T., Kremsner, P. G. & Velavan, T. P. A trivial role of STAT4 variant in chronic hepatitis B induced hepatocellular carcinoma. *Infect. Genet. Evol.* **18,** 257–261 (2013).

79. Diaz-Gallo, L. M., Palomino-Morales, R. J., Gomez-Garcia, M., Cardena, C., Rodrigo, L., Nieto, A., Alcain, G., Cueto, I., Lopez-Nevot, M. A. & Martin, J. STAT4 gene influences genetic predisposition to ulcerative colitis but not Crohn's disease in the Spanish population: a replication study. *Hum. Immunol.* **71,** 515–519 (2010).

80. Dieude, P., Guedj, M., Wipff, J., Ruiz, B., Hachulla, E., Diot, E., Granel, B., Sibilia, J., Tiev, K., Mouthon, L., Cracowski, J. L., Carpentier, P. H., Amoura, Z., Fajardy, I., Avouac, J., Meyer, O., Kahan, A., Boileau, C. & Allanore, Y. STAT4 is a genetic risk factor for systemic sclerosis having additive effects with IRF5 on disease susceptibility and related pulmonary fibrosis. *Arthritis Rheum.* **60,** 2472–2479 (2009).

81. Fan, Z. D., Wang, F. F., Huang, H., Huang, N., Ma, H. H., Guo, Y. H., Zhang, Y. Y., Qian, X. Q. & Yu, H. G. STAT4 rs7574865 G/T and PTPN22 rs2488457 G/C polymorphisms influence the risk of developing juvenile idiopathic arthritis in Han Chinese patients. *PLoS One* **10,** e0117389 (2015).

82. Glas, J., Seiderer, J., Nagy, M., Fries, C., Beigel, F., Weidinger, M., Pfennig, S., Klein, W., Epplen, J. T., Lohse, P., Folwaczny, M., Goke, B., Ochsenkuhn, T., Diegelmann, J., Muller-Myhsok, B., Roeske, D. & Brand, S. Evidence for STAT4 as a common autoimmune gene: rs7574865 is associated with colonic Crohn's disease and early disease onset. *PLoS One* **5,** e10373 (2010).

83. Han, J. W., Zheng, H. F., Cui, Y., Sun, L. D., Ye, D. Q., Hu, Z., Xu, J. H., Cai, Z. M., Huang, W., Zhao, G. P., Xie, H. F., Fang, H., Lu, Q. J., Xu, J. H., Li, X. P., Pan, Y. F., Deng, D. Q., Zeng, F. Q., Ye, Z. Z., Zhang, X. Y., Wang, Q. W., Hao, F., Ma, L., Zuo, X. B., Zhou, F. S., Du, W. H., Cheng, Y. L., Yang, J. Q., Shen, S. K., Li, J., Sheng, Y. J., Zuo, X. X., Zhu, W. F., Gao, F., Zhang, P. L., Guo, Q., Li, B., Gao, M., Xiao, F. L., Quan, C., Zhang, C., Zhang, Z., Zhu, K. J., Li, Y., Hu, D. Y., Lu, W. S., Huang, J. L., Liu, S. X., Li, H., Ren, Y. Q., Wang, Z. X., Yang, C. J., Wang, P. G., Zhou, W. M., Lv, Y. M., Zhang, A. P., Zhang, S. Q., Lin, D., Li, Y., Low, H. Q., Shen, M., Zhai, Z. F., Wang, Y., Zhang, F. Y., Yang, S., Liu, J. J. & Zhang, X. J. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1234–1237 (2009).

84. Hou, S., Yang, Z., Du, L., Jiang, Z., Shu, Q., Chen, Y., Li, F., Zhou, Q., Ohno, S., Chen, R., Kijlstra, A., Rosenbaum, J. T. & Yang, P. Identification of a susceptibility locus in STAT4 for Behcet's disease in Han Chinese in a genome-wide association study. *Arthritis Rheum.* **64,** 4104–4113 (2012).

85. Huang, X., Wang, Z., Jia, N., Shangguan, S., Lai, J., Cui, X., Wu, F. & Wang, L. Association between STAT4 polymorphisms and the risk of juvenile idiopathic arthritis in Han Chinese populations. *Clin. Exp. Rheumatol.* **37,** 333–337 (2019).

86. Jiang, D. K., Sun, J., Cao, G., Liu, Y., Lin, D., Gao, Y. Z., Ren, W. H., Long, X. D., Zhang, H., Ma, X. P., Wang, Z., Jiang, W., Chen, T. Y., Gao, Y., Sun, L. D., Long, J. R., Huang, H. X., Wang, D., Yu, H., Zhang, P., Tang, L. S., Peng, B., Cai, H., Liu, T. T., Zhou, P., Liu, F., Lin, X., Tao, S., Wan, B., Sai-Yin, H. X., Qin, L. X., Yin, J., Liu, L., Wu, C., Pei, Y., Zhou, Y. F., Zhai, Y., Lu, P. X., Tan, A., Zuo, X. B., Fan, J., Chang, J., Gu, X., Wang, N. J., Li, Y., Liu, Y. K., Zhai, K., Zhang, H., Hu, Z., Liu, J., Yi, Q., Xiang, Y., Shi, R., Ding, Q., Zheng, W., Shu, X. O., Mo, Z., Shugart, Y. Y., Zhang, X. J., Zhou, G., Shen, H., Zheng, S. L., Xu, J. & Yu, L. Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. *Nat. Genet.* **45,** 72–75 (2013).

87. Kim, L. H., Cheong, H. S., Namgoong, S., Kim, J. O., Kim, J. H., Park, B. L., Cho, S. W., Park, N. H., Cheong, J. Y., Koh, I., Shin, H. D. & Kim, Y. J. Replication of genome wide association studies on hepatocellular carcinoma susceptibility loci of STAT4 and HLA-DQ in a Korean population. *Infect. Genet. Evol.* **33,** 72–76 (2015).

88. Kobayashi, S., Ikari, K., Kaneko, H., Kochi, Y., Yamamoto, K., Shimane, K., Nakamura, Y., Toyama, Y., Mochizuki, T., Tsukahara, S., Kawaguchi, Y., Terai, C., Hara, M., Tomatsu, T., Yamanaka, H., Horiuchi, T., Tao, K., Yasutomo, K., Hamada, D., Yasui, N., Inoue, H., Itakura, M., Okamoto, H., Kamatani, N. & Momohara, S. Association of STAT4 with susceptibility to rheumatoid arthritis and systemic lupus erythematosus in the Japanese population. *Arthritis Rheum.* **58,** 1940–1946 (2008).

89. Korman, B. D., Alba, M. I., Le, J. M., Alevizos, I., Smith, J. A., Nikolov, N. P., Kastner, D. L., Remmers, E. F. & Illei, G. G. Variant form of STAT4 is associated with primary Sjogren's syndrome. *Genes Immun.* **9,** 267–270 (2008).

90. Lee, H. S., Park, H., Yang, S., Kim, D. & Park, Y. STAT4 polymorphism is associated with early-onset type 1 diabetes, but not with late-onset type 1 diabetes. *Ann. N. Y. Acad. Sci.* **1150,** 93–98 (2008).

91. Lee, H. S., Remmers, E. F., Le, J. M., Kastner, D. L., Bae, S. C. & Gregersen, P. K. Association of STAT4 with rheumatoid arthritis in the Korean population. *Mol. Med.* **13,** 455–460 (2007).

92. Li, Y., Zhang, K., Chen, H., Sun, F., Xu, J., Wu, Z., Li, P., Zhang, L., Du, Y., Luan, H., Li, X., Wu, L., Li, H., Wu, H., Li, X., Li, X., Zhang, X., Gong, L., Dai, L., Sun, L., Zuo, X., Xu, J., Gong, H., Li, Z., Tong, S., Wu, M., Li, X., Xiao, W., Wang, G., Zhu, P., Shen, M., Liu, S., Zhao, D., Liu, W., Wang, Y., Huang, C., Jiang, Q., Liu, G., Liu, B., Hu, S., Zhang, W., Zhang, Z., You, X., Li, M., Hao, W., Zhao, C., Leng, X., Bi, L., Wang, Y., Zhang, F., Shi, Q., Qi, W., Zhang, X., Jia, Y., Su, J., Li, Q., Hou, Y., Wu, Q., Xu, D., Zheng, W., Zhang, M., Wang, Q., Fei, Y., Zhang, X., Li, J., Jiang, Y., Tian, X., Zhao, L., Wang, L., Zhou, B., Li, Y., Zhao, Y., Zeng, X., Ott, J., Wang, J. & Zhang, F. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjogren's syndrome at 7q11.23. *Nat. Genet.* **45,** 1361–1365 (2013).

93. Lu, Y., Zhu, Y., Peng, J., Wang, X., Wang, F. & Sun, Z. STAT4 genetic polymorphisms association with spontaneous clearance of hepatitis B virus infection. *Immunol. Res.* **62,** 146–152 (2015).

94. Manthiram, K., Preite, S., Dedeoglu, F., Demir, S., Ozen, S., Edwards, K. M., Lapidus, S., Katz, A. E., Genomic Ascertainment, Cohort, Feder, H. M., Jr., Lawton, M., Licameli, G. R., Wright, P. F., Le, J., Barron, K. S., Ombrello, A. K., Barham, B., Romeo, T., Jones, A., Srinivasalu, H., Mudd, P. A., DeBiasi, R. L., Gul, A., Marshall, G. S., Jones, O. Y., Chandrasekharappa, S. C., Stepanovskiy, Y., Ferguson, P. J., Schwartzberg, P. L., Remmers, E. F. & Kastner, D. L. Common genetic susceptibility loci link PFAPA syndrome, Behcet's disease, and recurrent aphthous stomatitis. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 14405–14411 (2020).

95. Mirkazemi, S., Akbarian, M., Jamshidi, A. R., Mansouri, R., Ghoroghi, S., Salimi, Y., Tahmasebi, Z. & Mahmoudi, M. Association of STAT4 rs7574865 with susceptibility to systemic lupus erythematosus in Iranian population. *Inflammation* **36,** 1548–1552 (2013).

96. Mitchell, A. L., Macarthur, K. D., Gan, E. H., Baggott, L. E., Wolff, A. S., Skinningsrud, B., Platt, H., Short, A., Lobell, A., Kampe, O., Bensing, S., Betterle, C., Kasperlik-Zaluska, A., Zurawek, M., Fichna, M., Kockum, I., Nordling Eriksson, G., Ekwall, O., Wahlberg, J., Dahlqvist, P., Hulting, A. L., Penna-Martinez, M., Meyer, G., Kahles, H., Badenhoop, K., Hahner, S., Quinkler, M., Falorni, A., Phipps-Green, A., Merriman, T. R., Ollier, W., Cordell, H. J., Undlien, D., Czarnocka, B., Husebye, E. & Pearce, S. H. Association of autoimmune Addison's disease with alleles of STAT4 and GATA3 in European cohorts. *PLoS One* **9,** e88991 (2014).

97. Moon, C. M., Cheon, J. H., Kim, S. W., Shin, D. J., Kim, E. S., Shin, E. S., Kang, Y., Park, J. J., Hong, S. P., Nam, S. Y., Kim, T. I. & Kim, W. H. Association of signal transducer and activator of transcription 4 genetic variants with extra-intestinal manifestations in inflammatory bowel disease. *Life Sci.* **86,** 661–667 (2010).

98. Nordmark, G., Kristjansdottir, G., Theander, E., Appel, S., Eriksson, P., Vasaitis, L., Kvarnstrom, M., Delaleu, N., Lundmark, P., Lundmark, A., Sjowall, C., Brun, J. G., Jonsson,

M. V., Harboe, E., Goransson, L. G., Johnsen, S. J., Soderkvist, P., Eloranta, M. L., Alm, G., Baecklund, E., Wahren-Herlenius, M., Omdal, R., Ronnblom, L., Jonsson, R. & Syvanen, A. C. Association of EBF1, FAM167A(C8orf13)-BLK and TNFSF4 gene variants with primary Sjogren's syndrome. *Genes Immun.* **12,** 100–109 (2011).

99. Orozco, G., Alizadeh, B. Z., Delgado-Vega, A. M., Gonzalez-Gay, M. A., Balsa, A., Pascual-Salcedo, D., Fernandez-Gutierrez, B., Gonzalez-Escribano, M. F., Petersson, I. F., van Riel, P. L., Barrera, P., Coenen, M. J., Radstake, T. R., van Leeuwen, M. A., Wijmenga, C., Koeleman, B. P., Alarcon-Riquelme, M. & Martin, J. Association of STAT4 with rheumatoid arthritis: a replication study in three European populations. *Arthritis Rheum.* **58,** 1974–1980 (2008).

100. Palomino-Morales, R. J., Rojas-Villarraga, A., Gonzalez, C. I., Ramirez, G., Anaya, J. M. & Martin, J. STAT4 but not TRAF1/C5 variants influence the risk of developing rheumatoid arthritis and systemic lupus erythematosus in Colombians. *Genes Immun.* **9,** 379–382 (2008).

101. Piotrowski, P., Lianeri, M., Wudarski, M., Olesinska, M. & Jagodzinski, P. P. Contribution of STAT4 gene single-nucleotide polymorphism to systemic lupus erythematosus in the Polish population. *Mol. Biol. Rep.* **39,** 8861–8866 (2012).

102. Prahalad, S., Hansen, S., Whiting, A., Guthery, S. L., Clifford, B., McNally, B., Zeft, A. S., Bohnsack, J. F. & Jorde, L. B. Variants in TNFAIP3, STAT4, and C12orf30 loci associated with multiple autoimmune diseases are also associated with juvenile idiopathic arthritis. *Arthritis Rheum.* **60,** 2124–2130 (2009).

103. Rueda, B., Broen, J., Simeon, C., Hesselstrand, R., Diaz, B., Suarez, H., Ortego-Centeno, N., Riemekasten, G., Fonollosa, V., Vonk, M. C., van den Hoogen, F. H., Sanchez-Roman, J., Aguirre-Zamorano, M. A., Garcia-Portales, R., Pros, A., Camps, M. T., Gonzalez-Gay, M. A., Coenen, M. J., Airo, P., Beretta, L., Scorza, R., van Laar, J., Gonzalez-Escribano, M. F., Nelson, J. L., Radstake, T. R. & Martin, J. The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. *Hum. Mol. Genet.* **18,** 2071–2077 (2009).

104. Sabri, A., Grant, A. V., Cosker, K., El Azbaoui, S., Abid, A., Abderrahmani Rhorfi, I., Souhi, H., Janah, H., Alaoui-Tahiri, K., Gharbaoui, Y., Benkirane, M., Orlova, M., Boland, A., Deswarte, C., Migaud, M., Bustamante, J., Schurr, E., Boisson-Dupuis, S., Casanova, J. L., Abel, L. & El Baghdadi, J. Association study of genes controlling IL-12-dependent IFN-gamma immunity: STAT4 alleles increase risk of pulmonary tuberculosis in Morocco. *J. Infect. Dis.* **210,** 611–618 (2014).

105. Saevarsdottir, S., Stefansdottir, L., Sulem, P., Thorleifsson, G., Ferkingstad, E., Rutsdottir, G., Glintborg, B., Westerlind, H., Grondal, G., Loft, I. C., Sorensen, S. B., Lie, B. A., Brink, M., Arlestig, L., Arnthorsson, A. O., Baecklund, E., Banasik, K., Bank, S., Bjorkman, L. I., Ellingsen, T., Erikstrup, C., Frei, O., Gjertsson, I., Gudbjartsson, D. F., Gudjonsson, S. A., Halldorsson, G. H., Hendricks, O., Hillert, J., Hogdall, E., Jacobsen, S., Jensen, D. V., Jonsson, H., Kastbom, A., Kockum, I., Kristensen, S., Kristjansdottir, H., Larsen, M. H., Linauskas, A., Hauge, E. M., Loft, A. G., Ludviksson, B. R., Lund, S. H., Markusson, T., Masson, G., Melsted, P., Moore, K. H. S., Munk, H., Nielsen, K. R., Norddahl, G. L., Oddsson, A., Olafsdottir, T. A., Olason, P. I., Olsson, T., Ostrowski, S. R., Horslev-Petersen, K., Rognvaldsson, S., Sanner, H., Silberberg, G. N., Stefansson, H., Sorensen, E., Sorensen, I. J., Turesson, C., Bergman, T., Alfredsson, L., Kvien, T. K., Brunak, S., Steinsson, K., Andersen, V., Andreassen, O. A., Rantapaa-Dahlqvist, S., Hetland, M. L., Klareskog, L.,

Askling, J., Padyukov, L., Pedersen, O. B., Thorsteinsdottir, U., Jonsdottir, I., Stefansson, K., Members of the, Dbds Genomic Consortium, Danish, R. A. Genetics Working Group & Swedish Rheumatology Quality Register Biobank Study, Group. Multiomics analysis of rheumatoid arthritis yields sequence variants that have large effects on risk of the seropositive subset. *Ann. Rheum. Dis.* **81,** 1085–1095 (2022).

106. Shi, Z., Zhang, Q., Chen, H., Lian, Z., Liu, J., Feng, H., Miao, X., Du, Q. & Zhou, H. STAT4 Polymorphisms are Associated with Neuromyelitis Optica Spectrum Disorders. *Neuromolecular Med.* **19,** 493–500 (2017).

107. Sigurdsson, S., Nordmark, G., Garnier, S., Grundberg, E., Kwan, T., Nilsson, O., Eloranta, M. L., Gunnarsson, I., Svenungsson, E., Sturfelt, G., Bengtsson, A. A., Jonsen, A., Truedsson, L., Rantapaa-Dahlqvist, S., Eriksson, C., Alm, G., Goring, H. H., Pastinen, T., Syvanen, A. C. & Ronnblom, L. A risk haplotype of STAT4 for systemic lupus erythematosus is over-expressed, correlates with anti-dsDNA and shows additive effects with two risk alleles of IRF5. *Hum. Mol. Genet.* **17,** 2868–2876 (2008).

108. Stock, C. J. W., De Lauretis, A., Visca, D., Daccord, C., Kokosi, M., Kouranos, V., Margaritopoulos, G., George, P. M., Molyneaux, P. L., Nihtyanova, S., Chua, F., Maher, T. M., Ong, V., Abraham, D. J., Denton, C. P., Wells, A. U., Wain, L. V. & Renzoni, E. A. Defining genetic risk factors for scleroderma-associated interstitial lung disease : IRF5 and STAT4 gene variants are associated with scleroderma while STAT4 is protective against scleroderma-associated interstitial lung disease. *Clin. Rheumatol.* **39,** 1173–1179 (2020).

109. Tsuchiya, N., Kawasaki, A., Hasegawa, M., Fujimoto, M., Takehara, K., Kawaguchi, Y., Kawamoto, M., Hara, M. & Sato, S. Association of STAT4 polymorphism with systemic sclerosis in a Japanese population. *Ann. Rheum. Dis.* **68,** 1375–1376 (2009).

110. Xu, L., Dai, W. Q., Wang, F., He, L., Zhou, Y. Q., Lu, J., Xu, X. F. & Guo, C. Y. Association of STAT4 gene rs7574865G > T polymorphism with ulcerative colitis risk: evidence from 1532 cases and 3786 controls. *Arch. Med. Sci.* **10,** 419–424 (2014).

111. Yan, N., Meng, S., Zhou, J., Xu, J., Muhali, F. S., Jiang, W., Shi, L., Shi, X. & Zhang, J. Association between STAT4 gene polymorphisms and autoimmune thyroid diseases in a Chinese population. *Int. J. Mol. Sci.* **15,** 12280–12293 (2014).

112. Yi, J., Fang, X., Wan, Y., Wei, J. & Huang, J. STAT4 polymorphisms and diabetes risk: a meta-analysis with 18931 patients and 23833 controls. *Int. J. Clin. Exp. Med.* **8,** 3566–3572 (2015).

113. Yi, L., Wang, J. C., Guo, X. J., Gu, Y. H., Tu, W. Z., Guo, G., Yang, L., Xiao, R., Yu, L., Mayes, M. D., Assassi, S., Jin, L., Zou, H. J. & Zhou, X. D. STAT4 is a genetic risk factor for systemic sclerosis in a Chinese population. *Int. J. Immunopathol. Pharmacol.* **26,** 473–478 (2013).

114. Yin, H., Borghi, M. O., Delgado-Vega, A. M., Tincani, A., Meroni, P. L. & Alarcon-Riquelme, M. E. Association of STAT4 and BLK, but not BANK1 or IRF5, with primary antiphospholipid syndrome. *Arthritis Rheum.* **60,** 2468–2471 (2009).

115. Zervou, M. I., Goulielmos, G. N., Castro-Giner, F., Tosca, A. D. & Krueger-Krasagakis, S. STAT4 gene polymorphism is associated with psoriasis in the genetically homogeneous population of Crete, Greece. *Hum. Immunol.* **70,** 738–741 (2009).

116. Zervou, M. I., Mamoulakis, D., Panierakis, C., Boumpas, D. T. & Goulielmos, G. N. STAT4: a risk factor for type 1 diabetes? *Hum. Immunol.* **69,** 647–650 (2008).

117. Zhao, X., Jiang, K., Liang, B. & Huang, X. STAT4 gene polymorphism and risk of chronic hepatitis B-induced hepatocellular carcinoma. *Cell Biochem. Biophys.* **71,** 353–357 (2015).

Chapter 3: Context-aware deconvolution of cell-cell communication with Tensor-cell2cell

Cell interactions determine phenotypes, and intercellular communication is shaped by cellular contexts such as disease state, organismal life stage, and tissue microenvironment. Single-cell technologies measure the molecules mediating cell-cell communication, and emerging computational tools can exploit these data to decipher intercellular communication. However, current methods either disregard cellular context or rely on simple pairwise comparisons between samples, thus limiting the ability to decipher complex cell-cell communication across multiple time points, levels of disease severity, or spatial contexts. Here we present Tensor-cell2cell, an unsupervised method using tensor decomposition, which is the first strategy to decipher context-driven intercellular communication by simultaneously accounting for multiple stages, states, or locations of the cells. To do so, Tensor-cell2cell uncovers context-driven patterns of communication associated with different phenotypic states and determined by unique combinations of cell types and ligand-receptor pairs. As such, Tensor-cell2cell robustly improves upon and extends the analytical capabilities of existing tools. We show Tensor-cell2cell can identify multiple modules associated with distinct communication processes (e.g., participating cell-cell and ligand receptor pairs) linked to severities of Coronavirus Disease 2019 and to Autism Spectrum Disorder. Thus, we introduce an effective and easy-to-use strategy for understanding complex communication patterns across diverse conditions.

3.1 Introduction

Organismal phenotypes arise as cells adapt and coordinate their functions through cell-cell interactions within their microenvironments[1]. Variations in these interactions and the resulting phenotypes can occur because of genotypic differences (e.g. different subjects) or the transition from one biological state or condition to another[2] (e.g. from one life stage into another, migration from one location into another, and transition from health to disease states). These interactions are mediated by changes in the production of signals and receptors by the cells, causing changes in cell-cell communication (CCC). Thus, CCC is dependent on temporal, spatial and condition-specific contexts[3], which we refer to here as cellular contexts. "Cellular contexts" refer to variation in genotype, biological state or condition that can shape the microenvironment of a cell and therefore its CCC. Thus, CCC can be seen as a function of a context variable that is not necessarily binary and can encompass multiple levels (e.g. multiple time points, gradient of disease severities, different subjects, distinct tissues, etc.). Consequently, varying contexts trigger distinct strength and/or signaling activity[1,4–6] of communication, leading to complex dynamics (e.g. increasing, decreasing, pulsatile and oscillatory communication activities across contexts). Importantly, unique combinations of cell-cell and ligand-receptor (LR) pairs can follow different context-dependent dynamics, making CCC hard to decipher across multiple contexts.

Single-cell omics assays provide the necessary resolution to measure these cell-cell interactions and the ligand-receptor pairs mediating CCC. While computational methods for inferring CCC have been invaluable for discovering the cellular and molecular interactions underlying many biological processes, including organismal development and disease pathogenesis[5], current approaches cannot account for high variability in contexts (e.g., multiple time points or phenotypic states) simultaneously. Existing methods lose the correlation structure across contexts since they involve repeating analysis for each context separately, disregarding informative variation in CCC across such factors as disease severities, time points, subjects, or cellular locations[7]. Additional analysis steps are required to compare and compile results from pairwise comparisons[8–11], reducing the statistical power and hindering efforts to link phenotypes to CCC. Moreover, this roundabout process is computationally expensive, making analysis of large sample cohorts intractable. Thus, new methods are needed that analyze CCC while accounting for the correlation structure across multiple contexts simultaneously.

Tensor-based approaches such as Tensor Component Analysis[12] (TCA) can deconvolve patterns associated with the biological context of the system of interest. While matrix-based dimensionality reduction methods such as Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF), Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) can extract low-dimensional structures from the data and reflect important molecular signals[13,14], TCA is better suited to analyze multidimensional datasets obtained from multiple biological contexts or conditions[7] (e.g. time points, study subjects and body sites). Indeed, TCA outperforms matrix-based dimensionality reduction methods when recovering ground truth patterns associated with, for example, dynamic changes in microbial composition across multiple patients[15] and neuronal firing dynamics across multiple experimental trials[12]. TCA exhibits superior performance because it does not require the aggregation of datasets across varying contexts into a single matrix. It instead organizes the data as a tensor, the higher order generalization of matrices, which better preserves the underlying context-driven correlation structure by retaining mathematical features that matrices lack[16,17]. Thus, with the correlation structure retained, the use of TCA with expression data across many contexts allows one to gain a detailed understanding of how context shapes communication, as well as the specific molecules and cells mediating these processes.

Here, we introduce Tensor-cell2cell, a TCA-based strategy that deconvolves intercellular communication across multiple contexts and uncovers modules, or latent context-dependent patterns, of CCC. These data-driven patterns reveal underlying communication changes given the simultaneous interaction between contexts, ligand-receptor pairs, and cells. We first show that Tensor-cell2cell successfully extracts temporal patterns from a simulated dataset. We also illustrate that Tensor-cell2cell is broadly applicable, enabling the study of diverse biological questions associated with COVID-19 severity and Autism Spectrum Disorder (ASD). While our approach can simultaneously analyze more than two samples, we show that Tensor-cell2cell is faster, demands less memory and can achieve better accuracy in separating context-specific information than simpler analyses accessible to other tools. We further demonstrate that Tensor-cell2cell can leverage existing CCC tools by using their output communication scores to analyze multiple contexts. Thus, Tensor-cell2cell's easily interpretable output leverages existing tools, and enables quick identification of key mediators of cell-cell communication across contexts, both reproducing known results and identifying novel interactors.

3.2 Results

3.2.1 Deciphering context-driven communication patterns with Tensor-cell2cell

Organizing biological data through a tensor preserves the underlying correlation structure of the biological conditions of interest[12,15,17]. Extending this approach to infer cell-cell communication enables analysis of important ligand-receptor pairs and cell-cell interactions in a context-aware manner. Accordingly, we developed Tensor-cell2cell, a method based on tensor decomposition[17] that extracts context-driven latent patterns of intercellular communication in an unsupervised manner. Briefly, Tensor-cell2cell first generates a 4D-communication tensor that contains non-negative scores to represent cell-cell communication across different conditions (Fig. 3.3.1a-c). Then, a non-negative TCA[18] is applied to deconvolve the latent CCC structure of this tensor into low-dimensional components or factors (Fig. 3.3.1d-e). Thus, each of these factors can be interpreted as a module or pattern of communication whose dynamics across contexts is indicated by the loadings in the context dimension (Fig 3.1e).

To demonstrate how Tensor-cell2cell recovers latent patterns of communication, we simulated a system of 3 cell types interacting through 300 LR pairs across 12 contexts (represented in our simulation as time points) (Fig 3.2a). We built a 4D-communication tensor that incorporates a set of embedded patterns of communication that were assigned to certain LR pairs used by specific pairs of interacting cells, and represented through oscillatory, pulsatile, exponential, and linear changes in communication scores (Fig. 3.3.2a-f; see Appendix for further details of simulating and decomposing this tensor). Using Tensor-cell2cell, we found that four factors led to the decomposition that best minimized error (Fig. 3.6a), consistent with the number of introduced patterns (Fig 3.2f). This was robustly observed in multiple independent simulations (Fig. 3.7a).

Our simulation-based analysis further demonstrates that Tensor-cell2cell accurately detects context-dependent changes of communication, and identifies which LR pairs, sender cells, and receiver cells are important (Fig 3.2g). In particular, the context loadings of the TCA on the simulated tensor accurately recapitulate the introduced patterns (Fig. 3.3.2f-g), while ligand-receptor and cell loadings properly capture the ligand-receptor pairs, sender cells and receiver cells assigned as participants of the cognate pattern (Fig 3.2g). Indeed, we observed a concordance between the “ground truth” LR pairs assigned to a pattern and their respective factor loadings through Jaccard index and Pearson correlation metrics (Supplementary Tables 1-2**)**. Moreover, Tensor-cell2cell robustly recovered communication patterns when we added noise to the simulated tensor (Fig. 3.7 and Appendix).

Figure 3.1: Tensor representation and factorization of cell-cell communication.
In a given context (n-th context among N total contexts), cell-cell communication scores (see available scoring functions in REF[5]) are computed from the expression of the ligand and the receptor in a LR pair (k-th pair among K pairs) for a specific sender-receiver cell pair (i-th and j-th cells among I and J cells, respectively). This results in a communication matrix containing all pairs of sender-receiver cells for that LR pair (**a**). The same process is repeated for every single LR pair in the input list of ligand-receptor interactions, resulting in a set of communication matrices that generate a 3D-communication tensor (**b**). 3D-communication tensors are built for all contexts and are used to generate a 4D-communication tensor wherein each dimension represents the contexts (colored lines), ligand-receptor pairs, sender cells and receiver cells (**c**). A non-negative TCA model approximates this tensor by a lower-rank tensor equivalent to the sum of multiple factors of rank-one (R factors in total) (**d**). Each component or factor (r-th factor) is built by the outer product of interconnected descriptors (vectors) that contain the loadings for describing the relative contribution that contexts, ligand-receptor pairs, sender cells and receiver cells have in the factor (**e**). For interpretability, the behavior that context loadings follow represent a communication pattern across contexts. Hence, the communication captured by a factor is more relevant or more likely to be occurring in contexts with higher loadings. Similarly, ligand-receptor pairs with higher loadings are the main mediators of that communication pattern. By constructing the tensor to account for directional interactions (panels a-b), ligands and receptors in LR pairs with high loadings are mainly produced by sender and receiver cells with high loadings, respectively.

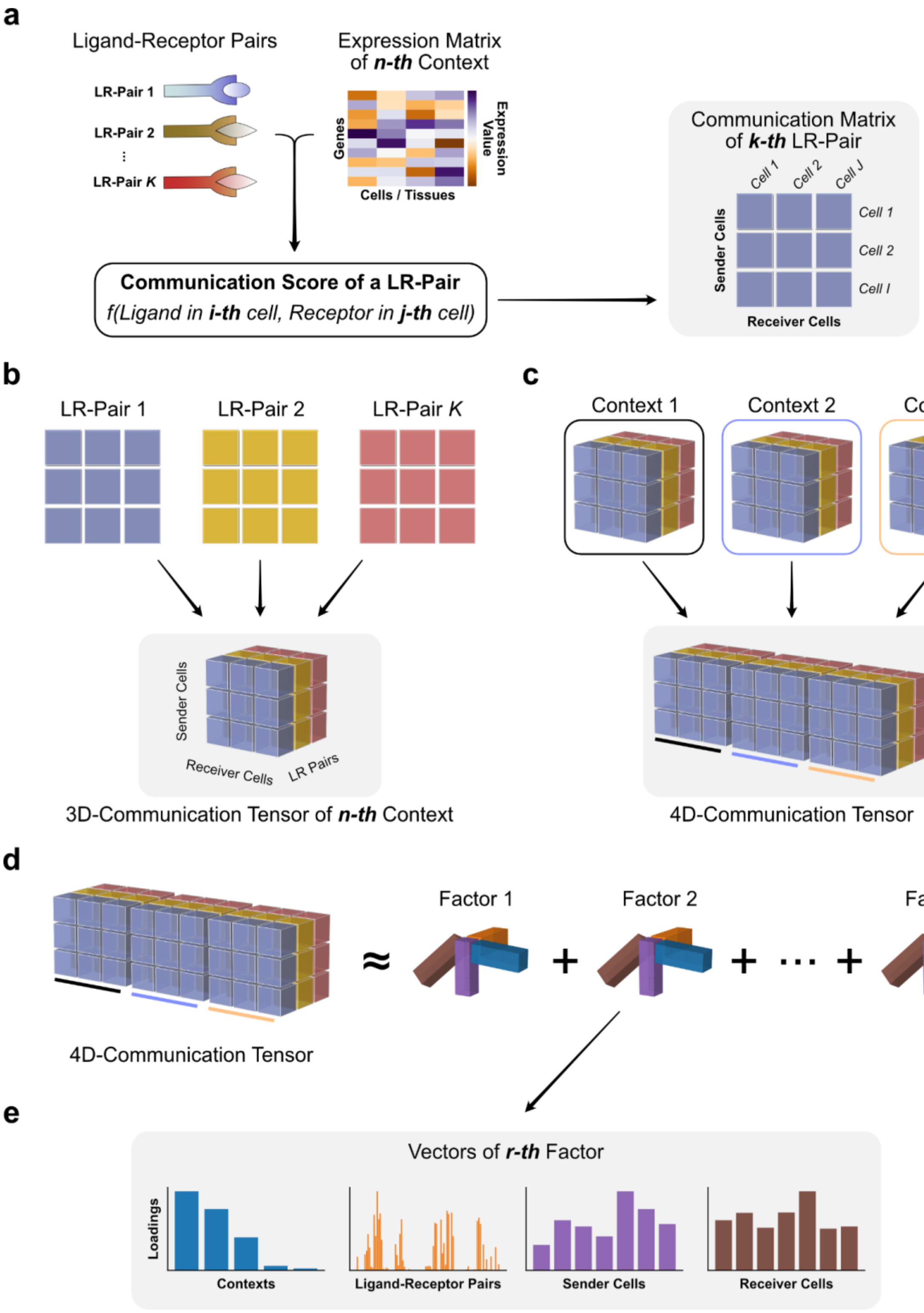
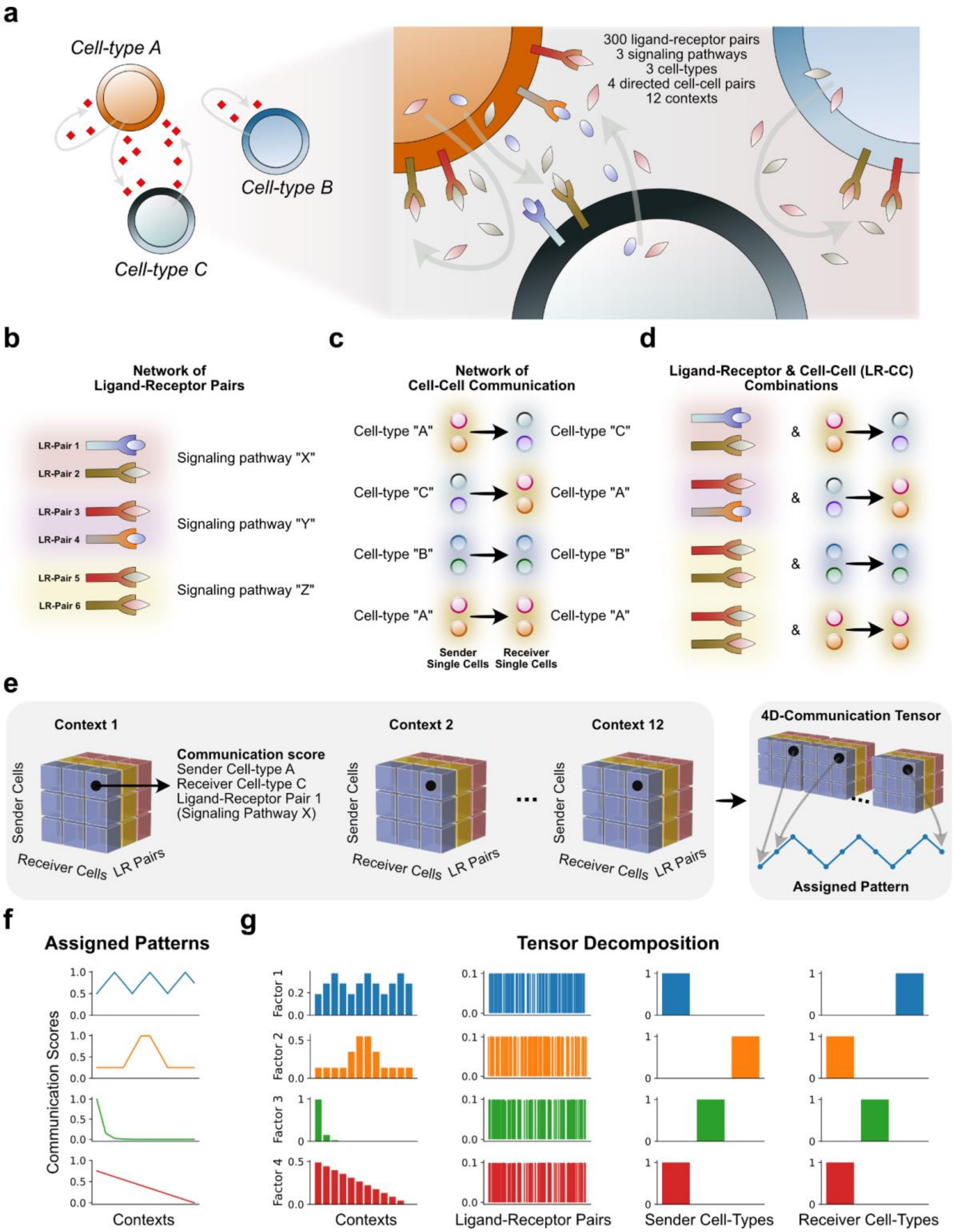

a
Ligand-Receptor Pairs
LR-Pair 1
LR-Pair 2
⋮
LR-Pair K
Expression Matrix
of n-th Context
Genes
Cells / Tissues
Expression
Value
Communication Score of a LR-Pair
f(Ligand in i-th cell, Receptor in j-th cell)
Communication Matrix
of k-th LR-Pair
Cell 1
Cell 2
Cell J
Cell 1
Cell 2
Cell I
Sender Cells
Receiver Cells
b
LR-Pair 1
LR-Pair 2
LR-Pair K
Sender Cells
Receiver Cells
LR Pairs
3D-Communication Tensor of n-th Context
c
Context 1
Context 2
Context N
4D-Communication Tensor
d
4D-Communication Tensor
≈
Factor 1
+
Factor 2
+ ··· +
Factor R
e
Vectors of r-th Factor
Loadings
Contexts
Ligand-Receptor Pairs
Sender Cells
Receiver Cells

Figure 3.2: Tensor-cell2cell recovers simulated communication patterns.
(**a**) Cell-cell communication scenario used for simulating patterns of communication across different contexts (here each a different time point). (**b**) Examples of specific ligand-receptor (LR) and (**c**) cell-cell pairs that participate in the simulated interactions. Individual LR pairs and cell pairs were categorized into groups of signaling pathways and cell types, respectively. In this simulation, signaling pathways did not overlap in their LR pairs, and each pathway was assigned 100 different LR pairs. (**d**) Distinct combinations of signaling pathways with sender-receiver cell type pairs were generated (LR-CC combinations). LR-CC combinations that were assigned the same signaling pathway overlap in the LR pairs but not in the interacting cell types. (**e**) A simulated 4D-communication tensor was built from each time point's 3D-communication tensor. Here, a communication score was assigned to each ligand-receptor and cell-cell member of a LR-CC combination. Each communication score varied across time points according to a specific pattern. (**f**) Four different patterns of communication scores were introduced to the simulated tensor by assigning a unique pattern to a specific LR-CC combination. From top to bottom, these patterns were an oscillation, a pulse, an exponential decay and a linear decrease. The average communication score (y-axis) is shown across time points (x-axis). This average was computed from the scores assigned to every ligand-receptor and cell-cell pair in the same LR-CC combination. (**g**) Results of running Tensor-cell2cell on the simulated tensor. Each row represents a factor, and each column a tensor dimension, wherein each bar represents an element of that dimension (e.g. a time point, a ligand-receptor pair, a sender cell or a receiver cell). Factor loadings (y-axis) are displayed for each element of a given dimension. Here, the factors were visually matched to the corresponding latent pattern in the tensor, and their loadings were normalized to unit Euclidean length. Assigned pattern scores and loading source data are provided in the Source Data file.


a
Cell-type A
Cell-type B
Cell-type C
300 ligand-receptor pairs
3 signaling pathways
3 cell-types
4 directed cell-cell pairs
12 contexts
b
Network of
Ligand-Receptor Pairs
LR-Pair 1
LR-Pair 2
Signaling pathway "X"
LR-Pair 3
LR-Pair 4
Signaling pathway "Y"
LR-Pair 5
LR-Pair 6
Signaling pathway "Z"
c
Network of
Cell-Cell Communication
Cell-type "A"
Cell-type "C"
Cell-type "C"
Cell-type "A"
Cell-type "B"
Cell-type "B"
Cell-type "A"
Cell-type "A"
Sender
Single Cells
Receiver
Single Cells
d
Ligand-Receptor & Cell-Cell (LR-CC)
Combinations
&
&
&
&
e
Context 1
Communication score
Sender Cell-type A
Receiver Cell-type C
Ligand-Receptor Pair 1
(Signaling Pathway X)
Sender Cells
Receiver Cells
LR Pairs
Context 2
Sender Cells
Receiver Cells
LR Pairs
...
Context 12
Sender Cells
Receiver Cells
LR Pairs
4D-Communication Tensor
...
Assigned Pattern
f
Assigned Patterns
Communication Scores
1.0
0.5
0.0
Contexts
g
Tensor Decomposition
Factor 1
Factor 2
Factor 3
Factor 4
0.0
0.2
0.5
1
0.1
0
Contexts
Ligand-Receptor Pairs
Sender Cell-Types
Receiver Cell-Types

3.2.2 Tensor-cell2cell robustly extends cell-cell communication analysis

To demonstrate the power of accounting for multiple contexts simultaneously, we compared the computational efficiency and accuracy of our method with respect to CellChat[10], the only tool that summarizes multiple pairwise comparisons in an automated manner (Table 3.1). Since CellChat cannot extract patterns of CCC across multiple contexts, we instead use the output of its joint manifold learning on pairwise-based changes in signaling pathways as a comparable proxy to the output of Tensor-cell2cell. Despite the use of these proxy comparisons, we emphasize that the conceptual outputs reported by Tensor-cell2cell are unique. Briefly, we found that Tensor-cell2cell is faster, uses less memory, and achieves higher accuracy when analyzing CCC of multiple samples (Fig. 3.8); using a GPU further increases computational speed of Tensor-cell2cell. See more details regarding this comparison in the Methods and *Tensor-cell2cell is fast and accurate* section of the Appendix.

A major advantage of Tensor-cell2cell is that it acts as a robust dimensionality reduction method for any communication scores arranged as a tensor. To illustrate this, we set out to harness the sample-wise communication scoring outputs of other tools. Tensor-cell2cell can restructure these outputs into a 4D-communication tensor (Fig 3.1), extending their capabilities to recover context-dependent patterns of communication. This generalizability enables users to employ any scoring method. Thus, we ran Tensor-cell2cell on communication scores generated by sample-specific analysis with CellPhoneDB[19], CellChat[10], NATMI[9], and SingleCellSignalR[20], as well as the built-in scoring of Tensor-cell2cell. Specifically, we analyzed twelve bronchoalveolar lavage fluid (BALF) samples from patients with different severities of COVID-19 (healthy, moderate and severe) with each method listed above. We assessed the consistency of decomposition between all five scoring methods by using the CorrIndex[21]. The CorrIndex value lies between 0 and 1, with a higher score indicating more dissimilar decomposition outputs; we thus report our similarity results as (1-CorrIndex). Our results indicate that Tensor-cell2cell can consistently identify context-dependent communication patterns independent of the initial communication scoring method (Fig 3.3a, Fig. 3.9), with a mean similarity score of 0.82. Furthermore, differences in decomposition results are driven at the ligand-receptor resolution, yet tend not to propagate to the cell- or context-resolution (Appendix and Fig. 3.10-6). While these results agree with previous reports regarding the inconsistency of scoring methods for ligand-receptor interactions[22], they also show the power of tensor decomposition to resolve these inconsistencies and identify biologically and conceptually consistent communication patterns.

Since Tensor-cell2cell requires the use of multiple conditions or samples, we also assessed biases that may have been introduced by batch effects during gene expression count transformation (e.g., normalization, batch correction, etc). Specifically, we assessed the impact of applying the log(CPM+1) and the fraction of non-zero cells as preprocessing methods[23], and ComBat[24] and Scanorama[25] as batch-effect correction. Here, we also used the BALF COVID-19 samples and built the 4D-tensors using the gene expression values obtained in each case. After running the tensor decomposition, these strategies generated results that seem biologically comparable, as measured with a mean similarity score of 0.86 (Fig 3.3b). As expected, using the raw counts leads to the most biased and different results in comparison to the other preprocessing methods; the mean similarity score between raw counts and all other approaches is 0.77. In contrast, the highest similarity was between the log(CPM+1) and the non-zero fraction of cells. This result is also expected since the non-zero fraction of cells is comparable to the log(CPM+1). However, the non-zero fraction performs better in comparisons of lowly expressed genes[23](e.g. receptors on the cell surface[26]), so we included this fraction as part of the Tensor-cell2cell built-in workflow. Thus, Tensor-cell2cell can detect consistent CCC signatures independent of the method by which gene expression is corrected, with the exception of raw counts, as indicated by the high similarities observed (Fig 3.3b).

Table 3.1: Methodological strategy and context-based analysis in available tools.

Tool	Communication Score	Context Evaluation	Simultaneous Contexts	Multimeric LR pairs	Data Resolution	Platform	Refs.
Tensor-cell2cell	Expression Mean, Expression Product and Geometric Mean	Builds a tensor with all contexts simultaneously and runs a tensor decomposition, accounting for the correlation structure across contexts	Unlimited	Yes	Bulk, Single Cell	Python	This work
CellChat	Mass-action-based probability	Runs separate analyses of each context, does pairwise comparisons and harmonizes them through a joint manifold learning	2	Yes	Single Cell	R	[10]
CellPhoneDB	Expression Mean	None	1	Yes	Single Cell	Python	[19]
CellTalker	Differential Combinations	Differential analysis between two contexts	2	No	Single Cell	R	[8]
Connectome	Modified Expression Product	Differential analysis between two contexts. An overall analysis of cell-type importance can be done for more contexts	2	No	Single Cell	R	[11]
ICELLNET	Expression Product	None	1	Yes	Bulk, Single Cell	R	[27]
iTalk	Differential Combinations	Differential analysis between two contexts	2	No	Single Cell	R	[28]
NATMI	Expression Product and Normalized Expression Product	None	1	No	Bulk, Single Cell	Python	[9]
NicheNet	Personalized-PageRank-based score	None	1	No	Bulk, Single Cell	R	[29]
scAgeCom	Geometric Mean	Differential analysis between two contexts	2	Yes	Single Cell	R	[30]
scTensor	Expression Product	None	1	No	Single Cell	R	[31]

Table 3.1: Methodological strategy and context-based analysis in available tools.

Tool	Communication Score	Context Evaluation	Simultaneous Contexts	Multimeric LR pairs	Data Resolution	Platform	Refs.
SingleCellSignalR	Regularized Expression Product	None	1	No	Single Cell	R	[20]

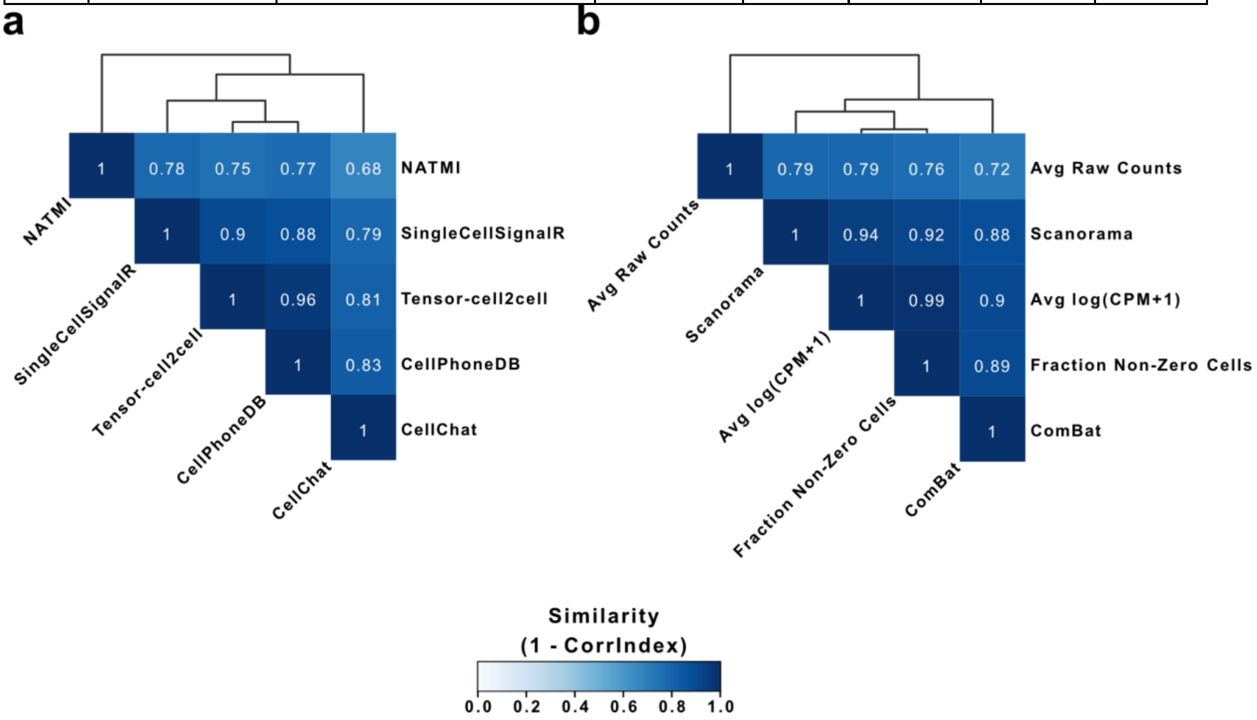


Figure 3.3: Comparison of tensor decompositions resulting from varying input values.
The similarity of tensor decompositions performed on 4D-communication tensors constructed from the single-cell dataset of BALF in patients with varying severities. For a given comparison, constructed tensors have the same elements in each dimension. (**a**) Similarity between tensor decompositions performed on 4D-communication tensors, each corresponding to communication scores computed from different tools for inferring cell-cell communication. The scoring functions correspond to those of CellChat[10], CellPhoneDB[19], NATMI[9], SingleCellSignalR[20] and the built-in methods in Tensor-cell2cell. (**b**) Similarity between tensor decompositions performed on 4D-communication tensors, each modifying the gene expression values by different preprocessing methods (log(CPM+1) and the fraction of non-zero cells[23]) or batch-effect correction methods (Combat[24] and Scanorama[25]), as well as using the raw counts. The communication scores in (**b**) were calculated as the mean expression between the partners in each LR pair, previously aggregating gene expression at the single-cell level into the cell-type level. In (**a**) and (**b**) similarity was measured as (1-CorrIndex), where the CorrIndex[21] is a distance metric for comparing different decompositions on tensors containing the same indices and its values range from 0 to 1 (more similar to more dissimilar). Assessed methods were hierarchically clustered by the similarities of their tensor decompositions. Similarity values are provided in the Source Data file.

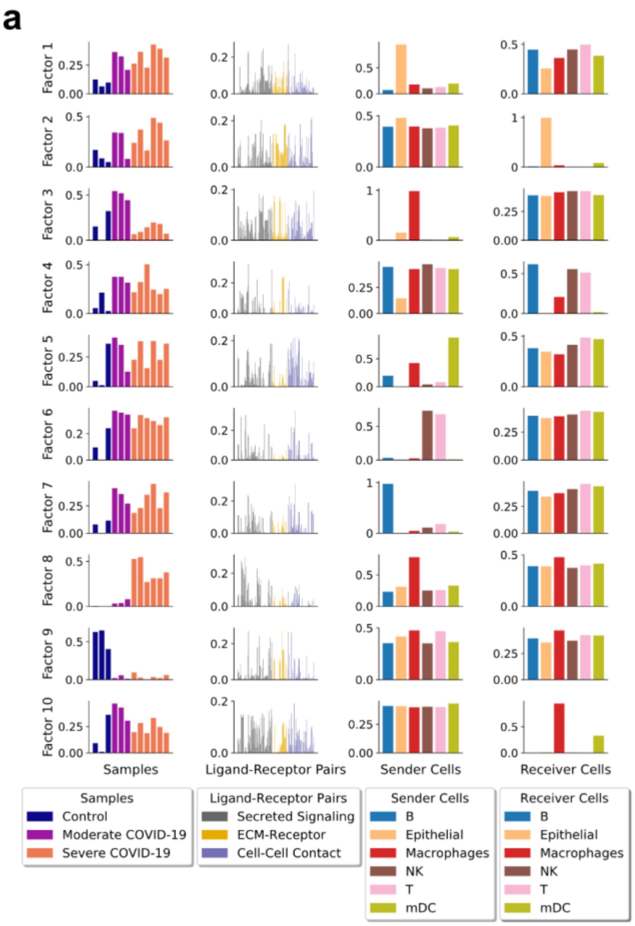
3.2.3 Tensor-cell2cell links intercellular communication with varying severities of COVID-19

Great strides have been made to unravel molecular and cellular mechanisms associated with SARS-CoV-2 infection and COVID-19 pathogenesis. Thus, we tested our method on a single-cell dataset of BALF samples from COVID-19 patients[32] to see how many cell-cell and LR pair relationships in COVID-19 could be revealed by Tensor-cell2cell. By decomposing the tensor associated with this dataset into 10 factors (Fig 3.4a and Fig. 3.6b), Tensor-cell2cell found factors representing communication patterns that are highly correlated with COVID-19 severity (Fig 3.4c) and other factors that distinguish features of the different disease stages (Fig. 3.12), consistent with the high performance that the classifier achieved for this dataset (Fig. 3.8f,h). Furthermore, these factors involve signaling molecules previously linked with severity in separate works (Table 3.4).

The first two factors capture CCC involving autocrine and paracrine interactions of epithelial cells with immune cells in BALF (Fig 3.4a). The sample loadings of these factors reveal a communication pattern wherein the involved LR and cell-cell interactions become stronger as severity increases (Spearman correlation of 0.72 and 0.61, Fig. 3.4c and Fig. 3.12). Although this observation was not reported in the original study, it is consistent with a previous observation of a correlation between COVID-19 severity and the airway epithelium-immune cell interactions[33]. Specifically, epithelial cells are highlighted by Tensor-cell2cell as the main sender cells in factor 1 (Fig 3.4a), and we further provide new details of the molecular mechanisms involving top ranked signals such as APP, MDK, MIF and CD99 (Fig 3.4b). These molecules have been reported to be produced by epithelial cells[34–40] and participate in immune cell recruiting[36–38,41], in response to mechanical stress in lungs[39] and regeneration of the alveolar barrier during viral infection[40]. In addition, epithelial cells act as the main receiver in factor 2 (Fig 3.4a), involving proteins such as PLXNB2, SDC4 and F11R (Fig 3.4b), which were previously determined important for tissue repair and inflammation during lung injury[42–44]. Remarkably, a new technology for experimentally tracing CCC revealed that SEMA4D-PLXNB2 interaction promotes inflammation in a diseased central nervous system[45]; our approach suggests a similar role promoting inflammation in severe COVID-19, specifically mediating the communication between immune and epithelial cells, as reflected in factor 2 (Fig 3.4b).

Figure 3.4: Deconvolution of intercellular communication in patients with varying severity of COVID-19.

(**a**) Factors obtained after decomposing the 4D-communication tensor from a single-cell dataset of BALF in patients with varying severities of COVID-19. 10 factors were selected for the analysis, as indicated in Fig. 3.6b. Here, the context corresponds to samples coming from distinct patients (12 in total, with three healthy controls, three moderate infections, and six severe COVID-19 cases). Each row represents a factor and each column represents the loadings for the given tensor dimension (samples, LR pairs, sender cells and receiver cells), normalized to unit Euclidean length. Bars are colored by categories assigned to each element in each tensor dimension, as indicated in the legend. (**b**) List of the top 5 ligand-receptor pairs ranked by loading for each factor. The corresponding ligands and receptors in these top-ranked pairs are mainly produced by sender and receiver cells with high loadings, respectively. Ligand-receptor pairs with supporting evidence (Table 3.4) for a relevant role in general immune response (black bold) or in COVID-19-associated immune response (red bold) are highlighted. (**c**) Coefficients associated with loadings of each factor: Spearman coefficient quantifying correlation between sample loadings and COVID-19 severity, and Gini coefficient quantifying the dispersion of the edge weights in each factor-specific cell-cell communication network (to measure the imbalance of communication). Important values are highlighted in red (higher absolute Spearman coefficients represent stronger correlations; while smaller Gini coefficients represent distributions with similar edge weights). Loadings and coefficients are provided in the Source Data file.



b

Top-5 Ligand-Receptor Pairs

Factor 1		Factor 2	
APP - CD74	0.268	SEMA4D - PLXNB2	0.212
MDK - NCL	0.246	SEMA4A - PLXNB2	0.203
MIF - CD74 & CD44	0.231	MDK - SDC4	0.202
MDK - ITGA4 & ITGB1	0.227	COL9A2 - SDC4	0.186
CD99 - CD99	0.220	F11R - F11R	0.186

Factor 3		Factor 4	
SIGLEC1 - SPN	0.194	MIF - CD74 & CD44	0.321
RETN - CAP1	0.182	LGALS9 - CD44	0.289
MDK - NCL	0.177	COL9A2 - CD44	0.245
FN1 - ITGA4 & ITGB1	0.176	LAMB3 - CD44	0.242
FN1 - ITGA4 & ITGB7	0.172	LAMB2 - CD44	0.234

Factor 5		Factor 6	
CD99 - CD99	0.213	CD99 - CD99	0.333
ITGB2 - CD226	0.211	CCL5 - CCR5	0.305
CD86 - CTLA4	0.210	CCL5 - CCR1	0.293
ITGB2 - ICAM2	0.201	GZMA - F2R	0.241
ITGB2 - ICAM1	0.192	PTPRC - MRC1	0.222

Factor 7		Factor 8	
CD99 - CD99	0.307	CCL2 - CCR2	0.297
MIF - CD74 & CXCR4	0.241	CCL3 - CCR1	0.275
CD22 - PTPRC	0.231	CCL8 - CCR1	0.261
MIF - CD74 & CD44	0.229	CCL3 - CCR5	0.238
SELL - MADCAM1	0.191	CCL3L1 - CCR1	0.238

Factor 9		Factor 10	
FN1 - CD44	0.274	CD99 - PILRA	0.191
MIF - CD74 & CD44	0.269	LGALS9 - HAVCR2	0.170
APP - CD74	0.266	ANXA1 - FPR1	0.163
PTPRC - MRC1	0.263	MDK - LRP1	0.154
RETN - CAP1	0.243	PTPRC - MRC1	0.153

c

Factor	Spearman Coefficient	Gini Coefficient
Factor 1	0.72	0.50
Factor 2	0.61	0.76
Factor 3	-0.26	0.75
Factor 4	0.39	0.48
Factor 5	0.40	0.59
Factor 6	0.25	0.65
Factor 7	0.51	0.68
Factor 8	0.92	0.24
Factor 9	-0.51	0.09
Factor 10	-0.02	0.74

Our strategy also elucidates communication patterns attributable to specific groups of patients according to disease severity (Fig 3.4a). For example, we found interactions that are characteristic of severe (factor 8) and moderate COVID-19 (factors 3 and 10), and healthy patients (factor 9) (adj. P-value < 0.05, Fig. 3.12). Factor 8 was the most correlated with severity of the disease (Spearman coefficient 0.92, Fig. 3.4c) and highlights macrophages playing a major role as pro-inflammatory sender cells. Their main signals include CCL2, CCL3 and CCL8, which are received by cells expressing the receptors CCR1, CCR2 and CCR5 (Fig 3.4b). Consistent with our result, another study of BALF samples[33] revealed that critical COVID-19 cases involve stronger interactions of cells in the respiratory tract through ligands such as CCL2 and CCL3, expressed by inflammatory macrophages[33]. Moreover, the inhibition of CCR1 and/or CCR5 (receptors of CCL2 and CCL3) has been proposed as a potential therapeutic target for treating COVID-19[33,46]. Tensor-cell2cell also deconvolved patterns attributable to moderate rather than severe COVID-19, also highlighting interactions driven by macrophages (factors 3 and 10; Fig. 4a). However, top-ranked molecules (Fig 3.4b) and gene expression patterns (Fig. 3.13) suggest that the intercellular communication is led by macrophages with an anti-inflammatory M2-like phenotype, in contrast to factor 8 (pro-inflammatory phenotype). Multiple top-ranked signals in factors 3 and 10 have been associated with an M2 macrophage phenotype acting in the immune response to SARS-CoV-2[47–52].

In contrast to severe and moderate COVID-19 patients, communication patterns associated with healthy subjects involve all sender-receiver cell pairs with a similar importance. In particular, factor 9 (Fig 3.4a) demonstrated the smallest Gini coefficient (0.09; Fig. 4c), which measures the extent to which edge weights between sender and receiver cells are evenly distributed in the factor-specific cell-cell communication network. Smaller Gini coefficients show more even distributions, i.e., more equally weighted potential of communication across sender and receiver cell pairs (see Methods). This indicates that the intercellular communication represented by factor 9 is ubiquitous across cell types. Thus, this conservation across cells may be an indicator of communication during homeostasis, since the context loadings for this factor are not associated with disease (Fig. 3.12). Interestingly, a top-ranked LR pair in factor 9 is MIF-CD74/CD44 (Fig 3.4b), which is consistent with ubiquitous expression of MIF across tissues and its protective role in normal conditions[40,53]. Thus, Tensor-cell2cell extracts communication patterns distinguishing one group of patients from another and detects known mechanisms of immune response during disease progression (Appendix), which is important for therapeutic applications.

3.2.4 Tensor-cell2cell elucidates communication mechanisms associated with Autism Spectrum Disorders

Dysregulation of neurodevelopment in Autism Spectrum Disorders (ASD) is associated with perturbed signaling pathways and CCC in complex ways[54]. To understand these cellular and molecular mechanisms, we analyzed single-nucleus RNA-seq (snRNA-seq) data from postmortem prefrontal brain cortex (PFC) from 13 ASD patients and 10 controls[55]. We built a 4D-communication tensor containing 16 cell types present in all samples, including neurons and non-neuronal cells, and 749 LR pairs; then we used Tensor-cell2cell to deconvolve their associated CCC into 6 context-driven patterns (Fig 3.5a and Fig. 3.6c). In these factors, we observe communication between all neurons (factor 1), as well as communication of specific neurons in the cortical layers I-VI (factors 2 and 3), interneurons (factor 4), astrocytes and oligodendrocytes (factor 5), and endothelial cells (factor 6).

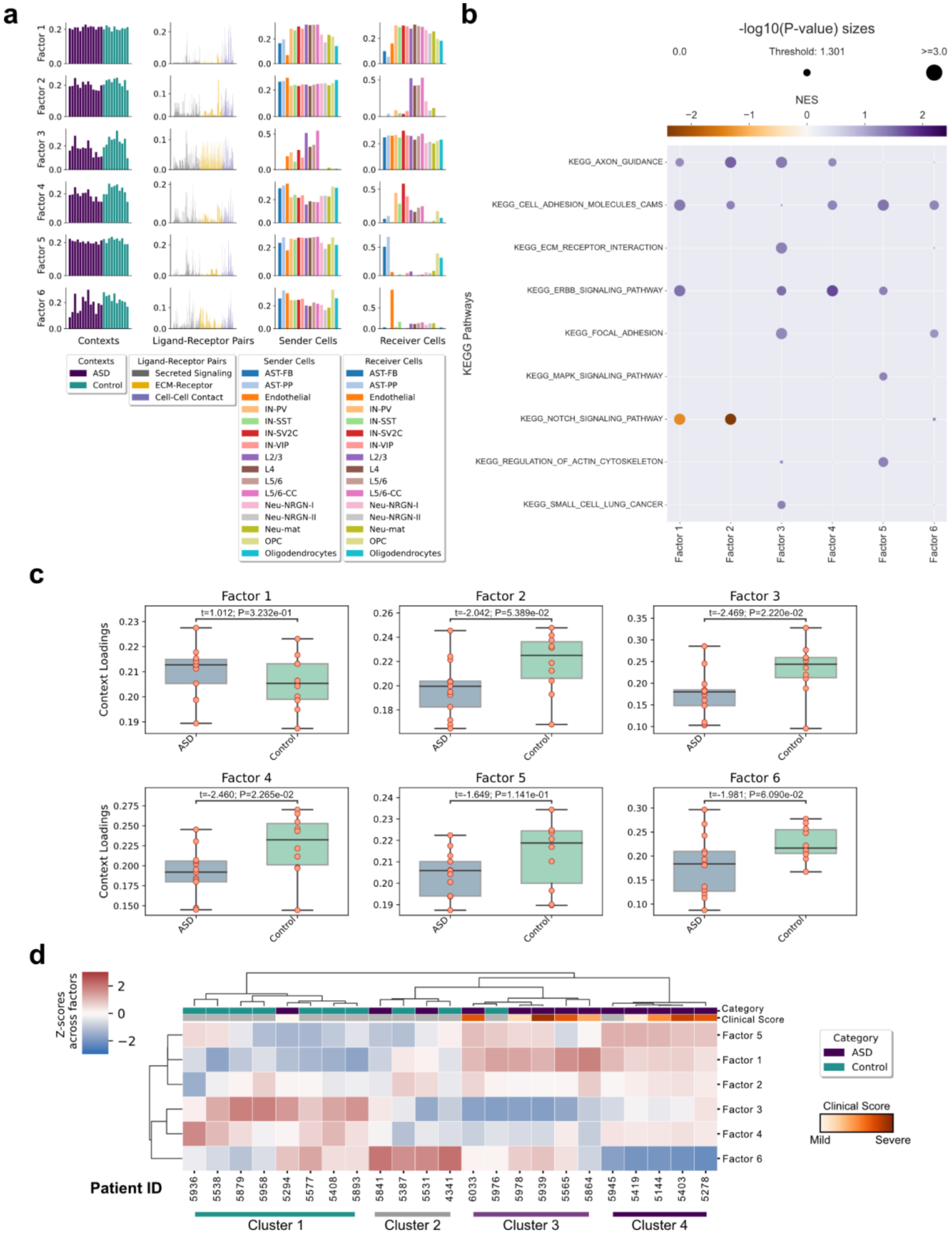
Tensor-cell2cell's outputs can be further dissected using downstream analyses with common approaches. To illustrate this, we ranked the LR pairs by their loadings in a factor-specific fashion, and ran Gene Set Enrichment Analysis[56] (GSEA) using LR pathway sets built from KEGG pathways[57] (see Methods). We observed that each factor was associated with different biological functions including axon guidance, cell adhesion, extracellular-matrix-receptor interaction, ERBB signaling, MAPK signaling, among others (Fig 3.5b). Dysregulation of axon guidance, synaptic processes and MAPK pathway have been previously linked to ASD from differential analysis[55,58], supporting our observations. Moreover, our results extend to other roles associated with extracellular matrix, focal adhesion of cells, regulation of actin cytoskeleton, and signaling through ErbB receptors, which involves Akt, PI3K, and mTOR pathways, as well as regulation of cell proliferation, migration, motility, differentiation, and apoptosis[59]. Thus, Tensor-cell2cell outputs can be used to assign macro-scale biological functions to each of the factors, extending the interpretation of factor-specific CCC.

After identifying main pathways involved in each factor, one can further use sample loadings to identify how these functions are associated with each sample group. By doing so, we found that factors 3 and 4 significantly distinguish ASD from typically-developing controls (Fig 3.5c). Neurons in cortical layers are the main sender cells in factor 3, while interneurons are key receiver cell types in factor 4 (Fig 3.5a and Fig. 3.14), with parvalbumin interneurons (IN-PV), and SV2C-expressing interneurons (IN-SV2C) as the top ranked cells, consistent with the previously reported cell types that are more affected in ASD condition[55] (i.e., with a greater number of dysregulated genes), and that correspond to neurons in the cortical layers I-VI, IN-SV2C and IN-PV. Thus, considering the overall decreased sample loadings in the ASD group, the GSEA results, and the factor-specific CCC networks built from the cell loadings (Fig. 3.14), our analysis suggests that there is a downregulation of axon guidance, cell adhesion, and ERBB signaling involving neurons in cortical layers I-VI and interneurons in ASD patients. See Appendix for further discussion.

Clustering methods can be applied for grouping samples in an unsupervised manner. Thus, we can assess the overall similarity between samples across all factors; considering combinations of factors can offer additional insights to the analysis as compared to considering one factor at a time. We use hierarchical clustering to group samples into four main clusters (Fig 3.5d). Cluster 1 mainly groups controls, cluster 2 is not associated with any category, cluster 3 mostly represents ASD patients, and cluster 4 is completely related to ASD condition. These clusters also reveal that combinations of factors separate samples by ASD and control groups. For example, samples in cluster 1 seem to have smaller loadings in factors 1 and 5, and higher loadings in factors 3 and 4. Interestingly, the only ASD sample present in this cluster had the smallest ASD clinical score, suggesting that CCC mechanisms are more similar to controls when the phenotype is mild. In contrast, cluster 3 shows an opposite CCC behavior to cluster 1. Cluster 4 also reveals that the combination of factor 6 with low sample loadings and factors 1 and 5 with high values is a strong marker of several ASD patients, even though factors 1, 5, and 6 did not show significant differences between sample groups (Fig 3.5c). Based on this, patients in cluster 4 had increased CCC through the NRXNs-NRLGs, CTNs-NRCAMs, and NCAMs-NCAMs interactions (synapse and cell adhesion) in neurons as senders and receivers, and astrocytes and oligodendrocytes as receivers, as well as a decreased CCC through VEGFs-FLT1, PTPRM-PTPRM, and PTN-NCL interactions (angiogenesis, neural migration and neuroprotection) related to endothelial cells as the main receivers (Table 3.5). Finally, both ASD-clusters seem to be slightly distinct in terms of phenotype, considering their mean clinical scores of 25.0 and 22.8, respectively for clusters 3 and 4, but without significant differences. Thus, downstream analyses reveal that multiple dysregulations of CCC patterns captured by Tensor-cell2cell occur simultaneously in ASD condition (Fig 3.5d), even though these patterns could not explain phenotypic differences when considered in isolation (Fig. 3.5c).

Figure 3.5: Application of Tensor-cell2cell to study mechanisms underlying intercellular communication in patients with ASD.

(**a**) Factors obtained after decomposing the 4D-communication tensor from a single-nucleus dataset of prefrontal brain cortex samples from patients with or without ASD. Six factors were selected for the analysis, as indicated in Fig. 3.6c. Here, the context corresponds to samples coming from distinct patients (23 in total, with thirteen ASD patients and ten controls). Each row represents a factor and each column represents the loadings for the given tensor dimension (samples, LR pairs, sender cells and receiver cells), normalized to unit Euclidean length. Bars are colored by categories assigned to each element in each tensor dimension, as indicated in the legend. Cell-type annotations are those used in REF[55]. (**b**) GSEA performed on the pre-ranked LR pairs by their respective loadings in each factor, and using KEGG pathways. Dot sizes are proportional to the negative logarithmic of the P-values, as indicated at the top of the panel. The threshold value indicates the size of a P-value=0.05. The dot colors represent the normalized enrichment score (NES) after the permutations performed by the GSEA, as indicated by the colorbar. P-values were obtained from the permutation step performed by GSEA, and adjusted with a Benjamini-Hochberg correction across all factors. (**c**) Boxplot representation for ASD (n=13) and control (n=10) groups of patients. Each panel represents the sample loadings, grouped by condition category, in each of the factors. Boxes represent the quartiles and whiskers show the rest of each distribution. Groups were compared by a two-sided independent t-test, followed by a Bonferroni correction. For each pairwise comparison, the exact values of the test statistics (t) and the adjusted P-values (P) are shown. (**d**) Heatmap of the standardized sample loadings across factors (z-scores) for each of the samples. Samples and factors were grouped by hierarchical clustering. Major clusters of the samples are indicated at the bottom. The category of each sample is colored on the top, according to the legend. A clinical score of each patient is also shown, according to the colorbar. Controls, and ASD samples without an assigned score, were colored gray. This clinical score summarizes the social interactions, communication, repetitive behaviors, and abnormal development of the patients, as indicated in REF[55]. Loadings, enrichment scores, and clinical scores are provided in the Source Data file.


a
b
c
d
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Contexts
Ligand-Receptor Pairs
Sender Cells
Receiver Cells
Contexts
ASD
Control
Ligand-Receptor Pairs
Secreted Signaling
ECM-Receptor
Cell-Cell Contact
Sender Cells
Receiver Cells
AST-FB
AST-PP
Endothelial
IN-PV
IN-SST
IN-SV2C
IN-VIP
L2/3
L4
L5/6
L5/6-CC
Neu-NRGN-I
Neu-NRGN-II
Neu-mat
OPC
Oligodendrocytes
-log10(P-value) sizes
0.0
Threshold: 1.301
>=3.0
NES
−2
−1
0
1
2
KEGG Pathways
KEGG_AXON_GUIDANCE
KEGG_CELL_ADHESION_MOLECULES_CAMS
KEGG_ECM_RECEPTOR_INTERACTION
KEGG_ERBB_SIGNALING_PATHWAY
KEGG_FOCAL_ADHESION
KEGG_MAPK_SIGNALING_PATHWAY
KEGG_NOTCH_SIGNALING_PATHWAY
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON
KEGG_SMALL_CELL_LUNG_CANCER
Context Loadings
t=1.012; P=3.232e-01
t=-2.042; P=5.389e-02
t=-2.469; P=2.220e-02
t=-2.460; P=2.265e-02
t=-1.649; P=1.141e-01
t=-1.981; P=6.090e-02
Z-scores across factors
Category
Clinical Score
Mild
Severe
Patient ID
5936
5538
5879
5958
5294
5577
5408
5893
5841
5387
5531
4341
6033
5976
5978
5939
5565
5864
5945
5419
5144
5403
5278
Cluster 1
Cluster 2
Cluster 3
Cluster 4

3.3 Discussion

Here we present Tensor-cell2cell, a computational approach that identifies modules of cell-cell communication and their changes across contexts (e.g., across subjects with different disease severity, multiple time points, different tissues, etc.). Our approach can rank LR pairs based on their contribution to each communication module and connect these signals to specific cell types and phenotypes. Tensor-cell2cell's ability to consider multiple contexts simultaneously to identify context-dependent communication patterns goes beyond state-of-the-art tools, which are either unaware of the context driving CCC[5,19,29,60] or require analysis of each context separately to perform pairwise comparisons in posterior steps[10,11]. Tensor-cell2cell is therefore a flexible method that can integrate multiple datasets and readily identify patterns of intercellular communication in a context-aware manner, reporting them through interconnected and easily interpretable scores.

Tensor-cell2cell robustly detects communication patterns using many other scoring methods[13]. Thus, our method is not only an improvement over other tools, but also greatly extends these tools, enabling new analyses with existing methods. One can choose any tool of interest, run it on each context separately, and use the resulting communication scores to build and deconvolve a 4D-communication tensor. Other tools, such as CellChat, allow the generation of scores at the signaling pathway level instead of LR pairs. This, combined with Tensor-cell2cell, could provide additional information about changes in signaling pathways. Thus, Tensor-cell2cell can also be used for analyzing any other score linking gene expression from cell pairs, beyond just scores based on protein-protein interactions. In this regard, our tool outputs consistent results regardless of the preprocessing and batch correction method we evaluated (Fig 3.3b). Nevertheless, it is best practice to employ integration/batch-correction methods to correct gene expression and annotate cell types before running Tensor-cell2cell to ensure this source of variation is controlled[61].

Tensor-cell2cell is faster for analyzing multiple samples than pairwise comparisons, providing a considerable improvement in running time and reduced memory requirements (Appendix). Tensor-cell2cell's runtime can be further accelerated when a GPU is available (Fig. 3.8a). It is also more accurate, resulting in 10-20% higher classification accuracy of subjects with COVID-19 when compared to CellChat (Fig. 3.8e-h). However, we note that benchmarking CCC prediction tools is challenging due to the lack of a ground truth[5], and it is hard to compare and evaluate tools because of the qualitative differences in their outputs[22] (Appendix). While pairwise comparisons can be informative about differential cellular and molecular mediators of communication, the results are less interpretable (Fig. 3.15), do not provide the multi-scale resolution available in Tensor-cell2cell (Fig. 3.4a and 5a), and do not identify context-dependent patterns.

Meaningful biology can be easily identified from Tensor-cell2cell. For example, a manual interpretation of the BALF COVID-19 decomposition found communication results not previously observed in the original study[32] and recapitulated findings spanning tens of peer-reviewed articles (Table 3.4). This included a correlation between the lung epithelium-immune cell interactions and COVID-19 severity[33] and molecular mediators that distinguished moderate and severe COVID-19 (see "Tensor-cell2cell elucidates molecular mechanisms distinguishing moderate from severe COVID-19" in the Appendix). Additionally, Tensor-cell2cell results can be coupled with downstream analysis methods to facilitate interpretation and provide further insights of underlying mechanisms. In our ASD case-study (Fig 3.5), such analyses included GSEA, clustering, visualization and statistical comparison of factors, and factor-specific analysis of sender-receiver communication networks (Fig. 3.14). In the ASD case-study, we found dysregulated CCC directly distinguished ASD patients from controls and was linked with a downregulation of axon guidance, cell adhesion, synaptic processes, and ERBB signaling in cortical neurons and interneurons (Fig. 3.5a,b), consistent with previous evidence[55,58,62,63]. Moveover, accounting for the combinatorial relationship of samples across factors demonstrated additional complex relationships of CCC dysregulation (Fig 3.5d).

A limitation to consider is the potential of missing communication scores in the tensor (e.g., when a rare cell type appears in only one sample). Although Tensor-cell2cell can handle cell types that are missing in some conditions, the implemented tensor decomposition algorithm can be further optimized for missing values. Since the implemented algorithm is not optimized for this purpose, we built a 4D-communication tensor that contains only the cell types that are shared across all samples in our COVID-19 and ASD study cases. Thus, further developments will facilitate analyses with missing values to include all possible members of communication (i.e., LR pairs and cell types that may be missing in certain contexts).

In addition to single cell data analyzed here, Tensor-cell2cell also accepts bulk transcriptomics data (an example of a time series bulk dataset of *C. elegans* is included in a Code Ocean capsule, see Methods), and it could further be used to analyze proteomic data. We demonstrated the application of Tensor-cell2cell in cases where samples correspond to distinct patients, but it can be applied to many other contexts. For instance, our strategy can be readily applied to time series data by considering time points as the contexts, and to spatial transcriptomic datasets, by previously defining cellular niches or neighborhoods as the contexts, given their spatial signatures[64]. We have included Tensor-cell2cell as a part of our previously developed tool cell2cell[65], enabling previous functionalities such as employing any list of LR pairs (including protein complexes), multiple visualization options, and personalizing the communication scores to account for other signaling effects such as the (in)activation of downstream genes in a signaling pathway[29,66,67]. Thus, these attributes make Tensor-cell2cell valuable for identifying key cell-cell and LR pairs mediating complex patterns of cellular communication within a single analysis for a wide range of studies.

3.4 Methods

3.4.1 RNA-seq data processing

RNA-seq datasets were obtained from publicly available resources. Datasets correspond to a large-scale single-cell atlas of COVID-19 in humans[68], a COVID-19 dataset of single-cell transcriptomes for BALF samples[32]. COVID-19 datasets were collected as raw count matrices from the NCBI's Gene Expression Omnibus[69] (GEO accession numbers GSE158055 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158055] and GSE145926 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145926], respectively), while the ASD dataset is available in the NCBI's BioProject under accession code PRJNA434002 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434002/], but we obtained the log2-transformed UMI counts from https://cells.ucsc.edu/autism/downloads.html. In total, the first dataset contains 1,462,702 single cells, the second 65,813 and the last one 104,559 single nuclei. The first dataset contains samples of patients with varying severities of COVID-19 (control, mild/moderate and severe/critical) and we selected just 60 PBMC samples among all different sample sources (20 per severity type). In the second dataset, we considered the 12 BALF samples of patients with varying severities of COVID-19 (3 control, 3 moderate and 6 severe) and preprocessed them by removing genes expressed in fewer than 3 cells, which left a total of 11,688 genes in common across samples. In the ASD dataset, PFC samples from 23 patients with and without ASD condition (13 ASD patients and 10 controls) were considered, and preprocessed similarly to the BALF dataset, resulting in a total of 24,298 genes in common across samples. In all datasets, we used the cell type labels included in their respective metadata. We aggregated the gene expression from single cells/nuclei into cell types by calculating the fraction of cells in the respective label with non-zero counts, as previously recommended for properly representing genes with low expression levels[23], as usually happens with genes encoding surface proteins[26].

3.4.2 Ligand-receptor pairs

A human list of 2,005 ligand-receptor pairs, 48% of which include heteromeric-protein complexes, was obtained from CellChat[10]. We filtered this list by considering the genes expressed in the PBMC and BALF expression datasets and that match the IDs in the list of LR pairs, resulting in a final list of 1639 and 189 LR pairs, respectively. While in the ASD dataset, 749 LR pairs that matched the gene IDs were considered.

3.4.3 Building the context-aware communication tensor

For building a context-aware communication tensor, three main steps are followed: 1) A communication matrix is built for each ligand-receptor pair contained in the interaction list from the gene expression matrix of a given sample. To build this communication matrix, a communication score[5] is assigned to a given LR pair for each pair of sender-receiver cells. The communication score is based on the expression of the ligand and the receptor in the respective sender and receiver cells (Fig 3.1a). 2) After computing the communication matrices for all LR pairs, they are joined into a 3D-communication tensor for the given sample (Fig 3.1b). Steps 1 and 2 are repeated for all the samples (or contexts) in the dataset. 3) Finally, the 3D-communication tensors for each sample are combined, each of them representing a coordinate in the 4th-dimension of the 4D-communication tensor (or context-aware communication tensor; Fig. 1c).

To build the tensor for all datasets, we computed the communication scores as the mean expression between the ligand in a sender cell type and cognate receptor in a receiver cell type, as previously described[19]. For the LR pairs wherein either the ligand or the receptor is a multimeric protein, we used the minimum value of expression among all subunits of the respective protein to compute the communication score. In all cases we further considered cell types that were present across all samples. Thus, the 4D-communication tensor for the PBMC, BALF and ASD datasets resulted in a size of 60 x 1639 x 6 x 6; 12 x 189 x 6 x 6, and 23 x 749 x 16 x 16 respectively (that is, samples x ligand-receptor pairs x sender cell types x receiver cell types).

3.4.4 Non-negative tensor component analysis

Briefly, non-negative TCA is a generalization of NMF to higher-order tensors (matrices are tensors of order two). To detail this approach, let χ represent a *C* x *P* x *S* x *T* tensor, where *C*, *P*, *S* and *T* correspond to the number of contexts/samples, ligand-receptor pairs, sender cells and receiver cells contained in the tensor, respectively. Similarly, let χ_{ijkl} denote the representative interactions of context *i*, using the LR pair *j,* between the sender cell *k* and receiver cell *l*. Thus, the TCA method underlying Tensor-cell2cell corresponds to CANDECOMP/PARAFAC[70,71], which yields the decomposition, factorization or approximation of χ through a sum of *R* tensors of rank-1 (Fig 3.1d):

$$\chi \approx \sum_{r=1}^{R} \quad c^r \otimes p^r \otimes s^r \otimes t^r \tag{1}$$

Where the notation ⊗ represents the outer product and $c^r, p^r, s^r and\ t^r$ are vectors of the factor *r* that contain the loadings of the respective elements in each dimension of the tensor (Fig 3.1e). These vectors have values greater than or equal to zero. Similar to NMF, the factors are permutable and the elements with greater loadings represent an important component of a biological pattern captured by the corresponding factor. Values of individual elements in this approximation are represented by:

$$\chi_{ijkl} \approx \sum_{r=1}^{R} \quad c_i^r \otimes p_j^r \otimes s_k^r \otimes t_l^r \tag{2}$$

The tensor factorization is performed by iterating the following objective function until convergence through an alternating least squares minimization[17,72]:

$$min_{\{c,p,s,t\}} \left\| \chi - \sum_{r=1}^{R} c^r \otimes p^r \otimes s^r \otimes t^r \right\|_F^2 \tag{3}$$

Where $\left\| . \right\|_F^2$represent the squared Frobenius norm of a tensor, calculated as the sum of element-wise squares in the tensor:

$$\left\| \chi \right\|_F^2 = \sum_{i=1}^{C} \sum_{j=1}^{P} \sum_{k=1}^{S} \sum_{l=1}^{T} \chi_{ijkl}{}^2 \tag{4}$$

All the described calculations were implemented in Tensor-cell2cell through functions available in Tensorly[73], a Python library for tensors.

3.4.5 Measuring the error of the tensor decomposition

Depending on the number of factors used for approximating the 4D-communication tensor, the reconstruction error calculated in the objective function can vary. To quantify the error with an interpretable value, we used a normalized reconstruction error as previously described[12]. This normalized error is on a scale of zero to one and is analogous to the fraction of unexplained variance used in PCA:

$$\frac{\left\| \chi - \sum_{r=1}^{R} c^r \otimes p^r \otimes s^r \otimes t^r \right\|_F^2}{\left\| \chi \right\|_F^2} \tag{5}$$

3.4.6 Running Tensor-cell2cell with communication scores from external tools

We assessed the similarity of tensor decomposition on the BALF dataset using different communication scoring methods (CellChat[10], CellPhoneDB[19], NATMI[9], SingleCellSignalR[20], and

Tensor-cell2cell's built-in scoring). To enable consistency between methods, we used the same ligand-receptor PPI database (CellChat – see "Ligand-receptor pairs") and ran each method via LIANA[22]. LIANA provides a number of advantages over running each tool separately, including consistent thresholding and parameters, interoperability between methods and LR databases, and modifications to allow methods that could not originally account for protein complexes to do so. We adjusted parameters to match those of Tensor-cell2cell's built-in scoring by not filtering for minimal proportions of expression by cell type or thresholding for differentially expressed genes.

As input to LIANA, we constructed a Seurat object with log(CPM+1) normalized counts for each sample. For each tool and sample, LIANA outputs an edge-list of communication scores for a given combination of sender and receiver cells, as well as ligand-receptor pairs. We extended Tensor-cell2cell's functionalities to restructure a set of these edge-lists, each associated with a sample, into a 4D-communication tensor (Fig 3.1). This functionality enables users to either provide input expression matrices and use Tensor-cell2cell's built-in scoring, or to run their communication scoring method of choice on each sample and provide the resultant edge-lists as input. To further ensure consistency, we subsetted each resultant tensor to the intersection of ligand-receptor pairs scored across all 5 methods. For each method, this resulted in a tensor consisting of 12 samples, 172 ligand-receptor pairs, and 6 sender and receiver cells.

3.4.7 Evaluating the effect of gene expression preprocessing and batch-effect correction on Tensor-cell2cell

To evaluate how gene expression preprocessing and batch-effect correction impact the results of Tensor-cell2cell, we assessed the similarity of tensor decomposition on the BALF dataset. To compute the communication scores for building the tensors (Fig 3.1a), we used different gene expression values, including the raw UMI counts, the preprocessed values with log(CPM+1) and the fraction of non-zero cells[23], and the batch-corrected values with ComBat[24] and Scanorama[25]. Except by the fraction of non-zero cells, which already aggregated single-cells into cell-types, other values were aggregated into the cell-type level by computing their average value for each gene across single cells with the same cell-type label. As the communication score, we used the expression mean of the interacting partners in each LR pair. Thus, we built 4D-communication tensors as mentioned for the BALF data in the Methods subsection "Building the context-aware communication tensor". The tensor decomposition resulting with the fraction of non-zero cells in this case corresponds to the same in Fig. 4.

3.4.8 Measuring the similarity between distinct tensor decomposition runs

To assess decomposition consistency between different scoring methods or preprocessing pipelines, we employed the CorrIndex[21]. The CorrIndex is a permutation- and scaling-invariant distance metric that enables consistent comparison of decompositions between tensors containing the same elements, without need to align the factors obtained in each case (separate tensor decompositions can output similar factors but in different order). The CorrIndex value lies between 0 and 1, with a higher score indicating more dissimilar decomposition outputs. To score tensor decompositions, the output factor matrices must first be vertically stacked. We implemented a modification that instead assesses each tensor dimension separately (see Supplementary Note for more details). While taking the minimal score between all dimensions tends to be more stringent, it disregards the combinatorial effects of all dimensions together. These combinatorial effects are important because they better reflect the goal of tensor decomposition and because similarity in those dimensions that are not the minimal one may be artificially inflated. To facilitate the use of the CorrIndex and its modified version, we wrote a Python implementation that is available on the Tensorly package[73].

3.4.9 Downstream analyses using the loadings from the tensor decomposition

We incorporate several downstream analyses of Tensor-cell2cell's decomposition outputs to further elucidate the underlying cell- and molecular- mediators of cell-cell communication. Each of these analyses are associated with a specific tensor dimension, and thus, a specific biological resolution. This includes 1) statistical, correlative, and clustering analyses to understand context associations for each factor, 2) gene set enrichment analysis of ligand-receptor loadings to identify granular signaling pathways associated with factors, 3) the generation of factor-specific cell-cell communication networks to represent the overall communication state of cells in that factor.

We can understand the context associations for a factor by comparing the loadings of samples associated with distinct contexts. For statistical significance, we conduct an independent t-test pairwise between each context group associated with the samples and use Bonferonni's correction to account for multiple comparisons. We use this for both the COVID-19 BALF dataset (Fig. 3.12 and 8) and the ASD dataset (Fig 3.5c). We also conduct correlative analyses – assuming ordinal contexts (i.e., healthy control < moderate COVID-19 < severe COVID-19), we take the Spearman correlation between the sample loadings and sample severity (Fig 3.4c). Finally, we also hierarchically cluster the samples using their loadings across all factors (Fig 3.5d). For this purpose, we use the normalized loadings resulting from the tensor decomposition, and standardize them across all factors. Then, we apply an agglomerative hierarchical clustering by using Ward's method and the Euclidean distance as a metric. Note that this type of clustering analysis can be applied to the other tensor dimensions.

We can use the LR-pair loadings of a factor to identify the signaling pathways associated with it, by using the Gene Set Enrichment Analysis[56] (GSEA). Before running the analysis, pathways of interest have to be assigned to a list of associated LR pairs. We do that by considering the KEGG gene sets available at http://www.gsea-msigdb.org/. We annotate a LR pair available in CellChat with the gene sets that contain all genes participating in that LR interaction. Then, by filtering LR pathway sets to those containing at least 15 LR pairs, we end up with 22 LR pathway sets. To run GSEA, we rank the LR pairs in each factor by their loadings, and use the PreRanked GSEA function in the package gseapy, by including the 22 LR pathway sets as input. As parameters of the "gseapy.prerank" function, we consider 999 permutations, gene sets (LR pathway sets here) with at least 15 elements, and a score weight of 1 for computing the enrichment scores[56].

Finally, we generate factor-specific cell-cell communication networks. To do so, for a factor *r*, we take the outer product between the sender-cell loadings vector, $\boldsymbol{s}^{\boldsymbol{r}}$, and the receiver-cell loadings vector, $\boldsymbol{t}^{\boldsymbol{r}}$. Conceptually, this outer product represents an adjacency matrix of a factor-specific cell-cell communication network, where each value is an edge weight representing the overall communication between a pair of sender-receiver cells (Fig. 3.14). We can further use this network to understand the communication distribution inequality between sender and receiver cells. We compute a Gini coefficient[74] ranging between 0 and 1 on the distribution of edge weights in the adjacency matrix (Fig 3.4c). A value of 1 represents maximal inequality of overall communication between cell pairs (i.e. one cell pair has a high overall communication value while the others have a value of 0) and 0 indicates minimal inequality (i.e. all cell pairs have the same overall communication values). More generally, the outer product between any two tensor dimension loadings for a given factor conceptually represents the joint distribution of the elements in those two dimensions and can be informative of how the specific elements are related.

3.4.10 Benchmarking of computational efficiency of tools

We measured the running time and memory demanded by Tensor-cell2cell and CellChat to analyze the COVID-19 dataset containing PBMC samples. Each tool was evaluated in two scenarios: either using each sample individually, or by first combining samples by severity (control, mild/moderate, and severe/critical) by aggregating the expression matrices. The latter was intended to favor CellChat by diminishing the number of pairwise comparisons to always be between three contexts; thus, increases in running time or memory demand in this case are not due to an exponentiation of comparisons (*n* samples choose 2). CellChat was run by following the procedures outlined in the "Comparison_analysis_of_multiple_datasets" vignette (https://github.com/sqjin/CellChat/tree/master/tutorial). Briefly, signaling pathway communication probabilities were first individually calculated for each sample or context. Next, pairwise comparisons between each sample or context were obtained by computing either a "functional" or a "structural" similarity. The functional approach computes a Jaccard index to compare the signaling pathways that are active in two cellular communication networks, while the structural approach computes a network dissimilarity[75] to compare the topology of two signaling networks (see REF[10] for further details). Finally, CellChat performs a manifold learning approach on sample similarities and returns UMAP embeddings for each signaling pathway in each different context (e.g. if CellChat evaluates 10 signaling pathways in 3 different contexts, it will return embeddings for 30 points) which can be used to rank the similarity of shared signaling pathways between contexts in a pairwise manner.

The analyses of computational efficiency were run on a compute cluster of 2.8GHz x2 Intel(R) Xeon(R) Gold 6242 CPUs with 1.5 TB of RAM (Micron 72ASS8G72LZ-2G6D2) across 32 cores. Each timing task was limited to 128 GB of RAM on one isolated core and one thread independently where no other processes were being performed. To limit channel delay, data was stored on the node where the job was performed, where the within socket latency and bandwidth are 78.9 ns and 46,102 MB/s respectively. For all timing jobs, the same ligand-receptor pairs and cell types were used. Furthermore, to make the timing comparable, all samples in the dataset were subsampled to have 2,000 single cells. In the case of Tensor-cell2cell, the analysis was also repeated by using a GPU, which corresponded to a Nvidia Tesla V100.

3.4.11 Training and evaluation of a classification model

A Random Forest[76] (RF) model was trained to predict disease status based on both COVID-19 status (healthy-control vs. patient with COVID-19) and severity (healthy-control, moderate symptoms, and severe symptoms). The RF model was trained using a Stratified K-Folds cross-validation (CV) with 3-Fold CV splits. On each CV split a RF model with 500 estimators was trained and RF probability-predictions were compared to the test set using the Receiver Operating Characteristic (ROC). The mean and standard deviation from the mean were calculated for the area under the Area Under the Curve (AUC) across the CV splits. This classification was performed on the context loadings of Tensor-cell2cell, and the two UMAP dimensions of the structural and functional joint manifold learning of CellChat, for both the BALF and PBMC COVID-19 datasets. All classification was performed through Scikit-learn (v. 0.23.2)[77].

3.4.12 Statistics and Reproducibility

No sample-size calculation was performed. Instead, we used the number of samples included in each of the previously published datasets that we used. The only data exclusion performed was for the PBMC COVID-19 datasets, which originally includes 284 samples. For running our benchmarking, we subsetted the dataset to only include 60 samples. These samples were randomly selected for each COVID-19 severity, with 20 corresponding to control patients, 20 to mild/moderate COVID-19 patients, and 20 to severe/critical COVID-19 patients. For reproducibility, we deposited all our analyses including data and exact versions of code and software in a Code Ocean capsule. Results can be exactly replicated by running the analyses in that capsule. Randomization and blinding do not apply to this work because we analyzed previously published and annotated datasets.

3.5 Appendix

3.5.1 Appendix A: Simulating 4D-communication tensors

Tensor simulation consisted of six steps 1) generating the protein-protein interaction (PPI) network for ligand-receptor (LR) pairs, 2) generating the cell-cell (CC) network, 3) separately labeling LR pairs and single cells with a metadata group or category (to represent signaling pathways and cell types, respectively), 4) randomly associating a subset of LR pairs in the same signaling pathway with a subset of sender-receiver cell type pairs (generating a LR-CC combination), 5) assigning a context-dependent pattern of communication scores to LR-CC combinations, and 6) filling the tensor with communication scores that follow assigned communication patterns across contexts.

The LR pairs are generated as a random, unweighted, bipartite network; each of the two node types represents either a ligand or a receptor. This network retained a scale-free property using *StabEco*'s (v0.1) *BiGraph* function, with the power law exponent value set to 2 and the average degree value set to 3. We confirmed that the degree distribution does fit the power-law using a maximum likelihood estimation method from *igraph* (v0.8.3), and proceeded to remove all disconnected nodes from the network. We next generated an edge list of all cell-cell interactions, which is simply all pairwise permutations of cells; permutations retain directionality, allowing a distinction between sender and receiver cells. In distinct cell-cell interactions, autocrine interactions (self-loops) are allowed. LR pairs are uniformly binned to one of three metadata categories. The same process is repeated for cells, which are then further condensed from individual cells to their category to represent bulk rather than single-cell data; autocrine interactions at the single-cell resolution are considered homotypic interactions at the bulk resolution. Conceptually, these categories can be thought of as analogous to biological groupings such as signaling types (e.g., paracrine and endocrine) for LR pairs or cell-types for cells.

We simulate changes to intercellular conditions across twelve contexts. In order to populate the tensor, we randomly assign four communication patterns--expected communication scores that change as a function of condition--to distinct LR-CC combinations. An LR-CC combination" consists of one LR pair label (signaling pathway) and two cell labels (cell types), one for the sender cell and another for the receiver cell. In assigning patterns, signaling pathways or pairs of cell types may be individually reused in different communication patterns, but their combinations must be unique. For example, take two signaling pathways with the labels "X" and "Z", and two cell types with the labels "A" and "B". Accounting for directionality, cells can form the following interactions: "AA", "AB", "BB", and "BA". Next, any of the following LR-CC combinations may be assigned a pattern: "X-AA", "X-AB", "X-BB", "X-BA", "Z-AA", "Z-AB", "Z-BB", "Z-BA". If "X-AB" is assigned to a pattern, then tensor coordinates for LR-pairs categorized under the signaling pathway "X", used by the sender cells categorized under cell type "A" and receiver cells categorized under cell type "B" will be filled with communication scores to reflect that pattern. LR-CC combinations not assigned to a pattern are considered the background. Biologically, the motivation for allowing redundancy in the LR-pair or CC-pair categories assigned to a pattern (e.g., “Z-AA” and “Z-AB”), is that the same groups of LR-pairs may dictate different types of interactions between different cell types (e.g., LR pairs under “Z” dictate a linear increase across contexts for “AA” cell interactions, but a pulse in “AB” cell interactions).

We then randomly initialized selected LR-CC combinations with one of four possible patterns (linear, oscillatory, pulsatile, and exponential) and their expected communication score in the first condition. Finally, we assign the expected communication score for each LR-CC combination in each proceeding context based on the respective pattern and starting communication score. All LR-CC combinations that were not chosen to have a pattern (the “background”) have an expected communication score of zero. Our final tensors have a size of 12 x 300 x 3 x 3 (contexts, LR pairs, sender cells and receiver cells, respectively).

3.5.2 Appendix B: Adding noise to simulations

To assess the robustness of our tensor factorization approach, we added noise to communication scores in each context and each pattern when filling the tensor. In the simulation, noise (represented as $\boldsymbol{n}$) is a parameter that can take on any value between 0 and 1. When $\boldsymbol{n} = 0$, all communication scores are set to their expected value $\boldsymbol{\mu}$. When $\boldsymbol{n} > 0$, we draw from a uniform truncated normal distribution with the minimum and maximum values set to 0 and 1, respectively, and the mean seat to the expected value. The standard deviation $\boldsymbol{\sigma}$ is a function of noise $\boldsymbol{n}$. For $\boldsymbol{\mu} > 0$, $\boldsymbol{\sigma} = \boldsymbol{c \cdot \mu \cdot n}$ where $\boldsymbol{c}$ is a scaling factor which we set to 1.1. We represent this distribution as N~($\mu = \boldsymbol{\mu}$, $\sigma = 1.1\boldsymbol{\mu n}$, min=0, max=1). This causes the communication score dispersion to increase with noise (Fig. 3.64a).

At $\boldsymbol{\mu} = 0$, as in the background and in cases where LR-CC combinations assigned to a pattern reach an expected value of 0, we must adjust $\boldsymbol{\sigma}$ to be a function of a different value (notated $\boldsymbol{\mu_L}$ instead of $\boldsymbol{\mu}$); otherwise $\boldsymbol{\sigma}$ would always equal 0 independently of the value of noise. Specifically, $\boldsymbol{\mu_L}$ represents the desired maximum average value of the communication scores at $\boldsymbol{\mu} = 0$, achieved when $\boldsymbol{n} = 1$. We refer to $\boldsymbol{\mu_L}$ as the maximum background noise. Thus, the distribution at $\boldsymbol{\mu} = 0$ is annotated as N~(μ= 0, $\sigma = \boldsymbol{\mu_L} * \boldsymbol{n}$, , min=0, max=1). The entire distribution is then scaled by a factor **c’** to ensure that the average value of the distribution at $\boldsymbol{n} = 1$, which we define as $\boldsymbol{\mu_B}$, is equal to $\boldsymbol{\mu_L}$**.** The need for scaling is apparent in its absence, or when setting **c’** = 1. In this case, $\boldsymbol{\mu_B}$ is consistently less than $\boldsymbol{\mu_L}$ at $\boldsymbol{n} = 1$ (Fig. 3.64b). In order to identify an appropriate value for **c’**, we calculate $\boldsymbol{\mu_B}$ at multiple values of $\boldsymbol{\mu_L}$ ranging between 0 and 1, setting **c’** = 1. Next we fit a piecewise function to this curve, defined by the linear function $\boldsymbol{\mu_B} = m\boldsymbol{\mu_L} + b$ when $\boldsymbol{\mu_L} < \boldsymbol{\mu}^*$ and the exponential function $\boldsymbol{\mu_B} = \left(\frac{\mu_L - d}{c}\right)^a$ when $\boldsymbol{\mu_L} \geq \boldsymbol{\mu}^*$. We use *scipy’s* (v1.6.0) *curve_fit* function to estimate the parameters *m*, *b*, *d*, *c*, *a*, and $\boldsymbol{\mu}^*$. This fit allows us to predict the average value of the distribution N~(μ= 0, $\sigma = \boldsymbol{\mu_L} * \boldsymbol{n}$, , min=0, max=1) when **c’** = 1 for a given $\boldsymbol{\mu_L}$. Annotating this prediction as $\boldsymbol{\mu'_B}$, we can then set **c’** = $\frac{\mu_L}{\mu'_B}$ to achieve the desired behavior of $\boldsymbol{\mu_B} = \boldsymbol{\mu_L}$ at $\boldsymbol{n} = 1$ (Fig. 3.64c). Finally, since this scaling approach results in communication scores > 1, we set all drawn values greater than 1 to 1 (Fig. 3.64d). By limiting $\boldsymbol{\mu_L}$ to values ≤ 0.25, we can avoid the bimodal distribution seen at $\boldsymbol{\mu_L} \geq 0.5$. We can change $\boldsymbol{\mu_L}$ for the background, which we refer to as the “maximum background score”, to any value between 0 and 0.25. For non-background contexts, when $\boldsymbol{\mu} = 0$, $\boldsymbol{\mu_L}$ is set to the smallest non-zero $\boldsymbol{\mu}$ for that specific LR-CC combination across all contexts. Allowing $\boldsymbol{\mu_L}$ for the background to be changed in the simulation permits the independent assessment of noise added to the patterns only, or both the patterns and the background.

In the example tensor (Fig. 1), we assigned a minimal baseline level of noise ($\boldsymbol{n} = 0.01$) to avoid any issues that may arise from sparsity due to assigning communication scores of 0 to the background. We scale communication scores of the background, i.e. LR-CC combinations that were not assigned to a pattern, in such a manner that the average value would be 0.05 when $\boldsymbol{n} = 1$ (i.e., we set $\boldsymbol{\mu}_{\mathbf{L}} = 0.05$).

3.5.3 Appendix C: Accuracy of the tensor decomposition in assigning ligand-receptor importance

To quantify the accuracy of the tensor decomposition in selecting important LR pairs for each factor, we used a Jaccard index. Here, we compared the LR loadings of the respective factor with the "ground truth" communication scores of LR pairs assigned to each pattern. To do so, we first binarized LR loadings in each factor by categorizing them into high or low loading equal-interval bins, under the assumption that the decomposition provides sufficient separation in loadings between those assigned to a pattern and those that are in the background. With these binarized scores, we can assume LR pairs in the high loading bin are influential for a particular factor. Finally, we computed the pairwise Jaccard index between LR pairs assigned to a given pattern from the simulation and "high-loading" LR pairs in each factor (Table 3.2). We expected a high Jaccard value exclusively between an assigned pattern and the factor that has loadings in the context dimension that recapitulate that pattern.

We also used a Pearson correlation metric to compare LR loadings with communication scores assigned to each pattern. Since all patterns are assigned to unique LR-CC combinations, we can reduce the tensor to a matrix confined to the contexts and LR pairs for any given assigned pattern by selecting only the sender and receiving cells assigned to that pattern. We expect that LR pairs with high loadings have high variance in communication scores across contexts, since this variance reflects context-dependent changes. Thus, we calculated the pairwise Pearson correlation between the context-driven variance in communication scores across all LR pairs that were assigned to a given pattern and the LR loadings of each factor (Table 3.3).

In the example simulated tensor (Fig. 2e-f), the assigned communication scores for each pattern form a bimodal distribution. The corresponding loadings of each factor resulting from tensor factorization similarly demonstrated bimodal distributions. Pairwise comparisons between communication scores and factor loadings resulted in Jaccard indices of 1 for LR pairs assigned to a pattern and 0 for others; similar results were obtained for Pearson correlations.

3.5.4 Appendix D: Assessing the robustness of the tensor decomposition to noise

To assess the robustness of Tensor-cell2cell to noise, we simulated tensors, as described previously, at varying levels of noise; we iterated through noise values ranging between 0 to 1 and measured the decomposition error in each iteration. In each iteration, much like an elbow analysis, we decomposed the simulated tensor to ranks of 4, 5, and 6. Between the three decompositions, we retain the one that resulted in the smallest error, with the caveat that additional ranks should decrease error by at least 1.1 log-fold change (LFC).

We ran the analysis 1000 times to evaluate the effect of different random seeds for generating noise while holding the maximum background score constant. We repeat this process for multiple maximum background score values. Next, at each maximal background score, we fit a locally weighted smoothing (LOESS) curve to the error and noise outputs using statsmodels' (v0.12.2) lowess function. This gave us predicted error measurements for each level of noise tested. To assess the amount of noise that needs to be added to the system to surpass a heuristic threshold of error--in this case 0.3--we then interpolated noise from the LOESS-predicted errors using scipy's interp1d function.

3.5.5 Appendix E: Tensor-cell2cell is robust to noise

We generated 133,000 simulated tensors using the same parameters for generating the example tensor (Fig. 1), but while also adding varying levels of noise to assess how the error of decomposition may be affected. To ensure that our assessment was independent of a specific interaction network, in each simulation, a new ligand-receptor PPI network was generated, and interaction patterns were assigned to new LR-CC combinations. Noise here was intended to represent non-biological variation that may arise in single-cell measurements due to a number of technical factors in RNA-sequencing, such as sample handling and library preparation. The addition of noise dampens the signal of each communication pattern in two ways: 1) by decreasing the change in communication score between conditions (resolution noise) and 2) by increasing the presence of interactions occurring by chance through randomly assigning higher communication scores to ligand-receptor and cell-cell pairs not assigned to a communication pattern (background noise).

When running Tensor-cell2cell on each of the simulated tensors, 85.1% of the decompositions resulted in a minimum error when the rank was set to equal the number of simulated patterns (r = 4) (Fig. 3.7a). Meanwhile, decompositions with ranks of 5 and 6 minimized error in 10.4% and 4.5% of cases, respectively. We assessed the level of noise required to surpass a heuristic error threshold of 0.3 at each level of maximum background noise (Fig. 3.7b). We found that this decomposition error was reached when adding a resolution noise between 0.26-0.37; the resolution noise needed to achieve an error of 0.3 monotonically increased with decreasing maximum background noise. Thus, Tensor-cell2cell is robust to both resolution and background noise. Taken together, these results indicate that Tensor-cell2cell is capable of handling noise in both the interacting and non-interacting cells when capturing various condition-dependent patterns.

3.5.6 Appendix F: Tensor-cell2cell is fast and accurate

We ran Tensor-cell2cell and CellChat on a single-cell transcriptome atlas of peripheral blood mononuclear cells (PBMCs) from COVID-19 patients with varying severity[68] to measure the time and memory demands of each tool when performing the context-driven CCC analysis (Fig. 3.8a-d). Considering the number of samples in this dataset, processing time of CellChat scales more rapidly with the high number of pairwise comparisons. To control this, we varied the number of samples and performed this benchmarking in two scenarios: 1) by considering every sample individually as a context, wherein one can obtain sample-specific signatures that may coincide with others of the same severity (Fig. 3.8a-b), and 2) by considering every severity (control, mild/moderate and severe/critical) as contexts by aggregating cognate samples (Fig. 3.8c-d), which keeps the number of pairwise comparisons constant at three comparisons but at the expense of losing sample-specific information.

For a fair comparison of the tools, we set them to perform the analysis using exactly the same list of LR pairs, shared cell types across samples, as well as to run only one CPU core. Tensor-cell2cell performed better in all cases we tested exclusively running in the CPU (without the GPU scenario, Fig. 3.8a-d). The most favorable scenario led to an ~89-fold improvement in running time, which occured when 60 samples were analyzed as individual contexts and CellChat comparisons were run under the "structural" method (computes a network topology dissimilarity[10]**)**. However, CellChat performed only half as fast as Tensor-cell2cell when using the “functional” comparison (Fig. 3.8a), which is based on a Jaccard similarity. Other cases where CellChat was comparable to Tensor-cell2cell were when samples were aggregated into major contexts (Fig. 3.8c); however, this demanded substantial increases in memory usage (Fig. 3.8d). The “structural” method can handle cell types that are not present in all contexts, which may explain the inferior speed than the “functional” method. However, the comparison between the “structural” method and Tensor-cell2cell is pertinent since our tool can also handle cell types that are not present in all contexts, despite that the current algorithm is not optimized for this purpose as other tensor decomposition methods do, and both tools were set to run on just cell types that are in all samples. In addition, part of the speed-up of Tensor-cell2cell over CellChat could be due to the language used to build each tool (CellChat runs on R, while Tensor-cell2cell runs on Python).

Importantly, the improvement of Tensor-cell2cell is achieved even though it runs a brute-force elbow analysis (that is, by computing the error for every rank in a range of values). In this regard, this step can be either omitted (for example, when a desired number of factors is used; No elbow case in Fig. 3.8a-d) or optimized (for example, when a binary search is used), multiplying the speed-up we reported here by ~10-fold and ~3-fold, respectively. Remarkably, the memory usage of Tensor-cell2cell never surpassed 16GB in any of the tested scenarios, even when using 60 samples as individual contexts (Fig. 3.8b); meanwhile, CellChat surpassed 16GB when aggregating 12 samples (Fig. 3.8d) or using 24 samples as individual contexts (Fig. 3.8b). Again, this could be due to the programming language that each tool uses. Nevertheless, in terms of what each user would have to deal with, Tensor-cell2cell is more efficient in both time and memory, indicating it can more readily be run on multiple contexts simultaneously in a personal computer or laptop. Moreover, Tensor-cell2cell can run on a GPU when available, which can substantially improve the computational time of the analysis (up to 19- and 790-fold faster than the “functional” and “structural” methods of CellChat, respectively, when analyzing 60 PBMC samples; Fig. 3.8a-d).

We next evaluated the accuracy of Tensor-cell2cell and CellChat in classifying individual samples after predicting context-driven CCC (Fig. 3.8e-h). It is important to consider that the outputs coming from each tool are extremely different due to the scope of their analyses, so a direct comparison is not feasible. Hence, we instead used an intermediary approach that uses a classification model to evaluate how well each tool separates contexts given their outputs. In particular, we measured how well each tool separates samples by COVID-19 severity (Fig. 3.8e-f) and disease state (Fig. 3.8g-h). For this, we trained a classifier to predict severity (control, mild/moderate vs severe/critical) and a disease state (healthy vs COVID-19) in two different COVID-19 datasets, one containing PBMC samples[68] and the other bronchoalveolar lavage fluid (BALF) samples[32]. We next measured their accuracy with the area under the receiver operating characteristic curve (AUC). Tensor-cell2cell outperformed CellChat when classifying PBMC samples by severity (Fig. 3.8e), and performed similarly when classifying samples by disease state (Fig. 3.8g). Moreover, Tensor-cell2cell performed better than CellChat in all classification tasks associated with BALF samples (Fig. 3.8f,h). Surprisingly, all methods performed better (highest AUC) when classifying BALF samples than when classifying PBMC samples, possibly due to a more evident severity-driven variation of the immune response in the infection site rather than in the periphery. Thus, these results show that Tensor-cell2cell can successfully find signatures of CCC that differentiate between contexts in a computationally efficient manner.

Importantly, the outputs and details offered by CellChat and Tensor-cell2cell differ. CellChat reports context-associated UMAP embeddings of signaling pathways, while Tensor-cell2cell outputs TCA embeddings for contexts, ligand-receptor pairs, and interconnected sender and receiver cells. By training classifiers that accept these differing outputs, in most cases, Tensor-cell2cell greatly outperformed CellChat (Fig. 3.8e,f,h). While in a few scenarios we observed qualitatively comparable performance (Fig. 3.8g, Tensor-cell2cell and the functional method of CellChat), Tensor-cell2cell always performed better quantitatively. Although context classification is a useful approach for comparison, this strategy cannot evaluate how well these methods infer CCC due to their distinct scopes and the vast differences in the tools' outputs. In this regard, the outputs of Tensor-cell2cell seem valuable for identifying specific molecular targets and involved cells of a context-dependent module of communication, encompassing information beyond the scope of CellChat, which largely focuses on pair-wise, context-specific differences between signaling pathways. Nevertheless, to make a fair comparison of both methods, we did not use all of the high dimensionality that Tensor-cell2cell outputs offer, and used just the context dimension loadings.

3.5.7 Appendix G: Tensor-cell2cell is robust to communication scoring inputs

To further compare decomposition results between communication scoring methods, we modified the CorrIndex metric to score each dimension separately, and retain the score of the most dissimilar dimension. While this approach disregards the combinatorial effects of patterns across dimensions that are accounted for in decomposition, it is more stringent and brings focus specifically to dissimilarities between decomposition. With this modified metric, we found the average similarity score decreased from 0.82 to 0.64. NATMI and CellChat, in particular, were more dissimilar from the other methods (Fig. 3.10), which agrees with the fact that these two exhibited the lowest similarity score of 0.68 using the unmodified CorrIndex.

Next, we asked whether a specific tensor dimension is driving the observed dissimilarity. We ran the Corrindex comparing each method on each tensor dimension separately (Fig. 3.11). While the sample, sender cell, and receiver cell dimensions all exhibited high similarity (average scores of 0.94, 0.92, and 0.93, respectively), the ligand-receptor dimension had a substantially lower similarity. In fact, we found that our modified scoring approach consistently identified the ligand-receptor dimension as the most dissimilar dimension across all comparisons. This indicates that the ligand-receptor dimension drives dissimilarity in decomposition outputs. This makes sense since the ligand-receptor dimension reflects the raw score output by each distinct method, as well as previous benchmarking reports concluding that the scoring method used has a substantial impact on the predicted interactions[22].

Given the high similarity of the other dimensions, as well as that of the unmodified CorrIndex, we conclude that Tensor-cell2cell is robust to differences at the ligand-receptor resolution, mitigating the propagation of differences within the ligand-receptor dimension to other dimensions and consistently identifying overarching communication patterns. We see that while the quantitative CorrIndex agrees visually with the qualitative consistency in factorization outputs (Fig. 3.64), we also see that there are some discrepancies at the cell-resolution, especially with NATMI. This is expected given differences in the ligand-receptor scores, which directly define the overall sender-receiver communication. The CorrIndex for sender- and receiver- dimensions is high because considering a single dimension independently of the other dimensions enables an optimal alignment that otherwise may not be possible; this pitfall is mitigated by the fact that the unmodified CorrIndex score is high. Interestingly, while CellChat is the most dissimilar at the ligand-receptor dimensions, NATMI is the most dissimilar for the other three dimensions, further reinforcing the importance of not only considering each dimension separately, but all dimensions in combination.

3.5.8 Appendix H: Tensor-cell2cell detects traditional mechanisms associated with immune response to infections

Factors computed by Tensor-cell2cell provide further insights about immune responses during SARS-CoV-2 infection. For example, factors 4, 6 and 7 are associated with lymphocytes (Fig. 4a), particularly the communication of NK and T cells (factors 4 and 6), and B cells (factors 4 and 7). CD44 is predicted to be an important receptor on lymphocytes (factor 4 in Fig. 4b), which consistently is key for cell migration[78] and resolving lung inflammation[79,80]. Among the top-ranked signaling molecules in NK and T cells as senders (factor 6 in Fig. 4b), we found CCL5 and GZMA (granzyme A), which are involved in immune cell activation and cytotoxic effector functions of CD8+ T cell and NK cells, events that are key in the control of viral infections[81–83], as well as the interaction of PTPRC (CD45) with MRC1 (CD206), which regulates T-cell functionality[84]. Similarly, factor 5 seems associated with antigen-presenting cells such as mDC, macrophages and B cells as senders, especially through known interactions facilitating antigen presentation[85–88] (e.g. CD99-CD99 as well as interactions between integrin ITGB2 and intercellular adhesion molecules ICAM1 and ICAM2). Therefore, our strategy can successfully detect meaningful biological processes and cell-cell interactions involved during disease progression.

3.5.9 Appendix I: Tensor-cell2cell elucidates molecular mechanisms distinguishing moderate from severe COVID-19.

Tensor-cell2cell recapitulated molecular findings such as the role of SEMA4D-PLXNB2 interaction promoting inflammation validated in another work[45], interaction that Tensor-cell2cell revealed to be stronger in cases with more lung inflammation (severe cases) (Fig. 4). Our method also associated macrophage CCC in severe cases with interactions between CCL2, CCL3, CCR1 and CCR5, that are main proinflammatory molecules in COVID-19 as observed in another work[33], wherein their importance as potential therapeutic targets for diminishing COVID-19 severity was proposed[33]. Additionally, we identified novel CCC patterns and mechanisms regarding COVID-19 pathogenesis. For example, Grant *et al.* reported that CD206^{hi} alveolar macrophages participate in the immune response to SARS-CoV-2 infection[51], but the underlying mechanisms mediating this response remain unclear. Factor 10 seems to extend the results presented by Grant *et al.* by showing that macrophage-expressed MRC1 (CD206) interacts with PTPRC (CD45) expressed by other cells (Fig. 4a-b). Interestingly, the MRC1-PTPRC interaction mediating macrophage communication can promote immune tolerance[84], which is consistent with factor 10 being associated with moderate cases, wherein anti-inflammatory macrophages (M2-like phenotype) seem to be characteristic. Remarkably, the source article of the BALF dataset[32] reported that M2-like (anti-inflammatory) macrophages were present with higher frequency than M1-like (pro-inflammatory) macrophages in healthy and moderate COVID-19 patients, while M1-like macrophages were more frequent in severe COVID-19 patients[32,89], supporting the results of Tensor-cell2cell. However, this work only detected differences in the cellular compositions detected by their markers, but did not provide a link with molecular mechanisms. Another example is a recent GWAS study that reported 13 significant loci associated with SARS-CoV-2 infection[90], wherein ICAM1 popped up as an involved gene. Remarkably, Tensor-cell2cell assigned a high loading to the ITGB2-ICAM1 interaction in a communication pattern that seems to be associated with antigen presentation (factor 5, Fig. 4a-b), providing further insights of its potential mechanism.

3.5.10 Appendix J: ASD pathogenesis could be explained by factors 3 and 4

A longstanding hypothesis for ASD pathogenesis is that in some subjects, neurons exhibit local hyperconnectivity, but deficits in longer range connections[91–93]. Neurons in cortical layers

2/3 and corticocortical projecting neurons in layers 5/6 are the main sender cells in factor 3 (Fig. 5a and Fig. 3.14), suggesting that factor 3 may relate to local overconnectivity in ASD. Two key regulators of neurodevelopment, Neuregulin 1 (NRG1) and Ephrin A5 (EFNA5), were the main ligands associated with sender cells. Erb-B2 Receptor Tyrosine Kinases (ERBB2-4) and ephrin receptors (EPHA3-5,7 and EPHB2), another class of receptor protein-tyrosine kinases, were the main receivers. Factor 3 receptors were broadly expressed across cell types. The neuregulin/ERB and ephrin systems play major roles in neuronal migration[94,95] and axon guidance[96], so reduced signaling in ASD may contribute to migration errors and local hyperconnectivity.

A second long-standing theory for ASD pathogenesis is that an imbalance between excitation and inhibition contributes to cortical dysfunction[97,98]. Inhibitory interneurons are the key receiver cell types in factor 4 (Fig. 5a and Fig. 3.14), with parvalbumin interneurons (IN-PV), and SV2C-expressing interneurons (IN-SV2C) as the top ranked cells, suggesting that factor 4 could relate to excitation-inhibition imbalance in ASD. Pleiotrophin (PTN), Protein Tyrosine Phosphatase Receptor Type M (PTPRM), and Heparin Binding EGF Like Growth Factor (HBEGF) were the top senders in factor 4. The main receivers were Anaplastic Lymphoma Receptor Tyrosine Kinase (ALK), PTPRM, and ERBB4. PTN is released by neural stem cells to promote the normal development of newborn neurons by binding to ALK, a receptor tyrosine kinase in the insulin receptor family[99]. The insulin/ insulin-like growth factor 1 (IGF-1) signaling pathway has previously been implicated as a potential therapeutic target for normalizing functional connectivity dysregulation in syndromic and idiopathic ASD[100–102] and a pilot clinical trial studying the effects of IGF-1 as a pharmacotherapeutic for ASD showed promising results[103]. These results suggest that Tensor-cell2cell could have utility as a tool for identifying novel pharmacotherapeutic targets.

3.5.11 Appendix K: Analysis of the BALF COVID-19 dataset with CellChat

We ran the analysis that CellChat offers for pairwise comparisons of contexts using the BALF dataset on the same conditions as used with Tensor-cell2cell. To simplify the analysis and interpretation, samples with the same severity were aggregated (Fig. 3.15-13). From the joint manifold learning analysis on signaling pathways (Fig. 3.15a), CellChat groups functionally similar molecules, especially in the immune response they participate in (e.g. cell adhesion and cytokines). By inspecting the pairwise comparison on signaling pathways (Fig. 3.15b), pathways such as GDF, OCLN, SELL, LAMININ and SPP1 seem to increase as severity increases. They are associated with moderate cases in the comparison of healthy vs moderate COVID-19, and associated with severe cases in the comparison of moderate vs severe. This suggests a potential correlation between these specific pathways and COVID-19 severity. Particularly, growth differentiation factors (GDFs) are the ones that Tensor-cell2cell did not detect, and they correspond to stress-, infection-, and inflammation-induced cytokines that can suppress immune responses[104], potentially explaining the severity association detected by CellChat. Similarly, Occludin interaction (OCLN-OCLN) was not detected by Tensor-cell2cell's built-in scoring function, an interaction that corresponds to an airway tight junction that is one target of viral infections[105]. SELL, Laminins and SPP1 involving integrins were captured by Tensor-cell2cell (Fig. 4). Although SPP1 specific interactions are not among the top-5 LR pairs, they are among the top-10 pairs; and other integrin interactions were detected by Tensor-cell2cell. Here, it is easily noticeable that autocrine interactions of macrophages, involving CCL2, CCL3, CCL7 and CCL8 with their respective receptors (CCR1, CCR2 and CCR5), are increased in more severe cases; result that is coherent with the factor 8 output by Tensor-cell2cell. Thus, CellChat can offer pathway-level details missed by Tensor-cell2cell, but without providing information about their specific cellular associations. Furthermore, CellChat misses other mechanisms that Tensor-cell2cell found (Fig. 4). Nevertheless, CellChat is still a powerful tool that can detect some of the results we presented with Tensor-cell2cell, such as the role of SELL and LAMININ pathways, the association of MIF and healthy patients, and the role of macrophages in more severe cases, potentially explaining its good performance in the classification benchmarking (Fig. 3.8e-h). Importantly, these discrepancies may be mitigated by extending Tensor-cell2cell to use other communication scoring methods as input, e.g. CellChat's communication probabilities (Fig. 3a and Fig. 3.9).

Finally, a key conceptual limitation of CellChat and other communication scoring tools, as compared to Tensor-cell2cell, is the need for pairwise comparisons between contexts, which prevents the identification of communication patterns across contexts and results in an exponential increase in the number of results with increasing samples/contexts, even when aggregating samples by context groups (Fig. 3.15-13). Another limitation is the inability to consider all biological scales (ligand-receptor, cell-cell, and context) simultaneously, both of which reduce the interpretability of results.

3.5.12 Appendix L: Additional Tables

Table 3.2: Jaccard index to evaluate the accuracy of Tensor-cell2cell on ranking ligand-receptor pairs.

Pattern	Decomposition Factor			
	Factor 1	Factor 2	Factor 3	Factor 4
Pulse	0.00	1.00	0.00	0.00
Oscillation	1.00	0.00	0.00	0.00
Linear*	0.00	0.00	1.00	1.00
Exponential*	0.00	0.00	1.00	1.00

* In this simulation, the same set of LR pairs was assigned to both the exponential and linear patterns (but in each case they were used by different sender-receiver cell pairs). Thus, when the exponential pattern has a high Jaccard index, the linear pattern is expected to have a high Jaccard index as well, and vice versa.

Table 3.3: Pearson correlation to evaluate the consistency between ground truth ligand-receptor pairs and LR loadings.

Pattern	Decomposition Factor			
	Factor 1	Factor 2	Factor 3	Factor 4
Pulse	-0.49	1.00	-0.50	-0.50
Oscillation	1.00	-0.49	-0.51	-0.51
Linear*	-0.51	-0.50	1.00	1.00
Exponential*	-0.51	-0.50	1.00	1.00

* In this simulation, the same set of LR pairs was assigned to both the exponential and linear patterns (but in each case they were used by different sender-receiver cell pairs). Thus, when the exponential pattern has a high Pearson correlation, the linear pattern is expected to have a high Pearson correlation as well, and vice versa.

Table 3.4: Literature support for the top-ranked ligand-receptor interactions of the COVID-19 study case.

Factor	Ligand*	Receptor*	Reported role in immune response and/or COVID-19	Refs.
Factor 1	CD99	CD99	CD99 is involved in the transendothelial migration of neutrophils and monocytes	106
	MIF	CD74 & CD44	During early infection in COVID-19, plasma MIF concentration is increased. Reported association between an early MIF response, organ malfunction and 28-day survival	107
			MIF-CD74 promotes wound healing and recovery during lung injury, including injuries associated with viral infections	40,108
			CD74 is involved in blocking the entry of coronaviruses through endosomes, including SARS-CoV-2	109
	MDK	NCL	Potential role of MDK in facilitating viral entry. Nucleolin (NCL) facilitates nucleocytoplasmic transport of MDK	39
		ITGA4 & ITGB1	During inflammation, MDK-integrin interactions facilitate neutrophil trafficking	110
Factor 2	SEMA4D	PLXNB2	SEMA4D-PLXNB2 interaction participates in epithelial wound repair	111
			SEMA4D regulates allergic inflammation in lungs. Participates in neutrophil activation	112
			SEMA4D-PLXNB2 interaction promotes inflammation and neurodegeneration in experimental autoimmune encephalomyelitis	45
	SEMA4A	PLXNB2	SEMA4A-PLXNB2 induces production of Th17 cytokines in CD4+ cells	113
	-	SDC4	Syndecan-4 (SDC4) contributes to the cell entry of SARS-CoV-2 and attenuates antiviral immunity	114

Table 3.4: Literature support for the top-ranked ligand-receptor interactions of the COVID-19 study case.

Factor	Ligand*	Receptor*	Reported role in immune response and/or COVID-19	Refs.
Factor 3	SIGLEC 1	SPN	Siglec-1 binds SPN (CD43) during viral infection and this interaction inhibits interferon gamma production in T cells	115
	RETN	CAP1	Resistin (RETN) cytokine signaling via its receptor, CAP1, activates multiple inflammatory signaling pathways in monocytes	116
			RETN is a marker of neutrophil activation, which has an increased production in patients with critical COVID-19	117
	FN1	-	Increased levels of FN1 in lungs of patients with lung fibrosis	118
		ITGA4 & ITGB1	α4β1 integrin (ITGA4 & ITGB1) is involved in the recruitment of leukocytes by activated endothelial cells, involving migration processes such as tethering, rolling, arrest and adhesion. Fibronectin is one of its ligands.	119
		ITGA4 & ITGB7	α4 integrins are a target of natalizumab, a drug for treating multiple sclerosis. This drug had a favorable outcome in a COVID-19 patient with multiple sclerosis	120
Factor 4	COL9A2	CD44	CD44 is a cellular adhesion molecule and receptor for laminin and collagen, among other ligands. Increased expression of CD44 in bronchial samples from patients with critical COVID-19. Increased collagen in lungs during viral infection. Laminin is increased in serum of COVID-19 patients. CD44 participates in resolution of lung inflammation.	33,80,121–123
	LAMB3			
	LAMB2			
	LGALS9	CD44	LGALS9-CD44 regulates the immune response. LGALS9 is expressed by induced regulatory T cells as a mediator of immune suppression, which can act on cells expressing CD44	124

Table 3.4: Literature support for the top-ranked ligand-receptor interactions of the COVID-19 study case.

Factor	Ligand*	Receptor*	Reported role in immune response and/or COVID-19	Refs.
Factor 5	ITGB2	CD226	When upregulated on CD8+ T cells, CD226 enhances their cytotoxic effector functions	125
			CD226 is upregulated on CD8+ T cells of COVID-19 patients	126
			CD226+ monocytes are increased in COVID-19 patients compared to healthy patients	127
	CD86	CTLA4	CTLA4 is upregulated in CD8+ T cells present in lungs of COVID-19 patients. CD86 is upregulated in macrophages with M2-like phenotype present in BALF samples of patients with severe COVID-19	48
	ITGB2	ICAM2	Expression of ICAM2 is upregulated in lung epithelium infected by SARS-CoV2	128
		ICAM1	High expression of ICAM1 in lung epithelium infected by SARS-CoV2	128
			Increased presence of ICAM1 in lungs of patients with COVID-19	129
			Increased presence of ICAM1 in serum of patients with COVID-19, and higher levels were observed in non-survivors than in survivors	130,131
			Neutrophil cytotoxicity is mediated by the interaction between ITGB2 (CD18) and ICAM1	132
			ICAM1 is associated with COVID-19 by a GWAS study	90

Table 3.4: Literature support for the top-ranked ligand-receptor interactions of the COVID-19 study case.

Factor	Ligand*	Receptor*	Reported role in immune response and/or COVID-19	Refs.
Factor 6	CCL5	-	Increased level of CCL5 at the early stage of infection in patients with mild COVID-19	[133]
		CCR5	CCR5 gene expression is increased in BALF samples of patients with COVID-19	[134]
			CCR5 Δ32 polymorphism is positively correlated with SARS-CoV2 infection and COVID-19 mortality rate	[135]
			CCL5-CCR5 interaction induces migration of macrophages and NK cells	[136]
		CCR1	Protective role of CCR1 and CCR5 in a MA15-SARS-CoV mouse model infection, and in human DCs infected with SARS-CoV.	[136]
			CCL5-CCR1 interaction induces migration of macrophages and NK cells	[136]
	GZMA	F2R	GZMA is involved in the response of CD8+ T and NK cells to SARS-CoV-2, helping to distinguish healthy patients from COVID-19 patients. Protease-activated receptor 1 (F2R) has been identified as a potential target for therapeutic purposes in COVID-19	[82,137]
Factor 7	SELL	-	L-selectin (SELL) regulates neutrophil trafficking to sites of inflammation	[138]
			High expression of SELL in a subpopulation of monocytes present in patients with COVID-19	[127]
		MADCAM1	SELL-mediated lymphocyte rolling on MADCAM1	[139]
	CD22	PTPRC	CD22 is involved in response of B cells to SARS-CoV-2. PTPRC (CD45) is an immune regulator associated with severity in COVID-19	[140,141]

Table 3.4: Literature support for the top-ranked ligand-receptor interactions of the COVID-19 study case.

Facto r	Ligand*	Receptor*	Reported role in immune response and/or COVID-19	Refs.
Factor 8	CCL2	CCR2	Presence of CCL2 in plasma of patients with critical COVID-19 is higher than in plasma of patients with mild COVID-19	142
			Increased expression of CCL8 in postmortem lungs from patients with COVID-19	143
			CCR2 gene expression is increased in BALF samples of patients with COVID-19	134
			CCL2-CCR2 interaction induces migration of inflammatory monocytes	136
	CCL3	CCR5	High expression of CCL3 in the respiratory tract of patients with COVID-19	144
			CCL3 gene expression is increased in BALF samples of patients with COVID-19	134
			CCL3-CCR5 interaction induces migration of macrophages and NK cells	136
	CCL8	-	Increased expression of CCL8 in postmortem lungs from patients with COVID-19	143
		CCR1	CCL8-CCR1 interaction induces migration of T cells involved in Th2 response	136
	CCL3L1	CCR1	High expression of CCL3L1 in BALF samples from patients with COVID-19	145

Table 3.4: Literature support for the top-ranked ligand-receptor interactions of the COVID-19 study case.

Factor	Ligand*	Receptor*	Reported role in immune response and/or COVID-19	Refs.
Factor 10	CD99	PILRA	PILRA is an inhibitory receptor expressed on macrophages. CD99 is a ligand of this receptor. PILRA regulates the recruitment of neutrophils in inflammatory responses. A knock-out mouse model of PILRA had neutrophils with enhanced transmigration	47,146
	LGALS9	HAVCR2	HAVCR2 (TIM-3), is involved in T cell exhaustion and tolerance. TIM-3 is involved in T cell exhaustion in a progressive way with the severity of COVID-19	147,148
			High expression of TIM-3 in CD8+ T cells of patients with severe COVID-19, its expression is higher in CD8+ T cells from BALF compared to CD8+ T cells from PBMCs	149
	ANXA1	-	ANXA1 orchestrates epithelial repair	150
			A mimetic peptide of ANXA1 has been proposed as a potential treatment of severe COVID-19	151
		FPR1	FPR1 promotes wound healing in the respiratory tract	152
	MDK	LRP1	LRP-1 is one of the main receptors of MDK. MDK promotes the recruitment of polymorphonuclear cells during an acute inflammatory response. A recent article suggested a potential role of MDK-LRP1 interaction in neutrophil infiltration and the neutrophil extracellular trap formation during COVID-19	39,110
	PTPRC	MRC1	MRC1 is expressed on surfaces of macrophages. PTPRC (CD45) and MRC1 (CD206) interaction promotes immune tolerance. CD206^{hi} macrophages are involved in immune response to SARS-CoV-2 infection. CD45 is an immune regulator associated with severity in COVID-19	51,84,141

* Some ligand-receptor interactions are repeated in different factors, but in this table they are mentioned only in their first appearance. Repeated LR pairs in different factors can be seen in Fig. 4b.

Table 3.5: Top-5 ligand-receptor pairs captured in each factor of the tensor decomposition of the ASD data set.

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
NEGR1 - NEGR1	LAMA2 - SV2B	NRG1 - ERBB2&ERBB4	PTN - ALK	NRG3 - ERBB4	VEGFA - FLT1
NRXN3 - NLGN1	LAMA4 - SV2B	NRG1 - ERBB2&ERBB3	PTPRM - PTPRM	NCAM1 - NCAM2	PTPRM - PTPRM
NRXN1 - NLGN1	LAMB2 - SV2B	NRG1 - ERBB3	HBEGF - ERBB4	CADM1 - CADM1	VEGFB - FLT1
CTN1 - NRCAM	LAMA1 - SV2B	NRG1 - ERBB4	NRG3 - ERBB4	NCAM1- NCAM1	NRXN1 - NLGN2
NCAM1 - NCAM1	LAMA3 - SV2B	EFNA5 - EPHB2	BTC - ERBB4	CTN1 - NRCAM	PTN - NCL

3.5.13 Appendix M: Additional Figures

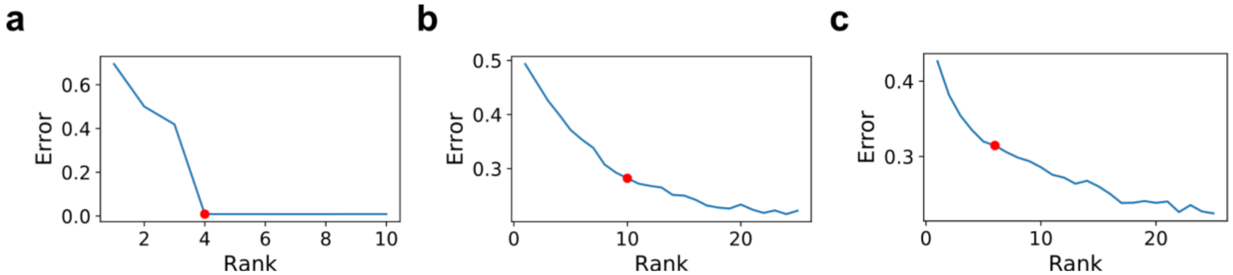


Figure 3.6: Rank selection for the tensor decomposition through an elbow analysis. A selection of the rank employed for the tensor decomposition (i.e. number of factors) is obtained by an elbow analysis on a normalized error associated with the reconstruction of the original tensor (see the Methods section). This analysis was performed on the tensors built from (**a**) the simulated tensor of communication scores, (**b**) the BALF-COVID-19, and (**c**) PFC-ASD datasets. The red dot indicates the rank selected in each case for performing the decomposition analysis. The criteria here were selecting an error value close to 0.3 (or smaller if possible) and to the area with no substantial changes when increasing the rank used in the factorization (elbow), while also keeping the decomposition rank as low as possible. Source data can be found in the Code Ocean capsule.

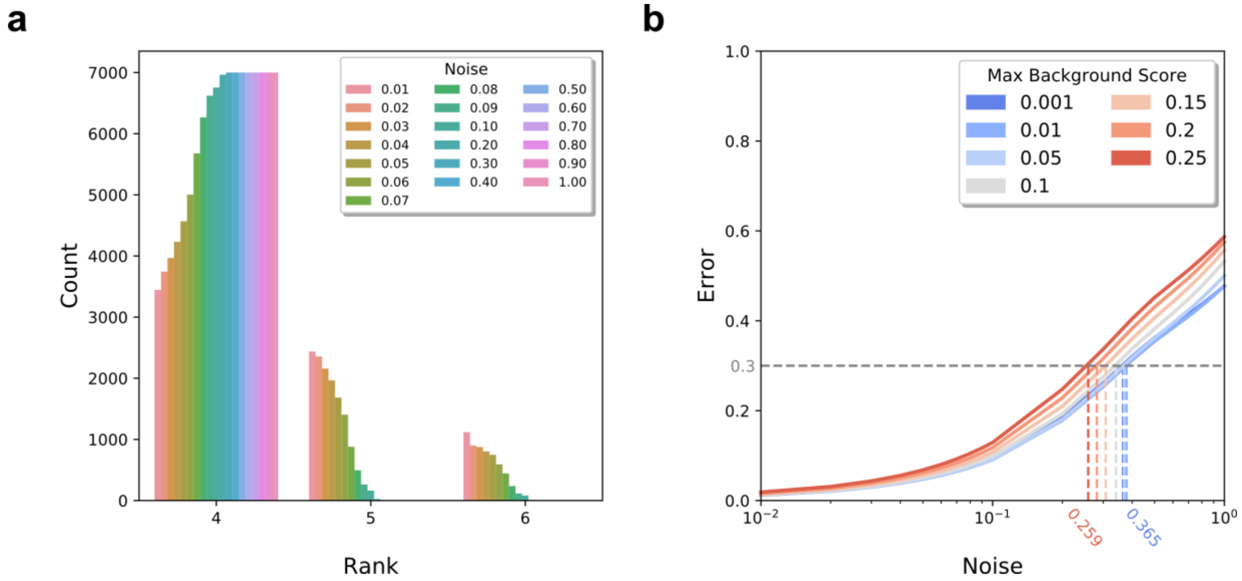
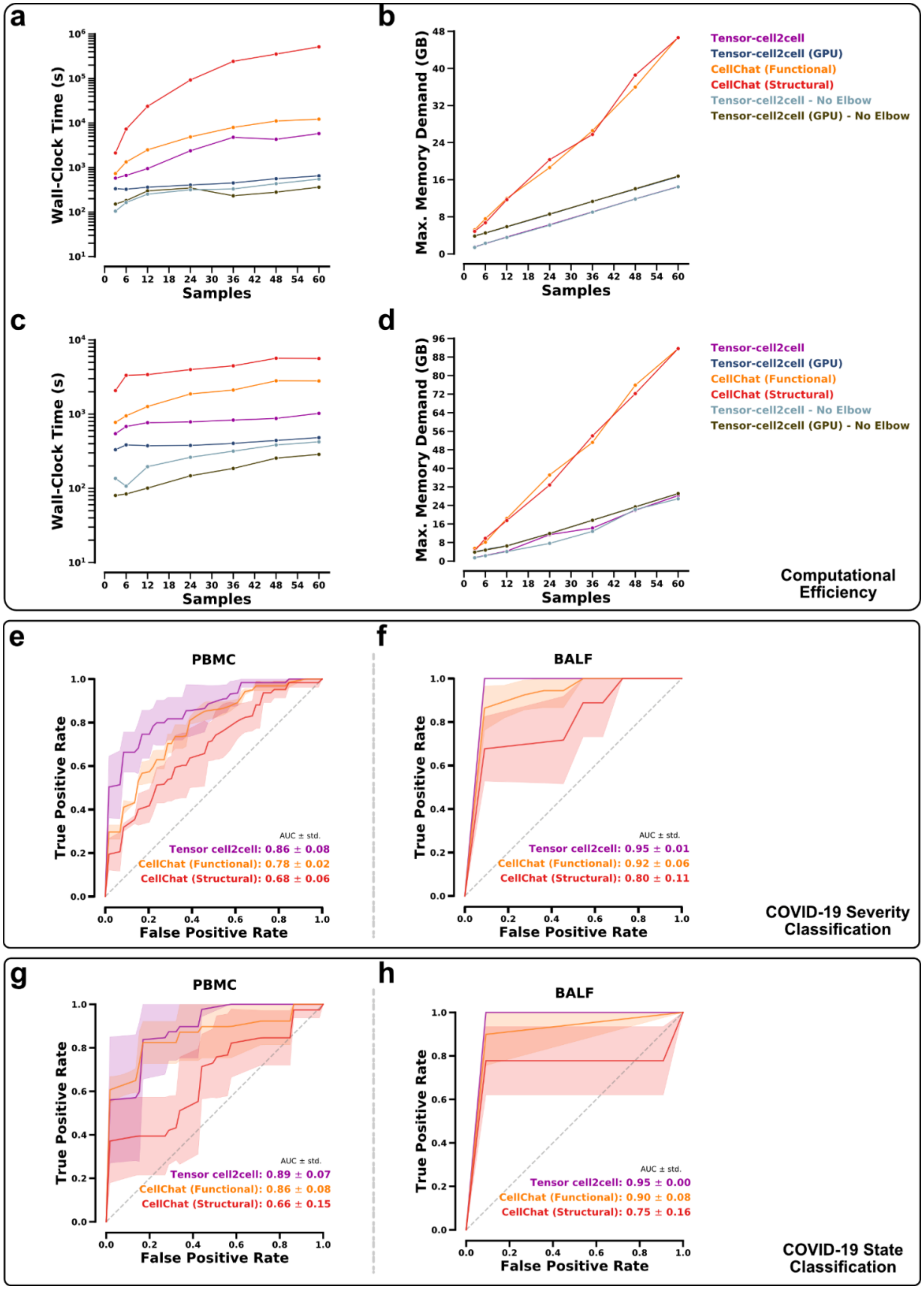


Figure 3.7: Tensor-cell2cell is robust to noise. (**a**) Across all simulations, the total counts (y-axis) of each decomposition rank (x-axis) selected for error minimization. Counts within each rank are stratified by the level of noise (**b**) Locally weighted smoothing curves (LOESS, solid lines) visualizing the relationship between decomposition error (y-axis) and noise added to the communication scores during tensor simulation (x-axis). The maximum average communication score of the background in each simulation was limited to values ranging between 0.001 and 0.25 (legend). The amount of noise needed to achieve the heuristic error threshold of 0.3 (dashed lines) are interpolated from the LOESS predicted values for error. Interpolated noise values are displayed for maximum background noise values of 0.001 and 0.25. Source data can be found in the Code Ocean capsule.

Figure 3.8: Benchmarking Tensor-cell2cell.

(**a**) Running time of Tensor-cell2cell and CellChat for analyzing CCC when the contexts correspond to individual patient samples. (**b**) Memory usage of each method to perform analyses in (a). (**c**) Running time of Tensor-cell2cell and CellChat for analyzing CCC when the contexts correspond to COVID-19 severity (individual samples were aggregated by severity: control, mild/moderate and severe/critical COVID-19). (**d**) Memory usage of each method to perform the analyses in (c). In (a) and (c), Tensor-cell2cell was benchmarked when running with or without a GPU, a feature unavailable in CellChat, and when considering or not the elbow analysis for selecting the number of factors. In addition, CellChat was benchmarked by using the two approaches it has for pairwise comparisons (functional and structural similarities, see Methods). (**e-h**) Receiver operating characteristic (ROC) curves of random forest models for classifying individual samples from the outputs of Tensor-cell2cell and the two CellChat approaches (functional and structural). These models predict specific severities of patients (control, mild/moderate or severe/critical) for (**e**) PBMC and (**f**) BALF samples, and disease state (healthy or COVID-19) for (**g**) PBMC and (**h**) BALF samples. For each classifier, the mean (solid line) ± standard deviation (transparent area) of the ROCs were computed from the 3-fold cross validations. Source data can be found in the Code Ocean capsule.


a
Wall-Clock Time (s)
10¹
10²
10³
10⁴
10⁵
10⁶
0 6 12 18 24 30 36 42 48 54 60
Samples
b
Max. Memory Demand (GB)
0 8 16 24 32 40 48
0 6 12 18 24 30 36 42 48 54 60
Samples
Tensor-cell2cell
Tensor-cell2cell (GPU)
CellChat (Functional)
CellChat (Structural)
Tensor-cell2cell - No Elbow
Tensor-cell2cell (GPU) - No Elbow
c
Wall-Clock Time (s)
10¹
10²
10³
10⁴
0 6 12 18 24 30 36 42 48 54 60
Samples
d
Max. Memory Demand (GB)
0 8 16 24 32 40 48 56 64 72 80 88 96
0 6 12 18 24 30 36 42 48 54 60
Samples
Tensor-cell2cell
Tensor-cell2cell (GPU)
CellChat (Functional)
CellChat (Structural)
Tensor-cell2cell - No Elbow
Tensor-cell2cell (GPU) - No Elbow
Computational Efficiency
e
PBMC
True Positive Rate
False Positive Rate
0.0 0.2 0.4 0.6 0.8 1.0
AUC ± std.
Tensor cell2cell: 0.86 ± 0.08
CellChat (Functional): 0.78 ± 0.02
CellChat (Structural): 0.68 ± 0.06
f
BALF
True Positive Rate
False Positive Rate
0.0 0.2 0.4 0.6 0.8 1.0
AUC ± std.
Tensor cell2cell: 0.95 ± 0.01
CellChat (Functional): 0.92 ± 0.06
CellChat (Structural): 0.80 ± 0.11
COVID-19 Severity Classification
g
PBMC
True Positive Rate
False Positive Rate
0.0 0.2 0.4 0.6 0.8 1.0
AUC ± std.
Tensor cell2cell: 0.89 ± 0.07
CellChat (Functional): 0.86 ± 0.08
CellChat (Structural): 0.66 ± 0.15
h
BALF
True Positive Rate
False Positive Rate
0.0 0.2 0.4 0.6 0.8 1.0
AUC ± std.
Tensor cell2cell: 0.95 ± 0.00
CellChat (Functional): 0.90 ± 0.08
CellChat (Structural): 0.75 ± 0.16
COVID-19 State Classification

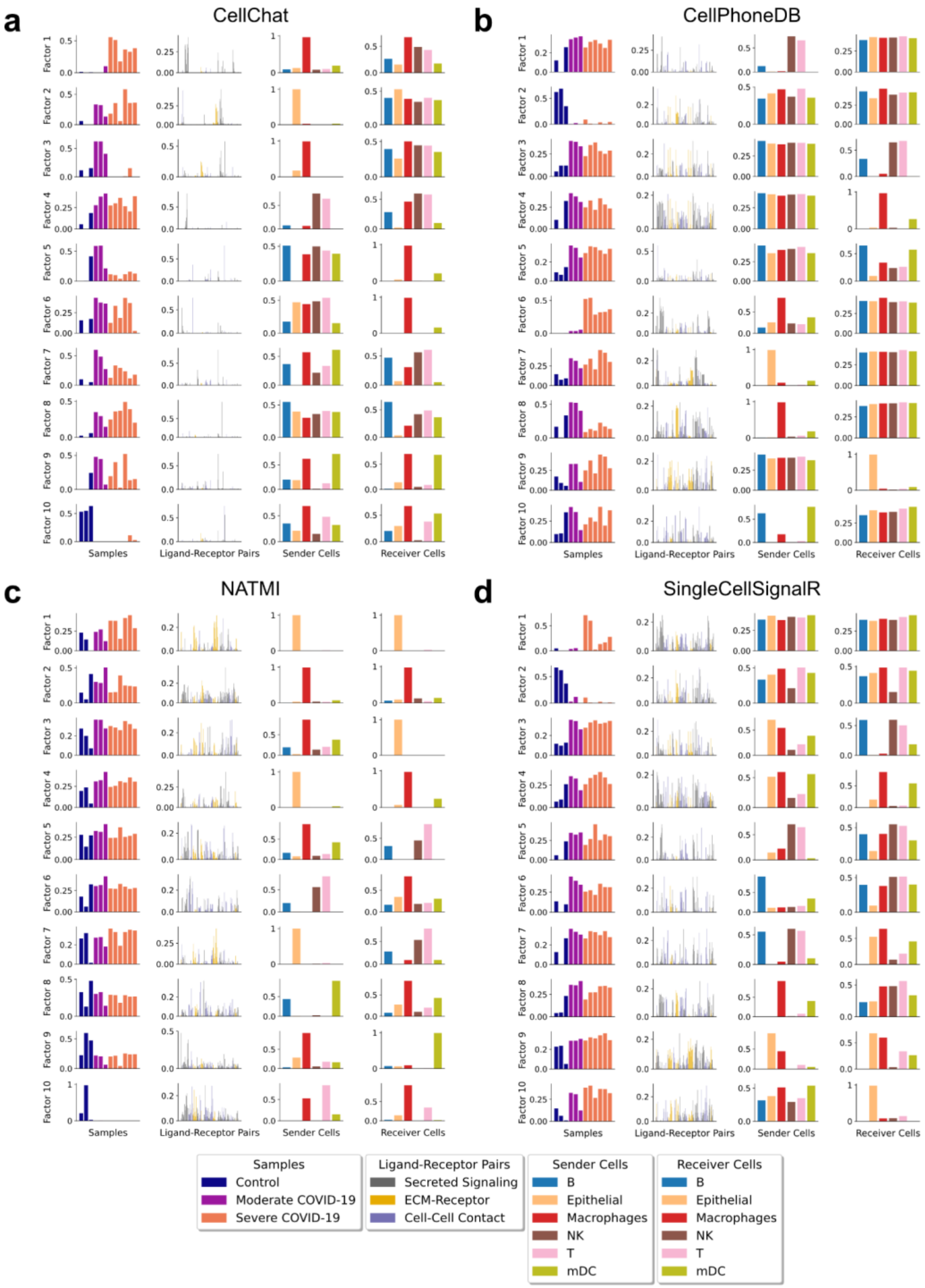


Figure 3.9: Tensor decompositions when using external tools to compute the communication scores.

Factors obtained after decomposing the 4D-Communication Tensors constructed from running external tools on each of the twelve BALF COVID-19 samples separately. The external tools used were (**a**) CellChat, (**b**) CellPhoneDB, (**c)** NATMI, and (**d**) SingleCellSignalR. For consistency, 10 factors (rank = 10) were selected for this analysis and 4D-Communication Tensors were subsetted to the same 176 ligand-receptor pairs, taken from CellChat's database, and filtered for those present in all samples and scored across all tools prior to running decomposition. Source data can be found in the Code Ocean capsule.

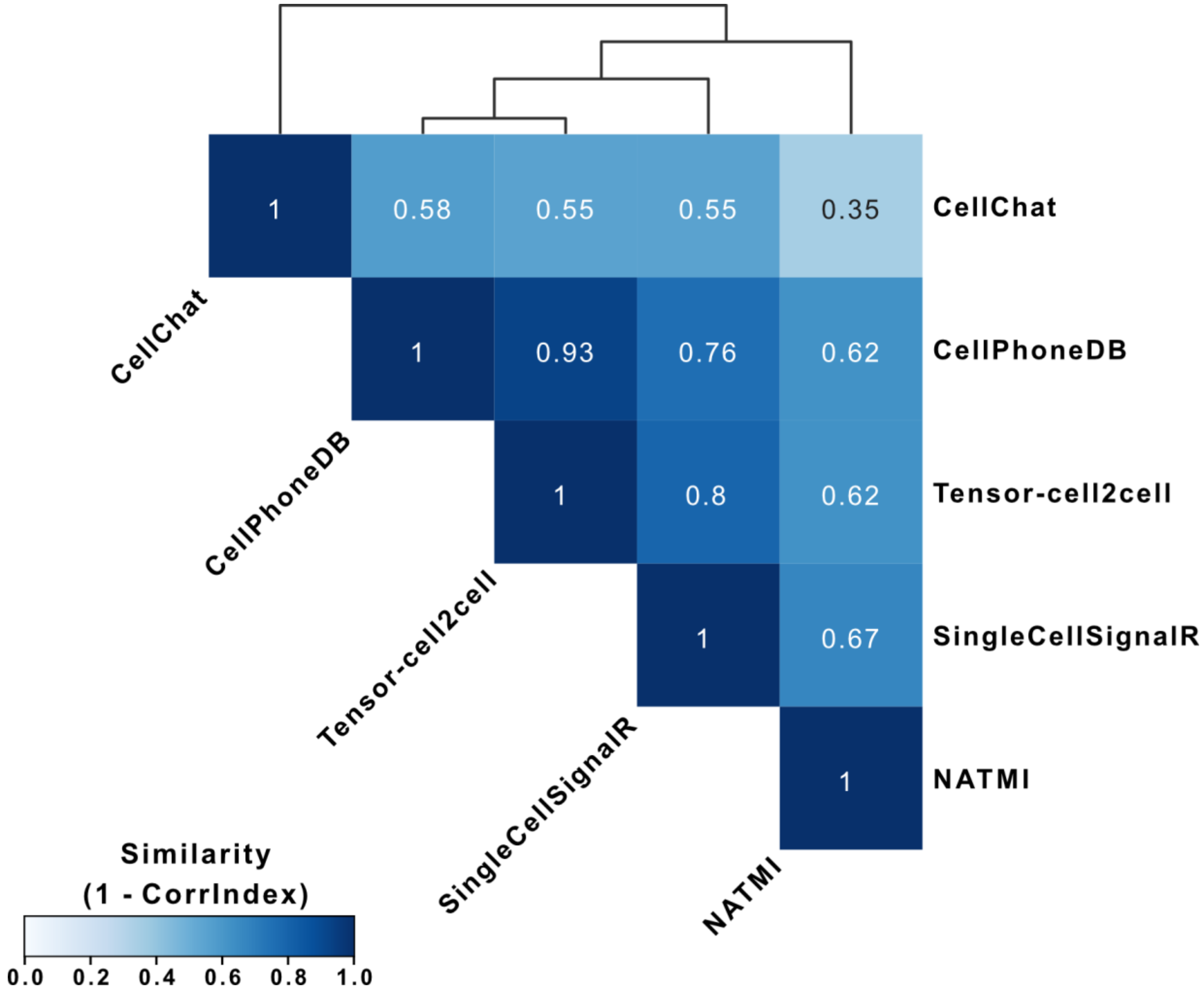


Figure 3.10: Conservative evaluation of decomposition similarity when using different communication scores.
Similarity (1 - CorrIndex) between tensor decompositions performed on the same 4D-communication tensor for a single-cell dataset of BALF in patients with varying severities, but with the communication scores computed from different tools for inferring cell-cell communication. Here, the CorrIndex score is modified for stringency by calculating it on each tensor dimension separately and subsequently selecting the maximal value (most dissimilar). Pairwise resultant similarities between all scoring methods are hierarchically clustered and visualized as a heatmap. Note that, since the ligand-receptor dimension was consistently the most dissimilar across all comparisons, this Fig. is the same as Fig. 3.11b. Source data can be found in the Code Ocean capsule.

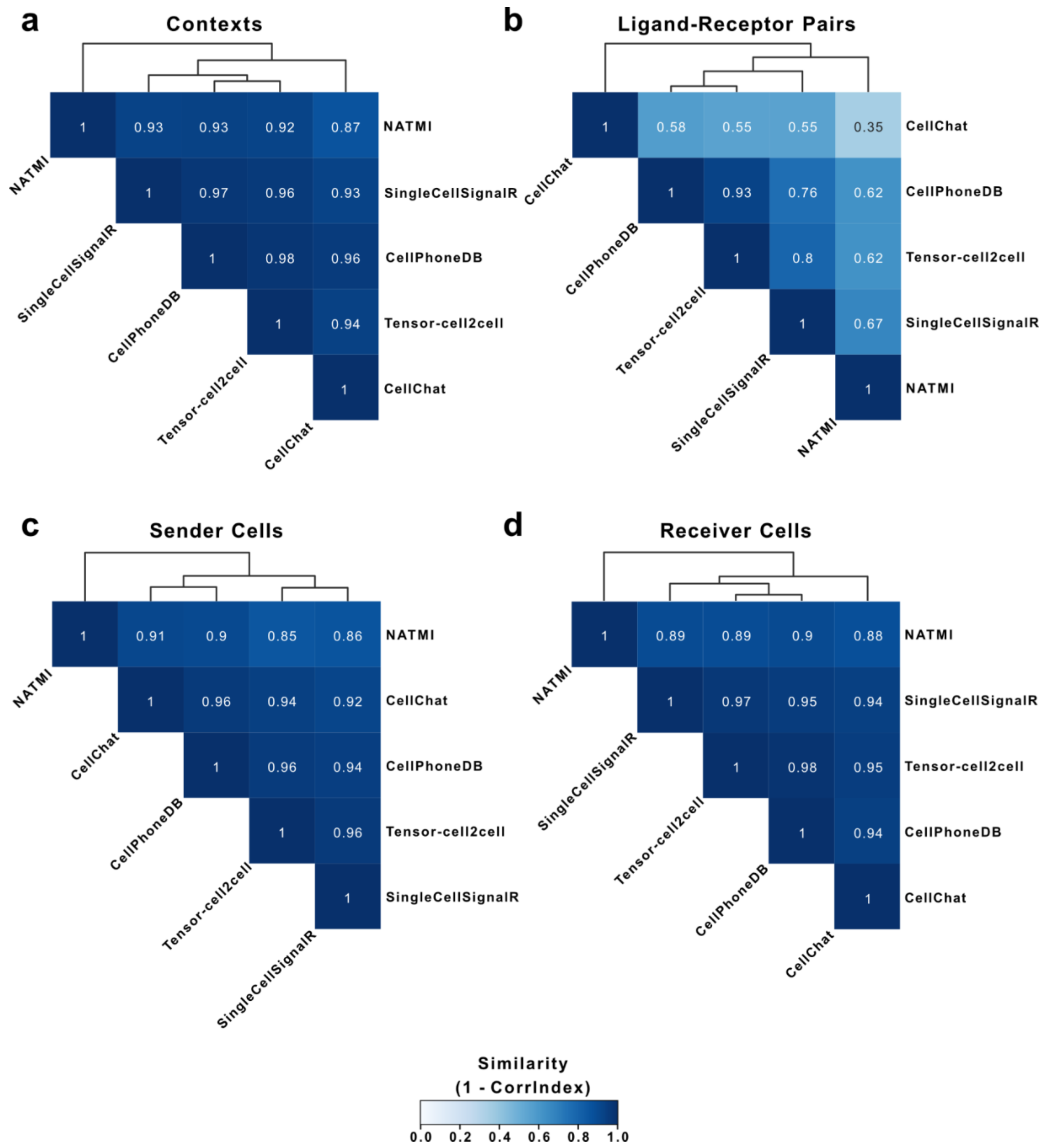


Figure 3.11: Evaluation of the main sources of dissimilarities when using different communication scores.

Similarity (1 - CorrIndex) between tensor decompositions performed on the same 4D-communication tensor for a single-cell dataset of BALF in patients with varying severities, but with the communication scores computed from different tools for inferring cell-cell communication. Here, the CorrIndex score is modified to consider each tensor dimension separately. Pairwise resultant similarities between all scoring methods for the (**a**) Context, (**b**) Ligand-Receptor Pair, (**c**) Sender Cell, (**d**) and Receiver Cell dimensions are hierarchically clustered and visualized as a heatmap. Source data can be found in the Code Ocean capsule.

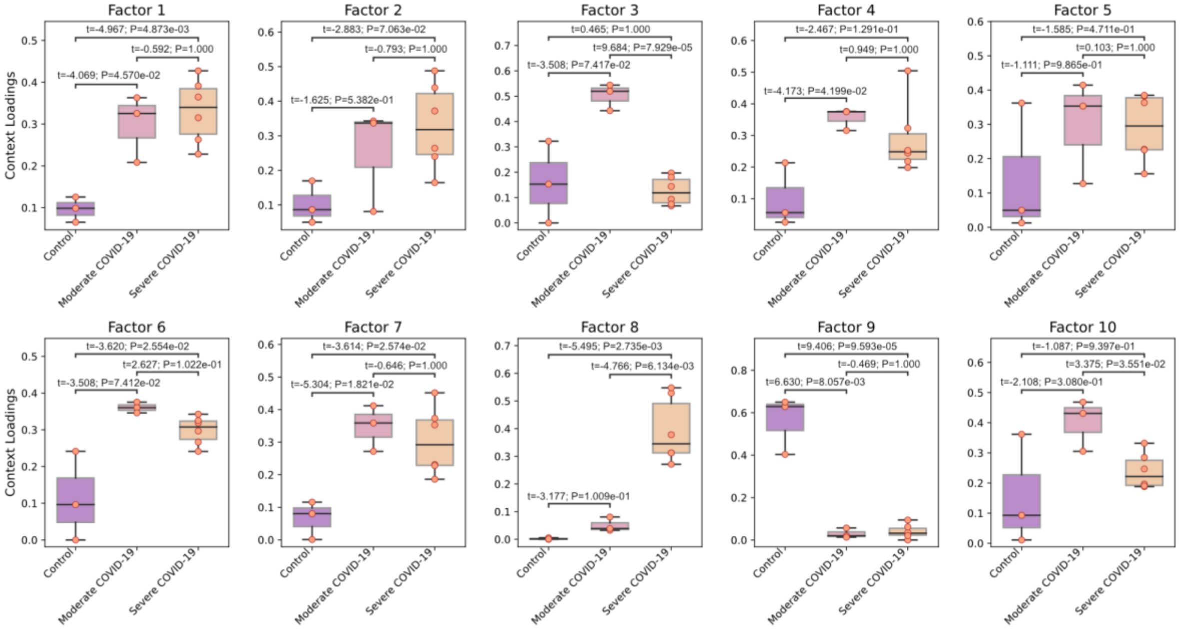
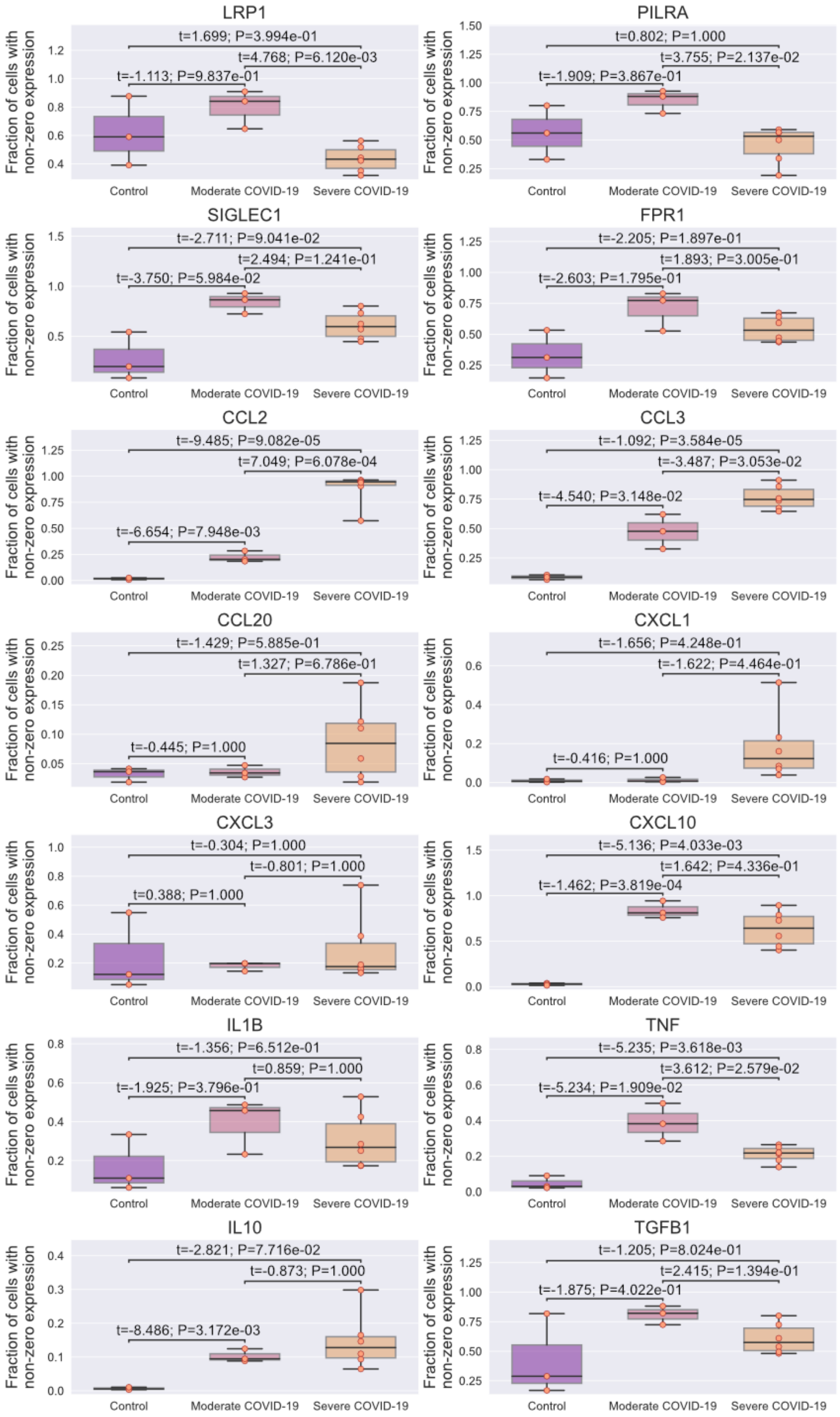


Figure 3.12: Boxplots of loadings for severities of COVID-19.

Boxplot representation for the different groups in the COVID-19 patients (as indicated in the x-axis; n=three healthy control patients, three patients with moderate disease, and six patients with severe disease). Each panel represents the sample loadings, grouped by disease condition, in each of the factors. Boxes represent the quartiles and whiskers show the rest of each distribution. A two-sided independent t-test was run to compare differences between the mean of the loadings of the groups, which was followed by a Bonferroni multiple test correction. For each pairwise comparison, the exact values of the test statistics (t) and the adjusted P-values (P) are shown. Context loadings employed in this Fig. are the same as in Fig. 4, and these data can be found in the Source Data file.

Figure 3.13: Gene expression associated with M1- and M2-like macrophages.
Gene expression of cytokines and/or receptors that are representative of M1- and M2-like phenotypes are shown for macrophages in each patient, grouped by COVID-19 severity (as indicated in the x-axis; n=three healthy control patients, three patients with moderate disease, and six patients with severe disease). Each subplot represents a different gene (as indicated in each title). In this case the gene expression values correspond to the fraction of cells with non-zero expression (y-axis) for the cluster of single cells annotated as macrophages. Boxes represent the quartiles and whiskers show the rest of each distribution. A two-sided independent t-test was run to compare differences between the mean of the loadings of the groups, which was followed by a Bonferroni multiple test correction. For each pairwise comparison, the exact values of the test statistics (t) and the adjusted P-values (P) are shown. Source data can be found in the Code Ocean capsule.


LRP1
t=1.699; P=3.994e-01
t=4.768; P=6.120e-03
t=-1.113; P=9.837e-01
PILRA
t=0.802; P=1.000
t=3.755; P=2.137e-02
t=-1.909; P=3.867e-01
SIGLEC1
t=-2.711; P=9.041e-02
t=2.494; P=1.241e-01
t=-3.750; P=5.984e-02
FPR1
t=-2.205; P=1.897e-01
t=1.893; P=3.005e-01
t=-2.603; P=1.795e-01
CCL2
t=-9.485; P=9.082e-05
t=7.049; P=6.078e-04
t=-6.654; P=7.948e-03
CCL3
t=-1.092; P=3.584e-05
t=-3.487; P=3.053e-02
t=-4.540; P=3.148e-02
CCL20
t=-1.429; P=5.885e-01
t=1.327; P=6.786e-01
t=-0.445; P=1.000
CXCL1
t=-1.656; P=4.248e-01
t=-1.622; P=4.464e-01
t=-0.416; P=1.000
CXCL3
t=-0.304; P=1.000
t=-0.801; P=1.000
t=0.388; P=1.000
CXCL10
t=-5.136; P=4.033e-03
t=1.642; P=4.336e-01
t=-1.462; P=3.819e-04
IL1B
t=-1.356; P=6.512e-01
t=0.859; P=1.000
t=-1.925; P=3.796e-01
TNF
t=-5.235; P=3.618e-03
t=3.612; P=2.579e-02
t=-5.234; P=1.909e-02
IL10
t=-2.821; P=7.716e-02
t=-0.873; P=1.000
t=-8.486; P=3.172e-03
TGFB1
t=-1.205; P=8.024e-01
t=2.415; P=1.394e-01
t=-1.875; P=4.022e-01
Fraction of cells with non-zero expression
Control
Moderate COVID-19
Severe COVID-19

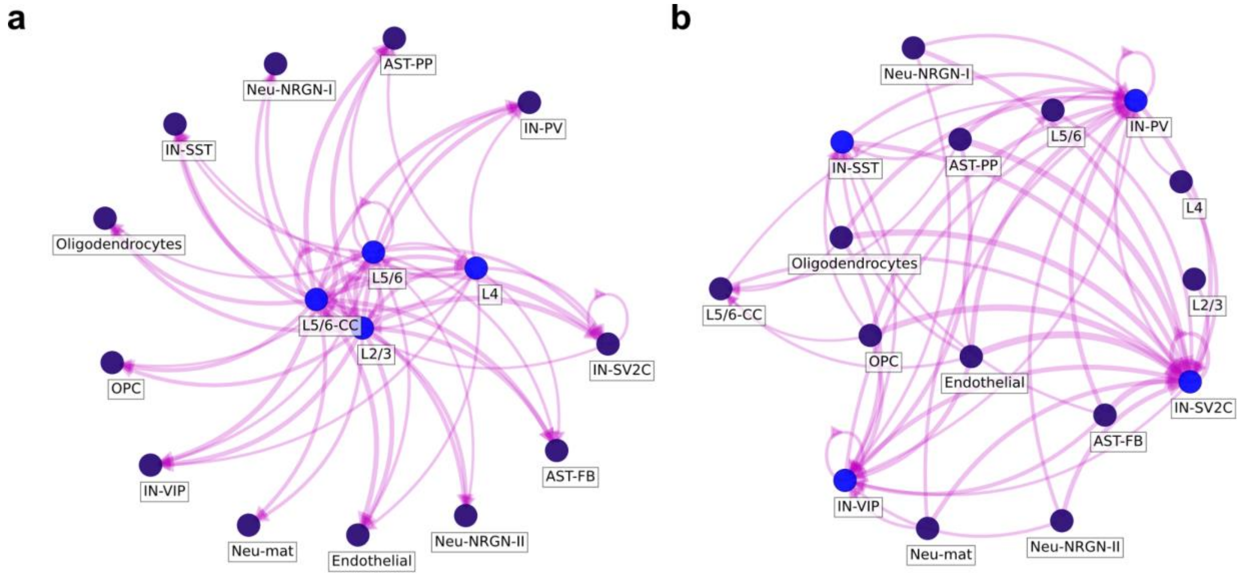


Figure 3.14: Factor-specific cell-cell communication networks that drive significant differences between ASD patients and controls.

Factor-specific cell-cell communication networks of (**a**) factor 3 and (**b**) factor 4 from the tensor decomposition of the ASD data set. These CCC networks representing the overall interactions between cells can be built for each of the factors by using the outer product between their respective sender-cell and receiver-cell normalized loadings (see ***Methods****)*. Resulting values represent edge weights. When building these networks, all cell types are connected to each other. To consider only biological meaningful interactions, we filter edges with a weight above a threshold of 0.075. In (**a**) and (**b**), nodes colored in blue represent those that are important as sender cells (factor 3) or receiver cells (factor 4). Edge widths are proportional to the edge weights. Source data can be found in the Code Ocean capsule.

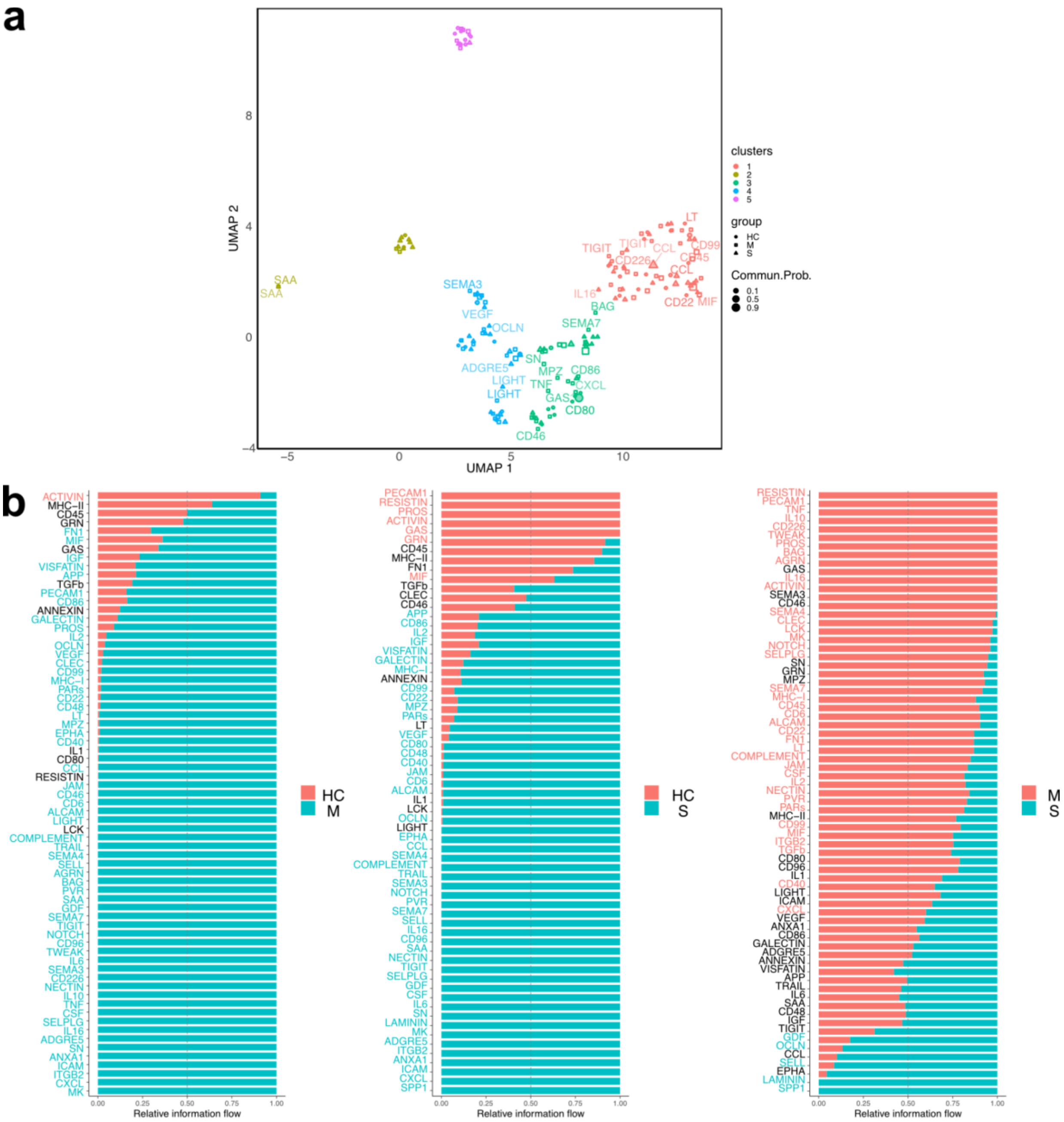
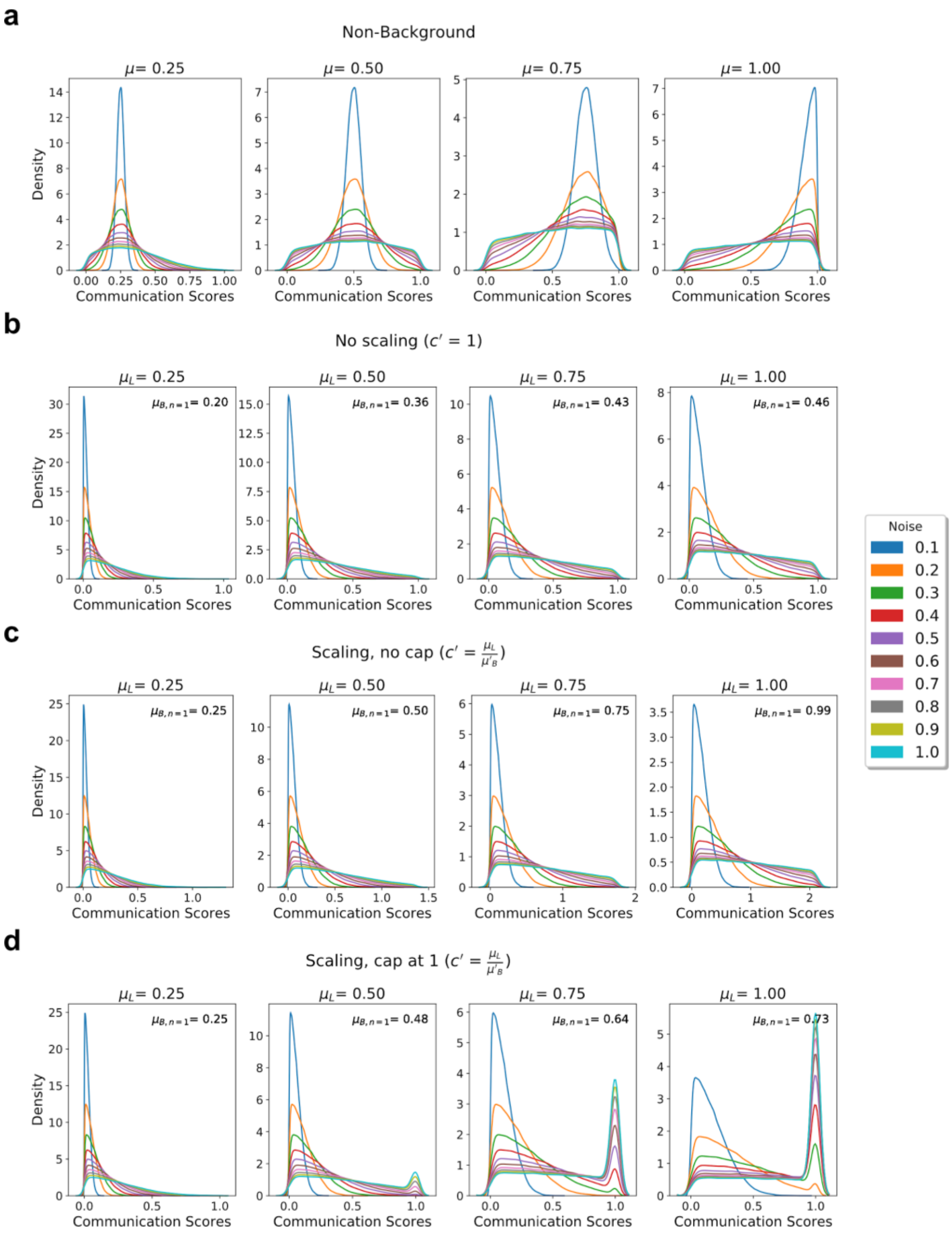


Figure 3.15: Downstream analyses of the BALF-COVID-19 dataset available to CellChat when considering multiple contexts.

For simplicity, sample expression matrices were aggregated by context (healthy control: *HC*, moderate COVID-19: *M*, and severe COVID-19: *S*) prior to running a CellChat functional analysis, as described in the ***Benchmarking of computational efficiency of tools*** section of the Methods. Analyses are conducted on the pairwise similarity between signaling pathways. (**a**) UMAP embeddings and clustering results computed from the pairwise similarity matrix between signaling pathways across all contexts. This manifold learning process summarizes multiple pairwise comparisons in an automated manner. **(b)** The ranked information flow (total communication probability between all pairs of cell groups) compared between each context for each signaling pathway. Source data are provided in the Source Data file.

Figure 3.16: Noise increases dispersion of communication scores.
(**a**) From left to right, the distribution of communication scores at increasing levels of expected value across noise. (**b**) From left to right, the distribution of communication scores at increasing maximal average value when noise is 1, without scaling the distribution, for an expected value of 0. (**c**) From left to right, the distribution of communication scores at increasing maximal average value when noise is 1, scaling the distributions such that the average value of the distribution when noise is 1 equals the maximal average value, for an expected value of 0. Scores are not capped to have a maximum value of 1. (**d**) From left to right, the distribution of communication scores at increasing maximal average value when noise is 1, scaling the distributions such that the average value of the distribution when noise is 1 equals the maximal average value, for an expected value of 0. Scores are capped to have a maximum value of 1. Panels b-d are annotated with the average value of the distribution when noise is 1. $\boldsymbol{\mu}$ represents the expected value of communication scores, $\boldsymbol{\mu}_{\mathbf{L}}$ is the desired maximum average value of the communication scores, $\boldsymbol{\mu}'_{\mathbf{B}}$ is a function of $\boldsymbol{\mu}_{\mathbf{L}}$ and **c'** is a scaling factor. All these parameters are described in more details in the Supplementary Notes (see Adding noise to simulations). Source data can be found in the Code Ocean capsule.



a
Non-Background
μ= 0.25
μ= 0.50
μ= 0.75
μ= 1.00
Density
0 2 4 6 8 10 12 14
0.00 0.25 0.50 0.75 1.00
0 1 2 3 4 5 6 7
0.0 0.5 1.0
0 1 2 3 4 5
0.0 0.5 1.0
0 1 2 3 4 5 6 7
0.0 0.5 1.0
Communication Scores
b
No scaling ($c' = 1$)
μ_L= 0.25
$\mu_{B,n=1}$= 0.20
μ_L= 0.50
$\mu_{B,n=1}$= 0.36
μ_L= 0.75
$\mu_{B,n=1}$= 0.43
μ_L= 1.00
$\mu_{B,n=1}$= 0.46
Density
0 5 10 15 20 25 30
0.0 2.5 5.0 7.5 10.0 12.5 15.0
0 2 4 6 8 10
0 2 4 6 8
0.0 0.5 1.0
Communication Scores
c
Scaling, no cap ($c' = \frac{\mu_L}{\mu'_B}$)
μ_L= 0.25
$\mu_{B,n=1}$= 0.25
μ_L= 0.50
$\mu_{B,n=1}$= 0.50
μ_L= 0.75
$\mu_{B,n=1}$= 0.75
μ_L= 1.00
$\mu_{B,n=1}$= 0.99
Density
0 5 10 15 20 25
0 2 4 6 8 10
0 1 2 3 4 5 6
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
0.0 0.5 1.0
0.0 0.5 1.0 1.5
0 1 2
0 1 2
Communication Scores
d
Scaling, cap at 1 ($c' = \frac{\mu_L}{\mu'_B}$)
μ_L= 0.25
$\mu_{B,n=1}$= 0.25
μ_L= 0.50
$\mu_{B,n=1}$= 0.48
μ_L= 0.75
$\mu_{B,n=1}$= 0.64
μ_L= 1.00
$\mu_{B,n=1}$= 0.73
Density
0 5 10 15 20 25
0 2 4 6 8 10
0 1 2 3 4 5 6
0 1 2 3 4 5
0.0 0.5 1.0
Communication Scores
Noise
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

3.5.14 Appendix N: Data availability

All input data used for the analyses in this work and the result-generated data are available online in a Code Ocean capsule (https://doi.org/10.24433/CO.0051950.v2). In particular, we used a single-cell atlas of COVID-19 in humans[68], previously deposited in the NCBI's Gene Expression Omnibus database under accession code GSE158055 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158055], a COVID-19 dataset of single-cell transcriptomes for BALF samples[32], previously deposited in the NCBI's Gene Expression Omnibus database under accession code GSE145926 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145926], and a single-nucleus ASD dataset previously deposited in the NCBI’s BioProject database under accession code PRJNA434002 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434002/]. The list of ligand-receptor interactions employed in our analyses corresponds to the database previously published with CellChat[10], and is available in a Compendium of Ligand-Receptor Pairs [https://github.com/LewisLabUCSD/Ligand-Receptor-Pairs/blob/master/Human/Human-2020-Jin-LR-pairs.csv] that we previously published[5]. The data generated in this study for the loadings resulting from the tensor decompositions of the simulated, COVID-19 and ASD datasets are available in the Source Data file. Source data that are not included in this file can be found and reproduced in the Code Ocean capsule.

3.5.15 Appendix O: Code availability

All the code used for the analyses in this work is available online in a Code Ocean capsule (https://doi.org/10.24433/CO.0051950.v2), which includes the exact version of all tools and software employed, and allows one to perform online a reproducible run of our analyses, outputting pertinent results. Tensor-cell2cell is implemented in our cell2cell suite[65], and its GitHub repository and full documentation can be found at http://lewislab.ucsd.edu/cell2cell/, which also includes comprehensive tutorials that go from raw UMI data to running Tensor-cell2cell, followed by downstream analyses using Tensor-cell2cell’s outputs. The code for benchmarking the computational efficiency should be run in a local computer, and is available in a GitHub repository (https://github.com/LewisLabUCSD/CCC-Benchmark).

3.5.16 Appendix P: Authors, Contributions, and Acknowledgements

Authors: Erick Armingol*, Hratch M. Baghdassarian*, Cameron Martino Araceli Perez-Lopez, Caitlin Aamodt, Rob Knight, Nathan E. Lewis

*contributed equally to work

E.A., H.M.B., and N.E.L. conceived the work. C.M. contributed important insights for creating Tensor-cell2cell. E.A. implemented Tensor-cell2cell and performed the analyses on the datasets of COVID-19 and ASD. H.M.B. designed and created the simulated 4D-communication tensor and performed the analyses on the simulated data. E.A., H.M.B., and C.M. performed benchmarking and statistical analyses. C.M. trained classifiers and compared Tensor-cell2cell to CellChat. H.B. performed benchmarking analyses using different external CCC tools. E.A. performed benchmarking analyses using different preprocessing and batch-correction methods. E.A. and H.M.B. developed downstream analyses. A.P.L. helped to interpret the COVID-19 results and researched literature. C.A. helped to interpret the ASD study case and researched literature. R.K. contributed to the benchmarking analyses. E.A. and H.M.B. wrote the paper and all authors carefully reviewed, discussed and edited the paper.

3.6 References

1. Hwang, S., Kim, S., Shin, H. & Lee, D. Context-dependent transcriptional regulations between signal transduction pathways. *BMC Bioinformatics* **12,** 19 (2011).

2. Shakiba, N., Jones, R. D., Weiss, R. & Del Vecchio, D. Context-aware synthetic biology by controller design: Engineering the mammalian cell. *Cell Syst* **12,** 561–592 (2021).

3. Rachlin, J., Cohen, D. D., Cantor, C. & Kasif, S. Biological context networks: a mosaic view of the interactome. *Mol. Syst. Biol.* **2,** 66 (2006).

4. Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M. J., Blüthgen, N. & Saez-Rodriguez, J. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9,** 20 (2018).

5. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* 1–18 (2020).

6. Griffiths, J. I., Wallet, P., Pflieger, L. T., Stenehjem, D., Liu, X., Cosgrove, P. A., Leggett, N. A., McQuerry, J. A., Shrestha, G., Rossetti, M., Sunga, G., Moos, P. J., Adler, F. R., Chang, J. T., Sharma, S. & Bild, A. H. Circulating immune cell phenotype dynamics reflect the strength of tumor–immune cell interactions in patients during immunotherapy. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 16072–16082 (2020).

7. Omberg, L., Golub, G. H. & Alter, O. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 18371–18376 (2007).

8. Cillo, A. R., Kürten, C. H. L., Tabib, T., Qi, Z., Onkar, S., Wang, T., Liu, A., Duvvuri, U., Kim, S., Soose, R. J., Oesterreich, S., Chen, W., Lafyatis, R., Bruno, T. C., Ferris, R. L. & Vignali, D. A. A. Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* **52,** 183–199.e9 Preprint at https://doi.org/10.1016/j.immuni.2019.11.014 (2020)

9. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11,** 1–11 (2020).

10. Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V. & Nie, Q. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12,** 1088 (2021).

11. Raredon, M. S. B., Yang, J., Garritano, J., Wang, M., Kushnir, D., Schupp, J. C., Adams, T. S., Greaney, A. M., Leiby, K. L., Kaminski, N., Kluger, Y., Levchenko, A. & Niklason, L. E. Computation and visualization of cell-cell signaling topologies in single-cell systems data using Connectome. *Sci. Rep.* **12,** 4187 (2022).

12. Williams, A. H., Kim, T. H., Wang, F., Vyas, S., Ryu, S. I., Shenoy, K. V., Schnitzer, M., Kolda, T. G. & Ganguli, S. Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98,** 1099–1115.e8 (2018).

13. Stein-O’Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., Xu, Y. & Fertig, E. J. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34,** 790–805 (2018).

14. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.* **20,** 269 (2019).

15. Martino, C., Shenhav, L., Marotz, C. A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., Morton, J. T., Jiang, L., Dominguez-Bello, M. G., Swafford, A. D., Halperin, E. & Knight, R. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39,** 165–168 (2021).

16. Anandkumar, A., Jain, P., Shi, Y. & Niranjan, U. N. Tensor vs. Matrix Methods: Robust Tensor Decomposition under Block Sparse Perturbations. in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (eds. Gretton, A. & Robert, C. C.) **51,** 268–276 (PMLR, 2016).

17. Rabanser, S., Shchur, O. & Günnemann, S. Introduction to Tensor Decompositions and their Applications in Machine Learning. *arXiv [stat.ML]* (2017). at <http://arxiv.org/abs/1711.10781>

18. Friedlander, M. P. & Hatz, K. Computing non-negative tensor factorizations. *Optim. Methods Softw.* **23,** 631–647 (2008).

19. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* (2020). doi:10.1038/s41596-020-0292-x

20. Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M. & Colinge, J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **48,** e55 (2020).

21. Sobhani, E., Comon, P., Jutten, C. & Babaie-Zadeh, M. CorrIndex: A permutation invariant performance index. *Signal Processing* **195,** 108457 (2022).

22. Dimitrov, D., Türei, D., Boys, C., Nagai, J. S., Ramirez Flores, R. O., Kim, H., Szalai, B., Costa, I. G., Dugourd, A., Valdeolivas, A. & Saez-Rodriguez, J. Comparison of Resources and Methods to infer Cell-Cell Communication from Single-cell RNA Data. *bioRxiv* 2021.05.21.445160 (2021). doi:10.1101/2021.05.21.445160

23. Booeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by log(x + 1)* or log(1 + x). *Bioinformatics* (2021). doi:10.1093/bioinformatics/btab085

24. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

25. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37,** 685–691 (2019).

26. Baccin, C., Al-Sabah, J., Velten, L., Helbling, P. M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L. M., Trumpp, A. & Haas, S. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat. Cell Biol.* **22,** 38–48 (2020).

27. Noël, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C., Kieffer, Y., Mechta-Grigoriou, F. & Soumelis, V. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.* **12,** 1089 (2021).

28. Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., Wang, M., Ajani, J., Futreal, A. & Wang, L. iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *Cancer Biology* (2019). doi:10.1101/507871

29. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* (2019). doi:10.1038/s41592-019-0667-5

30. Lagger, C., Ursu, E., Equey, A., Avelar, R. A., Pisco, A. O., Tacutu, R. & de Magalhães, J. P. scAgeCom: a murine atlas of age-related changes in intercellular communication inferred with the package scDiffCom. *bioRxiv* 2021.08.13.456238 (2021). doi:10.1101/2021.08.13.456238

31. Tsuyuzaki, K., Ishii, M. & Nikaido, I. Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. *bioRxiv* 566182 (2019). doi:10.1101/566182

32. Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., Liu, L., Amit, I., Zhang, S. & Zhang, Z. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26,** 842–844 (2020).

33. Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M. T., Kazmierski, J., Timmermann, B., Twardziok, S., Schneider, S., Machleidt, F., Müller-Redetzky, H., Maier, M., Krannich, A., Schmidt, S., Balzer, F., Liebig, J., Loske, J., Suttorp, N., Eils, J., Ishaque, N., Liebert, U. G., von Kalle, C., Hocke, A., Witzenrath, M., Goffinet, C., Drosten, C., Laudi, S., Lehmann, I., Conrad, C., Sander, L.-E. & Eils, R. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* (2020). doi:10.1038/s41587-020-0602-4

34. Schmitt, T. L., Steiner, E., Klingler, P., Lassmann, H. & Grubeck-Loebenstein, B. Thyroid epithelial cells produce large amounts of the Alzheimer beta-amyloid precursor protein (APP) and generate potentially amyloidogenic APP fragments. *J. Clin. Endocrinol. Metab.* **80,** 3513–3519 (1995).

35. Puig, K. L., Manocha, G. D. & Combs, C. K. Amyloid precursor protein mediated changes in intestinal epithelial phenotype in vitro. *PLoS One* **10,** e0119534 (2015).

36. Zemans, R. L., Colgan, S. P. & Downey, G. P. Transepithelial migration of neutrophils: mechanisms and implications for acute lung injury. *Am. J. Respir. Cell Mol. Biol.* **40,** 519–535 (2009).

37. Schenkel, A. R., Mamdouh, Z., Chen, X., Liebman, R. M. & Muller, W. A. CD99 plays a major role in the migration of monocytes through endothelial junctions. *Nat. Immunol.* **3,** 143–150 (2002).

38. Pasello, M., Manara, M. C. & Scotlandi, K. CD99 at the crossroads of physiology and pathology. *J. Cell Commun. Signal.* **12,** 55–68 (2018).

39. Sanino, G., Bosco, M. & Terrazzano, G. Physiology of Midkine and Its Potential Pathophysiological Role in COVID-19. *Front. Physiol.* **11,** 616552 (2020).

40. Farr, L., Ghosh, S. & Moonah, S. Role of MIF Cytokine/CD74 Receptor Pathway in Protecting Against Injury and Promoting Repair. *Front. Immunol.* **11,** 1273 (2020).

41. Weckbach, L. T., Muramatsu, T. & Walzog, B. Midkine in inflammation. *ScientificWorldJournal* **11,** 2491–2505 (2011).

42. Xia, J., Swiercz, J. M., Bañón-Rodríguez, I., Matković, I., Federico, G., Sun, T., Franz, T., Brakebusch, C. H., Kumanogoh, A., Friedel, R. H., Martín-Belmonte, F., Gröne, H.-J., Offermanns, S. & Worzfeld, T. Semaphorin-Plexin Signaling Controls Mitotic Spindle Orientation during Epithelial Morphogenesis and Repair. *Dev. Cell* **33,** 299–313 (2015).

43. Nikaido, T., Tanino, Y., Wang, X., Sato, S., Misa, K., Fukuhara, N., Sato, Y., Fukuhara, A., Uematsu, M., Suzuki, Y., Kojima, T., Tanino, M., Endo, Y., Tsuchiya, K., Kawamura, I., Frevert, C. W. & Munakata, M. Serum Syndecan-4 as a Possible Biomarker in Patients With Acute Pneumonia. *J. Infect. Dis.* **212,** 1500–1508 (2015).

44. Azari, B. M., Marmur, J. D., Salifu, M. O., Ehrlich, Y. H., Kornecki, E. & Babinska, A. Transcription and translation of human F11R gene are required for an initial step of atherogenesis induced by inflammatory cytokines. *J. Transl. Med.* **9,** 98 (2011).

45. Clark, I. C., Gutiérrez-Vázquez, C., Wheeler, M. A., Li, Z., Rothhammer, V., Linnerbauer, M., Sanmarco, L. M., Guo, L., Blain, M., Zandee, S. E. J., Chao, C.-C., Batterman, K. V., Schwabenland, M., Lotfy, P., Tejeda-Velarde, A., Hewson, P., Manganeli Polonio, C., Shultis, M. W., Salem, Y., Tjon, E. C., Fonseca-Castro, P. H., Borucki, D. M., Alves de Lima, K., Plasencia, A., Abate, A. R., Rosene, D. L., Hodgetts, K. J., Prinz, M., Antel, J. P., Prat, A. & Quintana, F. J. Barcoded viral tracing of single-cell interactions in central nervous system inflammation. *Science* **372,** (2021).

46. Zhang, F., Mears, J. R., Shakib, L., Beynor, J. I., Shanaj, S., Korsunsky, I., Nathan, A., Donlin, L. T. & Raychaudhuri, S. IFN- γ and TNF- α drive a CXCL10 + CCL2 + macrophage phenotype expanded in severe COVID-19 and other diseases with tissue inflammation. *bioRxiv* (2020). doi:10.1101/2020.08.05.238360

47. Kohyama, M., Matsuoka, S., Shida, K., Sugihara, F., Aoshi, T., Kishida, K., Ishii, K. J. & Arase, H. Monocyte infiltration into obese and fibrilized tissues is regulated by PILRα. *Eur. J. Immunol.* **46,** 1214–1223 (2016).

48. Saheb Sharif-Askari, N., Saheb Sharif-Askari, F., Mdkhana, B., Al Heialy, S., Alsafar, H. S., Hamoudi, R., Hamid, Q. & Halwani, R. Enhanced expression of immune checkpoint receptors during SARS-CoV-2 viral infection. *Mol Ther Methods Clin Dev* **20,** 109–121 (2021).

49. Martinez, F. O., Combes, T. W., Orsenigo, F. & Gordon, S. Monocyte activation in systemic Covid-19 infection: Assay and rationale. *EBioMedicine* **59,** 102964 (2020).

50. Ocaña-Guzman, R., Torre-Bouscoulet, L. & Sada-Ovalle, I. TIM-3 Regulates Distinct Functions in Macrophages. *Front. Immunol.* **7,** 229 (2016).

51. Grant, R. A., Morales-Nebreda, L., Markov, N. S., Swaminathan, S., Querrey, M., Guzman, E. R., Abbott, D. A., Donnelly, H. K., Donayre, A., Goldberg, I. A., Klug, Z. M., Borkowski, N., Lu, Z., Kihshen, H., Politanska, Y., Sichizya, L., Kang, M., Shilatifard, A., Qi, C., Lomasney, J. W., Argento, A. C., Kruser, J. M., Malsin, E. S., Pickens, C. O., Smith, S. B., Walter, J. M., Pawlowski, A. E., Schneider, D., Nannapaneni, P., Abdala-Valencia, H., Bharat, A., Gottardi, C. J., Budinger, G. R. S., Misharin, A. V., Singer, B. D., Wunderink, R. G. & NU SCRIPT Study Investigators. Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature* **590,** 635–641 (2021).

52. Matsuyama, T., Kubli, S. P., Yoshinaga, S. K., Pfeffer, K. & Mak, T. W. An aberrant STAT pathway is central to COVID-19. *Cell Death Differ.* **27,** 3209–3225 (2020).

53. Florez-Sampedro, L., Soto-Gamez, A., Poelarends, G. J. & Melgert, B. N. The role of MIF in chronic lung diseases: looking beyond inflammation. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **318,** L1183–L1197 (2020).

54. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22,** 345–361 (2016).

55. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H. & Kriegstein, A. R. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364,** 685–689 (2019).

56. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

57. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

58. Astorkia, M., Lachman, H. M. & Zheng, D. Characterization of Cell-cell Communication in Autistic Brains with Single Cell Transcriptomes. *bioRxiv* 2021.10.15.464577 (2021). doi:10.1101/2021.10.15.464577

59. Avraham, R. & Yarden, Y. Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Biol.* **12,** 104–117 (2011).

60. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell-cell communication through single-cell transcriptomics. *Current Opinion in Systems Biology* (2021). doi:10.1016/j.coisb.2021.03.007

61. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15,** e8746 (2019).

62. Abbasy, S., Shahraki, F., Haghighatfard, A., Qazvini, M. G., Rafiei, S. T., Noshadirad, E., Farhadi, M., Rezvani Asl, H., Shiryazdi, A. A., Ghamari, R., Tabrizi, Z., Mehrfard, R., Esmaili Kakroudi, F., Azarnoosh, M., Younesi, F., Parsamehr, N., Garaei, N., Abyari, S., Salehi, M., Gholami, M., Zolfaghari, P., Bagheri, S. M., Pourmehrabi, M., Rastegarimogaddam, E., Nobakht, E., Nobakht, E. & Partovi, R. Neuregulin1 types mRNA level changes in autism spectrum disorder, and is associated with deficit in executive functions. *EBioMedicine* **37,** 483–488 (2018).

63. Gazestani, V. H., Pramparo, T., Nalabolu, S., Kellman, B. P., Murray, S., Lopez, L., Pierce, K., Courchesne, E. & Lewis, N. E. A perturbed gene network containing PI3K-AKT, RAS-ERK and WNT-β-catenin pathways in leukocytes is linked to ASD genetics and symptom severity. *Nat. Neurosci.* **22,** 1624–1634 (2019).

64. Tanevski, J., Flores, R. O. R., Gabor, A., Schapiro, D. & Saez-Rodriguez, J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol.* **23,** 97 (2022).

65. Armingol, E., Ghaddar, A., Joshi, C. J., Baghdassarian, H., Shamie, I., Chan, J., Her, H.-L., O'Rourke, E. J. & Lewis, N. E. Inferring a spatial code of cell-cell interactions across a whole animal body. *bioRxiv* 2020.11.22.392217 (2022). doi:10.1101/2020.11.22.392217

66. Wang, S., Karikomi, M., MacLean, A. L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* **47,** e66 (2019).

67. Mishra, V., Re, D. B., Le Verche, V., Alvarez, M. J., Vasciaveo, A., Jacquier, A., Doulias, P.-T., Greco, T. M., Nizzardo, M., Papadimitriou, D., Nagata, T., Rinchetti, P., Perez-Torres, E. J., Politi, K. A., Ikiz, B., Clare, K., Than, M. E., Corti, S., Ischiropoulos, H., Lotti, F., Califano, A. & Przedborski, S. Systematic elucidation of neuron-astrocyte interaction in models of amyotrophic lateral sclerosis using multi-modal integrated bioinformatics workflow. *Nat. Commun.* **11,** 5579 (2020).

68. Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., Yang, Y., He, J., Ma, W., He, J., Wang, P., Cao, Q., Chen, F., Chen, Y., Cheng, X., Deng, G., Deng, X., Ding, W., Feng, Y., Gan, R., Guo, C., Guo, W., He, S., Jiang, C., Liang, J., Li, Y.-M., Lin, J., Ling, Y., Liu, H., Liu, J., Liu, N., Liu, S.-Q., Luo, M., Ma, Q., Song, Q., Sun, W., Wang, G., Wang, F., Wang, Y., Wen, X., Wu, Q., Xu, G., Xie, X., Xiong, X., Xing, X., Xu, H., Yin, C., Yu, D., Yu, K., Yuan, J., Zhang, B., Zhang, P., Zhang, T., Zhao, J., Zhao, P., Zhou, J., Zhou, W., Zhong, S., Zhong, X., Zhang, S., Zhu, L., Zhu, P., Zou, B., Zou, J., Zuo, Z., Bai, F., Huang, X., Zhou, P., Jiang, Q., Huang, Z., Bei, J.-X., Wei, L., Bian, X.-W., Liu, X., Cheng, T., Li, X., Zhao, P., Wang, F.-S., Wang, H., Su, B., Zhang, Z., Qu, K., Wang, X., Chen, J., Jin, R. & Zhang, Z. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184,** 1895–1913.e19 (2021).

69. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210 (2002).

70. Carroll, J. D. & Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. *Psychometrika* **35,** 283–319 (1970).

71. Harshman, R. A. & Others. Foundations of the PARAFAC procedure: Models and conditions for an' explanatory' multimodal factor analysis. (1970). at <https://www.psychology.uwo.ca/faculty/harshman/wpppfac0.pdf>

72. Anandkumar, A., Ge, R. & Janzamin, M. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv [cs.LG]* (2014). at <http://arxiv.org/abs/1402.5180>

73. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *arXiv [cs.LG]* (2016). at <http://arxiv.org/abs/1610.09555>

74. Farris, F. A. The Gini Index and Measures of Inequality. *Am. Math. Mon.* **117,** 851–864 (2010).

75. Schieber, T. A., Carpi, L., Díaz-Guilera, A., Pardalos, P. M., Masoller, C. & Ravetti, M. G. Quantification of network structural dissimilarities. *Nat. Commun.* **8,** 13928 (2017).

76. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

77. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Others. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12,** 2825–2830 (2011).

78. Dorschner, R. A., Lee, J., Cohen, O., Costantini, T., Baird, A. & Eliceiri, B. P. ECRG4 regulates neutrophil recruitment and CD44 expression during the inflammatory response to injury. *Sci Adv* **6,** eaay0518 (2020).

79. Kawana, H., Karaki, H., Higashi, M., Miyazaki, M., Hilberg, F., Kitagawa, M. & Harigaya, K. CD44 suppresses TLR-mediated inflammation. *J. Immunol.* **180,** 4235–4245 (2008).

80. Teder, P., Vandivier, R. W., Jiang, D., Liang, J., Cohn, L., Puré, E., Henson, P. M. & Noble, P. W. Resolution of lung inflammation by CD44. *Science* **296,** 155–158 (2002).

81. Kang, C. K., Han, G.-C., Kim, M., Kim, G., Shin, H. M., Song, K.-H., Choe, P. G., Park, W. B., Kim, E. S., Kim, H. B., Kim, N.-J., Kim, H.-R. & Oh, M.-D. Aberrant hyperactivation of cytotoxic T-cell as a potential determinant of COVID-19 severity. *Int. J. Infect. Dis.* **97,** 313–321 (2020).

82. Jiang, Y., Wei, X., Guan, J., Qin, S., Wang, Z., Lu, H., Qian, J., Wu, L., Chen, Y., Chen, Y. & Lin, X. COVID-19 pneumonia: CD8+ T and NK cells are decreased in number but compensatory increased in cytotoxic potential. *Clin. Immunol.* **218,** 108516 (2020).

83. Xiao, M., Noman, M. Z., Menard, L., Chevigne, A., Szpakowska, M., Bosseler, M., Ollert, M., Berchem, G. & Janji, B. Driving Cytotoxic Natural Killer Cells into Melanoma: If CCL5 Plays the Music, Autophagy Calls the Shots. *Crit. Rev. Oncog.* **23,** 321–332 (2018).

84. Schuette, V., Embgenbroich, M., Ulas, T., Welz, M., Schulte-Schrepping, J., Draffehn, A. M., Quast, T., Koch, K., Nehring, M., König, J., Zweynert, A., Harms, F. L., Steiner, N., Limmer, A., Förster, I., Berberich-Siebelt, F., Knolle, P. A., Wohlleber, D., Kolanus, W., Beyer, M., Schultze, J. L. & Burgdorf, S. Mannose receptor induces T-cell tolerance via inhibition of CD45 and up-regulation of CTLA-4. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 10649–10654 (2016).

85. Pata, S., Otáhal, P., Brdička, T., Laopajon, W., Mahasongkram, K. & Kasinrerk, W. Association of CD99 short and long forms with MHC class I, MHC class II and tetraspanin CD81 and recruitment into immunological synapses. *BMC Res. Notes* **4,** 293 (2011).

86. Lim, T. S., Goh, J. K. H., Mortellaro, A., Lim, C. T., Hämmerling, G. J. & Ricciardi-Castagnoli, P. CD80 and CD86 differentially regulate mechanical interactions of T-cells with antigen-presenting dendritic cells and B-cells. *PLoS One* **7,** e45185 (2012).

87. Pribila, J. T., Quale, A. C., Mueller, K. L. & Shimizu, Y. Integrins and T cell-mediated immunity. *Annu. Rev. Immunol.* **22,** 157–180 (2004).

88. Lebedeva, T., Dustin, M. L. & Sykulev, Y. ICAM-1 co-stimulates target cells to facilitate antigen presentation. *Curr. Opin. Immunol.* **17,** 251–258 (2005).

89. Jafarzadeh, A., Chauhan, P., Saha, B., Jafarzadeh, S. & Nemati, M. Contribution of monocytes and macrophages to the local tissue inflammation and cytokine storm in COVID-19: Lessons from SARS and MERS, and potential therapeutic interventions. *Life Sci.* **257,** 118102 (2020).

90. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* (2021). doi:10.1038/s41586-021-03767-x

91. Courchesne, E. & Pierce, K. Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. *Curr. Opin. Neurobiol.* **15,** 225–230 (2005).

92. Geschwind, D. H. & Levitt, P. Autism spectrum disorders: developmental disconnection syndromes. *Curr. Opin. Neurobiol.* **17,** 103–111 (2007).

93. Courchesne, E., Pramparo, T., Gazestani, V. H., Lombardo, M. V., Pierce, K. & Lewis, N. E. The ASD Living Biology: from cell proliferation to clinical phenotype. *Mol. Psychiatry* **24,** 88–107 (2019).

94. Anton, E. S., Ghashghaei, H. T., Weber, J. L., McCann, C., Fischer, T. M., Cheung, I. D., Gassmann, M., Messing, A., Klein, R., Schwab, M. H., Lloyd, K. C. K. & Lai, C. Receptor tyrosine kinase ErbB4 modulates neuroblast migration and placement in the adult forebrain. *Nat. Neurosci.* **7,** 1319–1328 (2004).

95. Conover, J. C., Doetsch, F., Garcia-Verdugo, J. M., Gale, N. W., Yancopoulos, G. D. & Alvarez-Buylla, A. Disruption of Eph/ephrin signaling affects migration and proliferation in the adult subventricular zone. *Nat. Neurosci.* **3,** 1091–1097 (2000).

96. López-Bendito, G., Cautinat, A., Sánchez, J. A., Bielle, F., Flames, N., Garratt, A. N., Talmage, D. A., Role, L. W., Charnay, P., Marín, O. & Garel, S. Tangential neuronal migration controls axon guidance: a role for neuregulin-1 in thalamocortical axon navigation. *Cell* **125,** 127–142 (2006).

97. Trakoshis, S., Martínez-Cañada, P., Rocchi, F., Canella, C., You, W., Chakrabarti, B., Ruigrok, A. N., Bullmore, E. T., Suckling, J., Markicevic, M., Zerbi, V., MRC AIMS Consortium, Baron-Cohen, S., Gozzi, A., Lai, M.-C., Panzeri, S. & Lombardo, M. V. Intrinsic excitation-inhibition imbalance affects medial prefrontal cortex differently in autistic men versus women. *Elife* **9,** (2020).

98. Courchesne, E., Gazestani, V. H. & Lewis, N. E. Prenatal Origins of ASD: The When, What, and How of ASD Development. *Trends Neurosci.* **43,** 326–342 (2020).

99. Tang, C., Wang, M., Wang, P., Wang, L., Wu, Q. & Guo, W. Neural Stem Cells Behave as a Functional Niche for the Maturation of Newborn Neurons through the Secretion of PTN. *Neuron* **101,** 32–44.e6 (2019).

100. Marchetto, M. C. N., Carromeu, C., Acab, A., Yu, D., Yeo, G. W., Mu, Y., Chen, G., Gage, F. H. & Muotri, A. R. A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. *Cell* **143,** 527–539 (2010).

101. Shcheglovitov, A., Shcheglovitova, O., Yazawa, M., Portmann, T., Shu, R., Sebastiano, V., Krawisz, A., Froehlich, W., Bernstein, J. A., Hallmayer, J. F. & Dolmetsch, R. E. SHANK3 and IGF1 restore synaptic deficits in neurons from 22q13 deletion syndrome patients. *Nature* **503,** 267–271 (2013).

102. Marchetto, M. C., Belinson, H., Tian, Y., Freitas, B. C., Fu, C., Vadodaria, K., Beltrao-Braga, P., Trujillo, C. A., Mendes, A. P. D., Padmanabhan, K., Nunez, Y., Ou, J., Ghosh, H., Wright, R., Brennand, K., Pierce, K., Eichenfield, L., Pramparo, T., Eyler, L., Barnes, C. C.,

Courchesne, E., Geschwind, D. H., Gage, F. H., Wynshaw-Boris, A. & Muotri, A. R. Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. *Mol. Psychiatry* **22,** 820–835 (2017).

103. Kolevzon, A., Bush, L., Wang, A. T., Halpern, D., Frank, Y., Grodberg, D., Rapaport, R., Tavassoli, T., Chaplin, W., Soorya, L. & Buxbaum, J. D. A pilot controlled trial of insulin-like growth factor-1 in children with Phelan-McDermid syndrome. *Mol. Autism* **5,** 54 (2014).

104. Pence, B. D. Growth differentiation factor-15 in immunity and aging. *Front. Aging* **3,** (2022).

105. Linfield, D. T., Raduka, A., Aghapour, M. & Rezaee, F. Airway tight junctions as targets of viral infections. *Tissue Barriers* **9,** 1883965 (2021).

106. Lou, O., Alcaide, P., Luscinskas, F. W. & Muller, W. A. CD99 is a key mediator of the transendothelial migration of neutrophils. *J. Immunol.* **178,** 1136–1143 (2007).

107. Bleilevens, C., Soppert, J., Hoffmann, A., Breuer, T., Bernhagen, J., Martin, L., Stiehler, L., Marx, G., Dreher, M., Stoppe, C. & Simon, T.-P. Macrophage Migration Inhibitory Factor (MIF) Plasma Concentration in Critically Ill COVID-19 Patients: A Prospective Observational Study. *Diagnostics (Basel)* **11,** (2021).

108. Marsh, L. M., Cakarova, L., Kwapiszewska, G., von Wulffen, W., Herold, S., Seeger, W. & Lohmeyer, J. Surface expression of CD74 by type II alveolar epithelial cells: a potential mechanism for macrophage migration inhibitory factor-induced epithelial repair. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **296,** L442–52 (2009).

109. Bruchez, A., Sha, K., Johnson, J., Chen, L., Stefani, C., McConnell, H., Gaucherand, L., Prins, R., Matreyek, K. A., Hume, A. J., Mühlberger, E., Schmidt, E. V., Olinger, G. G., Stuart, L. M. & Lacy-Hulbert, A. MHC class II transactivator CIITA induces cell resistance to Ebola virus and SARS-like coronaviruses. *Science* **370,** 241–247 (2020).

110. Weckbach, L. T., Gola, A., Winkelmann, M., Jakob, S. M., Groesser, L., Borgolte, J., Pogoda, F., Pick, R., Pruenster, M., Müller-Höcker, J., Deindl, E., Sperandio, M. & Walzog, B. The cytokine midkine supports neutrophil trafficking during acute inflammation by promoting adhesion via β2 integrins (CD11/CD18). *Blood* **123,** 1887–1896 (2014).

111. Witherden, D. A., Watanabe, M., Garijo, O., Rieder, S. E., Sarkisyan, G., Cronin, S. J. F., Verdino, P., Wilson, I. A., Kumanogoh, A., Kikutani, H., Teyton, L., Fischer, W. H. & Havran, W. L. The CD100 receptor interacts with its plexin B2 ligand to regulate epidermal γδ T cell function. *Immunity* **37,** 314–325 (2012).

112. Shanks, K., Nkyimbeng-Takwi, E. H., Smith, E., Lipsky, M. M., DeTolla, L. J., Scott, D. W., Keegan, A. D. & Chapoval, S. P. Neuroimmune semaphorin 4D is necessary for optimal lung allergic inflammation. *Mol. Immunol.* **56,** 480–487 (2013).

113. Carvalheiro, T., Affandi, A. J., Malvar-Fernández, B., Dullemond, I., Cossu, M., Ottria, A., Mertens, J. S., Giovannone, B., Bonte-Mineur, F., Kok, M. R., Marut, W., Reedquist, K. A., Radstake, T. R. & García, S. Induction of Inflammation and Fibrosis by Semaphorin 4A in Systemic Sclerosis. *Arthritis Rheumatol* **71,** 1711–1722 (2019).

114. Hudák, A., Letoha, A., Szilák, L. & Letoha, T. Contribution of Syndecans to the Cellular Entry of SARS-CoV-2. *Int. J. Mol. Sci.* **22,** (2021).

115. Jans, J., Unger, W. W. J., Vissers, M., Ahout, I. M. L., Schreurs, I., Wickenhagen, A., de Groot, R., de Jonge, M. I. & Ferwerda, G. Siglec-1 inhibits RSV-induced interferon gamma production by adult T cells in contrast to newborn T cells. *Eur. J. Immunol.* **48,** 621–631 (2018).

116. Lee, S., Lee, H.-C., Kwon, Y.-W., Lee, S. E., Cho, Y., Kim, J., Lee, S., Kim, J.-Y., Lee, J., Yang, H.-M., Mook-Jung, I., Nam, K.-Y., Chung, J., Lazar, M. A. & Kim, H.-S. Adenylyl cyclase-associated protein 1 is a receptor for human resistin and mediates inflammatory actions of human monocytes. *Cell Metab.* **19,** 484–497 (2014).

117. Meizlish, M. L., Pine, A. B., Bishai, J. D., Goshua, G., Nadelmann, E. R., Simonov, M., Chang, C.-H., Zhang, H., Shallow, M., Bahel, P., Owusu, K., Yamamoto, Y., Arora, T., Atri, D. S., Patel, A., Gbyli, R., Kwan, J., Won, C. H., Dela Cruz, C., Price, C., Koff, J., King, B. A., Rinder, H. M., Wilson, F. P., Hwa, J., Halene, S., Damsky, W., van Dijk, D., Lee, A. I. & Chun, H. J. A neutrophil activation signature predicts critical illness and mortality in COVID-19. *Blood Adv* **5,** 1164–1177 (2021).

118. Muro, A. F., Moretti, F. A., Moore, B. B., Yan, M., Atrasz, R. G., Wilke, C. A., Flaherty, K. R., Martinez, F. J., Tsui, J. L., Sheppard, D., Baralle, F. E., Toews, G. B. & White, E. S. An essential role for fibronectin extra type III domain A in pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **177,** 638–645 (2008).

119. Baiula, M., Spampinato, S., Gentilucci, L. & Tolomelli, A. Novel Ligands Targeting α4β1 Integrin: Therapeutic Applications and Perspectives. *Front Chem* **7,** 489 (2019).

120. Aguirre, C., Meca-Lallana, V., Barrios-Blandino, A., Del Río, B. & Vivancos, J. Covid-19 in a patient with multiple sclerosis treated with natalizumab: May the blockade of integrins have a protective role? *Mult. Scler. Relat. Disord.* **44,** 102250 (2020).

121. Ishii, S., Ford, R., Thomas, P., Nachman, A., Steele, G., Jr & Jessup, J. M. CD44 participates in the adhesion of human colorectal carcinoma cells to laminin and type IV collagen. *Surg. Oncol.* **2,** 255–264 (1993).

122. Wang, L., Cheng, W. & Zhang, Z. Respiratory syncytial virus infection accelerates lung fibrosis through the unfolded protein response in a bleomycin-induced pulmonary fibrosis animal model. *Mol. Med. Rep.* **16,** 310–316 (2017).

123. Shao, H., Qin, Z., Geng, B., Wu, J., Zhang, L., Zhang, Q., Wu, Q., Li, L. & Chen, H. Impaired lung regeneration after SARS-CoV-2 infection. *Cell Prolif.* **53,** e12927 (2020).

124. Wu, C., Thalhamer, T., Franca, R. F., Xiao, S., Wang, C., Hotta, C., Zhu, C., Hirashima, M., Anderson, A. C. & Kuchroo, V. K. Galectin-9-CD44 interaction enhances stability and function of adaptive regulatory T cells. *Immunity* **41,** 270–282 (2014).

125. Ayano, M., Tsukamoto, H., Kohno, K., Ueda, N., Tanaka, A., Mitoma, H., Akahoshi, M., Arinobu, Y., Niiro, H., Horiuchi, T. & Akashi, K. Increased CD226 Expression on CD8+ T Cells Is Associated with Upregulated Cytokine Production and Endothelial Cell Injury in Patients with Systemic Sclerosis. *J. Immunol.* **195,** 892–900 (2015).

126. Herrmann, M., Schulte, S., Wildner, N. H., Wittner, M., Brehm, T. T., Ramharter, M., Woost, R., Lohse, A. W., Jacobs, T. & Schulze Zur Wiesch, J. Analysis of Co-inhibitory Receptor Expression in COVID-19 Infection Compared to Acute Plasmodium falciparum Malaria: LAG-

3 and TIM-3 Correlate With T Cell Activation and Course of Disease. *Front. Immunol.* **11,** 1870 (2020).

127. Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., De Domenico, E., Wendisch, D., Grasshoff, M., Kapellos, T. S., Beckstette, M., Pecht, T., Saglam, A., Dietrich, O., Mei, H. E., Schulz, A. R., Conrad, C., Kunkel, D., Vafadarnejad, E., Xu, C.-J., Horne, A., Herbert, M., Drews, A., Thibeault, C., Pfeiffer, M., Hippenstiel, S., Hocke, A., Müller-Redetzky, H., Heim, K.-M., Machleidt, F., Uhrig, A., Bosquillon de Jarcy, L., Jürgens, L., Stegemann, M., Glösenkamp, C. R., Volk, H.-D., Goffinet, C., Landthaler, M., Wyler, E., Georg, P., Schneider, M., Dang-Heine, C., Neuwinger, N., Kappert, K., Tauber, R., Corman, V., Raabe, J., Kaiser, K. M., Vinh, M. T., Rieke, G., Meisel, C., Ulas, T., Becker, M., Geffers, R., Witzenrath, M., Drosten, C., Suttorp, N., von Kalle, C., Kurth, F., Händler, K., Schultze, J. L., Aschenbrenner, A. C., Li, Y., Nattermann, J., Sawitzki, B., Saliba, A.-E., Sander, L. E. & Deutsche COVID-19 OMICS Initiative (DeCOI). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **182,** 1419–1440.e23 (2020).

128. Fagone, P., Ciurleo, R., Lombardo, S. D., Iacobello, C., Palermo, C. I., Shoenfeld, Y., Bendtzen, K., Bramanti, P. & Nicoletti, F. Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies. *Autoimmun. Rev.* **19,** 102571 (2020).

129. Nagashima, S., Mendes, M. C., Camargo Martins, A. P., Borges, N. H., Godoy, T. M., Miggiolaro, A. F. R. D. S., da Silva Dezidério, F., Machado-Souza, C. & de Noronha, L. Endothelial Dysfunction and Thrombosis in Patients With COVID-19-Brief Report. *Arterioscler. Thromb. Vasc. Biol.* **40,** 2404–2407 (2020).

130. Tong, M., Jiang, Y., Xia, D., Xiong, Y., Zheng, Q., Chen, F., Zou, L., Xiao, W. & Zhu, Y. Elevated Expression of Serum Endothelial Cell Adhesion Molecules in COVID-19 Patients. *J. Infect. Dis.* **222,** 894–898 (2020).

131. Spadaro, S., Fogagnolo, A., Campo, G., Zucchetti, O., Verri, M., Ottaviani, I., Tunstall, T., Grasso, S., Scaramuzzo, V., Murgolo, F., Marangoni, E., Vieceli Dalla Sega, F., Fortini, F., Pavasini, R., Rizzo, P., Ferrari, R., Papi, A., Volta, C. A. & Contoli, M. Markers of endothelial and epithelial pulmonary injury in mechanically ventilated COVID-19 ICU patients. *Crit. Care* **25,** 74 (2021).

132. Barnett, C. C., Jr, Moore, E. E., Mierau, G. W., Partrick, D. A., Biffl, W. L., Elzi, D. J. & Silliman, C. C. ICAM-1-CD18 interaction mediates neutrophil cytotoxicity through protease release. *Am. J. Physiol.* **274,** C1634–44 (1998).

133. Zhao, Y., Qin, L., Zhang, P., Li, K., Liang, L., Sun, J., Xu, B., Dai, Y., Li, X., Zhang, C., Peng, Y., Feng, Y., Li, A., Hu, Z., Xiang, H., Ogg, G., Ho, L.-P., McMichael, A., Jin, R., Knight, J. C., Dong, T. & Zhang, Y. Longitudinal COVID-19 profiling associates IL-1RA and IL-10 with disease severity and RANTES with mild disease. *JCI Insight* **5,** (2020).

134. Xiong, Y., Liu, Y., Cao, L., Wang, D., Guo, M., Jiang, A., Guo, D., Hu, W., Yang, J., Tang, Z., Wu, H., Lin, Y., Zhang, M., Zhang, Q., Shi, M., Liu, Y., Zhou, Y., Lan, K. & Chen, Y. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. *Emerg. Microbes Infect.* **9,** 761–770 (2020).

135. Panda, A. K., Padhi, A. & Prusty, B. A. K. CCR5 Δ32 minorallele is associated with susceptibility to SARS-CoV-2 infection and death: An epidemiological investigation. *Clin. Chim. Acta* **510,** 60–61 (2020).

136. Khalil, B. A., Elemam, N. M. & Maghazachi, A. A. Chemokines and chemokine receptors during COVID-19 infection. *Comput. Struct. Biotechnol. J.* **19,** 976–988 (2021).

137. Rovai, E. S., Alves, T. & Holzhausen, M. Protease-activated receptor 1 as a potential therapeutic target for COVID-19. *Exp. Biol. Med.* **246,** 688–694 (2021).

138. Ivetic, A., Hoskins Green, H. L. & Hart, S. J. L-selectin: A Major Regulator of Leukocyte Adhesion, Migration and Signaling. *Front. Immunol.* **10,** 1068 (2019).

139. Berg, E. L., McEvoy, L. M., Berlin, C., Bargatze, R. F. & Butcher, E. C. L-selectin-mediated lymphocyte rolling on MAdCAM-1. *Nature* **366,** 695–698 (1993).

140. Ogega, C. O., Skinner, N. E., Blair, P. W., Park, H.-S., Littlefield, K., Ganesan, A., Ladiwala, P., Antar, A. A., Ray, S. C., Betenbaugh, M. J., Pekosz, A., Klein, S. L., Manabe, Y. C., Cox, A. L. & Bailey, J. R. Durable SARS-CoV-2 B cell immunity after mild or severe disease. *medRxiv* (2020). doi:10.1101/2020.10.28.20220996

141. Jin, M., Shi, N., Wang, M., Shi, C., Lu, S., Chang, Q., Sha, S., Lin, Y., Chen, Y., Zhou, H., Liang, K., Huang, X., Shi, Y. & Huang, G. CD45: a critical regulator in immune cells to predict severe and non-severe COVID-19 patients. *Aging* **12,** 19867–19879 (2020).

142. Yang, Y., Shen, C., Li, J., Yuan, J., Wei, J., Huang, F., Wang, F., Li, G., Li, Y., Xing, L., Peng, L., Yang, M., Cao, M., Zheng, H., Wu, W., Zou, R., Li, D., Xu, Z., Wang, H., Zhang, M., Zhang, Z., Gao, G. F., Jiang, C., Liu, L. & Liu, Y. Plasma IP-10 and MCP-3 levels are highly associated with disease severity and predict the progression of COVID-19. *J. Allergy Clin. Immunol.* **146,** 119–127.e4 (2020).

143. Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T. X., Oishi, K., Panis, M., Sachs, D., Wang, T. T., Schwartz, R. E., Lim, J. K., Albrecht, R. A. & tenOever, B. R. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181,** 1036–1045.e9 (2020).

144. Hue, S., Beldi-Ferchiou, A., Bendib, I., Surenaud, M., Fourati, S., Frapard, T., Rivoal, S., Razazi, K., Carteaux, G., Delfau-Larue, M.-H., Mekontso-Dessap, A., Audureau, E. & de Prost, N. Uncontrolled Innate and Impaired Adaptive Immune Responses in Patients with COVID-19 Acute Respiratory Distress Syndrome. *Am. J. Respir. Crit. Care Med.* **202,** 1509–1519 (2020).

145. Indari, O., Jakhmola, S., Manivannan, E. & Jha, H. C. An Update on Antiviral Therapy Against SARS-CoV-2: How Far Have We Come? *Front. Pharmacol.* **12,** 632677 (2021).

146. Wang, J., Shiratori, I., Uehori, J., Ikawa, M. & Arase, H. Neutrophil infiltration during inflammation is regulated by PILRα via modulation of integrin activation. *Nat. Immunol.* **14,** 34–40 (2013).

147. Tang, R., Rangachari, M. & Kuchroo, V. K. Tim-3: A co-receptor with diverse roles in T cell exhaustion and tolerance. *Semin. Immunol.* **42,** 101302 (2019).

148. Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., Chen, L., Li, M., Liu, Y., Wang, G., Yuan, Z., Feng, Z., Zhang, Y., Wu, Y. & Chen, Y. Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). *Front. Immunol.* **11,** 827 (2020).

149. Aghbash, P. S., Eslami, N., Shamekh, A., Entezari-Maleki, T. & Baghi, H. B. SARS-CoV-2 infection: The role of PD-1/PD-L1 and CTLA-4 axis. *Life Sci.* **270,** 119124 (2021).

150. Leoni, G., Alam, A., Neumann, P.-A., Lambeth, J. D., Cheng, G., McCoy, J., Hilgarth, R. S., Kundu, K., Murthy, N., Kusters, D., Reutelingsperger, C., Perretti, M., Parkos, C. A., Neish, A. S. & Nusrat, A. Annexin A1, formyl peptide receptor, and NOX1 orchestrate epithelial repair. *J. Clin. Invest.* **123,** 443–454 (2013).

151. Bonavita, A. G. Ac2-26 mimetic peptide of annexin A1 to treat severe COVID-19: A hypothesis. *Med. Hypotheses* **145,** 110352 (2020).

152. Jeong, Y. S. & Bae, Y.-S. Formyl peptide receptors in the mucosal immune system. *Exp. Mol. Med.* **52,** 1694–1704 (2020).

Chapter 4: Combining LIANA and Tensor-cell2cell to decipher cell-cell communication across multiple samples

In recent years, data-driven inference of cell-cell communication has helped reveal coordinated biological processes across cell types. While multiple cell-cell communication tools exist, results are specific to the tool of choice, due to the diverse assumptions made across computational frameworks. Moreover, tools are often limited to analyzing single samples or to performing pairwise comparisons. As experimental design complexity and sample numbers continue to increase in single-cell datasets, so does the need for generalizable methods to decipher cell-cell communication in such scenarios. Here, we integrate two tools, LIANA and Tensor-cell2cell, which combined can deploy multiple existing methods and resources, to enable the robust and flexible identification of cell-cell communication programs across multiple samples. In this protocol, we show how the integration of our tools facilitates the choice of method to infer cell-cell communication and subsequently perform an unsupervised deconvolution to obtain and summarize biological insights. We explain how to perform the analysis step-by-step in both Python and R, and we provide online tutorials with detailed instructions available at https://ccc-protocols.readthedocs.io/. This protocol typically takes ~1.5h to complete from installation to downstream visualizations on a GPU-enabled computer, for a dataset of ~63k cells, 10 cell types, and 12 samples.

4.1 Introduction

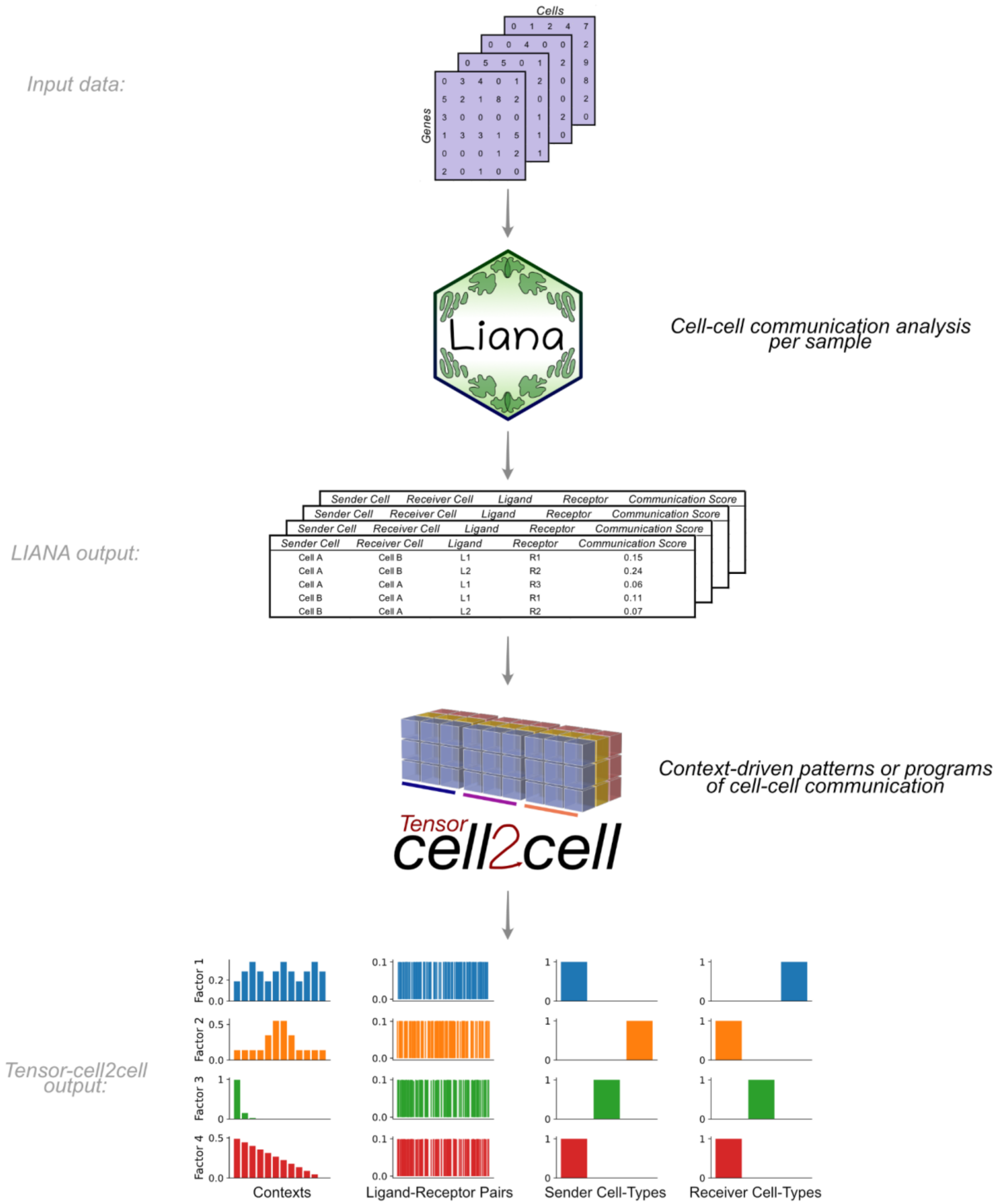
Cell-cell communication (CCC) coordinates higher-order biological functions in multicellular organisms[1,2], dictating phenotypes in response to different contexts such as disease state, spatial location, and organismal life stage. In recent years, many tools have been developed to leverage single-cell and spatial transcriptomics data to understand CCC events driving various biological processes. While each computational strategy contributes unique and valuable developments, many are tool-specific and challenging to integrate due to a plethora of different inference methods and resources housing prior knowledge[3]. Moreover, most tools do not account for the relationships of coordinated CCC events (CCC programs) across different contexts[4], either disregarding context altogether by analyzing samples individually or being limited to pairwise comparisons. Thus, as the ability to generate large single-cell and spatial transcriptomics datasets and the interest in studying CCC programs continues to increase[5–7], the need to robustly decipher CCC is becoming essential.

4.1.1 Development of the protocol

We combine two independent yet highly complementary tools that leverage existing methods to enable robust and hypothesis-free analysis of context-driven cell-cell communication programs (**Fig. 4.1**). LIANA[3] is a computational framework that implements multiple available ligand-receptor resources (i.e., database of ligand-receptor interactions) and methods to analyze CCC. In particular, the user can employ LIANA to select any method and resource of choice or combine multiple approaches simultaneously to obtain consensus predictions. Tensor-cell2cell[8] is a dimensionality reduction approach devised to uncover context-driven CCC programs across multiple samples simultaneously. Specifically, Tensor-cell2cell uses CCC scores inferred by any method and arranges the data into a 4D tensor to capture the coordinated relationship between ligand-receptor interactions, communicating cell type pairs, and samples. Together, LIANA and

Tensor-cell2cell unify existing approaches to enable researchers to easily use their preferred CCC resource and method and subsequently analyze any number of samples into biologically-relevant CCC insights without the additional complications of installing separate tools or reconciling discrepancies between them.

For this protocol, we adapted LIANA and Tensor-cell2cell to enable their smooth integration. Thus, our protocol demonstrates the concerted use of both tools, describes the insights they provide, and facilitates the interpretation of their outputs. We base this protocol on recent best practices for single-cell transcriptomics and CCC inference[9]. We begin by processing the key inputs of our tools. Then, we guide the selection of methods and prior-knowledge resources to score intercellular communication, using LIANA's consensus method and resource to infer the potential CCC events for each sample. We use Tensor-cell2cell to summarize the intercellular communication events across samples, and we describe key technical considerations to enable consistent decomposition results. We also showcase the robustness of Tensor-cell2cell to missing values across samples. Finally, we guide the interpretation of the decomposition results, and show multiple downstream analyses and visualizations to facilitate interpretation of the context-dependent CCC programs. For example, we illustrate how biologically-relevant results can be obtained by coupling the outputs with pathway-enrichment analyses. We also provide quickstart and in-depth online tutorials with detailed descriptions of all steps described in this protocol and their crucial parameters. All these materials are available in both Python and R at https://ccc-protocols.readthedocs.io/. Collectively, these materials provide a comprehensive and flexible playbook to investigate cell-cell communication from single-cell transcriptomics.



Sender Cell	Receiver Cell	Ligand	Receptor	Communication Score
Cell A	Cell B	L1	R1	0.15
Cell A	Cell B	L2	R2	0.24
Cell A	Cell A	L1	R3	0.06
Cell B	Cell A	L1	R1	0.11
Cell B	Cell A	L2	R2	0.07

Figure 4.1: Integration of LIANA and Tensor-cell2cell to identify context-driven programs of cell-cell communication.
LIANA and Tensor-cell2cell can be used together to infer the molecular basis of cell-cell interactions by running analysis across multiple samples, conditions or contexts. Given a method, resource, and expression data, LIANA outputs CCC scores for all interactions in a sample. We adapted both tools to be highly compatible with each other, so LIANA outputs can be directly passed to Tensor-cell2cell to detect the programs from the scores computed with LIANA. Tensor-cell2cell uses the communication scores generated for multiple samples to identify context-driven CCC programs.

4.1.2 Applications of the protocol

LIANA and Tensor-cell2cell have been used for diverse purposes. LIANA was initially used to compare and evaluate different ligand-receptor methods in diverse biological contexts. Tensor-cell2cell was originally applied to link CCC programs with different severities of COVID-19 and Autism Spectrum Disorder (ASD)[8]. Briefly, LIANA evaluated different methods and showed that they have limited agreement in terms of communication mechanisms[3,8], while Tensor-cell2cell revealed distinct CCC program dysregulations associated with severe COVID-19 specifically rather than moderate cases, as well as combinations of programs distinguishing ASD from neurotypical condition. Notably, LIANA provides a consensus resource and can aggregate multiple methods into consensus communication scores. Additionally, there is a natural complementarity between the two tools, as Tensor-cell2cell can use input scores from any CCC method (**Fig. 4.1**) and generates consistent decomposition results across methods. Thus, our tools are highly generalizable and applicable to the analysis of any single-cell transcriptomics datasets. For example, LIANA has been used for the analysis of myocardial infarction[10] and TGFβ signaling in breast cancer[11], among others. Our tools are also applicable to other data modalities containing potentially interacting cell populations. Specifically, one can adapt LIANA or use existing spatial tools[12] and combine their outputs with Tensor-cell2cell to generate spatially-informed CCC insights across contexts. Similarly, one can also obtain metabolite-mediated intercellular interactions[13,14], and decompose those into patterns across contexts with Tensor-cell2cell[15]. One can also apply Tensor-cell2cell to extract CCC programs occurring at specific tissues[16] or at a whole-body organism level[17]. In this protocol, we focus on how one can leverage the different CCC methods and resources, generalized by LIANA, to infer context-dependent CCC programs with Tensor-cell2cell from single-cell transcriptomics data.

4.1.3 Comparison with other methods

A plethora of ligand-receptor methods have emerged, most of which were published with their own resources[1,3,8]. Many of these provide distinct scoring functions to prioritize interactions, yet studies have reported low agreement between their predictions[3,18,19]. Due to the lack of a gold standard, the benchmark of these methods remains limited and it is challenging to choose the method that works best. To this end, in addition to providing multiple individual methods via LIANA, we also enable their consensus, which we use in this protocol, under the assumption that the wisdom of the crowd is less biased than any individual method.

While many methods exist to infer ligand-receptor interactions from a single sample, fewer approaches were designed to compare CCC interactions across conditions. These include CrossTalkeR[20], which utilizes network topological measures to compare communication patterns, CellPhoneDB[21], which accepts user-provided lists of differentially-expressed genes to return relevant ligand-receptor interactions, and scDiffCom, which uses a combined permutation approach across both cell types and conditions. Still, the aforementioned approaches are limited to pairwise comparisons. To our knowledge the only approach other than Tensor-cell2cell that can handle more than two conditions is CellChat[22]; however it is still based on pairwise comparisons, subsequently applying a manifold learning to summarize pathway-focused similarities of contexts. A key advantage of Tensor-cell2cell is that it considers all samples simultaneously while preserving the relationships between ligand-receptor interactions and communicating cell-type pairs. Thus, Tensor-cell2cell preserves higher-order CCC relationships and translates those into mechanistic CCC programs of potentially interacting ligands, receptors and communicating cell types.

4.1.4 Limitations

Although our tools provide robust and flexible solutions to infer CCC patterns across contexts, they inherit the limitations associated with inferring communication events from transcriptomics data. These include the assumption that gene co-expression is indicative of active signaling events, which are largely mediated by proteins and their interactions, while also disregarding any biological processes, such as protein translation, post-translational modifications, secretion, diffusion, and trigger of intracellular events that precede and follow the interaction itself. Moreover, the aggregation of single cells into cell groups is essential when inferring potential CCC events, which could occlude some signals in heterogeneous tissues, thereby biasing the insights that can be obtained. Finally, since the input of Tensor-cell2cell is a 4D-tensor, it requires that all elements are measured across all features and samples. Consequently, one should consider how to handle missing values caused by samples that do not present the same cell types and/or expressed genes when constructing this tensor. Deciding whether those reflect biologically-meaningful zeroes or a technical artifact may lead to variations in the resulting CCC patterns. We provide an extended discussion and analysis of the related parameter choices that may help users decide how to handle this challenge (Appendix 4.6.1).

4.1.5 Expertise needed to implement the protocol

Our protocol requires basic understanding of Python or R and single-cell data analysis. Yet, some of the detailed tutorials also touch on considerations that would be of interest to computational biologist power users.

4.2 Materials

4.2.1 Hardware

This protocol was run on a computer with the following specifications:

- CPU: AMD Ryzen Threadripper 3960x (24 cores)
- Memory: 128GB DDR4
- GPU: NVIDIA RTX A6000 48GB

However, the minimal requirements for running this protocol are:

- CPU: 64-bit Intel or AMD processor (4 cores)
- Memory: 16GB DDR3
- GPU: NVIDIA GTX 1050 Ti (Optional)
- Storage: At least 10GB available

4.2.2 Software

Table 4.1: Required packages for computational environment.

Package Name	Package Version	Language	Install With
jupyter			conda
ipywidgets			conda
pip	>=22	Python	conda
scanpy	>=1.9	Python	conda
*cuda-toolkit			conda
*pytorch-cuda	11.6		conda
*torchvision			conda
*torchaudio			conda
pytorch, *cuda enabled			conda
scvi-tools	>=0.18	Python	conda
scikit-misc	0.1.4	Python	conda
cell2cell	0.6.7	Python	pip
liana	0.1.7	Python	pip
decoupler	1.3.3	Python	pip
omnipath	1.0.6	Python	pip
singlecellexperiment		R	conda
remotes	>=2	R	conda
devtools	>=2	R	conda
seuratobject		R	conda
biocmanager	>=1.30	R	conda
seurat	>=4	R	conda
hd5r		R	conda
furrr		R	conda
textshape		R	conda
forcats		R	conda
rstatix		R	conda
ggpubr		R	conda
scater		R	conda
zellkonverter		R	conda
liana	Commit ID: ab70b34066f68df60e9ed0d0ce72b0d00f871b7e	R	remotes
seurat-disk	Commit ID: 9b89970eac2a3bd770e744f63c7763419486b14c	R	remotes
decoupleR	Commit ID: c17d635e0720c86f2386c39ad7dea8614df393f1	R	biocmanager

*: For GPU enabled use only

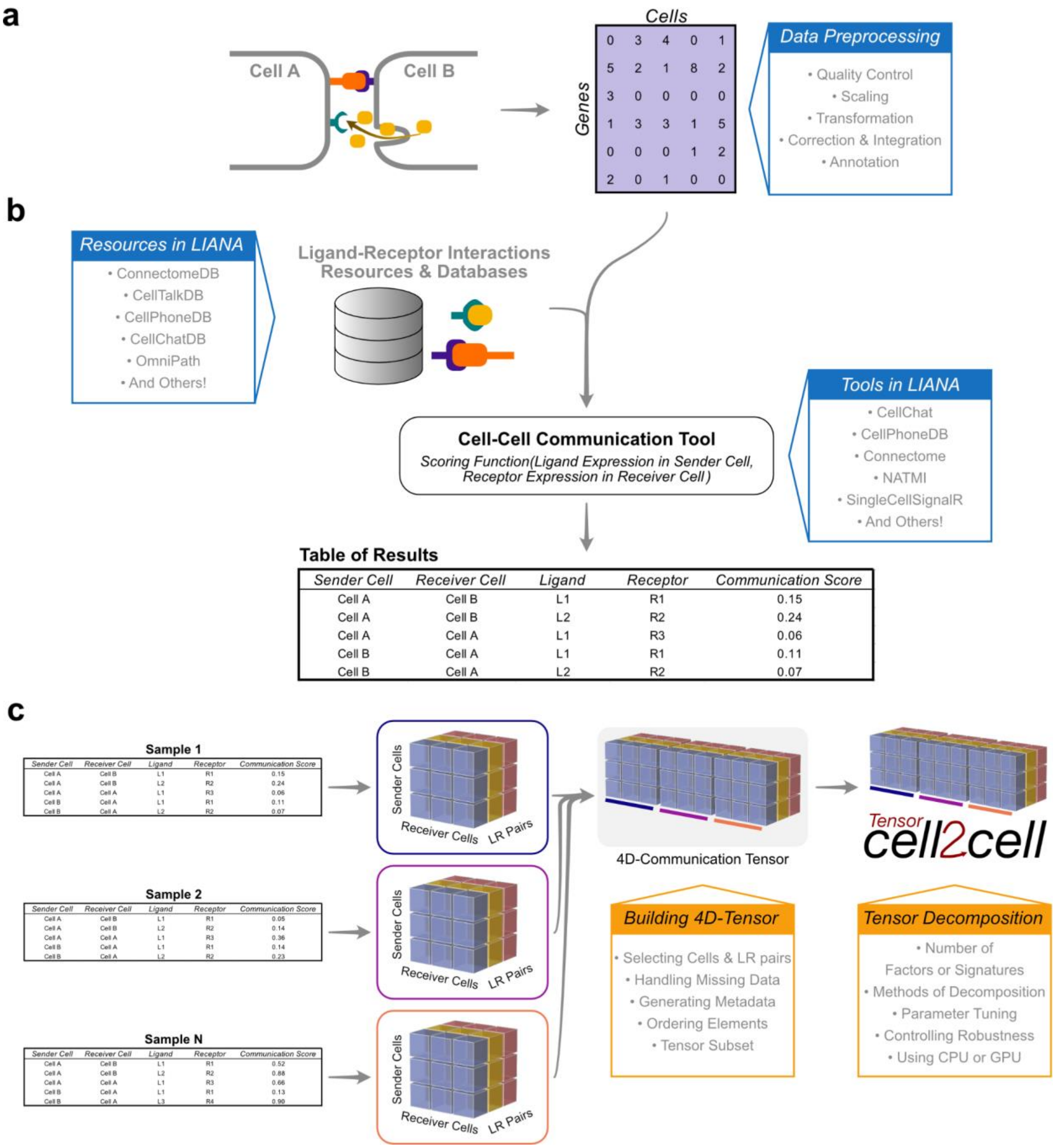
Python packages should always be installed. R language packages only need to be installed if planning to run the notebooks in R.

4.2.3 Equipment setup

To facilitate the setup of a virtual environment containing all required packages with their corresponding versions, we provide an executable `setup_env.sh` script together with instructions on a Github repository we prepared for this protocol: https://github.com/saezlab/ccc_protocols/tree/main/env_setup

4.3 Procedure

∆ **CRITICAL** In this section we introduce our protocol (**Fig. 4.2**) using Python. The same protocol is implemented in R and is available online at https://ccc-protocols.readthedocs.io/en/latest/notebooks/ccc_R/QuickStart.html.



0	3	4	0	1
5	2	1	8	2
3	0	0	0	0
1	3	3	1	5
0	0	0	1	2
2	0	1	0	0

Data Preprocessing

- Quality Control
- Scaling
- Transformation
- Correction & Integration
- Annotation

Resources in LIANA

- ConnectomeDB
- CellTalkDB
- CellPhoneDB
- CellChatDB
- OmniPath
- And Others!

Tools in LIANA

- CellChat
- CellPhoneDB
- Connectome
- NATMI
- SingleCellSignalR
- And Others!

Cell-Cell Communication Tool

Scoring Function(Ligand Expression in Sender Cell, Receptor Expression in Receiver Cell)

Table of Results

Sender Cell	*Receiver Cell*	*Ligand*	*Receptor*	*Communication Score*
Cell A	Cell B	L1	R1	0.15
Cell A	Cell B	L2	R2	0.24
Cell A	Cell A	L1	R3	0.06
Cell B	Cell A	L1	R1	0.11
Cell B	Cell A	L2	R2	0.07

Sample 1

Sender Cell	*Receiver Cell*	*Ligand*	*Receptor*	*Communication Score*
Cell A	Cell B	L1	R1	0.15
Cell A	Cell B	L2	R2	0.24
Cell A	Cell A	L1	R3	0.06
Cell B	Cell A	L1	R1	0.11
Cell B	Cell A	L2	R2	0.07

Sample 2

Sender Cell	*Receiver Cell*	*Ligand*	*Receptor*	*Communication Score*
Cell A	Cell B	L1	R1	0.05
Cell A	Cell B	L2	R2	0.14
Cell A	Cell A	L1	R3	0.36
Cell B	Cell A	L1	R1	0.14
Cell B	Cell A	L2	R2	0.23

Sample N

Sender Cell	*Receiver Cell*	*Ligand*	*Receptor*	*Communication Score*
Cell A	Cell B	L1	R1	0.52
Cell A	Cell B	L2	R2	0.88
Cell A	Cell A	L1	R3	0.66
Cell B	Cell A	L1	R1	0.13
Cell B	Cell A	L3	R4	0.90

Building 4D-Tensor

- Selecting Cells & LR pairs
- Handling Missing Data
- Generating Metadata
- Ordering Elements
- Tensor Subset

Tensor Decomposition

- Number of Factors or Signatures
- Methods of Decomposition
- Parameter Tuning
- Controlling Robustness
- Using CPU or GPU

Figure 4.2: Overview of the protocol for inferring cell-cell communication through LIANA and Tensor-cell2cell.

Main inputs, steps, resources and options are summarized for the distinct steps of this protocol: (**a**) A preprocessed gene expression matrix according to the best practices of single-cell analysis is expected as input (step 3 in the Procedure section). (**b**) This input data is integrated with the ligand-receptor resources available in LIANA to infer cell-cell communication using any of the methods implemented in LIANA (step 4 in the Procedure section). An output containing the cell-cell communication scores across all interactions per sample is generated. (**c**) The LIANA output is then directly passed to Tensor-cell2cell to build the respective communication tensor used by the tensor component analysis (steps 5.1-5.2 in the Procedure section). The output generated by Tensor-cell2cell can be later employed for other downstream analyses (steps 5.3 and 6 in the Procedure section).

4.3.1 Installation and Environment Setup

Install Anaconda or Miniconda through the official instructions at: https://docs.anaconda.com/anaconda/install/index.html. Then, open a terminal to create and activate a conda environment:

```
conda create -n ccc_protocols
conda activate ccc_protocols
```

If you will be using a GPU, install PyTorch using conda:

```
conda install pytorch torchvision torchaudio pytorch-cuda=11.6 -c pytorch -c nvidia
```

Install Tensor-cell2cell, LIANA, and decoupler using PyPI:

```
pip install cell2cell liana decoupler
```

For fully reproducible runs of our Tutorials in both Python and R, we have specified the required packages and their versions in **Table 4.1**. You can also follow instructions in the *Environment setup* section to install a clean virtual environment with all package requirements.

Notebooks to run this tutorial can be created by starting jupyter notebook:

```
jupyter notebook
```

4.3.2 Initial Setups

First, if you are using a NVIDIA GPU with CUDA cores, set `use_gpu=True` and enable PyTorch with the following code block. Otherwise, set `use_gpu=False` or skip this part.

```
use_gpu = True
if use_gpu:
import tensorly as tl
tl.set_backend('pytorch')
```

Then, import all the packages we will use in this tutorial:

```
import cell2cell as c2c
import liana as li
import pandas as pd
import decoupler as dc
import scanpy as sc
import matplotlib.pyplot as plt
%matplotlib inline
import plotnine as p9
import seaborn as sns
```

Afterwards, specify the data and output directories:

```
data_folder = '../../data/'
output_folder = '../../data/outputs/'
c2c.io.directories.create_directory(data_folder)
c2c.io.directories.create_directory(output_folder)
```

We begin by loading the single-cell transcriptomics data. For this tutorial, we will use a lung dataset of 63k immune and epithelial cells across three control, three moderate, and six severe COVID-19 patients[23]. We use a convenient function to download the data and store it in the AnnData format, on which the scanpy[26] package is built.

```
adata = c2c.datasets.balf_covid(data_folder + '/Liao-BALF-COVID-19.h5ad')
```

4.3.3 Data Preprocessing

Data preprocessing is crucial for the correct application of this (**Fig. 4.2**a). Here, we only highlight the essential steps. However, other aspects of data preprocessing should be considered and performed according to the best practices of single-cell analysis (https://github.com/theislab/single-cell-best-practices).

4.3.3.1 Quality Control ● TIMING < 5 min

The loaded data has already been pre-processed to a degree and comes with cell annotations. Nevertheless, we highlight some of the key steps. To mitigate noise, we filter non-informative cells and genes:

```
sc.pp.filter_cells(adata, min_genes=200)
sc.pp.filter_genes(adata, min_cells=3)
```

We additionally remove a high mitochondrial content:

```
adata.var['mt'] = adata.var_names.str.startswith('MT-')
sc.pp.calculate_qc_metrics(adata,
                           qc_vars=['mt'],
                           percent_top=None,
                           log1p=False,
                           inplace=True)
adata = adata[adata.obs.pct_counts_mt < 15, :]
```

Which is followed by removing cells with a high number of total UMI counts, potentially representing more than one single cell (doublets):

```
adata = adata[adata.obs.n_genes < 5500, :]
```

! CAUTION Here, we covered the absolute basics. We omit other common practice steps, such as the removal of cells with high ribosomal content and the correction of ambient RNA. Additionally, in certain scenarios, particularly in such where technical variation is expected to be notable, the application of quality control steps by sample is desirable.

4.3.3.2 Normalization ● TIMING < 2 min

We have now removed the majority of noisy readouts and we can proceed to count normalization, as most cell-cell communication tools typically use normalized count matrices as input. Normalized counts are usually obtained in two essential steps, the first being count depth scaling which ensures that the measured count depths are comparable across cells. This is then usually followed up with log1p transformation, which stabilizes the variance of the counts and enables the use of linear metrics downstream:

```
# Save the raw counts to a layer
adata.layers["counts"] = adata.X.copy()
# Normalize the data
sc.pp.normalize_total(adata, target_sum=1e4)
sc.pp.log1p(adata)
```

∆ CRITICAL A key parameter of this command is:

- **target_sum** ensures that after normalization each observation (cell) has a total count equal to that number.

These normalization steps ensure that the aggregation of cells into cell types, a common practice for CCC inference, is done on comparable cells with approximately normally-distributed feature values.

? TROUBLESHOOTING Expression matrices with nan or inf values causes errors. Users should stick to common normalization techniques, and any nan, negative or inf values must be filled to avoid errors.

4.3.4 Inferring cell-cell communication

Following preprocessing of the single-cell transcriptomics data, we proceed to the inference of potential CCC events (**Fig. 4.3b**). In this case, we will use LIANA to infer the ligand-receptor interactions for each sample. LIANA is available in Python and R, and supports Scanpy, SingleCellExperiment and Seurat objects as input. LIANA is highly modularized, and it natively implements the formulations of several methods, including CellPhoneDBv2[24], Connectome[25], log2FC, NATMI[26], SingleCellSignalR[27], CellChat[22], a geometric mean, as well as a consensus score in the form of a rank aggregate[28] from any combination of methods (**Fig. 4.3**). The high modularity of LIANA further enables the straightforward addition of any other ligand-receptor method.

LIANA classifies the scoring functions from the different methods into two categories: those that infer the "*Magnitude*" and "*Specificity*" of interactions. The "*Magnitude*" of an interaction is a measure of the strength of the interaction, and the "*Specificity*" of an interaction is a measure of how specific an interaction is to a given pair of cell groups. Generally, these categories are complementary, and the magnitude of the interaction is often in agreement with the specificity of the interaction. In other words, a ligand-receptor interaction with a high magnitude score in a given pair of cell types is likely to also be specific, and vice versa.

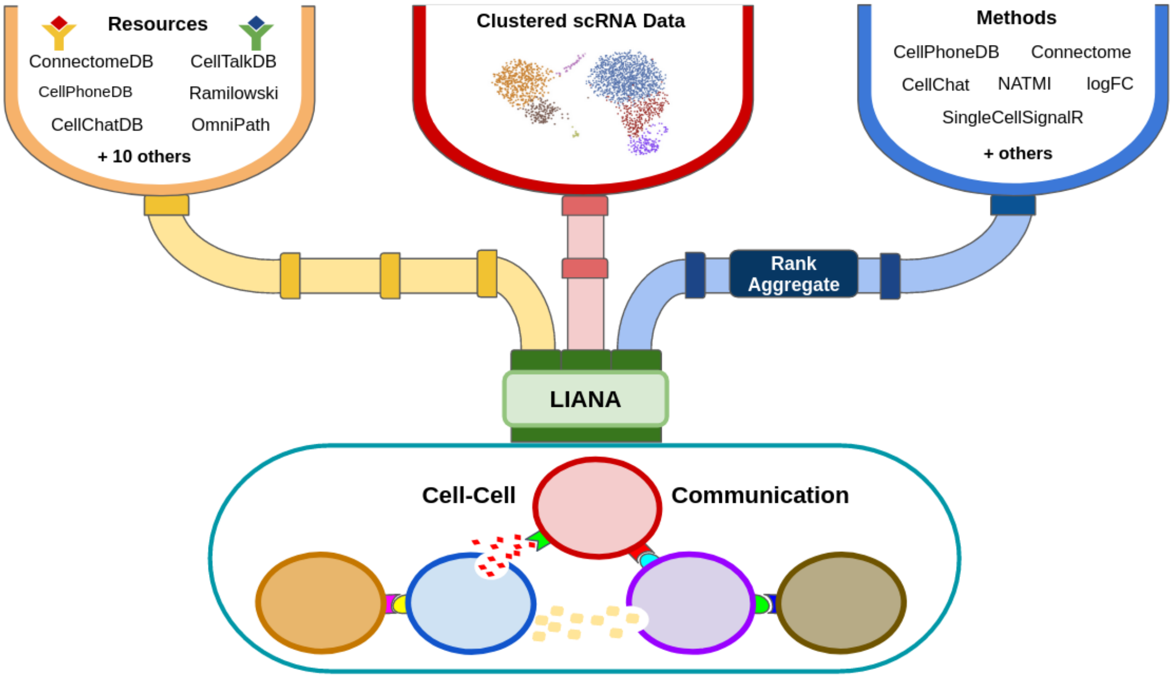


Figure 4.3: LIANA is a user-friendly and modular ligand-receptor analysis framework. LIANA provides a variety of methods and resources to infer cell-cell communication, making it easy to use multiple existing methods in a coherent manner. It also provides consensus scores and resources to provide generalized results. Figure adapted from[3].

4.3.4.1 Selecting a method to infer cell-cell communication

While there are many commonalities between the different methods implemented in LIANA, there also are many variations and different assumptions affecting how the magnitude and specificity scores are calculated. These variations can result in limited agreement in inferred predictions when using different CCC methods[3,18,19]. To this end, in LIANA we additionally provide a rank_aggregate score, that can be used to aggregate any of the scoring functions above into a consensus score.

By default, LIANA calculates an aggregate rank using a re-implementation of the RobustRankAggregate method[32], and generates a probability distribution for ligand-receptors that are ranked consistently better than expected under a null hypothesis. The consensus of ligand-receptor interactions across methods can therefore be treated as a P-value. We show in detail how LIANA's rank aggregate or any of the individual methods can be used to infer communication events from a single sample or context at “Python Tutorial 02 Infer-Communication-Scores” [https://ccc-protocols.readthedocs.io/en/latest/notebooks/ccc_python/02-Infer-Communication-Scores.html].

∆ **CRITICAL** When using LIANA with Tensor-cell2cell, we recommend selecting a scoring function that reflects the *Magnitude* of the interactions, as how the interactions *Specificity* relates to changes across samples is unclear. In this protocol, we will use the `*magnitude_rank*` scoring function from LIANA, under the assumption that ensemble approaches are potentially less biased than any single method alone[20].

? TROUBLESHOOTING The default decomposition method of Tensor-cell2cell is a non-negative Tensor Component Analysis, which, as implied, expects non-negative values as the inputs. Thus, when selecting the method of choice, make sure that you do not have negative CCC scores. If so, you can replace them by zeros or the minimum positive value.

4.3.4.2 Selecting ligand-receptor resources

When considering ligand-receptor prior knowledge resources, a common theme is the trade-off between coverage and quality, and similarly each resource comes with its own biases[3]. LIANA takes advantage of OmniPath[29], which includes expert-curated resources of CellPhoneDBv2, CellChat, ICELLNET[30], connectomeDB2020[26], CellTalkDB[31], as well as 10 others[3,29]. LIANA further provides a consensus expert-curated resource from the aforementioned five resources, along with some curated interactions from SignaLink[32]. In this protocol, we will use the consensus resource from LIANA, though any of the other resources are available via LIANA, and one can also use LIANA with their own custom resource.

Selecting any of the lists of ligand-receptor pairs in LIANA can be done through the following command:

```
lr_pairs = li.resource.select_resource('consensus')
```

Here ‘consensus’ indicates the use of LIANA’s consensus resource , but it can be replaced by any other available resource (e.g. ‘cellphonedb’, ‘cellchatdb’, ‘connectomeDB’, etc.).

? TROUBLESHOOTING Users should choose a resource with gene identifiers and organism that corresponds to that of their data. By default, LIANA uses human gene symbol identifiers, but additionally provides a murine resource as well as functionalities to convert via orthology to other organisms.

4.3.4.3 Running LIANA for each sample ● Timing 4 minutes

Here, we will run LIANA’s `rank_aggregate` with six methods (by default, CellPhoneDBv2, CellChat, SingleCellSignalR, NATMI, Connectome, log2FC) on all of the samples in the dataset.

```
li.mt.rank_aggregate.by_sample(adata,
                               sample_key='sample_new',
                               groupby='celltype',
                               resource_name='consensus',
                               expr_prop=0.1,
                               min_cells=5,
                               n_perms=100,
                               use_raw=False,
                               verbose=True,
                               inplace=True
                                      )
```

∆ CRITICAL Key parameters here are:

- **adata** stands for Anndata, the data format used by scanpy[33], and we pass here with an object with a single sample/context.
- **sample_key** corresponds to the sample identifiers, available as a column in the `adata.obs` dataframe.
- **groupby** corresponds to the cell group label stored in `adata.obs`.
- **resource_name** - name of any of the resources available via LIANA
- **expr_prop** is the expression proportion threshold (in terms of cells per cell type

expressing the protein) for any protein subunit involved in the interaction, according to which we keep or discard the interactions.

- **min_cells** is the minimum number of cells per cell type required for a cell type to be considered in the analysis
- **n_perms** is the number of permutations for P-value estimation
- **use_raw** is a boolean that indicates whether to use the `adata.raw` slot, here the log-normalized counts are assigned to `adata.X`, other options include passing the name of a layer via the `layer` parameter or using the counts stored in `adata.raw`**.**
- **verbose** is a Boolean value that indicates whether to print the progress of the function
- **inplace** indicates whether storing the results in place, i.e. to `adata.uns[“liana_res”]`.

∆ **CRITICAL** LIANA considers interactions as occurring only if the ligand and receptor, and all of their subunits, are expressed in at least 10% of the cells (by default) in both clusters involved in the interaction. Any interactions that do not pass these criteria are not returned by default. To return those, the user can use the `return_all_lrs` parameter. These results will later be used to generate a tensor of ligand-receptor interactions across contexts that will be decomposed into CCC patterns by Tensor-Cell2cell. Thus, how non-expressed interactions are handled is critical to consider when building the tensor later on (see “Python Tutorial 03 Generate-Tensor”.

4.3.4.4 Visualize output

One can visualize the output as a dotplot, but with the addition of the samples.

```
li.pl.dotplot_by_sample(adata=adata,
                        colour='magnitude_rank',
                        size='specificity_rank',
                        source_labels=["B", "pDC", "Macrophages"],
                        target_labels=["T", "Mast", "pDC", "NK"],
                        ligand_complex='B2M',
                        receptor_complex=['CD3D', 'KLRD1'],
                        sample_key='sample_new',
                        inverse_colour=True,
                        inverse_size=True,
                        figure_size=(9, 9),
                        size_range=(1, 6),
                        )
```

Key parameters here are:

- **source_labels** is a list containing the names of the sender cells of interest.
- **target_labels** is a list containing the names of the receiver cells of interest.
- **ligand_complex** is a list containing the names of the ligands of interest.
- **receptor_complex** is a list containing the names of the receptors of interest.
- **sample_key** is a string containing the column name where samples are specified.

■ **PAUSE POINT** We can export the liana results by sample to a csv, and save them for later use:

```
adata.uns['liana_res'].to_csv(output_folder + '/LIANA_by_sample.csv', index=False)
```

Alternatively one could just export the whole AnnData object, together with the ligand-receptor results stored at `adata.uns[‘liana_res’]`:

```
adata.write_h5ad(output_folder + '/adata_processed.h5ad', compression='gzip')
```

4.3.5 Comparing cell-cell communication across multiple samples

4.3.5.1 Building a 4D-communication tensor ● Timing <1 minute

First, we generate a list containing all samples from our AnnData object. For visualization purposes we sort them according to COVID-19 severity. Here, we can find the names of each of the samples in the 'sample_new' column of the adata.obs information:

```
sorted_samples = sorted(adata.obs['sample_new'].unique())
```

Tensor-cell2cell performs a tensor decomposition to find context-dependent patterns of cell-cell communication. It builds a 4D-communication tensor, which in this case is built from the communication scores obtained from LIANA for every combination of ligand-receptor and sender-receiver cell pairs per sample (**Fig. 4.2c**). For this protocol and associated tutorials, we implemented a function that facilitates building this communication tensor:

```
tensor = li.multi.to_tensor_c2c(liana_res=adata.uns['liana_res'],
                             sample_key='sample_new',
                             source_key='source',
                             target_key='target',
                             ligand_key='ligand_complex',
                             receptor_key='receptor_complex',
                             score_key='magnitude_rank',
                                     inverse_fun=lambda x: 1 - x,
                             how='outer',
                             outer_fraction=1/3.,
                             context_order=sorted_samples,
                              )
```

? TROUBLESHOOTING Since the `*magnitude_rank*` from LIANA represents a score where the values closest to 0 represent the most probable communication events, we need to invert the communication scores to use it with Tensor-cell2cell. See the parameter `inverse_fun` below for further details for transforming this score.

∆ CRITICAL Key parameters here are:

- **liana_res** is the dataframe containing the results from LIANA, usually located in `adata.uns['liana_res']. We can pass directly the AnnData object to the parameter adata to this function. If the AnnData object is passed, we do not need to specify the liana_res parameter.

- **sample_key**, **source_key**, **target_key**, **ligand_key**, **receptor_key**, and **score_key** are the column names in the dataframe containing the samples, sender cells, receiver cells, ligands, receptors, and communication scores, respectively. Each row of the dataframe contains a unique combination of these elements.
- **inverse_fun** is the function we use to convert the communication score before building the tensor. In this case, the 'magnitude_rank' score generated by LIANA considers low values as the most important ones, ranging from 0 to 1. In contrast, Tensor-cell2cell requires higher values to be the most important scores, so here we pass a function (lambda x: 1 - x) to adapt LIANA's magnitude-rank scores (subtracts the LIANA's score from 1). If None is passed instead, no transformation will be performed on the communication score. If using other scores coming from one of the methods implemented in LIANA, a similar transformation can be done depending on the parameters and assumptions of the scoring method.
- **how** controls which ligand-receptor pairs and cell types to include when building the tensor. This decision depends on whether the missing values across a number of samples for both ligand-receptor interactions and sender-receiver cell pairs are considered to be biologically-relevant. Options are:
    - ***'inner'*** is the most strict option since it only considers cell types and ligand-receptor pairs that are present in all contexts (intersection).
    - ***'outer'*** considers all cell types and ligand-receptor pairs that are present across contexts (union).
    - ***'outer_lrs'*** considers only cell types that are present in all contexts (intersection), while all ligand-receptor pairs that are present across contexts (union).
    - ***'outer_cells'*** considers only ligand-receptor pairs that are present in all contexts

(intersection), while all cell types that are present across contexts (union).

- **outer_fraction** controls the elements to include in the union scenario of the how options. Only elements that are present at least in this fraction of samples/contexts will be included. When this value is 0, the tensor includes all elements across the samples. When this value is 1, it acts as using how='inner'.
- **context_order** is a list specifying the order of the samples. The order of samples does not affect the results, but it is useful for posterior visualizations.

We can check the shape of this tensor to verify the number of samples, ligand-receptor pairs, sender cells, and receiver cells, respectively:

```
Tensor.shape
```

In addition, optionally we can generate the metadata for coloring the elements in each of the tensor dimensions (i.e., for each of the contexts/samples, ligand-receptor pairs, sender cells, and receiver cells) in posterior visualizations. This metadata corresponds to dictionaries for each of the dimensions, containing the elements and their respective major groups, such as a signaling categories for a ligand-receptor interactions, a hierarchically more granular cell type, or a disease condition for a sample. In cases where we do not account for such information, we do not need to generate such dictionaries.

For example, we can build a dictionary for the contexts/samples dictionary by using the metadata in the AnnData object. In this example dataset, we can find samples in the column ‘sample_new’, while their majors groups (representing COVID-19 severity) are found in the column ‘condition’:

```
context_dict = adata.obs.sort_values(by='sample_new') \
                      .set_index('sample_new')['condition'] \
                      .to_dict()
```

Then, the metadata can be generated with:

```
dimension_dicts = [context_dict, None, None, None]
meta_tensor = c2c.tensor.generate_tensor_metadata(interaction_tensor=tensor,
                                        metadata_dicts=dimension_dicts,
                                        fill_with_order_elements=True
                                        )
```

Notice that the None elements in the variable dimensions_dicts represent the dimensions where we are not including additional metadata. If you want to include metadata about major groups for those dimensions, you have to replace the corresponding None by a dictionary as described before.

■ **PAUSE POINT** We can export our tensor and its metadata for performing the tensor decomposition later:

```
c2c.io.export_variable_with_pickle(variable=tensor,
                           filename=output_folder + '/Tensor.pkl')
c2c.io.export_variable_with_pickle(variable=meta_tensor,
                              filename=output_folder + '/Tensor-Metadata.pkl')
```

Then, we can load them with:

```
tensor = c2c.io.read_data.load_tensor(output_folder + '/Tensor.pkl')
meta_tensor = c2c.io.load_variable_with_pickle(output_folder + '/Tensor-Metadata.pkl')
```

4.3.5.2 Running Tensor-cell2cell across samples ● Timing 5 minutes with a ‘regular’ run or 40 minutes with a ‘robust’ run (using a GPU in both cases)

Now that we have built the tensor and its metadata, we can run Tensor Component Analysis via Tensor-cell2cell with one simple command that we implemented for our unified tools:

```
c2c.analysis.run_tensor_cell2cell_pipeline(interaction_tensor=tensor,
                                 tensor_metadata=meta_tensor,
                                                    rank=None,
                                        tf_optimization='robust',
                                        random_state=0,
                                        device='cuda',
                                        output_folder=output_folder,
                              )
```

∆ **CRITICAL Key parameters of this command are:**

- **rank** is the number of factors or latent patterns we want to obtain from the analysis. You can either indicate a specific number or leave it as None to perform the decomposition with a suggested number from an elbow analysis.

- **tf_optimization** indicates whether running the analysis in the 'regular' or the 'robust' way. The regular way runs the tensor decomposition fewer times than the robust way to select an optimal result. Additionally, the former employs less strict convergence parameters to obtain optimal results than the latter, which is also translated into a faster generation of results.
- **random_state** is the seed for randomization. It controls the randomization used when initializing the optimization algorithm that performs the tensor decomposition. It is useful for reproducing the same result every time that the analysis is run. If None, a different randomization will be used each time.
- **device** indicates whether we are using the 'cpu' or a GPU with 'cuda' cores. See the Installation section of this tutorial for instructions to enable the use of GPU(s).
- **output_folder** is the full path to the folder where the results will be saved. Make sure that this folder exists before passing it here.

This command will output three main results: a figure with the elbow analysis for suggesting a number of factors (only if rank=None), a figure with the loadings assigned to each element within a tensor dimension per factor obtained, and an excel file containing the values of these loadings.

? TROUBLESHOOTING Elbow analysis returns a rank equal to one, or the curve is increasing instead of decreasing. This may be due to high sparsity in the tensor. The sparsity can be decreased by re-building the 4D tensor after re-running LIANA (**Step 4.3**) with a smaller \`expr_prop\` (e.g. \`expr_prop=0.05\`) or by only re-building the tensor (**Step 5.1**) with a higher \`outer_fraction\` (e.g. \`outer_fraction=0.8\`).

4.3.5.3 Downstream visualizations: Making sense of the factors ● Timing <2 minutes

The figure representing the loadings in each factor generated in the previous section can be interpreted by interconnecting all dimensions within a single factor. For example, if we take Factor 4 in **Fig. 4.4**, the CCC program here occurs in each sample in a manner proportional to their loadings, here correlated with COVID-19 severity. Relevant interactors can be interpreted according to their loadings (i.e. ligand-receptor pairs, sender cells, and receiver cells with high loadings play a more prominent role in an identified CCC program). Ligands in high-loading ligand-receptor pairs are sent predominantly by high-loading sender cells, and interact with the cognate receptors on the high-loadings receiver cells. In this factor, the program would be predominantly driven by changes in the receptor expression of receiver cells such as macrophages, neutrophils and myeloid dendritic cells.

We can access the loading values of samples in each of the factors with:

```
tensor.factors['Contexts']
```

In this case we obtain a dataframe where rows represent the samples, columns the factors generated by the decomposition, and entries are the loadings of each element within the corresponding factor. We can also access the loadings for the elements in the other dimensions by replacing ‘Contexts’ with ‘Ligand-Receptor Pairs’, ‘Sender Cells’, or ‘Receiver Cells’. Then, we can use these loadings to perform various downstream analyses (**Fig. 4.5**).

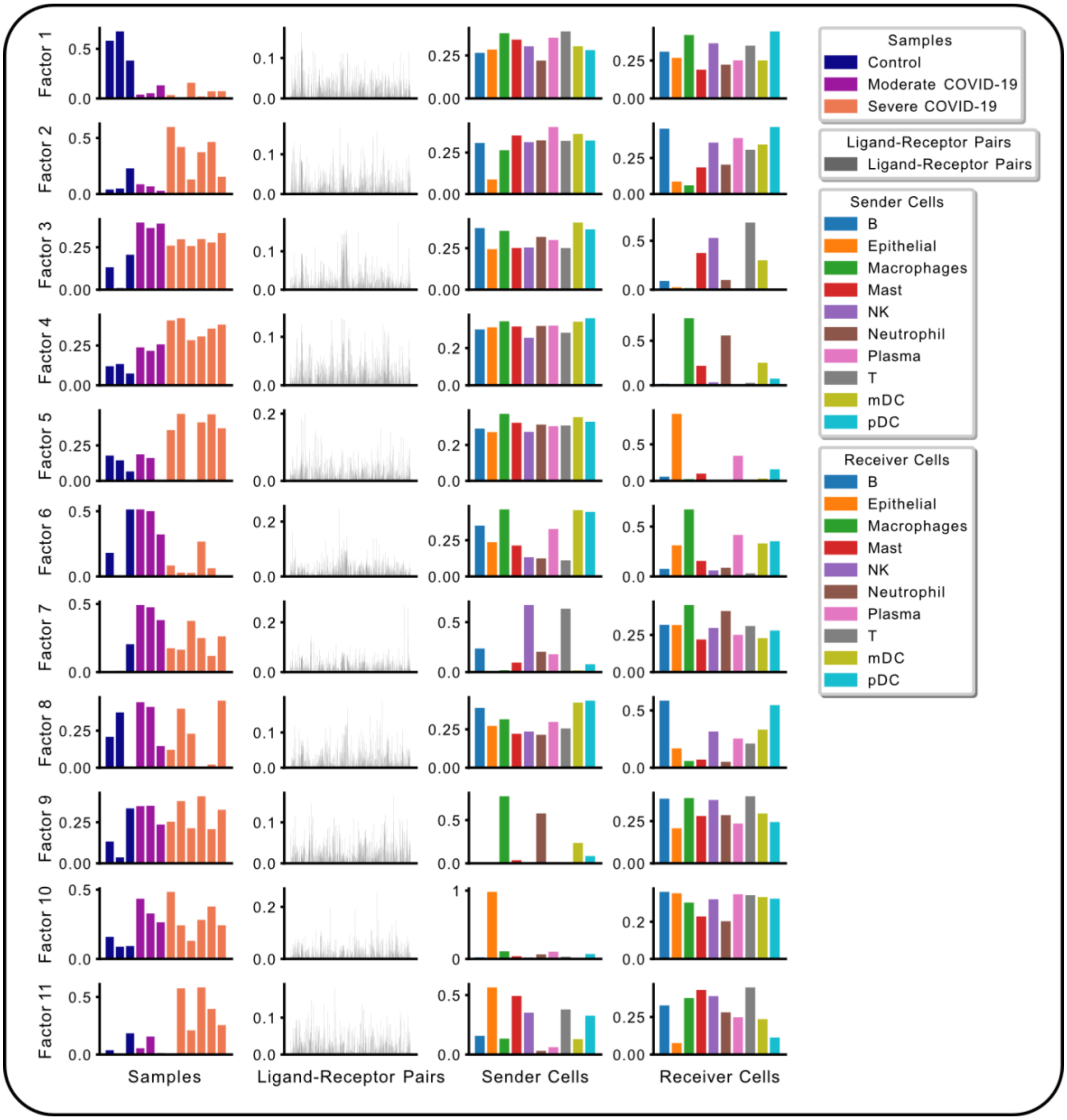
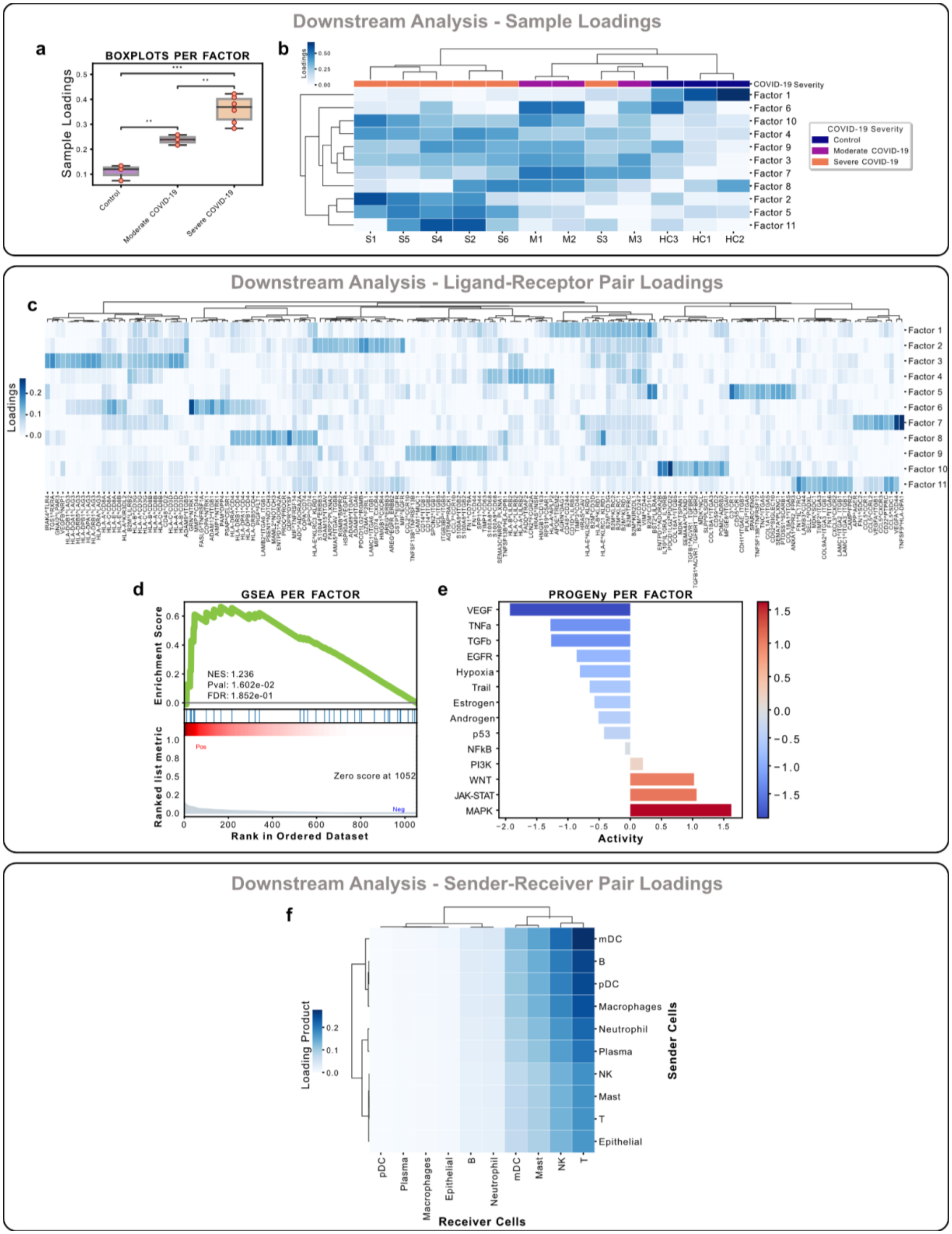


Figure 4.4: Cell-cell communication programs obtained by combining LIANA and Tensor-cell2cell. After inferring cell-cell communication with LIANA from the COVID-19 data, and running a Tensor Component Analysis with Tensor-cell2cell, 11 factors were obtained (rows here), each of which represents a different cell-cell communication program. Within a factor, loadings (y-axis) are reported for each element (x-axis) in every tensor dimension (columns). Elements here are colored by their major groups as indicated in the legend.

Figure 4.5: Examples of downstream analyses performed on the results from the LIANA and Tensor-cell2cell framework.

Downstream analyses can be performed by using the loadings of one of the tensor dimensions. Context or sample loadings (**a-b**) can be used to (**a**) compare statistically different condition groups within the same cell-cell communication program or (**b**) to group samples across all programs. (**c-e**) Similarly, ligand-receptor interactions can be analyzed from their loadings per or across factors. (**c**) Key ligand-receptor pairs whose loadings are above a threshold can be clustered depending on their importance across all cell-cell communication programs. They can also be ranked according to their loadings within a factor (factor-specific analyses), and this information can be used to run an enrichment analysis such as (**d**) GSEA or (**e**) PROGENy to associate each of the programs with different functions or pathways. (**f**) Finally, cell type loadings can be jointly used within a factor to have an overall representation of the cell-cell communication (i.e., a factor-specific network of communication).


Downstream Analysis - Sample Loadings
a
BOXPLOTS PER FACTOR
Sample Loadings
0.5
0.4
0.3
0.2
0.1
***
**
**
Control
Moderate COVID-19
Severe COVID-19
b
Loadings
0.50
0.25
0.00
COVID-19 Severity
Factor 1
Factor 6
Factor 10
Factor 4
Factor 9
Factor 3
Factor 7
Factor 8
Factor 2
Factor 5
Factor 11
S1
S5
S4
S2
S6
M1
M2
S3
M3
HC3
HC1
HC2
COVID-19 Severity
Control
Moderate COVID-19
Severe COVID-19
Downstream Analysis - Ligand-Receptor Pair Loadings
c
Loadings
0.2
0.1
0.0
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
Factor 8
Factor 9
Factor 10
Factor 11
d
GSEA PER FACTOR
Enrichment Score
0.6
0.4
0.2
0.0
NES: 1.236
Pval: 1.602e-02
FDR: 1.852e-01
Ranked list metric
1.0
0.8
0.5
0.2
0.0
Pos
Zero score at 1052
Neg
0
200
400
600
800
1000
Rank in Ordered Dataset
e
PROGENy PER FACTOR
VEGF
TNFa
TGFb
EGFR
Hypoxia
Trail
Estrogen
Androgen
p53
NFkB
PI3K
WNT
JAK-STAT
MAPK
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Activity
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Downstream Analysis - Sender-Receiver Pair Loadings
f
Loading Product
0.2
0.1
0.0
mDC
B
pDC
Macrophages
Neutrophil
Plasma
NK
Mast
T
Epithelial
Sender Cells
pDC
Plasma
Macrophages
Epithelial
B
Neutrophil
mDC
Mast
NK
T
Receiver Cells

For example, we can use loadings to compare groups of samples (**Fig. 4.5**a-b) with box plots and statistical tests:

```
groups_order = ['Control', 'Moderate COVID-19', 'Severe COVID-19']
fig_filename = output_folder + '/BALF-Severity-Boxplots.pdf'
_ = c2c.plotting.context_boxplot(context_loadings=tensor.factors['Contexts'],
                                 metadict=context_dict,
                                 nrows=3,
                                 figsize=(16, 12),
                                group_order=groups_order,
                                 statistical_test='t-test_ind',
                                 pval_correction='fdr_bh',
                                 cmap='plasma',
                                 verbose=False,
                                           filename=fig_filename
                                  )
```

∆ **CRITICAL** In this case, we can change the statistical test and the multiple-test correction with the parameters `statistical_test` and `pval_correction`. Here we used an independent t-test and a Benjamini-Hochberg correction. Additionally, we can set verbose=True to print exact test statistics and P-values.

We can also generate heatmaps for the elements with loadings above a certain threshold in a given dimension (**Fig. 4.5**b,c,f). Furthermore, we can cluster these elements by the similarity of their loadings across all factors:

```
fig_filename = output_folder + '/Clustermap-LRs.pdf'
_ = c2c.plotting.loading_clustermap(loadings=tensor.factors['Ligand-Receptor Pairs'],
                                 loading_threshold=0.1,
                                 use_zscore=False,
                                 figsize=(28, 8),
                                 filename=fig_filename,
                                    row_cluster=False
                                 )
```

? TROUBLESHOOTING Note that here we plot the loadings of the dimension representing the ligand-receptor pairs. In addition, we prioritize the pairs with high loadings using the parameter `loading_threshold=0.1`. In this case, the elements are included only if they are greater than or equal to that threshold in at least one of the factors. If we use `loading_threshold=0`, we would consider all of the elements. Considering all of the elements would require modifying the parameter `figsize` to enlarge the figure.

! CAUTION Changing the parameter `use_zscore` to **True** would standardize the loadings of one element across all factors. This is useful to compare an element across factors and highlight the factors in which that element is most important. Modifying `row_cluster` to **True** would also cluster the factors depending on the elements that are important in each of them.

4.3.6 Pathway Enrichment Analysis: Interpreting the context-driven communication

The decomposition of ligand-receptor interactions across samples into loadings associated with the conditions reduces the dimensionality of the inferred interactions substantially. Nevertheless, we are still working with 1,054 interactions across multiple factors associated with the disease labels. To this end, as is commonly done when working with omics data types, we can perform pathway enrichment analysis to identify the general biological processes of interest. By using the loadings for each ligand-receptor pair, we can rank them within each factor and use this ranking as input to enrichment analysis (**Fig. 4.5**d-e). Pathway enrichment thus serves two purposes; it further reduces the dimensionality of the inferred interactions, and it enhances the biological interpretability of the inferred interactions.

Here, we will show the application of classical gene set enrichment analysis on the ligand-receptor loadings. We will use GSEA[34] with KEGG Pathways[35], as well as a multivariate linear regression from decoupler-py[36] with the PROGENy pathway resource[37].

First, we assign ligand-receptor loadings to a variable:

```
lr_loadings = tensor.factors['Ligand-Receptor Pairs']
```

4.3.6.1 Classic Pathway Enrichment

For the pathway enrichment analysis, we use ligand-receptor pairs instead of individual genes. KEGG was initially designed to work with sets of genes, so first we need to generate ligand-receptor sets for each of its pathways. A ligand-receptor pair is assigned as part of a pathway set if all of the genes in the pair are part of the gene set of such pathway:

```
# Generate list with ligand-receptors pairs in DB
lr_list = ['^'.join(row) for idx, row in lr_pairs.iterrows()]
# Specify the organism and pathway database to use for building the LR set
organism = "human"
pathwaydb = "KEGG"
# Generate ligand-receptor gene sets
lr_set = c2c.external.generate_lr_geneset(lr_list,
                          complex_sep='_',
                              lr_sep='^',
                          organism=organism,
                           pathwaydb=pathwaydb,
                          readable_name=True,
                            output_folder=output_folder
                          )
```

Note that we use the `lr_pairs` database that we loaded in the ***Selecting ligand-receptor resources*** section.

∆ CRITICAL Key parameters of this command are:

- **complex_sep** indicates the symbol separating the gene names in the protein complex.
- **lr_sep** is the symbol separating a ligand and a receptor complex.
- **organism** is the organism matching the gene names in the single-cell dataset. It could be either “human” or “mouse”.
- **pathwaydb** is the name of the database to be loaded, provided with the cell2cell package. Options are “GOBP”, “KEGG”, and “Reactome”.

Run GSEA via cell2cell which calls the `gseapy.prerank` function internally ● **Timing** < 1 minute

```
pvals, scores, gsea_df = c2c.external.run_gsea(loadings=lr_loadings,
                                    lr_set=lr_set,
                                   output_folder=output_folder,
                                    weight=1,
                                    min_size=15,
                                    permutations=999,
                                    processes=6,
                                    random_state=6,
                                    significance_threshold=0.05,
                                    )
```

∆ CRITICAL Key parameters of this command are:

- **lr_set** is a dictionary associating pathways (keys) with ligand-receptor pairs (values).
- **weight** represents the original parameter p in GSEA. It is an exponent that controls the importance of the ranking values (loadings in our case).
- **min_size** indicates the minimum number of LR pairs that a set has to contain to be considered in the analysis.
- **permutations** indicates the number of permutations to perform to generate the null distribution.
- **random_state** is the reproducibility seed.
- **significance_threshold** is the P-value threshold to consider significance.

Now that we have obtained the normalized-enrichment scores (NES) and corresponding P-values from GSEA, we can plot those using the following function from cell2cell:

```
pathway_label = '{} Annotations'.format(pathwaydb)
fig_filename = output_folder + '/GSEA-Dotplot.pdf'
with sns.axes_style("darkgrid"):
         dotplot = c2c.plotting.pval_plot.generate_dot_plot(pval_df=pvals,
                                    score_df=scores,
                                         significance=0.05,
                                        xlabel='',
                                    ylabel=pathway_label,
                                    cbar_title='NES',
                                    cmap='PuOr',
                                    figsize=(5, 12),
                                    label_size=20,
                                     title_size=20,
                                    tick_size=12,
                                    filename=fig_filename
                                    )
```

4.3.6.2 Footprint enrichment analysis

In footprint enrichment analysis, instead of considering the genes whose products (proteins) are directly involved in a process of interest, we consider the genes affected by it - i.e. those that change downstream as a consequence of the process[38]. In this case, we will use the PROGENy resource to infer the pathways driving the identified context-dependent patterns of ligand-receptor pairs. PROGENy was built in a data-driven manner using perturbation data[37]. Consequently, it assigns different weights to each gene in its pathway genesets according to its importance. Thus, we need an enrichment method that can account for weights. To do so, we will use a multivariate linear regression implemented in decoupler-py[36].

As we did in GSEA using Tensor-cell2cell, we first have to generate ligand-receptor gene sets while also assigning a weight to each ligand-receptor interaction. This is done by taking the mean between the ligand and receptor weights. For ligand and receptor complexes, we first take the mean weight for all subunits. We keep ligand-receptor weights only if all the proteins in the interaction are sign-coherent and present for a given pathway.

Load the PROGENy genesets and then convert them to sets of weighted ligand-receptor pairs:

```
# We first load the PROGENy gene sets
net = dc.get_progeny(organism='human', top=5000)
# Then convert them to sets with weighted ligand-receptor pairs
lr_progeny = li.funcomics.generate_lr_geneset(lr_pairs, net, lr_separator="^")
```

Run footprint enrichment analysis using the `mlm` method from decoupler-py ● Timing < 1 minute:

```
estimate, pvals = dc.run_mlm(lr_loadings.transpose(),
                    lr_progeny,
                    source="source",
                    target="interaction",
                    use_raw=False
                    )
```

Here, `estimate` and `pvals` correspond to the t-values and P-values assigned to each pathway. Finally, we generate Heatmap for the 14 Pathways in PROGENy across all Factors:

```
fig_filename = output_folder + '/PROGENy.pdf'
_ = sns.clustermap(estimate, xticklabels=estimate.columns, cmap='coolwarm', z_score=4)
plt.savefig(fig_filename, dpi=300, bbox_inches='tight')
```

From the heatmap, we can also generate a Barplot for the PROGENy pathways for a specific factor:

```
selected_factor = 'Factor 10'
fig_filename = output_folder + '/PROGENy-{}.pdf'.format(selected_factor.replace(' ', '-'))
dc.plot_barplot(estimate,
        selected_factor,
        vertical=True,
              cmap='coolwarm',
        save=fig_filename)
```

Table 4.2: Troubleshooting.

Step	Problem	Possible reason	Solution
3 & 4	Error: Expression matrix contains non-finite values (nan or inf) Warning: Make sure that normalized counts are passed	Mishandling counts processing	Ensure that the matrix containing normalized counts is passed. Replace nan and inf values by zeros.
4.1	Negative values in LIANA outputs	Using preprocessed data with negative expression values.	Avoid using preprocessing methods that generate negative values (e.g. centering the data to the mean values, using batch-corrected expression values, etc.).
4.2	Not enough ligand-receptor pairs in the data for the analysis	Mismatched symbol IDs	LIANA by default uses a resource with gene symbol IDs. When working with e.g. Ensembl IDs users need to provide an external resource; see https://ccc-protocols.readthedocs.io/en/latest/notebooks/ccc_python/02-Infer-Communication-Scores.html
5.1	CCC scores representing opposed importance	When using ‘magnitude_rank’ scores from LIANA, lower values are more important. However, Tensor-cell2cell prioritizes high values as the important ones.	Build the 4D tensor using an \`inverse_fun\` to make lower values to be the most important scores.
5.2	Rank selection through the elbow analysis is not behaving properly	High sparsity or number of missing values in the tensor	Re-run LIANA with less stringent parameters (e.g. smaller expr_pror). Re-build the tensor with more strict how parameters (e.g. using how=’inner’ or increasing outer_fraction).
5.3	Visualization of loadings are not properly displayed in heatmaps	Too many or few elements in the dimension to visualize	To visualize all elements, use the parameter \`loading_threshold=0\` to create the heatmaps. If you have too many elements, you can prioritize those with high loadings, so a threshold can be set. E.g., \`loading_threshold=0.1\`

4.5 Anticipated Results

Deciphering cell-cell communication with LIANA yields all ligand-receptor interactions, defined in the prior knowledge resource, for every pair of cell types within the dataset. For each interaction, a set of statistics is assigned. These typically include a value that reflects the magnitude and specificity of interaction depending on the method of choice. The magnitude scores for each interaction in each sample are transformed into a 4D tensor that is then decomposed by Tensor-cell2cell. Prior to decomposition, it is recommended to estimate the optimal number of factors required to reconstruct the original tensor. For each output factor, we obtain four vectors that represent the sample, ligand-receptor interaction, sender cell type, and receiver cell type loadings. We can interpret the loadings as the relative importance of each element in each dimension of the original tensor. Together, the four vectors in a given factor constitute the CCC programs. The vectors are interconnected such that their combination across dimensions define a CCC program, with loadings in the sample dimension representing the context-dependence of the program and elements from each of the other dimensions (ligand-receptor interactions and cell types) with high loadings being key mediators of this program. By focusing on sample loadings associated with a given condition label, we can thus identify the cell types and interactions also associated with that label. To aid the interpretation of LIANA and Tensor-cell2cell results, we also provide a wide range of visualizations and strategies to summarize the interaction loadings into biologically-meaningful insights. We anticipate that our unified protocol will aid the scientific community in studying CCC using large single-cell datasets with a high number of samples and biological conditions.

4.6 Appendix

4.6.1 Appendix A: Benchmarking Missing Indices in Tensor-cell2cell

To help users make informed decisions regarding choices in their computational pipeline, we benchmarked two key factors that can influence Tensor-cell2cell's outputs, batch correction of expression data and missing tensor indices across samples.

4.6.1.1 Motivation

The purpose of these analyses was to determine how the tensor decomposition used by Tensor-cell2cell handles missing values. Cell-cell communication (CCC) inference tools output the following for each sample: communication scores associated with interactions (i.e., sender and receiver cell type pairs and ligand-receptor (LR) pairs)[2].

Tensor-cell2cell constructs the tensor by concatenating the output of these CCC tools across the context dimension. Consequently, cell types or LR pairs that are not present across all samples will result in missing values (at the "sample - sender cell type - receiver cell type - LR pair" tensor coordinate or index) in the samples they were absent from. Sample-specific "missing" data may be the result of any of the following:

1) Technical limitations in measuring a gene or cell.
2) Computational pipelines (data processing, negative expression counts or communication scores, thresholding parameters of CCC tools, etc.) that result in the exclusion of certain measurements.
3) The cell type or LR pair is absent from that sample due to biological reasons (a "true biological zero").

Tensor-cell2cell's decomposition will handle these missing values differently depending on how they are filled in during tensor construction. Within the tool, this is handled by the "how", "cell_fill", "lr_fill", and "outer_fraction" parameters discussed in the protocols and tutorials. Also note that using LIANA's "return_all" parameter can mitigate the number of missing values due to thresholding parameters in point #2 above.

If the missing values are filled with NaN (the default option in Tensor-cell2cell), Tensor-cell2cell's non-negative canonical polyadic decomposition will "mask" these values during decomposition. Technically, this means that during iteration of the Alternating Least Squares algorithm, masked indices are randomly initialized then updated in each iteration. The masked indices in the full tensor are updated with those imputed from the previous iteration, leading to a new optimization problem and new output set of masked values in each iteration. Conceptually, this is essentially an imputation of missing values. If missing values are filled with a floating point value, they will not be masked and be considered as the actual value they were filled in with (rather than imputed) during the decomposition. For example, filling missing indices with 0 will cause these indices to be treated as "true biological zeroes".

Thus, our goal is to determine the effect that the fraction of missing indices and the value with which they are filled, has on decomposition results. To do so, we simulate CCC across multiple samples and construct a gold-standard tensor with no missing indices. Next, we generate missing interactions in the dataset, and fill these values in the tensor with either NaN or 0. Finally, we compare the similarity of decomposition outputs between the tensor with missing indices and the gold-standard using the CorrIndex metric [https://doi.org/10.1016/j.sigpro.2022.108457]. Our expectations are as follows:

1) If Tensor-cell2cell is appropriately robust to missing indices due to technical limitations or computational pipelines that exclude measurements (Points #1 and 2 above), similarity between the gold-standard tensor decomposition output and that of the tensor with missing indices should

be high even at a high fraction of missing indices (i.e., when filled with NaN and masked, Tensor-cell2cell should be able to accurately impute the data).
2) If Tensor-cell2cell is appropriately sensitive to missing indices because those are truly absent in the sample, similarity between the gold-standard tensor decomposition output and that of the tensor with missing indices should be low as the fraction of missing indices increases (i.e., when filled with 0 and not masked, Tensor-cell2cell should be able to accurately distinguish between the gold-standard tensor and that with true biological zeroes).

4.6.1.2 Results

To determine the effect of missing values on Tensor-cell2cell's output, we created a gold-standard context-dependent CCC simulation with no missing indices, iteratively added more missing values, and compared the similarity (ranging from 0 to 1) between resultant decompositions.

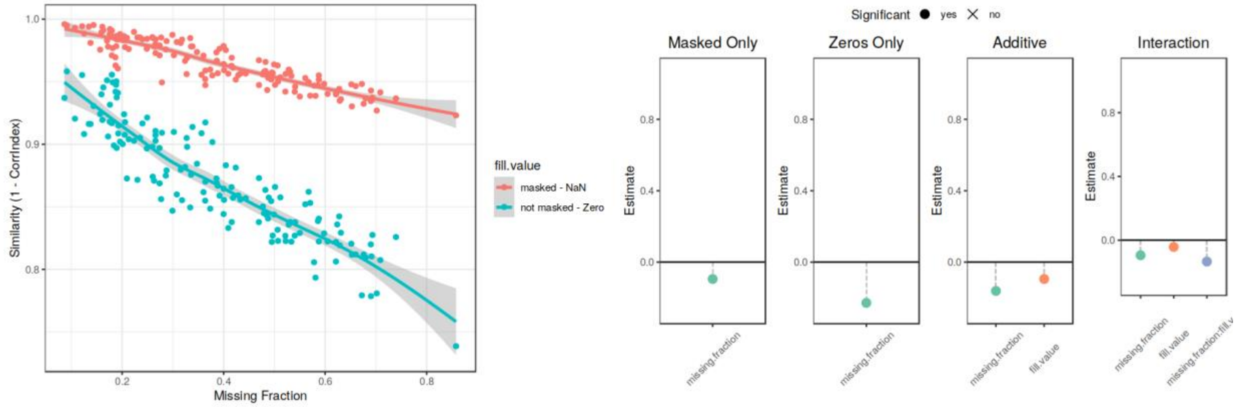


Figure 4.6: Tensor-cell2cell's output decreases as the fraction of missing elements in the tensor increases.
(a) Scatterplot comparing the change in similarity (1 - CorrIndex) (y-axis) as a function of the fraction of missing elements in the tensor (x-axis) for tensors that were filled with NaN and masked (red) or filled with zero and not masked (green). The solid lines show a local polynomial regression and the shaded regions show the 95% confidence interval of this fit. **(b)** Coefficient estimate of a linear regression estimating similarity from the fraction of missing elements in the tensor run only on iterations in which the tensor was masked. **(c)** Coefficient estimate of a linear regression estimating similarity from the fraction of missing elements in the tensor run only on iterations in which the tensor was not masked. **(d)** Coefficient estimate of a multivariate linear regression estimating similarity from the fraction of missing elements in the tensor and whether the tensor was masked (Similarity ~ Missing.Fraction + Fill.Value). **(e)** Coefficient estimate of a multivariate linear regression estimating similarity from the interaction between the fraction of missing elements in the tensor and whether the tensor was masked (Similarity ~ Missing.Fraction * Fill.Value). For all linear regression visualizations, estimates are on the y-axis, coefficients are on the x-axis, the horizontal solid black line represents 0 (no effect), and estimates are significant if the BH FDR <= 0.05. Intercepts are not visualized as all are ~1 and significant.

We found that there was a significant decrease in the similarity of Tensor-cell2cell's output with that of the gold-standard as the fraction of missing indices increased when filling both with NaN (masked) or zero (not masked). However, those that were not masked had a substantially larger decrease in similarity than those that were (Fig. 4.6b-e). Comparing the two filling methods independently, we see that the outputs without masking decreased in similarity at more than twice the rate of those that were masked (Fig. 4.6b-c). Consequently, the similarity for masked outputs at even very high missing indices (>85% of tensor elements missing) remained above 0.9. The lowest similarity for masked values was 0.923, occurring when 85.8% of the tensor elements were missing. In contrast, when filling with zero at the same fraction of missing values, the similarity was 0.739. When considering the two filling methods in combination with the missing fraction, we see that similarity is lower by 0.094 on average when filling with zero (Fig. 4.6d) and the decrease in similarity per unit increase in missing fraction goes from 0.094 to 0.228 when filling with zero rather than NaN (Fig. 4.6e). Altogether, these results indicate that Tensor-cell2cell is robust enough to impute missing values and sensitive enough to handle true biological zeros.

4.6.1.3 Methods

4.6.1.3.1 Single-Cell RNA Data Simulation

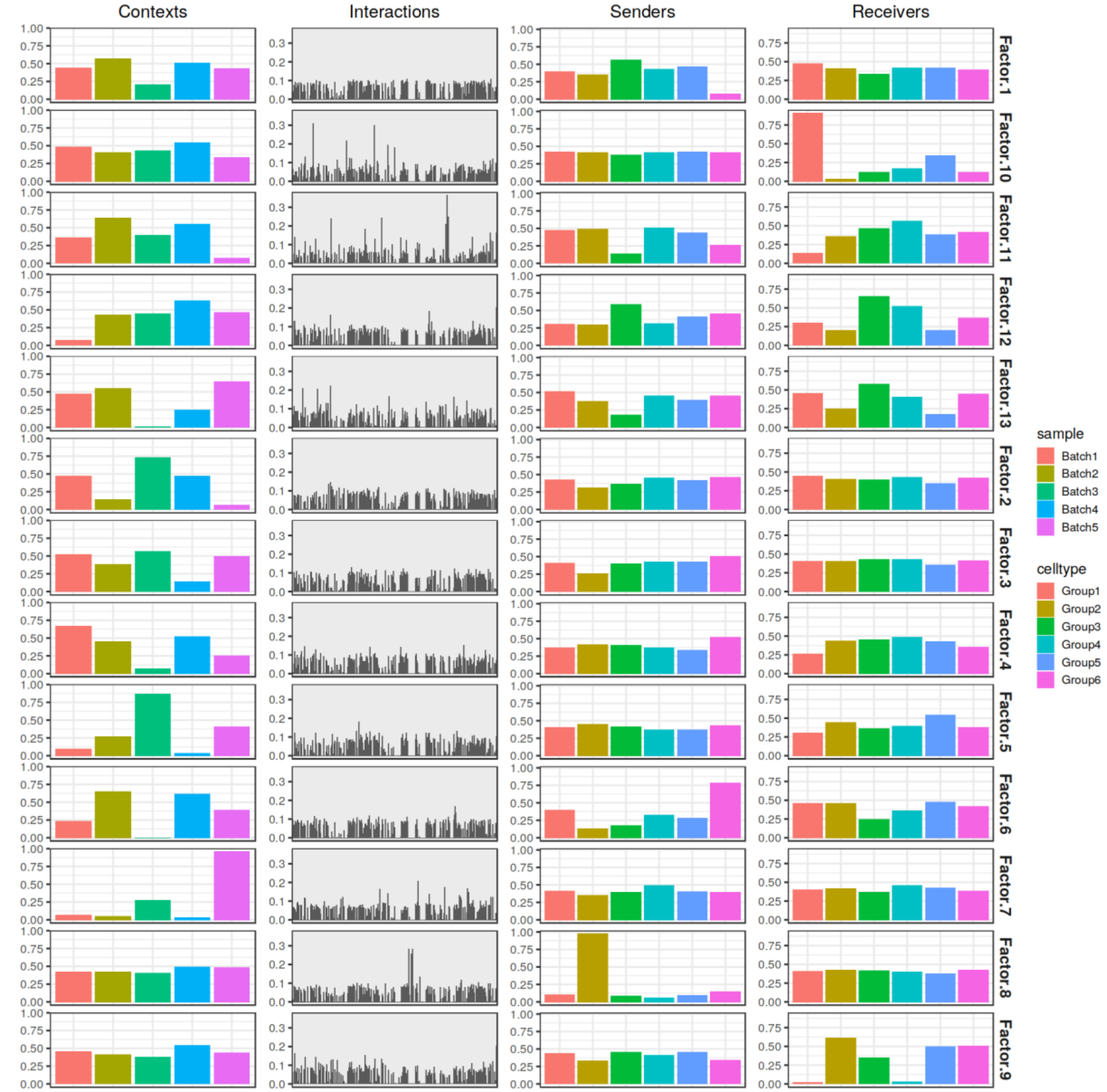


Figure 4.7: The 13 factors estimated by Tensor-cell2cell.

We simulated single-cell RNA-sequencing expression data using Splatter[39], adapting a previously described computational approach[40]. We generate a single-cell expression dataset containing 2000 genes and 5000 cells evenly distributed across 6 cell types and 5 samples. Samples are represented by their respective batch, and we introduce a small batch effect to the dataset by setting the Splatter params “batch.facLoc” and “batch.facScale” both to 0.125. This baseline batch effect ensures that average gene expression values across samples are not exactly the same, such that cell-cell communication is expected to change in a context-dependent manner and Tensor-cell2cell will identify multiple factors (rank > 1). To ensure that differences in cell type dominate over batch-effects we also adjust the Splatter params de.facLoc and de.facScale to be 0.3 and 0.6, respectively.

Next, we do some basic preprocessing of the data for each sample. We apply quality control filters to the cells and genes as implemented previously[40]. Briefly, low-quality cells were identified and disregarded using the scuttle package based on standard metrics (mitochondrial fraction, library size, and number of genes detected); genes detected in fewer than 1% of cells are discarded. Next, counts were normalized using scran pooling[41] and a $\log_{+1}$ transformation.

A random subset of 200 genes were chosen to simulate a ligand-receptor interaction network as previously described[8]. Briefly, we use StabEco's BiGraph function, with the power law exponent value set to 2 and the average degree value set to 3, to generate a scale-free, directed, bipartite network from the 200 genes. Half the genes are assigned to be ligands and the other half to be receptors. Not all genes were part of the connected network (70/200) and were excluded from downstream analyses. This ligand-receptor interaction network represents the custom resource input to LIANA's cell-cell communication scoring.

Finally, to generate missing indices in the 4D-Communication Tensor, we iteratively omit a random subset of genes or cell types from the expression data. Specifically, we iterate through combinations of the following two variables: the fraction of cell types to remove in a given sample (⅙, ⅓, ½, ⅔), the fraction of genes (within the 130 in the simulated LR interaction network) to remove in a given sample (1/10, 3/10, 1/2). We also set the fraction of samples to apply these omissions to (⅕, ⅖, ⅔) and whether the cell types and ligand-receptor pairs omitted should be the same across the samples in which the omission is applied to.

4.6.1.3.2 Communication Scoring

With simulations providing the necessary inputs of a log-normalized expression dataset and a ligand-receptor interaction resource with iteratively more missing values, we then use LIANA to score communication in each sample of each iteration.

To assess samples in a manner independent of the scoring method, we use LIANA's aggregate ranking approach to generate a consensus score across the magnitude scoring types. Thus, we score communication using the methods that output a magnitude score only (CellPhoneDB, SingleCellSignalR, and Connectome/NATMI which both output the same magnitude score). After obtaining the consensus score, we invert them using (1 - score) to give those interactions with more importance a higher value.

We assign the following non-default parameters to the "liana_wrap" function: we set "expr_prop" to 0.05 and "return_all" to True. Decreasing expr_prop from the default value of 0.1 allows for more interactions to be considered within a sample. This decreases the number of interactions that are present but thresholded out, thus allowing the assessment of missing tensor indices to be influenced more by their explicit exclusion. When the return_all parameter is set to True, the interactions that were present in the sample but did not receive a communication score due to thresholding are filled with the worst of the scored interactions. We set this to True because any missing values that were not explicitly simulated, i.e. thresholded out, are true biological zeros. Finally, we pass our simulated LR interaction network as the custom resource.

4.6.1.3.3 Tensor Building and Decomposition

4D-Communication tensors are built from the output of LIANA using the "liana_tensor_c2c" function with default parameters. Similarly, tensors are decomposed using the "decompose_tensor" function with default parameters, except that "tf_optimization" is set to "regular" and "init" is set to 'svd' when estimating the tensor rank. The rank is estimated on the gold-standard tensor using the automated elbow analysis described in the main text, and this rank is used to decompose tensors with omitted values (Fig. 4.7b).

The tensor built from the dataset with omitted values by default has the missing indices masked. The fraction of indices with missing values in the tensor is calculated by taking the total number of masked values and dividing it by the total number of interactions stored in the tensor. To assess the effect of filling these values with true biological zero in addition to NaN, we run the decomposition twice, once with these indices masked and a second time with them unmasked (and the communication scores as these indices being 0).

Finally, both of the tensors generated from the dataset with omitted values are compared to the gold-standard tensor using the CorrIndex as previously described[8]. Briefly, the CorrIndex represents a dissimilarity between decomposition outputs and lies between 0 and 1; we convert this to a similarity metric by using (1-CorrIndex).

4.6.2 Appendix B: Authors, Contributions, and Acknowledgments

Authors: Hratch Baghdassarian*, Daniel Dimitrov*, Erick Armingol*, Julio Saez-Rodriguez, Nathan E. Lewis

*contributed equally to work

H.B., D.D, and E.A. conceived the project, adapted the computational tools, developed the protocol, and wrote the initial version of the manuscript. J.S.R. and N.E.L. revised the manuscript and supervised the project. H.B., D.D, and E.A. contributed equally. J.S.R. and N.E.L are both corresponding authors and have contributed equally.

4.7 References

1. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell-cell communication through single-cell transcriptomics. *Curr Opin Syst Biol* **26,** 12–23 (2021).

2. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22,** 71–88 (2021).

3. Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P. L., Nagai, J. S., Boys, C., Ramirez Flores, R. O., Kim, H., Szalai, B., Costa, I. G., Valdeolivas, A., Dugourd, A. & Saez-Rodriguez, J. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13,** 3224 (2022).

4. Shakiba, N., Jones, R. D., Weiss, R. & Del Vecchio, D. Context-aware synthetic biology by controller design: Engineering the mammalian cell. *Cell Syst* **12,** 561–592 (2021).

5. Mitchel, J., Grace Gordon, M., Perez, R. K., Biederstedt, E., Bueno, R., Ye, C. J. & Kharchenko, P. V. Tensor decomposition reveals coordinated multicellular patterns of transcriptional variation that distinguish and stratify disease individuals. *bioRxiv* 2022.02.16.480703 (2022). doi:10.1101/2022.02.16.480703

6. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nat. Biotechnol.* **40,** 1467–1477 (2022).

7. Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B. & Saez-Rodriguez, J. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. *bioRxiv* 2023.02.23.529642 (2023). doi:10.1101/2023.02.23.529642

8. Armingol, E., Baghdassarian, H. M., Martino, C., Perez-Lopez, A., Aamodt, C., Knight, R. & Lewis, N. E. Context-aware deconvolution of cell-cell communication with Tensor-cell2cell. *Nat. Commun.* **13,** 3665 (2022).

9. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Schiller, H. B. & Theis, F. J. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24,** 550–572 (2023).

10. Kuppe, C., Ramirez Flores, R. O., Li, Z., Hayat, S., Levinson, R. T., Liao, X., Hannani, M. T., Tanevski, J., Wünnemann, F., Nagai, J. S., Halder, M., Schumacher, D., Menzel, S., Schäfer, G., Hoeft, K., Cheng, M., Ziegler, S., Zhang, X., Peisker, F., Kaesler, N., Saritas, T., Xu, Y., Kassner, A., Gummert, J., Morshuis, M., Amrute, J., Veltrop, R. J. A., Boor, P., Klingel, K., Van Laake, L. W., Vink, A., Hoogenboezem, R. M., Bindels, E. M. J., Schurgers, L., Sattler, S., Schapiro, D., Schneider, R. K., Lavine, K., Milting, H., Costa, I. G., Saez-Rodriguez, J. & Kramann, R. Spatial multi-omic map of human myocardial infarction. *Nature* **608,** 766–777 (2022).

11. Alečković, M., Cristea, S., Gil Del Alcazar, C. R., Yan, P., Ding, L., Krop, E. D., Harper, N. W., Rojas Jimenez, E., Lu, D., Gulvady, A. C., Foidart, P., Seehawer, M., Diciaccio, B., Murphy, K. C., Pyrdol, J., Anand, J., Garza, K., Wucherpfennig, K. W., Tamimi, R. M., Michor, F. & Polyak, K. Breast cancer prevention by short-term inhibition of TGFβ signaling. *Nat. Commun.* **13,** 7558 (2022).

12. Tanevski, J., Flores, R. O. R., Gabor, A., Schapiro, D. & Saez-Rodriguez, J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol.* **23,** 97 (2022).

13. Zheng, R., Zhang, Y., Tsuji, T., Gao, X., Wagner, A., Yosef, N., Chen, H., Zhang, L., Tseng, Y.-H. & Chen, K. MEBOCOST: Metabolite-mediated Cell Communication Modeling by Single Cell Transcriptome. *bioRxiv* 2022.05.30.494067 (2022). doi:10.1101/2022.05.30.494067

14. Zhao, W., Johnston, K. G., Ren, H., Xu, X. & Nie, Q. Inferring neuron-neuron communications from single-cell transcriptomics through NeuronChat. *bioRxiv* (2023). doi:10.1101/2023.01.12.523826

15. Armingol, E., Larsen, R. O., Cequeira, M., Baghdassarian, H. & Lewis, N. E. Unraveling the coordinated dynamics of protein- and metabolite-mediated cell-cell communication. *bioRxiv* 2022.11.02.514917 (2022). doi:10.1101/2022.11.02.514917

16. Zhang, Z., Qin, Y., Wang, Y., Li, S. & Hu, X. Integrated analysis of cell-specific gene expression in peripheral blood using ISG15 as a marker of rejection in kidney transplantation. *Front. Immunol.* **14,** 1153940 (2023).

17. Ghaddar, A., Armingol, E., Huynh, C., Gevirtzman, L., Lewis, N. E., Waterston, R. & O'Rourke, E. J. Whole-body gene expression atlas of an adult metazoan. *bioRxiv* 2022.11.06.515345 (2022). doi:10.1101/2022.11.06.515345

18. Liu, Z., Sun, D. & Wang, C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biol.* **23,** 218 (2022).

19. Wang, S., Zheng, H., Choi, J. S., Lee, J. K., Li, X. & Hu, H. A systematic evaluation of the computational tools for ligand-receptor-based cell-cell interaction inference. *Brief. Funct. Genomics* **21,** 339–356 (2022).

20. Nagai, J. S., Leimkühler, N. B., Schaub, M. T., Schneider, R. K. & Costa, I. G. CrossTalkeR: analysis and visualization of ligand-receptorne tworks. *Bioinformatics* **37,** 4263–4265 (2021).

21. Garcia-Alonso, L., Handfield, L.-F., Roberts, K., Nikolakopoulou, K., Fernando, R. C., Gardner, L., Woodhams, B., Arutyunyan, A., Polanski, K., Hoo, R., Sancho-Serra, C., Li, T., Kwakwa, K., Tuck, E., Lorenzi, V., Massalha, H., Prete, M., Kleshchevnikov, V., Tarkowska, A., Porter, T., Mazzeo, C. I., van Dongen, S., Dabrowska, M., Vaskivskyi, V., Mahbubani, K. T., Park, J.-E., Jimenez-Linan, M., Campos, L., Kiselev, V. Y., Lindskog, C., Ayuk, P., Prigmore, E., Stratton, M. R., Saeb-Parsy, K., Moffett, A., Moore, L., Bayraktar, O. A., Teichmann, S. A., Turco, M. Y. & Vento-Tormo, R. Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat. Genet.* **53,** 1698–1711 (2021).

22. Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V. & Nie, Q. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12,** 1088 (2021).

23. Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., Liu, L., Amit, I., Zhang, S. & Zhang, Z. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26,** 842–844 (2020).

24. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15,** 1484–1506 (2020).

25. Raredon, M. S. B., Yang, J., Garritano, J., Wang, M., Kushnir, D., Schupp, J. C., Adams, T. S., Greaney, A. M., Leiby, K. L., Kaminski, N., Kluger, Y., Levchenko, A. & Niklason, L. E. Computation and visualization of cell-cell signaling topologies in single-cell systems data using Connectome. *Sci. Rep.* **12,** 4187 (2022).

26. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11,** 5011 (2020).

27. Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M. & Colinge, J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **48,** e55 (2020).

28. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28,** 573–580 (2012).

29. Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., Korcsmáros, T. & Saez-Rodriguez, J. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17,** e9923 (2021).

30. Noël, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C., Kieffer, Y., Mechta-Grigoriou, F. & Soumelis, V. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.* **12,** 1089 (2021).

31. Shao, X., Liao, J., Li, C., Lu, X., Cheng, J. & Fan, X. CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. *Brief. Bioinform.* **22,** (2021).

32. Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I. J., Vellai, T., Csermely, P. & Korcsmáros, T. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst. Biol.* **7,** 7 (2013).

33. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).

34. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39,** (2023).

35. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49,** D545–D551 (2021).

36. Badia-I-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O. & Saez-Rodriguez, J. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform Adv* **2,** vbac016 (2022).

37. Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M. J., Blüthgen, N. & Saez-Rodriguez, J. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9,** 20 (2018).

38. Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Curr Opin Syst Biol* **15,** 82–90 (2019).

39. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18,** 174 (2017).

40. Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M. & Theis, F. J. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19,** 41–50 (2022).

41. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17,** 75 (2016).

Chapter 5: Human ME-Models Refine Resource Allocation Principles and Extend Prediction of Biological Processes

Genome-scale metabolic models use an optimality framework to provide unique mechanistic insights to intracellular activity based on nutrient and energy resources. Such models provide a knowledge base of metabolism and can predict context-specific intracellular fluxes for specific cell objectives. However, they do not explicitly and comprehensively account for machinery resources, which represent a major cost to the cell. ME-Models have addressed this by integrating metabolic models with expression reactions, but such models have only been developed for prokaryotes. Here, we present humanME, a Python tool to generate and analyze human ME-Models from input metabolic models. We demonstrate that the ME-Model has improved prediction accuracy of growth rate relative to metabolic models. Due to the additional constraints imposed by the ME-Model framework, we also identify more efficient and unique solutions for growth. Finally, we show that transcriptional fluxes can be used as a proxy for transcriptome measurements.

5.1 Introduction

A fundamental goal of systems biology is to quantitatively and accurately characterize how interactions between cellular components give rise to various phenotypes and physiological functions[1]. Integrating omics measurements with computational approaches to address this goal can reveal fundamental insights to biological function[2] and improve cell engineering approaches[3]. Furthermore, computational models that can correctly predict cell and organismal phenotypes reveal the accuracy of model assumptions and relevant incorporated biological processes.

Within systems biology, resource allocation provides a unique lens through which to view such relationships. Under the theoretical framework of resource allocation, cells are optimizing for a specific set of tasks (i.e., objectives) under resource constraints[4]. Cells integrate the information from their environment to determine their biological objective (e.g., growth[5], secretion[6], or cell migration[7]). The availability of resources such as extracellular nutrient, bioenergy, and macromolecular machinery inform pathway activity such that cells can most efficiently complete these objectives. When a cell encounters multiple objectives,because there is a shared pool of limiting resources[8], it undergoes trade-offs. Such trade-offs are typically analyzed using a Pareto analysis[9–11].

Genome-scale models (GEMs) of metabolism (M-Models) have high utility because they implement this optimality framework, consolidate much biological knowledge as a database, and integrate omics data to predict phenotypes[12]. Furthermore, they maintain high-resolution molecular and mechanistic details such that specific reaction fluxes are also simulated. GEMs have provided key insights into wide-ranging biological systems, such as metabolic phenotypes underlying Alzheimer's Disease[13] and polyamine metabolism in T-helper 17 cell pathogenicity.[14] However, while they explicitly account for nutrient and bioenergetic resources, they can only indirectly account for macromolecular machinery (e.g., via context extraction[15] or flux minimization[16,17]).

However, a comprehensive accounting of macromolecular machinery costs is crucial not only for improving model accuracy, but also to extend the mechanistic details of GEMs. This is because macromolecular synthesis costs represent a substantial portion of the overall cell resource budget[18] and are demonstrated to affect cell activity. For example, cells have evolved ribosomal features that maximize auto-catalytic activity[19], demonstrate reduced expression of energetically costly proteins[6], and change their activity to minimize machinery costs[20].

In prokaryotes, genome-scale models of metabolism and expression (ME-Models)[21,22] have explicitly accounted for machinery by incorporating transcription and translation of reaction-catalyzing enzymes[23] and coupling[24] these enzymes to their respective metabolic reactions (Fig. 5.1). ME-Models encompass the entirety of the M-Model (the "metabolic module"). Thus, they can conduct the same type of analyses of an M-Model, but the additional machinery constraints improve some of the issues resulting from the fact that M-models are underdetermined[20]. Because ME-Models further encode transcription and translation reactions (the "expression module"), they uniquely enable variable RNA and protein biomass components and mechanistic insights to gene expression[25,26].

To date, ME-Models have not been built for eukaryotes, largely due to their additional complexities, such as multiple subcompartments, slow growth rates[27,28], and non-growth objectives[20]. Here, we build on the approaches implemented in prokaryotes as well as a mammalian GEM that integrates metabolism with the secretory pathway[6,20] to build human ME-Models. We develop a tool that can take a user-provided, context-extracted human metabolic model as input and output a respective ME-Model. Our tool is implemented in Python and built on the COBRApy framework to enable user-friendly building and analysis of ME-Models. Like their prokaryotic counterparts, these ME-Models enable an exhaustive and direct accounting of machinery costs of metabolic activity. Beyond metabolism, they also simulate gene transcriptional, translational, and transport fluxes. Using the gene expression infrastructure, we've allowed the tool to flexibly produce user-specified non-machinery proteins in any cell compartment of interest, including secretion of proteins to the extracellular matrix. As such, the model can optimize for non-metabolic objectives such as migration.

We use our tool to build a context-specific ME-Model of the K562 cell line from NCI-60. We demonstrate that this ME-Model has improved prediction accuracy of growth rate relative to metabolic models. Due to the additional constraints imposed by the ME-Model framework, we also identify more efficient and unique solutions for growth. Next, we show that transcriptional fluxes can be used as a proxy for transcriptome measurements.

5.2 Results

5.2.1 A Tool for Building and Analyzing Human Genome-Scale Models of Metabolism and Gene Expression

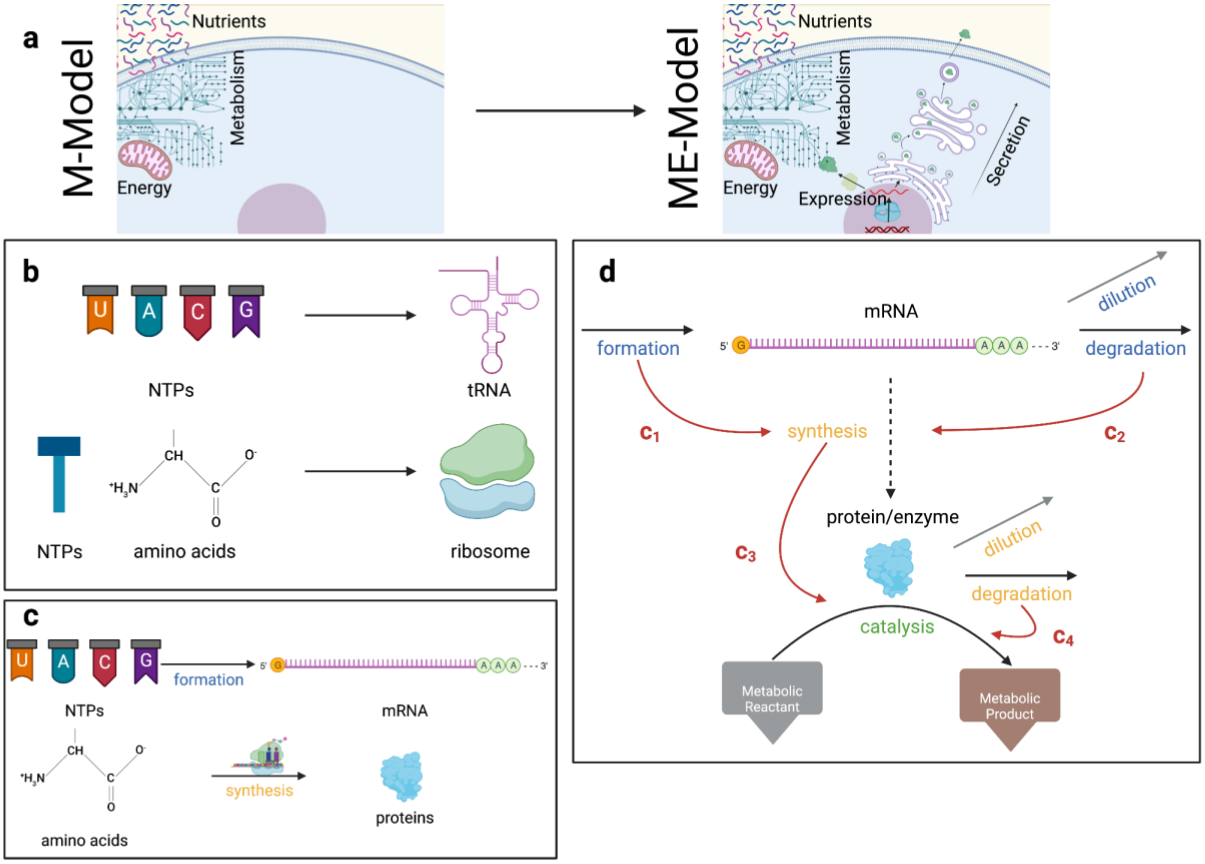


Figure 5.1: Key steps for building a ME-Model.
(a) An input M-model is provided, which can account for the metabolic module, as well as nutrient and energy resources. This input is used to construct a ME-Model, which can additionally account for machinery costs and simulate the synthesis of non-metabolic proteins through the expression module. **(b)** tRNA and ribosomes are synthesized from metabolic precursors. **(c)** Specific mRNAs and proteins are synthesized from metabolic precursors for each gene in the model. **(d)** A simplified reaction network of gene expression used to derive the coupling coefficients, i.e., parameters that link the different processes. Reactions for mRNA (blue captions) are coupled to reactions for protein (orange captions). Reactions for proteins are coupled to the respective metabolic (or expression) reactions that they catalyze (green caption). Black and gray arrows represent the gene expression reactions. The dashed arrow represents the fact that mRNA is not explicitly depleted by translation, but translation is dependent on mRNA concentration. The gray lines for dilution represent the fact that these reactions aren't explicitly modeled for each macromolecule; rather, they are accounted for in the biomass dilution reaction. The red lines indicate which reactions are coupled to each other and are labeled by their respective coupling coefficient.

The Python tool, “humanME”, takes as input a context-specific extracted M-Model, based on Recon2.2[29]. It subsequently returns a corresponding ME-Model with coupled reaction-catalyzing enzymes (Fig. 1a). It does so in 5 key steps as outlined below (for details, see Methods - Building the ME-Model). humanME also can conduct downstream analyses of both the metabolic and expression module of this output. Due to the additional machinery constraints imposed by the ME-Model, it is crucial that an accurately curated M-Model is used as input to avoid issues with feasibility when solving (see Appendix AA). The steps for generating the ME-Model are as follow:

Step 1: Generate expression reactions for ribosome biogenesis and tRNAs (Fig. 5.1b). While these are independent of the inputs, they are necessary for protein synthesis in the subsequent step (Fig. 5.1c).

Step 2: Generate gene-specific expression reactions (transcription, translation, and degradation) for each protein participating in reaction catalysis, using the gene-protein-reaction (GPR) rules specified by the model (Fig. 5.1c). These proteins are also transported to the appropriate subcellular compartment where the reaction is occuring, and form complexes according to the GPR. Note that, because expression reactions themselves are catalyzed by enzymes (auto-catalytic), we recursively generate corresponding expression reactions until no new expression module enzymes are introduced.

Step 3: Couple[24] mRNA to enzyme and enzyme to the metabolism in order to generate flux demand throughout the expression module (Fig. 5.1d, Methods - Reaction Coupling).

Step 4: Re-formulate the M-Model biomass reaction to allow for variable RNA and protein biomass components (see Methods - Formatting the Biomass Reaction).

Step 5: Conduct downstream analyses. This includes, for example, maximizing for growth rate and solving for non-growth objectives (by adapting FBA to include growth rate as a parameter), flux variability analysis (FVA), comparing the metabolic module to the M-Model, and assessing the expression reaction fluxes.

While humanME provides consensus gene features necessary for Steps 1-3 (termed the protein-specific information matrix (PSIM)), users may optionally provide their own PSIM to customize these features. Additionally, the tool can generate expression reactions for proteins that are not catalyzing any reactions in the M-Model, which enables exploration of such processes such as extracellular secretion. Overall, our tool provides a robust framework that allows users to generate a context-specific human ME-Model for their cell type or tissue of interest.

5.2.2 The ME-Model Metabolic Module is More Efficient and Unique Compared to its M-Model Counterpart

To assess the ME-Model, we adapted a previously generated[30] context-specific M-Model of the K-562 cell line to use as input (see Methods - Refining NCI-60 Cell Line M-Model Inputs). While the M-Model predicted a growth rate of 0.0559 hr^{-1}, the additional machinery constraints implemented in the ME-Model reduced the predicted growth rate 0.0266 hr^{-1}. While both models were within an order of magnitude of the reported experimental growth rate of 0.0354 hr^{-1} (https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm), the ME-Model demonstrated slightly more higher prediction accuracy according to percent change and fold-change metrics (Table 5.1).

Table 5.1: Relative discrepancy between predicted and experimental growth rate for K-562.

	M-Model	**ME-Model**
Percent Change (%)	58.041	-24.679
log$_2$-Fold-Change	0.659	-0.412

Next, we explored the differences between the ME-Model's metabolic module and the M-Model. First, we tested if the additional machinery constraints resulted in a more efficient solution, as quantified by the total absolute flux through all metabolic reactions (Fig. 5.2a). We expected that the additional biosynthetic costs of gene expression will force the ME-Model to use its enzymes as efficiently as possible, thus reducing the flux through metabolic reactions unless absolutely necessary to achieve its objective. This expectation is analogous to the conceptual motivation for parsimonious flux balance analysis (pFBA)[16,17] and the Max-Min Driving Force (MDF)[31]. To make this comparison fair, we assessed the efficiency of both models at the same growth rate (the maximum of the ME-Model, 0.0266 hr^{-1}). We found that the ME-Model is ~3 orders of magnitude lower than the FBA solution for the metabolic model (blue dashed line, Fig. 5.2a), with total fluxes of 36.497 mmol/gDW_{cell}/hr and 37,4034.383 mmol/gDW_{cell}/hr, respectively. This difference is significant given that it lies far outside the solution space of the M-Model (blue kernel density estimate, Fig. 5.2a).

We reasoned that, in part, this reduced flux is due to the elimination of thermodynamically infeasible cycles (type III loops[32]). Thus, we ran the M-Model FBA solution through the CycleFreeFlux algorithm[33] to eliminate thermodynamically infeasible loops from this solution. We found that this resulted in a much more comparable efficiency. Interestingly, the efficiency of the ME-Model lies between that of pFBA (16.313 mmol/gDW_{cell}/hr) and CycleFreeFlux (36.497 mmol/gDW_{cell}/hr). This indicates that beyond eliminating type III loops, the ME-Model is also accounting for the minimization of machinery costs. Given that pFBA also minimizes the flux through spontaneous reactions, it makes sense that the ME-Model total flux is higher than pFBA.

Finally, given that the underdetermined nature of M-Models[20] results in a large solution space[34], we asked whether the addition of machinery constraints could help identify a more constrained solution. Thus, we conducted flux variability analysis (FVA) of M-Model and the ME-Model at each of their respective maximum growth rates. Due to computational limitations in solving the ME-Model, we ran FVA for reactions from three selected pathways representing different network topologies: a linear pathway (glycolysis), a cyclic pathway (the TCA cycle), and a branching pathway (purine biosynthesis). Comparing the flux ranges between the two models, it is apparent that the ME-Model solution is substantially more unique than the M-Model across all reactions (Fig. 5.2b).

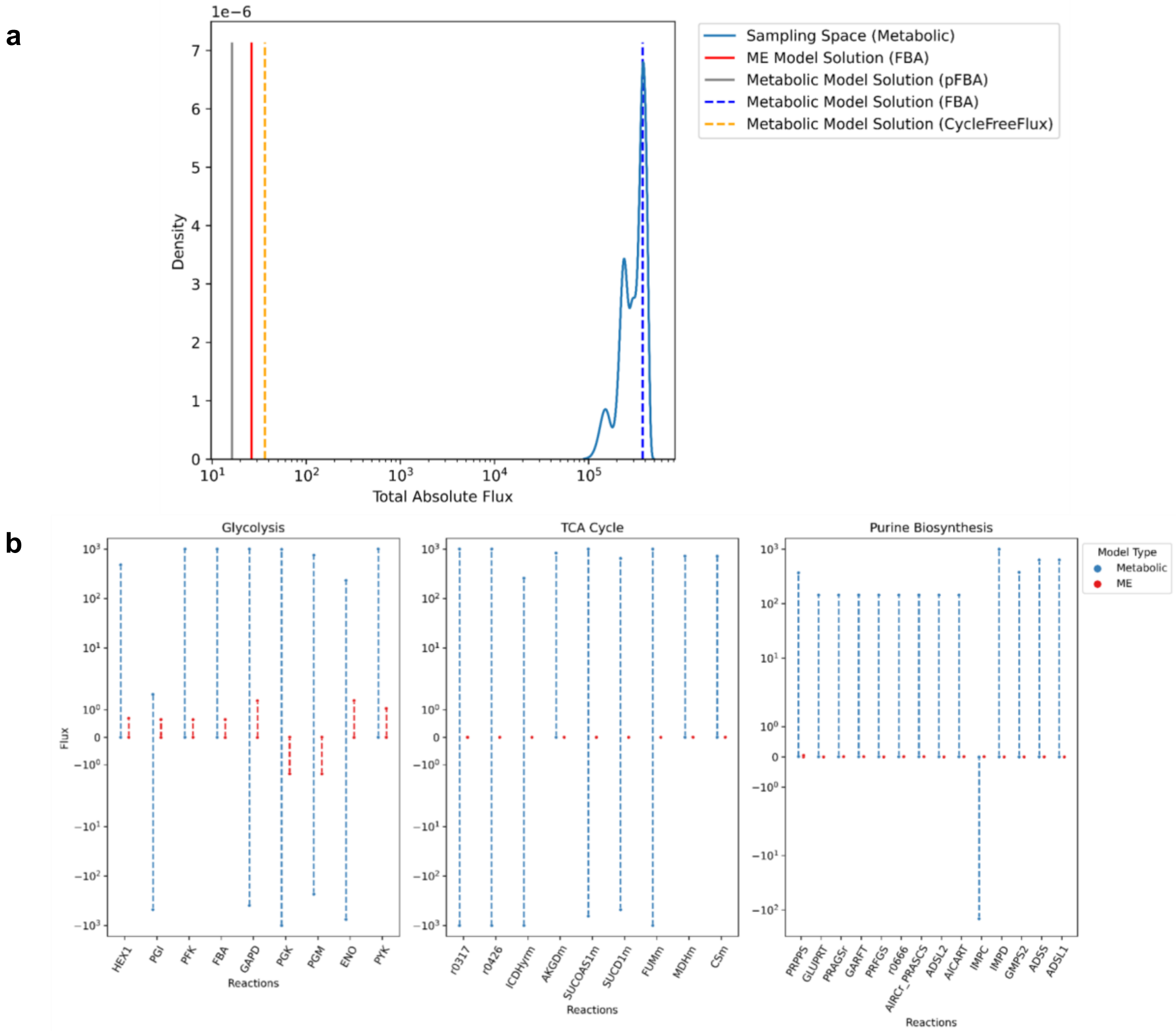


Figure 5.2: Comparison of the ME-Model's metabolic module with the M-Model.
(a) Efficiency of the flux balance analysis (FBA) solution at a growth rate of 0.0266 hr^{-1}, quantified by the total absolute flux (x-axis). The kernel density estimate (y-axis) of the M-Model from sampling the solution space (blue distribution) is shown across 1000 samples. Vertical lines represent solutions for FBA of the M-Model (blue dashed line), FBA output of the M-Model run through the CycleFreeFlux[33] algorithm (orange dashed line), FBA of the ME-Model (red line), and pFBA[16,17] of the M-Model (gray line). **(b)** Flux variability analysis (FVA) of the ME-Model (red) and the M-Model (blue) for reactions in glycolysis (left panel), the TCA cycle (middle panel), and purine biosynthesis (right panel). Individual reactions (x-axis) are plotted against minimum and maximum fluxes (y-axis, symlog scale) that maintain the maximum growth rate identified by each model.

5.2.3 The ME-Model Predicts Transcriptional Fluxes Indicative of Enzyme-Mediated Bottlenecks of Oxidative Phosphorylation

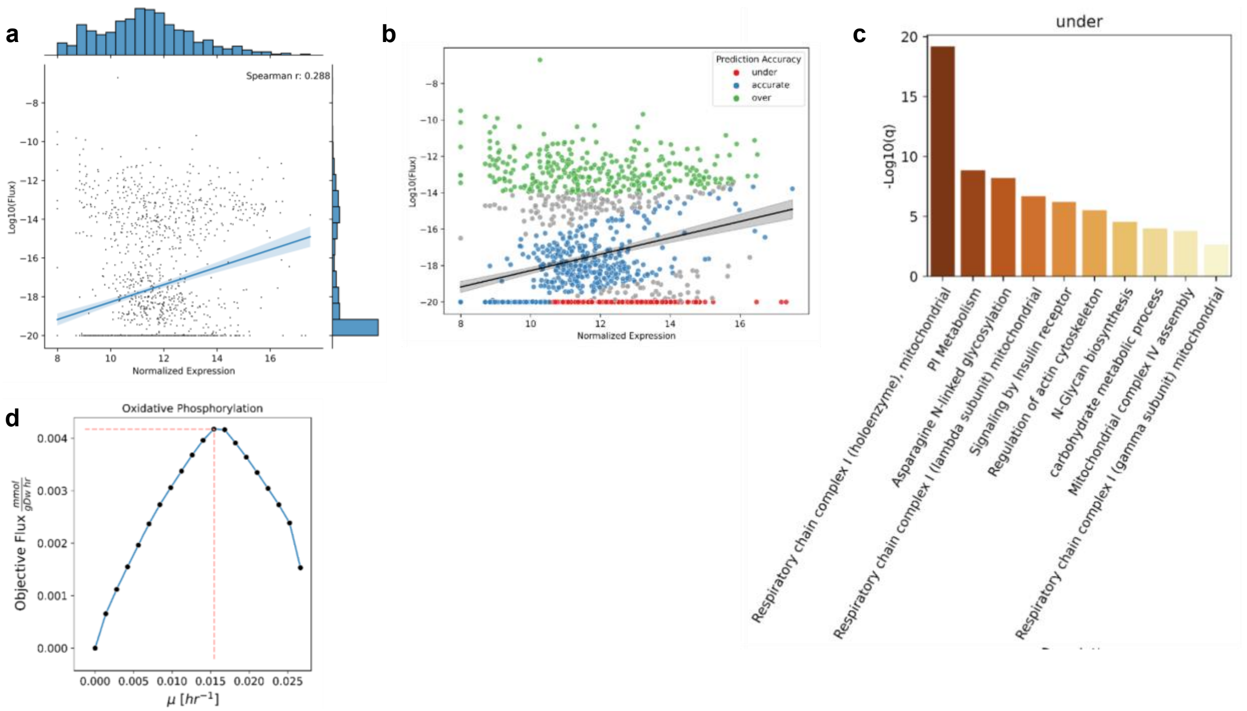


Figure 5.3: Comparison of simulated transcriptional fluxes with transcriptomic abundance. **(a)** Scatterplot of RNA-Seq abundance and $\log_{10}$(mRNA flux). Blue line is an ordinary least squares regression with 95% confidence interval. Histograms for either variable are displayed. **(b)** Same as (a), but colored by over- (Pearson residual ≥ 0.8 and $\log_{10}$(flux) ≥ -14, green), accurately- (absolute value of regression Pearson residual is ≤ 0.7, blue), and under- (Pearson residual ≤ 0.8 and flux ~ 0, red) predicted genes. **(c)** Top 10 terms (x-axis) as indicated by -$\log_{10}$(q-value) (y-axis) from over-representation analysis of under-predicted genes. **(d)** Line-plot of maximum flux value for Oxidative Phosphorylation (y-axis) at each corresponding growth rate ranging from 0 to the maximum feasible growth rate (0.0266 hr^{-1}). Red dashed line identifies the growth rate at which Oxidative Phosphorylation is maximized.

Because the ME-Model simulates gene expression fluxes, using the same K-562 ME-Model, we asked the extent to which simulated mRNA transcriptional fluxes could serve as a proxy for relative macromolecular abundance. For each gene expressed by the ME-Model, we retrieved the net mRNA flux by taking the flux for mRNA formation (i.e., transcription, processing, and export) and subtracting the flux for mRNA degradation. Comparing these simulated fluxes to transcriptomic abundance (“Supplementary Data 1” from ref[35]), we found a Spearman correlation of 0.288 (Fig. 5.3a). While this demonstrates a moderately positive effect size, we found that ~⅓ of the genes are expressed but have little to no flux (Fig. 5.3b) ("under-predicted"), decreasing the overall correlation.

Since our ME-Model solution was maximizing growth, we thought that these under-predicted genes may be related to non-growth associated tasks. However, upon running an over-representation analysis through Metascape[36], we found that four of the ten top enriched terms were related to the electron transport chain (ETC) (Fig. 5.3c). However, when setting the objective to reactions associated with ETC, the correlation values did not improve substantially (data not shown). Alternative reasons for the model inaccuracy of under-predicted genes could include inaccuracies in the transcriptomic data or in the model assumptions. With regards to transcriptomic data, this may include 1) noise in measurements for this pathway or 2) mRNA abundance not serving as a good proxy for enzymatic activity (e.g., due to PTMs). With regards to the model, it may 1) not be accounting for cell hedging for other objectives or 2) have lenient coupling constraints reducing expression demand. These are all potential avenues to explore in the future.

Of note, when testing the various ETC objectives, we consistently observed that their optimal values changed as a parabolic function of growth rate, increasing up to a growth rate of ~0.015 hr^{-1} and decreasing again through to the maximum growth rate (Fig. 5.3d). Given that the model was set up with unlimited oxygen transport and no forced glucose uptake or lactic acid secretion, this indicates that machinery constraints alone can emulate the Warburg effect, and is also worthy of future investigation.

5.3 Discussion

Here we present humanME, a Python tool that enables one to build ME-Models from context-specific human metabolic models. We demonstrate that, when compared to the M-Model, the additional machinery constraints imposed in the ME-Model find more efficient solutions that eliminate thermodynamically infeasible loops and that they reduce the solution space. We also demonstrate that, with additional fine-tuning, simulated gene expression fluxes may serve as a good proxy for macromolecular abundance.

The ME-Model has limitations that may be addressed in the future. These include limitations in the metabolic models that it builds on, such as the fact that these models do not incorporate regulatory networks and, when using FBA, cannot account for dynamic contexts in which metabolite concentrations are not at steady-state. Furthermore, because the ME-Model adds many reactions, solving times begin to get lengthy, particularly when trying to run whole-model analyses such as FVA of all reactions. Finally, in developing the gene expression reactions, a number of assumptions were made (e.g., 1:1 stoichiometric ratios of enzyme complex subunits) that can be refined to improve prediction accuracy.

Altogether, the ME-Model represents an advance towards whole-cell modeling in mammalian cells. In the future, this can be used to explore detailed mechanisms outside of metabolism, e.g., the relationship between gene expression and phenotypes such as growth rate[5] or expanding the proteostasis network to explore protein folding in different contexts[37]. The ME-Model could also serve as a building block to multicellular tissue modeling, by adapting methods used for multicellular M-Models[13,38] or integrating with other modeling approaches such as those for deciphering cell-cell interactions[39].

5.4 Methods

5.4.1 Building the ME-Model

5.4.1.1 Protein-Specific Information Matrix (PSIM)

We construct a gold-standard PSIM containing all the gene features needed for the expression of genes in Recon2.2 and expression module machinery (Table 5.2). While we include other genes (e.g., for expression of non-machinery), we only ensure completeness of features for machinery genes. For sequence features, cDNA and protein sequences were taken from MANE Select[40]. In instances where MANE did not contain information for a given gene, Refseq Select[41] was used in its place. Refseq Select is a superset of MANE Select. Refseq and MANE Select together gave specific isoform protein and transcript sequences for 18495 unique genes. Of the 2353 combined genes in the metabolic and expression module, 78 were not covered by this. For these remaining 78, we use APPRIS[42]. We choose the isoform with the highest "PRINCIPAL" rating. The number of exons and gene sequences were downloaded from the ENSEMBL rest API. pre-mRNA and mRNA sequences were obtained by transcribing the gene and cDNA sequences respectively. Details on other features in Table 5.2 are provided in corresponding subsections within "Building the ME-Model" in which those features are used.

Table 5.2: Gene features in the PSIM.

Name	**Description**	**Default Value (if not present)**	**Source**
HGNC_ID[a]	The gene ID in HGNC format (HGNC:####). There should be an entry for all genes that are included in the M_Model GPR and in non-machinery.		
PREMRNA_SEQ[b]	The gene pre-mRNA sequence.	Technically none, but preprocess.correct_inputs.correct_psim will fill incorrect values with the gold-standard PSIM values. Requirements include that values can only include 'A', 'C', 'G', 'U', the sequence length must be >= mrna sequence length, and the sequence length must be >= 3*protein sequence length.	[40–42]
MRNA_SEQ[b]	The gene mRNA sequence (isoform specific).	Technically none, but preprocess.correct_inputs.correct_psim will fill incorrect values with the gold-standard PSIM values.	[40–42]
PROTEIN_SEQ[b]	The gene protein sequence (isoform specific).	Technically none, but preprocess.correct_inputs.correct_psim will fill incorrect values with the gold-standard PSIM values. Requirements include values can only include one-letter amino-acid codes and the sequence length <= (mrna sequence length/3)	[40–42]
POLYA_LENGTH[c]	The length of the mature mRNA polyA tail.	If not provided, randomly draws from a johnsonsu distribution	[43]
N_EXONS[c]	The number of exons in the premrna (isoform specific). Use to estimate the number of introns (as # of exons - 1).	Estimated as (pre-mRNA sequence length)/6700	
TMD[d]	The number of transmembrane domains contained in the sequence.	0	[6]
DSB[d]	The number of disulfide bonds in the protein.	0	[6]

Table 5.2: Gene features in the PSIM.

Name	Description	Default Value (if not present)	Source
GPI[d]	Whether a GPI anchor is present in the protein. 0 if not present, 1 otherwise	0	[6]
OG[d]	The number of utilized O-linked glycosylation sites in the protein.	0	[6]
ALPHA_M[c]	The mRNA degradation/turnover rate (hrs^{-1}). Used in calculating coupling constraints.	0.06 hrs^{-1} (median value from Source)	[44]
ALPHA_P[d]	The protein degradation/turnover rate (hrs^{-1}). Used in calculating coupling constraints.	0.02 hrs^{-1} (median value from Source)	[45,46]
PTR[c]	A gene-specific, tissue-independent protein to RNA ratio. Used in calculating the coupling constraints.	65163 (median value from Source)	[47]
LOCATION	The final location of the protein. Required for non-machinery, disregarded for machinery (pipeline infers location from the reaction compartments).		

a: default values unavailable, must be provided by user (or from gold-standard PSIM)
b: default values unavailable, but if not provided or incorrect, will fill in with the gold-standard PSIM values when available (otherwise will error out)
c: standard default values used when not provided
d: these are only used for proteins that will be processed via the secretory pathway (ER, Golgi, extracellular membrane, plasma membrane, lysosome).

We also include additional information in the gold-standard PSIM that is not used for building the ME-Model. This includes a “Machinery” column indicating whether a protein is considered machinery according to the full Recon2.2 ('Metabolic'), the GPRs for expression reactions ('Expression'), both ('Both'), or neither ('Non-Machinery') and “Source” indicating which database the isoform sequences were attained.

5.4.1.2 Preprocessing Inputs

Two inputs to the ME-Model building are preprocessed. The first is an M-Model obtained from a context-extraction of Recon2.2 (i.e., it should be a subset of Recon2.2), which is a required input. The second is the PSIM, which defaults to the gold-standard if not provided by the user.

For the M-Model, first we check whether any reaction GPRs have the same enzyme repeated more than once. If they do, we remove them. In Recon2.2, this is the case for the following reactions: ACCOAC, OIVD1m, OIVD2m, OIVD3m, PFK, PI5P3K. We also ensure that exchange reactions are formatted as in Recon2.2, meaning that they are spontaneous and introduce metabolites to the model in two steps: 1) bounded input flux into the boundary compartment, and 2) an unbounded, reversible flux from the boundary compartment to the extracellular matrix. If present, we remove the gene “HGNC:4686”, which is involved in catalysis of the reaction “GUACYC”, because it is a pseudogene with no affiliated protein sequence. If present, we re-format genes with the format “HGNC:HGNC:#” to “HGNC:#”. In Recon2.2, genes formatted this way are HGNC:HGNC:987 and HGNC:HGNC:2898. Next, there is a set of required metabolites for generating the gene expression reactions. If the M-Model does not contain these metabolites, first, we check whether it is present in another compartment and if it is, we add a transport reaction. If it’s not present at all in the M-Model, we add it to the model using a sink reaction. Note that this results in the metabolite being generated at no resource cost to the model. We also add transport of hydrogen between the nucleus and cytosol, which is not present in Recon2.2. Finally, since the biomass reaction has to be reformatted (see Methods - Formatting the Biomass Reaction), we remove the biomass reaction as well as the biomass component formation reactions.

For a user-provided PSIM, preprocessing will ensure that the gene sequence features are present. The gold-standard PSIM is used to fill in missing values as well as any missing machinery genes. For other features, default values (Table 5.2) are used to fill in missing values.

5.4.1.3 tRNA Expression

Currently, tRNA reactions and molecules are constant. However, we note that humanME is constructed in a manner to accommodate custom mature tRNA, 5' leader 3' trailer, and intronic sequences[48]. The consensus tRNA sequence length was set to 72 bps for the mature tRNA, 6 bps for the 5' leader, and 9 bps for the 3' trailer, the most frequent values from “Table S4” of ref[49]. From this table, a position-independent consensus sequence was obtained from the frequency of each base in its relative position (i.e., normalized to total sequence length).

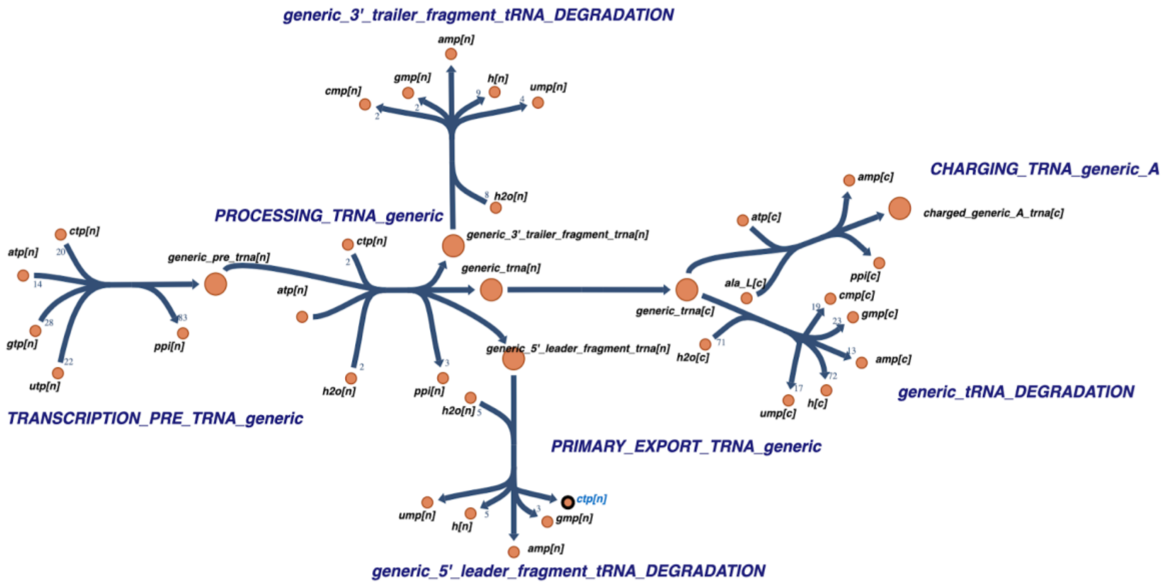


Figure 5.4: Reaction network for mature tRNA synthesis.

Reactions for tRNA synthesis[50,51] (Fig. 5.4) include nuclear pre-tRNA synthesis by RNAP3[52], nuclear tRNA processing (e.g., 5' leader cleavage by RNase P[53], 3' trailer cleavage by RNase Z[54], addition of CCA[55], and splicing[56]), export to the cytosol [57], degradation of mature cytosolic tRNA[51], and tRNA charging[58].

5.4.1.4 Ribosome Biogenesis

Ribosomal proteins are expressed[59] according to the gold-standard PSIM as all other genes (see Methods - Gene Expression). RPL40 and RPS27A also have cleavage of ubiquitin by UCHL3[60].

rRNA sequences are obtained from NCBI. rRNA endonucleolytic cut sites are identified from ref[61,62] and scaled according to the NCBI sequence lengths. 5S rRNA is transcribed[63], processed[59,64], complexed with RPL5 and RPL11, and exported to the nucleus[65]. 18s, 5.8S, and 47S rRNA expression[61,62,64] includes 47S transcription[66] processing of intermediate pre-rRNAs.

pre-RNA processing occurs alongside formation of ribonucleoprotein complexes [64,67,6862,64,67–70] and nucleocytosolic export[71], ultimately resulting in 60S and 40S subunits that are joined[72] to form an active ribosome. Complex formation steps are irreversible, but the final active ribosome in the cytosol can dissociate into its individual subunits. rRNAs can be degraded in the cytosol by the exosome upon ribosome complex dissociation.

5.4.1.5 Gene Expression: RNA

Gene expression reactions to produce mRNA and proteins that are involved in catalyzing each model reaction are generated. First, nuclear pre-mRNAs are elongated from nuclear ribonucleoside triphosphates. Next, pre-mRNAs are processed, including addition of a 5' cap[73], 3' polyA tail[74,75], and splicing.

polyA tail lengths were obtained from the HeLA and iPSC organoid values in "Dataset 3" of ref[43]. Taking the average between these datasets, we fit a number of distributions to the lengths and found that the "johnsonsu" distribution had the best fit (lowest SSE and only distribution with a KS test p-value > 0.05). If the polyA tail length is not provided in the PSIM, it is randomly drawn from the fitted distribution. If it is provided, the polyA tail will be randomly drawn from a normal distribution, with the mean as the provided value and the standard deviation calculated as described by an ordinary-least squares regression. Specifically, to get an expected standard deviation at a given length, we fit the regression predicting the standard deviation from the mean of polyA tails (both of these values are provided in "Dataset 3" for each gene). While the polyA tail is degraded over lifetime, we assume it to be degraded in one single reaction.

In instances where the number of exons is provided by the PSIM, the number of introns is estimated to be the number of exons - 1 because there is an average of 1 fewer introns than exons per gene[76,77]. Otherwise, we estimate the number of introns as a function of the pre-mRNA sequence length (1 intron per 6.7 kbp[77]). The number of introns is used to calculate the total ATPs hydrolyzed during splicing, at a rate of 10 ATP per intron[18]. Lariats are degraded with the number of phosphodiester hydrolyzation events equal to the number of introns.

Next, processed products are exported to the cytosol via TREX[78–81]. Since transcription, processing, and export reactions are linear, there is an option to merge them into one reaction to reduce the total number of reactions and reduce solving time. Finally, mRNA is degraded[82] to ribonucleoside monophosphates in the 5' to 3' direction, including decapping by Nudix[83]. We note that humanME has the 3' to 5' degradation mechanisms implemented as well, but defaults to the 5' to 3' direction to have a single degradation reaction to use in the coupling (see Methods - Reaction Coupling).

5.4.1.6 Gene Expression: Protein

Once cytosolic mRNA is produced, the protein expression reactions depend on the final location of the protein. Proteins are transported to the final location wherein the reaction occurs. If the reaction occurs across more than one compartment, one is selected. If the extracellular matrix is one of those compartments, the protein location is the plasma membrane. Otherwise, if the cytosol is one of those compartments, it is disregarded. Next, the compartment with the most metabolites is assigned as the final location. If there is a tie, one is randomly selected. We categorize those proteins with final locations in the cytosol, nucleus, mitochondrial matrix, mitochondrial intermembrane space, and peroxisome as "Cytosolic Transport" (i.e., they are transported to those compartments directly from the cytosol[84]). We categorize those proteins with final locations in the endoplasmic reticulum (ER), Golgi apparatus, lysosome, plasma membrane, and extracellular matrix as "Secretory Transport" (i.e., they are transported to those compartments via the secretory pathway).

For proteins that undergo Cytosolic Transport, they are translated[85] in the cytosol. For proteins that undergo Secretory Transport, they undergo co-translational translocation. We estimate 1 GTP hydrolysis event per amino acid during translation[85]. Proteins destined for the cytosol, peroxisome[86,87], and nucleus that are longer than 100 amino acids[88] undergo irreversible post-translational folding by HSP70 and HSP40[89–91] in the cytosol. Proteins destined for the mitochondria undergo folding during transport. They are first translocated to the mitochondrial matrix[92], and those destined for the intermembrane space are further transported via the OXA complex[93]. Proteins destined for the nucleus can either undergo passive diffusion (if < 40kDa[84]) or classical nuclear import[94].

Cytosolic proteins undergo degradation via the ubiquitin-proteasome pathway (Fig. 5.5), since this accounts for 70% of protein degradation[95]. Ubiquitin is expressed like other cytosolic proteins for UBC and UB genes, without folding and converted to monomers using the deubiquitinase USP5[96]. Proteins targeted for degradation are polyubiquitinated by 4 monomers via the E1-E3 ligases. 1 ATP is consumed per monomer[97]. Next, proteins are degraded by the 26S proteasome[98], with 2 ATP hydrolysis events per amino acid[97]. This ATP hydrolysis rate is also assumed for all other degradation and translocation reactions unless otherwise specified. Proteins destined for the cytosol, peroxisome, ER and Golgi via retro-translocation and ERAD, and nuclear proteins that can undergo passive nucleocytoplasmic diffusion can be degraded by this mechanism. Nuclear proteins can otherwise be degraded by a nuclear proteasome. Mitochondrial matrix and intermembrane proteins are degraded by LON protease[99], with 2 ATP hydrolyzed per amino acid, and i-AAA protease[100], respectively. Peroxisomal proteins can also be degraded directly in the peroxisome via LONP2[86], with 2 ATP hydrolyzed per amino acid.

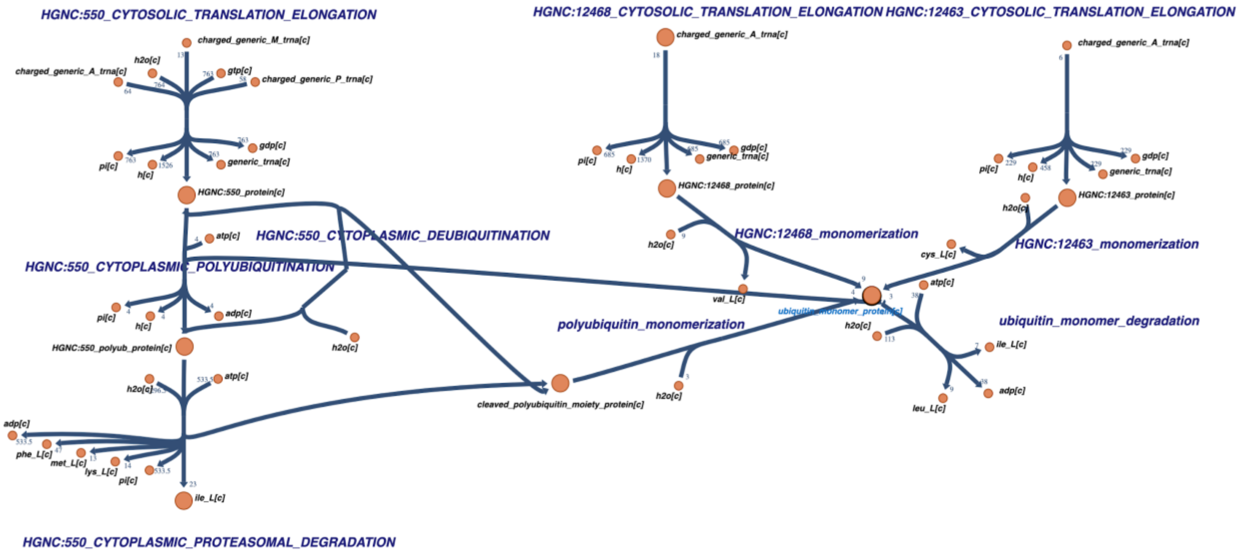

HGNC:550_CYTOSOLIC_TRANSLATION_ELONGATION
charged_generic_M_trna[c]
h2o[c]
charged_generic_A_trna[c]
gtp[c]
charged_generic_P_trna[c]
gdp[c]
pi[c]
h[c]
generic_trna[c]
HGNC:550_protein[c]
atp[c]
HGNC:550_CYTOPLASMIC_DEUBIQUITINATION
HGNC:550_CYTOPLASMIC_POLYUBIQUITINATION
pi[c]
h[c]
adp[c]
HGNC:550_polyub_protein[c]
h2o[c]
h2o[c]
atp[c]
adp[c]
phe_L[c]
met_L[c]
lys_L[c]
pi[c]
ile_L[c]
HGNC:550_CYTOPLASMIC_PROTEASOMAL_DEGRADATION
cleaved_polyubiquitin_moiety_protein[c]
HGNC:12468_CYTOSOLIC_TRANSLATION_ELONGATION
charged_generic_A_trna[c]
pi[c]
h[c]
gdp[c]
generic_trna[c]
HGNC:12468_protein[c]
h2o[c]
HGNC:12468_monomerization
val_L[c]
polyubiquitin_monomerization
h2o[c]
ubiquitin_monomer_protein[c]
h2o[c]
HGNC:12463_CYTOSOLIC_TRANSLATION_ELONGATION
charged_generic_A_trna[c]
pi[c]
h[c]
gdp[c]
generic_trna[c]
HGNC:12463_protein[c]
h2o[c]
cys_L[c]
HGNC:12463_monomerization
atp[c]
ubiquitin_monomer_degradation
ile_L[c]
adp[c]
leu_L[c]

Figure 5.5: Reaction network for cytosolic protein degradation.

Transport reactions for proteins that undergo secretory transport are adapted from ref[6], with the addition of degradation reactions. Proteins that are destined for the Golgi and ER undergo degradation via ERAD[101,102]. They are irreversibly "misfolded", retro-ranslocated, and then undergo cytosolic degradation. Plasma membrane proteins and lysosomal proteins undergo lysosomal degradation via cathepsins. Plasma membrane proteins are targeted to the lysosome by first undergoing ubiquitination[103] and then endocytosis[104].

Once transported to their final location, proteins that are subunits of a complex (i.e., GPRs that contain "AND" boolean logic) are joined via non-covalent interactions formed in a spontaneous reaction. The reversibility of complex formation reactions is a user-provided input, defaulting to being reversible. The reasoning behind making this reversible is for the ME-Model to be more resource efficient, as the monomeric subunits can then individually be degraded or, if they participate in other reactions as part of other enzymes, they can be re-used without having to be completely synthesized again. The individual subunits and the complexes have associated degradation reactions. For enzyme complexes, any parameters used in coupling (see Methods -

Reaction Coupling) are taken as the median across all subunits in the complex. Currently, since GPRs do not specify complex stoichiometry, complexes are assumed to form at a 1:1 ratio; however, humanME is set up to accept specified subunit stoichiometries in the future.

5.4.1.7 Reaction Coupling

After creating the gene expression reactions, we must use reaction coupling to create demand for gene expression reaction fluxes. mRNA reactions must be coupled to protein reactions, and protein reactions must be coupled to metabolic reactions (Fig. 5.1d). Coupling constraints not only provide a way to link biological layers within the GEM framework, but they can also improve accuracy of models by explicitly linking dependent biological processes.

Let's say reaction 1 looks like this: $A + B \rightarrow C$, and reaction 2 looks like this: $D \rightarrow E + F$.

In general the flux through reaction 1, v_1, can be coupled to the flux through reaction 2, v_2 such that $v_1 = cv_2$, where c is "coupling coefficient." This can be done in one of two methods. The first method combines the two reactions with the coupling coefficient: $A + B + cD \rightarrow C + cE + cF$. The second method is to use one of the reaction products (or introduce a "proxy" metabolite product with no mass) in reaction 1 as a substrate in reaction 2, scaled by the coupling coefficient. Reaction 1 would be $A + B \rightarrow C + P$, where P is the proxy metabolite. Reaction 2 would be $cP + D \rightarrow E + F$. In humanME, we always use method 2.

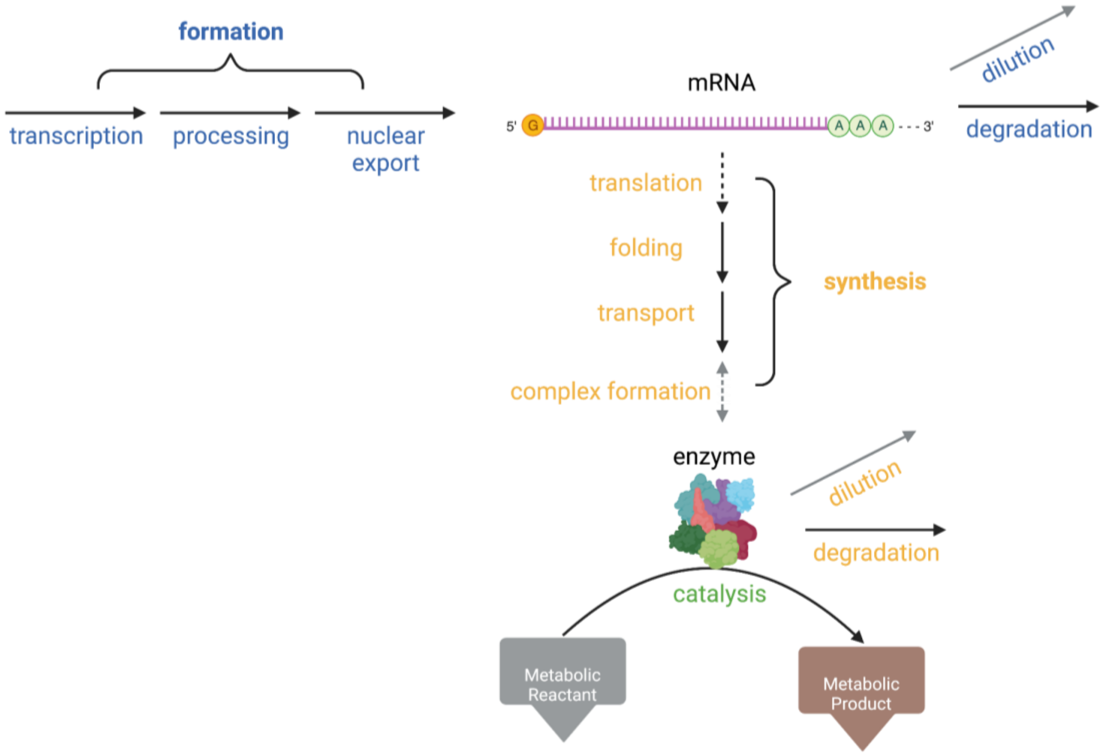


Figure 5.6: The expression reaction network for a given gene in the ME-Model.
All labels as described in Fig. 5.1. Brackets represent a series of typically linear reactions that are collapsed into a single reaction for simplification during coupling derivation. The gray dashed line indicates that complex formation only exists when the GPR contains "AND" boolean logic, otherwise the final monomeric protein product that is transported to its final location is the enzyme.

To derive coupling coefficients, we first simplify the reaction network (from that shown in Fig. 5.6 to that shown in Fig. 5.1d). This simplification merges linear series of reactions and allows for a simpler derivation of coupling constraints under steady-state. Variations of this reaction network (Fig. 5.1d) are commonly used to characterize genome-scale central dogma rates [27,105–107].

Using this network, we derive two coupling coefficients that couple the mRNA layer to the protein layer(Fig. 5.1d). Specifically, c_1 couples mRNA formation to protein synthesis and c_2 couples mRNA degradation to protein synthesis. Similarly, we derive two coupling coefficients that couple the protein layer to the respective reactions that they catalyze. Specifically, c_3 couples enzyme formation to the catalyzed reaction and c_4 couples enzyme degradation to the catalyzed reaction. Note that c_4 is not implemented in prokaryotic ME-Models that assume growth rates are much larger than protein degradation rates. Reaction GPRs are parsed to identify isozymes ("OR" boolean logic). During building of the ME-Model, each isozyme is separately expressed, and separately coupled to the reaction it is catalyzing. For example, if a catalysis reaction is associated with two isozymes, the ME-Model will contain two versions of the catalysis reaction. Each version will contain the same metabolites and stoichiometry, except that they will be coupled with the c_3 and c_4 coefficients associated with the respective isozymes.

Typically, the final reaction in the series of linear reactions is the reaction that is coupled. For example, if the active enzyme is a complex, the complex formation reaction is coupled to the catalyzed reaction in c_3; alternatively, if the enzyme is a monomer, the final transport reaction to the cell compartment is coupled to the catalyzed reaction in c_3,. The exception is in the protein synthesis reaction (denominator, Equations (11) and (12)) for c_1 and c_2. In this case, it is the first protein synthesis reaction (translation or co-translational translocation for Cytosolic and Secretory Transport enzymes, respectively) that is coupled to the final mRNA formation reaction. Logically, this makes sense because we are limiting the coupling to one branch point (i.e., a multi-localizing protein that is both cytosolically translated and co-translationally translocated).

For orphan reactions, i.e., those without a GPR, we assign a cytosolic dummy protein to catalyze them ("deorphaning"). We reason that this is a more accurate representation of machinery resource costs than allowing the reaction to proceed spontaneously. The dummy protein represents a typical gene calculated from the features in the PSIM, limited to genes in Recon2.2. pre-mRNA and mRNA sequence lengths are calculated as the median across those in the PSIM. Protein sequence lengths are ⅓ this mRNA length. The base pair or amino acid in each position is assigned according to the relative frequency in that position as done for the consensus tRNA sequences (see Methods - tRNA Expression). Other features are calculated as the median of the values in the PSIM. Reactions that are not orphaned include: exchange reactions, demand reactions, expression module reactions that we curated and deliberately left without a GPR (e.g., complex formation), and some transport reactions. For transport reactions, we do not de-orphan those where all transported metabolites are less than 504 kDa[108] or if the reaction name contains “via diffusion”.

Table 5.3: Term definitions for coupling constraints.

Term/Notation	Definition	Units
[X]	Concentration of metabolite X	mmol/gDW$_{cell}$
i	mRNA of a specific gene	
j	Protein of a specific gene	
$\tau_{1/2,\ i}$	mRNA half-life	hr
$\alpha_{m,i}$	First-order mRNA degradation constant Calculated as ln(2)/$\tau_{1/2,\ i}$	hr^{-1}
$\alpha_{p,j}$	First-order protein degradation constant	hr^{-1}
$k_{p,j}$	First order rate constant for protein synthesis	hr^{-1}
MW$_j$	Molecular weight of protein *j*	kDa
SASA$_j$	Protein solvent accessible surface area Estimated as MW$_j^{0.75}$	
$k_{cat,j}$	Enzyme catalytic rate	hr^{-1}
PTR$_{i,j}$	The protein-to-rna ratio for a given gene.	
μ	Cell growth rate	hr^{-1}

To improve biological accuracy, we derive coupling constraints from first principles according to the simplified network above. They include a growth-dependent (dilution) and growth-independent (degradation) term. Growth-dependent terms account for dilution of molecules due to cell division. Since the biomass dilution reaction is a sink, these components do not contribute to consumption and are not mass-balanced (see Methods - Formatting the Biomass Reaction for technical details).

For *c₁* and *c₂*, we wanted to derive coupling coefficients using context-independent parameters, i.e., those that we can expect to be fairly consistent across conditions. To this end, we utilize $\mathrm{PTR}_{i,j}$, which, while being gene-specific, is demonstrated to be tissue-independent[109,110]. We also formulate coupling coefficients in terms of degradation rates rather than production rates, as degradation rates don't change a genes' distribution in Crick space, have a smaller dynamic range, and smaller contribution to the overall steady-state protein levels[106].

5.4.1.7.1 Coupling Coefficient Derivations

Coupling coefficients c_1 and c_2 are derived as follows:

From steady-state:

$$v_{\mathrm{formation,i}} = v_{\mathrm{dilution,i}} + v_{\mathrm{degradation,i}} \text{ (1)}$$

$$v_{synthesis,j} = v_{dilution,j} + v_{degradation,j} \text{ (2)}$$

From first-order reaction laws:

$$v_{\mathrm{dilution,i}} = \mu * [mRNA_i] \text{ (3)}$$

$$v_{\mathrm{dilution,j}} = \mu * [protein_j] \text{ (4)}$$

$$v_{\mathrm{degradation,i}} = \alpha_{m,i} * [mRNA_i] \text{ (5)}$$

$$v_{\mathrm{degradation,j}} = \alpha_{p,j} * [protein_j] \text{ (6)}$$

Additionally:

$$PTR_{i,j} = \frac{[protein_j]}{[mrna_i]} \text{ (7)}$$

Substituting (4) and (6) into (2):

$$v_{synthesis,j} = (\alpha_{p,j} + \mu)[protein_j] \text{ (8)}$$

Substituting (7) into (8):

$$v_{synthesis,j} = (\alpha_{p,j} + \mu)PTR_{i,j}[mrna_i] \text{ (9)}$$

Substituting (3) and (5) into (1):

$$v_{formation,i} = (\alpha_{m,i} + \mu)[mrna_i] \text{ (10)}$$

Finally, if we define:

$$c_1 = \frac{v_{formation,i}}{v_{synthesis,j}} \text{ (11)}$$

$$c_2 = \frac{v_{degradation,i}}{v_{synthesis,j}} \text{ (12)}$$

Then, we can substitute (9) and (10) into the denominator and numerator of (11), respectively:

$$c_1 = \frac{(\alpha_{m,i} + \mu)[mRNA_i]}{(\alpha_{p,j} + \mu)PTR_{i,j}[mRNA_i]} = \frac{(\alpha_{m,i} + \mu)}{(\alpha_{p,j} + \mu)PTR_{i,j}} \text{ (13)}$$

And we can substitute (9) and (5) into the denominator and numerator of (12), respectively:

$$c_2 = \frac{\alpha_{m,i}[mRNA_i]}{(\alpha_{p,j} + \mu)PTR_{i,j}[mRNA_i]} = \frac{\alpha_{m,i}}{(\alpha_{p,j} + \mu)PTR_{i,j}} \text{ (14)}$$

Coupling coefficients *c_3* and *c_4* are derived as follows:

Assuming Michaelis-Menten kinetics:

$$v_{\mathrm{catalysis,j}} = \frac{v_{\max}[S]}{K_{m,j} + S} = \frac{k_{\mathrm{cat,j}}[protein_j][S]}{K_{m,j} + S} \text{ (15)}$$

assuming K_m << [S]:

$$v_{\mathrm{catalysis,j}} = k_{\mathrm{cat,j}}[protein_j] \text{ (16)}$$

If we define:

$$c_3 = \frac{v_{\text{synthesis,j}}}{v_{\text{catalysis,j}}} \quad (17)$$

$$c_4 = \frac{v_{\text{degradation,j}}}{v_{\text{catalysis,j}}} \quad (18)$$

Then, we can substitute (8) and (16) into the numerator and denominator of (17), respectively:

$$c_3 = \frac{(\mu + \alpha_{\text{p,j}})[protein_j]}{k_{\text{cat,j}}[protein_j]} = \frac{(\mu + \alpha_{\text{p,j}})}{k_{\text{cat,j}}} \quad (19)$$

And we can substitute (6) and (16) into the denominator and numerator of (18), respectively:

$$c_4 = \frac{\alpha_{\text{p,j}}[protein_j]}{k_{\text{cat,j}}[protein_j]} = \frac{\alpha_{\text{p,j}}}{k_{\text{cat,j}}} (20)$$

5.4.1.7.2 Coupling Coefficient Parameter Values

The right-hand side of (13), (14), (19), and (20) are all parameters that are reported in the PSIM or can be calculated from the PSIM. $\text{PTR}_{i,j}$ is obtained from ref[47], and for genes that do not have a reported value, the median of 65163 is used. $\alpha_{m,i}$ is obtained from ref[44], and for genes that do not have a reported value, the median of 0.061 hr^{-1} is used. $\alpha_{p,j}$ is obtained from ref[111]. Without replacing any intersecting genes, additional values for proteins reported as short-lived were obtained from “Table S2” of ref[46]. For genes that do not have a reported value, the median of 0.02 hr^{-1} is used.

Since *in vitro* measurements of $k_{\text{cat,j}}$ are not widely available, we estimate the value from the SASA_j as in COBRAme[21]. Specifically, we can set:

$$k_{cat,j} = \text{SASA}_j \frac{M_j(k_{cat})}{M_j(\text{SASA})} (21)$$

Where M_j represents the median. The median SASA is calculated by taking all unique metabolic enzymes (unique complexes or single-proteins). Across 5, 785 Recon2.2 enzymes, including complexes, the median SASA is 25.85. To get a median enzyme catalytic rate, we query the BRENDA database[112] for human enzymes. We disregard proteins reported as "mutants" in the comments. When a range was provided for a particular enzyme, the average was taken. Altogether, we retrieved 1638 catalytic rates from 238 unique EC-classes and 259 unique EC-class-protein pairs. For each EC-class-protein pair, we took the average across all reported values. We arrive at $M_j(k_{cat}) = 14338.8\ hr^{-1}$.

5.4.1.8 Formatting the Biomass Reaction

The ME-Model has a different formulation for the biomass reaction than the M-Model (Appendix B) to enable variable RNA and protein mass fractions. This can be directly modified from the M-Model biomass reaction as formatted in Appendix B (equations B-1 and B-5). In the ME-Model, each biomass component has a separate 1:1 input to the total biomass (see Table 4 for notation):

$$\text{for each } i \in n, \text{ biomass}_i \rightarrow \text{biomass}_{\text{tot}} \quad (22)$$

This 1:1 stoichiometry can be interpreted as follows: Each gram of the biomass component *i* produced contributes to 1 g of total biomass. This allows RNA and protein to be input to biomass at variable proportions. With this 1:1 ration, using the same logic equation (B-4), the units of flux through (22) reactions simplify to hr^{-1}. (22) has a lower bound of zero and an upper bound of 1000, but is constrained by μ as explained below. Given this 1:1 ratio, to enforce mass fraction of biomass components other than RNA and protein, each biomass component formation reaction must be reformatted from (B-5) to:

$$p_i s_{1,i} \mathrm{X}_{1,i} + p_i s_{2,i} \mathrm{X}_{2,i} + \ldots + p_i s_{m,i} \mathrm{X}_{m,i} \rightarrow p_i \text{biomass}_i \quad (23)$$

The flux bounds of (23) are constrained to growth (i.e., lower bound = μ and upper bound = μ). Units of p_i and $(p_i{*}s_{j,i})$ are the same as in Appendix B. Since $\text{biomass}_{\text{tot}}$ directly determines growth rate (see description of biomass dilution reaction below), constraining the flux bounds of (23) by μ ensures that only p_i counts of biomass_i can contribute to $\text{biomass}_{\text{tot}}$ in (22). In other words, when the flux through (23) = μ, the production rate of $\text{biomass}_{,}$ = p_iμ. Note, (B-7) mass balance for (B-5) can be rewritten as below for (23):

$$\sum_{j=1}^{m} p_i s_{j,i} * \mathrm{MW}(\mathrm{X}_{j,i}) = p_i \quad (24)$$

For the variable biomass components (protein and RNA), instead of (23), biomass is produced (and consumed) directly in the gene expression reactions. For example, let's imagine a generic protein synthesis reaction produces *k* units of protein_A, where *k* is the stoichiometric coefficient (Reaction A: amino acids → (*k*)protein_A). Considering the units of *k* (Table 4 row 1) and molecular weight (Table 4 row 3), this can be converted to the amount of protein biomass produced by scaling *k* by the molecular weight of the protein:

$$k * \mathrm{MW}(\text{protein}_A) = \text{biomass}_{\text{protein}} \quad (25)$$

This matches the units of (B-3):

$$\frac{mmol}{gDW_{cell}} * \frac{g}{mmol} = \frac{g}{gDW_{cell}} \quad (26)$$

Thus, we can add $\text{biomass}_{\text{protein}}$ as a product in Reaction A (amino acids → (*k*)protein_A + (k)(MW(protein_A))$\text{biomass}_{\text{protein}}$). Each reaction that forms or degrades RNA or protein, excluding coupled macromolecules (see Methods - Reaction Coupling section for reasoning), produces or consumes biomass in this manner. For consumption, the biomass term is in the substrates rather than the products. Finally, we can create a biomass dilution reaction that is analogous to the biomass consumption reaction in Appendix B, with the only difference being that in the ME-Model, this is bounded by μ. This biomass dilution reaction serves as the objective in the linear program (see Methods - Solving the ME-Model section for details).

We note that RNA and protein can only vary within their total mass fraction. Let's say that all other biomass components sum up to a mass fraction of *F*, where F < 1. Then, while each of p_{RNA} and $p_{protein}$ can vary, they are constrained by $p_{RNA} + p_{protein} = 1 - F$ (from equation B-2).

Many biomass objectives include an ATP hydrolysis that represents GAM energetic costs[113]. In M-Models, the GAM ATP hydrolysis is often included in the protein biomass component formation reaction, as in Recon2.2. Thus, the elimination of the (B-5) format of the RNA and protein biomass components generally will eliminate the ATP hydrolysis for growth-associated maintenance (GAM). However, GAM is typically implemented in the protein biomass component formation reaction because protein synthesis costs represent a large portion of GAM costs and thus can be used as a proxy for GAM (see step 32 of ref[113]). While there are other energetic costs which are not accounted for in the ME Model, e.g. error-checking and[114] replication, we reason that removal of the explicit GAM ATP hydrolysis term by the ME-Model should not substantially underestimate the GAM costs, since the ME model explicitly accounts for transcript and protein expression costs which constitute an overwhelming majority of GAM costs[18].

5.4.2 Solving the ME-Model

Given the large order of magnitude differences between standard stoichiometric coefficients and coupling coefficients, we needed to use a high-precision solver. Thus, we implemented the qMINOS solver[115] through the QMINOS class in solveME[116]. Since some parameters in the ME-Model are a function of the variable μ (e.g., coupling constraints and biomass reactions), to create a Linear Program (LP), we must first substitute in a floating value for this variable. Once a value of μ is assigned, the model can be solved as in the M-Model; in other words, we can optimize for an objective of interest at a particular value of growth rate. In this case, if the objective of interest is growth rate, the solution will identify a flux equal to the assigned value. Thus, to maximize growth, we must identify the boundary at which the ME-Model becomes infeasible. To do this, we implement a binary search algorithm as described in COBRAme[21].

Finally, we implement an algorithm to maximize for an objective other than growth rate (as in Fig. 5.3d). To do so, we first need to identify the maximum growth rate μ*. Next, using growth rate values at a specified number of intervals within the feasible growth range [0, μ*], we can solve the LP and store the objective value. Next, we estimate the objective value as a function of μ using the scipy function "interp1D". Finally, we estimate the objective value across 1000 growth rate values specified at even intervals within the feasible growth range. We take the maximum estimated objective value across these 1000 growth rate values as the optimized objective value. We must do this rather than directly solving across 1000 growth rate values because the time it takes to run the LP is computationally limiting.

5.4.3 Refining NCI-60 Cell Line M-Model Inputs

Input M-models were adapted from those generated in ref[30]. Specifically, we took the mCADRE[117]-extracted M-models (not "Protected") and made the following adaptations: 1) changes to the biomass reaction, 2) changes to the exchange reaction bounds, and 3) addition of missing reactions that prevented model feasibility (see Appendix A for details on point 2-3).

Reasoning for changes to the biomass reaction can be found in Appendix B. First, the biomass reaction was re-formatted from the net reaction to reactions for formation of each biomass component separately, matching the format of Recon 2.2. In ref[30], the biomass reaction stoichiometric coefficients were calculated from "Table S1" from the "Supplementary Information" of ref'[15]. We combined the "resolved" and "unresolved" lipid mass fractions, representing mass fractions of 7.98e-2 and 2.27e-2, respectively, into a single lipid component by taking the sum (lipid mass fraction = 1.025e-1). Additionally, ATP hydrolysis for growth-associated maintenance (GAM) ("Growth-associated" and "Unresolved other components" from "Table S1"[15].) has a mass fraction of 4.37e-2. However, GAM ATP hydrolysis should not have a direct mass fraction. Since

Recon2.2 incorporates GAM ATP hydrolysis into the formation reaction for the protein biomass component, we do the same here. We add the associated substrates to the protein formation component, and we add the 4.37e-2 mass fraction to the existing specified mass fraction for protein, 7.21e-1, for a final protein mass fraction of 7.647e-1. We set the "Resolved small molecules" from "Table S1"[15] as the "other" biomass component. Next, for each biomass component formation reaction, if equation (B-7) does not hold, substrate coefficients were re-scaled to do so. Here, *i* represents a substrate in the biomass component formation reaction, *s* is the stoichiometric coefficient of that substrate, and *MW* is the molecular weight of that substrate in kDa. For the carbohydrate biomass component specifically, the molecular weight of glycogen specified in "Table S1" is less than that of a glucose monomer; so, we use the Recon2.2 molecular weight while re-scaling.

In ref[30], exchange reaction bounds were specified using measurements reported in "Table S2". Exchange reactions that were removed during mCADRE model-extraction but experimentally reported in this table were not retained. Here, we added those reactions back, assuming that if they were experimentally measured, they were present in the cell line despite results of the model extraction step.

5.4.4 K-562 Cell Line ME-Model Analysis

To compare M-Model fluxes with ME-Model fluxes of the metabolic module, we aggregated certain reactions from the ME-Model solution. During ME-Model building, isozymes (reactions with an "OR" in the GPR) are each split into a separate reaction. Additionally, reversible reactions are split into their forward and reverse directions, separately. Reactions that were split due to isozymes were aggregated by taking the sum of their fluxes. Next, reversible reactions that were split into their forward and reverse directions were aggregated by subtracting the flux for the reaction in the reverse direction from that in the forward direction.

To compare model efficiency, all solving was run by setting the objective to the biomass reaction and the biomass reaction was bounded to be the maximum growth rate of the ME-Model (0.0266 hr^{-1}). In this manner, discrepancies in total absolute flux were not due to one solution generating more overall biomass. FBA solutions for the M-Model and ME-Model were identified using the qMINOS solver. The M-Model solution without thermodynamically infeasible loops was found by taking the FBA solution and running it through the CycleFreeFlux algorithm[33] via cobrapy's "loopless_solution" function. The pFBA solution was also identified using cobrapy's "pfba" function.The solution space distribution was identified by sampling it 1000 times using cobrapy's "sample" function.

FVA reaction bounds of the M-Model were identified using cobrapy's "flux_variability_analysis" function while holding the biomass reaction at its maximum value (0.0559 hr^{-1}). FVA reaction bounds of the ME-Model were found by iterating through each reaction and maximizing and minimizing the upper and lower bounds, respectively, while holding the biomass reaction at its maximum value (0.0266 hr^{-1}). In this case, growth rates were held at their respective maximum values, rather than being the same between models, to ensure that flux ranges were limited by the respective constraints in each model and did not have additional flexibility available to fluxes by not performing optimally.

For over-representation analysis via Metascape[36], the background was set to be all genes in the context-specific M-Model.Finally, the "Oxidative Phosphorylation" objective (Fig. 5.3d) was set to be the sum of the reactions for each step in the electron transport chain (Recon2.2 reaction IDs: "NADH2_u10m", "FADH2ETC", "CYOR_u10m", "CYOOm2"), the proton pump (Recon2.2 reaction ID: "ATPS4m"), and a few additional associated reactions (Recon2.2 reaction IDs: "SUCD1m_F", "ETFQO", "FCLTm").

5.5 Appendix

5.5.1 Appendix A: Additional Machinery Constraints Identify Missing Key Reactions in M-Model

When first constructing the ME-Model for the K-562 cell line, we observed that it was not feasible, even at very low growth rates. The additional machinery constraints imposed by the ME-Model caused infeasibility due to key reactions being missing. Due to fewer constraints in the M-Model, these were not identified in the original M-Model. Reactions included those in the glycerol-3-phosphate shuttle for co-factor recycling, transport of a number of metabolites, exchange reactions (see Methods - Refining NCI-60 Cell Line M-Model Inputs for details), and proline biosynthesis. Upon adding these reactions that were originally present in Recon2.2 but not the mCADRE extracted M-Model, the ME-Model became feasible. This demonstrates that the ME-Model can help in further refining the metabolic reactions that should be included in analyses.

Additionally, in instances where the M-Model forced flux through reactions, we relaxed the boundaries. In other words, if the lower bound was ≥ 0, we set it to 0, and if the upper bound was ≤, we set it to 0. This change to flux bounds was applied to exchange reactions as well, which somewhat deviates from *in vivo* accuracy since these bounds were set according to experimental measurements as reported in “Table S2” of ref[30], but was necessary for model feasibility.

5.5.2 Appendix B: M-Model Biomass Reaction

Here, we outline how the M-Model biomass reaction is formulated, which is important for appropriately re-formatting it for the ME-Model (see Methods - Formatting the Biomass Reaction).

Table 5.4: Term definitions for biomass.

Term/Notation	Definition	Units
[X]	Concentration of metabolite X Held by the stoichiometric coefficient	mmol/gDW$_{cell}$
v	Reaction flux Metabolite consumption/production rate	mmol/gDW$_{cell}$/hr
MW	metabolite/macromolecule molecular weight	kDa = g/mmol
µ	Cell growth rate	hr^{-1}
i	Indexing of biomass components	
biomass$_i$	Biomass component *i* Typically: DNA, RNA, Lipid, Carbohydrate, Lipid, Other	
n	The total number of biomass components	
p$_i$	Mass fraction of biomass component *i* (of the total biomass of the cell, what relative proportion does biomass component *i* represent?)	g biomass$_i$/gDW$_{cell}$ (see equation B-3)
j	Indexing of metabolite substrates that form biomass$_i$	
m	The total number of metabolite substrates that form biomass$_i$	

The biomass reaction is formulated as the sum of its components, each with a stoichiometric coefficient representing its mass fraction:

$$\text{For } i \in n,\ p_{i=1} * \text{biomass}_{i=1} + p_{i=2} * \text{biomass}_{i=2} + \ldots + p_{i=n} * \text{biomass}_{i=n} \rightarrow \text{biomass}_{\text{tot}} \qquad \text{(B-1)}$$

Given the definition of p$_i$ (Table 4) and for mass balance we have:

$$\sum_{i=1}^{n} p_i = 1 \qquad \text{(B-2)}$$

Unlike other reactions, the biomass reaction flux has special units. Since the biomass$_{tot}$ is the same as the cell dry weight (DW$_{cell}$), 1g of biomass$_{tot}$ must be produced per 1gDW$_{cell}$. Next, given the definition of p$_i$, the units must be:

$$\frac{\text{g biomass}_i}{\text{1g biomass}_{\text{tot}}} = \frac{\text{g biomass}_i}{\text{1g DW}_{\text{cell}}} \qquad \text{(B-3)}$$

Given (B-2) and (B-3), we get the following flux units for the biomass reaction (i.e., the total biomass production rate):

$$\frac{\text{g biomass}_{\text{tot}}}{\text{g DW}_{\text{cell}} * \text{hr}} = \text{hr}^{-}1 \qquad \text{(B-4)}$$

Each biomass$_i$ has its own component formation reaction:

$$\text{For } j \in m,\ s_{j=1,i}\text{X}_{j=1,i} + s_{j=2,i}\text{X}_{j=2,i} + \ldots + s_{j=m,i}\text{X}_{j=m,i} \rightarrow \text{biomass}_i \text{(B-5)}$$

Note, (B-5) and (B-1) are often combined to create a net biomass reaction as follows:

$$[(p_1 * s_{1,1})\text{X}_{1,1} + (p_1 * s_{2,1})\text{X}_{2,1} + \ldots + (p_1 * s_{m,1})\text{X}_{m,1}] + [(p_2 * s_{1,2})\text{X}_{1,2} + (p_2 * s_{2,2})\text{X}_{2,2} + \ldots + (p_2 * s_{m,2})\text{X}_{m,2})] + \ldots + [(p_n * s_{1,n})\text{X}_{1,n} + (p_n * s_{2,n})\text{X}_{2,n} + \ldots + (p_n * s_{m,n})\text{X}_{m,n})] \rightarrow \text{biomass}_{\text{tot}} \qquad \text{(B-6)}$$

For the ME-Model, the M-Model input should be formulated as (B-1) and (B-5), rather than (B-6). Thus, if your M-Model is formulated as (B-6), divide each coefficient by p_i to get the coefficients for (B-5). We know the units of p_i from (B-3) and the units of p_i*$s_{j,i}$ from Table 4 (row 1). Thus, the units of $s_{j,i}$ should be mmol/(g biomass$_i$). To achieve mass balance analogous to (B-2), we have:

$$\sum_{j=1}^{m} s_{j,i} * \text{MW}(\text{X}_{j,i}) = 1 \qquad \text{(B-7)}$$

(B-7) tells us that with the stoichiometric amount ($s_{j,i}$) of $\text{X}_{j,i}$, (B-5) produces 1 g biomass$_i$. Finally, there is a biomass consumptions reaction to consume biomass$_{\text{tot}}$ and prevent it from accumulating, generating mass balance (biomass$_{\text{tot}}$ →).

5.5.3 Appendix C: Data Availability

Files needed for building the ME-Model can be found at https://github.com/hmbaghdassarian/human_me_data.

5.5.4 Appendix D: Authors, Contributions, and Acknowledgements

Authors: Hratch M. Baghdassarian, Juan Tibocha-Bonilla, Erick Armingol, Laurence Yang, Nathan E. Lewis.

H.M.B., and N.E.L. conceived the work. All authors provided important insights for formulating and analyzing the ME-Model. J.T.B. and L.Y. provided input on deriving the coupling coefficients. J.T.B provided input on formulating the biomass reaction. L.Y. provided input on implementing the solver. E.A. provided input on improving model feasibility and implemented code to parse reaction GPRs. H.B. implemented the humanME package and performed analyses of the K-562 cell line. H.M.B. wrote the paper and all authors carefully reviewed, discussed and edited the paper.

5.6 References

1. Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr, Assad-Garcia, N., Glass, J. I. & Covert, M. W. A whole-cell computational model predicts phenotype from genotype. *Cell* **150,** 389–401 (2012).

2. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* **26,** 721–744 (2010).

3. Samoudi, M., Masson, H. O., Kuo, C.-C., Robinson, C. M. & Lewis, N. E. From omics to Cellular mechanisms in mammalian cell factory development. *Curr. Opin. Chem. Eng.* **32,** (2021).

4. Basan, M. Resource allocation and metabolism: the search for governing principles. *Curr. Opin. Microbiol.* **45,** 77–83 (2018).

5. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330,** 1099–1102 (2010).

6. Gutierrez, J. M., Feizi, A., Li, S., Kallehauge, T. B., Hefzi, H., Grav, L. M., Ley, D., Baycin Hizal, D., Betenbaugh, M. J., Voldborg, B., Faustrup Kildegaard, H., Min Lee, G., Palsson, B. O., Nielsen, J. & Lewis, N. E. Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nat. Commun.* **11,** 68 (2020).

7. Zanotelli, M. R., Rahman-Zaman, A., VanderBurgh, J. A., Taufalele, P. V., Jain, A., Erickson, D., Bordeleau, F. & Reinhart-King, C. A. Energetic costs regulated by cell mechanics and confinement are predictive of migration path during decision-making. *Nat. Commun.* **10,** 4185 (2019).

8. Jones, R. D., Qian, Y., Siciliano, V., DiAndreth, B., Huh, J., Weiss, R. & Del Vecchio, D. An endoribonuclease-based feedforward controller for decoupling resource-limited genetic modules in mammalian cells. *Nat. Commun.* **11,** 5690 (2020).

9. Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K. & Alon, U. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336,** 1157–1160 (2012).

10. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336,** 601–604 (2012).

11. Adler, M., Korem Kohanim, Y., Tendler, A., Mayo, A. & Alon, U. Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. *Cell Syst* **8,** 43–52.e5 (2019).

12. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biol.* **20,** 121 (2019).

13. Lewis, N. E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M. P., Cheng, J. K., Patel, N., Yee, A., Lewis, R. A., Eils, R., König, R. & Palsson, B. Ø. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.* **28,** 1279–1285 (2010).

14. Wagner, A., Wang, C., Fessler, J., DeTomaso, D., Avila-Pacheco, J., Kaminski, J., Zaghouani, S., Christian, E., Thakore, P., Schellhaass, B., Akama-Garren, E., Pierce, K., Singh, V., Ron-Harel, N., Douglas, V. P., Bod, L., Schnell, A., Puleston, D., Sobel, R. A., Haigis, M., Pearce, E. L., Soleimani, M., Clish, C., Regev, A., Kuchroo, V. K. & Yosef, N. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity. *Cell* **184,** 4168–4185.e21 (2021).

15. Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C. & Lewis, N. E. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst* **4,** 318–329.e6 (2017).

16. Holzhütter, H.-G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271,** 2905–2922 (2004).

17. Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D. & Palsson, B. Ø. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6,** 390 (2010).

18. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15690–15695 (2015).

19. Reuveni, S., Ehrenberg, M. & Paulsson, J. Ribosomes are optimized for autocatalytic production. *Nature* **547,** 293–297 (2017).

20. Schinn, S.-M., Morrison, C., Wei, W., Zhang, L. & Lewis, N. E. Systematic evaluation of parameters for genome-scale metabolic models of cultured mammalian cells. *Metab. Eng.* **66,** 21–30 (2021).

21. Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O'Brien, E. J., Liu, J. K. & Palsson, B. O. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput. Biol.* **14,** e1006302 (2018).

22. Dahal, S., Zhao, J. & Yang, L. Recent advances in genome-scale modeling of proteome allocation. *Current Opinion in Systems Biology* **26,** 39–45 (2021).

23. Thiele, I., Jamshidi, N., Fleming, R. M. T. & Palsson, B. Ø. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5,** e1000312 (2009).

24. Thiele, I., Fleming, R. M. T., Bordbar, A., Schellenberger, J. & Palsson, B. Ø. Functional characterization of alternate optimal solutions of Escherichia coli's transcriptional and translational machinery. *Biophys. J.* **98,** 2072–2081 (2010).

25. Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K. & Palsson, B. O. In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3,** 929 (2012).

26. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9,** 693 (2013).

27. Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E. H., Fields, A. P., Schwartz, S., Raychowdhury, R., Mumbach, M. R., Eisenhaure, T., Rabani, M., Gennert, D., Lu, D., Delorey, T., Weissman, J. S., Carr, S. A., Hacohen, N. & Regev, A. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347,** 1259038 (2015).

28. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2,** e270 (2014).

29. Swainston, N., Smallbone, K., Hefzi, H., Dobson, P. D., Brewer, J., Hanscho, M., Zielinski, D. C., Ang, K. S., Gardiner, N. J., Gutierrez, J. M., Kyriakopoulos, S., Lakshmanan, M., Li, S., Liu, J. K., Martínez, V. S., Orellana, C. A., Quek, L.-E., Thomas, A., Zanghellini, J., Borth, N., Lee, D.-Y., Nielsen, L. K., Kell, D. B., Lewis, N. E. & Mendes, P. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12,** 109 (2016).

30. Richelle, A., Chiang, A. W. T., Kuo, C.-C. & Lewis, N. E. Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. *PLoS Comput. Biol.* **15,** e1006867 (2019).

31. Noor, E., Bar-Even, A., Flamholz, A., Reznik, E., Liebermeister, W. & Milo, R. Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS Comput. Biol.* **10,** e1003483 (2014).

32. Price, N. D., Famili, I., Beard, D. A. & Palsson, B. Ø. Extreme pathways and Kirchhoff's second law. *Biophys. J.* **83,** 2879–2882 (2002).

33. Desouki, A. A., Jarre, F., Gelius-Dietrich, G. & Lercher, M. J. CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. *Bioinformatics* **31,** 2159–2165 (2015).

34. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–248 (2010).

35. Klijn, C., Durinck, S., Stawiski, E. W., Haverty, P. M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., Pau, G., Reeder, J., Cao, Y., Mukhyala, K., Selvaraj, S. K., Yu, M., Zynda, G. J., Brauer, M. J., Wu, T. D., Gentleman, R. C., Manning, G., Yauch, R. L., Bourgon, R., Stokoe, D., Modrusan, Z., Neve, R. M., de Sauvage, F. J., Settleman, J., Seshagiri, S. & Zhang, Z. A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33,** 306–312 (2015).

36. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C. & Chanda, S. K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10,** 1523 (2019).

37. Chen, K., Gao, Y., Mih, N., O'Brien, E. J., Yang, L. & Palsson, B. O. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 11548–11553 (2017).

38. Bordbar, A., Feist, A. M., Usaite-Black, R., Woodcock, J., Palsson, B. O. & Famili, I. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst. Biol.* **5,** 180 (2011).

39. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22,** 71–88 (2021).

40. Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J. M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V. S., Kay, M., Kodali, V. K., McGarvey, K. M., McMahon, A., Mudge, J. M., Murphy, D. N., Murphy, M. R., Rajput, B., Rangwala, S. H., Riddick, L. D., Thibaud-Nissen, F., Threadgold, G., Vatsan, A. R., Wallin, C., Webb, D., Flicek, P., Birney, E., Pruitt, K. D., Frankish, A., Cunningham, F. & Murphy, T. D. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604,** 310–315 (2022).

41. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D. & Pruitt, K. D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–45 (2016).

42. Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A. & Tress, M. L. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41,** D110–7 (2013).

43. Legnini, I., Alles, J., Karaiskos, N., Ayoub, S. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16,** 879–886 (2019).

44. Gregersen, L. H., Schueler, M., Munschauer, M., Mastrobuoni, G., Chen, W., Kempa, S., Dieterich, C. & Landthaler, M. MOV10 Is a 5' to 3' RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3' UTRs. *Mol. Cell* **54,** 573–585 (2014).

45. Cambridge, S. B., Gnad, F., Nguyen, C., Bermejo, J. L., Krüger, M. & Mann, M. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J. Proteome Res.* **10,** 5275–5284 (2011).

46. Li, J., Cai, Z., Vaites, L. P., Shen, N., Mitchell, D. C., Huttlin, E. L., Paulo, J. A., Harry, B. L. & Gygi, S. P. Proteome-wide mapping of short-lived proteins in human cells. *Mol. Cell* **81,** 4722–4735.e5 (2021).

47. Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallström, B. M., Uhlén, M., Asplund, A., Pontén, F., Wieland, T., Hopf, T., Hahne, H., Kuster, B. & Gagneur, J. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol. Syst. Biol.* **15,** e8513 (2019).

48. Betat, H., Long, Y., Jackman, J. E. & Mörl, M. From end to end: tRNA editing at 5'- and 3'-terminal positions. *Int. J. Mol. Sci.* **15,** 23975–23998 (2014).

49. Gogakos, T., Brown, M., Garzia, A., Meyer, C., Hafner, M. & Tuschl, T. Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell Rep.* **20,** 1463–1475 (2017).

50. Chatterjee, K., Nostramo, R. T., Wan, Y. & Hopper, A. K. tRNA dynamics between the nucleus, cytoplasm and mitochondrial surface: Location, location, location. *Biochim. Biophys. Acta Gene Regul. Mech.* **1861,** 373–386 (2018).

51. Phizicky, E. M. & Hopper, A. K. tRNA biology charges to the front. *Genes Dev.* **24,** 1832–1860 (2010).

52. Turowski, T. W. & Tollervey, D. Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochem. Soc. Trans.* **44,** 1367–1375 (2016).

53. Walker, S. C. & Engelke, D. R. Ribonuclease P: the evolution of an ancient RNA enzyme. *Crit. Rev. Biochem. Mol. Biol.* **41,** 77–102 (2006).

54. Rossmanith, W. Localization of human RNase Z isoforms: dual nuclear/mitochondrial targeting of the ELAC2 gene product by alternative translation initiation. *PLoS One* **6,** e19152 (2011).

55. Lizano, E., Schuster, J., Müller, M., Kelso, J. & Mörl, M. A splice variant of the human CCA-adding enzyme with modified activity. *J. Mol. Biol.* **366,** 1258–1265 (2007).

56. Paushkin, S. V., Patel, M., Furia, B. S., Peltz, S. W. & Trotta, C. R. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3’ end formation. *Cell* **117,** 311–321 (2004).

57. Li, S. & Sprinzl, M. Interaction of immobilized human exportin-t with calf liver tRNA. *RNA Biol.* **3,** 145–149 (2006).

58. Banik, S. D. & Nandi, N. Chirality and protein biosynthesis. *Top. Curr. Chem.* **333,** 255–305 (2013).

59. Bohnsack, K. E. & Bohnsack, M. T. Uncovering the assembly pathway of human ribosomes and its emerging links to disease. *EMBO J.* **38,** e100278 (2019).

60. Grou, C. P., Pinto, M. P., Mendes, A. V., Domingues, P. & Azevedo, J. E. The de novo synthesis of ubiquitin: identification of deubiquitinases acting on ubiquitin precursors. *Sci. Rep.* **5,** 12836 (2015).

61. Henras, A. K., Plisson-Chastang, C., O’Donohue, M.-F., Chakraborty, A. & Gleizes, P.-E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip. Rev. RNA* **6,** 225–242 (2015).

62. Pirouz, M., Munafò, M., Ebrahimi, A. G., Choe, J. & Gregory, R. I. Exonuclease requirements for mammalian ribosomal RNA biogenesis and surveillance. *Nat. Struct. Mol. Biol.* **26,** 490–500 (2019).

63. Ciganda, M. & Williams, N. Eukaryotic 5S rRNA biogenesis. *Wiley Interdiscip. Rev. RNA* **2,** 523–533 (2011).

64. Aubert, M., O’Donohue, M.-F., Lebaron, S. & Gleizes, P.-E. Pre-Ribosomal RNA Processing in Human Cells: From Mechanisms to Congenital Diseases. *Biomolecules* **8,** (2018).

65. Murdoch, K., Loop, S., Rudt, F. & Pieler, T. Nuclear export of 5S rRNA-containing ribonucleoprotein complexes requires CRM1 and the RanGTPase cycle. *Eur. J. Cell Biol.* **81,** 549–556 (2002).

66. Russell, J. & Zomerdijk, J. C. B. M. The RNA polymerase I transcription machinery. *Biochem. Soc. Symp.* 203–216 (2006).

67. Sloan, K. E., Bohnsack, M. T., Schneider, C. & Watkins, N. J. The roles of SSU processome components and surveillance factors in the initial processing of human ribosomal RNA. *RNA* **20,** 540–550 (2014).

68. Preti, M., O'Donohue, M.-F., Montel-Lehry, N., Bortolin-Cavaillé, M.-L., Choesmel, V. & Gleizes, P.-E. Gradual processing of the ITS1 from the nucleolus to the cytoplasm during synthesis of the human 18S rRNA. *Nucleic Acids Res.* **41,** 4709–4723 (2013).

69. Montellese, C., Montel-Lehry, N., Henras, A. K., Kutay, U., Gleizes, P.-E. & O'Donohue, M.-F. Poly(A)-specific ribonuclease is a nuclear ribosome biogenesis factor involved in human 18S rRNA maturation. *Nucleic Acids Res.* **45,** 6822–6836 (2017).

70. Tafforeau, L., Zorbas, C., Langhendries, J.-L., Mullineux, S.-T., Stamatopoulou, V., Mullier, R., Wacheul, L. & Lafontaine, D. The Complexity of Human Ribosome Biogenesis Revealed by Systematic Nucleolar Screening of Pre-rRNA Processing Factors. *Mol. Cell* **51,** 539–551 (2013).

71. Rouquette, J., Choesmel, V. & Gleizes, P.-E. Nuclear export and cytoplasmic processing of precursors to the 40S ribosomal subunits in mammalian cells. *EMBO J.* **24,** 2862–2872 (2005).

72. Eliseev, B., Yeramala, L., Leitner, A., Karuppasamy, M., Raimondeau, E., Huard, K., Alkalaeva, E., Aebersold, R. & Schaffitzel, C. Structure of a human cap-dependent 48S translation pre-initiation complex. *Nucleic Acids Res.* **46,** 2678–2689 (2018).

73. Toczydlowska-Socha, D., Zielinska, M. M., Kurkowska, M., Astha, Almeida, C. F., Stefaniak, F., Purta, E. & Bujnicki, J. M. Human RNA cap1 methyltransferase CMTr1 cooperates with RNA helicase DHX15 to modify RNAs with highly structured 5' termini. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373,** (2018).

74. Clerici, M., Faini, M., Aebersold, R. & Jinek, M. Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. *Elife* **6,** (2017).

75. Rüegsegger, U., Blank, D. & Keller, W. Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol. Cell* **1,** 243–253 (1998).

76. Sakharkar, M. K., Chow, V. T. K. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol.* **4,** 387–393 (2004).

77. Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. & Vitale, L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database* **2016,** (2016).

78. Zhu, J., He, F., Wang, D., Liu, K., Huang, D., Xiao, J., Wu, J., Hu, S. & Yu, J. A novel role for minimal introns: routing mRNAs to the cytosol. *PLoS One* **5,** e10144 (2010).

79. Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N. & Reed, R. Recruitment of the human TREX complex to mRNA during splicing. *Genes Dev.* **19,** 1512–1517 (2005).

80. Heath, C. G., Viphakone, N. & Wilson, S. A. The role of TREX in gene expression and disease. *Biochem. J* **473,** 2911–2935 (2016).

81. Viphakone, N., Hautbergue, G. M., Walsh, M., Chang, C.-T., Holland, A., Folco, E. G., Reed, R. & Wilson, S. A. TREX exposes the RNA-binding domain of Nxf1 to enable mRNA export. *Nat. Commun.* **3,** 1006 (2012).

82. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8,** 113–126 (2007).

83. van Dijk, E., Cougot, N., Meyer, S., Babajko, S., Wahle, E. & Séraphin, B. Human Dcp2: a catalytically active mRNA decapping enzyme located in specific cytoplasmic structures. *EMBO J.* **21,** 6915–6924 (2002).

84. Bauer, N. C., Doetsch, P. W. & Corbett, A. H. Mechanisms Regulating Protein Localization. *Traffic* **16,** 1039–1061 (2015).

85. Dever, T. E. & Green, R. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4,** a013706 (2012).

86. Walker, C. L., Pomatto, L. C. D., Tripathi, D. N. & Davies, K. J. A. Redox Regulation of Homeostasis and Proteostasis in Peroxisomes. *Physiol. Rev.* **98,** 89–115 (2018).

87. Kim, P. K. & Hettema, E. H. Multiple pathways for protein transport to peroxisomes. *J. Mol. Biol.* **427,** 1176–1190 (2015).

88. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M. & Hartl, F. U. Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* **82,** 323–355 (2013).

89. Vabulas, R. M., Raychaudhuri, S., Hayer-Hartl, M. & Hartl, F. U. Protein folding in the cytoplasm and the heat shock response. *Cold Spring Harb. Perspect. Biol.* **2,** a004390 (2010).

90. Radons, J. The human HSP70 family of chaperones: where do we stand? *Cell Stress Chaperones* **21,** 379–404 (2016).

91. Ohtsuka, K. & Hata, M. Mammalian HSP40/DNAJ homologs: cloning of novel cDNAs and a proposal for their classification and nomenclature. *Cell Stress Chaperones* **5,** 98–112 (2000).

92. Künkele, K. P., Heins, S., Dembowski, M., Nargang, F. E., Benz, R., Thieffry, M., Walz, J., Lill, R., Nussberger, S. & Neupert, W. The preprotein translocation channel of the outer membrane of mitochondria. *Cell* **93,** 1009–1019 (1998).

93. Stiller, S. B., Höpker, J., Oeljeklaus, S., Schütze, C., Schrempp, S. G., Vent-Schmidt, J., Horvath, S. E., Frazier, A. E., Gebert, N., van der Laan, M., Bohnert, M., Warscheid, B., Pfanner, N. & Wiedemann, N. Mitochondrial OXA Translocase Plays a Major Role in Biogenesis of Inner-Membrane Proteins. *Cell Metab.* **23,** 901–908 (2016).

94. Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E. & Corbett, A. H. Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J. Biol. Chem.* **282,** 5101–5105 (2007).

95. Sha, Z., Zhao, J. & Goldberg, A. L. Measuring the Overall Rate of Protein Breakdown in Cells and the Contributions of the Ubiquitin-Proteasome and Autophagy-Lysosomal Pathways. *Methods Mol. Biol.* **1844,** 261–276 (2018).

96. Kimura, Y. & Tanaka, K. Regulatory mechanisms involved in the control of ubiquitin homeostasis. *J. Biochem.* **147,** 793–798 (2010).

97. Gong, X., Du, D., Deng, Y., Zhou, Y., Sun, L. & Yuan, S. The structure and regulation of the E3 ubiquitin ligase HUWE1 and its biological functions in cancer. *Invest. New Drugs* **38,** 515–524 (2020).

98. Bard, J. A. M., Goodall, E. A., Greene, E. R., Jonsson, E., Dong, K. C. & Martin, A. Structure and Function of the 26S Proteasome. *Annu. Rev. Biochem.* **87,** 697–724 (2018).

99. Gur, E. & Sauer, R. T. Recognition of misfolded proteins by Lon, a AAA(+) protease. *Genes Dev.* **22,** 2267–2277 (2008).

100. Patron, M., Sprenger, H.-G. & Langer, T. m-AAA proteases, mitochondrial calcium homeostasis and neurodegeneration. *Cell Res.* **28,** 296–306 (2018).

101. Bonifacino, J. S. & Weissman, A. M. Ubiquitin and the control of protein fate in the secretory and endocytic pathways. *Annu. Rev. Cell Dev. Biol.* **14,** 19–57 (1998).

102. Hoseki, J., Ushioda, R. & Nagata, K. Mechanism and components of endoplasmic reticulum-associated degradation. *J. Biochem.* **147,** 19–25 (2010).

103. MacGurn, J. A. Garbage on, garbage off: new insights into plasma membrane protein quality control. *Curr. Opin. Cell Biol.* **29,** 92–98 (2014).

104. Babst, M. Quality control: quality control at the plasma membrane: one mechanism does not fit all. *J. Cell Biol.* **205,** 11–20 (2014).

105. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

106. Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat. Commun.* **10,** 1–15 (2019).

107. Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis. *Cell Systems* **9,** 569–579.e7 (2019).

108. Matsson, P. & Kihlberg, J. How Big Is Too Big for Cell Permeability? *J. Med. Chem.* **60,** 1662–1664 (2017).

109. Edfors, F., Danielsson, F., Hallström, B. M., Käll, L., Lundberg, E., Pontén, F., Forsström, B. & Uhlén, M. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **12,** (2016).

110. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

111. Cambridge, S. B., Gnad, F., Nguyen, C., Bermejo, J. L., Krüger, M. & Mann, M. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J. Proteome Res.* **10,** 5275–5284 (2011).

112. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D. & Schomburg, D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* **49,** D498–D508 (2021).

113. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5,** 93–121 (2010).

114. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13,** 344–349 (2010).

115. Ma, D., Yang, L., Fleming, R. M. T., Thiele, I., Palsson, B. O. & Saunders, M. A. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. *Sci. Rep.* **7,** 40863 (2017).

116. Yang, L., Ma, D., Ebrahim, A., Lloyd, C. J., Saunders, M. A. & Palsson, B. O. solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinformatics* **17,** 391 (2016).

117. Wang, Y., Eddy, J. A. & Price, N. D. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* **6,** 153 (2012).