UNIVERSITY OF CALIFORNIA

Los Angeles

Mechanistic and data-driven antibody response modeling strategies

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

by

Cyrillus Tan

2024

ABSTRACT OF THE DISSERTATION

Mechanistic and data-driven
antibody response modeling strategies

by

Cyrillus Tan
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2024
Professor Aaron S. Meyer, Chair

Antibodies are essential to adaptive immunity and therapeutic development. IgG antibodies coordinate immune effector responses by selectively binding to target antigens and interacting with various effector cells via Fcγ receptors. In this study, I explore two computational strategies for modeling antibody responses. First, I extend and employ a mechanistic model to analyze mixed Fc IgG binding measurements. This multivalent binding model efficiently predicts interactions between mixtures of multiple multivalent ligands and multiple cell surface receptors. Applied to experimental data, this model quantitatively matches mixed FcγR binding measurements, refines affinity estimates, and predicts antibody-mediated immune effector cell responses. Notably, it highlights IgG2's binding capabilities to FcγRI, contrary to previous nonbinding estimations. Second, I adopt a data-driven approach using tensor-based methods to deconvolute systems serology data. Given the complexity of recent biological research characterized by measurements in multiple degrees of variation, I provide an overview of applying tensor methods to high-throughput biological datasets. Applied these principles to HIV- and SARS-CoV-2- infected patients' serum sample data, tensor methods reveal consistent patterns and outperform traditional methods in data reduction and prediction accuracy, emphasizing their efficacy in identifying immune functional responses and disease status. Overall, this study demonstrates how mechanistic and data-driven approaches can be effectively applied to analyze antibody-mediated immunity, showcasing their distinct roles in computational biology.

The dissertation of Cyrillus Tan is approved.

Alexander Hoffmann

Jingyi Li

Elaine F. Reed

Aaron S. Meyer, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1. Introduction

Chapter 2. A general model of multivalent binding with ligands of heterotypic subunits and multiple surface receptors

Chapter 3. Mixed IgG Fc immune complexes exhibit blended binding profiles and refine FcR affinity estimates

Chapter 4. **The structure is the message: preserving experimental context through tensor decompositions**

Chapter 5. **Tensor-structured decomposition improves systems serology analysis**

LIST OF TABLES

Chapter 3. Mixed IgG Fc immune complexes exhibit blended binding profiles and refine FcR affinity estimates

Chapter 4. **The structure is the message: preserving experimental context through tensor decompositions**

Chapter 5. **Tensor-structured decomposition improves systems serology analysis**

ACKNOWLEDGMENTS

First and foremost, I want to express my profound gratitude to my advisor, Aaron Meyer. His unwavering support for my project progressions and career development has been pivotal. With his mentorship, I have not only gained profound insights into scientific research but also learned effective work practices. I deeply appreciate his trust in me and the opportunities he has provided. From detailed guidance to well-conceived projects, his mentorship has been instrumental to my scholarly growth. His accessibility and flexibility have created a nurturing laboratory.

I am very fortunate to be part of an exceptional scientific community. My committee has provided not only scientific guidance but also substantial support for my professional endeavors. My collaborators, who are leading experts in antibody immunity, have demonstrated remarkable professionalism. I am also grateful for the invaluable connections I made through the bioinformatics program, where my cohort has maintained strong camaraderie, fellow students have formed a vibrant social network, and alumni have generously offered abundant career support.

The past few years have been a transformative journey of personal growth. Throughout, I have learned to invest in wellness, lead a well-rounded life, and embrace previously daunting experiences. You win some, you learn some, making it essential to foster self-acceptance, resilience in challenging times, and perseverance through setbacks. It is evident that the curiosity for knowledge and the love of man have driven my path hitherto and beyond.

I have been sustained by the perpetual companionship of Anthony, Robert, Nick, Jack, Ryan, Max, Jian, and many other friends, which has continually uplifted my spirit. My time at the University of California and in Los Angeles has introduced me to some incredible people who have brought me immense joy and adventure, and whom I intend to hold close for years to come.

I thank my family for their support of my pursuit.

Chapter 2 is adapted from “A general model of multivalent binding with ligands of heterotypic subunits and multiple surface receptors”, written by myself and A.S. Meyer, published in *Mathematical Biosciences* 342 (2021): 108714.

Chapter 3 is adapted from “Mixed IgG Fc immune complexes exhibit blended binding profiles and refine FcR affinity estimates”, written by myself along with A. Lux, M. Biburger, P. Varghese, S. Lees, F. Nimmerjahn, and A.S. Meyer, published in *Cell Reports* 42.7 (2023).

Chapter 4 is adapted from “The structure is the message: preserving experimental context through tensor-based analyses”, written by myself and A.S. Meyer, published in *Cell Systems* 15.8 (2024): 679-693.

Chapter 5 is adapted from “Tensor-structured decomposition improves systems serology analysis”, written by myself along with M.C. Murphy, H.S. Alpay, S.D. Taylor, and A.S. Meyer, published in *Molecular Systems Biology* 17.9 (2021): e10243.

VITA

Education

2014 – 2018	B.S. in Biology and B.S. in Computer Science, *magna cum laude*
	Emory University, Atlanta, GA
2018 – present	Ph.D. candidate in Bioinformatics
	University of California, Los Angeles, Los Angeles, CA

Publications

Tan, ZC, MC Murphy, HS Alpay, SD Taylor, and AS Meyer. "Tensor-structured decomposition improves systems serology analysis." *Molecular Systems Biology* 17.9 (2021): e10243.

Tan, ZC, and AS Meyer. "A general model of multivalent binding with ligands of heterotypic subunits and multiple surface receptors." *Mathematical Biosciences* 342 (2021): 108714.

Tan, ZC, BT Orcutt-Jahns, and AS Meyer. "A quantitative view of strategies to engineer cell-selective ligand binding." *Integrative Biology* 13.11 (2021): 269-282.

Mitchell, S, R Tsui, ZC Tan, A Pack, and A Hoffmann. "The NF-κB multi-dimer system model: a knowledge base to explore diverse biological contexts" *Science Signaling* 16.776 (2023): eabo28.

Tan, ZC, A Lux, M Biburger, P Varghese, S Lees, F Nimmerjahn, AS Meyer. "Mixed IgG Fc immune complexes exhibit blended binding profiles and refine FcR affinity estimates" *Cell Reports* 42.7 (2023): 112734.

Chin, JL, ZC Tan, LC Chan, …, A Hoffmann, VG Fowler Jr, EF Reed, MR Yeaman, AS Meyer, MRSA Systems Immunobiology Group. "Tensor modeling of MRSA bacteremia cytokine and transcriptional patterns reveals coordinated, outcome-associated immunological programs." *PNAS Nexus* 3.5 (2024): pgae185.

Tan, ZC, and AS Meyer. "The structure is the message: preserving experimental context through tensor-based analyses" *Cell Systems* 15.8 (2024): 679-693.

Chapter 1

Introduction

The unexamined life is not worth living.

Socrates

1.1 Background

Antibodies are essential components of adaptive immunity and versatile platforms for therapeutic development. They coordinate immune effector responses by selectively binding to target antigens with their antigen-binding (Fab) regions and interacting with various immune effector cells via their fragment-crystallizable (Fc) regions[1]. To comprehensively understand an antibody's role in promoting immunity, it is crucial to evaluate both its antigen-binding capabilities and its Fc properties. This study aims to develop computational tools to model antibody responses by integrating these aspects.

The diversity among antibody Fc regions is substantial, encompassing different antibody types, subclasses, and glycosylation patterns. Immunoglobulin G (IgG), the most abundant and therapeutically capable antibody, directs effector responses by binding to Fcγ receptors (FcγRs) through its Fc region. FcγR activation begins with the engagement of several antibodies on an antigen target, forming an immune complex (IC). The response triggered by this IC is determined by the specific Fc regions and Fc receptors involved, each differing in signaling effects and expression patterns[2]. Given that physiological antibody responses are typically polyclonal, consisting of a mixture of antibodies with diverse antigen specificities and Fc regions, analyzing the effects of Fc mixtures is crucial for a comprehensive understanding of antibody-mediated immunology.

While pieces of Fc receptor pathways have been well characterized[3,4], the diversity of Fc regions and Fc receptors and their complexity of binding interactions pose a significant challenge in analyzing antibody Fc mixtures. Fully understanding these interactions requires evaluating the collective impact of all possible interactions and accounting for multiple and mixed ligands, multiple receptors, and multivalent binding formed by ICs. To analyze the experimental measurements of mixed Fc antibodies, I propose to extend and employ a mechanistic binding model. This model aims to provide an integrated view, allowing us to identify gaps in our current understanding.

Recent technological advancements in systems serology promise to broaden our understanding of antibody-mediated protection[5]. This multiplex Fc assay aims to profile the humoral immune response by jointly quantifying sample serum antibodies' antigen-binding and Fc biophysical properties in parallel. These measurements have proven highly predictive of effector cell-elicited responses and overall antibody-elicited immune protection. However, analyzing systems serology data presents significant challenges. I propose representing the data as a three-dimensional tensor to separate the antigen and receptor-level variations. This tensor can be effectively analyzed using tensor decomposition, a data-driven, unsupervised dimensionality reduction method[6].

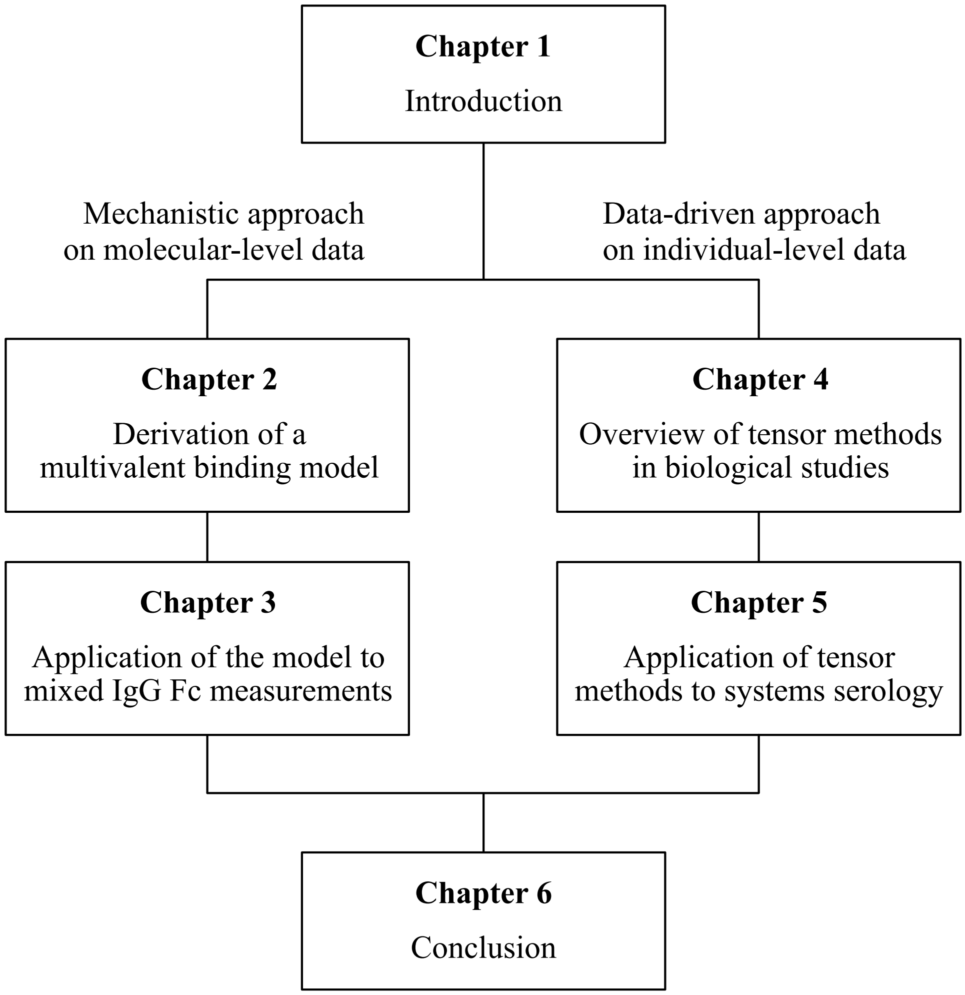

Chapter 1
Introduction
Mechanistic approach on molecular-level data
Data-driven approach on individual-level data
Chapter 2
Derivation of a multivalent binding model
Chapter 4
Overview of tensor methods in biological studies
Chapter 3
Application of the model to mixed IgG Fc measurements
Chapter 5
Application of tensor methods to systems serology
Chapter 6
Conclusion

Figure 1.1. Structure of this dissertation

Mechanistic and data-driven models are based on different epistemic philosophies and are employed for molecular and individual-level data. Each has distinct uses in studying antibody responses and independently derives unique insights. In this dissertation, I present both approaches, aiming to provide a more comprehensive picture of antibody-mediated immunity.

1.2 Overview of this dissertation

In this dissertation, I employ two distinct computational approaches to dissect the effects of antibody Fc mixtures: a bottom-up mechanistic approach analyzing molecular-level experiments and a top-down data-driven approach scrutinizing individual-level measurements. The structure of this dissertation is illustrated in Fig. 1.1.

The first approach, detailed in **Chapters 2 and 3**, involves developing a biophysical model to analyze mixed Fc IgG binding measurements in controlled experiments. This begins with advancing computational ligand-receptor binding modeling capabilities. Multivalent cell surface receptor binding is common in biology, with significant functional and therapeutic implications. In **Chapter 2**, I extend a general mechanistic multivalent binding model to accommodate a large number of interactions between multiple ligands and receptors. The model enables large-scale predictions of mixture binding and the binding space of a ligand, offering an efficient framework for analyzing multivalent binding.

In **Chapter 3**, I apply this binding model to experimental measurements of FcγR binding to mixed Fc ICs. I find that the binding of these mixtures aligns along a continuum between pure cases and quantitatively matches my model, except for several low-affinity interactions primarily involving IgG2. The model can provide refined estimates of these affinities. Finally, I demonstrate that the model can also predict effector cell-elicited platelet depletion in humanized mice. This chapter establishes an *in vitro* to *in vivo* quantitative framework for modeling mixed IgG Fc-effector cell regulation.

The second approach, outlined in **Chapters 4 and 5**, applies a tensor-based, data-driven machine learning method to deconvolute systems serology measurements. In **Chapter 4**, I review the application of tensor methods in biological studies and explore their optimal applications. Tensor methods can find wide applications in current biological studies with the rise of multiplex and high-throughput assays, which profile cell responses across various experimental parameters. I review how tensor-based analyses and decompositions can preserve multivariate experimental structures, arguing that tensor methods are poised to become integral to the biomedical data sciences toolkit.

In **Chapter 5**, I employ tensor methods to analyze systems serology measurements. Using measurements from studies of HIV- and SARS-CoV-2-infected subjects as exemplars, tensor methods outperform standard methods like principal component analysis in data reduction while maintaining equivalent prediction of immune functional responses and disease status. Model interpretation improves through effective data reduction, separation of Fc and antigen-binding effects, and recognition of consistent patterns across individual measurements. Thus, tensor methods are effective strategies for data exploration in systems serology.

Finally, in **Chapter 6**, I provide a high-level overview of the relative merits of mechanistic and statistical approaches and summarize their roles in computational biology studies beyond antibody response modeling.

Chapter 2

A general model of multivalent binding with ligands of heterotypic subunits and multiple surface receptors

What I cannot create, I do not understand.

Richard Feynman

2.1 Introduction

Binding to extracellular ligands is among the most fundamental and universal activities of a cell. Many important biological activities, and cell-to-cell communication in particular, are based on recognizing extracellular molecules via specific surface receptors. For example, multivalent ligands are common extracellular factors in the immune system[7], and computational models have been applied to study IgE-FcεRI[8], MHC-T cell receptor[9], and IgG-FcγR interactions[10]. However, these models are specific to their biological applications, limited to a single homogenous ligand and receptor[11], or fail to scale with valency[12].

Multivalent binding to various receptors on a cell can be accounted for by the kinetics of individual association reactions between each monomer-receptor pair. However, when the complexes contain multiple ligand monomers of either the same or different kinds, and when there is a mixture of complexes with either the same or different valencies, the system becomes complicated: different binding orders of units on a complex create a combinatorically large amount of possible reactions, and the competition among different kinds of ligands and complexes impedes intuitive understanding. In this case, enumerating all binding configurations and reactions becomes impractical.

In this chapter, we extend a simple two-step, multivalent binding model to cases involving multiple receptors and ligand subunits[9,11,13–15]. By harnessing the power of combinatorics via applying the multinomial theorem and focusing on macrostates, we can predict the amount of binding for each ligand and receptor at equilibrium. We derive macroscopic quantities for both specifically arranged and randomly assorted complexes and demonstrate how this model enables large-scale predictions on mixture binding and the binding space of a ligand.

Our model provides both generality and computational efficiency, allowing large-scale predictions such as characterizing synergism using a mixture of ligands and depicting the general binding behavior of a compound. The compactness and elegance of the formulae enable both analytical and numerical analyses, in turn allowing for the construction of higher-level computational tools. We expect this binding model will be widely applicable to many biological contexts.

2.2 Preliminaries

Vector and matrix notation

In this chapter, we denote a vector in boldface letter and its entry in the same letter but with subscript and not in boldface, e.g. $\mathbf{C} = [C_1, C_2, \ldots, C_n]$. The sum of elements for a vector is denoted as $|\mathbf{C}| = \sum_{i=1}^{n} C_i$.

For any matrix (A_{ij}) of size $m \times n$, we denote the vector formed by its i-th row as $\mathbf{A_{i\bullet}} = [A_{i1}, A_{i2}, \cdots, A_{in}]$, and the vector formed by its j-th column as $\mathbf{A_{\bullet j}} = [A_{1j}, A_{2j}, \cdots, A_{mj}]$. The row sums of matrix (A_{ij}), therefore, can be written as $|\mathbf{A_{1\bullet}}|$, $|\mathbf{A_{2\bullet}}|$, $\cdots$, $|\mathbf{A_{m\bullet}}|$, and column sums $|\mathbf{A_{\bullet 1}}|$, $|\mathbf{A_{\bullet 2}}|$, $\cdots$, $|\mathbf{A_{\bullet n}}|$.

In this chapter, multinomial coefficients such as n choose $k_1, k_2, \cdots, k_n$ will be written as

$$\binom{n}{\mathbf{k}} = \begin{pmatrix} & & n & \\ k_1 & k_2 & \cdots & k_n \end{pmatrix} = \frac{n!}{k_1!\,k_2!\cdots k_n!}.$$

The implicit assumption here is that $|\mathbf{k}| = n$, and each $k_i \in \mathbb{N}$.

Some useful theorems in combinatorics

From the binomial theorem, we know that

$$\sum_{i=0}^{f} \binom{f}{i} \Phi^i = (1+\Phi)^f.$$

Differentiating both sides by Φ, we get

$$\sum_{i=0}^{f} i \binom{f}{i} \Phi^i = f\Phi(1+\Phi)^{f-1}.$$

We can derive similar property from the multinomial theorem. Assume the elements of a nonnegative integer vector $\mathbf{q}$ add up to f, or $|\mathbf{q}| = f$. Given another nonnegative vector $\boldsymbol{\varphi}$ with sum of elements $|\boldsymbol{\varphi}|$, we have

$$\sum_{|\mathbf{q}|=f} \binom{f}{\mathbf{q}} \prod_i \varphi_i{}^{q_i} = |\boldsymbol{\varphi}|^f.$$

Differentiate both sides by φ_m where φ_m can be any entry of $\boldsymbol{\varphi}$, we have

$$\sum_{|\mathbf{q}|=f} \binom{f}{\mathbf{q}} q_m \prod_i \varphi_i{}^{q_i} = \varphi_m f |\boldsymbol{\varphi}|^{f-1}.$$

We can multiply two independent multinomial theorem equations together, too. Let $\mathbf{u}$ and $\mathbf{v}$ be two nonnegative integer vectors, $\mathbf{a}$ and $\mathbf{b}$ be two nonnegative vectors, and $|\mathbf{u}| = m$, $|\mathbf{v}| = n$, we have

$$\sum_{\substack{|\mathbf{u}|=m\\|\mathbf{v}|=n}} \binom{m}{\mathbf{u}}\binom{n}{\mathbf{v}}\prod_i a_i^{u_i}\prod_j b_j^{v_j} = |\mathbf{a}|^m|\mathbf{b}|^n.$$

Throughout this chapter, we consolidate multiple summation symbols into one. In this case, we use $\sum_{|\mathbf{u}|=m,|\mathbf{v}|=n}$ as a shorthand for $\sum_{|\mathbf{u}|=m}\sum_{|\mathbf{v}|=n}$. From the combinatorics property, we can derive the sum of a linear combination of two exponents from each multinomial term as

$$\sum_{\substack{|\mathbf{u}|=m\\|\mathbf{v}|=n}} \binom{m}{\mathbf{u}}\binom{n}{\mathbf{v}}\Big(\underbrace{k_1u_p + k_2v_q}_{\text{linear combination}}\Big)\prod_i a_i^{u_i}\prod_j b_j^{v_j}$$

$$= k_1a_pm|\mathbf{a}|^{m-1}|\mathbf{b}|^n + k_2|\mathbf{a}|^mb_qn|\mathbf{b}|^{n-1}$$

$$= \left[\frac{k_1ma_p}{|\mathbf{a}|} + \frac{k_2nb_q}{|\mathbf{b}|}\right]|\mathbf{a}|^m|\mathbf{b}|^n,$$

where k_1 and k_2 can be any constant.

We can extend this to the product of N multinomial equations. Let $\mathbf{q_1},\cdots,\mathbf{q_N}$ be N nonnegative integer vectors, each with $|\mathbf{q_i}| = \theta_i$, and $\boldsymbol{\psi}_\mathbf{1},\cdots,\boldsymbol{\psi}_\mathbf{N}$ be N nonnegative vectors. The sum of any linear combination of exponent terms $\sum_r k_r\, q_{s_rt_r}$, where k_r's can be any constant, and each $q_{s_rt_r}$ is the t_r-th element of $\mathbf{q}_{\mathbf{s_r}}$, can be calculated as

$$\sum_{\substack{|\mathbf{q_1}|=\theta_1\\ \cdots\\ |\mathbf{q_N}|=\theta_N}} \left(\sum_r k_r\, q_{s_rt_r}\right)\prod_{i=1}^{N}\binom{\theta_i}{\mathbf{q_i}}\left(\prod_j \psi_{ij}^{q_{ij}}\right) = \left[\sum_r \frac{k_r\theta_{s_r}\psi_{s_rt_r}}{|\boldsymbol{\psi}_{\mathbf{s_r}}|}\right]\prod_{i=1}^{N}|\boldsymbol{\psi}_\mathbf{i}|^{\theta_i}.$$

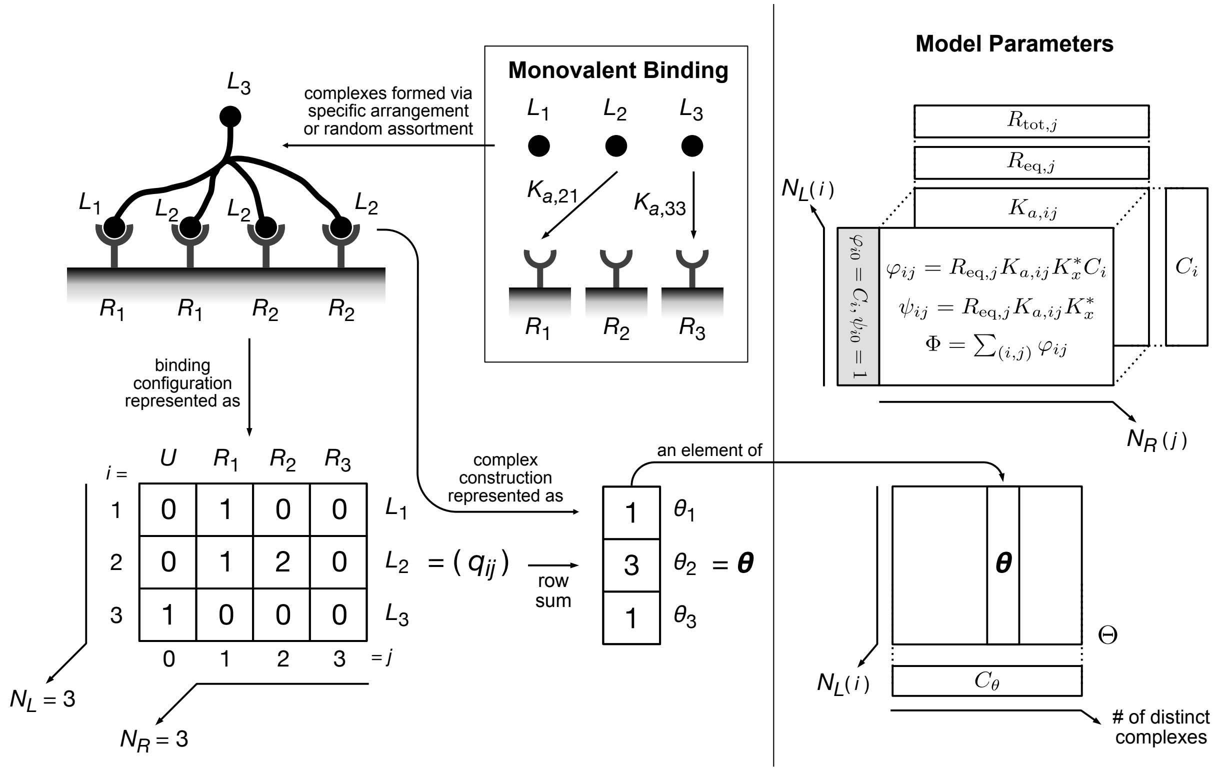


Figure 2.1 General setup of the model.

In this study, we investigate the binding behavior of complexes formed by monomer ligands in either a specific arrangement or by random assortment. We propose that the binding configuration between a complex and several receptors on a cell can be described as a matrix (q_{ij}). The construction of a complex can be written as a vector $\boldsymbol{\theta}$. The figure shows the dimensions of the model's parameters: C_i, the monomer compositions, are in a vector of length N_L; $R_{\text{tot},j}$ and $\mathrm{R}_{\mathrm{eq,j}}$, the receptor expression and equilibrium levels are in vectors of length N_R; the binding affinities, $K_{a,ij}$, are in a matrix of dimension $N_L \times N_R$; φ_{ij} and ψ_{ij} are in the matrices of dimension $N_L \times (N_R + 1)$. Θ is a set of all possible $\boldsymbol{\theta}$'s, with C_θ as each of their compositions. Each $\boldsymbol{\theta}$ is a vector of length N_L, and $C_{\boldsymbol{\theta}}$ should be in a vector of the same size as Θ.

2.3 Model setup

Parameters and notations

In this chapter, we investigate the binding between multivalent ligand complexes and a cell expressing various surface receptors. As shown in Fig. 2.1, we consider N_L types of distinct monomer ligands, namely L_1, L_2, ..., L_{N_L}, and N_R types of distinct receptors expressed on a cell, namely R_1, R_2, ..., R_{N_R}. The monovalent binding association constant between L_i and R_j is defined as $K_{a,ij}$. A ligand complex consists of one or several monomer ligands, and each of them can bind to a receptor independently. Its construction can be described by a vector $\mathbf{\theta} = [\theta_1, \theta_2, \ldots, \theta_{N_L}]$, where each entry θ_i represents how many L_i this complex contains. The sum of elements of vector $\mathbf{\theta}$, $|\mathbf{\theta}|$, is f, the valency of this complex.

The binding configuration at equilibrium between an individual complex and a cell expressing various receptors can be described as a matrix (q_{ij}) with N_L rows and $(N_R + 1)$ columns. For example, the complex bound as shown on the top left corner in Fig. 2.1 can be described as the matrix below it. q_{ij} represents the number of L_i to R_j binding, and q_{i0}, the entry on the 0-th column, is the number of unbound L_i on that complex in this configuration. This matrix can be unrolled into a vector form $\mathbf{q} = [q_{10}, q_{11}, \ldots, q_{1N_R}, q_{20}, \ldots, q_{2N_R}, q_{30}, \ldots, q_{N_L N_R}]$ of length $N_L(N_R + 1)$. Note that this binding configuration matrix (q_{ij}) only records how many L_i-to-R_j pairs are formed, regardless of which exact ligand on the complex binds. For example, in Fig. 2.1, swapping the two L_2's binding to R_2's will give us the same configuration matrix. Therefore, we will need to account for this combinatorial factor when applying the law of mass action.

We know from the conservation of mass that for this complex, $\theta_i = q_{i0} + q_{i1} + q_{i2} + \cdots + q_{iN_R} = |\mathbf{q_{i\bullet}}|$ must hold for all i's. Mathematically, vector $\mathbf{\theta}$ is the row sums of matrix (q_{ij}). The corresponding $\boldsymbol{\theta}$ of a binding configuration $\mathbf{q}$, $\boldsymbol{\theta}(\mathbf{q})$, written in the format of a function of $\mathbf{q}$, can be determined by this relationship. Also, the sum of elements in $\mathbf{q}$, $|\mathbf{q}| = f$, will be the valency. The concentration of complexes in the solution is L_0 (not to be confused with L_i, the name of ligands, when $i = 1,2,\cdots,N_L$). It is the concentration of all ligands at the equilibrium state. It is approximately the same as the initial ligand concentration if the amount of ligands is much greater than that of the receptors and binding event does not significantly deplete the ligand concentration. On the receptor side, $R_{\mathrm{tot},i}$ is the total number of R_i expressed on the cell surface. This usually can be measured experimentally. $R_{\mathrm{eq},i}$ is the number of unbound R_i on a cell at the equilibrium state during the ligand complex-receptor interaction, and usually must be calculated from $R_{\mathrm{tot},i}$ as we will explain later.

The binding of a ligand complex, a large molecule, is complicated. To simplify the matter, we will need to make some key thermodynamic assumptions. In this model, we make two assumptions on the binding dynamics:

1. The initial binding of L_i on a free (unbound) complex to a surface receptor R_j has the same affinity (association constant, $K_{a,ij}$) as that of a monomer ligand L_i;
2. In order for the detailed balance to hold, the association constant of any subsequent binding event on the surface of a cell after the initial interaction must be proportional to their corresponding monovalent affinity. We assume the subsequent binding affinity in multivalent interactions between L_i and R_j to be $K_x^* K_{a,ij}$.

K_x^* is a term coined as the crosslinking constant. It captures the difference between free and multivalent ligand-receptor binding, including but not limited to steric effects and local receptor clustering[16]. In practice this term is often fit to apply this model to a specific biological context.

We create two last variables that will help to simplify our equations. For all i in $\{1,2,\dots,N_L\}$, we define $\psi_{ij} = R_{\text{eq},j}K_{a,ij}K_x^*$ and $\varphi_{ij} = R_{\text{eq},j}K_{a,ij}K_x^*C_i$ where $j = \{1,2,\dots,N_R\}$, and we define $\psi_{i0} = 1$, $\varphi_{i0} = C_i$. Therefore, $\varphi_{ij} = \psi_{ij}C_i$ holds for all i and j. Then we define the sum of this new matrix (φ_{ij}) as $\sum_{i=1}^{N_L}\sum_{j=1}^{N_R}\varphi_{ij} = \Phi$, and $\sum_{i=1}^{N_L}\sum_{j=0}^{N_R}\varphi_{ij} = \Phi + \sum_{i=1}^{N_L}C_i = 1 + \Phi$. The rationale of these definitions will become clear in future sections.

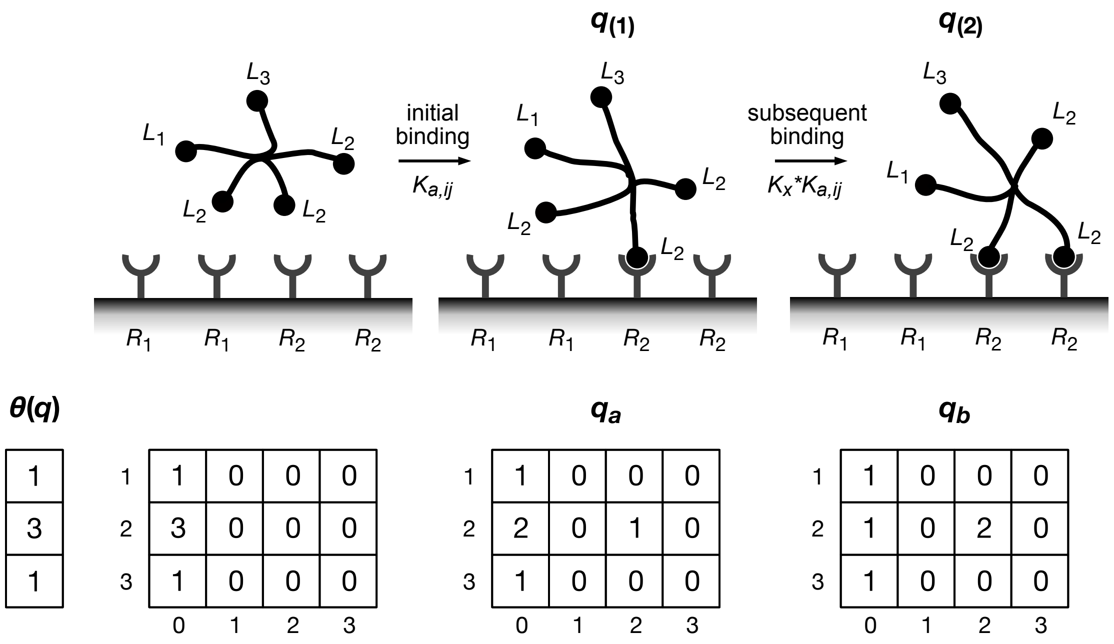
The amount of a specific binding configuration

With the definitions of our model we can now derive the amount of complexes bound with the configuration described as $\mathbf{q}$ on a cell at equilibrium, $v_{\mathbf{q}}$. We know that the composition of any complex can be described by a vector $\boldsymbol{\theta}$ of length N_L, where each entry θ_i represents the number of monomers L_i within the complex. We can enumerate all possible binding configurations of $\boldsymbol{\theta}$ complex by filling the matrix (q_{ij}) with any nonnegative integer values so long as its row sums equal $\boldsymbol{\theta}$. Conversely, for a certain binding configuration, $\mathbf{q}$, the construction of the complex involved must be its row sum, $\boldsymbol{\theta}(\mathbf{q})$, and the concentration of this complex is $L_0 C_{\boldsymbol{\theta}(\mathbf{q})}$. If the corresponding complex $\boldsymbol{\theta}(\mathbf{q})$ does not exist in the solution, we set $C_{\boldsymbol{\theta}(\mathbf{q})} = 0$. Since $\boldsymbol{\theta}(\mathbf{q})$ is defined only by the ligand concentration at equilibrium, it will remain $L_0 C_{\boldsymbol{\theta}(\mathbf{q})}$.

Initial binding

We start with the initial binding reaction of a complex, L_i-to-R_j. As shown in Fig. 2.2, the reactants are the free complexes and the free receptors R_j (in this case R_2), and the product are the L_i-to-R_j (in this case L_2-R_2) monovalently bound complexes $\mathbf{q}_{(\mathbf{1})}$. We denote the amount of this new complex as $v_{\mathbf{q}_{(\mathbf{1})}}$. The concentration of free complexes is $L_0 C_{\boldsymbol{\theta}(\mathbf{q}_{(\mathbf{1})})}$. The equilibrium constant for this reaction is $K_{a,ij}$. Therefore, we have

$$v_{\mathbf{q}_{(\mathbf{1})}} = L_0 C_{\boldsymbol{\theta}(\mathbf{q}_{(\mathbf{1})})} R_{\text{eq},j} K_{a,ij}.$$



$$\theta(q) = \begin{array}{|c|}\hline 1 \\ \hline 3 \\ \hline 1 \\ \hline\end{array} \qquad \begin{array}{c|c|c|c|c|} \cline{2-5} 1 & 1 & 0 & 0 & 0 \\ \cline{2-5} 2 & 3 & 0 & 0 & 0 \\ \cline{2-5} 3 & 1 & 0 & 0 & 0 \\ \cline{2-5} \multicolumn{1}{c}{} & \multicolumn{1}{c}{0} & \multicolumn{1}{c}{1} & \multicolumn{1}{c}{2} & \multicolumn{1}{c}{3} \end{array}$$

$$q_a = \begin{array}{c|c|c|c|c|} \cline{2-5} 1 & 1 & 0 & 0 & 0 \\ \cline{2-5} 2 & 2 & 0 & 1 & 0 \\ \cline{2-5} 3 & 1 & 0 & 0 & 0 \\ \cline{2-5} \multicolumn{1}{c}{} & \multicolumn{1}{c}{0} & \multicolumn{1}{c}{1} & \multicolumn{1}{c}{2} & \multicolumn{1}{c}{3} \end{array} \qquad q_b = \begin{array}{c|c|c|c|c|} \cline{2-5} 1 & 1 & 0 & 0 & 0 \\ \cline{2-5} 2 & 1 & 0 & 2 & 0 \\ \cline{2-5} 3 & 1 & 0 & 0 & 0 \\ \cline{2-5} \multicolumn{1}{c}{} & \multicolumn{1}{c}{0} & \multicolumn{1}{c}{1} & \multicolumn{1}{c}{2} & \multicolumn{1}{c}{3} \end{array}$$

Figure 2.2 The scheme of cell-complex binding step by step.

We assume the initial binding event has the same affinity as monomer binding, $K_{a,ij}$, while subsequent binding has an association constant scaled by K_x^*, the crosslinking constant. Each binding configuration scheme above can be described by the **q** right below, if we ignore the statistical factors. $\boldsymbol{\theta}(\mathbf{q})$ is the construction of the complex and can be implied from **q**.

While the binding configuration of $\mathbf{q}_{(\mathbf{1})}$ can be described by $\mathbf{q}_{\mathbf{a}}$, the total amount of complexes that bind as described as $\mathbf{q}_{\mathbf{a}}$ may not be the same as $v_{\mathbf{q}_{(\mathbf{1})}}$, since $\mathbf{q}_{\mathbf{a}}$ does not consider the number of ways this binding L_i can be chosen. An equivalent explanation is that, $\mathbf{q}_{(\mathbf{1})}$ is only one possible microstate to achieve the $\mathbf{q}_{\mathbf{a}}$ configuration, and we need to count the total number of possible microstates for $\mathbf{q}_{\mathbf{a}}$. Accounting for this statistical factor, we have

$$v_{\mathbf{q}_{\mathbf{a}}} = v_{\mathbf{q}_{(1)}} \binom{\theta_i}{1} = v_{\mathbf{q}_{(1)}} \binom{\theta_i}{\mathbf{q}_{\mathbf{a},\mathbf{i}\bullet}},$$

$$v_{\mathbf{q}_{\mathbf{a}}} = L_0 C_{\boldsymbol{\theta}(\mathbf{q}_{\mathbf{a}})} R_{\mathrm{eq},j} K_{a,ij} \binom{\theta_i}{\mathbf{q}_{\mathbf{a},\mathbf{i}\bullet}},$$

since $\boldsymbol{\theta}(\mathbf{q}_{(1)}) = \boldsymbol{\theta}(\mathbf{q}_{\mathbf{a}})$. $\mathbf{q}_{\mathbf{a},\mathbf{i}\bullet}$ is a vector formed by the i-th row of $\mathbf{q}_{\mathbf{a}}$. For example, in Fig. 2.2, $\mathbf{q}_{\mathbf{a},\mathbf{2}\bullet} = [2,0,1,0]$. Conceptually, $\binom{\theta_i}{\mathbf{q}_{\mathbf{a},\mathbf{i}\bullet}}$ can be understood as the number of ways to split θ_i L_i's into q_{i0} of unbound, q_{i1} of R_1-bound, q_{i2} of R_2-bound, ..., and q_{iN_R} of R_{N_R}-bound units. In the initial binding reaction, only q_{i0} and q_{ij} will be nonzero, with $q_{i0} = \theta_i - 1$ and $q_{ij} = 1$, so it is effectively the same as $\binom{\theta_i}{1}$. However, the multinomial coefficient expression can be generalized to subsequent binding steps.

Subsequent binding

For a subsequent binding between L_i and R_j (i and j are not necessarily the same as in initial binding), we have the reactants as a bound complex, $\mathbf{q}_{(\mathbf{1})}$, and a free receptor R_j (in the case shown by Fig. 2.2, R_2), while the product is another bound complex, $\mathbf{q}_{(\mathbf{2})}$. The equilibrium constant is $K_x^* K_{a,ij}$, then

$$v_{\mathbf{q}_{(2)}} = v_{\mathbf{q}_{(1)}} R_{\mathrm{eq},j} K_x^* K_{a,ij}.$$

To account for the statistical factors of $v_{\mathbf{q_b}}$, we have $v_{\mathbf{q_b}} = v_{\mathbf{q}_{(2)}} \binom{\theta_i}{\mathbf{q_{b,i\bullet}}}$. For example, in Fig. 2.2, $\mathbf{q_{b,2\bullet}} = [1,0,2,0]$. Putting these together, we have

$$v_{\mathbf{q_b}} = v_{\mathbf{q_a}} R_{\mathrm{eq},j} K_x^* K_{a,ij} \frac{\binom{\theta_i}{\mathbf{q_{b,i\bullet}}}}{\binom{\theta_i}{\mathbf{q_{a,i\bullet}}}}.$$

By recursion, we can solve $v_{\mathbf{q}}$ for any $\mathbf{q}$ from these equations. It is

$$v_{\mathbf{q}} = \frac{L_0 C_{\boldsymbol{\theta}(\mathbf{q})}}{K_x^*} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} (R_{\mathrm{eq},j} K_x^* K_{a,ij})^{q_{ij}} \prod_{i=1}^{N_L} \binom{\theta_i}{\mathbf{q_{i\bullet}}}$$

$$= \frac{L_0 C_{\boldsymbol{\theta}(\mathbf{q})}}{K_x^*} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}} \prod_{i=1}^{N_L} \binom{\theta_i}{\mathbf{q_{i\bullet}}}$$

$$= \frac{L_0 C_{\boldsymbol{\theta}(\mathbf{q})}}{K_x^*} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}} \prod_{i=1}^{N_L} \binom{\theta_i}{\mathbf{q_{i\bullet}}}$$

if we define $\psi_{ij} = R_{\mathrm{eq},j} K_x^* K_{a,ij}$ for $j = 1,2,\cdots,N_R$ and $\psi_{i0} = 1$ for all i. $\prod_{(i,j)=(1,0)}^{(N_L,N_R)}$ is a shorthand for $\prod_{i=1}^{N_L} \prod_{j=0}^{N_R}$. In the next section, we will use this formula repeatedly.

Notice that this equation is only for surface-bound complexes, not suitable for calculating the concentration of unbound $\mathbf{q}$, when every nonzero values are on its 0-th column. The concentration of unbound ligands should always be $L_0 C_{\boldsymbol{\theta}(\mathbf{q})}$. However, for algebraic convenience, we allow such definition but only to subtract them later, and will name it $v_{0,\mathrm{eq}}$ which equals $L_0 C_{\boldsymbol{\theta}(\mathbf{q})} / K_x^*$.

2.4 Macroscopic equilibrium predictions

From here we will investigate the macroscopic properties of binding, such as the total amount of ligand bound and receptor bound on a cell surface at equilibrium. We consider two different ways complexes in the solution can be formed. First, complexes may come in a specific arrangement. In this case, the structure and exact concentration for each complex are designed and known. Alternatively, ligand monomers of known proportion can congregate into complexes with a fixed valency f. Through random assortment, any combination of f monomer ligands can form a complex, and their concentration will follow a multinomial distribution. We will explore these two cases separately.

Complexes formed in a specific arrangement

When complexes are specifically arranged, the structure and proportion of each kind are well-defined. To formulate this mathematically, we assume that we have various kinds of complexes, and each of them can be described by a vector $\boldsymbol{\theta}$ of length N_L, with each entry θ_i as the number of L_i in this complex. The valency of each complex may be different, and for complex $\boldsymbol{\theta}$ its valency is $|\boldsymbol{\theta}|$. The proportion of $\boldsymbol{\theta}$ among all complexes is defined as $C_{\boldsymbol{\theta}}$, and the concentration of each $\boldsymbol{\theta}$ complex will be $L_0 C_{\boldsymbol{\theta}}$. For example, if we create a mixture of 20% of bivalent L_1 and 80% of bispecific $L_1 - L_2$, then $\boldsymbol{\theta_1} = [2,0]$, $\boldsymbol{\theta_2} = [1,1]$, $C_{\boldsymbol{\theta_1}} = 20\%$, and $C_{\boldsymbol{\theta_2}} = 80\%$. If the mixture solution has a total concentration of 10 nM, then the concentration of $\boldsymbol{\theta_1}$ is 2 nM, and the concentration of $\boldsymbol{\theta_2}$ is 8 nM.

We further conceptualize that Θ is a set of all existing $\boldsymbol{\theta}$'s. By this setting, we should have $\sum_{\boldsymbol{\theta} \in \Theta} C_{\boldsymbol{\theta}} = 1$. These complexes will bind in various configurations which can all be described as a $\mathbf{q}$. We define Q as a set of all possible $\mathbf{q}$'s, and we borrow the notation $\mathbf{q} \subseteq \boldsymbol{\theta}$ to indicate any binding configuration $\mathbf{q}$ that can be achieved by complex $\boldsymbol{\theta}$. This is equivalent to $|\mathbf{q_{i\bullet}}| = \theta_i$ for all i, or $\boldsymbol{\theta}$ is the row sum of (q_{ij}).

<u>*Solve the amount of free receptors*</u>

A remaining problem in the model setup is that in practice, only $R_{\text{tot},j}$, the total receptor expressions of each kind of a cell, can be experimentally measured, while the amount of free receptors at equilibrium, $R_{\text{eq},j}$, though being used extensively, is unknown. To find $R_{\text{eq},j}$, we first need to derive the amount of bound receptors of each kind, $R_{\text{bound},j}$, then use conservation of mass to solve $R_{\text{eq},j}$ numerically.

To calculate the amount of bound receptor $R_{\text{bound},n}$, we can simply add up all entries in the n-th column for every $\mathbf{q}$'s:

$$
\begin{aligned}
R_{\text{bound},n} &= \sum_{\mathbf{q}\in Q} |\,\mathbf{q}_{\bullet\mathbf{n}}| v_{\mathbf{q}} = \sum_{\mathbf{q}\in Q} |\,\mathbf{q}_{\bullet\mathbf{n}}| \frac{L_0 C_{\boldsymbol{\theta}(\mathbf{q})}}{K_x^*} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}} \prod_{i=1}^{N_L} \binom{\theta_i}{\mathbf{q}_{\mathbf{i}\bullet}} \\
&= \frac{L_0}{K_x^*} \sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \sum_{\mathbf{q}\subseteq\boldsymbol{\theta}} |\,\mathbf{q}_{\bullet\mathbf{n}}| \prod_{(i,j)=(1,0)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}} \prod_{i=1}^{N_L} \binom{\theta_i}{\mathbf{q}_{\mathbf{i}\bullet}} \\
&= \frac{L_0}{K_x^*} \sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \left[\frac{\psi_{1n}}{|\boldsymbol{\psi}_{\mathbf{1}\bullet}|}\theta_1 + \cdots + \frac{\psi_{N_L n}}{|\boldsymbol{\psi}_{\mathbf{N_L}\bullet}|}\theta_{N_L} \right] \prod_{i=1}^{N_L} |\boldsymbol{\psi}_{\mathbf{i}\bullet}|^{\theta_i} \\
&= \frac{L_0}{K_x^*} \sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \left[\sum_{i=1}^{N_L} \frac{\psi_{in}}{|\boldsymbol{\psi}_{\mathbf{i}\bullet}|}\theta_i \right] \prod_{i=1}^{N_L} |\,\boldsymbol{\psi}_{\mathbf{i}\bullet}|^{\theta_i},
\end{aligned}
$$

where $|\mathbf{q}_{\bullet\mathbf{n}}| = \sum_{m=1}^{N_L} q_{mn}$, and $|\boldsymbol{\psi}_{\mathbf{i}\bullet}| = \sum_{j=0}^{N_R} \psi_{ij}$.

By the conservation of mass, we have

$$
R_{\text{tot},n} = R_{\text{eq},n} + R_{\text{bound},n} = R_{\text{eq},n} + \frac{L_0}{K_x^*} \sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \left[\sum_{i=1}^{N_L} \frac{\psi_{in}}{|\boldsymbol{\psi}_{\mathbf{i}\bullet}|}\theta_i \right] \prod_{i=1}^{N_L} |\,\boldsymbol{\psi}_{\mathbf{i}\bullet}|^{\theta_i}.
$$

In this equation, $R_{\text{tot},n}$ are known, and any $|\boldsymbol{\psi}_{\mathbf{i}\bullet}|$ is a function of every $R_{\text{eq},j}$, $j = 1,2,\cdots,N_R$, so all $R_{\text{eq},j}$ need to be solved together. This system of equations usually does not have a closed form and must be solved numerically. When implementing, we suggest taking the logarithm of both sides of these equations so the exponents can be eliminated and the range is restricted to positive numbers.

As a side note, the total amount of bound receptors regardless of kind is

$$R_{\text{bound}} = \sum_{n=1}^{N_R} R_{\text{bound},n} = \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \sum_{n=1}^{N_R}\left[\frac{\psi_{1n}}{|\boldsymbol{\psi}_{\mathbf{1\bullet}}|}\theta_1 + \cdots + \frac{\psi_{N_L n}}{|\boldsymbol{\psi}_{\mathbf{N_L\bullet}}|}\theta_{N_L}\right]\prod_{i=1}^{N_L} |\,\boldsymbol{\psi}_{\mathbf{i\bullet}}|^{\theta_i}$$

$$= \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}}\left[\left(1-\frac{\psi_{10}}{|\boldsymbol{\psi}_{\mathbf{1\bullet}}|}\right)\theta_1 + \cdots + \left(1-\frac{\psi_{N_L 0}}{|\boldsymbol{\psi}_{\mathbf{N_L\bullet}}|}\right)\theta_{N_L}\right]\prod_{i=1}^{N_L} |\,\boldsymbol{\psi}_{\mathbf{i\bullet}}|^{\theta_i}$$

$$= \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}}\left[|\boldsymbol{\theta}| - \sum_{i=1}^{N_L}\frac{\theta_i}{|\boldsymbol{\psi}_{\mathbf{i\bullet}}|}\right]\prod_{i=1}^{N_L} |\,\boldsymbol{\psi}_{\mathbf{i\bullet}}|^{\theta_i}.$$

The amount of bound ligand complexes

Our model makes many macroscopic predictions readily accessible. For example, the amount of ligand bound at equilibrium is a useful quantity when measuring the overall quantity of tagged ligand. To compute this number, we can add up all $v_{\mathbf{q}}$ except the **q**'s that only have nonzero values on the 0-th column, $v_{0,\text{eq}}$. Consequently, the model prediction of bound ligand at equilibrium is

$$L_{\text{bound}} = \sum_{\mathbf{q}\in Q} v_{\mathbf{q}} - v_{0,\text{eq}} = \sum_{\mathbf{q}\in Q}\frac{L_0 C_{\boldsymbol{\theta}(\mathbf{q})}}{K_x^*}\prod_{(i,j)=(1,0)}^{(N_L,N_R)} {\psi_{ij}}^{q_{ij}}\prod_{i=1}^{N_L}\binom{\theta_i}{\mathbf{q}_{\mathbf{i\bullet}}} - \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}}$$

$$= \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}}\left[\sum_{\mathbf{q}\subseteq\boldsymbol{\theta}}\prod_{(i,j)=(1,0)}^{(N_L,N_R)} {\psi_{ij}}^{q_{ij}}\prod_{i=1}^{N_L}\binom{\theta_i}{\mathbf{q}_{\mathbf{i\bullet}}} - 1\right]$$

$$= \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}}\left[|\boldsymbol{\psi}_{\mathbf{1\bullet}}|^{\theta_1}|\boldsymbol{\psi}_{\mathbf{2\bullet}}|^{\theta_2}\ldots|\boldsymbol{\psi}_{\mathbf{N_L\bullet}}|^{\theta_{N_L}} - 1\right] = \frac{L_0}{K_x^*}\sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}}\left[\prod_{i=1}^{N_L} |\,\boldsymbol{\psi}_{\mathbf{i\bullet}}|^{\theta_i} - 1\right]$$

when $|\boldsymbol{\psi}_{\mathbf{i\bullet}}| = \sum_{j=0}^{N_R}\psi_{ij}$, and the predicted amount of bound complex $\boldsymbol{\theta}$ (complex of each kind) is

$$L_{\text{bound},\boldsymbol{\theta}} = \frac{L_0 C_{\boldsymbol{\theta}}}{K_x^*}\left[\prod_{i=1}^{N_L} \mid \boldsymbol{\psi}_{\mathbf{i}\bullet}|^{\theta_i} - 1\right]$$

The amount of fully bound ligands

In multivalent complexes like bispecific antibodies, drug activity may require that all subunits be bound to their respective targets[17]. The predicted amount of ligand full-valently bound can be calculated as

$$v_{\text{full,eq}} = \sum_{\boldsymbol{\theta}\in\Theta} \sum_{q_{10},\dots,q_{N_L 0}=0} \frac{L_0 C_{\boldsymbol{\theta}}}{K_x^*} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}} \binom{\theta_1}{\mathbf{q}_{\mathbf{1}\bullet}^*} \cdots \binom{\theta_{N_L}}{\mathbf{q}_{\mathbf{N_L}\bullet}^*} = \frac{L_0}{K_x^*} \sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \prod_{i=1}^{N_L} \left(\sum_{j=1}^{N_R} \psi_{ij}\right)^{\theta_i}$$

$$= \frac{L_0}{K_x^*} \sum_{\boldsymbol{\theta}\in\Theta} C_{\boldsymbol{\theta}} \prod_{i=1}^{N_L} (|\boldsymbol{\psi}_{\mathbf{i}\bullet}| - 1)^{\theta_i},$$

with $\mathbf{q}_{\mathbf{i}\bullet}^* = (q_{i1}, \dots, q_{iN_R})$, the $\mathbf{q}_{\mathbf{i}\bullet}$ vector without q_{i0}. In this equation, the multinomial coefficient $\binom{\theta_i}{\mathbf{q}_{\mathbf{i}\bullet}^*}$ describes the number of ways one can allocate θ_i receptors to any position in the i-th row of the (q_{ij}) matrix except the 0-th row which stands for unbound.

In fact, the predicted amount of any specific-valently bound ligands can be derived in such manner. For example, the amount of ligands that bind monovalently can be calculated as

$$v_{1,\text{eq}} = \sum_{\boldsymbol{\theta}\in\Theta} \frac{L_0 C_{\boldsymbol{\theta}}}{K_x^*} \sum_{i=1}^{N_L} \sum_{j=1}^{N_R} \psi_{ij}^{q_{ij}} \binom{\theta_i}{1} = \sum_{\boldsymbol{\theta}\in\Theta} \frac{L_0 C_{\boldsymbol{\theta}}}{K_x^*} \sum_{i=1}^{N_L} \mid \boldsymbol{\psi}_{\mathbf{i}\bullet}|\theta_i.$$

This can be used for estimating the amount of multimerized ligands, $L_{\text{multi}} = L_{\text{bound}} - v_{1,\text{eq}}$, and multimerized receptors, $R_{\text{multi}} = R_{\text{bound}} - v_{1,\text{eq}}$.

Complexes formed through random assortment

Another common mode of forming multivalent complexes in biology, such as in the formation of antibody-antigen complexes[10], is the stochastic assembly of monomer units to a common scaffold. Instead of a specific arrangement, we provide binding compounds of a fixed valency f and a mixture of monomer ligands, and complexes can form through random assortment. The concentration of these complexes, therefore, will follow a multinomial distribution.

To formulate this mathematically, we denote the proportion of L_i as C_i, and $\sum_{i=1}^{N_L} C_i = 1$. For example, say we have 40% L_1 and 60% L_2 in the solution to form dimers ($f = 2$), then $C_1 = 40\%$, $C_2 = 60\%$. As complex formation follows a binomial distribution, there will be 16% bivalent L_1, 36% bivalent L_2, and 48% $L_1 - L_2$ complex. In general, the probability of complexes formed as described by $\boldsymbol{\theta}$ in random assortment is

$$C_{\boldsymbol{\theta}} = \binom{f}{\boldsymbol{\theta}} C_1^{\theta_1} C_2^{\theta_2} \dots C_{N_L}^{\theta_{N_L}} = \binom{f}{\boldsymbol{\theta}} \prod_{i=1}^{N_L} C_i^{\theta_i}.$$

Since $\sum_{i=1}^{N_L} C_i = 1$, we know that

$$\sum_{\boldsymbol{\theta} \in \Theta} C_{\boldsymbol{\theta}} = \sum_{\boldsymbol{\theta} \in \Theta} \binom{f}{\boldsymbol{\theta}} \prod_{i=1}^{N_L} C_i^{\theta_i} = (C_1 + C_2 + \cdots + C_{N_L})^f = 1.$$

Plugging this relationship between $C_{\boldsymbol{\theta}}$ and C_i into $v_{\mathbf{q}}$ we previously derived for the amount of a specific binding configuration, we have

$$v_{\mathbf{q}} = \frac{L_0}{K_x^*} \binom{f}{\boldsymbol{\theta}(\mathbf{q})} \prod_{i=1}^{N_L} C_i^{\theta_i(\mathbf{q})} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}} \prod_{i=1}^{N_L} \binom{\theta_i(\mathbf{q})}{\mathbf{q}_{\mathbf{i}\bullet}} = \frac{L_0}{K_x^*} \binom{f}{\mathbf{q}} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} C_i^{q_{ij}} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} \psi_{ij}^{q_{ij}}$$

$$= \frac{L_0}{K_x^*} \binom{f}{\mathbf{q}} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} \varphi_{ij}^{q_{ij}},$$

where $\varphi_{ij} = R_{\text{eq},j} K_{a,ij} K_x^* C_i$ and $\varphi_{i0} = C_i$.

<u>*Solve the amount of free receptors*</u>

Similar to the specific arrangement case, we still need to solve $R_{\text{eq},n}$ numerically from $R_{\text{tot},n}$. We first derive the amount of bound receptors of each kind at equilibrium as

$$R_{\text{bound},n} = \sum_{\mathbf{q}\in Q} |\,\mathbf{q_{\bullet n}}| v_{\mathbf{q}} = \sum_{\mathbf{q}\in Q} |\,\mathbf{q_{\bullet n}}| \binom{f}{\mathbf{q}} \frac{L_0}{K_x^*} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} \varphi_{ij}^{q_{ij}} = \frac{L_0}{K_x^*} |\boldsymbol{\varphi}_{\bullet\mathbf{n}}| f \left[\sum_{(i,j)=(1,0)}^{(N_L,N_R)} \varphi_{ij} \right]^{f-1}$$

$$= \frac{L_0 f}{K_x^*} |\boldsymbol{\varphi}_{\bullet\mathbf{n}}| (1+\Phi)^{f-1}.$$

Then by the conservation of mass, we can numerically solve $R_{\text{eq},n}$ as

$$R_{\text{tot},n} = R_{\text{eq},n} + R_{\text{bound},n} = R_{\text{eq},n} + \frac{L_0 f}{K_x^*} |\boldsymbol{\varphi}_{\bullet\mathbf{n}}| (1+\Phi)^{f-1}.$$

Again, since $\Phi = \sum_{j=1}^{N_R} |\,\boldsymbol{\varphi}_{\bullet\mathbf{j}}|$ is a function of every $R_{\text{eq},n}$, all $R_{\text{eq},n}$ need to be solved together.

<u>*The amount of k-valently bound complexes*</u>

For randomly assorted complexes, we first derive the amount of ligands that bind k-valently. As we will show, it has a useful expression that can used to find many other quantities conveniently. First, let's break $\mathbf{q}$ into two separate vectors, $\mathbf{q} = (\mathbf{q_{\bullet 0}}, \mathbf{q_{\bullet x}})$. We define the vector formed by the 0-th column of $\mathbf{q}$ which stand for unbound as $\mathbf{q_{\bullet 0}}$, and the one formed by the other elements as $\mathbf{q_{\bullet x}}$. By the model setup, $|\mathbf{q}| = f$, $|\mathbf{q_{\bullet x}}| = k$, and $|\mathbf{q_{\bullet 0}}| = f - k$. We then have

$$v_{k,\mathrm{eq}} = \sum_{\substack{|\mathbf{q}_{\bullet\mathbf{x}}|=k \\ |\mathbf{q}_{\bullet\mathbf{0}}|=f-k}} v_{\mathbf{q}} = \sum_{\substack{|\mathbf{q}_{\bullet\mathbf{x}}|=k \\ |\mathbf{q}_{\bullet\mathbf{0}}|=f-k}} \binom{f}{\mathbf{q}_{\bullet\mathbf{x}} \quad \mathbf{q}_{\bullet\mathbf{0}}} \frac{L_0}{K_x^*} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} {\varphi_{ij}}^{q_{ij}}$$

$$= \sum_{\substack{|\mathbf{q}_{\bullet\mathbf{x}}|=k \\ |\mathbf{q}_{\bullet\mathbf{0}}|=f-k}} \binom{f}{k} \binom{k}{\mathbf{q}_{\bullet\mathbf{x}}} \binom{f-k}{\mathbf{q}_{\bullet\mathbf{0}}} \frac{L_0}{K_x^*} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} {\varphi_{ij}}^{q_{ij}} \prod_{i=1}^{i=N_L} {C_i}^{q_{i0}}$$

$$= \frac{L_0}{K_x^*} \binom{f}{k} \left[\sum_{|\mathbf{q}_{\bullet\mathbf{x}}|=k} \binom{k}{\mathbf{q}_{\bullet\mathbf{x}}} \prod_{(i,j)=(1,1)}^{(N_L,N_R)} {\varphi_{ij}}^{q_{ij}} \right] \left[\sum_{|\mathbf{q}_{\bullet\mathbf{0}}|=f-k} \binom{f-k}{\mathbf{q}_{\bullet\mathbf{0}}} \prod_{i=1}^{i=N_L} {C_i}^{q_{i0}} \right]$$

$$= \frac{L_0}{K_x^*} \binom{f}{k} \left[\sum_{(i,j)=(1,1)}^{(N_L,N_R)} {\varphi_{ij}}^{q_{ij}} \right]^k \left[\sum_{i=1}^{i=N_L} C_i \right]^{f-k} = \frac{L_0}{K_x^*} \binom{f}{k} \Phi^k .$$

The amount of total bound ligands and receptors

Many macroscopic properties can be derived from $v_{k,\mathrm{eq}}$. For example, the amount of total bound ligands is simply the sum of ligands bound monovalently to fully, and can be simplified to

$$L_{\mathrm{bound}} = \sum_{k=1}^{f} v_{k,\mathrm{eq}} = \sum_{k=0}^{f} v_{k,\mathrm{eq}} - v_{0,\mathrm{eq}} = \sum_{k=0}^{f} \frac{L_0}{K_x^*} \binom{f}{k} \Phi^k - \frac{L_0}{K_x^*} \binom{f}{0} \Phi^0 = \frac{L_0}{K_x^*} [(1+\Phi)^f - 1].$$

Similarly, the total bound receptors should be

$$R_{\mathrm{bound}} = \sum_{k=1}^{f} k\, v_{k,\mathrm{eq}} = \sum_{k=1}^{f} k \frac{L_0}{K_x^*} \binom{f}{k} \Phi^k = \frac{L_0}{K_x^*} f \Phi (1+\Phi)^{f-1}.$$

As we show here, these quantities all have elegant closed form solutions, and they are only dependent on Φ, a single value that incorporate all information about receptor amounts, monomer ligand compositions, and binding affinities. Φ was previously defined as $\Phi = \sum_{i=1}^{N_L} \sum_{j=1}^{N_R} \varphi_{ij} = \sum_{i=1}^{N_L} \sum_{j=1}^{N_R} R_{\mathrm{eq},j}\, K_{a,ij} K_x^* C_i$.

The number of cross-linked receptors

In some biological contexts such as T cell receptor-MHC[9] or antibody-Fc receptor[10] interactions, signal transduction is driven by receptor cross-linking due to multivalent binding. The amount of total cross-linked receptors can be derived from $v_{k,\text{eq}}$ as

$$R_{\text{multi}} = \sum_{k=2}^{f} k\, v_{k,\text{eq}} = R_{\text{bound}} - v_{1,\text{eq}} = \frac{L_0}{K_x^*} f\Phi(1+\Phi)^{f-1} - \frac{L_0}{K_x^*}\binom{f}{1}\Phi$$

$$= \frac{L_0}{K_x^*} f\Phi\left[(1+\Phi)^{f-1} - 1\right].$$

To find the number of crosslinked receptors of a specific kind, R_n, requires extra consideration. Similar to how $v_{k,\text{eq}}$ was found, we break $\mathbf{q}$ into three separate vectors, $\mathbf{q} = (\mathbf{q}_{\bullet\mathbf{0}}, \mathbf{q}_{\bullet\mathbf{n}}, \mathbf{q}_{\bullet\mathbf{x}})$. $\mathbf{q}_{\bullet\mathbf{0}}$ is the vector formed by the 0-th column of $\mathbf{q}$, $\mathbf{q}_{\bullet\mathbf{n}}$ is the vector formed by the n-th column of $\mathbf{q}$, and $\mathbf{q}_{\bullet\mathbf{x}}$ contains all others. If the complex is s-valently bound, then $|\mathbf{q}_{\bullet\mathbf{0}}| = f - s$. We further assume that $|\mathbf{q}_{\bullet\mathbf{n}}| = t$, then $|\mathbf{q}_{\bullet\mathbf{x}}| = s - t$. By this setup, we have

$$R_{\text{multi},n} = \sum_{s=2}^{f}\sum_{t=0}^{s} t \sum_{\substack{|\mathbf{q}_{\bullet\mathbf{x}}|=s-t \\ |\mathbf{q}_{\bullet\mathbf{n}}|=t \\ |\mathbf{q}_{\bullet\mathbf{0}}|=f-s}} v_{\mathbf{q}} = \sum_{s=2}^{f}\sum_{t=0}^{s} t \sum_{\substack{|\mathbf{q}_{\bullet\mathbf{x}}|=s-t \\ |\mathbf{q}_{\bullet\mathbf{n}}|=t \\ |\mathbf{q}_{\bullet\mathbf{0}}|=f-s}} \binom{f}{\mathbf{q}} \frac{L_0}{K_x^*} \prod_{(i,j)=(1,0)}^{(N_L,N_R)} {\varphi_{ij}}^{q_{ij}}$$

$$= \sum_{s=2}^{f}\sum_{t=0}^{s} \frac{tL_0}{K_x^*} \binom{f}{s-t \quad t \quad f-s} \left[\sum_{|\mathbf{q}_{\bullet\mathbf{x}}|=s-t} \binom{s-t}{\mathbf{q}_{\bullet\mathbf{x}}} \prod_{\substack{(i,j)=(1,0) \\ j\neq n}}^{(N_L,N_R)} {\varphi_{ij}}^{q_{ij}} \right]$$

$$\left[\sum_{|\mathbf{q}_{\bullet\mathbf{n}}|=t} \binom{t}{\mathbf{q}_{\bullet\mathbf{n}}} \prod_{i=1}^{N_L} {\varphi_{in}}^{q_{in}} \right] \left[\sum_{|\mathbf{q}_{\bullet\mathbf{0}}|=f-s} \binom{f-s}{\mathbf{q}_{\bullet\mathbf{0}}} \prod_{i=1}^{N_L} {C_i}^{q_{i0}} \right]$$

$$= \sum_{s=2}^{f}\sum_{t=0}^{s} \frac{tL_0}{K_x^*} \binom{f}{s-t \quad t \quad f-s} (\Phi - |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|)^{s-t} |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|^{t} \left(\sum_{i=1}^{N_L} C_i \right)^{f-s}$$

$$= \sum_{s=2}^{f} \frac{L_0}{K_x^*} \left[\sum_{t=0}^{s} t \binom{s}{t} \left(\frac{|\boldsymbol{\varphi}_{\bullet\mathbf{n}}|}{\Phi - |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|} \right)^{t} \right] \binom{f}{s} (\Phi - |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|)^{s}$$

$$= \sum_{s=2}^{f} \frac{L_0}{K_x^*} s \left(\frac{|\boldsymbol{\varphi}_{\bullet\mathbf{n}}|}{\Phi - |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|} \right) \left(\frac{\Phi}{\Phi - |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|} \right)^{s-1} \binom{f}{s} (\Phi - |\boldsymbol{\varphi}_{\bullet\mathbf{n}}|)^{s}$$

$$= \frac{L_0}{K_x^*} \left[\sum_{s=2}^{f} s \binom{f}{s} \Phi^{s} \right] \frac{|\boldsymbol{\varphi}_{\bullet\mathbf{n}}|}{\Phi} = \frac{L_0}{K_x^*} \frac{|\boldsymbol{\varphi}_{\bullet\mathbf{n}}|}{\Phi} [f\Phi(1+\Phi)^{f-1} - f\Phi]$$

$$= \frac{L_0 f}{K_x^*} |\boldsymbol{\varphi}_{\bullet\mathbf{n}}| [(1+\Phi)^{f-1} - 1].$$

This formula can useful when investigating the role of each receptor in a pathway that requires multimerized binding.

Of course, the macroscopic predictions provided in this section cannot exhaust many biological quantities one may wish to study, but with the ideas demonstrated here, the readers can derive their own formulae as needed.

Numerical implementation notes

In this model, most predictions can be calculated directly by closed form formulae after $R_{\mathrm{eq},n}$'s have been solved. For solving $R_{\mathrm{eq},n}$'s numerically, we have not found issues with deriving numerical solutions generally, except at extreme affinities (e.g., less than 1 pM K_d) combined with very high valency (e.g., greater than 64), where floating point errors can cause problems with the termination conditions of the root-finding operation. We have learned through experience that one need not set bounds on the root-finding, as the function is monotonic with a single root. While we use auto-differentiation of the model through the Python package Jax or Julia package ForwardDiff.jl during root-finding (both packages available on GitHub), one can use numerical differencing with identical results. Root-finding of $R_{\mathrm{eq},n}$ is by far the slowest part of the calculations.

Ligand concentration handling

Our model requires the total concentration of ligands at equilibrium, L_0. There are at least four approaches one could take to rectify some measurement of ligand concentration with the model: (1) First, one could apply an assumption of no ligand depletion. This is an extreme assumption in many cases but can be applicable in *in vitro* experiments where it is known that the ligand amount is many orders of magnitude greater than that of the receptors. (2) Alternatively, one might know the concentration of ligand in solution after some or all of any ligand depletion has occurred. (3) If one has an estimate of the absolute number of receptors and ligand units before binding, L_0 can be solved numerically by conservation of mass along with $R_{\mathrm{eq},n}$'s. We have

$$L_{\mathrm{init}}V = L_0 V + L_{\mathrm{bound}},$$

$$L_{\mathrm{init},\boldsymbol{\theta}}V = L_0 C_{\boldsymbol{\theta}} V + L_{\mathrm{bound},\boldsymbol{\theta}},$$

where V is the effective volume of the ligand solution, L_{init} is the total initial ligand concentration, and $L_{\mathrm{init},\boldsymbol{\theta}}$ is the initial concentration of complex $\boldsymbol{\theta}$. L_{bound} and $L_{\mathrm{bound},\boldsymbol{\theta}}$ may be found at the previously derived L_{bound} and $L_{\mathrm{bound},\boldsymbol{\theta}}$, depending on the occasion. (4) Finally, certain aspects of the system may not be sensitive to ligand concentration as an input parameter, or one could treat concentration as an unknown.

2.5 Application examples

In previous sections, we have shown how all macroscopic predictions made in this chapter can be written in closed form formulae. Therefore, some computationally expensive analyses are enabled by the efficiency of our model. Here, we provide two examples to demonstrate the utility of large-scale predictions made possible by this model.

Mixture binding prediction

Leveraging a synergistic effect among two or more drugs is of great interest in pharmaceutical development. A challenge in investigating synergy is to identify its underlying source. Most biological pathways follow a similar pattern: when the drug binds to certain surface receptors of a cell, a downstream pathway in the cell is initiated, leading to some actions. Therefore in general, synergism can come from either the initial binding events themselves or downstream signal transduction interactions. Binding-level synergy means that merely using a combination of ligands boosts the amount of binding to the important receptors and thus intensifies the overall effect. Downstream effect synergy indicates that the benefit of using mixtures arises from other cellular regulatory mechanisms two ligands can bring about. The binding model we propose here can help to investigate this issue by offering accurate predictions for the binding of multivalent complex mixtures.

In Fig. 2.3, we provide an example of mixture binding predictions (Fig. 2.3a). We investigate a mixture of three types of ligand complexes, bivalent L_1 ($\mathbf{\theta_1} = [2{,}0]$), bispecific $L_1 - L_2$ ($\mathbf{\theta_2} = [1{,}1]$), and monovalent L_1 ($\mathbf{\theta_3} = [1{,}0]$). The crosslinking constant is set to be $K_x^* = 10^{-12}$ cell · M, similar to previous results[10]. We predict the amount of binding of this mixture to a cell expressing three types of receptors, with $\mathbf{R_{tot}} = [2.5 \times 10^4, 3 \times 10^4, 2 \times 10^3]$ cell^{-1}. The affinity constants of L_1 to these three receptors are $\mathbf{K_{a,1\bullet}} = [1 \times 10^8, 1 \times 10^5, 6 \times 10^5]$ M^{-1}, and of L_2, $\mathbf{K_{a,2\bullet}} = [3 \times 10^5, 1 \times 10^7, 1 \times 10^6]$ M^{-1}. Here, we investigate the changing concentration of $\mathbf{\theta_1}$ and $\mathbf{\theta_2}$, while holding the amount of $\mathbf{\theta_3}$ constant at 0.2 nM. Fig. 2.3 shows the predicted total ligand bound (Fig. 2.3b) and R_3 bound (Fig. 2.3c) for only $\boldsymbol{\theta_1}$ or $\boldsymbol{\theta_2}$ with concentration from 0 to 0.8 nM, and their mixtures in every possible composition with total concentration 0.8 nM.

Mixture binding prediction can help us identify the source of synergy. To connect model predictions to experimental measurements, ligand binding might be measured by fluorescently-tagged ligands, while the number of bound receptors of a specific type might associate with an indirect measurement such as cellular response. After making a series of measurements for different compositions of mixtures, we can fit the 100% of one complex cases (numbers on the two ends on the plot) first and then compare the mixture measurements to the predictions. Determining whether the downstream effect contributes to the observed synergy (or antagonism) can be framed as a hypothesis testing problem:

$\mathbf{H_0}$: *The synergism of the mixture can be explained solely by binding.*

The uncertainty of mixture binding prediction comes from measurement errors of receptor abundance and binding affinities. Usually, the receptor expression of a cell population has an empirical distribution which can be measured. The confidence intervals in Fig. 2.3 were drawn with the assumption that receptor expression fluctuates up and down by 10% (coinciding with log-normally distributed amounts). Also, due to the measurement technique, the binding affinities may be over- or underestimated[18]. The confidence interval of mixture prediction can be determined by the model with all these considered, and a p-value can be derived.

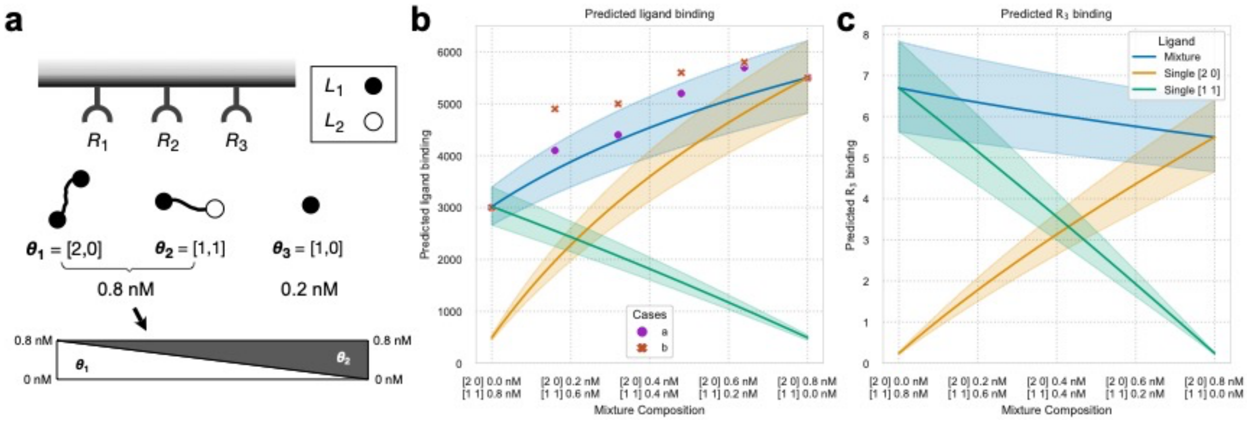


Figure 2.3. Prediction on mixture binding of $\boldsymbol{\theta_1} = [2, 0]$, $\boldsymbol{\theta_2} = [1, 1]$, and $\boldsymbol{\theta_3} = [1, 0]$.

(a) A schematic of the binding scenario;

(b) The predicted total ligand binding;

(c) The amount of bound R_3 at equilibrium.

Shaded areas in (b), (c) are simulated confidence intervals by varying the receptor levels up and down by 10%. The points on (b) represent possible experimental results, arbitrarily drawn for demonstration. In case a (purple circles), since most data points are inside the confidence interval, we can assume the measurement error can explain these variations. In case b (orange crosses), however, the synergism of these complexes is beyond the binding level.

We arbitrarily drew some possible experimental results on Fig. 2.3b for demonstration. If most mixture measurements fall within the confidence interval of the predictions (such as case *a* annotated by the purple circles in Fig. 2.3b), the synergy will very likely come from binding only. However, if the measurements are obviously beyond the confidence interval (case *b*, the orange crosses), it is reasonable to suspect a synergistic (or antagonistic) effect beyond binding alone. With the flexibility of the binding model, this method can also be extended to a mixture of more than two compounds.

Binding space of a ligand

When a dose of ligand (drug, hormone, cytokine, etc.) is released into the circulation system of an individual due to either physiological response or exogenous administration, the compound will spread and bind to many cell populations to varying extents. An essential question in pharmacology is how much a compound will bind to their intended target populations compared to off-target ones. This question is important for understanding basic biology as well as developing new therapeutics. For example, hormones and cytokines are important signaling molecules, and having a quantitative prediction of on- and off-target binding can help us understand their mechanism greatly. For drug development, binding prediction can guide optimization to improve specificity toward the intended targets. A cell population can be defined by the receptors they express. Therefore, given the parameters of the dose and the receptor profile of each cell population, our model can make all the predictions discussed previously.

From the perspective of this binding model, there is nothing special about any specific cell population. If the local concentration is constant everywhere, our model can map any cell with a certain receptor expression to the amount of binding induced by this dose. If the biological activity of this compound on a cell is related to the quantity of binding to a certain ligand or receptor, the effect of this dose can be written as a function f, with

$$\mathbf{R_{tot}} \in \mathbb{R}_+^{N_R} \mapsto f(\mathbf{R_{tot}}) \in \mathbb{R}_+,$$

where $\mathbf{R_{tot}}$ is a vector of nonnegative entries that describes the cell's expression of N_R receptors, and $f(\mathbf{R_{tot}})$ is the amount of binding. Here, we define the binding behavior of this dose (or any compound) as its binding space.

In Fig. 2.4, we plot the binding space of a bivalent L_1 ligand $\boldsymbol{\theta} = [2,0]$ with concentration 1 nM. The binding affinities are the same as described in the last subsection. In this binding space, we consider three receptors, R_1, R_2, and R_3. We plot how the amount of binding relates to the cell expression profile, $\mathbf{R_{tot}}$. Here, the amount of R_1 and R_2 varies with the two axes, while R_3 is held constant at 2.0×10^3 cell^{-1}. In this plot, each point represents a cell with a distinct expression profile, as some examples drawn on Fig. 2.4a. Then we use colors and contour lines to show the amount of binding. From these two plots, we can see that although both ligand binding and R_2 binding increase with more receptors, ligand binding is more sensitive to R_1 amounts, and R_2 binding R_2 amounts. To consider any specific cell population, one only needs to determine where its expression profile falls on the plot and read the predictions from the contour line. For example, on Fig. 2.4b, the red cell population will have about $e^{5.2} = 181$ bound ligands per cell. The number of contour lines a population ride on can also show intrapopulation variation. In this case, we expect the variation in ligand binding to fall between $e^{4.3} = 74$ and $e^{6.0} = 403$.

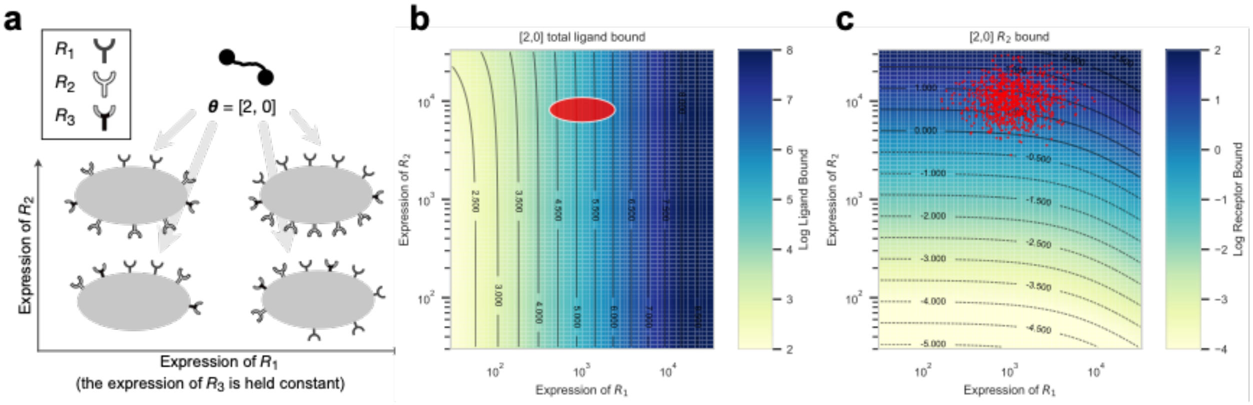


Figure 2.4. The binding space of 1 nM $\theta = [2, 0]$.

(a) A schematic diagram of the ligand and four examples of receptor-expressing cells represented by the coordinate;

(b) The amount of total ligand bound;

(c) Receptor R_2 bound predictions.

The x- and y-axis show the expression of R_1 and R_2, while the expression of R_3 is a constant, 2.0×10^3 cell^{-1}, and not shown. Any cell population can be drawn on the binding space. For example, the red ellipse on (b) represents a cell population with receptor expression of roughly $\mathbf{R_{tot}} = [1.0, 10.0, 2.0] \times 10^3$ cell^{-1}. We can alternatively project points of experimental single cell expression data onto a binding space, as shown on (c) (the points were generated arbitrarily assuming a population of log-normally distributed R_1 and R_2 expression for demonstration purpose.

The binding space can provide ample information about the compound. It is an intrinsic property of a ligand given its concentration and other ligand it mixes with, independent of any specific cell. The biological process of drug diffusion to a certain cell is analogous to sampling a point from this binding space. Its gradient indicates in which direction the binding level increases the fastest, as well as to which receptor amount it is the most sensitive. An inactive antagonist that introduces binding competition with the ligand can distort its binding space, and we can visualize it by the change of shape in the contour lines. This plot can also intuitively demonstrate intrapopulation binding variance and interpopulation cell specificity of the compound. With the development of high-throughput single-cell methods such as flow cytometry, the expression profiles of a collection of cells can be identified en masse, and we can overlap their results onto a binding space plot (as in Fig. 2.4c). This shows the promise of applying our model to single-cell data. Although we can only visualize two receptors in a plot, binding space applies to any N_R types of receptors. Theoretically, the concept of the binding space of a ligand is only complete when all relevant surface receptors are considered.

2.6 Discussion

In this chapter, I propose a mechanistic multivalent binding model that accounts for the interaction among multiple receptors and a mixture of ligand complexes formed by binding monomers. The flexible framework allows a mixture of both homogeneous and heterogeneous ligand complexes, even when they don't have the same valency. I first derive the amount of ligand of a specific binding configuration at equilibrium through the law of mass action. Using this formula, I make macroscopic predictions by applying the multinomial theorem strategically. Our predictions cover cases where complexes are formed by specific arrangement or random assortment. Finally, I provide two practical examples of how this model can help with biological research.

Compared with previous approaches, the model here is a uniquely scalable and elegant approach to multivalent binding when considering multivalent complexes of heterogeneous monomer composition and/or multiple receptors. Scalability to higher valency complexes is essential as rule-based computational models can become impractical due to a combinatorial explosion of binding states. By contrast, our model can make a large number of predictions easily, enabling mixture synergy analysis and binding space calculations across individual cells. The mathematical elegance of the model welcomes analytical studies and incorporating it into higher-level computational frameworks. For example, we apply auto-differentiation to ensure accuracy in the root-finding operation when solving for unbound receptor. We have similarly used auto-differentiation to solve for the gradients of the model with respect to input quantities when fitting it to data points. One could even feasibly derive analytical forms of the gradients. This enables one to build more complex computational models on top of this binding framework, such as inferring the composition of multivalent complexes in solution from indirect high-throughput assays. While differentiation of differential equation models is possible through adjoint state methods, solving can be sensitive to the parameters of the system, is much less efficient, and requires trade-offs in accuracy for performance.

The assumptions made in this model may compromise its accuracy in some cases. Our setup has a single crosslinking constant, K_x^*, to reflect the multivalency effect. In practice, this model has worked well in predicting experimental binding results[10,19,20]. However, the steric effects of a multivalent ligand can be more complicated and context-dependent. The complication of multivalency effect comes from the geometry of ligand complexes that introduced steric effect as well as the distribution of receptors on the cell surface. For instance, the length of the hinge region is needed to estimate the radius of area a molecule can reach[21]. Receptor clustering can be play a big role in the behavior of ligand binding as well[22]. Accounting for these effects requires more in-depth studies than just measuring the monomer binding kinetics. Some other computational approaches investigate steric effects more meticulously, but inevitably introduce some added complexity. For example, previous work has conducted a case-by-case exploration of how ligands bind when distributed randomly or ordered, arranged as a lattice, ring, or chain to give a better hindrance factor estimation[16]. When the actual situation is not known, our model can serve as an adequate starting point.

Although this model is very general purpose, it mainly focuses on the binding dynamics on a cell surface, similar to the previous work on which it is based[11,13,14]. For ligands discordant with the two-step binding process shown in Fig. 2.2, other model constructions might be necessary. For example, some previous work focuses on scaffold proteins as intracellular multivalent complexes[23], but these often lack independence between the individual monomer binding events. In this case, various alternative computational models have been developed[24–26].

Surface receptor binding is a universal event in biology. A prevalent question calls for a general solution. I expect this model to be successfully applied to many contexts. Previously, we have used a simpler version of the random assortment model to accurately predict IgG antibody-FcγR interactions[10], and also applied it to fit epithelial cell adhesion molecule binding data[19,20]. We are also working on applying the model to IL-2 immunocomplexes[27], for optimization of high-valency cytokines with specific cell targeting, design of cytokine-antibody bispecific antibody fusions, and as a factorization kernel in dimensionality reduction of systems serology data[5,28]. With the arise of multispecific drugs in the recent decade[29], I expect this model to apply even more widely, exhibit its full competence and facilitate both basic scientific research and new therapy development.

Data and software availability

A Python package of this model and the code for the plots can be found at https://github.com/meyer-lab/valentBind/. I also provide a Julia package of the model at https://github.com/meyer-lab/polyBindingModel.jl/.

Chapter 3

Mixed IgG Fc immune complexes exhibit blended binding profiles and refine FcR affinity estimates

Don't let the perfect be the enemy of the good.

3.1 Introduction

Antibodies are both a core component of adaptive immunity and a versatile platform for developing therapies. An antibody's role in promoting immunity is defined by its selectivity toward a target antigen, as determined by its variable region, and its ability to elicit effector cell responses, defined by the composition of its constant, fragment crystallizable (Fc) region. Antibodies of the IgG type direct effector responses by binding to Fcγ receptors (FcγRs) via their Fc region. FcγR activation is initiated through IgG-mediated clustering, which in turn is caused by the engagement of several antibodies on an antigen target, forming an immune complex (IC). Depending upon the receptors included, this interaction may promote or prevent an effector response. This clustering mechanism ensures that more than one IgG is present whenever effector responses occur.

The immune response triggered by an IgG IC consisting of a specific Fc form, including subclass or glycosylation, is defined by its binding to specific FcγRs, each of which differs in signaling effect and expression patterns[2]. Consequently, accurate estimates of IgG Fc-FcγR affinities are essential to understanding their effect. Most existing FcγR affinity measurements have been performed by surface plasmon resonance (SPR) using monovalent IgG[30,31]. SPR accurately assesses protein-protein binding kinetics, but many antibody-Fc receptor interactions are weak enough to fall outside the assay's quantitative range when assessed in monovalent form. Clustering leads to avidity effects, wherein even weak interactions can cooperatively lead to strong binding[19]. Indeed, avidity is widely employed in natural and engineered systems to promote binding through low-affinity interactions[32]. Therefore, direct measurement of IC binding might more accurately quantify IgG Fc properties, particularly for low-affinity interactions. Measuring Fc binding as multivalent ICs additionally resembles the relevant *in vivo* context of effector responses[33].

Physiological antibody responses universally involve Fc mixtures. For instance, during the course of infection, the composition of IgG subclasses shifts dynamically to different subclasses due to class switching[34]. Even when recombinantly manufacturing monoclonal therapeutic antibody preparations, heterogeneity exists in the glycosylation forms derived, and this glycan heterogeneity likely exists during endogenous antibody production as well[35,36]. With mixtures of antibodies of varied Fc composition but identical antigen binding, there might be an additive combination of effects, or a minor species (e.g., glycosylation variant) might present an outsized effect promoting or preventing effector responses. Therefore, knowledge of how these different forms influence the behavior of one another would allow one to modulate immune responses by adjusting subclass composition. With respect to therapeutic monoclonal antibody preparations, this would help guide the evaluation of biosimilars by determining whether glycosylation forms present at small fractions influence overall therapeutic efficacy[37].

After binding to Fc receptors, effector cell-elicited responses to IgG include several different functionally distinct mechanisms, including antibody-dependent cell cytotoxicity (ADCC) and phagocytosis (ADCP). Effector responses are coordinately regulated by the cell types present within a tissue[38,39], the FcγRs expressed on those effector cells[40], the Fc regions present within an immune complex[2], and properties of antigen engagement[41,42]. Regulation at the Fc receptor and cell population level is a challenge to engineering antibodies with desirable cell-killing functions, as well as understanding both productive and pathogenic immune responses. Furthermore, it has become clear that, in addition to NK cells, tissue-resident macrophages and bone marrow-derived monocytes participate in cytotoxic antibody-dependent target cell clearance. In contrast to NK cells (expressing only one activating FcγR, FcγRIIIA), these myeloid cell subsets express a broader set of activating FcγRs and the inhibitory FcγRIIB[40]. Thus, mixed IC may trigger all or specific subsets of activating/inhibitory FcγRs, resulting in further complexity. Despite the presence of and capacity to bind to multiple activating FcγRs on myeloid effector cells, our previous studies have demonstrated that individual IgG subclasses, such as mIgG2a/c, may mediate their activity through select activating FcγRs, indicating that there may be specialization in FcγR signaling[33].

Our team recently demonstrated that a model of IC-FcγR binding accurately captured and could predict *in vitro* binding across various IgG isotypes[10]. Further, it could accurately predict antibody-elicited tumor cell killing in mice across antibodies of varied isotype, glycosylation status, and FcγR knockouts[10]. Directly quantifying and predicting cell clearance makes it possible to accurately anticipate and optimize for antibody-mediated therapeutic effects. However, it is still unclear whether such a modeling strategy can accurately predict the response of human immune cells, particularly given the divergent properties between the murine and human receptors[43–45], and whether this modeling strategy can extend to ICs of mixed composition.

Here, we examined the binding properties of ICs with mixed IgG Fc composition. We quantified the binding of these ICs to each individual FcγR and observed that mixed-composition ICs resulted in a continuum of binding responses. A multivalent binding model extended to hetero-valent immune complex mixtures captured binding overall. However, surprisingly, it did not match certain low-affinity interactions[46]. Investigating the source of this discrepancy allowed us to improve the estimates of these interactions' affinities. We additionally demonstrate that the binding model can be used to both predict *in vivo* effector responses in humanized mice and infer the cell types responsible for these responses. Thus, while antibody effector responses operate through a complex milieu of antibody species, Fc receptors, and cell types, IC profiling paired with modeling provides a framework to reason about the role of each molecular and cellular element.

3.2 Profiling the binding effects of mixed-composition immune complexes

To determine the effect of having multiple Fc forms present within an immune complex (IC), we developed a controlled and simplified *in vitro* system. Like in previous work, we employed a panel of CHO cell lines expressing one of six individual human FcγRs[10] (Fig. 3.1a). ICs were formed by immobilizing anti-2,4,6-trinitrophenol (TNP) human IgG on conjugates of TNP and bovine serum albumin (TNP-BSA) with an average valency of 4 or 33. IgG binding was then quantified after incubation with the cells, using a constant IC concentration of 1 nM (Fig. 3.S1). In contrast to our previous work using a single IgG isotype, we assembled ICs from mixtures of each IgG isotype pair[10]. For each pair of IgGs, ICs were formed with a spectrum of six compositions of the IgG pair, including 100%/0%, 90%/10%, and 67%/33% mixtures. Combinations of 6 FcγRs, 2 valencies, 6 IgG pairs, and 6 IgG compositions resulted in 432 distinct experimental conditions. One-way ANOVA showed that more than 70% variance in the data was between experimental conditions rather than within them, indicating that more than 70% of the variance could be explained by biological differences (Tbl. 3.S1). This suggests that, within each condition, measurements were consistent.

Inspection of the resulting binding data revealed several expected patterns. Among the conditions with only one IgG present, the measured binding showed a strong, positive correlation with the documented IgG-FcγR interaction affinities (Fig. 3.1b). The higher valency ICs universally showed greater binding signal compared to their matching lower-valency counterparts, and there is an obvious negative trend between documented affinities and the ratio between the 33-valent and 4-valent complex binding (Fig. 3.1c). This trend is expected since, although complexes of both valencies can bind densely with high-affinity units, only high-valent complexes compensate for low affinity through avidity[19]. Therefore, while high-affinity complexes result in greater binding, low-affinity complexes have greater intervalency binding ratios. Finally, mixtures spanning 100% of one IgG isotype to another generally showed a monotonic shift with composition (Fig. 3.S1). These patterns, along with their reproducibility (Tbl. 3.S1), gave us confidence in the quality of the binding measurements.

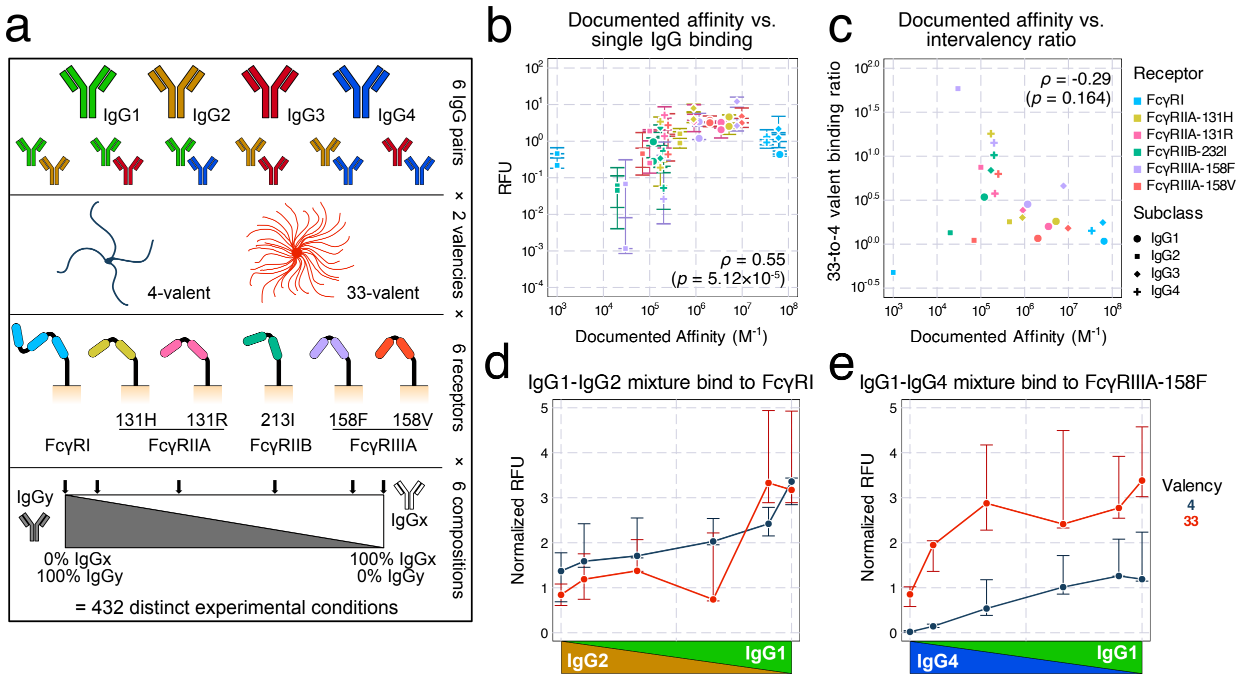


Figure 3.1. **Profiling the binding effects of mixed-composition immune complexes**.

(a) Schematic of the immune complex (IC) binding experiment. Individual or mixtures of IgG subclasses are immobilized as multivalent TNP complexes. The binding of these complexes to CHO cells expressing a single type of Fc receptor is then quantified.

(b) Measured binding in relative fluorescence units (RFU) versus the previously reported affinity of each interaction. Only single subclass conditions are plotted. Each condition has 3–5 technical replicates. Error bars represent the interquartile range of the measurements.

(c) The ratio of median binding quantified between valency 33 and 4, versus the reported affinity of the interaction. ρ represents the Spearman correlation coefficient. Significance testing was performed using the t-statistic under the null hypothesis that $\rho = 0$.

(d) IgG1-IgG2 mixture binding to FcγRI shows appreciable binding, even though IgG2-FcγRI is documented to be non-binding. The RFU level was normalized to match the FcγRI expression to the FcγRIIIA-158F expression (shown in e) on CHO cells.

(e) IgG1-IgG4 mixture binding to FcγRIIIA-158F.

In (d) and (e), each error bar represents the interquartile range of the three technical replicates in the respective condition.

We also observed several unexpected trends among the binding measurements. There was appreciable binding from IgG2-FcγRI interactions, despite this combination being reported as non-binding[31] (Fig. 3.1d). We also saw an increase in binding along the shift from IgG4 to IgG1 with FcγRIIIA-158F, even though these two isotypes are documented to have identical affinities[31] (Fig. 3.1e). These two observations are consistent with previous binding measurements using the same TNP-based IC system[10].

To better visualize the binding of these experimental conditions, we performed principal component analysis (PCA) on the median measurement of each condition, with each isotype mixture and valency as a sample and each receptor as a feature. The first principal component (PC1) explains more than 86% of the variance, and the first two components (PC1 and PC2) explain 93% (Fig. 3.2a). Inspecting the scores, we found that the 33-valent measurements are more broadly distributed, consistent with their greater expected binding (Fig. 3.2b-g). PC1 mostly separates IgG3 binding from other isotypes, reflecting that IgG3 has the greatest binding among IgG subclasses (Fig. 3.S1). PC2 separated the genotype variants of FcγRIIA and FcγRIIIA and associated most strongly with IgG3 and IgG4 (Fig. 3.2h), reflecting that these two subclasses showed larger differences in binding with genotype (Fig. 3.S1).

In all, these data support that TNP-assembled ICs provide a controlled *in vitro* system in which we can profile the effects of mixed IC composition on binding to effector cell populations. Quantifying binding using ICs may, in fact, provide more precise quantification of IgG-FcγR interaction affinities, particularly for lower affinity pairs, and mixed Fc composition ICs showed binding between that of the corresponding single Fc cases.

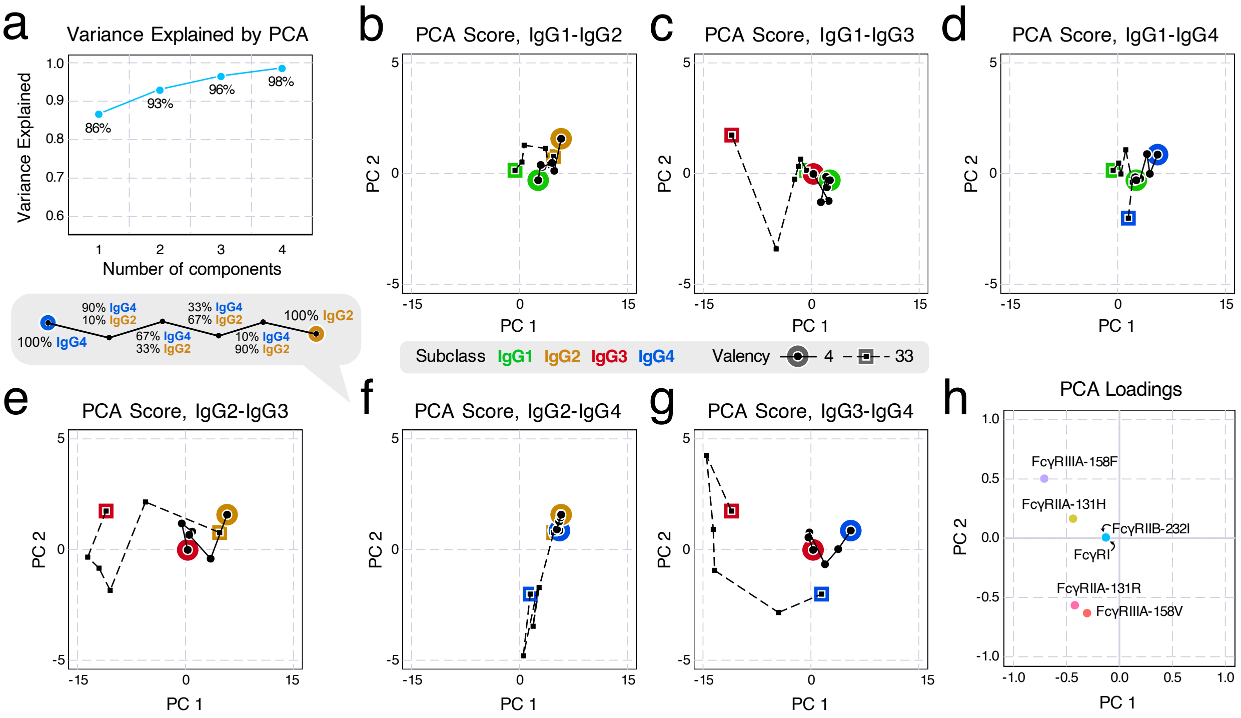


Figure 3.2. Principal component analysis (PCA) visualizes the variance in mixture binding measurements and their associated factors.

(a) Variance explained by each number of components in PCA. Two principal components (PC) explained greater than 93% of the measurement variance.

(b–g) PCA scores for immune complexes of each valency and pair of IgG subclasses.

(h) PCA loadings. The FcγRI and FcγRIIB-232I points overlap.

3.3 A multivalent binding model accurately predicts *in vitro* IgG mixture binding and updates Fc-FcγR affinities

To model the effects of polyclonal antibody responses, we extended a simple, equilibrium binding model that we have previously used to model antibody effector response[10,46]. Briefly, immune complexes are assumed to bind to FcγRs on the cell surface with monovalent binding kinetics, and then can engage additional receptors with a propensity proportional to their affinity (Fig. 3.3a). Though additional assumptions are not required for modeling ICs of mixed isotype composition, this extension leads to a large combinatorial expansion in the number of binding configurations. Through some properties of combinatorics, we derived simplified expressions for many macroscopic quantities to allow this model to scale to multi-ligand, multi-receptor, and multivalent situations[46].

We first used the measured receptor expression (Tbl. 3.S2) and documented affinities[31] with the model and obtained reasonable agreement overall (Fig. 3.3c). While the predicted values mostly agreed with the measurements, there were several notable outliers, most prominently an underestimate of IgG2-FcγRI binding (Fig. 3.3c, red circle). To improve the measurement fit, we reversed the estimation process and used the measured binding to infer the interaction affinities via Markov chain Monte Carlo (MCMC) (Fig. 3.3b). We first created a baseline fit quality by fitting all but the affinities (e.g., receptor abundance and the crosslinking parameter K_x^*, Fig. 3.3d). Although the fit improved, outliers persisted (circled in red in Fig. 3.3d). Therefore, we next performed the fitting while allowing the Fc-FcR affinities to vary. Although we only used the single-IgG measurements to infer the Fc affinities (Fig. 3.3e), we obtained much more accurate predictions for all measurements of both single and mixed IgG compositions (Fig. 3.3f).

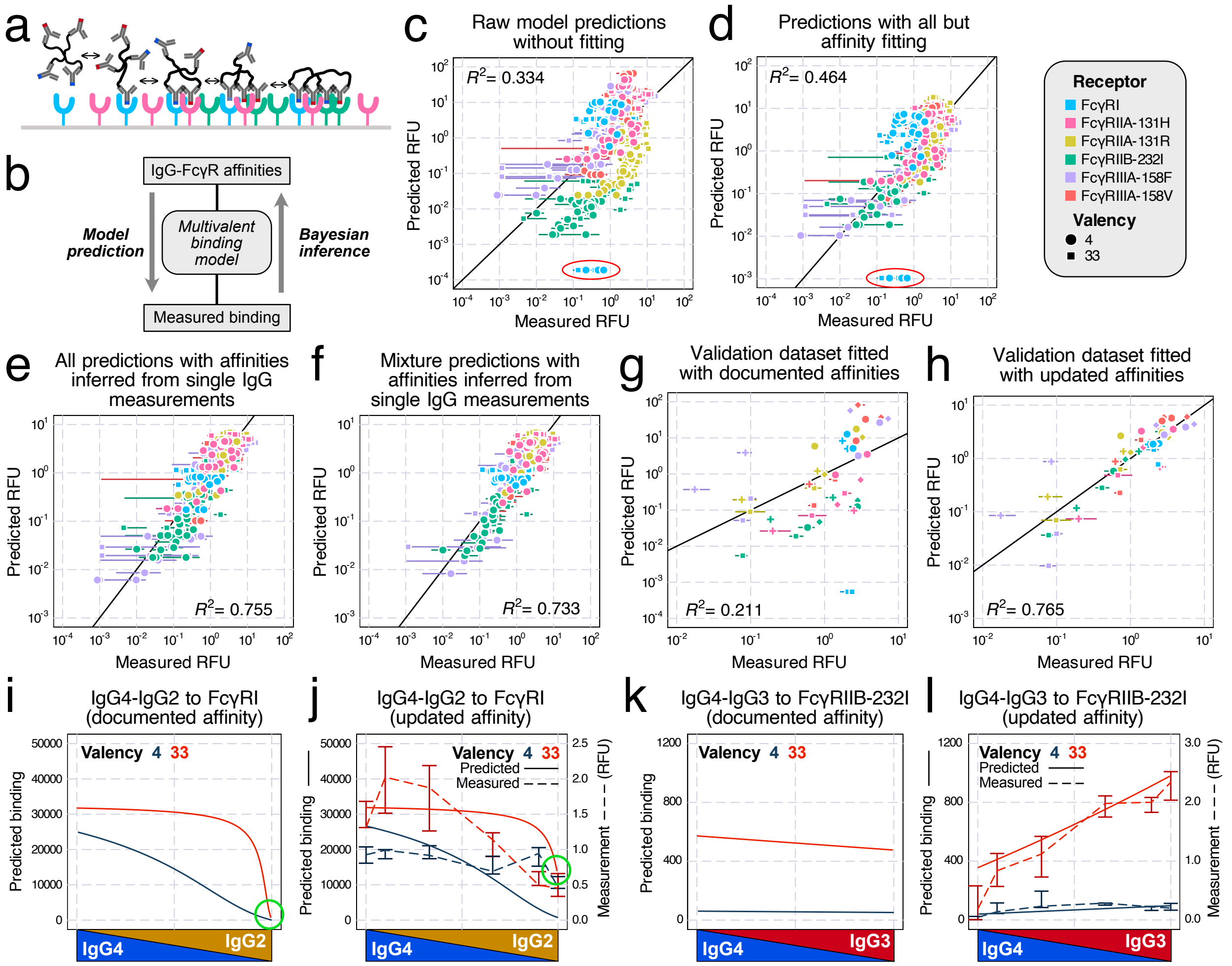


Figure 3.3. A multivalent binding model accurately accounts for *in vitro* binding of IgG mixtures.

(a) Schematic for the multivalent binding model.

(b) Schematic of the process of predicting binding with documented affinities and inferring affinities from measurements.

(c) Measured versus predicted binding by the binding model without fitting. Points also vary in the IgG subclass used, which is not indicated.

(d) Binding model prediction with all parameters but affinities fitted by Markov chain Monte Carlo (MCMC).

In (c) and (d), the IgG2-FcγRI outliers were circled in red. Since this interaction was previously reported as nonbinding, the actual predictions were all 0, but were clipped to a nonzero value (1/10 of the next smallest value) to be plotted on the log scale.

(e) Binding model prediction of all measurements (single and mixed IgG) with affinity inferred from the single IgG measurements.

(f) Binding model prediction of mixture IgG measurements with affinities updated using the single IgG measurements.

In (c–f), error bars represent the interquartile range of 3–5 technical replicates.

(g, h) Validation of the updated affinities with a separate dataset[10] by predicting the binding with either documented (g) or updated (h) affinities The error bars represent the interquartile range of four technical replicates for each condition.

(i–l) Predicted binding of IgG4-IgG2 mixture to FcγRI (i, j) and IgG4-IgG3 mixture to FcγRIIB-232I (k,l), with either documented (i, k) or updated affinities (j, l, solid line and left axis) compared with measured binding (j, l, dashed line and right axis). Error bars in (j) and (l) represent the interquartile range of 3–5 technical replicates.

To further confirm the generality of these updated affinities, we validated these new affinity estimates with an independent dataset collected in a previous study[10]. This previous study independently measured the binding of TNP-BSA complexes *in vitro* with two distinct average valencies (4 and 26), but only the binding of single IgG isotypes. We set the Fc affinities to either documented or updated values and let MCMC fit the other parameters. The new affinities resulted in a vastly improved agreement with the data (Fig. 3.3g/h).

To illustrate the impact of the affinity changes, we compared the binding predictions with two sets of affinities (Fig. 3.S2/S3) to their corresponding measurements (Fig. 3.S1). For FcγRI binding to IgG2-IgG4 mixtures, the experiment indicated that there was still notable binding with mostly or 100% IgG2, while IgG2-FcγRI was documented as non-binding[31]. The updated values amended the prediction and reflected this interaction, especially for the 33-valent complex (Fig. 3.3i/j, green circle). For FcγRIIB-232I binding to IgG3-IgG4, the documented affinities indicated there should be more binding to IgG4 compared to IgG3, contrary to our observation (Fig. 3.3k). The updated affinities instead accurately predicted the binding of all mixtures at both valencies (Fig. 3.3l). These examples demonstrate that the affinity adjustments greatly improved agreement with the binding measurements.

As our Fc affinity inference was constructed in a Bayesian fashion, both the prior (documented) and the posterior (updated) affinity values are represented as distributions accounting for uncertainty. Inspecting these updated distributions (Fig. 3.4a–d; Tbl. 3.S3), we noted several trends. The model made the largest adjustments to the Fc affinities of IgG2 (Fig. 3.4b), followed by IgG4 (Fig. 3.4d). Most IgG1 (Fig. 3.4a) and IgG3 (Fig. 3.4c) affinities remained unmodified, except for a slight increase in their FcγRIIB-232I affinities. The most notable update occurred to IgG2-FcγRI. Previously reported as nonbinding, FcγRI was revised to be the highest affinity receptor for IgG2, consistent with the receptor's high affinity to other human IgG subclasses. This discrepancy was reflected in the model prediction before affinity fitting, where the IgG2-FcγRI binding was the striking outlier (Fig. 3.3d/g). Another significant adjustment occurred with IgG3-FcγRIIB-232I. Although FcγRIIB-232I has a low affinity for all IgG subclasses, our update led to IgG3 being the strongest-binding subclass (Fig. 3.4c, S2 & S3). More subtle differences can be observed from specific model predictions (Fig. 3.S2 & S3). The revised affinities showed a similar overall correlation with binding overall (Fig. 3.4e). The intervalency binding ratios show a more prominent negative correlation, however, due to the movement of the IgG2-FcγRI outlier (Fig. 3.4f).

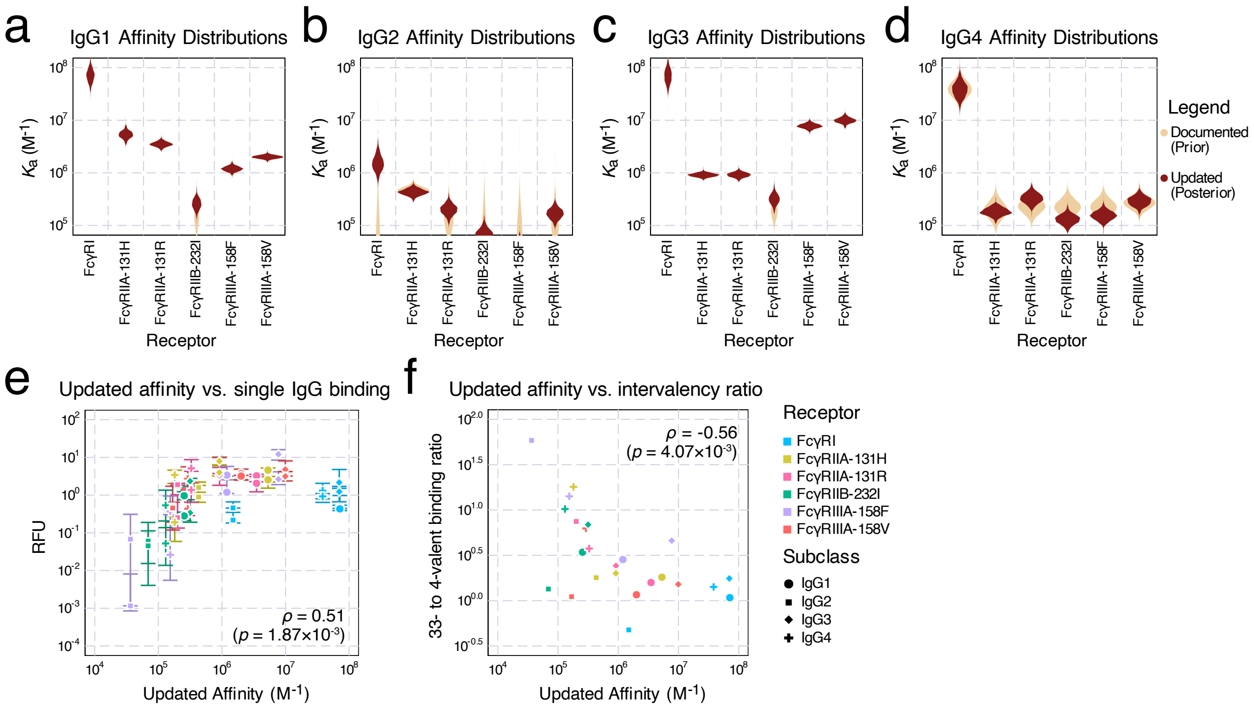


Figure 3.4. Inferred affinities from the binding data.

(a–d) The prior (documented) distributions of binding affinities (assuming all follow log-normal distributions) and posterior (updated) affinities of IgG1 (a), IgG2 (b), IgG3 (c), and IgG4 (d).

(e) Updated affinities plot against the binding measurements of single IgGs. Error bars represent the interquartile range of the 3–5 technical replicates.

(f) Updated affinities plot against the ratio of median binding between valency 33 and 4 complexes.

In (e) and (f), ρ represents the Spearman correlation coefficient. Significance testing was performed using the t-statistic under the hypothesis that $\rho = 0$.

3.4 Multivalent binding predicts antibody-elicited effector responses in humanized mice

We next sought to link the binding of ICs to their effects on the clearance of antigen targets *in vivo*. To quantify the antibody-driven activity of each effector cell, we first measured the binding of each human IgG subclass to immune effector cells collected from the peripheral blood of human donors *in vitro* in IC of two valencies, 4 and 33 (Fig. 3.5a–d). The measurements show that the binding amounts of IgG1 and IgG3 were generally about 10-fold higher in magnitude than those of IgG2 and IgG4. For the latter two subclasses, their 4-valent complex binding was almost negligible. In all cases except IgG2, neutrophils had more binding than classical and nonclassical monocytes.

We predicted the same quantities of IC binding by the multivalent binding model with either the previously documented[31] or updated affinities (Tbl. 3.S4), and the quantification of FcγR abundance[40] (Tbl. 3.S6, Fig. 3.5e–h). These estimated binding amounts broadly aligned with the measurements (Fig. 3.5i/j). Between the two sets of affinities, the predictions for IgG1 and IgG3 remained almost identical (Fig. 3.5e/g), while more differences were reflected in IgG2 and IgG4 (Fig. 3.5f/h), consistent with the affinities changing more for IgG2 and IgG4 (Fig. 3.4a–d). The predictions with documented and updated affinities were generally comparable in their concordance with the measurements (Fig. 3.5i/j). However, the predicted binding to nonclassical and classical monocytes was adjusted to be much higher for 33-valent IgG2 (Fig. 3.5f), better matching the measured values (Fig. 3.5b/i/j). Both sets of affinities predicted the binding of IgG4 to classical monocytes to be much higher than the measurements (Fig. 3.5i/j). These changes indicate that the updated affinities better predict IgG IC binding to effector cells, suggesting that they may also help improve the estimation of *in vivo* cell response.

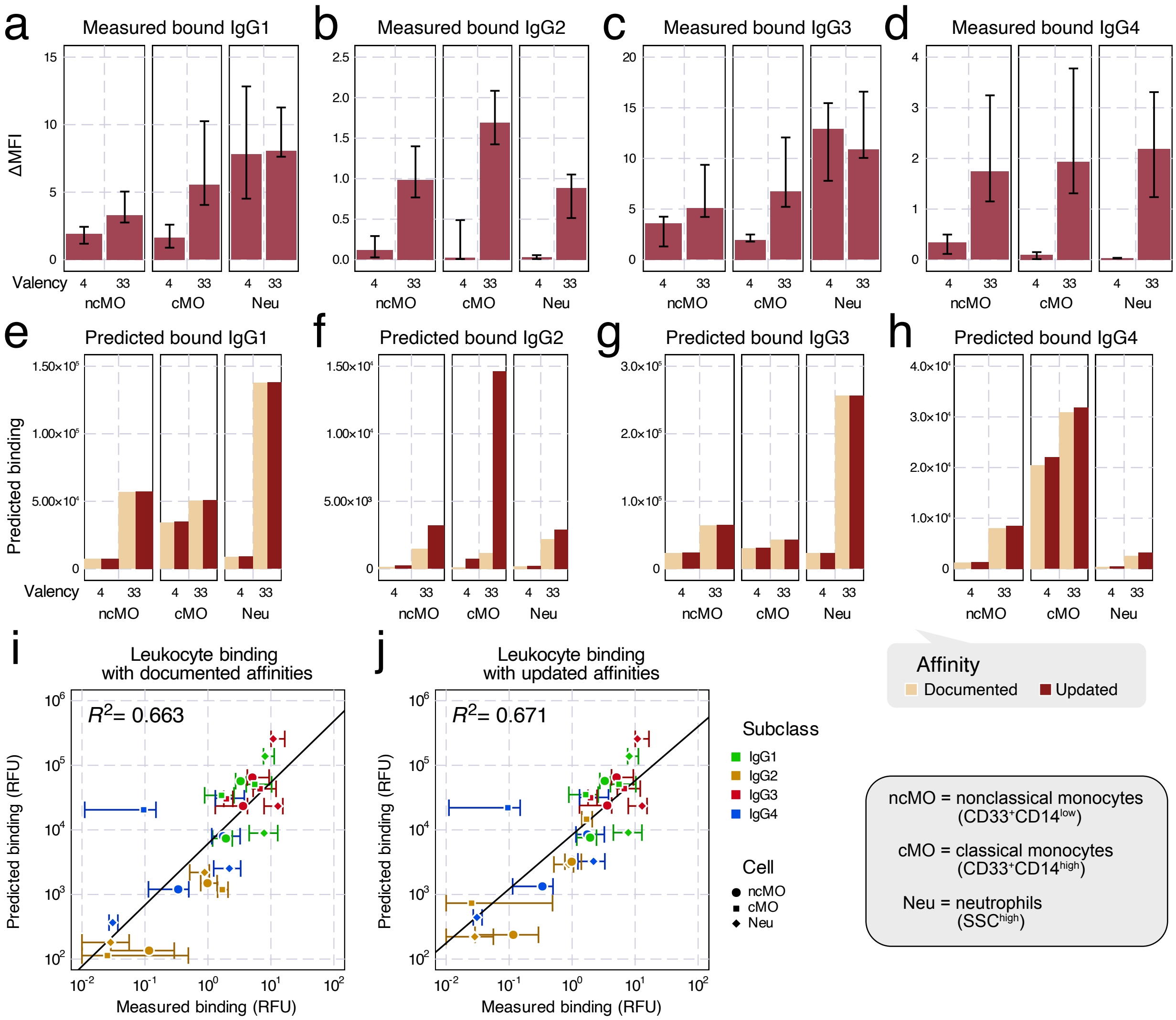


Figure 3.5. Predicting IgG effector cell binding with the multivalent binding model.

(a–d) Measured *in vitro* binding of IgG1 (a), IgG2 (b), IgG3 (c), and IgG4 (d) IC of either 4- or 33-valency to selective immune effector cells from human donors, classical (cMO) or nonclassical (ncMO) monocytes and neutrophils (Neu).

(e–h) Model-predicted IgG1 (e), IgG2 (f), IgG3 (g), and IgG4 (h) IC of 4- or 33- valent binding on each effector cell type under documented versus updated affinities.

(i, j) Measured versus predicted effector leukocyte binding under documented (i) or updated (j) affinities.

The error bars in (a–d) and (i, j) represent the interquartile range of six biological replicates.

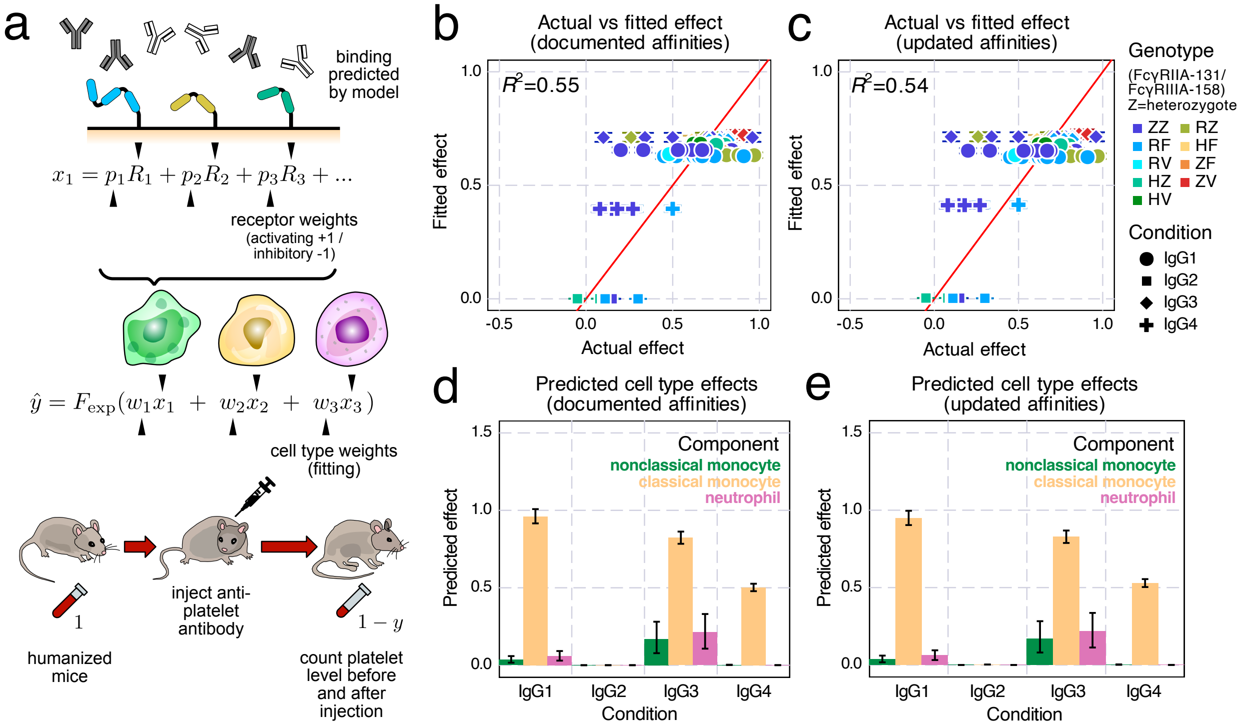


Figure 3.6. *In vivo* target cell depletion regression in humanized mice.

a) Schematic of *in vivo* platelet depletion regression. To predict the percentage decrease of platelet abundance after antibody injection in mice, we combined the binding model predictions with the Fc receptor and effector cell type weights, then transformed the sum into depletion percentage with an exponential distribution cumulative density function.

(b–e) Results of regression run using the documented (b, d) and updated (c, e) affinities.

(b, c) Actual versus predicted depletion of platelets.

(d, e) Predicted effector cell type effects.

Error bars indicate the interquartile range from MCMC sampling.

Next, we used the multivalent binding model with regression to predict *in vivo* antibody effector cell-driven platelet depletion in humanized mice. In the process of extending our previous model, we elected to use the cumulative density function of the exponential distribution as the link function in our generalized linear regression model to link the overall cell activity to the amount of target (e.g., platelet) depletion (Fig. 3.6a). Since the cell depletion effects have a limited range—one cannot deplete an antibody target of more than 100% or less than 0%—we must use a non-linear link function to transform the linear combination. While many functions provide this general relationship (such as the hyperbolic tangent function used before[10]), we realized that the extent of target cell depletion can be thought of as a form of survival analysis. In other words, given a certain antibody activity, a target cell has a certain probability of being cleared within the given timescale of the experiment. Assuming all target cells have an equal propensity of being cleared dictates an exponential relationship for the link function[47].

Having refined the cell clearance model, we applied it to a previously-collected dataset examining *in vivo* platelet depletion in humanized mice[48]. After fitting the cell type weighting, we found the model fit the experiments well, especially considering the experiment-to-experiment variability due to donor graft variation and other sources of experimental uncertainty (Fig. 3.6b/c). The fitting was almost identical when using documented (Fig. 3.6b) or updated (Fig. 3.6c) FcγR affinities.

A benefit of the generalized linear regression model is that it provides an easy interpretation of each component. Inspecting the inferred cell type effects, we found that classical monocytes were inferred to be the predominant effector cell type (Fig. 3.6d/e). IgG2 had some binding to each effector cell type, but no activity was inferred whatsoever (Fig. 3.5f, 6d/e). As the affinity updates are most relevant to IgG2, and this isotype had no *in vivo* effect, it is reasonable that these changes had little effect on agreement with the data (Fig. 3.5f, 6e). While neutrophils, not classical monocytes, had the greatest binding, classical monocytes were inferred to exert the greatest impact on platelet depletion across isotypes (Fig. 3.5a–d). This demonstrates that the most bound cell type does not equate to the most potent effector. One explanation may be that, in these humanized models, there are relatively low numbers of human neutrophils upon reconstitution[49]. The regression model can incorporate the molecular level binding estimation and the depletion outcome to provide insights into the overall potency of each cell type. Overall, we found that the binding model could predict antibody-elicited effector responses *in vivo* in humanized mice.

3.5 Discussion

In this chapter, we explored the binding properties of ICs with mixed IgG Fc composition and linked their *in vitro* effects to *in vivo* effector cell-elicited platelet depletion. To quantify the binding of mixed IgG ICs *in vitro*, we measured every human IgG subclass pair across a range of compositions multimerized at two different valencies (Fig. 3.1). Fitting these measurements to a model of multivalent interactions using documented affinities for each interaction, our model accurately captured the overall binding trends, with some outliers (Fig. 3.3). We uncovered that the model discrepancies could be explained by inaccurate estimates of especially low-affinity Fc receptor interactions, most prominently involving IgG2. We validated revised affinities within an independent dataset and found it greatly improved concordance with the data there as well. Finally, we used measurements of binding to effector cell populations to predict *in vivo* antibody-driven depletion of platelets in humanized mice (Fig. 3.5 & 3.6). While the updated affinities did not change the agreement of the model with the observed depletion, it did change the interpretation of IgG2's small effect on depletion—rather than not binding to classical monocytes, IgG2 binds strongly when in a larger IC, but platelets might provide insufficient avidity to observe sufficient engagement (Fig. 3.6).

Considering that polysaccharide antigens present during bacterial infections or upon vaccination efficiently trigger IgG2 responses[50], our data would support the notion that FcγR-dependent effector functions such as phagocytosis of opsonized bacteria may contribute to protective IgG responses in humans more than expected. Conversely, autoreactive IgG2 responses observed during many autoimmune diseases may contribute to autoimmune pathology via FcγRs, which may warrant to develop therapeutic interventions blocking this pathway also in IgG2-dominated autoimmune diseases[51]. Finally, with respect to the use of human IgG2 antibody formats as immunomodulatory antibodies for the therapy of cancer, our results would support strategies to engineer IgG2 variants with reduced binding to activating FcγRs and optimized binding for the inhibitory FcγRIIB, which has been shown to be critical for immunomodulatory IgG activity to further improve their therapeutic activity and reduce unwanted side-effects[52].

IgG subclasses and glycan variants are defined by their differing affinity toward each Fc receptor[31,45,53]. Therefore, accurate measurements of each Fc receptor affinity are critical to understanding the differences in immune responses to each IgG. Using a mechanistic multivalent binding model alongside *in vitro* binding fluorescence measurements, we were able to derive a new set of Fc affinities refined from those measured by SPR. Due to the heightened avidity, multivalent ICs were better at detecting low-affinity IgG-Fc receptor interactions (Fig. 3.4b). Examining binding through ICs also better simulates the relevant structure of Fc-FcR interactions *in vivo*. Harnessing avidity to overcome the low affinity of interactions is a common theme in immunology and its experimental characterization. For instance, tetramers are routinely used for isolating antigen-selective T cells[9]. Here, we additionally show that these complexes can be used alongside quantitative models to infer properties of these systems.

This framework can be extended to study other aspects of IgG biology such as IgG allotypes. Due to a large number of variants and their implication in ADCC, IgG3 allotypes are of particular interest, with their immunogenicity directly related to their FcγR affinities[54]. In comparison to SPR, our multivalent strategy may allow us to better distinguish the subtle differences in allotype Fc affinities, while the binding model can predict their NK cell responses for different IC valencies. IgG polymorphism also presents in forms independent of Fc binding affinities, such as half-life and hinge length. Having a computational model that can accurately quantify the binding effect may help with separating the affinity-dependent and independent factors, guiding optimal biologic designs.

Our results suggest that, within ICs comprised of several distinct subclasses or glycosylation variants, the Fc interaction effects are a blend of the constituent species' properties. This means that ICs' most extreme binding and effector responses should predominantly arise from whichever species is most potent in eliciting binding or a response. It also should provide some encouragement that the effector responses elicited from therapeutic antibodies should vary roughly in proportion to their relative composition; small contaminants of alternative Fc subclasses or glycosylation can only have a substantial effect if those species differ extremely in their responses alone. One caveat of this observation is that we only examined mixtures of antibodies with differing Fcs but identical antigen binding—polyclonal mixtures of antibodies will have still other interaction effects because antigens can form a higher valency complex when they are present in combination[55]. While in this chapter we only demonstrated Fc subclass mixtures, the same lessons likely apply to glycosylation mixtures, both *in vitro* and *in vivo*, since different subclasses and glycosylation variants exert their effect through divergent affinities toward Fc receptors.

Fc receptor-mediated effects are central to protection from both endogenously produced and therapeutic antibodies. Our work demonstrates that computational methods greatly facilitate reasoning about the complex signaling of the Fcγ receptor pathway quantitatively and at both cellular and organismal levels. This chapter extends our previous modeling to humanized mice and expands its application to the depletion of platelets[10]. We anticipate that mechanistic models of antibody-mediated protection, such as the one here, will continue to grow in their utility for studying model systems such as humanized mice. In fact, as other features of antibodies are incorporated, such as variation in antigen specificity, it may become possible to connect behavior *in vitro* all the way to protection in human subjects[5,56].

Limitations of the study

Although the updated affinities performed better in predicting the binding to human lymphocytes, there were still discrepancies in the IgG4 predictions (Fig. 3.5d/h/j). The predicted binding to classical monocytes was higher than measured, especially with 4-valent complexes. This measurement might have been underestimated, as the measurement was close to 0, while the 33-valent binding was comparable to or higher than those of nonclassical monocytes. IgG4-neutrophil binding was also underestimated. The most expressed FcγR on neutrophils is FcγRIIIB, but IgG4 was previously reported as nonbinding to this receptor. Although they were not included in our subclass mixture study, from measurements of single IgG subclass complex binding to FcγRIIIB-expressing CHO cells, we inferred that the IgG2 and IgG4 affinities are both much lower than 10^{-5} M^{-1}, supporting the documented nonbinding estimation (Fig. 3.S4; Tbl. 3.S3 & 3.S4).

Alternatively, evidence exists that neutrophils also express a low level of FcγRIIIA which has adequate binding to IgG4[57].

To investigate the *in vivo* implication of our revised FcγR affinity updates, we elected to use humanized mice as a model system. This is both a strength and a limitation of this study. Humanized mice serve as an ideal surrogate for understanding human immunity[58]. However, this model system is complicated by graft-to-graft differences, including the level of humanization and genetic heterogeneity of human stem cell donors[58]. The depletion data reflected these complications, with high donor-to-donor and mouse-to-mouse variation, limiting our ability to observe subtle changes (Fig. 3.6b/c)[48].

3.6 Material and Methods

Experimental Model and Study Participant Details

Unique reagents generated and key resources used in this chapter can be found in the table in the Star Methods section of the published work[59].

Aiming to investigate IC binding to primary human leukocytes, blood was drawn from six healthy volunteers with the informed consent of the donor and the local ethical committee.

Experiment Details

Chinese hamster ovary (CHO) cell FcγR expression quantitation

Human FcγR expression on stably transfected CHO cells was quantified by determining the antibody binding capacity (ABC) for antibodies specific to the respective Fcγ receptor (Tbl. 3.S2)[40]. Quantum Simply Cellular (QSC) anti-mouse beads (Bangs Laboratories Ltd.) with known binding capacities for mouse IgG were used according to manufacturer's instructions. Subsequently, a reference curve was generated by correlating the fluorescence intensity (caused by the respective anti-FcγR antibody) and the number of antibody binding sites of the different QSC beads. This reference curve was established in each experiment for all FcγR-specific antibodies of interest (PE-conjugated clone 10.1 to detect FcγRI, clone AT10 to detect FcγRIIA/B and clone 3G8 to detect FcγRIIIA, all from Biolegend) and used to calculate receptor numbers based on fluorescence intensity of FcγR staining on CHO cells. Samples were measured on a FACSCantoII flow cytometer and analyzed with FACSDiva software.

Immune Complex Binding Measurement

CHO cells stably expressing human FcγRs were used to assess IgG-IC binding to hFcγRs as previously described[33]. Briefly, ICs were generated by coincubation of 10 μg/ml anti-TNP human IgG subclasses (clone 7B4, produced in-house) and 5 μg/ml BSA coupled with either an average of 4 or 33 TNP molecules (Biosearch Technologies) to mimic low or high valency ICs, respectively, for 3 h with gentle shaking at room temperature. To address the impact of distinct subclass combinations on binding to hFcγRI, hFcγRIIA-131H/R, FcγRIIB and FcγRIIIA-158F/V, human IgG1 through IgG4 subclasses were mixed at specific conditions (100%, 90%, 66%, 33%, 10% of one subclass filled up to 10 μg/ml with the respective second subclass) before the addition of TNP-BSA. CHO cells stably expressing FcγRIIIB NA1 and NA2 variants were generated for this study and employed to determine IgG subclass binding for 100% IgG1-4 immune complexes of low and high valency. ICs were subsequently incubated with 100,000 FcγR expressing or untransfected control CHO cells for 1 h under gentle shaking at 4°C. Bound ICs were detected using a PE-conjugated goat anti-human IgG F(ab')$_2$ fragment at 0.5 μg/ml (Jackson ImmunoResearch Laboratories) on a BD FACSCanto II flow cytometer. To calculate the fluorescence signal intensity (median fluorescence intensity, MFI) of specific immune complex binding, the background fluorescence intensity of anti-human IgG F(ab')$_2$ stained control cells was subtracted (ΔMFI). The measured IC fluorescence intensities were between 1,000 to 15,000, far from the equipment saturation level which occurred at around 260,000. Each experimental condition had 3–5 technical replicates. The relative fluorescence unit of each IC binding was normalized so that measurements of each day had geometric means of 1.0.

Alternatively, binding to human primary peripheral blood leukocytes co-expressing specific FcγRs was studied. Blood was drawn from healthy volunteers and erythrocytes were lysed by the addition of ddH_2O for 30 sec at room temperature to obtain total leukocytes. Immune complexes were generated as described above and incubated with 200,000 leukocytes. Leukocyte subpopulations were identified by staining cell-type-specific surface markers. Fluorescently labeled antibodies PE/Cy7-conjugated anti-CD19, PerCP-conjugated anti-CD3, APC-conjugated anti-CD33, Brilliant Violet 510 conjugated anti-CD14, FITC-conjugated anti-CD56 and APC/Fire 750 conjugated anti-CD45 were obtained from Biolegend. Immune complex binding was quantified upon staining with PE-conjugated goat anti-human IgG $F(ab')_2$ fragment at 0.5 μg/ml (Jackson ImmunoResearch Laboratories) and data acquisition on a BD FACSCanto II flow cytometer.

The cell identification strategy was as follows: aggregates of cells were excluded by their forward light scatter (FSC) characteristics (area vs. height) and dead cells based on staining with DAPI. Leukocytes were identified by expression of common leukocyte marker CD45. Among those, neutrophils were gated based on high side light scatter (SSC) characteristics and lack of surface CD14, and classical monocytes were based on intermediate SSC and expression of CD14. Within the $CD14^{-}SSC^{low}$ cells, B and T cells were gated by expression of CD19 or CD3, respectively. Staining of CD56 was used to distinguish NK cells. The remaining CD33-expressing cells were gated as nonclassical monocytes. ΔMFI of bound immune complexes was calculated by subtracting the background fluorescence intensity of PBS-treated leukocytes.

Data were analyzed with FlowJo or FACSDiva Flow Cytometry Analysis Software. Six (6) biological replicates were measured for each IC valency, IgG subclass, and leukocyte cell type combination. All measurements were normalized so that the daily geometric means were 1.0.

Quantification and Statistical Analysis

All statistical and computational analyses in this chapter were implemented by Julia v1.8.

Principal component analysis on mixture binding measurements

Principal component analysis on the IgG mixture binding measurement was performed with the package MultivariateStats.jl. The variance explained by principal component analysis was defined as $1-\frac{\|X-\hat{X}\|_F^2}{\|X\|_F^2}$, where $\|\cdot\|_F$ indicates the Frobenius norm.

Generalized multi-ligand, multi-receptor multivalent binding model

To model polyclonal antibody-antigen immune complexes (ICs), I employed a multivalent binding model to account for ICs of mixed IgG composition previously developed and detailed[46]. In this case, we define N_L as the number of distinct monomer Fc's and N_R as the number of FcRs, and the association constant of monovalent Fc-FcR binding between Fc i and FcR j as $K_{a,ij}$. Multivalent binding interactions after the initial interaction are assumed to have an association constant of $K_x^* K_{a,ij}$, proportional to their corresponding monovalent affinity. The concentration of complexes is L_0, and the complexes consist of random ligand monomer assortments according to their relative proportion.

Immune Complex Binding Analysis

Fitting the parameters in the binding quantification was performed by Markov chain Monte Carlo (MCMC) implemented by Turing.jl[60].

At first, we plugged the documented values into the binding model for all parameters without fitting, thus the geometric means of CHO cell receptor expression (Tbl. 3.S2), documented affinities[31], nominal valencies (4 and 33), and K_x^* as $6.31 \times 10^{-13}\text{cell} \cdot \text{M}$[10], as estimated in previous work (Fig. 3.3c)[10]. To examine the role of affinity fitting, we used MCMC to fit all parameters except (Fig. 3.3d) and including (Fig. 3.3e) affinities. CHO receptor prior distributions were inferred from their measured values through maximal likelihood estimation (MLE) in Distributions.jl[61] for both IgG mixture dataset (Tbl. 3.S2) and validation dataset[10] (Tbl. 3.S5). The affinity priors were inferred from documented Fc affinities and standard errors following several assumptions: (1) each prior follows a log-normal distribution; (2) the mode of the distribution is the documented value, and the interquartile range of the distribution is the standard error; (3) if the values of mode or standard errors are too small, the mode was clipped to $1 \times 10^4\ \text{M}^{-1}$, and the interquartile range was clipped to $1 \times 10^5\ \text{M}^{-1}$ to deal with recorded nonbinding cases[31,62]. The priors of the effective valency and crosslinking constant were:

$$f_4 \sim \log N(\mu = \log(4)\,, \sigma = 0.2),$$

$$f_{33} \sim \log N(\mu = \log(33)\,, \sigma = 0.2),$$

$$K_x^* \sim \log N(\mu = \log(6.31 \times 10^{-13})\,, \sigma = 2.0).$$

MCMC was initialized with the maximum a posteriori estimation (MAP) optimized by a limited-memory BFGS algorithm implemented by Optim.jl[63], then sampled through a No U-Turn Sampler (NUTS) implemented by Turing.jl[60].

In vivo Regression Model

We extended the *in vivo* antibody-elicited target cell depletion regression model with both cell type weights and FcγR weights (Fig. 3.6a). Depletion, y, was represented as the percent reduction in the number of target cells.

To quantify the activity of each effector cell, we first used the multivalent binding model to predict the amount of multimerized FcγR of each kind, $R_{\text{multi},i}$, assuming each IC is 4-valent. Then the activity of this cell type is assumed to be a linear combination of these predictions and a set of cell type weights, p_i, that are set to either +1 or -1 for activating or inhibitory receptors, respectively, clipped to 0 if it is negative:

$$x_n = \max\left(p_1 R_{\text{multi},1} + p_2 R_{\text{multi},2} + \cdots, 0\right).$$

To determine how these cell types bring the depletion effect at the organism level, we combine their estimated effects, x_n, with a weighted sum, where we introduce another set of weights, w_n, that are specific to each cell type. To convert the activities to a limited range of depletion (i.e., one cannot have a reduction over 100%), the regression was transformed by an exponential linker function (the cumulative density function of exponential distribution) such that the predicted effectiveness: $\hat{y} = F_{\exp}(wx) = 1 - \exp(-wx)$ so that $\lim_{X \to \infty} F_{\exp}(X) = 1$. Together, we defined the estimated depletion as

$$\hat{y} = F_{\exp}(w_1 x_1 + w_2 x_2 + \cdots).$$

We did not estimate the amount of each cell type in an individual, nor did we include them in the model, because the weights, w_n, are meant to absorb these quantities, while requiring effector cell abundance would limit the application of this model to organs where the tissue resident cell abundance has been accurately quantified.

The regression against *in vivo* effectiveness of IgG treatments was performed via MCMC implemented by Turing.jl[60]. For the multivalent binding model, the ligand concentration was assumed to be 1 nM and valency to be 4. The receptor expression level was set to the geometric means of the values measured in previous work (Tbl. 3.S6)[40]. For the receptor weights, p_i, we set the weight of the only inhibitory receptor, FcγRIIB, as -1.0 and every activating receptor to be

$+1.0$. The predicted cell type effects were estimated by multiplying the cell type weights by their predicted activities, the weighted sum of multimerized receptor from the binding model.

MCMC was initialized with MAP optimized by a limited-memory BFGS algorithm implemented by Optim.jl[63], then sampled through NUTS implemented by Turing.jl[60].

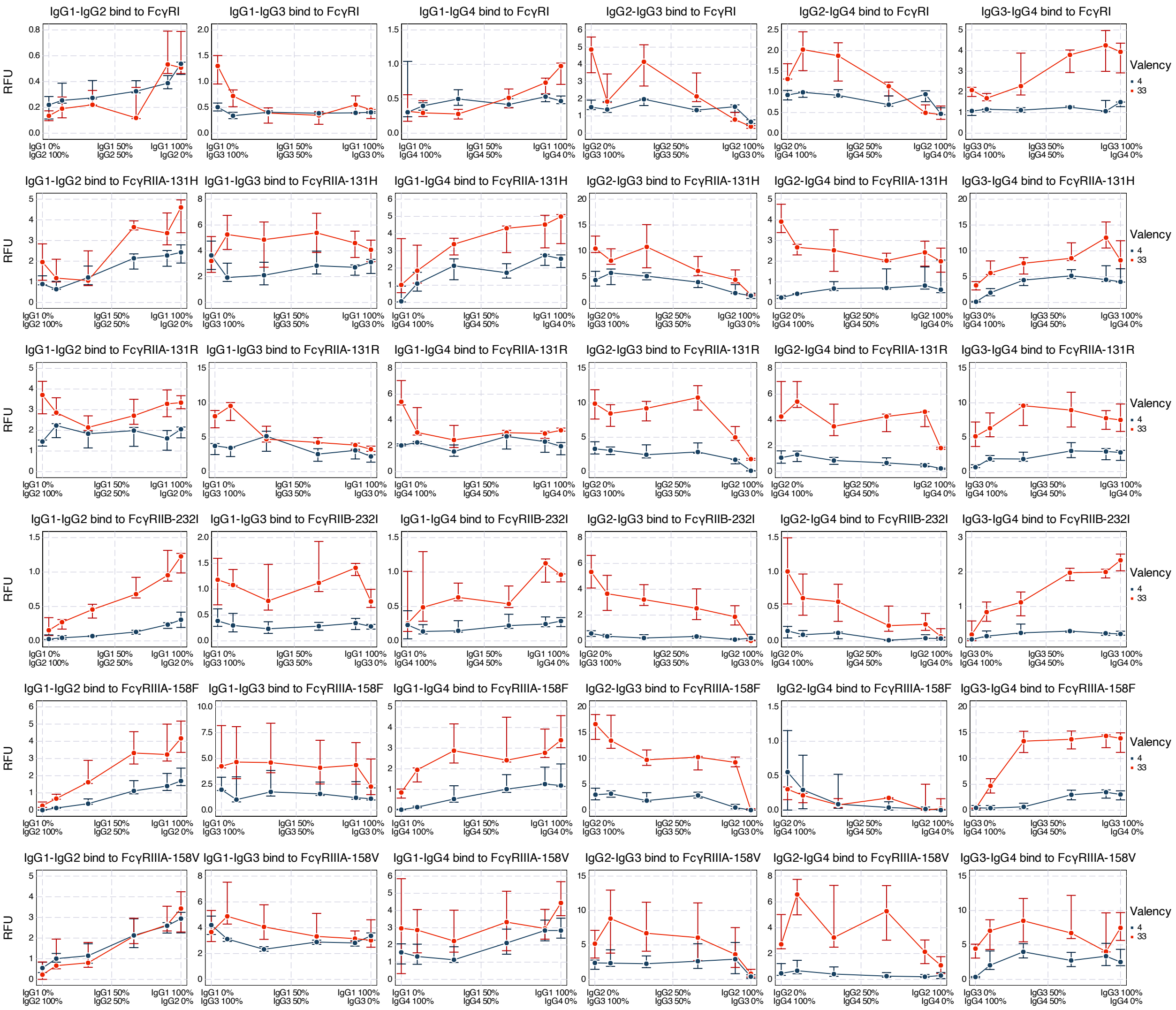


Figure 3.S1. Experimental IC mixture binding data.

Quantification of human IgG subclass pairs TNP-4-BSA and TNP-33-BSA IC binding to CHO cells expressing the indicated hFcγRs. Relative fluorescent units (RFU) of different multivalent immune complexes consisting of various IgG mixtures binding to different human immune cell receptors. Error bars indicate the 3-5 technical replicates from experiments (see methods for details). Fluorescent values were normalized so that the daily geometric average measurements are 1.

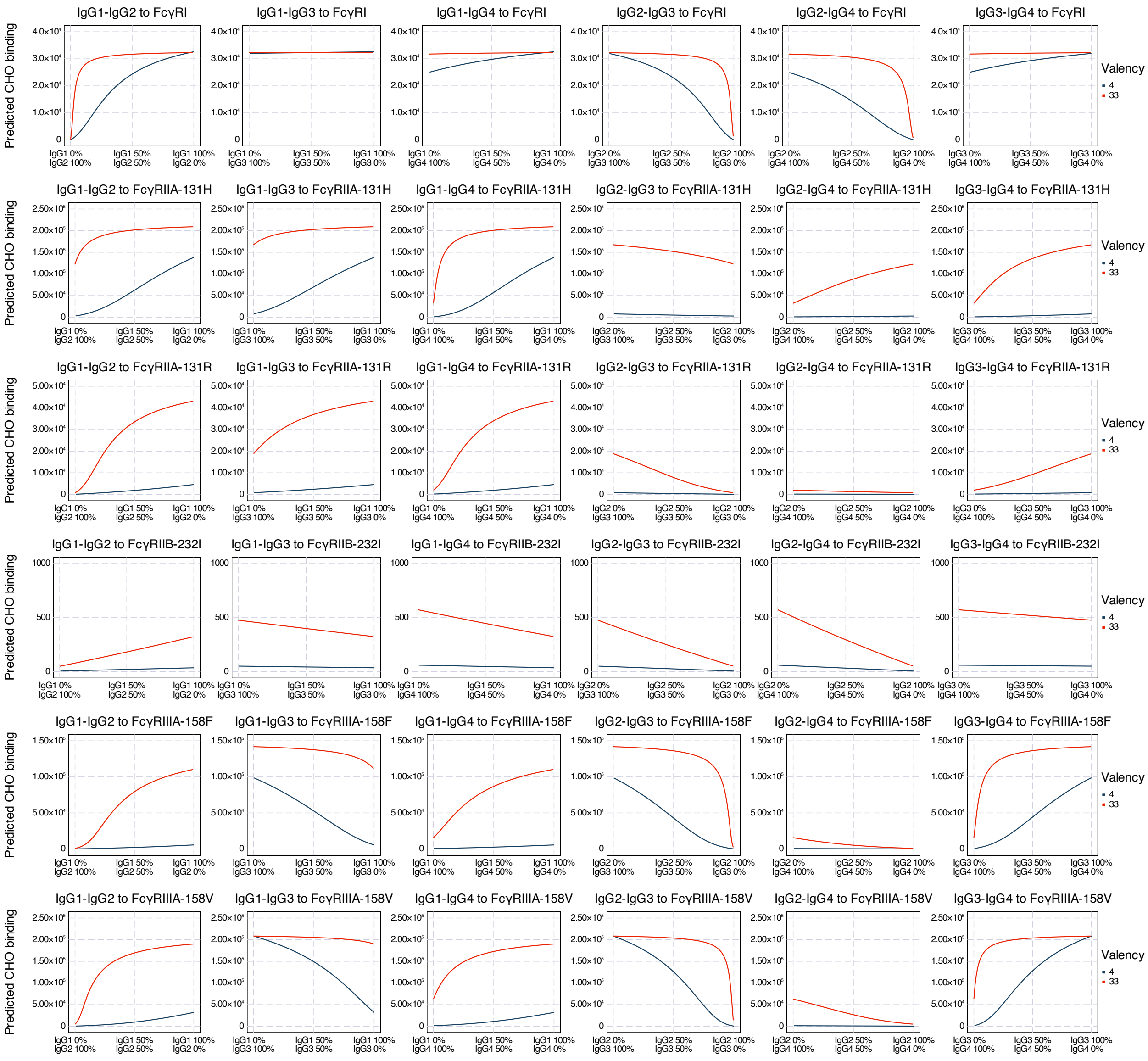


Figure 3.S2. Predicted binding of IgG subclass mixtures with documented affinities.

Amount of binding for complexes of each IgG subclass pairs binding to each CHO cell predicted by the multivalent binding model with the documented affinities. The receptor abundances were as geometric means of measurement.

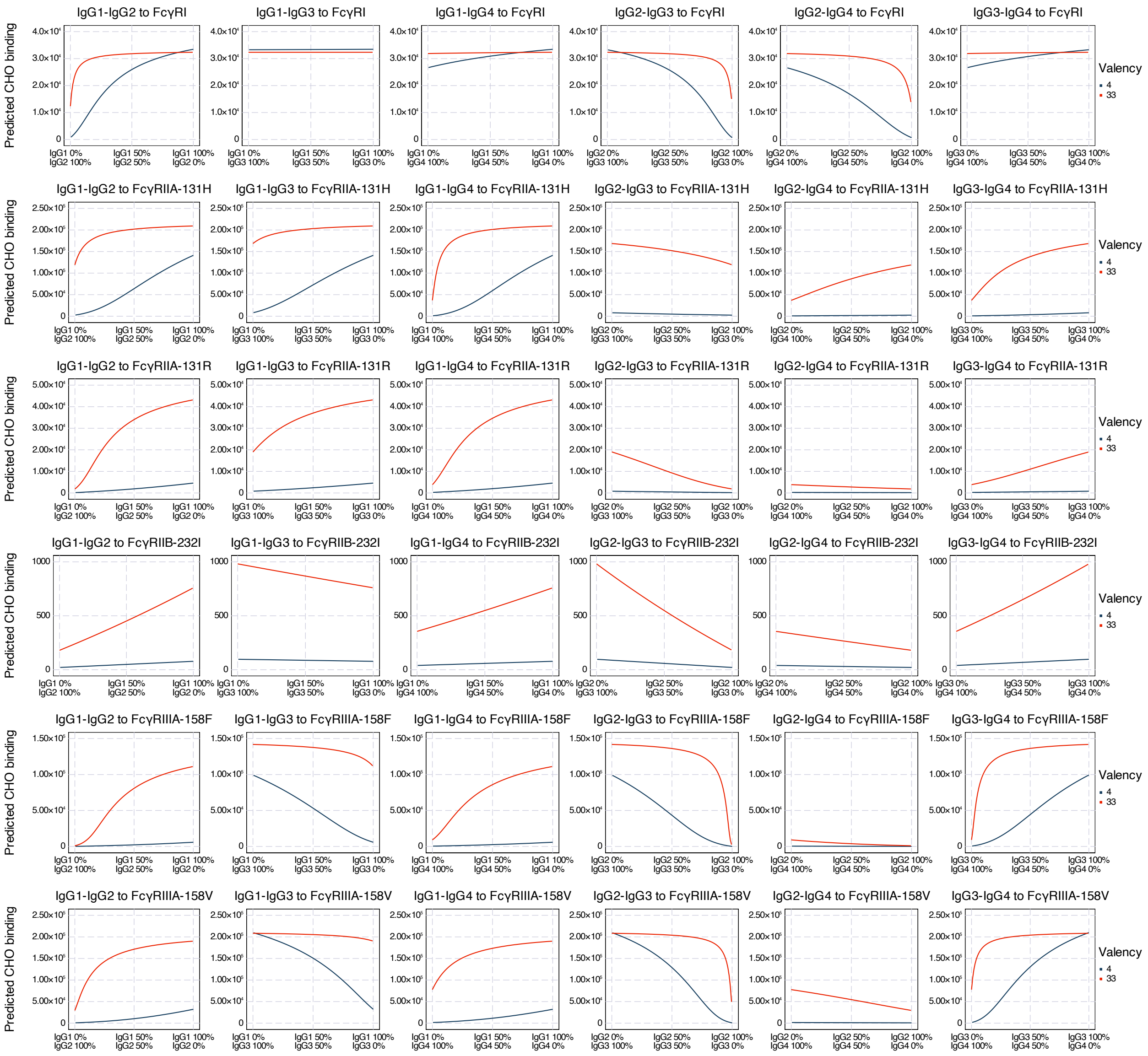


Figure 3.S3. Predicted binding of IgG subclass mixtures with updated affinities.

Amount of binding for complexes of each IgG subclass pairs binding to each CHO cell predicted by the multivalent binding model with the updated affinities. The receptor abundances were as geometric means of measurement.

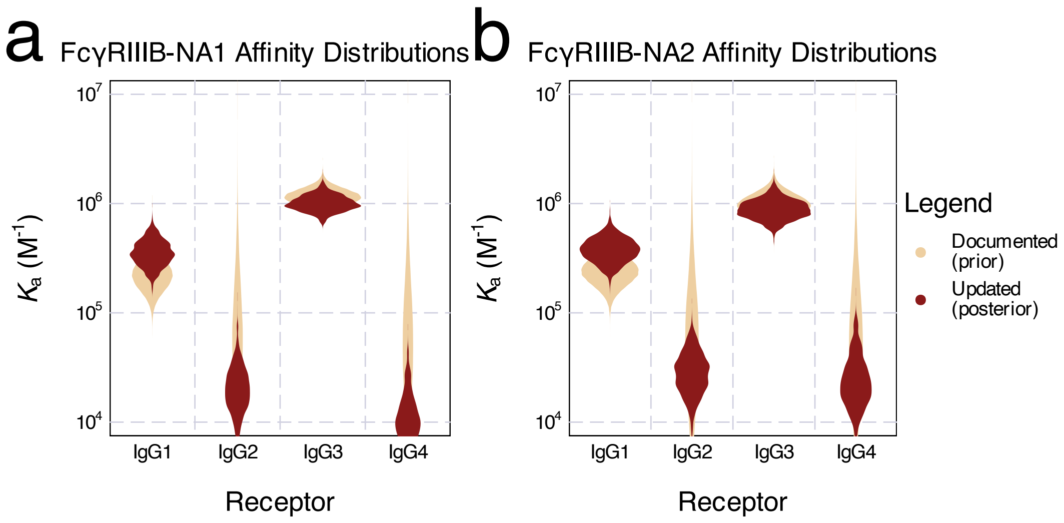


Figure 3.S4. Inferred affinities of IgG subclasses to FcγRIIIB variants.

The prior (assume all follow log-normal distributions) and posterior (updated) distributions of IgG binding affinities to FcγRIIIB-NA1 (a) and NA2 (b) variants. The binding FcγRIIIB affinities were inferred from pure IgG subclass immune complexes of 4 or 33 valencies binding to CHO cells stably expressing the NA1 and NA2 variant of FcγRIIIB, respectively. Data represents a separately measured dataset from the mixture binding fitting presented in Fig. 3.3 & 4. Immune complex binding was assessed in six independent experiments as described in the method section. Notice that previously IgG2 and IgG4 are both reported nonbinding to either variant of FcγRIIIB[31]. For better MCMC fitting, the prior distributions for their affinities were inflated to have medians 10^4 M^{-1} and interquartile ranges 10^5 M^{-1}.

Source	DF	SS	MSS	*F*	*p*
Condition	431	11110.23	25.778	6.00915	5.9926×10^{-125}
Residuals	1066	4572.88	4.290		
Total	1497	15683.11			

$R^2 = 0.7084$

Table 3.S1. One-way ANOVA performed on the measurements.

It indicates that the majority of variance in measurements comes from between conditions.

Receptor	Geometric mean	Inferred distribution
FcγRI	101494	logN(μ=11.53, σ=0.26)
FcγRIIA-131H	1006300	logN(μ=13.82, σ=0.27)
FcγRIIA-131R	190433	logN(μ=12.16, σ=1.46)
FcγRIIB-232I	75085	logN(μ=11.23, σ=1.40)
FcγRIIIA-158F	634324	logN(μ=13.36, σ=0.70)
FcγRIIIA-158V	979452	logN(μ=13.80, σ=0.38)

Table 3.S2. Geometric mean and inferred prior distribution of FcR abundance.

Antibody binding capacity on CHO cells from measurements used in IgG mixture *in vitro* binding experiment. The geometric means were calculated from primary data published in previous work[40]. logN represents a log-normal distribution.

K_a (M^{-1})	IgG1	IgG2	IgG3	IgG4
FcγRI	5.809~8.635×10^7	1.189~1.911×10^6	5.525~8.730×10^7	3.012~4.770×10^7
FcγRIIA-131H	4.780~5.929×10^6	3.930~4.724×10^5	8.622~9.621×10^5	1.625~2.057×10^5
FcγRIIA-131R	3.277~3.758×10^6	1.719~2.339×10^5	8.521~9.861×10^5	2.833~3.864×10^5
FcγRIIB-232I	2.166~3.053×10^5	5.931~8.134×10^4	2.714~3.702×10^5	1.131~1.504×10^5
FcγRIIIA-158F	1.106~1.288×10^6	2.938~4.405×10^4	7.243~8.329×10^6	1.337~1.764×10^5
FcγRIIIA-158V	1.913~2.107×10^6	1.433~1.966×10^5	0.919~1.073×10^7	2.509~3.348×10^5
FcγRIIIB-NA1*	2.839~4.227×10^5	1.375~2.635×10^4	0.891~1.088×10^6	0.692~1.332×10^4
FcγRIIIB-NA2*	3.243~4.474×10^5	1.886~3.574×10^4	0.787~1.016×10^6	1.627~3.138×10^4

Table 3.S3. Updated Fc affinities' interquartile ranges from their posterior distributions.

* FcγRIIIB affinities were inferred from single-subclass immune complexes binding to CHO cells. They were fitted separately from the other receptors. See Fig. 3.S4 for distribution details.

K_a (M^{-1})	IgG1	IgG2	IgG3	IgG4
FcγRI	7.140×10^{7}	1.496×10^{6}	6.998×10^{7}	3.835×10^{7}
FcγRIIA-131H	5.314×10^{6}	4.306×10^{5}	9.128×10^{5}	1.808×10^{5}
FcγRIIA-131R	3.503×10^{6}	2.005×10^{5}	9.178×10^{5}	3.291×10^{5}
FcγRIIB-232I	2.549×10^{5}	6.911×10^{4}	3.157×10^{5}	1.305×10^{5}
FcγRIIIA-158F	1.195×10^{6}	3.639×10^{4}	7.741×10^{6}	1.536×10^{5}
FcγRIIIA-158V	2.006×10^{6}	1.678×10^{5}	9.941×10^{6}	2.906×10^{5}
FcγRIIIB-NA1*	3.430×10^{5}	1.906×10^{4}	9.806×10^{5}	9.381×10^{3}
FcγRIIIB-NA2*	3.740×10^{5}	2.617×10^{4}	8.908×10^{5}	2.168×10^{4}

Table 3.S4. Updated Fc affinities median values.

* FcγRIIIB affinities were inferred from single-subclass immune complexes binding to CHO cells. They were fitted separately from the other receptors. See Fig. 3.S4 for distribution details.

Receptor	Geometric mean	Inferred distribution
FcγRI	232872	logN(μ=12.36, σ=0.25)
FcγRIIA-131H	318819	logN(μ=12.67, σ=0.26)
FcγRIIA-131R	1605372	logN(μ=14.29, σ=0.14)
FcγRIIB-232I	394556	logN(μ=12.89, σ=0.45)
FcγRIIIA-158F	4677645	logN(μ=15.36, σ=0.22)
FcγRIIIA-158V	3680708	logN(μ=15.12, σ=0.25)

Table 3.S5. Geometric mean and inferred prior distribution of FcR abundance.

Antibody binding capacity on CHO cells used in the validation binding dataset[10]. The geometric means were calculated from primary data published in previous work[40]. logN represents a log-normal distribution.

Receptor	Non-classical monocyte	Classical monocyte	Neutrophil
FcγRI	6326	84559	1847
FcγRIIA	82542	96646	158228
FcγRIIB	7140	5167	2351
FcγRIIIA	200213	19533	0*
FcγRIIIB	0*	0*	1299166

Table 3.S6. Geometric means of measured FcγR expression.

FcγR expression means. the number of quantified binding sites for the respective anti-FcR antibodies on effector cells calculated from the primary data published in previous work[40].

* Although one cannot have a geometric mean as 0, these values were consistently measured as non-expressed, so we used 0 as the value.

Chapter 4

The structure is the message: preserving experimental context through tensor decomposition

With no hindrances, there is no fear;
Freed from all distortion and delusion,
Ultimate nirvana is reached.

Heart Sutra

4.1 Introduction

Multiplex and high-throughput assays now enable the exploration of cell responses in unprecedented scale and detail. Consequently, studies of biological systems have increasingly focused on profiling biological systems across multiple contexts (Tbl. 4.1). For instance, a panel of candidate therapies might be profiled using cell samples derived from multiple organs, with several features of their response measured over time (Fig. 4.1a). Identifying how responses are shared or distinct across multiple cellular contexts and experimental conditions reveals more about the biological mechanism and enhances the generalizability of the results. At the same time, measuring cell lines and tissues across multiple parameters generates data with multiple dimensions (e.g. cell line, time, experimental conditions), which necessitates reevaluating how we represent and analyze such information.

Representing multivariate data in a tabular form can sacrifice the ultimate insight that can be derived. It is not uncommon that studies with several dimensions are still laid out in rows and columns with some dimensions merged. For the example in Fig. 4.1a, when the experiment is repeated over time, the columns must expand to combine two experimental parameters, drug and time point, such as “alfazumab – 1 hr,” “alfazumab – 3 hr,” “bravociclib – 1 hr,” “bravociclib – 3 hr”, etc. In this format, one may instinctively apply familiar off-the-shelf statistical approaches, such as principal component analysis (PCA), because the data appears to be in matrix form.

So, what is the problem with this? As communication philosopher Marshall McLuhan famously stated[64], “The medium is the message.” The choice of data structure influences its analysis and the subsequent insights. A tabular form implicitly treats each column and row as separated from one another, while merged dimensions diverge from this assumption. For instance, “alfazumab – 1 hr” and “alfazumab – 3 hr” share the same treatment, and “alfazumab – 1 hr” and “bravociclib – 1 hr” share the same timing; however, “bravociclib – 1 hr” and “alfazumab – 3 hr” differ in two distinct ways (Fig. 4.1b). When flattening a multidimensional dataset into a table, we compromise this property.

To devise a more effective approach, the “medium”, or structure, of the experiment must be incorporated. The example experiment varies across three degrees of freedom: organ, drug, and time; this is best represented by a three-dimensional array or tensor (Fig. 4.1c). A tensor representation aligns entries with shared meaning. For instance, when examined from the perspective of an organ (e.g., thymus, the green cubes), we find the pharmacokinetics profiles of all drugs on this organ; when viewed from a drug (e.g., foxtrotolol, the pink cubes), we find its impact on all organs over time (Fig. 4.1c).

Ref.	Brief description	Data modality	Contexts
[65]	Gene expression in *S. cerevisiae* cultures	DNA microarray	Genes, Time points, Conditions
[66]	Gene expression in across multiple human tissues	RNA sequencing	Individuals, Genes, Tissues
[67]	Metabolite profiles across cancer cell lines in Cancer Cell Line Encyclopedia	liquid chromatography–mass spectrometry	Cell lines, Metabolites, Genes
[68]	Synovial fibroblasts cytokine secretion after exposed to drug perturbations	Luminex assay	Samples, Stimuli, Inhibitors
[69]	Human Lung Cell Atlas	single-cell RNA sequencing	Cell types, Individuals, Gene, Anatomical locations
[70]	Metagenome data in Human Microbiome Projects	metagenomic whole genome shotgun sequencing	Subjects, Time point, Body sites
[71]	Protein expression change in human mammary epithelial cell after perturbation	reverse phase protein array	Proteins, Treatments, Time
[72]	Roadmap Epigenomics data from ENCODE project[73]	various epigenomics data	Cell types, Assays, Genomic positions
[74]	Height and weight-related traits from UK Biobank	physiological data	Individuals, Traits, Time points

[75]	Neuron recordings across time and trials in rodents and monkeys	neuronal firing rate	Neurons, Trials, Time

Table 4.1. Some examples of multivariate biological datasets.

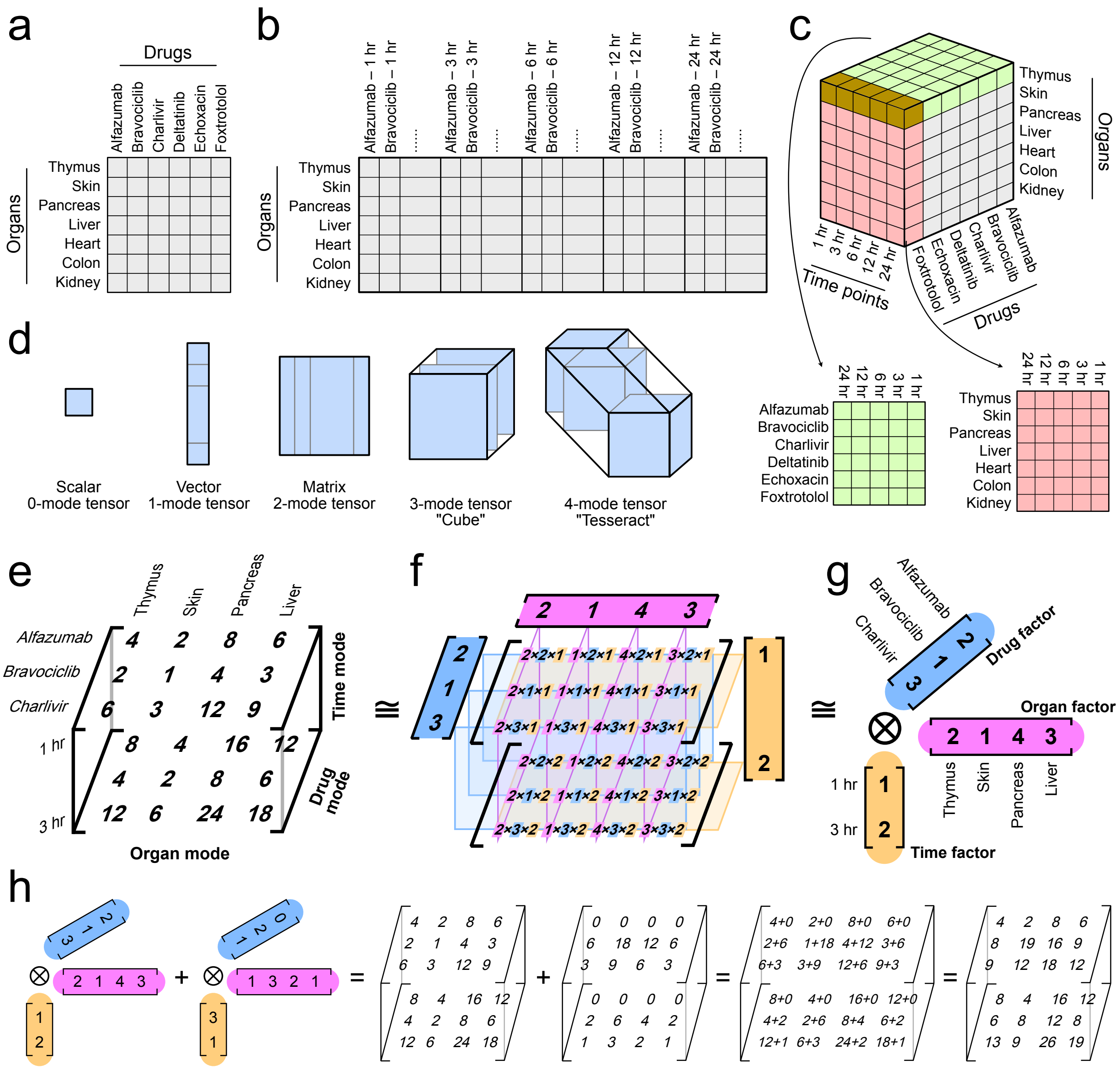


Figure 4.1. Basic concepts of tensor-structured data

(a) A dataset of cells collected from different organs responding to various drug treatments can be documented by a table.

(b) When the measurements are performed over multiple time points, the original columns in the table can be expanded into multilevel indices, recording both drug and time. This nonetheless breaks the assumption that all columns are equally related.

(c) Alternatively, the same data can be recorded as a three-dimensional array, with organ, drug, and time as three separate degrees of freedom. Here, the pink represents how every cell responds to foxtrotolol over time, and the green represents the pharmacokinetic profile of cells from the thymus over all treatments. The brown is shared by pink and green.

(d) Tensors are multidimensional arrays. A dimension of a tensor is a mode. 0, 1, and 2-mode arrays are known as scalar, vector, and matrix.

(e) An example of a rank-one tensor. A subset of the drug response dataset on cells from four organs responding to three drugs over two time points and has dimensions 4×3×2, organ by drug by time.

(f) Rank-one tensors are those whose every entry can be written as the product of a few numbers, one from each mode-specific vector, from their corresponding coordinates.

(g) A rank-one tensor can be written as multiple mode-specific factors joined by the vector outer product, $\otimes$.

(h) Even written as sets of vectors, these rank-one tensors should still be understood as arrays with numbers in every entry. Adding two tensors of the same shape is to add their corresponding position together.

In this chapter, I aim to provide an overview of how tensor methods have been applied and benefited systems biology studies, and how they can be deployed for wider research fields. I propose that tensor methods should and will become an established part of the basic biomedical data sciences toolbox.

Defining tensors and tensor decomposition

Tensors are nothing more than multidimensional arrays[6,76,77]. Zero-, one- and two-dimensional tensors are scalars, vectors, and matrices, respectively (Fig. 4.1d). To avoid conflicting definitions of "dimension" in linear algebra, "mode" or "order" are used—three-dimensional, three-mode, and third-order tensors are all the same concepts. A matrix has two modes—columns and rows—but tensors over three modes do not have mode-specific names. When structuring biological data into a tensor, each mode ideally relates to a varied parameter of the experiment, such as samples, genes, cell lines, treatments, concentrations, or time points.

Tensors can be analyzed by tensor decomposition. Before describing this, it is helpful to introduce the concept of rank-one tensors, the building block for tensor decomposition. Like with matrices, even large data tensors can be decomposed into a series of simple patterns, known as rank-one tensors. Unlike the concept of tensor mode, which is directly associated with the data dimensionality, the rank of a tensor is a separate and less evident concept that requires examining its entries. As an illustrative example, consider a smaller dataset with the response of cells from four organs to three drugs over two time points. By stacking the measurements at 1 hr (a 4×3 matrix) on top of the measurements at 3 hr (another 4×3 matrix), we obtain a 4×3×2 tensor with organ, drug, and time modes (Fig. 4.1e). In this contrived example, along the drug mode, every vector is a multiple of [2, 1, 3]. This indicates that all eight samples have the same drug-reaction profile. The measurements collected at 3 hr mark are double those at 1 hr, suggesting that all measured effects increase to twice the magnitude from 1 hr to 3 hr. The organ factor is [2, 1, 4, 3], indicating the ratio of the four organs' reaction magnitude: cells from the thymus react twice as much as cells from the skin, while cells from the pancreas and liver exhibit effects of four and three times as cells from the skin, respectively. Every entry in this tensor can be precisely computed by multiplying three numbers, each from the organ, drug, and time factor with their positions corresponding to its position in the tensor (Fig. 4.1f). To describe this property, we define this tensor as the outer product of these three vectors (Fig. 4.1g). Tensors that can be expressed as the outer product of a vector set are known as rank-one tensors. The number of vectors within the set is the order of this rank-one tensor; therefore, a rank-one tensor can have any number of modes. Rank-one tensors exhibit a single pattern association with each mode, enabling straightforward interpretation.

Most tensors are more complex than rank-one tensors. Nonetheless, by expressing them as the sum of rank-one tensors (Fig. 4.1h), interpretation becomes significantly easier, since they can be understood as the combination of these rank-one individual patterns. Even if we do not represent the original tensor exactly, if a small number of patterns can closely approximate the original tensor and capture essential information, we can still gain insights into the overall trends. This process of breaking down a complex tensor into the sum of a few patterns is known as tensor decomposition or tensor factorization.

4.2 A step-by-step guide on tensor decomposition

Structuring the data into a tensor format

Organizing a dataset into a tensor requires recognizing the structure defined by the experiment. In the example presented in Fig. 4.1c, it is natural to use a three-mode tensor with organ, drug, and time modes[78]. Tensor order can extend beyond three dimensions if, for instance, each organ, drug, and time combination was performed across multiple assays (e.g., measurement of many genes or proteins).

Measurements can only be separated into a distinct mode when the mode's labels relate to a common experimental entity across which the data can be grouped accordingly[79]. For example, should multiple technical replicates for each condition be grouped in a separate mode? No, because the "Sample 1" replicate of cells from the liver does not signify the same replicate as "Sample 1" of cells from the skin. We may either average these replicates during reformatting if their variation is not of particular interest, or apply resampling strategies to preserve replicate variances[80]. However, if these samples represent a common set of patients—"Sample 1" is the same for all cell types indicating that they came from the same individual—this justifies the inclusion of a corresponding mode. Similarly, single-cell measurements from different samples inherently come from different cells; therefore, single cells cannot form a distinct tensor mode when attempting to parallelize samples. They may only form a tensor mode when multi-omic assays are performed on identical cells. As another example, in single-cell analysis, when combining runs from different backgrounds, whether the clusters with the same label should be aligned depends on whether each cluster label holds the same meaning across backgrounds. If the cluster labels are assigned randomly (e.g., in *k*-means), they are not equivalent between runs, and therefore cannot form a separate "Cluster" mode. However, if the clusters can be identified based on cell surface markers and "Cluster 1" consistently represents the same cell type, this cell type mode is justified.

In a tensor format, the items representing positions along a mode are treated separately. Therefore, the order of items on a mode is inconsequential to the tensor decomposition. For instance, switching the positions of "3 hr" and "12 hr" on the time mode in the tensor in Fig. 4.1c does not affect its decomposition results. For longitudinal measurements where sometimes the time points cannot be aligned perfectly, compromises may have to be made. One approach can be binning, where similar time points of different samples are grouped into one category. For instance, if one individual only has samples at "3 hr", while another only has "4 hr", a binned "3-4 hr" category may be created to align them. Sometimes, several positions in the tensor may be left empty to maintain the data's logical structure (see "Missing data and imputation" on decomposition with missing entries).

Tensor decomposition can benefit from appropriate data preprocessing, such as centering, scaling, and transformation. Centering and scaling operations are always associated with a specific mode, so they become more complex when data has multiple modes[81]. For a three-mode tensor, chords are an extension of columns in a matrix, whereas slices are all values associated with a specific position along one mode (Fig. 4.2a). For example, chord-wise and slice-wise operations across organ mode, respectively, correspond to one type of measurement across all organs, and all numbers aligned to one organ. Centering is performed to make the data ratio-scaled, as tensor decomposition models assume. This means that a zero value indicates a true zero effect, making multiplications meaningful (i.e., doubling the number always equals twice the effect). Scaling is used to adjust the scale differences among variables to avoid larger values overshadowing the variation of interest, which helps maintain numerical stability during solving. A common preprocessing choice is to mean-center across the subject/sample mode and then scale the standard deviation to one within other modes. Transformation is another technique that usually applies to nonlinear data to ensure the measurements are ratio-scaled before the decomposition.

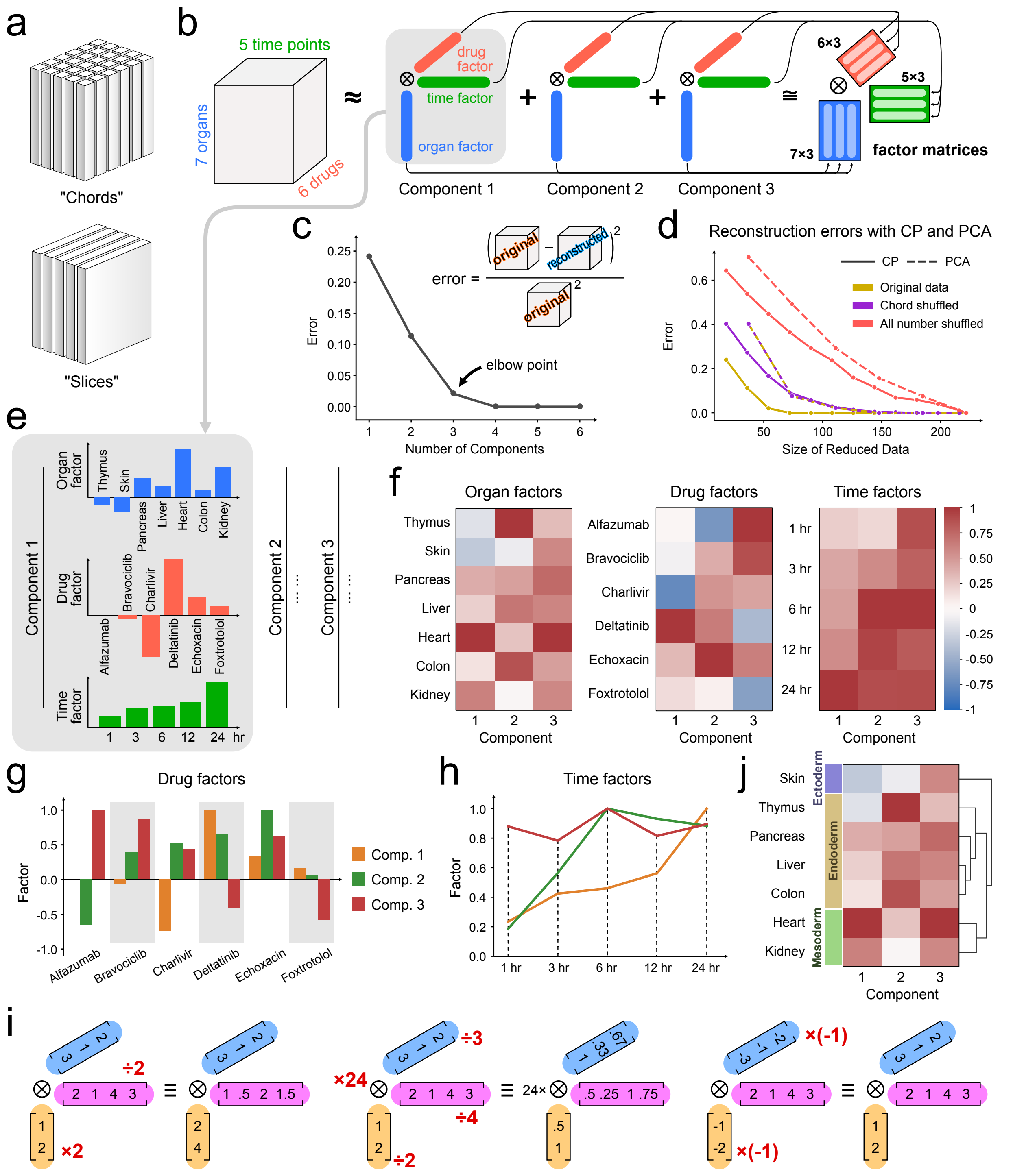


Figure 4.2. Fundamentals of Canonical Polyadic (CP) tensor decomposition

(a) For a three-dimensional array, a chord is the entries across all labels on one single mode, and slices are entries across two modes.

(b) CP decomposition approximates a complicated tensor as the sum of a few rank-one tensors. In the example here, for a drug response tensor of 7×6×5, organ by drug by time, after being decomposed into 3 components, we will have 3 factors for each of the three modes. Organizing them into matrices, we will have three factor matrices with shapes 7×3, 6×3, and 5×3 for organ, drug, and time, respectively.

(c) Plotting the number of components against the error. Error is defined as the sum of squared differences normalized by the sum of squares of the original tensor. An optimal component number may be attained at the elbow point on the plot (in this example, 3 components), or the point at which an acceptable error is reached.

(d) Sizes of reduced data plotted against their reconstruction errors using CP or PCA. CP decomposition represents the original dataset more concisely than PCA, whereas the chord-shuffled and all-number-shuffled versions reduce the advantage of CP, indicating the underlying data structure influences data compression.

(e) Plotting every factor separately to visualize tensor decomposition results. Here, three bar plots demonstrate the three mode-specific factors of Component 1. The factors of other components are omitted but can be shown similarly.

(f) Heatmaps to visualize the factor matrices compactly. We can both inspect all factors of a component across modes for its interpretation and/or compare factors within a mode to distinguish their differences.

(g) Factors of a discrete variable mode (such as drug mode here) can be visualized with a bar plot.

(h) Factors of a continuous variable mode (such as time point mode here) can be visualized with a line plot.

(i) Demonstration of factors' scale indeterminacy. Scaling the factors coordinately (left), factoring the weights to a separate scalar (middle), or negating factors in pairs (right) all yield equivalent factorizations, as they all reconstruct to the identical rank-one tensor.

(j) Organ factor heatmap reordered by hierarchical clustering on the factorization results. Other information, like the organs' biological grouping, can be labeled next to the heatmap to identify their association with the factors.

Programming Language	Package	Decomposition methods		Constraints Implemented
		Reviewed in this chapter	Other methods	
Python	TensorLy[82]	CP, Tucker, PARAFAC2, CMTF, CP partial least squares	Partial Tucker, Tensor Train, CP/Tucker regression	Nonnegativity, Symmetry, Regularization
MATLAB	Tensor Toolbox[83]	CP, Tucker		Symmetry, Sparsity, Orthogonality
R	rTensor[84]	CP, Tucker	3-mode tensor SVD, multilinear PCA	
	Multiway[85]	CP, Tucker, PARAFAC2	Simultaneous Component Analysis	Nonnegativity

Table 4.2. Selected tensor decomposition packages and the methods they implement.

CP: Canonical Polyadic Decomposition (also called PARAFAC or CANDECOMP);

CMTF: Coupled Matrix-Tensor Factorization;

SVD: Singular Value Decomposition;

PCA: Principal Component Analysis.

Tucker decomposition here can be higher-order SVD (HOSVD), truncated HOSVD, or higher-order orthogonal iteration.

Performing the decomposition

The decomposition method I review here is known as canonical polyadic (CP), parallel factors (PARAFAC), or canonical decomposition (CANDECOMP). Implementations of this method are available in software packages for various programming languages (Tbl. 4.2).

CP decomposition requires a data tensor and the desired number of components. The component number is the number of rank-one tensors used to approximate the original data (Fig. 4.2b). For each mode, the factors of each component can be regrouped into a factor matrix (Fig. 4.2b, right), in which the first columns of each matrix represent the first factor, the second columns the second factor, and so on. Thus, if we take the outer product of the first columns (Factor 1) in the three factor matrices, we will obtain the first decomposed rank-one tensor, Component 1. Repeating this process for each component and summing them up, we can reconstruct a tensor that approximates the original data (Fig. 4.2b). To summarize, the decomposed factors can be either grouped by mode into factor matrices, or by the factor indices into components. The goal of the decomposition algorithm is to make the reconstructed tensor match the original one as closely as possible.

The number of components

With CP decomposition, one must choose the number of components. Too few components will miss essential trends, while too many will lead to redundant factors, noise (overfitting), and poorer interpretability.

To quantify how well a decomposition with the chosen number of components fairly represents the original data, one can quantify the difference between the reconstructed tensor and the original data, or the reconstruction error. This value is calculated as the sum of squared differences between these two tensors, usually normalized by the sum of squares of the original data (Fig. 4.2c). Smaller errors indicate a better fit. While the error can range from 0 to any positive number, a successful fit should result in an error below 1 when normalized. The reconstruction error consistently decreases with a greater number of components, with diminishing returns where each additional component improves the fit to a lesser degree (Fig. 4.2c). Achieving a perfect fit to the data is typically not the goal of tensor decomposition. While this is technically feasible by setting the component number equal to the tensor's theoretical rank[6], in practice, this number is almost always too high for any practical use. To choose the optimal number of components, one may identify where the benefit of adding more components diminishes. This sometimes corresponds to the kink (or elbow) point on the error plot. However, such a transition point is not always evident.

The process described above resembles selecting component numbers in PCA but with a few distinctions. Tensor decomposition is not a recursive process: the components of a 3-component decomposition are not necessarily a subset of the 4-component decomposition. On that account, one must experiment with every candidate component number to identify an optimal choice. Components are also not guaranteed to be ordered[86]. Therefore, to create an error plot, the decomposition must be run for each number of candidate components (Fig. 4.2c).

The choice of component number directly relates to the data compression efficiency and fidelity trade-off (Fig. 4.2d). Since a tensor can be approximated with its factorization results which consist of fewer numbers, tensor factorization effectively compresses it. The smaller the size of the reduced data, the better the data compression ratio. However, this comes with the cost of a worse approximation (i.e. a larger reconstruction error) of the original data. For example, for a tensor with 7×6×5=210 values, a 4-component decomposition will compress it down to

4×(7+6+5)=72 numbers; if using 3 components, only 3×(7+6+5)=54 numbers, but with a greater reconstruction error.

The reconstruction error also depends on whether the underlying data structure can be well-approximated by lower ranks. As illustrated in Fig. 4.2d, CP decomposition can represent the drug response dataset more concisely than PCA—achieving a smaller representation under the same fidelity or comparable reduced sizes with lower error—since its underlying structure can be approximated well by the sum of multiple rank-one tensors. However, shuffling the chords in the data disrupts this low-rank tensor structure, causing the performance of CP to deteriorate. Further shuffling all numbers in the dataset eliminates the underlying low-rank matrix structure, thereby degrading even PCA's performance.

Another consideration is the inherent noise present in biological measurements. With too many components, tensor decomposition starts to fit trivial patterns which are more likely to be noise. In principle, we should cease adding more components when the algorithm begins to overfit (fit the original data too closely but lose generality), is prone to excessive local minima, or starts to violate the properties of CP. These situations may be assessed respectively through imputation tests (see "Missing data and imputation"), factor similarity tests (see "Optimization algorithms"), or core consistency diagnostics (see "Tucker decomposition").

Visualizing and interpreting the results

After validating the decomposition, the resulting factors can be inspected for biological insight. To provide a concrete example, we inspect our decomposition results shown in plots (Fig. 4.2e–h).

To visualize the results, one should design plots that describe how each factor is associated with the labels along each mode. Therefore, one can have one subplot for each factor (one from each mode) for each component, repeated for all components (Fig. 4.2e). In these plots, the x-axis indicates the labels, and the y-axis shows the factor weights. For a more concise visualization, one can also plot each factor matrix made from factors from all components as a heatmap with colors representing the weights (Fig. 4.2f). To visualize factors within a specific mode, bar plots and point plots generally work well for discrete labels such as samples, cell lines, or molecules (Fig. 4.2g), while line plots are more suitable for continuous labels such as time or concentration (Fig. 4.2h). Overall, visualization should optimally serve the presentation of the insights; there is no fixed rule.

The initial phase of interpretation involves delineating the meaning of each component pattern. This requires reading the plots across all modes. For instance, consider Component 1 (Fig. 4.2e). Within the organ factor, the largest signal originates from cells collected from the heart, followed by smaller weights from cells from the kidney. The same information can also be captured from the first column of the organ factor heatmap (Fig. 4.2f, left). Along the drug mode, the strongest signals appear on deltatinib in the positive direction and on charlivir in the negative. This can be read out from the heatmap (Fig. 4.2f, middle) or the drug factor bar plot (Fig. 4.2g) too. The time mode factor, on the other hand, has increasing values over time in Component 1. The orange line in the time factor plot best represents this trend (Fig. 4.2h). Putting this information together, one concludes that Component 1 mostly delineates a temporally increasing impact of deltatinib, positively, and charlivir, negatively, on cells from the heart (then the kidney). In practice, one can choose whichever plot best depicts the trend. Following the same logic, we see that Component 2 unveils an effect of mostly echoxacin, positively, and alfazumab, negatively, on cells collected from the thymus, peaking at 6 hours. Component 3 mostly indicates an effect of alfazumab and bravociclib on cells from mostly the heart that persists over time.

A specific mathematical intricacy, scale indeterminacy, can hinder clarity (Fig. 4.2i). As the effect along each mode is multiplied together, scaling these factors in an opposing way, i.e., doubling one factor and halving another within a component, yields equivalent results (Fig. 4.2i, left). This indicates that only the relative ratios of weights within a factor are certain, not the absolute values. Therefore, we should not compare the absolute weights between factors of different components, only the relative composition. To avoid ambiguity, one typically normalizes all factors to a defined scale, storing the weighting as a separate scalar (Fig. 4.2i, middle). The issue of indeterminacy extends to negative factors: by the same logic, negating two factors simultaneously also yields equivalent results (Fig. 4.2i, right). This is sometimes called sign indeterminacy[87]. One approach to avoid ambiguity is to make most modes positive by negating the factor vectors in pairs, ensuring that at most only one mode harbors factors with an overall negative effect (Fig. 4.2i, right).

One can also compare across components within a single mode. Within the organ mode, for instance, Components 1 and 3 assign similar factors to cells from the heart and kidney, unveiling shared localization in drug effect (Fig. 4.2f, left). In the drug mode, each drug has different factors, suggesting that they have divergent interaction profiles (Fig. 4.2f, middle, 2g). Each time factor also has a distinct trend, ranging from stable (Component 3) to increasing over time (Component 1) and peaking (Component 2) (Fig. 4.2h). To better identify similar entries (e.g. drugs or organs) on a mode discovered by tensor decomposition, one can also perform hierarchical clustering on the factor matrix and reorder the entries accordingly (Fig. 4.2j). This juxtaposes entries of similar factor weights, helping to reveal groupings of comparable entries. Additional known information, such as cell categories, sample classes, and patient statuses, can also be labeled next to the heatmap to help identify associations between the factors and their known groupings (Fig. 4.2j).

4.3 Details and considerations of tensor decomposition

The previous section presented an overview of employing tensor decomposition. However, several details of the procedure may help in certain circumstances.

Optimization algorithms

Solving tensor decomposition is, in its essence, an optimization problem. The objective is to find a set of factor matrices that, when multiplied, render a reconstructed tensor with minimal error (Fig. 4.3a). Common mathematical optimization algorithms, such as gradient descent or the Newton-Raphson method, can be employed here[88]. This "direct optimization" approach offers the advantage of versatility, since many optimization methods allow additional constraints, making it possible to develop new decomposition schemes. However, its performance relies heavily on the chosen method and initialization values, since a substantial number of parameters must be simultaneously solved.

As an alternative approach, we can first notice that the factor matrices exhibit symmetry: swapping mode orders does not change the solving. Also, if we know the correct factors of all other modes, solving for one mode can be converted into an ordinary least squares problem. Thus, we can tackle one mode at a time using least squares while treating the others as constant, then repeat this for every mode (Fig. 4.3b). We keep iterating until these factors converge. Over time, we can expect a monotonic decrease in the reconstruction error. This approach is called alternating least squares (ALS). Besides its efficiency, ALS often benefits from more stable and reproducible performance[89].

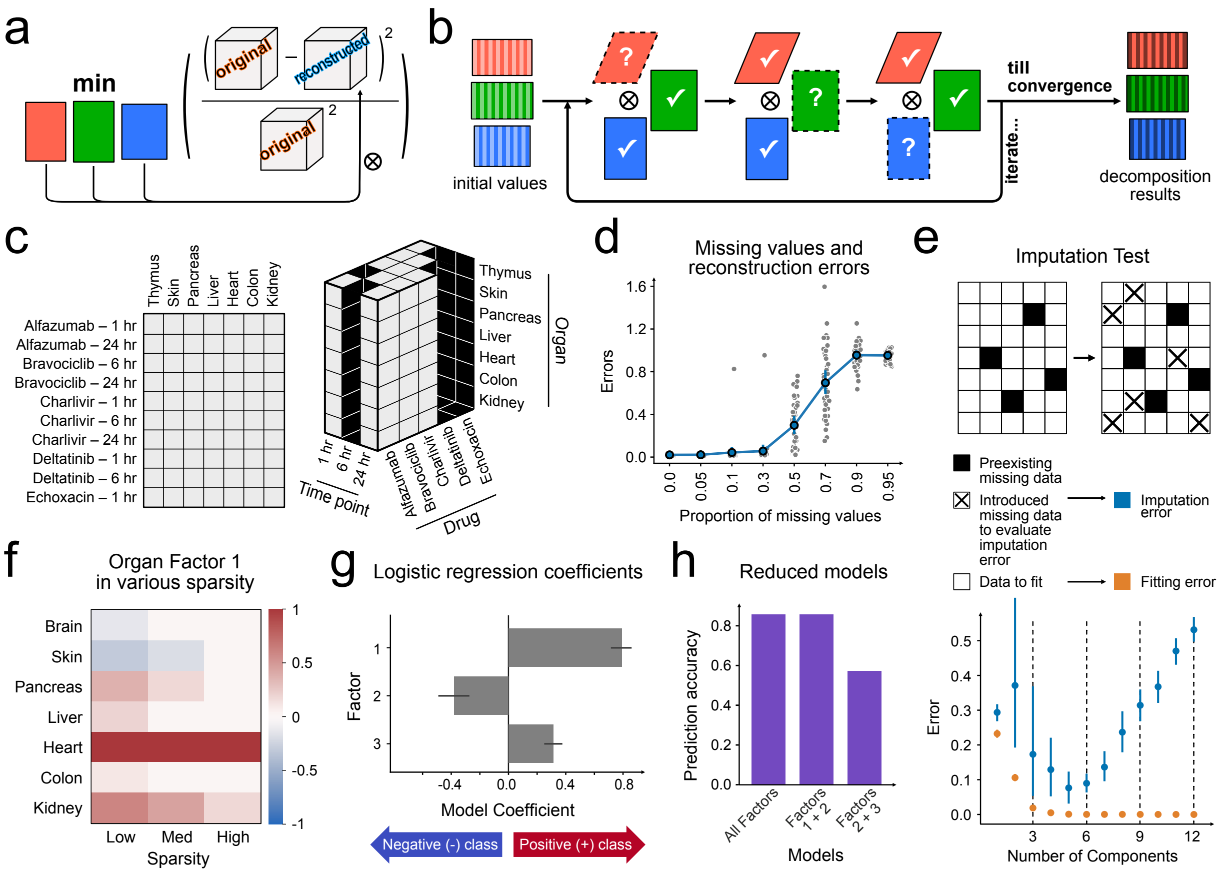


Figure 4.3. Technical details on applying CP to biological data

(a) Solving tensor decomposition is an optimization problem aiming to minimize the reconstruction error by adjusting the numbers in the factor matrices.

(b) Alternating least squares (ALS) is another strategy besides direct optimization. Starting from a set of initial values, it optimizes one factor matrix at a time with linear least squares while holding the others constant. This process is repeated on each factor matrix until convergence is reached.

(c) A demonstration of how structuring data into tensor format may create missing values. Although the original table on the left does not contain any missing values, since not all drug-time pairs are measured, the reformatted three-mode tensor contains missing chords.

(d) Various proportions of missing data were introduced to the tensor to evaluate how missingness impacts reconstruction errors. Each gray point represents one of 40 runs with random missing patterns. The blue points and error bars show the average reconstruction errors and 95% confidence intervals, respectively.

(e) Demonstration of the imputation test. Ignoring the preexisting missing data, we arbitrarily introduce more missing positions, use the remaining data to fit the decomposition, and then compare the reconstructed (i.e. imputed) values with the original values at the positions we removed. Plotting against the number of components, the fitting errors should decrease monotonically with more components. However, the imputation error will eventually increase with excessive components due to overfitting.

(f) Sparsity in tensor factors. These organ Factors 1 are in the order of increasing sparsity.

(g) Tensor decomposition factors can be used for response prediction when combined with regression. The coefficient of each factor indicates their association with the sample classes.

(h) For classification, the model may be reduced to using a subset of the factors.

Both methods require initial factor values. While a random initialization may be sufficient, a more informed estimation can expedite convergence. One such estimation involves using the principal components from a flattened version of the original tensor. This approach, known as SVD (singular value decomposition) initialization, usually yields more stable results and reduces the likelihood of a suboptimal solution (i.e. local minimum). However, neither initialization guarantees the best solution.

When the resulting factors are highly dependent on the starting point of the fitting, it can indicate that the optimization problem is ill-formed, suggesting that the chosen number of components is too large or that additional constraints would be helpful. The factor similarity test exploits this property to determine the appropriate component number[75,90]. In essence, this test quantifies to what extent different starting points change the resulting factors, helping determine up to how many components the factorization algorithm remains stable.

Missing data and imputation

Missing measurements frequently arise from experimental limitations. These omissions are not necessarily a result of oversight; certain measurements may be intentionally missing. This issue becomes particularly pronounced with tensors, as complete tensors require all possible combinations of all modes. Consequently, missing data can emerge simply from transforming a dataset into a tensor, even if the original data appears complete (Fig. 4.3c). For instance, the example dataset in Fig. 4.3c does not contain any missing values, but because the impacts of deltatinib and echoxacin after 6 hours were not measured, the reformatted tensor contains missing chords (Fig. 4.3c, right).

Tensor decomposition can be performed even with missing values in a tensor. This can be achieved either through ignoring the missing positions and only fitting the existing ones in direct optimization, or prefilling them with placeholders in the hope of updating these values iteratively through repeated factorization and reconstruction, or employing some form of censoring in ALS[91,92]. Note that zeros in a tensor will still be fit by the tensor decomposition algorithm, unlike explicitly missing values, so replacing missing values with zeros is incorrect. While some optimization methods can function even with a high proportion of missing values, the resulting factors can significantly deviate from those obtained with complete data. The extent of this deviation can vary widely depending on the underlying data structure and the specific missingness, but generally, a greater portion of missing data leads to larger reconstruction errors (Fig. 4.3d).

Tensor decomposition also provides an avenue to impute the missing values of a tensor. Since a full tensor can be reconstructed from the resulting decomposed factors (Fig. 4.1g, 2b), one can use these reconstructed values from tensor decomposition to replace the missing positions, effectively imputing them[92]. Compared with matrices, higher-order tensors benefit from the additional information from more shared coordinates. Tensor imputation through decomposition is not foolproof; it remains an area of ongoing research. Like matrix completion, it relies on inherent assumptions. If the original data cannot be approximated as lower-rank tensors (Fig. 4.2b), the imputed values can significantly deviate from their true values. Other factors, such as the quantity and distribution of the missing values and the chosen component numbers and decomposition method, can also influence the accuracy of imputation. A tensor cannot be missing all its values across a slice. Thus, in situations where there are very few non-missing values, it may be advantageous to consider discarding one position along a mode.

One can use imputative performance to assess the reliability of decomposition on a tensor or to determine its appropriate number of components. In an imputation test, one intentionally introduces additional missing values in the data (Fig. 4.3e, top). Following decomposition, the entire tensor is reconstructed from the factors, and the left-out values are compared against their reconstructed versions. A substantial disparity indicated by a high imputation error indicates an unsuccessful decomposition, attributable to either an ill-suited dataset or an excessively high number of components. While the fitting error monotonically decreases with more components, the imputation error often shows an optimum at an intermediate number of components (Fig. 4.3e, bottom).

Constraints on the factors

The optimization processes reviewed so far only aim to best fit the data. However, their results may suffer from low interpretability, overfitting, and instability. Numerical constraints on the factors can help with these issues. Although they may impact the goodness of fit, reasonable constraints can enhance the model's ability to reveal meaningful patterns, leading to more insightful discoveries. For example, one goal of constraints is to achieve sparsity, where a factor has nonzero values in only a few positions and renders others nearly or exactly zeros. This helps establish direct associations between factors and their effects[66]. For instance, in the hypothetical organ Factors 1 in Fig. 4.3f, the low-sparsity factor has weights on almost all organs, making interpretation more complicated. The high-sparsity version only has weights on the heart and the kidney, better indicating that this factor has the greatest association with these two organs. Regularization is commonly used to achieve sparsity in factors.

Nonnegativity is the most commonly used tensor decomposition constraint[93]. It aligns intuitively with the expectation that certain quantities in biology are inherently nonnegative: a cell cannot secrete a negative number of molecules, and a gene cannot be expressed at a negative level. Nonetheless, enforcing nonnegative factors may limit the tensor factors from modeling negative effects in biology, such as an upstream pathway that suppresses molecule secretion or inhibits gene expression. Another rationale for the nonnegativity constraint is to foster sparsity within the factors and avoid overfitting. Decompositions allowing negative factor values can yield degenerate components, where one component is strongly positive and another is strongly negative, mostly canceling each other out[86]. Enforcing all values in the factor matrices to be nonnegative obviates such occurrences, as the impact of any component cannot be counteracted by another. Nonnegative factorization often leads to minimal sacrifices in model error, solidifying its application in practice[75].

Constraints can also be used to enforce biological knowledge in a decomposition[94]. For instance, in neuroscience, one may postulate minimal crosstalk among different brain regions and limit the brain region factors to be a diagonal matrix[95]. In molecular biology, one may employ orthogonalization of the factors to enforce a clean delineation between components and traits[96]. This usually lacks a standardized approach, as biological contexts vary, and may require customized solving[90].

Subsequent analysis

While tensor analysis often serves as an important step for distilling data into significant patterns, further analysis beyond the factor plots (Fig. 4.2e–h) is often required to learn what component patterns indicate about biology. The factor matrices serve as efficient summaries for individual patterns linked to their respective modes. Consequently, each matrix can be isolated for a detailed analysis of the variation within a specific mode of interest. For instance, the components associated with genes or molecules of particular interest from prior knowledge can be further examined to validate their agreement with known mechanisms.

The decomposed factors can also be used as reduced data to predict responses or sample classes when combined with regression. The scale and sign of the weights for each factor indicate its effect on the regressed quantity (Fig. 4.3g). If only a subset of factors contributes to the effect of interest or the regression model can achieve comparable accuracy with fewer factors, the prediction model may use only a subset of them (Fig. 4.3h). For example, in Fig. 4.3h, prediction using only two factors, Factors 1 and 2, performs just as effectively as all factors.

4.4 Advanced tensor methods beyond CP

In this section, I cover more advanced tensor decomposition methods. For more complex biological data, it is particularly crucial to choose a method that best reflects the structure of the expected patterns.

Tucker decomposition: allowing all factors to interact

In CP decomposition, especially when there are more components, some factors may start to look similar within one mode. For example, in Fig. 4.2f, the organ Factors 1 and 3 appear similar. This redundancy arises from the inherent constraint of CP decomposition, where factors may only interact within the same component (Fig. 4.2b). In other words, because CP does not allow interaction between drug Factor 1 and time Factor 3, a repetitive drug factor must be present in Component 3 to capture a similar effect on organs. CP permits the existence of two identical factors in one mode, as long as their corresponding factors in other modes remain distinct. Therefore, the factors along a mode in CP decomposition may not succinctly summarize the trends in this mode.

Tucker decomposition is a different tensor decomposition model from CP with a more flexible construct[97,98]. It permits varying numbers of factors for each mode, and all factors across modes interact. For example, here we perform a (4,3,2)-rank Tucker decomposition on the 7×6×5 drug response data tensor, in which the organ mode has 4 distinct factors, the drug mode 3, and the time mode 2 (Fig. 4.4a). Consequently, there are 4×3×2=24 factor interactions. Each interaction can be understood as a component in CP (Fig. 4.4a, right). The magnitude of each interaction is characterized by its corresponding weight, and these 24 weights can be arranged into a 4×3×2 core tensor (Fig. 4.4b, left). The outcomes of a Tucker decomposition include a core tensor that models the factor interactions and three-factor matrices that represent the principal components (the major trends) along those three modes (Fig. 4.4a, middle)[6].

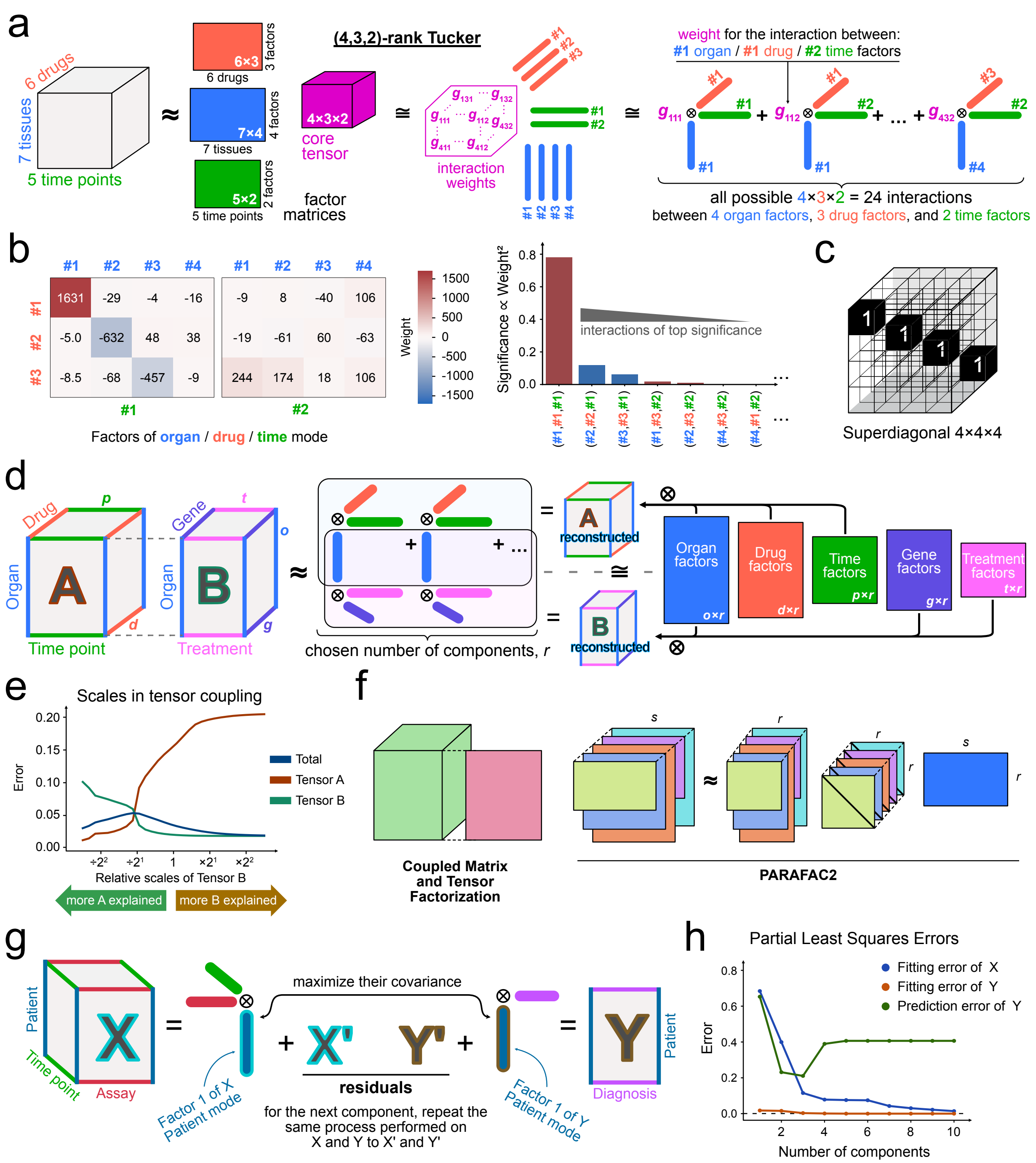


Figure 4.4. Tensor methods beyond CP: Tucker decomposition, coupling, and partial least squares

(a) Schematic of Tucker decomposition. This (4,3,2)-rank Tucker decomposition on the previous 7×6×5 drug response tensor allows all distinct 4 organ factors, 3 drug factors, and 2 time factors to interact. The weights of these 24 interactions are organized into a 4×3×2 core tensor. The results of a Tucker decomposition are a factor matrix for each mode and this core tensor.

(b) The core tensor of a Tucker decomposition. It can be visualized by showing the numbers in each slice. The significance of an interaction is proportional to its weight squared.

(c) A superdiagonal 4×4×4 tensor. CP is a special case of Tucker decomposition where the core tensor is superdiagonal.

(d) Schematic of coupled tensor decomposition. Here, two three-mode tensors, A and B, are coupled on the organ mode. Therefore, the organ factors are shared, while the drug and time factors are private to Tensor A, and gene and treatment factors Tensor B. The dimensionalities of them are indicated by the lowercase letters.

(e) The scaling issue in coupled tensor decomposition. When one of the coupled tensors has values with greater total variance, the factorization explains more variance in it if without proper scaling, leading to an uneven representation of the two datasets.

(f) Some other examples of tensor coupling: coupled matrix and tensor factorization (CMTF, left) and PARAFAC2 (right).

(g) Schematics of tensor partial least squares. Partial least squares is performed on two tensors, X and Y, with one aligned mode. During solving, the two separated X and Y factors of the aligned mode (patient mode in the example case) yield the maximal correlations. Partial least squares components are solved sequentially, as the next component is found by repeating the same process on the residuals, X' and Y', from the last round. Therefore, the components are in decreasing order of covariance explained.

(h) The performance of partial least squares can be evaluated by calculating the fitting errors of X and Y and the prediction errors of Y. Both fitting errors should decrease with more components, while the prediction errors (from cross-validation) of Y should initially decrease but eventually increase due to overfitting.

There are two ways to utilize the Tucker results. The factor matrices, consisting of the eigenvectors defined in higher dimensions, can be used as summaries of the modes. They can be visualized similarly to Fig. 4.2f. One can also analyze the core tensor to identify the top interactions by significance, which is defined as proportional to their weights squared (Fig. 4.4b, right). For example, here, ~78% of the variance can be explained by the interactions among Factor 1s of organ, drug, and time.

CP decomposition is equivalent to a specific instance of the Tucker decomposition, wherein factors associated with different components are non-interacting, thus the core tensor assumes a superdiagonal form, signifying that all off-diagonal positions are zeros (Fig. 4.4c). This superdiagonal property has been harnessed to test whether a CP decomposition is correctly implemented, known as core consistency diagnostics[99]. Specifically, after acquiring the CP factor matrices, if adding off-superdiagonal interactions to the retrofitted core tensor can improve the fitting considerably, the number of components may be inappropriate, or Tucker may be a better model than CP for this dataset.

Tucker decomposition offers a better mode-specific summary and more flexible analysis, which opens many possibilities for method development[100]. Many variants of Tucker decomposition, including higher-order SVD (HOSVD), have been applied to biological datasets[65,101].

Coupling: sharing factors across multiple tensors

The integration of (epi-)genomic, transcriptomic, and proteomic data, either in bulk or at the single-cell resolution, has provided opportunities for an integrated understanding of cellular processes. More broadly, biologists often encounter data fusion challenges when attempting to identify shared patterns among multiple data sources[102]. The joint analysis of several datasets can be formulated as coupling of tensors[103].

Coupling arises when two or more datasets are collected with differing dimensions, but all tensors share at least one “coupled” mode (Fig. 4.4d, left). Commonly coupled modes include samples or patients that are shared across multiple assays. For instance, there may be another dataset on cells from the same groups of organs measured in the previous dataset (Fig. 4.4d, Tensor A); this new dataset contains the gene expression of the cells from these organs under various treatments (Fig. 4.4d, Tensor B). In this case, the organ mode is shared, while each tensor has other uncoupled modes, such as drugs and genes. In a coupled decomposition, a shared mode will have a common factor matrix that is used by all tensors that comprise this mode (Fig, 4d, right). In this way, this factor matrix succinctly reflects the trends across these two coupled tensors.

Visualizing and interpreting the results of a coupled tensor factorization operate like with CP (Fig. 4.2f). Each tensor is decomposed into a series of rank-one components, and any coupled mode will have a single set of factors shared among all the tensors using it (Fig, 4d, middle). All other modes will still have their own factor matrices (Fig, 4d, right). In addition to examining components within a tensor, one can also compare the uncoupled private modes between two tensors to assess their associations. A unique advantage of coupling arises from missing data. If a certain tensor has missing entries, other tensors can share information through the coupled factors to improve imputation.

Coupling introduces a new issue. Because factorization minimizes the overall reconstruction error, the relative scaling among coupled tensors influences the priority in explaining patterns from each dataset. As the total variance of values can differ significantly across datasets collected from various assays, the decomposed factors can be dominated by one source if the data is not appropriately scaled. Typically, a range of scaling should be explored, and the overall and tensor-specific errors evaluated (Fig. 4.4e). If the factor matrices are used to predict some outcomes, the prediction accuracy can also be used to compare various scalings and determine an optimal scaling.

Overall, coupling offers remarkable flexibility for data integration. Although we refer to the methods as coupled tensor factorization, matrices (2-way tensors) are also included. For example, many applications have used coupled matrix and tensor factorization to jointly analyze a tensor and a matrix (Fig. 4.4f, left)[28,104]. Coupling also expands the applicability of tensor methods to more irregularly shaped data, as illustrated by PARAFAC2[105]. PARAFAC2 is a method that decomposes a series of matrices, where one mode is shared while another is unaligned and variable in size (Fig. 4.4f, right). This forms a ragged tensor to which CP or Tucker cannot be applied. PARAFAC2 projects the variable modes into a latent, uniform shared mode, identifying patterns not only on the shared mode but also across these matrices, effectively harnessing the benefits of coupling. Tensor coupling is an active field of method development, including combining it with other decomposition strategies (such as Tucker or partial least squares).

Partial least squares: informing decomposition by effects

Many scientific questions involve identifying how a series of measurements associate with a specific phenotype or outcome of interest. For example, one might associate patients' blood panel tensor with their diagnosis. In statistical terms, we have explanatory variables (X) and outcomes (Y), and our goal is to reveal only the patterns in X that uniquely associate with Y. This approach differs from simply coupling them where the joint variance of both X and Y is considered. Instead, the objective is to only capture the trends in X when they exhibit correlation with Y.

As mentioned previously, tensor decomposition factors can be combined with linear regression models. This two-step approach bears a resemblance to principal component regression: first, the data is decomposed using tensor decomposition without considering the effects (Y); then, regression is applied to capture correlations between the decomposed factors and their effects. However, as the first step is performed without the knowledge of Y, the decomposed X factors are not guaranteed to associate with Y. To address these challenges, partial least squares (PLS) methods have been developed, in both classification form (PLS discriminant analysis) and regression form (PLS regression)[106].

Tensor PLS is designed to uncover relationships between two tensors, X and Y, for predictors and responses, wherein one mode is aligned (Fig. 4.4g). For instance, consider tensor X representing medical tests on a group of patients over time, while matrix Y (a two-way tensor) records their diagnosis. The result of tensor PLS is analogous to performing two separate CPs on both X and Y simultaneously with the same number of components. After decomposition, they will each have a distinct patient factor matrix. However, PLS decomposes both datasets with the goal of maximizing the correlations between these two patient factors (Fig. 4.4g). The factors of the other non-aligned modes in X and Y come after obtaining the patient factors, and are defined to maximally capture variance within each dataset[107]. While the intricacies of the solving algorithm extend beyond the scope of this review, one helpful property to note is that tensor PLS is solved component-by-component. Each additional component is solved upon the residuals of X and Y (X’ and Y’) which are the original tensors subtracted by the solved components (Fig. 4.4g), meaning that components are ordered by the covariance they explain. Therefore, in a correctly performed tensor PLS, the fitting errors of both X and Y should decrease monotonically as more components are added (Fig. 4.4h). However, as a supervised learning method, tensor PLS does not always predict unseen samples better with more components due to the risk of overfitting. The optimal number of components can be determined through cross-validation, where a portion of the samples is left out during fitting to test the model's performance on them. It is expected that the prediction error of Y in cross-validation would initially decrease if the optimal number of components is greater than one (which is usually the case if Y is a matrix rather than a vector) and then increase after reaching the optimum (Fig. 4.4h).

Overall, PLS has unique advantages when focused on a particular response. Since it is designed to specifically discover those patterns associated with a prediction of interest, PLS can predict the effect with fewer components compared with CP. Tensor PLS can be combined with Tucker decomposition and coupling in explanatory (X) tensors, and techniques are available to handle missing values[108].

4.5 Biological insights from tensor-based methods

Tensor decompositions have applications in virtually all fields of biological data analysis. In this section, I summarize several notable examples.

Applications in bioinformatics

In bioinformatic studies, multi-omics data may contain tens of thousands of genes and millions of genomic positions. Tensor methods can simplify these large datasets generated by high-throughput techniques into a succinct set of components and do so more efficiently than matrix-based counterparts. These reduced latent structures group genes based on their common patterns revealed by the data, easing the scale of effect prediction.

Hore et al. illustrated how tensor methods can be applied to condense genes in RNA-seq data across multiple tissues into associated factors to reduce the scale of statistical testing and to strengthen their statistical power[66]. To reveal gene networks, they structured the gene expression levels into a gene by individual by tissue tensor. After applying the tensor method, the data was reduced into around two hundred components, a great reduction from the tens of thousands of genes they originally dealt with. These components grouped the genes by activities and indicated in what tissues they were active. Using individual scores as genotypes for genome-wide scanning on SNPs, they discovered the components that were significantly associated with *trans*-expression quantitative trait loci (eQTLs) and revealed their specific pathway or epigenomic regulation.

Using tensor factors to cluster genes in transcriptome is further exemplified by Wang et al[87]. With the increasing scale of multi-tissue datasets, classical clustering methods struggle to extract information from multi-way interactions in the transcriptome. To fully extract the three-way interactions between individuals, genes, and tissues, they applied constrained CP to RNA-seq and microarray measurements. Besides being able to run on three-dimensional data where traditional methods failed to reveal true patterns in simulated data, this tensor-based clustering method was shown to better test for differentially expressed genes with improved statistical power compared with single-tissue tests.

Durham et al.[72], on the other hand, applied tensor methods to large epigenome projects such as Encyclopedia of DNA Elements (ENCODE)[73] and the Roadmap Epigenomics Project[109]. In these massive datasets, many cell type and assay pairs were not measured due to time and funding constraints. Therefore, the imputation of these data has been extensively studied[110]. Organizing the ENCODE data into a three-mode tensor, they found that tensor-based imputation outperformed alternative approaches, demonstrating that structuring the data in tensor form helps model and explain variation across the data.

Other tensor methods have been applied to epigenomic data too. For example, a variant of Tucker decomposition has been applied to model spatial association within topologically associating domains[111]. The decomposed factors directly link epigenomic state and chromosomal topology. Tensor decomposition can be also combined with machine learning methods. For example, extending the work of Durham et al., the same group inputted the concatenated tensor factors from three different genomic resolutions into a feed-forward deep neural network to predict the epigenomic signals, allowing a multi-scale view of the genome[112].

Applications in neuroscience

Neuroscience is among the earliest fields to employ tensor methods[113,114]. As electroencephalography and functional magnetic resonance imaging data are collected over time, any experiment involving more than one electrode and trial is guaranteed to be at least three-dimensional. Conventionally, the data has been converted into matrices by averaging multiple trials, inevitably losing information about trial-to-trial variation. Therefore, tensor methods, including both CP and Tucker decomposition, have been attractive to the neural signal processing community[115].

Williams et al. presented a clean framework for applying tensor component analysis on large-scale neural data across time and trials[75]. Before running on the actual data, they demonstrated that tensor decomposition works well on simulated linear model neural networks and nonlinear recurrent neural networks, separating positive and negative cells with almost perfect accuracy. With the same simulations, PCA and independent component analysis failed to recover the right signal. They then applied the method to their experiments on mice's prefrontal activity and primate motor cortex. Nonnegative tensor decomposition was shown to cleanly separate neurons that were activated in various periods and associated with specific movements.

Applications in systems biology

Systems biology makes repeated measurements over different times, tissues, or spatial structures, so the data are naturally in tensor structure. These measurements may include sequencing, flow cytometry, or quantitative cell imaging, requiring solutions for data integration. Two specific concerns here are avoiding overfitting, as the datasets are often limited in size, and incorporating heterogeneous information. Therefore, nonnegative decomposition, imputation tests, coupling, and partial least squares have been used.

Tensor methods offer unique advantages for the study of systems biology by enabling concurrent comparison of multiple contexts and extracting their shared trends. For instance, Armingol et al. employed tensor decomposition to study cell-to-cell communication from RNA-seq data[116]. Contrary to many previous studies that cannot handle more than two cellular contexts simultaneously, by embedding communication matrices[117] into a four-mode tensor, they were able to characterize variation in cell-to-cell communication across several contexts coordinately.

The benefit of tensor decomposition in analyzing repeated measurements simultaneously can also be extended to compositional data in microbiology. Microbiome studies often take multiple samples from the same individual either longitudinally or spatially, but there is a lack of methods to account for both biological change and interindividual variability in them. Martino et al. took the tensor approach to deconvolute gut microbial sequencing data[118]. They demonstrated that unsupervised tensor decomposition can identify differentially abundant microbes, accounting for the high-dimensional, sparse, and compositional nature of microbiome data.

Tensor partial least squares can also be helpful in systems biology[119]. Netterfield et al. recently applied it in a study of DNA damage response[120]. They systematically profiled a human cell line with the treatment of DNA double-strand break-inducing drugs over time and concentrations, using tensor PLS to directly associate signaling to response, both as three-mode tensors, separating the time mode from drug concentrations. This allowed them to identify signals with time-dependent correlations with senescence and apoptosis. They also observed that tensor PLS required fewer parameters to predict the response than the conventional unfolded version.

4.7 Conclusion

In this chapter, I review the application of tensor decomposition to biological data analysis. The paramount lesson of this chapter is the profound influence of the chosen data representation, the "medium," on our comprehension of the data itself and the analytical approach. The selection of data representation should be driven by the natural structure of the underlying data and experiment rather than mere mathematical expediency. Approaching this analysis appropriately improves on the insights one can derive from the data through better accuracy, more evident interpretation, and an enhanced ability to integrate data across studies and scales. While tensor methods have gained increased prominence, they have much broader potential yet[121]. Part of the field's maturation will arise from a broader appreciation and understanding of these techniques.

Nevertheless, tensor decomposition, in its current form, is not without limitations. First, it is still fundamentally linear, so it may fail on datasets of nonlinearity characteristics. This does not forbid it from being an adequate baseline model though. Furthermore, the existing solving algorithms continue to grapple with numerical issues such as nonuniqueness in factors, instability when addressing missing data values, and challenges in hyperparameter tuning. These issues will be resolved by new theories and a broader appreciation of these techniques.

Chapter 5

Tensor-structured decomposition improves systems serology analysis

There is a crack, a crack in everything,
That's how the light gets in.
Leonard Cohen

5.1 Introduction

Whether during a natural infection, therapeutic vaccination, or an exogenously administered antibody therapy, antibody-mediated protection is a central component of the immune system. The unique property of antibodies is conceptually simple—they undergo affinity enrichment toward specific antigens—but the mechanisms of resulting protection are mediated through a network of interactions[122]. Therapies are often optimized based upon the titer or neutralizing capacity of the antibodies they deliver. However, many of the mechanisms for antibody-mediated protection occur through secondary interactions with the immune system via an antibody's fragment-crystallizable (Fc) region. While more challenging to quantify and identify as the mechanism of protective immunity, these immune system responses, such as antibody-dependent cellular cytotoxicity (ADCC)[123,124], complement deposition (ADCD)[125], cellular phagocytosis (ADCP)[126], and respiratory burst[127] are known to be just as or more important in many diseases.

A suite of recent technologies promises to broaden our view of antibody-mediated protection as the microarray did for gene expression. Systems serology aims to broadly profile the humoral immune response by jointly quantifying both the antigen-binding and Fc biophysical properties of antibodies in parallel[128]. In these assays, antibodies are first separated based on their binding to a panel of disease-relevant antigens[129,130]. Next, the binding of the immobilized antibodies to a panel of immune receptors is quantified. Other molecular properties of the disease-specific antibody fraction that affect immune engagement, such as glycosylation, may be quantified in parallel in an antigen-specific or -generic manner[129–131]. By accounting for the two necessary events for effector response—antigen binding and immune receptor engagement—these measurements have proven to be highly predictive of effector cell-elicited responses and overall antibody-elicited immune protection[5,132,133].

Although systems serology provides a major advancement in our ability to analyze the antibody-elicited immune response, analysis of these data is often challenging. Standard machine learning methods, such as regularized regression, principal components analysis (PCA), and partial least squares regression (PLSR) have been effective in identifying highly predictive immune correlates of protection[132,134]. However, identifying specific molecular changes or programs that give rise to protection is more difficult. First, because many of the measurements are overlapping in the molecules they quantify, or measure co-dependent processes, much of the data is highly inter-correlated[5,135]. Particularly when analyzing polyclonal antibody responses such as those which arise in vaccination or natural infection, protection may arise through single or combinations of molecular species and features within the antibody response, through either individual or combinations of antigens[136,137]. One successful approach in serology analysis has been to collapse molecular features into summary statistics, such as Fc breadth or polyfunctionality, though this requires pre-defined descriptors of these quantities[136]. Alternatively, patterns of interest can be experimentally derived, such as with blocking experiments, but this is labor-intensive and requires pre-existing monoclonal antibodies to define each pattern[55]. Unsupervised approaches that explicitly integrate patterns across both antigen and Fc properties will help to mechanistically characterize immune protection.

While systems serology measurements include a variety of different assays to quantify humoral response, a common overall structure exists to the data. Most of the measurements quantify the extent to which an antibody bridges all pairs of target antigen and receptor panels, across a set of individuals[128]. Binding to target antigen involves the antigen binding fragment (Fab) of an antibody, while immune receptor interactions occur through its crystallizable fragment (Fc) region. Thus, it is natural to split them up as they entail different regulatory processes. Along with the dimension of individuals, these measurements, therefore, can be thought of as a three-dimensional dataset, where every number in this "cube" of data represents a single measurement (Fig. 5.1a/b). Then, separately from these measurements, some properties of the humoral response, such as antibody glycosylation, may be assessed but without separation across different antigens[131,138]. With data of three or more dimensions, tensor decompositions, a family of unsupervised dimensionality reduction methods for higher-order tensors, provide a generalization of matrix decomposition techniques[6]. These methods are especially effective at data reduction when measurements have meaningful multi-dimensional features, such as time-course measurements[118]. Like PCA, tensor decomposition methods, when appropriately matched to the structure of data, help to visualize its variation, reduce noise, impute missing values, and reduce dimensionality[101].

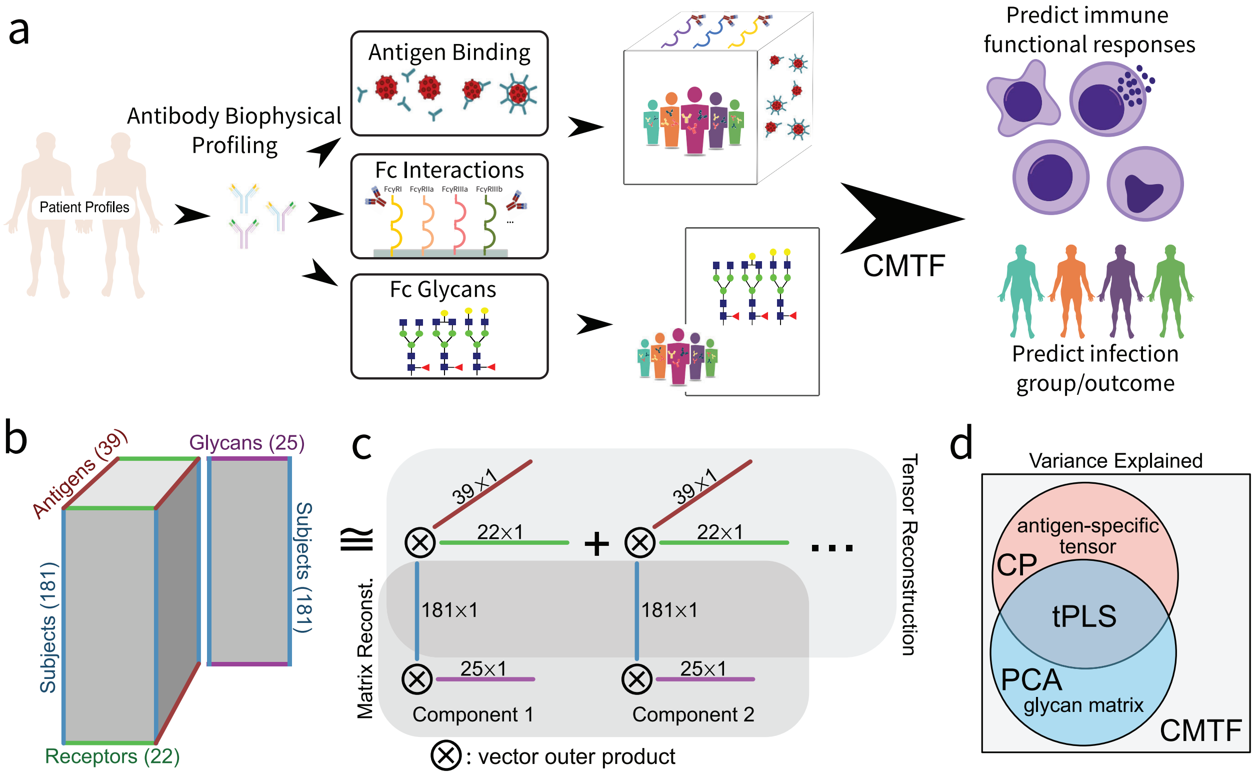


Figure 5.1. Systems serology measurements have a consistent multimodal structure.

(a) General description of the data. Antibodies are first separated based on their binding to a panel of disease-relevant antigens. Next, the binding of those immobilized antibodies to a panel of immune receptors is quantified. Other molecular properties of the disease-specific antibody fraction that affect immune engagement, such as glycosylation, may be quantified in parallel in an antigen-specific or -generic manner. These measurements have been shown to predict both disease status (see methods) and immune functional properties—ADCD, ADCC, antibody-dependent neutrophil phagocytosis (ADNP), and natural killer cell activation measured by IFNγ, CD107a, and MIP1β expression.

(b) Overall structure of the data under the CMTF framework. Antigen-specific measurements can be arranged in a three-dimensional tensor wherein one dimension each indicates subject, antigen, and receptor. In parallel, non-antigen-resolved measurements such as quantification of glycan composition can be arranged in a matrix with each subject along one dimension, and each glycan feature along the other. Although the tensor and matrix differ in their dimensionality, they share a common subject dimension.

(c) The data are reduced by identifying additively separable components represented by the outer product of vectors along each dimension. The subject dimension is shared across both the tensor and matrix reconstruction.

(d) Venn diagram of the variance explained by each factorization method. Canonical polyadic (CP) decomposition can explain the variation present within the tensor on its own[101], analogous to principal component analysis (PCA) on the glycan matrix. Tensor partial least squares regression (tPLS) allows one to explain the shared covariation between the matrix and tensor[139,140] . In contrast, here, we wish to explain the total variation across both the tensor and matrix[141]. This is accomplished with CMTF (see Materials and Methods).

As the structure of systems serology data is well-suited to tensor decomposition, we take advantage of this to reap the above benefits. As examples, we analyze two separate studies wherein systems serology measurements were shown to predict both functional immune responses and disease status within HIV- and SARS-CoV-2-infected subjects[132,133]. We first adapt a tensor decomposition approach—coupled matrix-tensor factorization (CMTF)—to reduce these measurements into consistent patterns across subjects, immunologic features, and antigen targets. Inspecting these factors reveals interpretable patterns in the humoral response, and these patterns' abundance across subjects predicts subjects' functional immune responses and infection state. Importantly, CMTF greatly improves the interpretability of these predictions compared with methods that do not recognize the structure of these data. This approach, therefore, provides a very general data-driven strategy for improving systems serology analysis.

5.2 Systems serology measurements can be arranged in tensor form for greater dimensionality reduction

We first sought to determine whether the structure of systems serology measurements could inform better data reduction strategies (Fig. 5.1). As an array-based measurement, wherein the amount of signal is dependent upon the quantity of both antigen and Fc interactions, we surmised that upon arranging measurements according to the antigen or Fc feature assessed we might more effectively identify patterns within the data (see detailed justification in methods). We started by restructuring the HIV infection serology data[132]. To integrate the antigen-specific array and gp120-exclusive glycan measurements, we used a form of tensor-based dimensionality reduction, coupled matrix-tensor factorization (CMTF) (Fig. 5.1b/c). By concatenating both the unfolded tensor and matrix during the alternating least squares (ALS) solve for the subject dimension, we achieve the optimal low-rank approximation for both datasets (Fig. 5.1d, see methods). This structure is like canonical polyadic (CP) decomposition on a single tensor, or principal component analysis (PCA) on a single matrix (Fig. 5.1d). The approximation aims to explain the maximal variance across both datasets, in contrast to partial least squares regression in matrix or tensor form (tPLS), which would explain only the shared variance (Fig. 5.1d).

To determine the extent of data reduction possible, we examined the reconstruction error upon decomposition with varying numbers of components (Fig. 5.2a). As the datasets were formatted into a 3-mode (i.e., axis) 181×22×39 tensor and a 181×25 matrix, we start with a structure of 159,823 entries, of which 95,484 or ~60% were filled with measurements (Fig. 5.1b). After factorization with 6 components, we are left with four matrices of 181×6, 22×6, 39×6, and 25×6. Therefore, we reduce the dataset to ~1.7% of the size ((181 + 22 + 39 + 25) × 6 = 1,602 numbers), while preserving 62% of its variation (Fig. 5.2a). For comparison, Fc array assays where these measurements came from with sufficient dynamic range reproduce roughly 80% of the variance across replicates[129]. Therefore, we are capturing the majority but not quite all true variation across subjects and measurements.

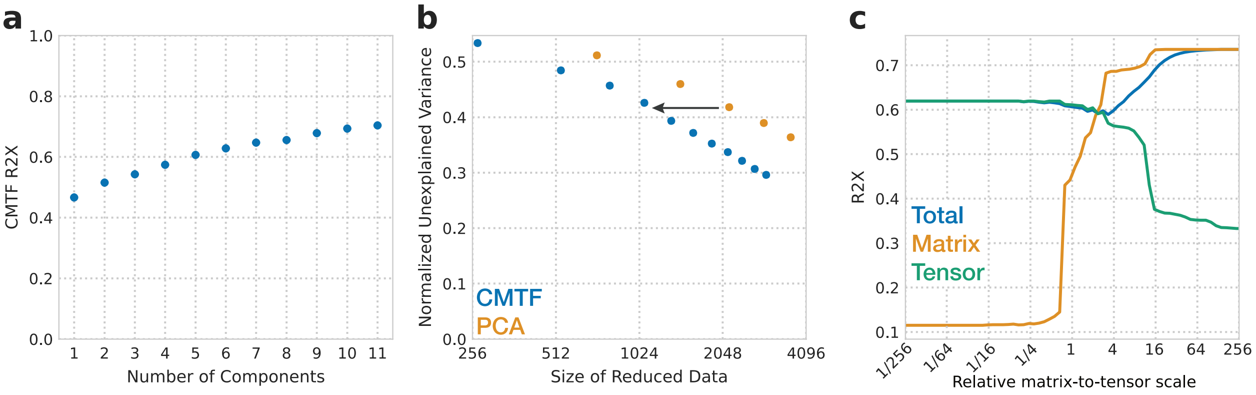


Figure 5.2. CMTF improves data reduction of systems serology measurements.

(a) Percent variance reconstructed (R^2X) versus the number of components used in CMTF decomposition.

(b) CMTF reconstruction error compared with PCA over varying sizes of the resulting factorization. The unexplained variance is normalized to the starting variance. Note the log scale on the x-axis. CMTF consistently led to a similar variance explained with half the resulting factorization size compared with PCA. For example, as indicated by the arrow, to obtain a normalized unexplained variance of 0.45, PCA required ~2,048 values, and CMTF needed only ~1,024 values.

(c) The overall and matrix- or tensor-specific R^2X with varied relative scaling.

We compared this to the data reduction possible with PCA with the data organized in a flattened matrix form. CMTF consistently led to a similar variance explained with half the resulting factorization size compared to PCA (Fig. 5.2b). For example, as indicated by the arrow, CMTF led to a normalized unexplained variance of 0.45 at ~1,024 values within the factorization, while PCA required ~2,048 to do the same. The difference between PCA and CMTF must arise from the latter's ability to "reuse" antigen patterns across receptors or vice versa. For example, if a component includes an increase in FcγRIII binding overall, PCA would still need to represent this increase in the loadings for every FcγRIII-antigen measurement. Thus, PCA is not able to "group" interaction effects across the two dimensions. The difference cannot arise through relaxing orthogonality; CMTF is still "hyper-orthogonal" (i.e., full rank across all tensor modes), and linearly dependent components would only reduce the total variance explained[6,140]. Overall, highly effective dimensionality reduction gave us confidence that this structured factorization identifies patterns of meaningful variation.

As CMTF aims to maximize the explained variances across both datasets, their relative scale influences the balance of the decomposition (Fig. 5.2c). We standardized the data during preprocessing by scaling the matrix so that it contains the same amount of variance as the tensor. In this case, CMTF explains ~62% of the tensor and ~40% of the matrix variance (R^2X). When the matrix is scaled to relatively larger variance, CMTF can achieve ~72% matrix R^2X, at the expense of the tensor R^2X dropping below 35%. Conversely, a smaller matrix does not increase the tensor R^2X over 65% but causes the matrix R^2X to decrease sharply. Our approach of equal variance scaling tuned the factorization to the range where it was responsive to both datasets.

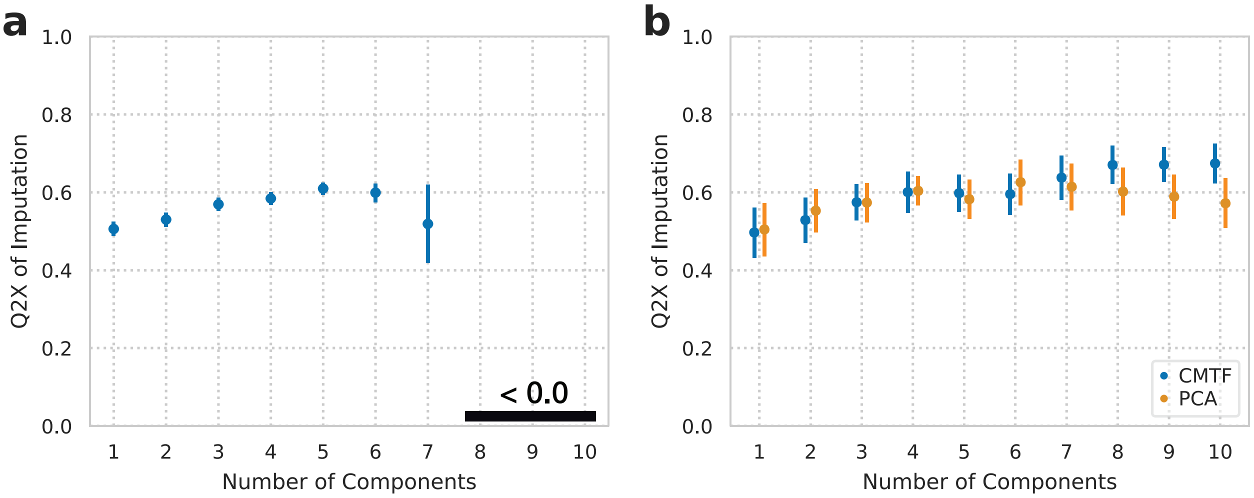


Figure 5.3. CMTF accurately imputes missing values.

(a) Percent variance predicted (Q^2X) versus the number of components used for imputation of 15 randomly held-out receptor–antigen pairs. Error bars indicate standard error of the mean from repeatedly held-out pairs (N = 20).

(b) Percent variance predicted (Q^2X) versus the number of components used for 15 randomly held-out individual values. Error bars indicate standard error of the mean from repeatedly held-out values (N = 10).

5.3 Factorization accurately imputes missing values

By rearranging the measurements into tensor form, our data structure created an entry for every combination of antigen, subject, and Fc property. However, since not every quantity represented by these entries was measured in the dataset, this tensor was left with empty positions, or missing values. To demonstrate that CMTF was robust with missing values, we benchmarked its ability to impute them.

Missing data is not uncommon to biological research. In an experiment, subject samples can be limited or only be available for a small set of measurements, or a subset of measurements can be prioritized by investigators based on prior knowledge. Incapable of handling missing values, one may have to exclude incomplete measurements. In the HIV serology data, gp120-specific glycan measurements were available for only half of the subjects. Consequently, models using the glycan measurements required a smaller patient cohort, and when they were included the prediction performance reduced[132]. Good imputation performance can not only potentially eliminate such tradeoff but also help to infer unknown information. Moreover, factorization accurately imputing missing values further supports that this approach identifies biologically meaningful and consistent patterns.

To evaluate the imputation performance of factorization, we first artificially introduced additional missing values by randomly removing *entire* receptor-antigen pairs *across* all subjects (see methods). We then performed CMTF which effectively filled these in and calculated the Q^2X of the inferred values compared to the left-out data (Fig. 5.3a). Factorization imputed these values with similar accuracy to the variance explained within observed measurements up to 6 components (Fig. 5.2a), supporting that it can identify meaningful patterns even in the presence of missing measurements. As we were effectively leaving out entire columns of data when arranged in flattened matrix form, we could not compare this performance to PCA. Using the average along the receptor or antigen dimensions led to Q^2X values very close to 0. As a less stringent imputation task, we left out batches of individual values and evaluated our ability to impute them. CMTF showed similar or slightly better performance when imputing individual values compared to PCA (Fig. 5.3b). This provides additional evidence that the patterns identified through factorization are a meaningful representation of the data.

5.4 Tensor decomposition accurately predicts functional measurements and subject classes

We next evaluated whether our reduced factors could predict the functional responses of immune cells and subject classes. Functional responses included antibody-dependent complement deposition (ADCD); cellular cytotoxicity (ADCC); neutrophil phagocytosis (ADNP); and the level of natural killer (NK) cell activation represented by the expression of IFNγ, CD107a, and MIP1β. Subject classes included whether subjects were able to control their infection and whether they were viremic at the time of study collection.

To predict the functional responses, we applied elastic net to the decomposed factors (see methods), and their prediction accuracies were defined as the Pearson correlation between measured and predicted values (Fig. 5.4a/c). To predict the subject classes, we applied logistic regression (see methods), and accuracy was defined as the percent classified correctly (Fig. 5.4b/d). To evaluate prediction, we implemented a 10-fold cross-validation strategy. Briefly, in each fold we used 90% of the subjects to learn the relationship between the data and the given prediction and then evaluated these predictions on the remaining 10%. The average performance of each approach was evaluated with every subject eventually held out in one of the 10 folds.

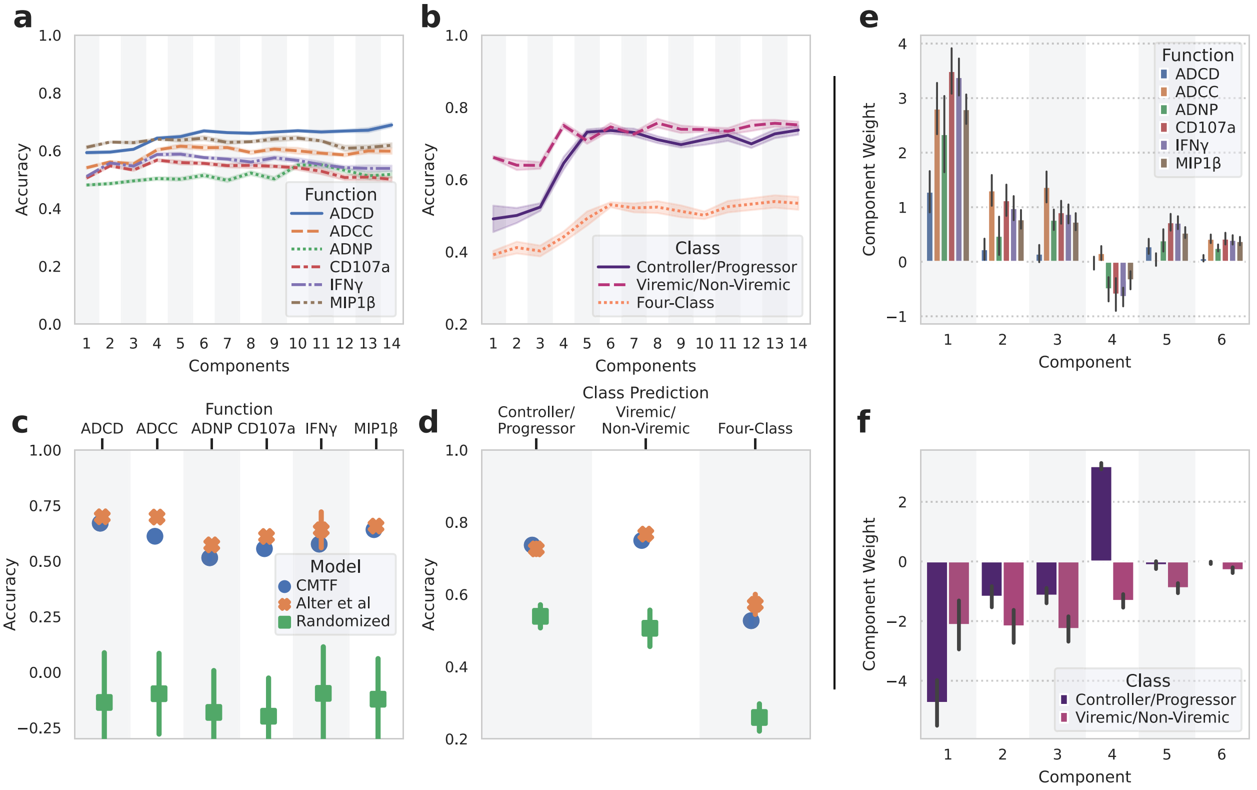


Figure 5.4. CMTF-reduced factors accurately predict functional measurements and subject classes.

(a) Accuracy (defined as the Pearson correlation coefficient) of functional response predictions with different numbers of components.

(b) Percent of subject classes predicted accurately with different numbers of components.

(c) Prediction accuracy for different functional response measurements using six components.

(d) Fraction predicted correctly for subject viral and controller status using six components.

(e, f) Model component weights for each function (e) and subject class (f) prediction.

The shaded area/error bars in (a–d) come from repeating a 10-fold cross-validation (with 10 differently shuffled folds) 10 times (n = 10), and the error bars in (e, f) come from bootstrapping 20 times (n = 20). All error bars indicate the standard deviation from repeated resampling.

To determine the optimal number of components, we first evaluated the prediction accuracies of CMTF with 1 to 14 components (Fig. 5.4a/b). With more components, functional response prediction accuracies improved marginally, and mostly plateaued after 6 components. Subject class predictions saw a leap from 3 to 4 components, especially for controller-progressor classification, and all class prediction accuracies plateaued after 6 components. We therefore concluded that 6 components were generally sufficient for good predictions.

For comparison, we reimplemented the elastic net-based immune functionality and subject predictions previously applied to these data (Fig. 5.4c/d, orange crosses)[132]. We observed similar performance to that reported. Differences from reported results could be explained by adjustments we made to the cross-validation strategy to prevent over-fitting (see methods). Broadly, we saw overall our method performed similarly to the previous method in predicting immune functional responses and subject classes (Fig. 5.4c/d, blue circles). While lower at 6 components, our prediction accuracy increased slightly for ADCD and ADNP at higher numbers of components (Fig. 5.4a). CMTF also had similar prediction accuracy for subject classes with 6 components (Fig. 5.4d, 5.S1). Importantly, in all cases, randomizing the subjects' classes completely removed the ability to make these predictions (Fig. 5.4c/d, green squares).

As both functions and subject classes were predicted with linear models, we plotted the component weights for these regression results (Fig. 5.4e/f). All the NK activation measurements (IFNγ, CD107a, and MIP1β) were highly correlated (Pearson correlation >0.85) and unsurprisingly had very similar model weights, while ADCD, ADCC, and ADNP differed more (Fig. 5.4e). To quantify the stability of these component-function and component-class relationships, we performed bootstrapping by resampling the subjects with replacement, and included error bars representing the standard deviation of the model weights (Fig. 5.4e/f). In every case most of the model weights varied little across samples. By contrast, bootstrapping the elastic net model of ADCD based on the original measurements themselves, as an example, led to entirely different model weights (Fig. 5.S2). We overall concluded that CMTF preserves sufficient information to predict these important features. Data reduction enables one to identify patterns that are associated with functional responses and subject classes, and component associations generalize more robustly upon resampling.

5.5 Factor components represent consistent patterns in the HIV humoral immune response

We plotted the results of our factorization in four factor plots to inspect the composition of each component across each factor dimension (Fig. 5.5). After ALS, components were ordered by their variance, with component 1 having the greatest variance and component 6 having the least. Since the effect of a component is the product of weights on three modes, the original tensor is invariant to coordinated sign flipping or scaling. We enforced that the receptor and antigen factors are positive on average by cancelling out negative effects along two factor modes. Factor components were also scaled to fall within the range of -1–1, and their scaling factors were 29.3, 12.4, 7.4, 7.2, 14.0, and 3.9 respectively. We elected to not scale the glycan factors on a per-component basis so that the relative scaling is evident in the plot itself (Fig. 5.5d). Every component must be distinct along at least one factor matrix due to hyper-orthogonality, so no component was redundant.

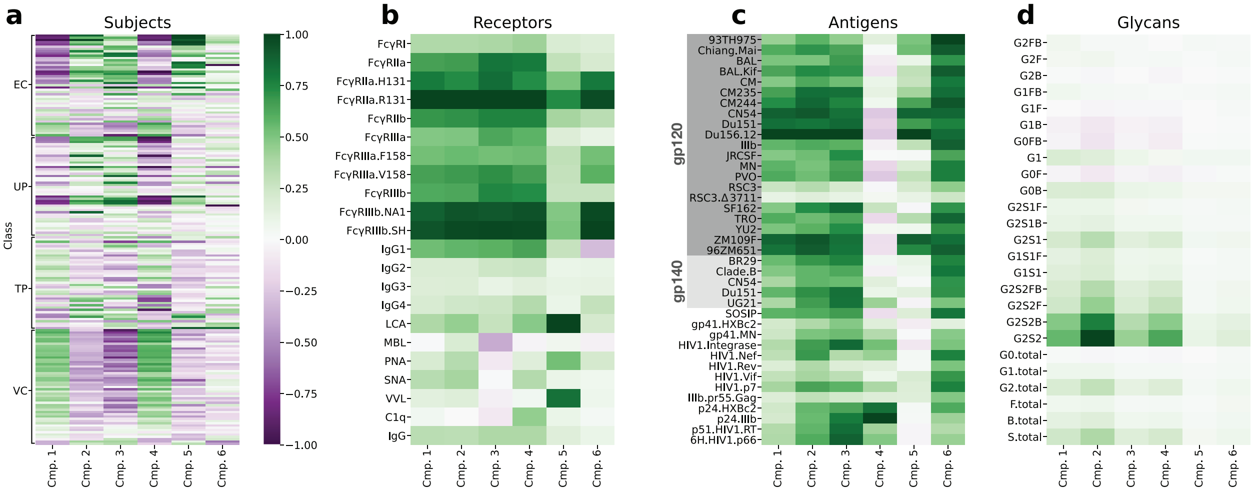


Figure 5.5. Factor components represent consistent patterns in the HIV humoral immune response.

(a-d) Decomposed components along subjects (a), receptors (b), antigens (c), and glycans (d). EC: elite controller, TP: treated progressor, UP: untreated progressor, VC: viremic controller (see Methods). All plots are shown on a common color scale after scaling each factor component within the range −1 to 1. Antigen names indicate both the protein (e.g., gp120, gp140, gp41, Nef, and Gag) and strain (e.g., Mai and BR29). Descriptions of each receptor name can be found in Tbl. 5.1.

Receptor	Description
FcgRI	FcγRI[130]
FcgRIIa	FcγRIIa[130]
FcgRIIa.H131	FcγRIIa.H131[130]
FcgRIIa.R131	FcγRIIa.R131[130]
FcgRIIb	FcγRIIb[130]
FcgRIIIa	FcγRIIIa[130]
FcgRIIIa.F158	FcγRIIIa.F158[130]
FcgRIIIa.V158	FcγRIIIa.V158[130]
FcgRIIIb	FcγRIIIb[130]
FcgRIIIb.NA1	FcγRIIIb.NA1[130]
FcgRIIIb.SH	FcγRIIIb.SH[130]
IgG1	Mouse anti-Human IgG1[129]
IgG2	Mouse anti-Human IgG2[129]
IgG3	Mouse anti-Human IgG3[129]
IgG4	Mouse anti-Human IgG4[129]
LCA	Lens Culinaris Agglutinin[130]
MBL	Mannan binding lectin[130]
PNA	Peanut Agglutinin[130]
SNA	Sambucus Nigra Lectin[130]
VVL	Vicia Villosa Lectin[130]
C1q	Human C1q[130]
IgG	Mouse anti-Human pan-IgG[129]

Table 5.1. Descriptions of the receptor detections found within the tensor analysis.

The resulting factor plots can be read in two ways. First, one can trace the effect of a component across different dimensions by looking at that component within each plot. For instance, component 4 represents a subset of unique variation in the data that is higher in viremic controllers (Fig. 5.5a), broadly covarying across FcγRs (Fig. 5.5b), and increases p24/decreases gp120 antigen binding (Fig. 5.5c). In an alternative view of the factorization results, one can ask how components are different in the variance they explain within a single factor mode. For instance, components 2 and 4 are very similar in their receptor interactions (Fig. 5.5b), but unique in their antigen binding specificity (Fig. 5.5c). Finally, the product of subject (Fig. 5.5a) and glycan (Fig. 5.5d) factors reconstructs the glycan measurements.

Components 1 and 2 explained the most variance and had broad receptor (Fig. 5.5b) and antigen (Fig. 5.5c) weighting, indicating that they represent overall titers in a general manner. Some difference exists within both components in their antigen specificity—component 1 is weighted toward surface antigens, while component 2 is more uniform in its antigen weights (Fig. 5.5c). Component 1 (along with component 4) was also uniquely high in viremic controllers compared to other groups (Fig. 5.5a). Component 3 represents a similar antigen specificity to component 2 (Fig. 5.5c), and similar receptor set except for most of the lectin-binding proteins (MBL, PNA, SNA, VVL) and C1q (Fig. 5.5b). Component 4 displayed similar receptor specificity to component 1 (Fig. 5.5b) but with unique antigen specificity that was positive for intracellular antigens and negative for surface ones (Fig. 5.5c). Component 5 was surface antigen-specific (Fig. 5.5c) and strongly specific for LCA, PNA, and VVL (Fig. 5.5b). Finally, component 6 was weighted toward genotype specific FcγR measurements over all others (Fig. 5.5b), with broad antigen specificity (Fig. 5.5c). As these were (1) the most sensitive measurements as indicated by their generally higher fluorescence signal before normalization and (2) the component's variation was greatest for the subjects that were low on component 1 (Fig. 5.5a), we took this to indicate the component explained variation specific to low-titer subjects.

We were surprised to find little unique variation in the glycan matrix factor along each component (Fig. 5.5d). The weights within each component were proportional to the dynamic range of each measurement (most for G2S2 and less for total G0 as an example; $r^2 = 0.82$). We took this to indicate that there is little variation explained in the glycan data beyond an overall increase or decrease. As independent evidence of this, a one-component PCA decomposition of just the glycan matrix could explain >70% of the variation in the glycan data, even after centering.

5.6 CMTF extensively reduces and visualizes dynamic responses to SARS-CoV-2 infection

To demonstrate the general benefit of tensor methods in systems serology data analysis, we applied them to a separate dataset on acute SARS-CoV-2 infection[133]. In this dataset, samples from SARS-CoV-2 negative and infected subjects were collected over the course of infection for about 4 weeks. Antibodies were tested for their antigen and Fc receptor engagement. We restructured the data into a three-mode tensor according to the sample, antigen, and receptor measured. In doing so we obtained a tensor of size 438×6×11 (Fig. 5.6a). In this form, the tensor contains no missing values. After log-transforming and centering the data on a per-antigen-receptor basis, two components could explain 74% of the variance with 0.3% the size of the original dataset ((438 + 6 + 11) × 2 = 910 numbers) (Fig. 5.6b).

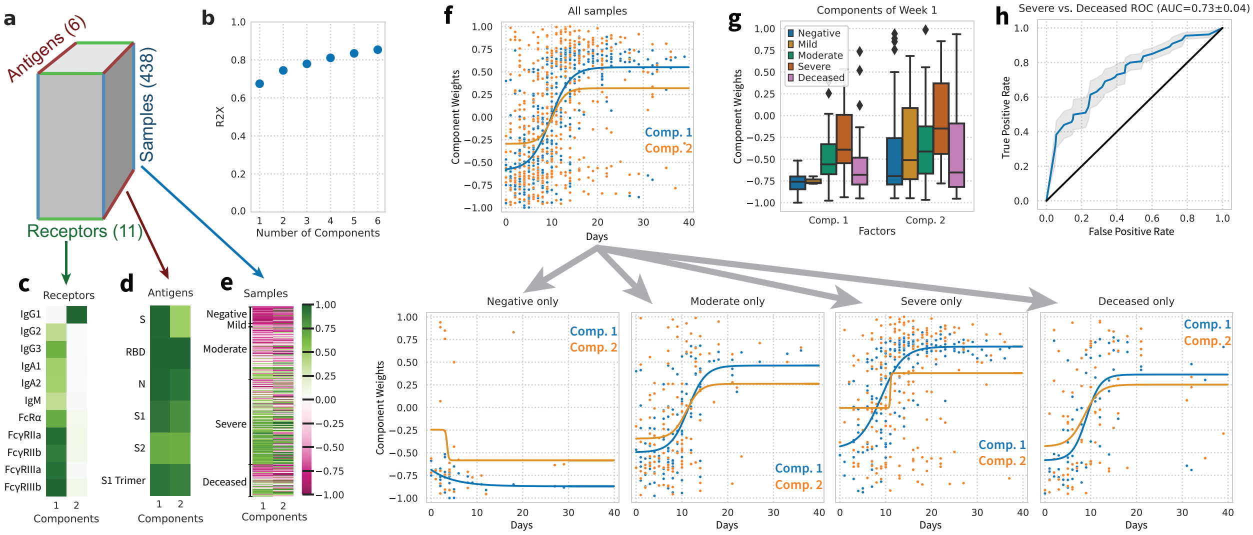


Figure 5.6. Application of tensor factorization to SARS-CoV-2 systems serology measurements.

(a) Schematic of the data tensor. Measurements were arranged according to samples, target antigen, and receptor detection.

(b) Percent variance reconstructed (R^2X) versus the number of components.

(c-e) Decomposed factor components along samples (c), antigens (d), and receptors (e).

(f) Subject component weights plotted according to the sample time after symptom onset, together and separated into PCR-negative subjects along with moderate, severe, and deceased cases.

(g) Boxplot of subject component weights for just samples within the first week after symptom onset, separated by subject group (negative N = 33, mild N = 7, moderate N = 122, severe N = 196, deceased N = 74). Each point represents a distinct biological sample. The three bands in each box represent the first, second, and third quartiles of the weights, from the bottom to the top, respectively; the whiskers extend up to 1.5 times the interquartile range beyond the box range; any outliers beyond the whisker ends are plotted as single points.

(h) ROC curve of logistic regression classifier for predicting severe disease versus deceased outcome. Model is built using the two component weights of the subject factors. The shaded area indicates the standard deviation from a 10-fold cross-validation.

The resulting factors clearly separated into a clear acute (e.g., IgG3, IgM, IgA) or long-term (IgG1-specific) response pattern[34] (Fig. 5.6c), with the abundance of each program in each sample indicated by the sample factors (Fig. 5.6e). Both components 1 and 2 generally shared broad specificity across antigens with slight differences (Fig. 5.6d). As expected, component 1 also represented stronger association with FcγR and FcαR immune receptors[31] (Fig. 5.6c).

We proceeded similarly to earlier analysis[133] and plotted each sample by the collection time after symptom onset, separated by the outcome of infection (SARS-CoV-2 negative, moderate disease, severe disease, and deceased) (Fig. 5.6f). A sigmoidal curve was fit to each temporal profile as a summary of the data. In contrast to the earlier analysis, we were able to plot these along the two components summarizing all the data, instead of the 66 individual measurements. Overall, samples showed a time-dependent increase in factor values (Fig. 5.6f). Interestingly, a subset of PCR-negative subjects showed positive weights specific to component 2, indicating some IgG1-specific pre-existing immunity. As previously observed, severe cases displayed a component 1 response that on average had a higher initial and final quantity compared to either moderate or deceased cases (Fig. 5.6f/g). A logistic regression classifier using just the week 1 data predicted severe versus deceased outcome with AUC of 0.73 (Fig. 5.6g/h), comparable to a random forest classifier in previous analysis[133] (AUC 0.71). Factorization with more components than 2 did not improve classification accuracy.

5.7 Discussion

We show here that tensor-structured data decomposition can improve our view of systems serology measurements. Specifically, this approach recognizes that antibody variation takes place across the distinct and separable antigen binding and Fc/receptor dimensions. Using this property, we identify that these measurements can be reduced more efficiently (Fig. 5.2), this reduction is robust to missing values (Fig. 5.3), and that properties of the immune system and infection can be accurately predicted (Fig. 5.4). Most critically, this form of dimensionality reduction provides a clearer interpretation of the resulting models (Fig. 5.5), as it accounts for the high degree of inter-correlation across each dimension. Finally, reducing the data into patterns enables robust associations between the biophysical parameters of antibodies and functional responses or immunological status (Fig. 5.4, 5.S2).

The resulting factors and their association with infection state extend prior knowledge regarding changes in humoral immunity in HIV. One of the clearest patterns is an association of progression status with components 1 and 4, representing an antigen shift between surface and intracellular antigens (Figs. 5.4f, 5.5c, 5.S1b). Abundance of p24 antigen and its antibody titer has been proposed as an effective marker of HIV progression[142], and predictive of death[143], although it correlates strongly with viral RNA and CD4+ counts[144]. As we observe with component 4, viremic controllers have been characterized as having especially high p24-specific IgG1 and IgG2 driving phagocytic responses[145]. The negative association with component 1 likely reflects a decrease in antibody titers overall which has separately been found to predict progression[146]. Therefore, while features of p24 abundance or antibody titers may have an incomplete and complex relationship with progression, a p24/gp120 ratio may be more predictive (Fig. 5.S3). Viremia status was predicted through an even decrease in many of the components (Fig. 5.4f), generally opposite the component weights predicting functional responses (Fig. 5.4E). This broad difference matches perfectly with previous observations that viral control is associated with polyfunctionality, rather than a specific molecular program[136]. Predicting viremia using the viral RNA quantities, rather than classifying groups based on a threshold, could reveal more specific regulatory changes since, for example, elite controllers were heterogeneous as a group, and this variation may correlate with viral RNA amounts[147] (Fig. 5.5a).

The functional predictions and their component contributions matched expected patterns. All three NK activation measurements (CD107a, IFNγ, and MIP1β) had very similar weights to be expected given their high correlation (Fig. 5.4e). The only component with a negative weight with respect to gp120/gp140, component 4, showed a negative or negligible contribution to functional predictions (Fig. 5.4e, 5c). More specifically, ADCC was predicted by positive association with all the components that included both FcγRIII and surface (gp120/gp140) antigen binding (components 1, 2, 3, and 6)(Fig. 5.4e, 5b, 5c). ADCD had only two consistently positive component weights—1 and 5—which can be taken to reflect probably overall titers and lectin-pathway complement activation, respectively, as component 5 had unusually strong weights for the glycan-binding probes LCA, PNA, and VVL[148] (Fig. 5.5b). In contrast, while sparse models can predict these functions as accurately[132], one cannot assign significance to the individual model weights as they change upon resampling the dataset (Fig. 5.S2), a challenge when modeling highly correlated measurements such as these[149–151]. Establishing links between specific molecular factors and functional responses could be further improved by experimentally introducing less correlated variation into the data, such as by measuring samples after enzymatic glycosylation modifications or depletion of certain isotypes[152,153]. This also highlights the need for multi-variate serological profiling, as single-factor studies are likely to find indirect associations.

Although the glycan measurements had insufficient variation to link them to specific molecular programs beyond variation in amounts overall, future refinements to these measurements or analysis may reveal more precise regulation[125]. Glycan measurements across a panel of antigens might reveal more specific regulation, particularly as glycans are known to be tuned in an antigen-specific manner[152,154,155]. Paired glycan and biophysical measurements in acute infection may also reveal more drastic glycan variation, especially given links with outcomes such as between severe COVID-19 and IgG fucosylation[133,155]. A tensor partial least squares regression approach would also reveal variation specifically associated with glycan changes by specifically focusing on variation shared in both datasets[139].

A more recent study examining SARS-CoV-2 infection allowed us to explore whether tensor-structured dimensionality reduction has benefits that extend to the serology of other disease and in longitudinal studies (Fig. 5.6). Surprisingly, we found that just two patterns within these data could explain 74% of the variance (Fig. 5.6b). While we were able to replicate the difference in dynamics between severe and deceased cases (Fig. 5.6f/g/h), the sufficiency of just two patterns argues for quantitative differences in these two patterns, rather than detailed qualitative changes in the immune response[133]. Perceived differences between individual measurements could arise in part from these two component patterns being combined in each measurement. As evidence of this, the reported measurements that differed in dynamics between severe and deceased subjects were almost exclusively those we observed to be weighted on both components 1 and 2, while those specific to component 1 showed no difference between outcomes[133] (Fig. 5.6e/f). It is also difficult to draw conclusions on a measurement-by-measurement basis, even in large studies such as this, due to large subject-to-subject variability and strong correlations between measurements (e.g., Fig. 5.S2). On the other hand, it is possible that immunologically significant patterns remain in the unexplained variance that are drowned out by the most drastic changes[155]. There is also a challenge in separating pre-existing partial immunity from prior exposures or cross-reactivity with other coronaviruses; some PCR-negative cases showed positive weights on component 2, presumably indicative of long-term humoral responses, possibly from cross-reactivity with other coronaviruses[156] (Fig. 5.6f). Longer-term longitudinal studies of acute infection would allow one to observe the transition from acute immune response to lasting protection, and potentially better resolve the dynamics of class switching alongside its functional consequences[34].

Other tensor arrangements of serology data will help to reveal new patterns within these data. Indeed, here we have arranged data both with subject, antigen, and receptor modes, in either coupled form (Fig. 5.1) or a single tensor (Fig. 5.6a). With longitudinal data in which timepoints can be aligned, one could create a mode representing the contribution of time[119]. While each antigen is treated similarly along one dimension, antigenic mutants or strains could also be separated into separate tensor modes before decomposition. This could lead to further data reduction (e.g., both strains of p24 and gp41 antigens share a similar signature; Fig. 5.5c) and simplify comparisons between strains. Differences in the weights would also essentially serve as an unsupervised prediction for competition experiments to reveal differences in the binding targets of polyclonal serum[55]. As compared to traditional blocking or mutational experiments, antigens in these measurements are multiplexed across identifiable beads[157,158]. Therefore, making measurements across a wider antigen panel requires just small amounts of each antigen, and can be scaled to hundreds of antigens without increased sample requirements. Finally, CMTF could be used to link other types of immune response measurements besides glycan quantitation to serology, such as cytokines and gene expression.

More effective dimensionality reduction in turn enables new ways of viewing antibody-mediated protection. Thinking of these measurements as akin to the microarray for gene expression data suggestions new possibilities in leveraging this data. One valuable property of CMTF is that it separates the immune receptor and antigen-binding patterns within the data. This will enable surveys for common Fc response patterns across diseases and studies because these different datasets would still share this axis. This “transfer learning” could therefore help to identify common patterns of immune dysregulation. With more extensive profiling of the various glycosylation and isotype Fc forms, it would be possible to fix the receptor axis of the decomposition, in effect matching new measurements to specific known immunologic patterns. These pattern-matching approaches would be much like gene set enrichment analysis for expression data[159]. The binding interactions of antibodies, while they produce combinatorial complexity, are a simple set of antigen and receptor binding. Ultimately, one should be able to apply multivalent binding models to mechanistically model the interactions within serum[10,13,46]. This might allow separation of avidity versus affinity in binding and integration with extensive prior characterization of Fc properties, such as the biophysical properties of individual glycoforms and isotypes[30,31]. A mechanistic view could also help to guide more advanced multi-modality therapeutic interventions, like inhibitors or enhancers of antibody response that cooperate with the cocktail of endogenous antibodies[154,160].

Ultimately, a comprehensive view of immunity needs advancements in measurements that are complementary to systems serology. Much like how systems serology has served to profile antibody-mediated protection, profiling methods are helping to characterize T cell-mediated immunity[161]. These technologies, alongside more traditional technologies to profile cytokine response, gene expression, and other molecular features, promise to provide a truly comprehensive view of immunity. Integrating these data will require dimensionality reduction techniques that recognize the structure of these data alone and in combination. Factorization methods, especially those operating on tensor structures, will be a natural solution to this challenge, due to their scalability, flexibility, and amenability to interpretation[101].

Software	Source	Version
TensorLy Python Library	http://tensorly.org/	v0.6.0
SciPy Python Library	https://www.scipy.org/	v1.7.0
NumPy Python Library	https://numpy.org/	v1.21.0
Pandas Python Library	https://pandas.pydata.org/	v0.2.5
Seaborn Python Library	https://seaborn.pydata.org/	v0.11.1
Python	https://www.python.org	v3.9.5

Table. 5.2. Software and packages used in this chapter

5.8 Materials and Methods

Subject cohort, antibody purification, effector function assays, and glycan analysis

All experimental measurements were collected from prior work[132,133]. Measurements were clipped to be at least 0.1 (HIV glycan), 1.0 (HIV biophysical) or 10.0 (COVID biophysical); log-transformed; and then centered on a per-measurement basis across subjects. The thresholds before log-transformation were determined to be well below the level of noise in the assays using the negative controls for each. Two antigens, gp140.HXBc2 and HIV1.Gag, were identified to only have one and two receptor measurements, respectively, making their factor values unstable because almost all measurements were missing. These were removed on import during the tensor-based analysis. HIV subjects were classified into four categories: untreated progressors, who failed to control viremia without combined anti-retroviral therapy (cART); treated progressors, who similarly failed to control viremia without cART but were on it for the study measurements; viremic controllers, who possessed a viral load between 50 and 2,000 RNA copies/mL without cART; and elite controllers, who had less than 50 copies/mL without cART. These were then grouped into two classifications: controllers (EC and VC) versus progressors (UP and TP); and viremic (UP and VC) versus non-viremic (TP and EC).

Coupled Matrix-Tensor Factorization

We decomposed the systems serology measurements into a reduced series of Kruskal-formatted factors. Tensor operations were performed using Tensorly[82]. Most measurements were made across specific antigens, and we structured them into a 3-mode tensor, $\mathcal{X}$, whose modes represent subjects, receptors, and antigens. Separately, gp120-associated antibody glycosylation was measured for half of the HIV subjects. These measurements were structured into a matrix, Y, representing the quantities for each subject (Fig. 5.1).

Shaping antigen-specific data into a 3-mode tensor recognizes that measurements of the same receptor or antigen should share variation within each component. However, since not all receptor-antigen pairs were measured, the constructed tensor contained missing values from the perspective of this data structure. Throughout the factorization algorithm, we used censored least squares solving, with rows corresponding to missing values removed.

In preprocessing, we scaled the matrix so that it contained the same amount of variance as the tensor. To perform CMTF, we assumed the subject mode was shared between the tensor and the matrix:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a_r} \circ \mathbf{b_r} \circ \mathbf{c_r} = \widehat{\mathcal{X}}$$

$$Y \approx \sum_{r=1}^{R} \mathbf{a_r} \circ \mathbf{d_r} = \hat{Y}.$$

Here, "∘" represents the vector outer product, and R is the total number of components in the factorization. The original tensor is approximated as a sum of R rank-one tensors constructed by the vector outer product along each mode. The original matrix is represented by the sum of R rank-one matrices formed by the outer product of row and column vectors. For the r-th component, $\mathbf{a_r}$, $\mathbf{b_r}$, and $\mathbf{c_r}$ are vectors indicating variation along the subject, receptor, and antigen dimensions, respectively, and $\mathbf{d_r}$ is a vector indicating variation along glycan forms within the glycan matrix.

Decomposition was initialized using singular value decomposition of the unfolded data along each mode, with missing values imputed by a one-component PCA model and entirely missing columns removed. We then optimized the decomposition using an alternating least squares (ALS) scheme[6] for up to 2000 iterations. In each ALS iteration, linear least squares solving was performed on each mode separately[92]:

$$\min_{A} \left\| \left[X_{(1)}\ Y \right] - A[(C \odot B)^{\mathrm{T}}\ D^{\mathrm{T}}] \right\|^2$$

$$\min_{B} \left\| X_{(2)} - B[(C \odot A)^{\mathrm{T}}\ ] \right\|^2$$

$$\min_{C} \left\| X_{(3)} - C[(B \odot A)^{\mathrm{T}}\ ] \right\|^2$$

$$\min_{D} \| Y - AD^{\mathrm{T}} \|^2$$

where $X_{(1)}$, $X_{(2)}$, and $X_{(3)}$ are the tensor unfoldings of $\mathcal{X}$ along each mode, and "$\odot$" represents Khatri-Rao product. The R^2X was checked on each even iteration and decomposition was terminated early if the change was found to be less than 10^{-5}.

Justification of Multiplicative Factor Interactions

Kruskal-formatted tensors are structured such that each factor component (receptors, antigens, subjects) should be multiplied together to reconstruct the data. This structure is simply a higher-dimensional generalization of matrix decomposition techniques like PCA or non-negative matrix factorization, in which scores and loadings matrices are multiplied together to reconstruct the data. An expectation of these approximations is that variation within the tensor occurs in a pattern that can be localized to each tensor slice, which is justified by the nature of the measurements being considered. These measurements are made in an array format, wherein plasma samples from subjects are incubated with individually identifiable beads covalently conjugated with antigens[129,130]. The conjugated antigens isolate IgG fractions specific to those targets. After washing, the beads are incubated with fluorescently labelled detection reagents that bind to the isolated IgG depending upon their properties. Thus, in essence, the assay is a bead-based sandwich ELISA, in which the IgG is the sandwiched target. Given the format, the amount of fluorescence measured on a given bead should be proportional to both (1) the amount of IgG isolated on the bead, and (2) the fluorescence signal obtained per isolated IgG (two of the tensor modes), supporting their multiplication. A multiplicative relationship allows each factor to contribute positively, negatively, or not at all to the variation represented by a component by a positive, negative, or zero weighting. A strong validation of this structure in tensor form is the large extent to which the data can be reduced without loss of information. It also fits with the biological expectation that antibody Fab variation influences the data along the antigenic slices, while Fc variation influences the data along the receptor ones.

Reconstruction Fidelity

To calculate the fidelity of our factorization results, we calculated the percent variance explained, R^2X. First, the total variance was calculated by summing the variance in both the antigen-specific tensor and the glycan matrix, where included:

$$v_{\text{total}} = \|\mathcal{X}\|^2 + \|Y\|^2$$

Variance was defined as the sum of each element squared, or the square of the norm. Any missing values were ignored in the variance calculation throughout. Then, the remaining variance after taking the difference between the original data and its reconstruction was calculated:

$$v_{r,\text{antigen}} = \|\mathcal{X} - \widehat{\mathcal{X}}\|$$

$$v_{r,\text{glycosylation}} = \|Y - \widehat{Y}\|$$

Finally, the fraction of variance explained was calculated:

$$R^2X = 1 - \frac{v_{r,\text{antigen}} + v_{r,\text{glycosylation}}}{v_{\text{total}}}$$

Where indicated as Q^2X instead, this quantity was calculated only for values left out to assess the fidelity of imputation.

Logistic Regression / Elastic Net

The data was centered and variance-normalized prior to model assembly. Logistic regression and elastic net were performed using LogisticRegressionCV and ElasticNetCV implemented within scikit-learn[162]. Both methods used 10-fold cross-validation to select the regularization strength with smallest cross-validation error, and a fraction of L1 regularization equal to 0.8 to match previous results[132]. Logistic regression used the SAGA solver[163].

Cross-Validation

We employed a 10-fold cross-validation strategy to evaluate each prediction model. Subjects were randomly assigned to folds to prevent the influence of subject ordering in the dataset. We found that sharing the cross-validation fold structure between hyperparameter selection and model benchmarking led to consistent overfitting. Therefore, we used a nested scheme in which the folds were assigned differently for hyperparameter selection and model performance quantification.

Principal Components Analysis

Principal components analysis was performed using the implementation within the Python package statsmodels and the SVD algorithm. Missing values were handled by an expectation-maximization approach wherein they were filled in repeatedly by PCA. This filling step was performed up to 100 iterations or until convergence as determined by a tolerance of 1×10^{-5}.

Missingness Imputation

To evaluate the ability of factorization to impute missing data, we introduced new missing values by removing (1) entire receptor-antigen pairs or (2) individual values from the antigen-specific tensor as indicated, and then quantifying the variance explained on reconstruction (Q^2X). More specifically, in the first situation, fifteen randomly selected receptor-antigen pairs were entirely removed (2,715 values) and marked as missing across all subjects, leaving ~93,000 values for training. In the second, fifteen randomly selected individual values were removed, leaving ~96,000 training values. CMTF decomposition was performed in each trial as described above, and the left-out data were compared to the reconstructed values. 20 or 10 trials were performed in each imputation situation, respectively. Varying numbers of components were used for decomposition and a Q^2X was calculated for each. In the second case we compared CMTF with PCA-based imputation with the dataset flattened into matrix form.

Fitting Sigmoidal Curves

The sigmoidal curves in Fig. 5.6f were fit to $y = A/[1 + \exp(-k(x - x_0))] + C$ using the Levenberg-Marquardt algorithm as implemented within curve_fit in the Python package scipy. The initial optimization point was set so that A was 0.6 of the y range; C was the smallest y; x_0 was the median x; and k was either 0.5 or -0.5, depending on whether the mean of the first half of y's was larger or smaller than that of the latter half.

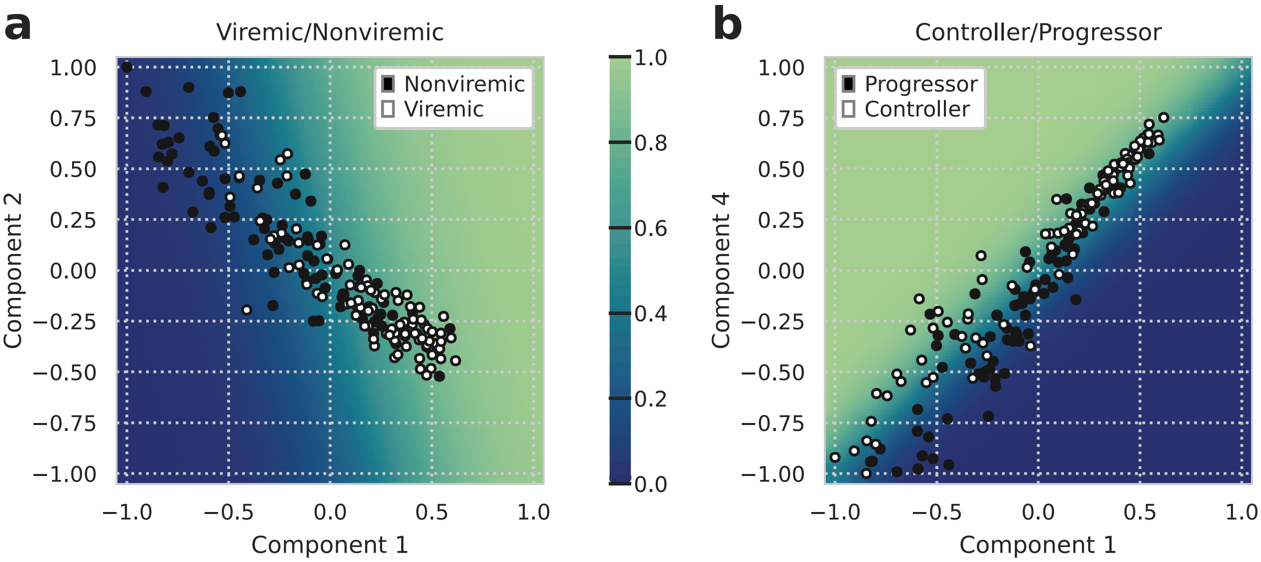


Figure 5.S1: Decision boundaries of subject classification.

(a) Viremic/non-viremic decision boundary, plotted as the probability of being viremic.

(b) Controller/progressor decision boundary, plotted as the probability of being a controller.

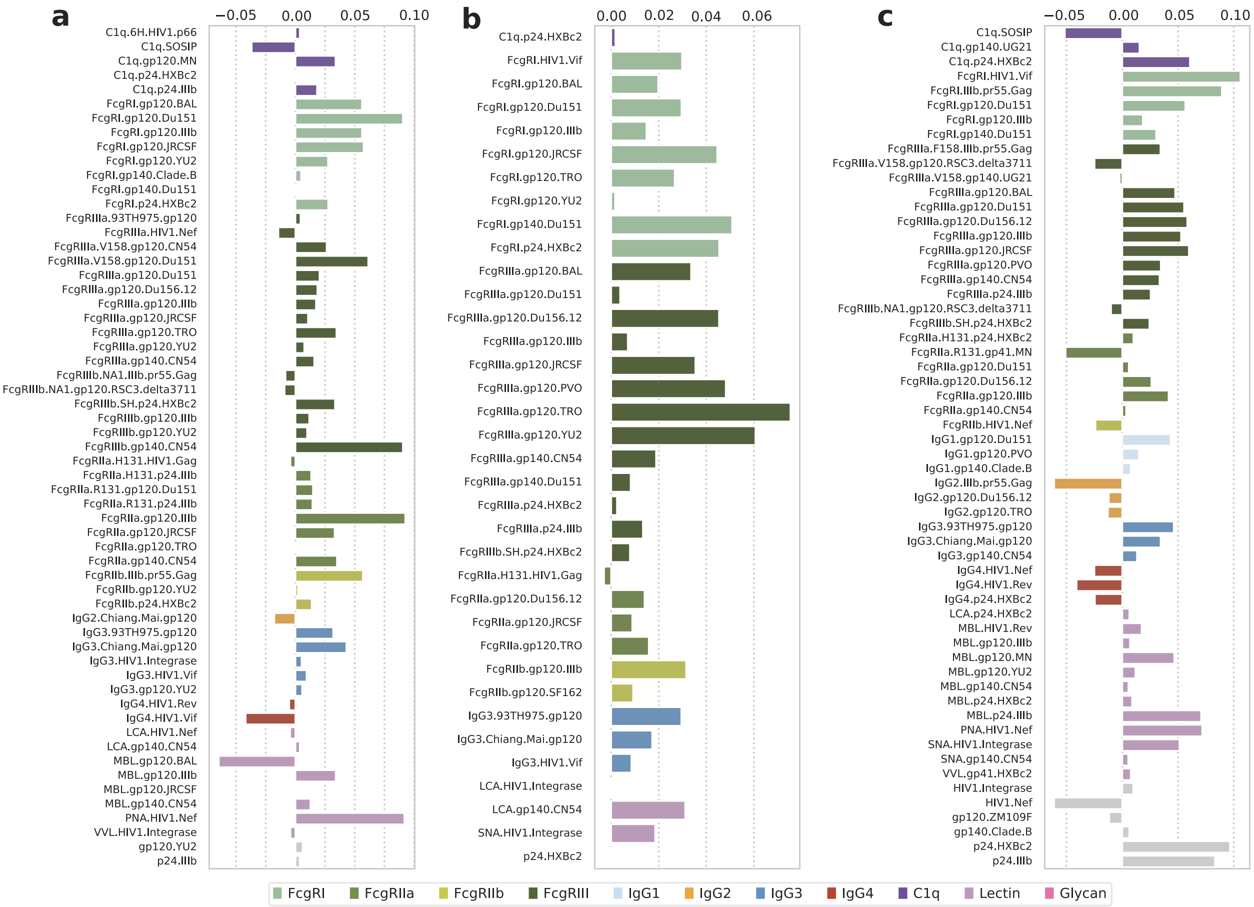


Figure 5.S2: Three runs of bootstrapped samples of elastic net model predicting ADCD.

(a-c) Each subplot shows the feature weights from an independent run of the elastic net model with bootstrapping predicting ADCD. Bars are colored according to the category of each input variable. Bootstrapping was performed by resampling subjects with replacement and then rebuilding the elastic net model predicting ADCD.

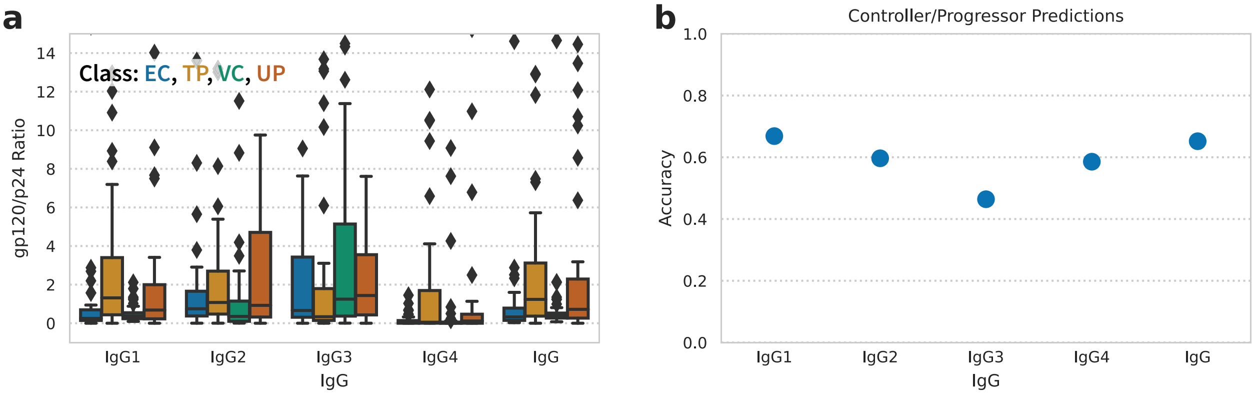


Figure 5.S3: A simple gp120/p24 antigen ratio predicts progression to varying degrees.

(a) Raw gp120/p24 IgG measurement ratios, separated by subject class. Data plotted as boxplots. Each point represents a distinct subject (EC N = 45, TP N = 44, VC N = 51, UP N = 41). The three bands in each box represent the first, second, and third quartiles of the ratios, from the bottom to the top, respectively; the whiskers extend up to 1.5 times the interquartile range beyond the box range; any outliers beyond the whisker ends are plotted as single points.

(b) Controller/progressor prediction accuracy using gp120/p24 ratios of each IgG isotype. Predictions were performed using logistic regression as described in methods. Accuracy is defined as the percent correctly classified.

Chapter 6

Conclusion

He is blessed by heaven.

Good fortune.

Nothing that does not further.

I Ching

In this dissertation, I present two distinct computational modeling strategies for antibody response: mechanistic-based and data-driven. These approaches embody two philosophical and epistemic pathways prevalent in contemporary biological studies[164–166]. My study showcases common traits in these approaches beyond antibody research. While each chapter has discussed specific biological implications, this chapter focuses on these two methodologies themselves and their roles in computational biology.

The mechanistic approach is typically employed to gain a deeper understanding of well-characterized systems. In traditional biology research, researchers study one pathway in great detail, examining each component meticulously. Mechanistic modeling is a quantitative extension of this approach, following a reductionist logic[167]: if we know all the details of a pathway, we should be able to make accurate predictions, and vice versa. Under this umbrella, there are ordinary differential equations (ODE) models[168], rule-based models[169], logical-network-based models[170], graph-based models[171], and others. A mechanistic model can integrate existing discoveries quantitatively, allowing for fine-tuning of details and hypothesis testing. For example, from Fig. 3.1b, we observe a significant discrepancy in FcγRI-IgG2 measurements compared to documented affinities, but without the model, quantifying it is challenging due to the nonlinear nature of the binding. This binding model can be easily applied to the binding of glycosylation mutants[172] and Fc optimization[173]. In such cases, the model can serve as an *in silico* lab to efficiently test the effects of unmeasured antibody mixtures[19,174].

However, the mechanistic approach has clear challenges. It requires that each component of the system be well-characterized. For instance, the binding model employed in Chapter 3 necessitates knowledge of different IgG subclasses, FcγRs, and their monovalent affinities. Consequently, mechanistic models build on existing knowledge, limiting their application in exploratory studies. These models also face scalability issues. The model developed in Chapter 2 demonstrates good scalability, representing a significant advancement. However, it relies on simplistic assumptions (such as a single K_x^* for steric effect) and only addresses a specific kind of cell surface binding interaction without additional processes (e.g., receptor trafficking or clustering). Moreover, cascading a specific pathway's effect to an individual level is challenging. In Fig. 3.6, I attempted this, but the results were still constrained since numerous factors remain unconsidered, including pharmacodynamics, individual-level diversity, and the impact of other pathways.

Conversely, the data-driven model operates on statistical reasoning. This approach has gained great popularity, as it is empowered by the recent advancements in big data and machine learning. It measures possible factors and effects directly, hoping that the statistical model reveals their association. This exploratory approach, as seen in Fig. 5.4f, allowed the discovery of components most correlated with patient classes, previously unknown. Statistical models' ability to generate new insights independent of controlled experiments makes them attractive to the biological community. This is particularly true when the number of potential factors is extensive, such as genome-wide association studies[175].

However, statistical approaches often oversimplify questions, with feature selection highly dependent on the data provided and the chosen model. They treat complex subjects, like cells or individuals, as black boxes[176], relying on statistical tools for insights. Nonetheless, correlations do not imply causations, especially in these complicated biological systems. Additionally, each model is based on different assumptions, leading to divergent discoveries. For instance, using the same dataset, tensor method approach presented in this study (Fig. 5.4f) produced a different set of features from a previous study using sparse regression (Fig. 5.4.S2). While model performance can be evaluated (e.g., *p*-values, cross-validation accuracy, model stability), definitive verification of discoveries is challenging without additional information. Hence, independent experimental validation cannot be replaced solely by statistical approaches.

In Chapter 4, I extensively discuss the potential applications of tensor methods in biological studies. While not claiming tensor methods are always superior, I argue that many modern biological measurements involve multiple variations, which tensor structures can more faithfully record. As some statisticians insightfully noted, "All models are wrong, but some are useful." The "usefulness" here must relate to how well the model reflects the measurement's intrinsic structure. My vision for future multivariate biological data analysis involves developing methods designed for inherently higher-dimensional tensor-shaped data beyond linear factorization.

To summarize, my study exemplifies the value of both mechanistic and data-driven approaches in antibody research[177]. The mechanistic model excels in systems with known mechanisms, aiming to refine specific understandings, while the data-driven approach shows promise in exploratory studies, identifying significant signals amidst numerous variables. Ideally, these approaches would converge, but numerous factors need to be considered to bridge all gaps. Efforts have been made to create hybrid models, such as data-driven models that incorporate known mechanisms, but these are often bespoke solutions. As we enter this new era of biology, groundbreaking technologies are emerging to gather vast amounts of data, enabling comprehensive mapping of biological systems. This evolution requires a reimagining of biology through quantitative lenses, the redesign of experiments to deeply incorporate statistical analyses, and the innovative application of recent advancements in artificial intelligence.

REFERENCES

1. Effector Responses: Cell-and Antibody-Mediated Immunity. in *Kuby Immunology* 415–450 (W. H. Freeman and Company, New York, 2013).
2. Nimmerjahn, F. & Ravetch, J. V. Divergent Immunoglobulin G Subclass Activity Through Selective Fc Receptor Binding. *Science* **310**, 1510–1512 (2005).
3. Bournazos, S., Gupta, A. & Ravetch, J. V. The role of IgG Fc receptors in antibody-dependent enhancement. *Nat Rev Immunol* **20**, 633–643 (2020).
4. Nimmerjahn, F. & Ravetch, J. V. Fcγ receptors as regulators of immune responses. *Nat Rev Immunol* **8**, 34–47 (2008).
5. Chung, A. W. *et al.* Dissecting Polyclonal Vaccine-Induced Humoral Immunity against HIV Using Systems Serology. *Cell* **163**, 988–998 (2015).
6. Kolda, T. G. & Bader, B. W. Tensor Decompositions and Applications. *SIAM Rev.* **51**, 455–500 (2009).
7. Paar, J. M., Harris, N. T., Holowka, D. & Baird, B. Bivalent Ligands with Rigid Double-Stranded DNA Spacers Reveal Structural Constraints on Signaling by FcεRI1. *The Journal of Immunology* **169**, 856–864 (2002).
8. Hlavacek, W. S. *et al.* Quantifying Aggregation of IgE-FcϵRI by Multivalent Antigen. *Biophysical Journal* **76**, 2421–2431 (1999).
9. Stone, J. D., Cochran, J. R. & Stern, L. J. T-Cell Activation by Soluble MHC Oligomers Can Be Described by a Two-Parameter Binding Model. *Biophysical Journal* **81**, 2547–2557 (2001).
10. Robinett, R. A. *et al.* Dissecting FcγR Regulation through a Multivalent Binding Model. *Cell Systems* **7**, 41-48.e5 (2018).

11. Perelson, A. S. Receptor clustering on a cell surface. III. theory of receptor cross-linking by multivalent ligands: description by ligand states. *Mathematical Biosciences* **53**, 1–39 (1981).

12. Errington, W. J., Bruncsics, B. & Sarkar, C. A. Mechanisms of noncanonical binding dynamics in multivalent protein–protein interactions. *Proceedings of the National Academy of Sciences* **116**, 25659–25667 (2019).

13. Perelson, A. S. & DeLisi, C. Receptor clustering on a cell surface. I. theory of receptor cross-linking by ligands bearing two chemically identical functional groups. *Mathematical Biosciences* **48**, 71–110 (1980).

14. Perelson, A. S. Receptor clustering on a cell surface. II. theory of receptor cross-linking by ligands bearing two chemically distinct functional groups. *Mathematical Biosciences* **49**, 87–110 (1980).

15. Macken, C. A. & Perelson, A. S. *Branching Processes Applied to Cell Surface Aggregation Phenomena*. (Springer Science & Business Media, 2013).

16. Hlavacek, W. S., Posner, R. G. & Perelson, A. S. Steric Effects on Multivalent Ligand-Receptor Binding: Exclusion of Ligand Sites by Bound Cell Surface Receptors. *Biophysical Journal* **76**, 3031–3043 (1999).

17. Piccione, E. C. *et al.* A bispecific antibody targeting CD47 and CD20 selectively binds and eliminates dual antigen expressing lymphoma cells. *mAbs* **7**, 946–956 (2015).

18. Hunter, S. A. & Cochran, J. R. Cell-Binding Assays for Determining the Affinity of Protein–Protein Interactions: Technologies and Considerations. in *Methods in Enzymology* (ed. Pecoraro, V. L.) vol. 580 21–44 (Academic Press, 2016).

19. Tan, Z. C., Orcutt-Jahns, B. T. & Meyer, A. S. A quantitative view of strategies to engineer cell-selective ligand binding. *Integrative Biology* **13**, 269–282 (2021).

20. Csizmar, C. M. *et al.* Multivalent Ligand Binding to Cell Membrane Antigens: Defining the Interplay of Affinity, Valency, and Expression Density. *J. Am. Chem. Soc.* **141**, 251–261 (2019).

21. Rhoden, J. J., Dyas, G. L. & Wroblewski, V. J. A Modeling and Experimental Investigation of the Effects of Antigen Density, Binding Affinity, and Antigen Expression Ratio on Bispecific Antibody Binding to Cell Surface Targets *. *Journal of Biological Chemistry* **291**, 11337–11347 (2016).

22. Caré, B. R. & Soula, H. A. Impact of receptor clustering on ligand binding. *BMC Syst Biol* **5**, 48 (2011).

23. Chapman, S. A. & Asthagiri, A. R. Quantitative effect of scaffold abundance on signal propagation. *Molecular Systems Biology* **5**, 313 (2009).

24. Yang, J. & Hlavacek, W. S. Scaffold-mediated nucleation of protein signaling complexes: Elementary principles. *Mathematical Biosciences* **232**, 164–173 (2011).

25. Levchenko, A., Bruck, J. & Sternberg, P. W. Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proceedings of the National Academy of Sciences* **97**, 5818–5823 (2000).

26. Wu, Y., Vendome, J., Shapiro, L., Ben-Shaul, A. & Honig, B. Transforming binding affinities from three dimensions to two with application to cadherin clustering. *Nature* **475**, 510–513 (2011).

27. Spangler, J. B. *et al.* Antibodies to Interleukin-2 Elicit Selective T Cell Subset Potentiation through Distinct Conformational Mechanisms. *Immunity* **42**, 815–825 (2015).

28. Tan, Z. C., Murphy, M. C., Alpay, H. S., Taylor, S. D. & Meyer, A. S. Tensor-structured decomposition improves systems serology analysis. *Molecular Systems Biology* **17**, e10243 (2021).

29. Deshaies, R. J. Multispecific drugs herald a new era of biopharmaceutical innovation. *Nature* **580**, 329–338 (2020).

30. Dekkers, G. *et al.* Affinity of human IgG subclasses to mouse Fc gamma receptors. *mAbs* **9**, 767–773 (2017).

31. Bruhns, P. *et al.* Specificity and affinity of human Fcγ receptors and their polymorphic variants for human IgG subclasses. *Blood* **113**, 3716–3725 (2009).

32. Deal, B. R. *et al.* Engineering DNA-Functionalized Nanostructures to Bind Nucleic Acid Targets Heteromultivalently with Enhanced Avidity. *J. Am. Chem. Soc.* **142**, 9653–9660 (2020).

33. Lux, A., Yu, X., Scanlan, C. N. & Nimmerjahn, F. Impact of Immune Complex Size and Glycosylation on IgG Binding to Human FcγRs. *J Immunol* **190**, 4315–4323 (2013).

34. Collins, A. & Jackson, K. A Temporal Model of Human IgE and IgG Antibody Function. *Frontiers in Immunology* **4**, (2013).

35. Higel, F., Seidl, A., Sörgel, F. & Friess, W. N-glycosylation heterogeneity and the influence on structure, function and pharmacokinetics of monoclonal antibodies and Fc fusion proteins. *European Journal of Pharmaceutics and Biopharmaceutics* **100**, 94–100 (2016).

36. Olivova, P., Chen, W., Chakraborty, A. B. & Gebler, J. C. Determination of N-glycosylation sites and site heterogeneity in a monoclonal antibody by electrospray quadrupole ion-mobility time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **22**, 29–40 (2008).

37. Mimura, Y. *et al.* Glycosylation engineering of therapeutic IgG antibodies: challenges for the safety, functionality and efficacy. *Protein Cell* **9**, 47–62 (2018).

38. Biburger, M. *et al.* Monocyte Subsets Responsible for Immunoglobulin G-Dependent Effector Functions In Vivo. *Immunity* **35**, 932–944 (2011).

39. Montalvao, F. *et al.* The mechanism of anti-CD20–mediated B cell depletion revealed by intravital imaging. *J Clin Invest* **123**, 5098–5103 (2013).

40. Kerntke, C., Nimmerjahn, F. & Biburger, M. There Is (Scientific) Strength in Numbers: A Comprehensive Quantitation of Fc Gamma Receptor Numbers on Human and Murine Peripheral Blood Leukocytes. *Frontiers in Immunology* **11**, (2020).

41. Bakalar, M. H. *et al.* Size-Dependent Segregation Controls Macrophage Phagocytosis of Antibody-Opsonized Targets. *Cell* **174**, 131-142.e13 (2018).

42. Tang, Y. *et al.* Regulation of Antibody-Dependent Cellular Cytotoxicity by IgG Intrinsic and Apparent Affinity for Target Antigen. *The Journal of Immunology* **179**, 2815–2823 (2007).

43. Lux, A. & Nimmerjahn, F. Of mice and men: the need for humanized mouse models to study human IgG activity in vivo. *J Clin Immunol* **33 Suppl 1**, S4-8 (2013).

44. Crowley, A. R. & Ackerman, M. E. Mind the Gap: How Interspecies Variability in IgG and Its Receptors May Complicate Comparisons of Human and Non-human Primate Effector Function. *Frontiers in Immunology* **10**, (2019).

45. Bruhns, P. Properties of mouse and human IgG receptors and their contribution to disease models. *Blood* **119**, 5640–5649 (2012).

46. Tan, Z. C. & Meyer, A. S. A general model of multivalent binding with ligands of heterotypic subunits and multiple surface receptors. *Mathematical Biosciences* **342**, 108714 (2021).

47. Epstein, B. The Exponential Distribution and Its Role in Life Testing. (1958).

48. Schwab, I., Lux, A. & Nimmerjahn, F. Pathways Responsible for Human Autoantibody and Therapeutic Intravenous IgG Activity in Humanized Mice. *Cell Reports* **13**, 610–620 (2015).

49. Coughlan, A. M. *et al.* Myeloid Engraftment in Humanized Mice: Impact of Granulocyte-Colony Stimulating Factor Treatment and Transgenic Mouse Strain. *Stem Cells and Development* **25**, 530–541 (2016).

50. Siber, G. R., Schur, P. H., Aisenberg, A. C., Weitzman, S. A. & Schiffman, G. Correlation between Serum IgG-2 Concentrations and the Antibody Response to Bacterial Polysaccharide Antigens. *New England Journal of Medicine* **303**, 178–182 (1980).

51. Volkov, M. *et al.* Comprehensive overview of autoantibody isotype and subclass distribution. *J Allergy Clin Immunol* **150**, 999–1010 (2022).

52. Nimmerjahn, F. & Ravetch, J. V. Translating basic mechanisms of IgG effector activity into next generation cancer therapies. *Cancer Immun* **12**, 13 (2012).

53. Dekkers, G. *et al.* Decoding the Human Immunoglobulin G-Glycan Repertoire Reveals a Spectrum of Fc-Receptor- and Complement-Mediated-Effector Activities. *Front. Immunol.* **8**, 877 (2017).

54. de Taeye, S. W. *et al.* FcγR Binding and ADCC Activity of Human IgG Allotypes. *Frontiers in Immunology* **11**, (2020).

55. Georgiev, I. S. *et al.* Delineating Antibody Recognition in Polyclonal Sera from Patterns of HIV-1 Isolate Neutralization. *Science* **340**, 751–756 (2013).

56. Lemke, M. M. *et al.* A systems approach to elucidate personalized mechanistic complexities of antibody-Fc receptor activation post-vaccination. *Cell Reports Medicine* **2**, 100386 (2021).

57. Golay, J. *et al.* Human neutrophils express low levels of FcγRIIIA, which plays a role in PMN activation. *Blood* **133**, 1395–1405 (2019).

58. Lux, A. *et al.* A Humanized Mouse Identifies the Bone Marrow as a Niche with Low Therapeutic IgG Activity. *Cell Reports* **7**, 236–248 (2014).

59. Tan, Z. C. *et al.* Mixed IgG Fc immune complexes exhibit blended binding profiles and refine FcR affinity estimates. *Cell Reports* **42**, (2023).

60. Ge, H., Xu, K. & Ghahramani, Z. Turing: A Language for Flexible Probabilistic Inference. in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* vol. 84 1682–1690.

61. Besançon, M. *et al.* Distributions.jl: Definition and Modeling of Probability Distributions in the JuliaStats Ecosystem. *Journal of Statistical Software* **98**, 1–30 (2021).

62. Whaley III, D. L. The interquartile range: Theory and estimation. (East Tennessee State University, 2005).

63. Mogensen, P. & Riseth, A. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software* **3**, (2018).

64. McLuhan, M. The Medium is the Message. in *Understanding Media: The Extensions of Man* (McGraw-Hill, New York, 1964).

65. Omberg, L. *et al.* Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Mol Syst Biol* **5**, 312 (2009).

66. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat Genet* **48**, 1094–1100 (2016).

67. Li, H. *et al.* The landscape of cancer cell line metabolism. *Nat Med* **25**, 850–860 (2019).

68. Jones, D. S. *et al.* Profiling drugs for rheumatoid arthritis that inhibit synovial fibroblast activation. *Nat Chem Biol* **13**, 38–45 (2017).

69. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat Med* **29**, 1563–1577 (2023).

70. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).

71. Gross, S. M. *et al.* A multi-omic analysis of MCF10A cells provides a resource for integrative assessment of ligand-mediated molecular and phenotypic responses. *Commun Biol* **5**, 1–20 (2022).

72. Durham, T. J., Libbrecht, M. W., Howbert, J. J., Bilmes, J. & Noble, W. S. PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. *Nat Commun* **9**, 1402 (2018).

73. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

74. Kemper, K. E. *et al.* Genetic influence on within-person longitudinal change in anthropometric traits in the UK Biobank. *Nat Commun* **15**, 3776 (2024).

75. Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98**, 1099-1115.e8 (2018).

76. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171 (1997).

77. Rabanser, S., Shchur, O. & Günnemann, S. Introduction to Tensor Decompositions and their Applications in Machine Learning. Preprint at https://doi.org/10.48550/arXiv.1711.10781 (2017).

78. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

79. Yahyanejad, F., Albert, R. & DasGupta, B. A survey of some tensor analysis techniques for biological systems. *Quant Biol* **7**, 266–277 (2019).

80. Caulk, A. W. & Janes, K. A. Robust latent-variable interpretation of in vivo regression models by nested resampling. *Sci Rep* **9**, 19671 (2019).

81. Bro, R. & Smilde, A. K. Centering and scaling in component analysis. *Journal of Chemometrics* **17**, 16–33 (2003).

82. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research* **20**, 1–6 (2019).

83. Bader, B. W. & Kolda, T. G. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Trans. Math. Softw.* **32**, 635–653 (2006).

84. Li, J., Bien, J. & Wells, M. T. rTensor: An R Package for Multidimensional Array (Tensor) Unfolding, Multiplication, and Decomposition. *Journal of Statistical Software* **87**, 1–31 (2018).

85. Helwig, N.E. multiway: Component Models for Multi-Way Data. (2019).

86. Krijnen, W. P., Dijkstra, T. K. & Stegeman, A. On the Non-Existence of Optimal Solutions and the Occurrence of 'Degeneracy' in the CANDECOMP/PARAFAC Model. *Psychometrika* **73**, 431–439 (2008).

87. Wang, M., Fischer, J. & Song, Y. S. Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The Annals of Applied Statistics* **13**, 1103–1127 (2019).

88. Acar, E., Kolda, T. & Dunlavy, D. *An Optimization Approach for Fitting Canonical Tensor Decompositions*. (2009) doi:10.2172/978916.

89. Tomasi, G. & Bro, R. A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics & Data Analysis* **50**, 1700–1734 (2006).

90. Roald, M. *et al.* An AO-ADMM Approach to Constraining PARAFAC2 on All Modes. *SIAM Journal on Mathematics of Data Science* **4**, 1191–1222 (2022).

91. Tomasi, G. & Bro, R. PARAFAC and missing values. *Chemometrics and Intelligent Laboratory Systems* **75**, 163–180 (2005).

92. Acar, E., Dunlavy, D. M., Kolda, T. G. & Mørup, M. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems* **106**, 41–56 (2011).

93. Ahn, M. *et al.* On Large-Scale Dynamic Topic Modeling with Nonnegative CP Tensor Decomposition. in *Advances in Data Science* (eds. Demir, I., Lou, Y., Wang, X. & Welker, K.) vol. 26 181–210 (Springer International Publishing, Cham, 2021).

94. Huang, F. *et al.* Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Briefings in Bioinformatics* **22**, bbaa140 (2021).

95. Sen, B. & Parhi, K. K. Extraction of common task signals and spatial maps from group fMRI using a PARAFAC-based tensor decomposition technique. in *2017 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1113–1117 (2017). doi:10.1109/ICASSP.2017.7952329.

96. Lyu, H., Wan, M., Han, J., Liu, R. & Wang, C. A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining. *Computers in Biology and Medicine* **89**, 264–274 (2017).

97. Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311 (1966).

98. De Lathauwer, L., De Moor, B. & Vandewalle, J. A Multilinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.* **21**, 1253–1278 (2000).

99. Bro, R. & Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics* **17**, 274–286 (2003).

100. Sankaranarayanan, P., Schomay, T. E., Aiello, K. A. & Alter, O. Tensor GSVD of Patient- and Platform-Matched Tumor and Normal DNA Copy-Number Profiles Uncovers Chromosome Arm-Wide Patterns of Tumor-Exclusive Platform-Consistent Alterations Encoding for Cell Transformation and Predicting Ovarian Cancer Survival. *PLoS ONE* **10**, e0121396 (2015).

101. Omberg, L., Golub, G. H. & Alter, O. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences* **104**, 18371–18376 (2007).

102. Acar, E., Bro, R. & Smilde, A. K. Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations. *Proceedings of the IEEE* **103**, 1602–1620 (2015).

103. Acar, E., Kolda, T. G. & Dunlavy, D. M. All-at-once Optimization for Coupled Matrix and Tensor Factorizations. Preprint at https://doi.org/10.48550/arXiv.1105.3422 (2011).

104. Chin, J. L. *et al.* Tensor modeling of MRSA bacteremia cytokine and transcriptional patterns reveals coordinated, outcome-associated immunological programs. *PNAS Nexus* **3**, pgae185 (2024).

105. Schenker, C., Wang, X. & Acar, E. PARAFAC2-based Coupled Matrix and Tensor Factorizations. doi:10.1109/ICASSP49357.2023.10094562.

106. Kreeger, P. K. Using Partial Least Squares Regression to Analyze Cellular Response Data. *Science Signaling* **6**, tr7–tr7 (2013).

107. Bro, R. Multiway calibration. Multilinear PLS. *J. Chemometrics* **10**, 47–61 (1996).

108. Folch-Fortuny, A., Arteaga, F. & Ferrer, A. PLS model building with missing data: New algorithms and a comparative study. *Journal of Chemometrics* **31**, e2897 (2017).

109. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045–1048 (2010).

110. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**, 364–376 (2015).

111. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* **7**, 10812 (2016).

112. Schreiber, J., Durham, T., Bilmes, J. & Noble, W. S. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biology* **21**, 81 (2020).

113. Hunyadi, B., Dupont, P., Van Paesschen, W. & Van Huffel, S. Tensor decompositions and data fusion in epileptic electroencephalography and functional magnetic resonance imaging data. *WIREs Data Mining and Knowledge Discovery* **7**, e1197 (2017).

114. Cong, F. *et al.* Tensor decomposition of EEG signals: A brief review. *Journal of Neuroscience Methods* **248**, 59–69 (2015).

115. Cong, F. *et al.* Multi-domain feature extraction for small event-related potentials through nonnegative multi-way array decomposition from low dense array EEG. *Int. J. Neur. Syst.* **23**, 1350006 (2013).

116. Armingol, E. *et al.* Context-aware deconvolution of cell–cell communication with Tensor-cell2cell. *Nat Commun* **13**, 3665 (2022).

117. 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).

118. Martino, C. *et al.* Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat Biotechnol* **39**, 165–168 (2021).

119. Chitforoushzadeh, Z. *et al.* TNF-insulin crosstalk at the transcription factor GATA6 is revealed by a model that links signaling and transcriptomic data tensors. *Science Signaling* **9**, ra59–ra59 (2016).

120. Netterfield, T. S. *et al.* Biphasic JNK-Erk signaling separates the induction and maintenance of cell senescence after DNA damage induced by topoisomerase II inhibition. *Cell Systems* **14**, 582-604.e10 (2023).

121. Mor, U. *et al.* Dimensionality Reduction of Longitudinal 'Omics Data using Modern Tensor Factorization. (2021) doi:10.1371/journal.pcbi.1010212.

122. Lu, L. L., Suscovich, T. J., Fortune, S. M. & Alter, G. Beyond binding: antibody effector functions in infectious diseases. *Nat Rev Immunol* **18**, 46–61 (2018).

123. Bournazos, S. *et al.* Broadly Neutralizing Anti-HIV-1 Antibodies Require Fc Effector Functions for In Vivo Activity. *Cell* **158**, 1243–1253 (2014).

124. Hessell, A. J. *et al.* Fc receptor but not complement binding is important in antibody protection against HIV. *Nature* **449**, 101–104 (2007).

125. Lofano, G. *et al.* Antigen-specific antibody Fc glycosylation enhances humoral immunity via the recruitment of complement. *Science Immunology* **3**, eaat7796 (2018).

126. Osier, F. H. *et al.* Opsonic phagocytosis of Plasmodium falciparummerozoites: mechanism in human immunity and a correlate of protection against malaria. *BMC Med* **12**, 108 (2014).

127. Joos, C. *et al.* Clinical Protection from Falciparum Malaria Correlates with Neutrophil Respiratory Bursts Induced by Merozoites Opsonized with Human Serum Antibodies. *PLOS ONE* **5**, e9871 (2010).

128. Arnold, K. B. & Chung, A. W. Prospects from systems serology research. *Immunology* **153**, 279–289 (2018).

129. Brown, E. P. *et al.* High-throughput, multiplexed IgG subclassing of antigen-specific antibodies from clinical samples. *Journal of Immunological Methods* **386**, 117–123 (2012).

130. Brown, E. P. *et al.* Multiplexed Fc array for evaluation of antigen-specific antibody effector profiles. *Journal of Immunological Methods* **443**, 33–44 (2017).

131. Mahan, A. E. *et al.* A method for high-throughput, sensitive analysis of IgG Fc and Fab glycosylation by capillary electrophoresis. *Journal of Immunological Methods* **417**, 34–44 (2015).

132. Alter, G. *et al.* High-resolution definition of humoral immune response correlates of effective immunity against HIV. *Mol Syst Biol* **14**, (2018).

133. Zohar, T. *et al.* Compromised Humoral Functional Evolution Tracks with SARS-CoV-2 Mortality. *Cell* **183**, 1508-1519.e12 (2020).

134. Choi, I. *et al.* Machine Learning Methods Enable Predictive Modeling of Antibody Feature:Function Relationships in RV144 Vaccinees. *PLOS Computational Biology* **11**, e1004185 (2015).

135. Pittala, S., Morrison, K. S. & Ackerman, M. E. Systems serology for decoding infection and vaccine-induced antibody responses to HIV-1. *Current Opinion in HIV and AIDS* **14**, 253 (2019).

136. Ackerman, M. E. *et al.* Polyfunctional HIV-Specific Antibody Responses Are Associated with Spontaneous HIV Control. *PLOS Pathogens* **12**, e1005315 (2016).

137. Ackerman, M. E. *et al.* Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. *Nat Med* **24**, 1590–1598 (2018).

138. Ackerman, M. E. *et al.* Natural variation in Fc glycosylation of HIV-specific antibodies impacts antiviral activity. *J Clin Invest* **123**, 2183–2192 (2013).

139. Zhang, X. & Li, L. Tensor Envelope Partial Least-Squares Regression. *Technometrics* **59**, 426–436 (2017).

140. Tan, Z. C. & Meyer, A. S. The structure is the message: Preserving experimental context through tensor decomposition. *Cell Systems* **15**, 679–693 (2024).

141. Choi, D., Jang, J.-G. & Kang, U. S3CMTF: Fast, accurate, and scalable method for incomplete coupled matrix-tensor factorization. *PLOS ONE* **14**, e0217316 (2019).

142. Schüpbach, J. *et al.* Use of HIV-1 p24 as a sensitive, precise and inexpensive marker for infection, disease progression and treatment failure. *Int J Antimicrob Agents* **16**, 441–445 (2000).

143. Rubio Caballero, M. *et al.* The prognostic markers of survival and progression in HIV-1 infection. A study of CD4+ lymphocytes, antigen p24 and viral load during 3 years in a cohort of 251 patients. *An Med Interna* **17**, 533–537 (2000).

144. Sabin, C. A. *et al.* Relationships among the Detection of p24 Antigen, Human Immunodeficiency Virus (HIV) RNA Level, CD4 Cell Count, and Disease Progression in HIV-Infected Individuals with Hemophilia. *The Journal of Infectious Diseases* **184**, 511–514 (2001).

145. Tjiam, M. C. *et al.* Viremic HIV Controllers Exhibit High Plasmacytoid Dendritic Cell–Reactive Opsonophagocytic IgG Antibody Responses against HIV-1 p24 Associated with Greater Antibody Isotype Diversification. *The Journal of Immunology* **194**, 5320–5328 (2015).

146. Tsoukas, C. M. & Bernard, N. F. Markers predicting progression of human immunodeficiency virus-related disease. *Clinical Microbiology Reviews* **7**, 14–28 (1994).

147. Côrtes, F. H. *et al.* HIV Controllers With Different Viral Load Cutoff Levels Have Distinct Virologic and Immunologic Profiles. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **68**, 377 (2015).

148. Merle, N. S., Church, S. E., Fremeaux-Bacchi, V. & Roumenina, L. T. Complement System Part I – Molecular Mechanisms of Activation and Regulation. *Front. Immunol.* **6**, (2015).

149. Candes, E. & Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351 (2007).

150. Efron, B. Prediction, Estimation, and Attribution. *International Statistical Review* **88**, S28–S59 (2020).

151. Tansey, W., Veitch, V., Zhang, H., Rabadan, R. & Blei, D. M. The Holdout Randomization Test for Feature Selection in Black Box Models. *Journal of Computational and Graphical Statistics* **31**, 151–162 (2022).

152. Albert, H., Collin, M., Dudziak, D., Ravetch, J. V. & Nimmerjahn, F. In vivo enzymatic modulation of IgG glycosylation inhibits autoimmune disease in an IgG subclass-dependent manner. *PNAS* **105**, 15005–15009 (2008).

153. Chung, A. W. *et al.* Polyfunctional Fc-Effector Profiles Mediated by IgG Subclass Selection Distinguish RV144 and VAX003 Vaccines. *Science Translational Medicine* **6**, 228ra38-228ra38 (2014).

154. Kaneko, Y., Nimmerjahn, F. & Ravetch, J. V. Anti-Inflammatory Activity of Immunoglobulin G Resulting from Fc Sialylation. *Science* **313**, 670–673 (2006).

155. Larsen, M. D. *et al.* Afucosylated IgG characterizes enveloped viral responses and correlates with COVID-19 severity. *Science* **371**, eabc8378 (2021).

156. Ng, K. W. *et al.* Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. *Science* **370**, 1339–1343 (2020).

157. Angeletti, D. *et al.* Defining B cell immunodominance to viruses. *Nat Immunol* **18**, 456–463 (2017).

158. Sesterhenn, F. *et al.* Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen. *PLOS Biology* **17**, e3000164 (2019).

159. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**, 15545–15550 (2005).

160. Pagan, J. D., Kitaoka, M. & Anthony, R. M. Engineered Sialylation of Pathogenic Antibodies In Vivo Attenuates Autoimmune Disease. *Cell* **172**, 564-577.e13 (2018).

161. Birnbaum, M. E. *et al.* Deconstructing the Peptide-MHC Specificity of T Cell Recognition. *Cell* **157**, 1073–1087 (2014).

162. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* (2011).

163. Defazio, A., Bach, F. & Lacoste-Julien, S. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. in *Advances in Neural Information Processing Systems* vol. 27 (Curran Associates, Inc., 2014).

164. Ellis, J. L. *et al.* Synergy between mechanistic modelling and data-driven models for modern animal production systems in the era of big data. *animal* **14**, s223–s237 (2020).

165. Bekisz, S. & Geris, L. Cancer modeling: From mechanistic to data-driven approaches, and from fundamental insights to clinical applications. *Journal of Computational Science* **46**, 101198 (2020).

166. Procopio, A. *et al.* Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review. *Computer Methods and Programs in Biomedicine* **240**, 107681 (2023).

167. Brigandt, I. & Love, A. Reductionism in Biology. in *The Stanford encyclopedia of philosophy* (2008).

168. Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. (CRC Press, Boca Raton, 2019). doi:10.1201/9780429492563.

169. Chylek, L. A., Harris, L. A., Faeder, J. R. & Hlavacek, W. S. Modeling for (physical) biologists: an introduction to the rule-based approach. *Phys. Biol.* **12**, 045007 (2015).

170. Wynn, M. L., Consul, N., Merajver, S. D. & Schnell, S. Logic-based models in systems biology: a predictive and parameter-free network analysis method. *Integrative Biology* **4**, 1323–1337 (2012).

171. Aittokallio, T. & Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics* **7**, 243–255 (2006).

172. Bournazos, S. *et al.* Antibody fucosylation predicts disease severity in secondary dengue infection. *Science* **372**, 1102–1105 (2021).

173. Bournazos, S., Corti, D., Virgin, H. W. & Ravetch, J. V. Fc-optimized antibodies elicit CD8 immunity to viral respiratory infection. *Nature* **588**, 485–490 (2020).

174. Mitchell, S., Tsui, R., Tan, Z. C., Pack, A. & Hoffmann, A. The NF-κB multidimer system model: A knowledge base to explore diverse biological contexts. *Science Signaling* **16**, eabo2838 (2023).

175. Mills, M. C., Barban, N. & Tropf, F. C. *An Introduction to Statistical Genetic Data Analysis*. (MIT Press, 2020).

176. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **7**, 154096–154113 (2019).

177. Brigandt, I. Systems biology and the integration of mechanistic explanation and mathematical explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **44**, 477–492 (2013).