UNIVERSITY OF CALIFORNIA SAN DIEGO

# Understanding genetic variation in the proteome: a multi-scale structural systems biology toolkit

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

in

Bioinformatics and Systems Biology

by

Nathan Da-Wei Mih

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Philip E. Bourne, Co-Chair
Professor Robert K. Naviaux
Professor Ross C. Walker
Professor Sheng Zhong

2018

The Dissertation of Nathan Da-Wei Mih is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2018

DEDICATION

To Chia-An

To Mom & Dad, and my brothers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

Supplemental File 2.1. Expanded methods and results text and figures detailing GEM-PRO construction, molecular modeling simulations, and systems modeling.

Supplemental File 2.2. GEM-PRO of the human erythrocyte and related pharmacogenomics files.

Supplemental File 2.3. Parameters used in molecular modeling simulations.

ACKNOWLEDGEMENTS

I would like to acknowledge a great number of people for helping me get to this point in my academic career, as it absolutely would not have been possible without them.

First, I would like to thank Professor Bernhard Palsson and his never-ending enthusiasm for scientific discovery and impactful work. My experience in the Systems Biology Research Group (SBRG) has been one of continuous learning and fruitful collaborations. I also owe many thanks to Liz Brunk. As an office mate and structural biology partner-in-crime, we embarked on a journey to not only apply what we were interested in on a larger scale but also introduce many of our lab mates to a three-dimensional world. Liz was an incredible mentor who influenced both my professional and personal development. I would also like to thank Ke Chen for her sage advice and inspiring attention to detail, and Eddie Catoiu for carrying on the torch of structural work in the lab. To the additional members of the "ssbio" group – Roger Chang, Jonathan Monk, Anand Sastry, Justin Tan – this journey would have been much lonelier without their help. To members and associates of the SBRG – Laurence Yang, Ali Ebrahim, Erol Kavvas, Patrick Phaneuf, Marta Matos, Yara Seif, Bin Du, Xin Fang, Colton Lloyd, JC Lachance, Edward O'Brien, Jared Broddrick, James Yurkovich, Aarash Bordbar, Sonal Choudhary, David Heckmann, Zachary King, Daniel Zielinski, Ye Gao, Gabriela Guzman, Andreas Drager, Nathan Lewis, Marc Abrams, Helder Balelo, and everyone else of course – each of them made life much less stressful in many ways.

Many thanks goes to the undergraduates that led my ambitious plan for a senior design project, Richard Xie, Rohan Grover, Winnie Shi, Ruoqing Cheng, Liangyu Zhao, Kritin Karkare, and Ify Aniefuna. The project would also not have come to fruition without the help of Professors Art Olson and David Goodsell, along with much patience from Brett Barbaro. To Professor Phil

Bourne, who took me on in my early years of graduate school and provided me with a smooth transition into Professor Palsson's lab. To my colleagues at the Bourne lab, the Protein Data Bank, and the Godzik lab – Spencer Bliven, Crystal Chia-Yu Ku, Andreas Prlic, Peter Rose, Alexander Rose, and Professor Adam Godzik, thank you for providing incredible structural bioinformatics advice and help along the way. To an unexpected group of colleagues I met in my fifth year as part of the GrAdvantage program – Jenny Kressel, Benjamin Tien, Thomas Brenner, Pratima Rawat, and Riley Peacock. It was an incredibly unique and exciting experience to work with these talented people in starting up a program for dialogue on the UCSD campus.

I would like to thank my dissertation committee – Professors Ross Walker, Robert Naviaux, and Sheng Zhong – for their incredible support and receptive manner over the years. This work also would not have been possible without being funded by the Novo Nordisk Foundation Center for Biosustainability (NNF10CC1016517) and the National Institutes of Health (grant T32GM008806), along with computational resources at the National Energy Research Scientific Computing Center and a travel grant from the Graduate Student Association that allowed me to wander off to Germany for a month.

Finally, a big thanks goes to my family and friends. My parents have always been there to support me every step of the way. Special thanks to my brothers who have inspired me and kept me company wherever we've been. And most of all, thanks to Chia-An Yen (and Lily, Spaghetti, Tangerine, and Potato) – for being my best friend, lifeline, fellow animal lover, party buddy, and much, much more.

Chapter 1, in full, is a reprint of the material as it appears in: Mih N, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, Monk JM, Zhang Z, Palsson BO. 2018. ssbio: A Python Framework

for Structural Systems Biology. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/bty077. The dissertation author was the primary author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Mih N*, Brunk E*, Bordbar A, Palsson BO. 2016. A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism. PLoS Comput. Biol. 12:e1005039. http://dx.doi.org/10.1371/journal.pcbi.1005039. The dissertation author was the primary author (equally contributing with Elizabeth Brunk) of this paper.

Chapter 3, in full, is a reprint of the material as it appears in: Brunk E*, Mih N*, Monk J, Zhang Z, O'Brien EJ, Bliven SE, Chen K, Chang RL, Bourne PE, Palsson BO. 2016. Systems biology of the structural proteome. BMC Syst. Biol. 10:26. http://dx.doi.org/10.1186/s12918-016-0271-6. The dissertation author was the primary author (equally contributing with Elizabeth Brunk) of this paper.

Chapter 4, in full, is a reprint of the material as it appears in: Mih N, Monk JM, Fang X, Catoiu E, Heckmann D, Yang L, Palsson BO. Adaptations of *Escherichia coli* strains to oxidative stress are reflected in properties of their structural proteomes. Submitted. The dissertation author was the primary author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in: Yang L, Mih N, Park JH, Tan J, Seo S, Kim D, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Feist AM, Palsson BO. 2018. Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms. Submitted. The dissertation author was a contributing author of this paper.

VITA

2012 Bachelor of Science, University of California, Irvine

2018 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

**Mih N**, Monk JM, Fang X, Catoiu E, Heckmann D, Yang L, Palsson BO. Adaptations of *Escherichia coli* strains to oxidative stress are reflected in properties of their structural proteomes. Submitted.

Fang X, Monk J, **Mih N**, Du B, Sastry AV, Kavvas E, Seif Y, Smarr L, Palsson BO. *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of human intestinal mucosa. Submitted.

Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Dräger A, Chen K, Sastry AV, Fang X, **Mih N**, Yang L, Eichner J, Cho BK, Kim D, Palsson BO. Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655. Submitted.

Yang L, **Mih N**, Park JH, Tan J, Seo S, Kim D, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Feist AM, Palsson BO. 2018. Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms. Submitted.

Choudhary KS, **Mih N**, Monk J, Kavvas E, Yurkovich JT, Sakoulas G, Palsson BO. 2018. The *Staphylococcus aureus* Two-Component System AgrAC Displays Four Distinct Genomic Arrangements That Delineate Genomic Virulence Factor Signatures. Submitted.

Kavvas E, Catoiu E, **Mih N**, Yurkovich JT, Seif Y, Dillon N, Heckmann D, Anand A, Yang L, Nizet V, Monk JM, Palsson BO. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. Submitted.

Feher VA, Schiffer JM, Mermelstein D, **Mih N**, Pierce LCT, McCammon JA, Amaro RE. Mechanisms for Benzene Dissociation through the Excited State of T4 Lysozyme L99A. Submitted.

Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, **Mih N**, Gatto F, Nilsson A, Preciat Gonzalez GA, Aurich MK, Prlić A, Sastry A, Danielsdottir AD, Heinken A, Noronha A, Rose PW, Burley SK, Fleming RMT, Nielsen J, Thiele I, Palsson BO. 2018. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol*. http://dx.doi.org/10.1038/nbt.4072.

**Mih N**, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, Monk JM, Zhang Z, Palsson BO. 2018. ssbio: A Python Framework for Structural Systems Biology. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/bty077.

Chen K, Gao Y, **Mih N**, O'Brien EJ, Yang L, Palsson BO. 2017. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proceedings of the National Academy of Sciences.* 114:11548–11553. http://www.pnas.org/content/114/43/11548.abstract.

Monk JM, Lloyd CJ, Brunk E, **Mih N**, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, Feist AM, Palsson BO. 2017. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol*. 35:904–908. http://dx.doi.org/10.1038/nbt.3956.

Fang X, Sastry A, **Mih N**, Kim D, Tan J, Yurkovich JT, Lloyd CJ, Gao Y, Yang L, Palsson BO. 2017. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences.* http://www.pnas.org/content/early/2017/08/30/1702581114.abstract.

Broddrick JT, Rubin BE, Welkie DG, Du N, **Mih N**, Diamond S, Lee JJ, Golden SS, Palsson BO. 2016. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proceedings of the National Academy of Sciences.* 113:E8344–E8353. http://dx.doi.org/10.1073/pnas.1613446113.

**Mih N***, Brunk E*, Bordbar A, Palsson BO. 2016. A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism. *PLoS Comput. Biol.* 12:e1005039. http://dx.doi.org/10.1371/journal.pcbi.1005039.

Brunk E*, **Mih N***, Monk J, Zhang Z, O'Brien EJ, Bliven SE, Chen K, Chang RL, Bourne PE, Palsson BO. 2016. Systems biology of the structural proteome. *BMC Syst. Biol.* 10:26. http://dx.doi.org/10.1186/s12918-016-0271-6.

Bruegger J, Haushalter B, Vagstad A, Shakya G, **Mih N**, Townsend CA, Burkart MD, Tsai S-C. 2013. Probing the Selectivity and Protein·Protein Interactions of a Nonreducing Fungal Polyketide Synthase Using Mechanism-Based Crosslinkers. *Chem. Biol.* 20:1135–1146. http://www.cell.com/article/S1074552113002792/abstract.

ABSTRACT OF THE DISSERTATION

Understanding genetic variation in the proteome: a multi-scale structural systems biology toolkit

by

Nathan Da-Wei Mih

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2018

Professor Bernhard Ø. Palsson, Chair
Professor Philip E. Bourne, Co-Chair

The computational representation of metabolic networks has traditionally abstracted the representation of enzymatic components that catalyze their associated reactions. However, the long history of reductionism in studying biological components, specifically the three-dimensional structure of proteins, has resulted in a rich knowledgebase of experimental data and computational tools capable of describing atomic-level biochemical interactions and inferring functional consequences. The challenge of integrating such information into large-scale computational models of cells results in the intersection of structural and systems biology. This dissertation aims

to explore and address the methodological issues that arise from the integration of these two fields, centered around the creation of a computational framework to answer the question of how best to utilize structural data in genome-scale models of metabolism. I present a software framework, *ssbio*, and an associated pipeline to simplify the integration process and allow the construction of high-quality genome-scale models with protein structures (GEM-PROs). The framework allows for the rigorous consideration and consolidation of the often repetitive or incomplete data of proteins in 3D space. Next, I explore a deep integration of molecular modeling tools into metabolic models to understand the impact of non-synonymous mutations on protein-ligand interactions within the human erythrocyte. I then show how these models provide utility in a comparative analysis of *Escherichia coli* and *Thermotoga maritima* by elucidating the impact of the structural proteome on the temperature dependence of growth, the distribution of protein fold families, and substrate specificity. Finally, I extend this framework to understand if hypothesized adaptations to oxidative stress are reflected in the structural proteomes of multiple strains of *E. coli*, along with the incorporation of a probabilistic model of oxidative damage based on selected structural features into a model of metabolism and protein synthesis. Overall, these studies represent the utility of a multi-scale approach to understanding how changes to the structural proteome can impact the system they participate in, along with providing the basis for future computational studies in structural systems biology.

# Chapter 1.

# ssbio: A Python Framework for Structural Systems Biology

## 1.1 Abstract

Working with protein structures at the genome-scale has been challenging in a variety of ways. Here, we present *ssbio*, a Python package that provides a framework to easily work with structural information in the context of genome-scale network reconstructions, which can contain thousands of individual proteins. The *ssbio* package provides an automated pipeline to construct high quality genome-scale models with protein structures (GEM-PROs), wrappers to popular third-party programs to compute associated protein properties, and methods to visualize and annotate structures directly in Jupyter notebooks, thus lowering the barrier of linking 3D structural data with established systems workflows. *ssbio* is implemented in Python and available to download under the MIT license at http://github.com/SBRG/ssbio. Documentation and Jupyter notebook tutorials are available at http://ssbio.readthedocs.io/en/latest/. Interactive notebooks can be launched using Binder at https://mybinder.org/v2/gh/SBRG/ssbio/master?filepath=Binder.ipynb.

## 1.2 Introduction

Merging the disciplines of structural and systems biology remains promising in a variety of ways, but differences in the fields present a learning curve for those looking toward this

integration within their own research. Beltrao et al. stated it best, that "apparently structural biology and systems biology look like two different universes" [1]. A great number of software tools exist within the structural bioinformatics community [2-6], and with recent advances in structure determination techniques, the number of experimental structures in the Protein Data Bank (PDB) continues to steadily rise [7]. The challenges of integrating external data and software tools into systems analyses have been detailed [8], and structural information is no exception to the norm. At the systems-level, curated network models such as genome-scale metabolic models (GEMs) provide a context for molecular interactions in a functional cell [9]. Recently, GEMs integrated with protein structures (GEM-PROs) have extended these models to explicitly utilize 3D structural data alongside modeling methods to substantiate a number of hypotheses, as we explain below.

Here, we present *ssbio*, a Python package designed with the goal of lowering the learning curve associated with efforts in structural systems biology. *ssbio* directly integrates with and builds upon the COBRApy toolkit [10] allowing for seamless integration with existing GEMs. The core functionality of *ssbio* is additionally extended by hooks to many popular third-party structural bioinformatics algorithms, such as DSSP, MSMS, SCRATCH, I-TASSER, and others (see Tables 1 and 2 for a full list) [11-14].

Table 1.1. A summary of currently available functionalities for a protein sequence contained in ssbio.
External software or web servers are listed if the functionality is dependent on it. If there are software listed as an alternate, then that means that there exists either another third-party program or a Python-based method that carries out the same function. For example, the Biopython pairwise2 function carries out a pairwise sequence alignment, but can also be run using the EMBOSS needle package. This table can also be found in the documentation at: http://ssbio.readthedocs.io/en/latest/sequence.html

| **Analysis type** | **Function** | **Description** | **Internal ssbio class or function** | **External software** | **Web server** | **Alternate external software** |
|---|---|---|---|---|---|---|
| Sequence-based predictions | Secondary structure and solvent accessibilities | Predictions of secondary structure and relative solvent accessibilities per residue | scratch module | SCRATCH [13] | | |
| | Thermostability | Free energy of unfolding (ΔG), adapted from Oobatake [15] and Dill [16] | thermostability module | | | |
| | Transmembrane domains | Prediction of transmembrane domains from sequence | tmhmm module | TMHMM [17] | | |
| | Aggregation propensity | Consensus method to predict the aggregation propensity of proteins, specifically the number of aggregation-prone segments on an unfolded protein sequence | aggregation propensity module | | AMYLPRED2 [18] | |
| Sequence-based calculations | Various sequence properties | Basic properties of the sequence, such as percent of polar, non-polar, hydrophobic or hydrophilic residues. | Biopython ProteinAnalysis [19] sequence residues module | | | EMBOSS pepstats [20] |
| | Sequence alignment | Basic functions to run pairwise or multiple sequence alignments | Biopython pairwise2 [19] alignment module | | | EMBOSS needle [20] |

Table 1.2. A summary of currently available functionalities for a protein structure contained in ssbio.
Residue-level properties from structures are linked to the correct residue numbering schemes in mapped sequences, or vice-versa. As in Table 1, external software, web servers and alternates are provided in the rightmost columns. This table can also be found in the documentation at: http://ssbio.readthedocs.io/en/latest/structure.html

| **Analysis type** | **Function** | **Description** | **Internal ssbio class or function** | **External software** | **Web server** | **Alternate external software** |
|---|---|---|---|---|---|---|
| Sequence-based prediction | Homology modeling | Preparation scripts and parsers for executing homology modeling algorithms | `itasserprep module` `itasserprop module` | I-TASSER [14] | | |
| Structure-based prediction | Transmembrane orientation | Prediction of transmembrane domains and orientation in a membrane | `opm module` | | OPM [21] | |
| | Kinetic folding rate | Prediction of protein folding rates from amino acid sequence | `kinetic folding rate module` | | FOLD-RATE [22] | |
| Structure-based calculation | Secondary structure | Calculations of secondary structure | Biopython DSSP [3] `dssp module` `stride module` | DSSP [11], | | STRIDE [23] |
| | Solvent accessibilities | Calculations of per-residue absolute and relative solvent accessibilities | Biopython DSSP [3] `dssp module` `freesasa module` | DSSP [11], | | FreeSASA [24] |
| | Residue depths | Calculations of residue depths | Biopython ResidueDepth [3] `msms module` | MSMS [12] | | |
| | Structural similarity | Pairwise calculations of 3D structural similarity | `fatcat module` | FATCAT [25] | | |
| | Quality | Custom functions to allow ranking of structures by percent identity to a defined sequence, structure resolution, and other structure quality metrics | `set representative structure function` | | | |
| | Various structure properties | Basic properties of the structure, such as distance measurements between residues or number of disulfide bridges | `structure residues module` | Biopython Struct [3] | | |
| Structure-based function | Structure cleaning, mutating | Custom functions to allow for the preparation of structure files for molecular modeling, with options to remove hydrogens/waters/heteroatoms, select specific chains, or mutate specific residues. | Biopython Select [3] `cleanpdb module` `mutatepdb module` | | | AmberTools [26] |

## 1.3 Functionality

### 1.3.1 Protein class

*ssbio* adds a `Protein` class as an attribute to a COBRApy `Gene` and is representative of the gene's translated polypeptide chain (Figure 1A). A `Protein` holds related amino acid sequences and structures, and a single representative sequence and structure can be set from these. This simplifies network analyses by enabling the properties of all or a subset of proteins to be computed and subsequently queried for. For details on these properties, as well as installation and execution instructions for the third-party software used to compute them, please see the documentation. Additionally, proteins with multiple structures available in the PDB can be subjected to QC/QA based on set cutoffs such as sequence coverage and X-ray resolution. Proteins with no structures available can be prepared for homology modeling through platforms such as I-TASSER [14]. Biopython representations of sequences (`SeqRecord` objects) and structures (`Structure` objects) are utilized to allow access to analysis functions available for their respective objects (Figure 1B) [19]. Finally, all information contained in a `Protein` (or in the context of a network model, multiple proteins) can be saved and shared as a JavaScript Object Notation (JSON) file.

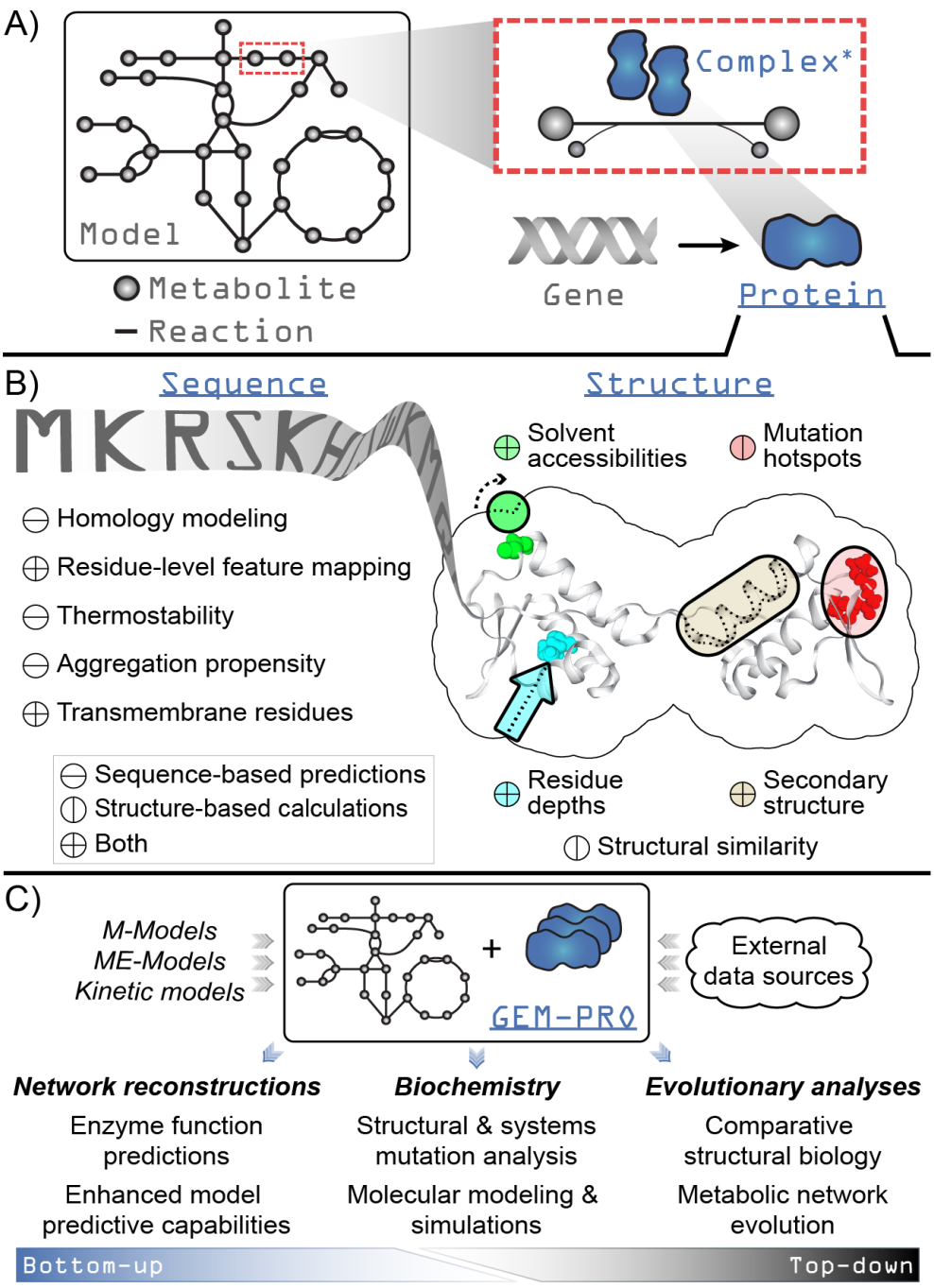



Figure 1.1. Overview of the design and functionality of ssbio.
Underlined fixed-width text in blue indicates added functionality to COBRApy for a genome-scale model loaded using ssbio. A) A simplified schematic showing the addition of a Protein to the core objects of COBRApy (fixed-width text in gray). A gene is directly associated with a protein, which can act as a monomeric enzyme or form an active complex with itself or other proteins (the asterisk denotes that methods for complexes are currently under development). B) Summary of functions available for computing properties on a protein sequence or structure. C) Uses of a GEM-PRO, from the bottom-up and the top-down. Once all protein sequences and structures are mapped to a genome-scale model, the resulting GEM-PRO has uses in multiple areas of study, as noted in the main text.

### 1.3.2 GEM-PRO pipeline

The objectives of the GEM-PRO pipeline have previously been detailed [27]. A GEM-PRO directly integrates structural information within a curated GEM (Figure 1C), and streamlines identifier mapping, representative object selection, and property calculation for a set of proteins. The pipeline provided in *ssbio* functions with an input of a GEM, but alternatively works with a list of gene identifiers or their protein sequences if network information is unavailable.

The added context of manually curated network interactions to protein structures enables different scales of analyses. For instance, from the top-down, global non-variant properties of protein structures such as the distribution of fold types can be compared within or between organisms [27-29]. From the bottom-up, structural properties predicted from sequence or calculated from structure can be utilized to guide a metabolic reconstruction [30] or to enhance model predictive capabilities [31-34]. Looking forward, applications to multi-strain modelling techniques [35-37] would allow strain-specific changes to be investigated at the molecular level, potentially explaining phenotypic differences or strain adaptations to certain environments.

### 1.3.3 Scientific analysis environment

We provide a number of Jupyter notebook tutorials to demonstrate analyses at different scales (i.e. for a single protein sequence or structure, set of proteins, or network model). These notebooks can be launched in a virtual environment through the Binder project (https://mybinder.org/), with most third-party software pre-installed so users can immediately run through tutorials and experiment with them. Certain data can be represented as Pandas DataFrames [38], enabling quick data manipulation and graphical visualization. These notebooks are further extended by visualization tools such as NGLview for interacting with and annotating 3D structures [39, 40], and Escher for constructing and viewing biological pathways [41].

### 1.3.4 Attribute access

In the context of a COBRApy `Model` loaded with *ssbio* (`my_model` in the example below), the protein and its main attributes can be accessed simply by:

```
from ssbio.pipeline.gempro import GEMPRO
my_model = GEMPRO(gem_name='si_demo', gem_file_path='/path/to/gem.xml',
                  gem_file_type='sbml')
my_gene_id = 'geneA'
my_protein = my_model.genes.get_by_id(my_gene_id).protein
my_protein.sequences  # Contains a list of stored amino acid sequences
my_protein.structures  # Contains a list of stored 3D structures
my_protein.representative_sequence  # Single representative amino acid sequence
my_protein.representative_structure  # Single representative 3D structure
```

### 1.3.5 Package organization

*ssbio* is organized into the following submodules for defined purposes. A table of currently available functionalities for protein sequences and structures can be found in Tables 1 and 2.

1. `ssbio.databases`, modules that heavily depend on the Bioservices package [42] and custom code to enable pulling information from web services such as UniProt, KEGG, and the PDB, and to directly convert that information into sequence and structure objects to load into a protein.
2. `ssbio.core`, modules which represent objects that make up some of the contents of a cell systems model (genes, proteins and protein complexes).
3. `ssbio.protein.sequence`, modules which allow a user to execute and parse sequence-based utilities such as sequence alignment algorithms or structural feature predictors.
4. `ssbio.protein.structure`, modules that mirror the sequence module but instead work with structural information to calculate properties, and also to streamline the generation of homology models as well as to prepare structures for molecular modeling tools such as docking or molecular dynamics.

5. `ssbio.pipeline.gempro`, a pipeline that simplifies the execution of these tools per protein while placing them into the context of a genome-scale model.
6. `ssbio.viz`, modules which focus on the 3D visualization of protein structures and network representations of systems models.

### 1.3.6 Cached files

Files such as sequences, structures, alignment files, and property calculation outputs can optionally be cached on a user's disk to minimize calls to web services, limit recalculations, and provide direct inputs to common sequence and structure algorithms which often require local copies of the data. For a GEM-PRO project, files are organized in the following fashion once a root directory and project name are set:

```
<ROOT_DIR>
└── <PROJECT_NAME>
      ├── data  # General directory for pipeline outputs
      ├── model  # SBML and GEM-PRO models are stored in this directory
      └── genes  # Per gene information
            └── <gene_id1>  # Specific gene directory
                  └── <protein_id1>  # Protein directory
                        ├── sequences  # Protein sequence files, alignments
                        └── structures  # Protein structure files, calculations
```

## 1.4 Conclusion

*ssbio* provides a Python framework for systems biologists to start thinking about detailed molecular interactions and how they impact their models, and enables structural biologists to scale up and apply their expertise to multiple enzymes working together in a system. Towards a vision of whole-cell *in silico* models, structural information provides invaluable molecular-level details, and integration remains crucial.

## 1.5 Funding

This work was supported by the Novo Nordisk Foundation Center for Biosustainability [NNF10CC1016517 to N.M., K.C., E.C., and A.S.]; the Swiss National Science Foundation [p2elp2_148961 to E.B.]; and the National Institute of General Medical Sciences of the National Institutes of Health [U01-GM102098 to B.O.P., 1-U01-AI124316-01 to J.M.M. and E.K.].

## 1.6 Acknowledgements

We would like to thank Dr. Zachary King, Patrick Phaneuf, Marta Matos, and Colton Lloyd for valuable discussions in software development, and Dr. Laurence Yang, Yara Seif, JC Lachance, and Jared Broddrick for insight into desired functionalities of the package. We would also like to thank Marc Abrams for proofreading of the manuscript.

Chapter 1, in full, is a reprint of the material as it appears in: Mih N, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, Monk JM, Zhang Z, Palsson BO. 2018. ssbio: A Python Framework for Structural Systems Biology. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/bty077. The dissertation author was the primary author of this paper.

# Chapter 2.

# A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism

## 2.1 Abstract

Progress in systems medicine brings promise to addressing patient heterogeneity and individualized therapies. Recently, genome-scale models of metabolism have been shown to provide insight into the mechanistic link between drug therapies and systems-level off-target effects while being expanded to explicitly include the three-dimensional structure of proteins. The integration of these molecular-level details, such as the physical, structural, and dynamical properties of proteins, notably expands the computational description of biochemical network-level properties and the possibility of understanding and predicting whole cell phenotypes. In this study, we present a multi-scale modeling framework that describes biological processes which range in scale from atomistic details to an entire metabolic network. Using this approach, we can understand how genetic variation, which impacts the structure and reactivity of a protein, influences both native and drug-induced metabolic states. As a proof-of-concept, we study three enzymes (catechol-O-methyltransferase, glucose-6-phosphate dehydrogenase, and

glyceraldehyde-3-phosphate dehydrogenase) and their respective genetic variants which have clinically relevant associations. Using all-atom molecular dynamic simulations enables the sampling of long timescale conformational dynamics of the proteins (and their mutant variants) in complex with their respective native metabolites or drug molecules. We find that changes in a protein's structure due to a mutation influences protein binding affinity to metabolites and/or drug molecules, and inflicts large-scale changes in metabolism.

## 2.2 Introduction

Synergistic advances in pharmacogenomics, genome-wide association studies (GWAS) and next-generation sequencing bring promise to future applications of personalized medicine. Exploring the mechanistic link between human sequence variation and responses to drug therapy is likely to shed light on why certain drugs show a reduced or even harmful effect on specific individuals. For example, if an individual has a specific polymorphism or rare variant, the consequences of administering a given drug are potentially immense if a life-threatening gene-drug association has not yet been identified [43]. While numerous harmful gene-drug associations have been identified from GWAS (and those with significant side effects now have warnings on pharmaceutical labels [44]), screening genome-wide associations across the broad scope of available pharmaceutical compounds is currently limited by both the cost of carrying out such studies [45] as well as a lack of statistical power due to the rarity of deleterious mutations.

To address these limitations, a number of recent studies have developed mechanistic, computational analyses and the construction of omics-based workflows that identify, for example, the mode of action of common drug side effects [46]. Genome-scale modeling enables the analysis of disease-causing mutations in mechanistic detail. Genome-scale models of metabolism (GEMs)

encompass the known interactions of diverse biological components, or the reactome of a target organism, into a unified, functional framework. This framework contains all known metabolic reactions, the genes that encode each enzyme, and all metabolites in a given organism and therefore provides a direct mapping from genes, to gene products, to the phenotypic responses of cellular activity. Mapping sequence variations in a gene to changes in the biological states of an entire metabolic network enables characterizing the effects of sequence variation in simplified cellular systems, such as the human erythrocyte [47, 48]. Furthermore, a recently updated version of the erythrocyte metabolic model (*i*AB-RBC-283), based on the global reconstruction of the human metabolic network (Recon 2) [49] has been used to study the response of the cell to deleterious single nucleotide polymorphisms (SNPs) as well as drugs with known targets [47, 50, 51].

Predicting the wide range of possible effects that SNPs and single nucleotide variations (SNVs) can have on structure-function relationships in proteins requires extending a systems-level description to include details from physics-based approaches, such as molecular dynamics simulations. To this end, three-dimensional structures of proteins provide complementary data for further elucidating changes in drug-protein interaction networks. Much attention has been placed on developing bioinformatics tools for the statistical analysis of large-scale data sets, (which contain information on non-synonymous, exonic mutations on individual proteins), and generating hypotheses that explain how mutations affect stability, protein-protein interactions, ligand binding, or catalytic function [52]. Atomistic simulations have been used as a complement to experimental methods to assess changes in relative binding affinities of potential lead compounds to key enzymatic targets [53]. While these approaches are rich in molecular-level details, they are limited in their ability to address *how* significant the observed changes are in the context of an entire biochemical pathway or, ultimately, a whole cell. This limitation thus motivates the need to

develop novel workflows that integrate systems-level and molecular-level details to characterize biological processes at graded levels of chemical detail [54-56].

The growing field of structural systems biology brings promise to the integration of systems and molecular sciences, enabling applications in personalized medicine [55, 57-59], drug discovery [60-62], understanding off target binding [31, 63, 64] or mechanisms of action, [65-67] and also to enhance pharmacokinetic/pharmacodynamic models [68]. Here, we build upon previous studies which integrate protein structural information into GEMs [27, 31, 64], by developing a multi-scale framework to analyze the effects of sequence variation on drug responses in human erythrocyte metabolism (Figure 1). Using genome-scale modeling approaches, we identify key proteins in erythrocyte metabolism that are perturbed in the presence of (i) pharmaceutical drugs and (ii) sequence variants. Using atomistic simulations, we characterize changes in structure and function relationships for different metabolic proteins in the form of drug or metabolite binding differences resulting from reported sequence variants. Finally, we integrate the knowledge gained from these simulations into a detailed genome-scale model of the erythrocyte, allowing for both constraint-based and kinetic methods of analysis to understand the systems-wide effect of these variants.

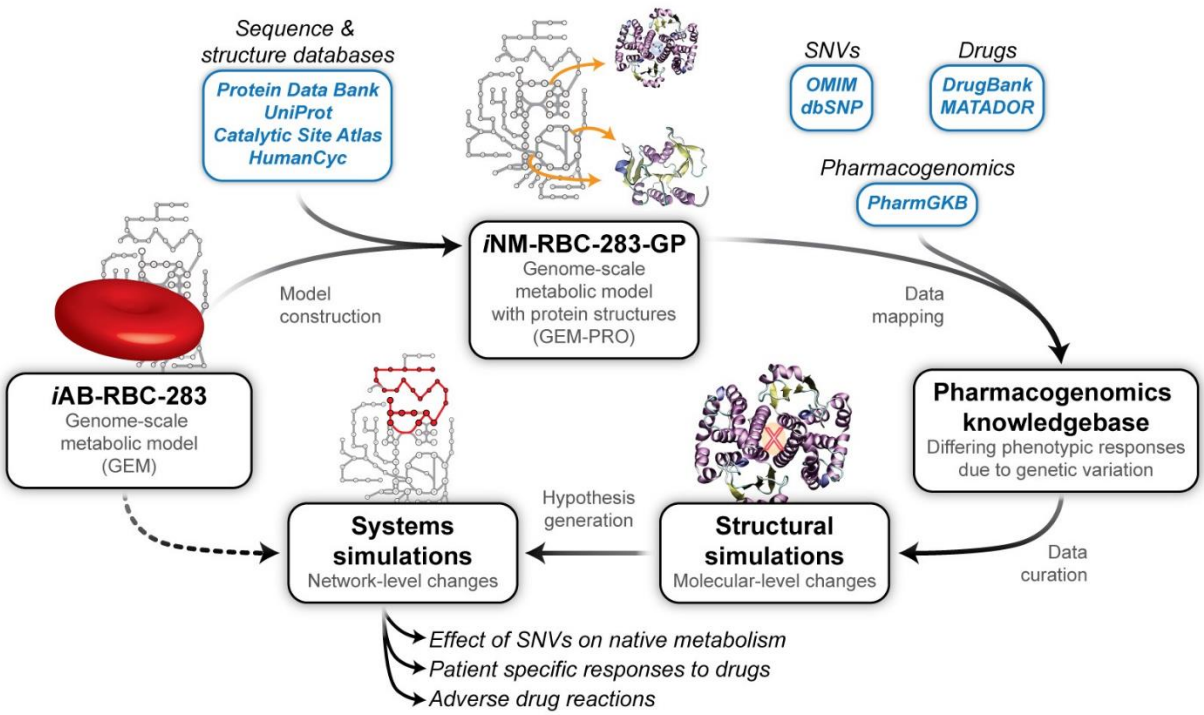



Figure 2.1. A novel workflow for advancing systems pharmacology.
Starting from the genome-scale model of human erythrocyte metabolism (*i*AB-RBC-283 [50]), we integrate information from sequence and structure databases, such as UniProt [69] and the Protein Data Bank (PDB) [70]. Using information from the PDB, experimental protein structures are linked to their respective encoding genes and interacting partners in the metabolic networks. Using homology modeling, representative templates are used to build structural models of target proteins when existing experimental structural information is sparse or missing. The resulting GEnome-scale model of Metabolism with PROtein structures, GEM-PRO, (referred to as *i*NM-RBC-283-GP), presents all of this information in a single database and can be used to generate hypotheses related to cell function in the presence of environmental perturbations. Using other external databases such as the PharmGKB [71], information about known SNPs, drug-related effects, and pharmacogenomic data is used to find promising protein targets that are characterized at the molecular level. Finally, the information gained from structural simulations (e.g. substrate docking and molecular dynamics simulations) can be used as input to guide systems modeling and test hypotheses related to drug-induced effects on metabolism.

## 2.3 Results and Discussion

### 2.3.1 Pharmacogenomics in the human erythrocyte

We were interested in quantifying the number of proteins in the human erythrocyte metabolism that (i) are known pharmaceutical targets and (ii) have been documented with both disease and non-disease causing mutations (Figure 2(a)). The erythrocyte presents a valuable and

tractable model system for studying the effects of human genetic variation on drug metabolism. First, it is widely appreciated that the erythrocyte possesses drug metabolizing capabilities such that extracts of erythrocyte enzymes are commonly used as a general measure of enzyme activity [72, 73]. Second, genetic changes that occur in cells other than the erythrocyte are often manifested in the erythrocyte, assuming correct isoforms and similar genetic control [74-77]. The ease of collection of human erythrocyte samples and subsequent purification of enzymes of interest motivates the study of the erythrocyte as an *in silico* model that can be tested against. Lastly, the erythrocyte outnumbers any other cell type in the human body (85% of the total cell count) [78].

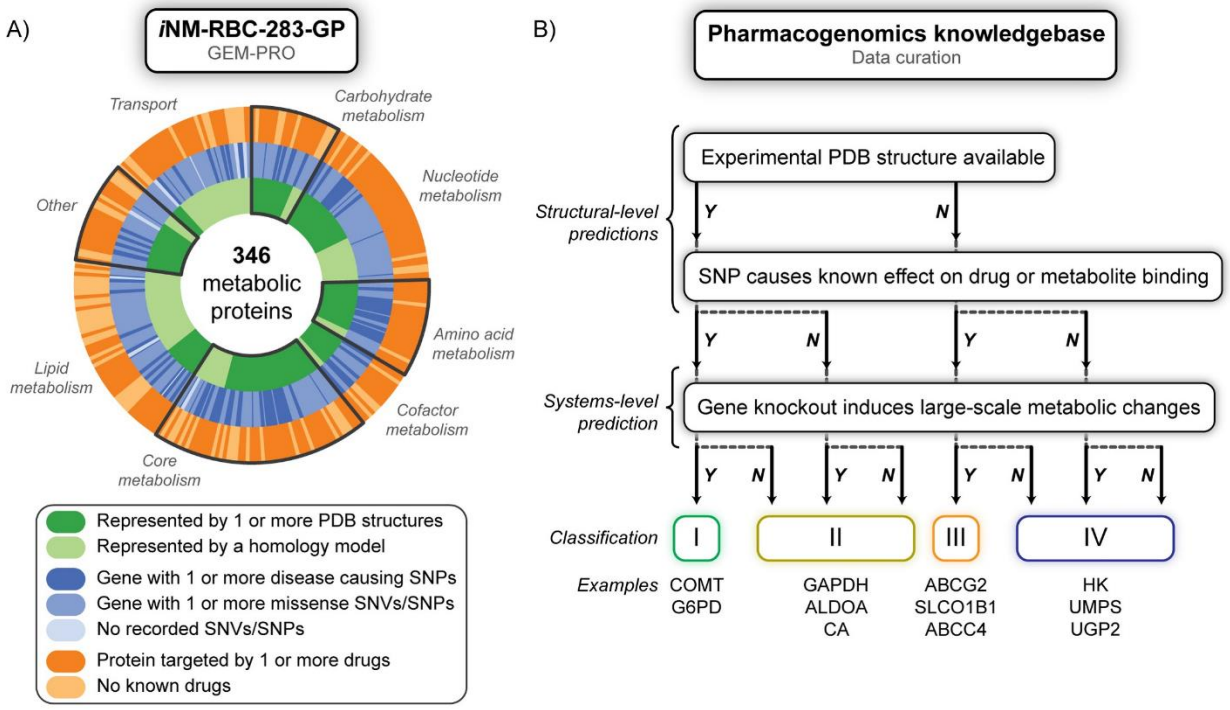



Figure 2.2. GEM-PRO and pharmacogenomics knowledgebase for the human erythrocyte.
In a), coverage of structural and pharmacogenomics information for the human erythrocyte. The metabolic network is based on 346 proteins, and each narrow slice of the pie chart represents one protein. The innermost circle represents structural coverage by an experimental structure (dark green) or by a homology model (light green). The middle circle indicates if the gene is known to contain at least one disease causing SNP (dark blue), at least one missense SNV or SNP (blue), or no recorded SNVs/SNPs (light blue). The outermost circle includes information from various drug databases, and indicates if that protein is known to be a drug or drug metabolite target (dark orange) or if no drugs target that protein (light orange). Basic subsystems of erythrocyte metabolism are highlighted as regions of the chart. For a full chart of numeric counts for each category and subsystem division see Fig C in Supplemental File 2.1. In b), pharmacogenomics knowledge base generation. Our knowledge base includes information on: drugs or metabolites that are predicted to bind to/are metabolized by a protein; known associations between a drug and variation within a population; all variation sites that alter the sequence of the protein target. Targets are filtered into four classes based on if there is a protein structure available, if a SNP causes known effects on drug or metabolite catalysis or binding, and finally if the protein itself is important within the context of the import and export of metabolites in the erythrocyte from gene knockout simulations and flux variability analysis (FVA). Included at the bottom are examples of genes that match these classes of information.

Starting from the set of metabolic genes in the genome-scale model, *i*AB-RBC-283 [50], we mapped gene identifiers to cross-referenced information from dbSNP [79], OMIM [80], and UniProt [81]. We find that for 6800 exon coding SNPs in genes which are expressed in the erythrocyte, the majority (>90%) are missense SNPs as opposed to frameshift or insertion/deletion variations. These SNPs map to 247 of the 281 genes (88%) in the erythrocyte model. The majority of these annotated as "disease-causing" map to enzymes within the heme biosynthesis, glycolysis, and galactose metabolism pathways, which is consistent with hemolytic dysfunction. Other non-disease causing SNPs, (or SNPs with unknown associations), occur in nucleotide metabolism. Harmful mutations also tend to alter the type of amino acid much more than non-disease causing SNPs. For instance, mutations from a hydrophobic residue to another hydrophobic residue are quite common, but disease causing SNPs greatly increase this type of amino acid change to a polar, non-polar, or positive amino acid (Fig D in Supplemental File 2.1).

Our pipeline also identifies variants that potentially influence drug-binding capabilities of respective proteins. Of the metabolic proteins in the erythrocyte, 143 are found to be potential targets for pharmaceutical action. We find 343 drugs (approved, experimental, withdrawn drugs, or drug metabolites) that bind to different proteins in the model [82, 83]. In addition, mapping to the PharmGKB database, we find 274 deleterious SNP-drug associations, or documented adverse reactions (i.e., pharmaceutical complications) in patients (referred to herein as SNP-drug association). To summarize, our systems pharmacological database provides details on all documented missense SNPs in erythrocyte metabolism, whether they are causal for disease or cause pharmaceutical complications in a significant percentage of the human population with a sequence variation [71]. In addition, our dataset contains information on drug-binding capabilities of all proteins in the model. This combined source of information for genetic and pharmacological

information within the erythrocyte allows for the selection of interesting targets to further analyze with both molecular and systems simulations.

### 2.3.2 Mapping protein structures to the metabolic network of the human erythrocyte

To address the structural implications of changes to sequence or drug-binding capacity, we were interested in mapping all protein-encoding genes within the metabolic network of the erythrocyte to their three-dimensional (3D) macromolecular structures. Integration of protein structural data and GEMs has previously been described through the construction of GEnome-scale models of Metabolism with PROtein structures (GEM-PRO). The established pipelines for constructing a GEM-PRO have been recently updated [27]. Applying this procedure for the human erythrocyte metabolic model, we start from the existing GEM, *i*AB-RBC-283 [50], and the final outcome is a mapping of all protein-encoding genes to the 3D structures of their catalyzing enzymes. The selected protein structures have been quality-controlled and ranked to ensure the highest quality structures are retained. The new GEM-PRO model, *i*NM-RBC-283-GP, initially contained structural coverage for 181 of the 346 proteins in the metabolic network (Figure 2(a)), and includes a total of 1766 unique PDB entries (the original GEM is comprised of 281 genes which encode 346 unique proteins). In addition, 312 homology models were obtained for proteins from existing homology model databases [84], using the I-TASSER suite of programs [14].

Our QC/QA pipeline identifies experimental structures and homology models that can be used with high confidence in molecular modeling simulations [27]. Several quality metrics are used to rank-order structures, including: (i) coverage of the wild-type amino acid sequence (with a wild-type being defined as the canonical UniProt sequence); (ii) X-ray structure resolution; (iii) number of missing or unresolved parts of the structure. The final QC/QA statistics indicate that 36% of proteins in the GEM model (125/346) have high quality structural information, whereas

the remaining 64% (221/346 proteins) can be represented by template-based and *ab initio* generated homology models (see Fig C in Supplemental File 2.1 for detailed statistics on subsystem coverage).

Interestingly, when we combine the structural data and the pharmacogenomic data, we are able to assess SNP data in the context of protein structural information and derive new association. For example, we find that, on average, disease causing SNPs are 4 Å closer to annotated enzyme active sites than non-disease causing SNPs. All structural annotations, mapped database information, and quality statistics are included as a supplementary database (Supplemental File 2.2).

### 2.3.3 Identifying signature proteins with disease phenotypes

One of the main advantages of assembling a structural systems pharmacological dataset for the erythrocyte is that it can be used to address questions requiring multi-scale perspectives, such as "Can mutating a single amino acid in a protein influence network-level perturbations, and, ultimately lead to disease phenotypes?" Considering the availability of information (pharmacogenomic and structural) that emerged from our mapping efforts, we were interested in focusing on several specific cases that could be studied in greater molecular detail, using a combined systems and molecular modeling approach.

To this end, we assessed the available experimental, pharmacogenomic, protein structural and metabolic information available for all proteins in the erythrocyte model. Given the data collected from publicly available datasets (described above), we classified proteins based on: (i) availability of experimental protein structure, drug or metabolite binding information, (ii) known harmful gene-drug associations and (iii) if the knockout of this gene within the context of erythrocyte caused significant changes in metabolite import and export (see Methods), resulting in

four different classes of proteins based on these criteria (Figure 2(b)). This categorization mainly aids in the next steps of our contributed workflow, in studying the effects of SNVs on metabolite and drug binding using all-atom molecular simulations.

As shown in Figure 2(b), Class I targets have the most information available, including 3D protein structures (some in complex with a metabolite, drug or analogue), known drug-protein interactions, gene-drug associations, and clinically relevant phenotypic responses to a drug therapy. This group of proteins includes six proteins: catechol-O-methyltransferase (COMT), aldehyde dehydrogenase (ALDH3A1), adenosine deaminase (ADA), glucose-6-phosphate dehydrogenase (G6PD), glutathione peroxidase 1 (GPX1), and uridine 5'-monophosphate synthase (UMPS). Class II targets provide case-studies amenable to experimental testing SNV or drug-induced effects. Class III & IV targets are proteins found to be important in the genome-scale model, but do not have other sources (structural or pharmacogenomic) of information available, and therefore constitute examples of where our molecular modeling framework is useful for filling in missing information (Table B in Supplemental File 2.2).

Here, we focus the rest of this study on three distinctive proteins in erythrocyte metabolism (Figure 3): (i) catechol-O-methyltransferase (COMT), a class I protein (according to our above classification scheme); (ii) glucose-6-phosphate dehydrogenase (G6PD), a class I protein; (iii) glyceraldehyde-3-phosphate dehydrogenase (GAPDH), a class II protein. For the purpose of validation, we study the class I proteins, which have ample experimental, structural and pharmacological data associated with their roles in metabolism. To assess the predictive value of this workflow, we study the class II protein, a rare variant where population data was not available to understand the impact of documented sequence variants. Such an example serves as a demonstration for how this structural systems biology framework can be used in the absence of

experimental and pharmacological data. The targets chosen for this study and their pharmacogenomic importance are outlined in Table 1.

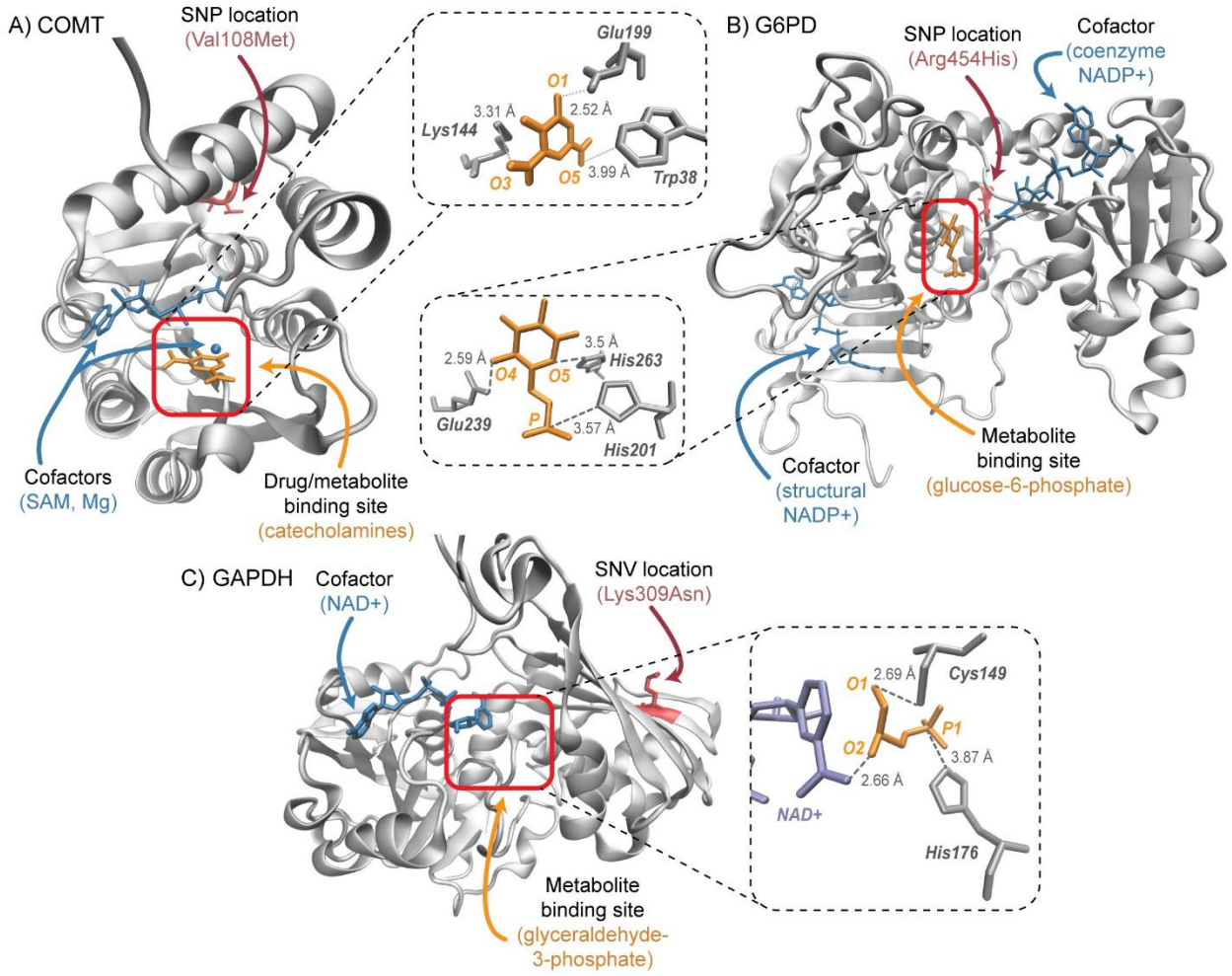



Figure 2.3. Protein targets studied in-depth using molecular simulations.
a) Protein structure of COMT (WT) from PDB entry 3BWM. In orange - crystallized position of an inhibitor analog, dinitrocatechol (DNC). In blue, cofactors needed for catalysis, S-adenosyl-methionine (SAM) and magnesium (Mg). In red, the position of the SNP (contained in PDB entry 3BWY). Zoom in - shows the active site of the enzyme with the crystallized DNC bound. b) Protein structure of G6PD (WT) from PDB entry 2BH9. In orange - crystallized position of the metabolite glucose-6-phosphate (G6P). In blue, the cofactor NADP+. In red, the position of the SNP. Zoom in - shows the active site of the enzyme with G6P bound. c) Protein structure of GAPDH (WT) from PDB entry 1U8F. The orange arrow indicates the known binding site of the metabolite glyceraldehyde-3-phosphate (G3P), which was not crystallized in the experimental structure. In blue, the cofactor NAD+. In red, the position of the SNV. Zoom in - binding site interactions of G3P in *E. coli* structure 1DC4.

Table 2.1. Signature proteins that impact erythrocyte metabolism and drug-induced phenotypes.

| Gene | Protein name | Native metabolite | Cofactors | WT PDB | Mutant PDB | SNP/SNV | Drugs known to bind | Number of exonic SNPs |
|---|---|---|---|---|---|---|---|---|
| COMT | Catechol-O-methyltransferase | Dopamine, epinephrine, norepinephrine | SAM, Mg | 3BWM, 3A7E | 3BWY | Val108Met | Tolcapone, entacapone, | 113 |
| G6PD | Glucose-6-phosphate dehydrogenase | Glucose 6-phosphate | $NADP^+$ | 2BH9 2BHL | Numerous | Arg454His | Phenobarbital metabolites | 206 |
| GAPDH | Glyceraldehyde-3-phosphate dehydrogenase | Glyceraldehyde 3-phosphate | $NAD^+$ | 1U8F | Modeled | Lys309Asn | N/A | 82 |

### 2.3.4 Molecular effects of sequence variation in protein-drug interactions

The next stage of our proposed workflow builds on previous methods [31, 64, 85, 86] and leverages systems modeling with molecular dynamics (MD) simulations. How SNPs/SNVs affect structure/function relationships proteins is a question that requires analysis beyond a comparison of crystal structures. Here, we take advantage of using an ensemble of protein conformations, generated from explicit solvent MD simulations, to study the effects of clinically relevant SNVs on drug and/or native metabolite binding (Figure 4(a)).

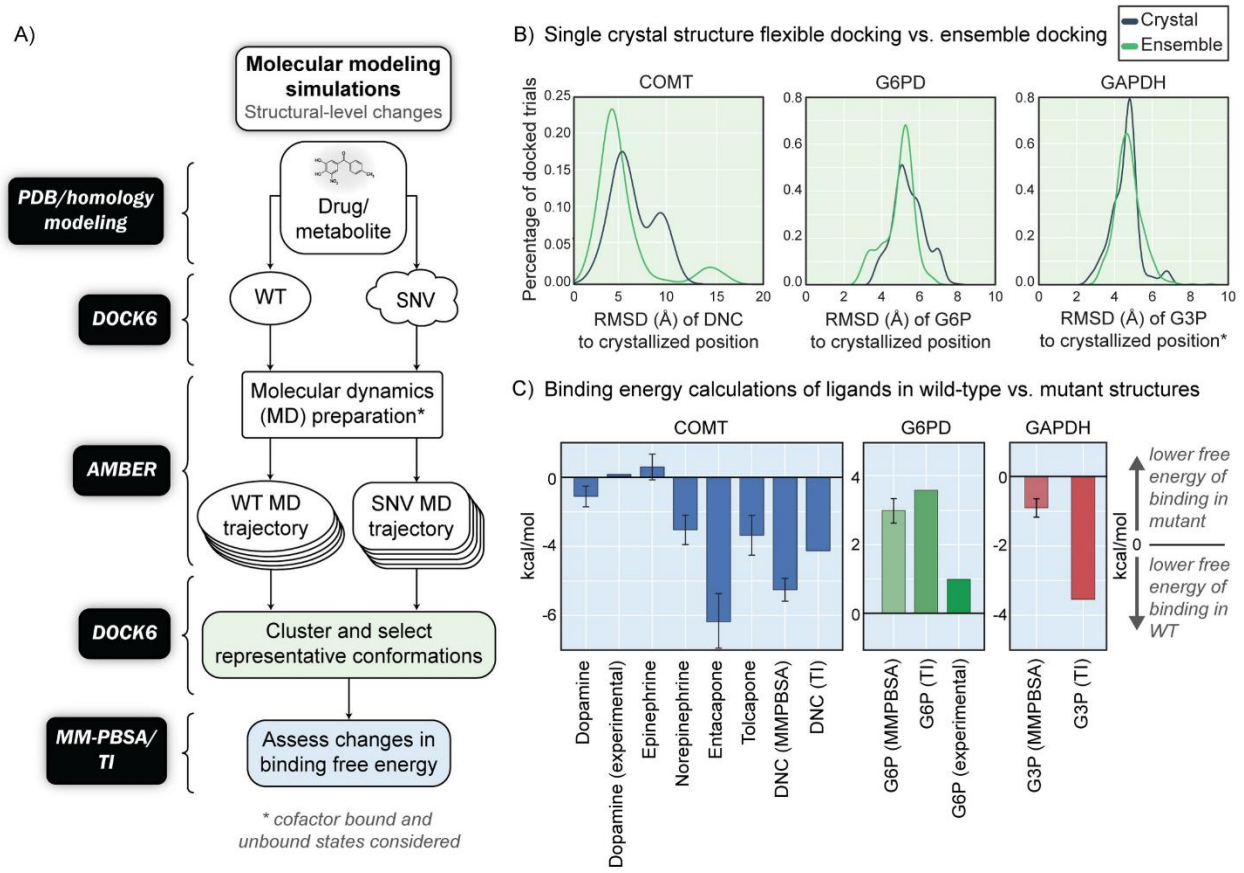



Figure 2.4. Molecular modeling framework and simulation results.
a) Molecular modeling frameworks used for molecular simulations of metabolite and drug binding differences between wild-type and mutant (SNV/SNP) proteins. In the first step, docking is first carried out on experimental or modeled protein structures. From molecular dynamics simulations, an ensemble of structures is generated from the long-time sampling of conformations that cannot be studied from a single, static structure (e.g. crystallographic structure). These ensemble structures provide multiple thermodynamic states of the protein that enable docking and analysis of binding free energy estimates. The overall goal of using these molecular modeling frameworks is to quantify the relative differences in the binding affinity of metabolites and drugs to wild-type and mutant proteins. Once these differences are computed, the ratios will be used to guide systems-level simulations. b) RMSD of predicted ligand poses of DNC to the original crystallized position based on docking trials to only the crystal structure (blue) versus utilizing an ensemble of structures (green). c) Differences in binding free energies from MM-PBSA calculations in wild-type vs. mutant proteins. A negative value indicates a lower predicted binding free energy to the wild-type protein, which corresponds to a higher binding affinity.

### *Catechol-O-methyltransferase (COMT)*

The activity of COMT in the erythrocyte, along with the inheritance of specific polymorphisms, is often used as a biomarker for different diseases, such as Parkinson's disease or schizophrenia [87-89]. COMT plays a critical role in the degradation of catecholamines, a class of

chemicals that mostly function as neurotransmitters in the human body [90], making it a prime target for further elucidating the effects of this SNP on protein-drug interactions. Further, COMT plays a key role in the erythrocyte, and its relationship to pharmacogenomic implications is likely to be applicable in other systems in the human body [91]. Of particular interest is the missense mutation, Val108Met, (i.e. Val158Met in the membrane bound version; dbSNP ID rs4680), which may cause changes in the protein's response to drug inhibitors [92]. While the crystallographic structures for both wild-type and SNP variants are available, minimal structural changes between the protein backbone of these two proteins are observed (i.e. they align with a 0.2 Å root mean squared deviation, RMSD (Fig E in Supplemental File 2.1)) [93].

We were interested in characterizing the binding mechanism of COMT, when it is in complex with either its native substrates (i.e. dopamine, epinephrine, norepinephrine) or known inhibitors/inhibitor analogs (e.g. dinitrocatechol, tolcapone, entacapone). Flexible molecular docking of dinitrocatechol (DNC), which is co-crystallized in both PDB structures, to the crystal structures of both wild-type and SNP variants gave a RMSD of less than 1 Å (of the drug backbone with respect to the original co-crystallized position) (Fig F in Supplemental File 2.1). We find that docking without the presence of the cofactors, (i.e., S-adenosyl methionine and a magnesium ion), slightly increases the RMSD (by 1.5 Å) of the binding pose, as expected due to the stabilizing features and steric constraints of these cofactors [94, 95]. Similar to DNC, docking of the native metabolites to the crystallographic structures retrieved binding poses within a RMSD of 2 Å of the original bound position (comparing equivalent atoms of the co-crystallized inhibitor DNC) (Fig G in Supplemental File 2.1). The best docked poses of the two drug molecules (tolcapone and entacapone) were initially reported about 10 Å away from the known binding site, which motivated

ensemble docking of both wild-type and variant proteins to understand the conformational space which was not represented in the crystal structure.

To generate an ensemble of conformations, we performed MD on both the wild-type and SNP variant proteins, in complex with their cofactors. We find that docking DNC to an ensemble of representative structures provides an increased accuracy in the final binding pose (Fig 4(B), COMT panel), compared to docked poses to only the single crystallographic structure, consistent with previous studies [96-100]. Ensemble docking of the catechol-like drugs and metabolites retrieved binding poses to within 5 Å of the original crystallized position in 72% of clustered snapshots from an MD trajectory. Furthermore, ensemble docking to wild-type COMT has a higher frequency of reproducing the crystallized binding orientation compared to the SNP variant (Fig H in Supplemental File 2.1). Following clustering of promising binding poses, we performed molecular dynamics for each of the proteins in the ligand-bound states and computed the free energy difference between the wild-type and variant proteins using MM-PBSA [101]. As shown in Figure 4(c), we find that for the majority of cases, the mutation V108M leads to a decreased affinity (i.e. an increase in relative binding free energy between wild-type and mutant protein, resulting in a negative $\Delta\Delta G$ ($\Delta G_{WT}$ - $\Delta G_{SNP}$)) to a majority of the native metabolites and drug molecules (excluding epinephrine; see Table K in Supplemental File 2.1). These findings are consistent with experiments that find the SNP variant to be less stable and less active, along with variant human subjects that respond less to drug therapy [92, 102, 103], yet no significant experimental differences were found with dopamine binding to the mutant [102].

*Glucose-6-phosphate dehydrogenase (G6PD)*

G6PD catalyzes the oxidation of glucose-6-phosphate (G6P) to 6-phospho-gluconolactone (6PG) within the pentose phosphate pathway, while maintaining the global concentration of

NADPH in the erythrocyte [104], required for protecting the cell from oxidative damage. There are more than four hundred sequence variants [105], of which many are implicated in hemolytic anemia and can be heavily influenced by drug side effects or a compromised immune system [106]. One particular missense mutation, referred to as the "Andalus" SNP (Arg454His; dbSNP ID rs137852324), has been classified as causing chronic nonspherocytic hemolytic anemia [107]. Structurally, the mutation occurs 21 Å from the substrate binding site (Figure 3(b)) and is expected to impact a salt bridging interaction with Asp286, potentially destabilizing the local structure of the protein. Notably, this arginine residue is highly conserved throughout organisms, reinforcing its structural and functional importance [108].

Similar to COMT, docking trials were carried out on wild-type G6PD and SNP variant structures. The wild-type structure was modified to generate the SNP variant and structural changes resulting from the sequence change were monitored during a 100 nanosecond trajectory. As expected, the salt bridging interaction between the mutated residue and Asp286 was eliminated (Fig I in Supplemental File 2.1). We performed ensemble docking simulations of various substrates to representative structures from the MD trajectory and found that, in 95% of the docking trials, G6P binds within 5 Å of the known active site of G6PD (Figure 4(b)). Although we do not observe large-scale differences in the docking poses of G6P in wild-type versus SNP variant proteins (Fig J in Supplemental File 2.1), binding free energy calculations indicate that the SNP variant has an increased binding affinity to the native substrate: G6P binds to the SNP variant with a $\Delta\Delta G = 3.00 \pm 0.68$ kcal/mol ($\Delta G_{WT} - \Delta G_{SNP}$). We find that this value is consistent when comparing to higher accuracy methods (e.g., from thermodynamic integration (TI), we find $\Delta\Delta G = 3.59$ kcal/mol) (Figure 4(c), G6PD panel). Experiments demonstrate that this mutation markedly increases the binding affinity of the native metabolite G6P in the variant while radically decreasing the overall

turnover rate ($K_M^{WT}$ = 52 ± 4 µM, $K_M^{SNP}$ = 9.71 ± 0.67 µM, calculated ΔΔG = 0.99 kcal/mol) [107]. Additionally, we find the SNP variant to have an increased binding affinity to the product of the reaction, 6PG (ΔΔG = 5.81 ± 3.11 kcal/mol), and a drastically decreased binding affinity to the cofactor NADP+ (ΔΔG = -13.057 ± 2.58 kcal/mol) at the secondary "structural" cofactor binding location [106, 109]. These may also be factors that contribute to the decreased turnover rate, such as due to a slower product release compared to wild-type behavior or enzyme instabilities caused due to a lower population of bound $NADP^+$.

*Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)*

GAPDH is an enzyme within the glycolytic pathway that catalyzes the conversion of glyceraldehyde 3-phosphate (G3P) to glycerate 1,3-bisphosphate, utilizing the cofactor $NAD^+$. It operates as a homotetramer, and a conserved cysteine residue (Cys149) is essential for its catalytic function [110]. Designated in this study as a Class II pharmacogenomic enzyme, it does not have any recent documented variants with phenotypic data, but from HapMap population sequencing data, a missense mutation, Lys309Asn (dbSNP ID rs11549334) was identified and predicted (using PolyPhen2 and SIFT) to be deleterious and/or disruptive (Table J in Supplemental File 2.1) [111, 112]. This mutant is found to occur 19 Å away from the binding site (Figure 3(c)). As with much of the sequence of GAPDH, this residue is conserved throughout eukaryotic organisms [113], and thus observed changes are rare and likely marked as deleterious according to these predictive algorithms.

The Lys309Asn mutant structure was generated by modifying the sequence of the experimental wild-type structure and monitoring structural changes during a 100 nanosecond trajectory. In this case, ensemble docking resulted in a slight trend for more correct binding poses in the WT ensembles when compared to the mutant (Figure 4(b), Fig K in Supplemental File 2.1).

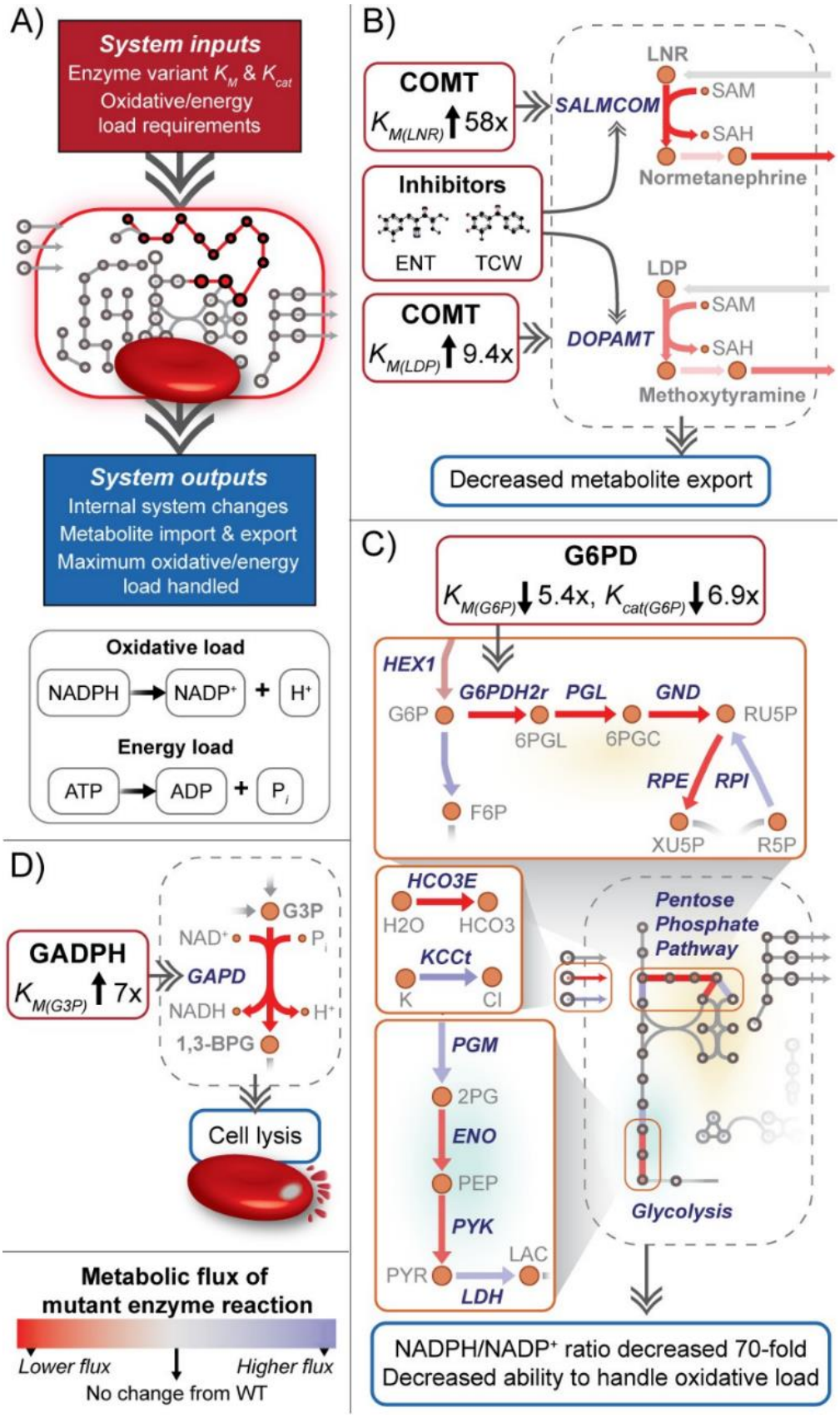
Clustering of the docked poses was carried out based on ligand-protein interactions obtained from literature (Table I in Supplemental File 2.1) [114-116]. Computing the free energy binding difference for G3P between wild-type GAPDH and mutant protein, we confirm that the wild-type binding affinity is stronger than that of the mutant variant ($\Delta\Delta G = -3.55 \pm 0.6$ kcal/mol) (Figure 4(c), GAPDH panel). The binding of the cofactor ($NAD^+$) was found to have similar binding affinities in both forms of the enzyme ($\Delta\Delta G = -0.5504 \pm 1.8$ kcal/mol). Due to the highly conserved nature of this specific residue and the suggestion of a decreased binding affinity to the native metabolite G3P, our predictions are consistent in that it may be a cause of enzymopathy.

### 2.3.5 Systems-level effects of sequence variation related to drug responses

While understanding protein-drug interactions provides information on *how* sequence variation changes protein structure and reactivity, evaluating the downstream effects of these changes requires a systems-level perspective (Figure 5(a)). Changes in metabolic networks can be assessed using a variety of systems methods including constraint-based and kinetic modeling techniques [9, 47, 117, 118]. To test the susceptibility of the metabolic network of the human erythrocyte to the harmful variants detailed above, we utilized both constraint-based modeling of the *i*AB-RBC-283 model [50] and a recently developed *in silico* kinetic rate law model derived from the Mass Action Stoichiometric Simulation (MASS) approach [119, 120]. For a number of proteins, disease causing mutations can cause systemic changes within the metabolic network or in the transport of certain metabolites [50, 121]. With regards to the erythrocyte, understanding these differences in metabolite transport can be correlated with changes in metabolite concentrations within biofluids, which potentially expands the use of this model as a diagnostic tool for human disease. Similar perturbations can also be linked to the specific phenotypic

responses of the erythrocyte, such as to drug treatments, or the ability to respond to changes in oxidative (rate of NADPH use in order to combat oxidants) or energy (rate of ATP use) load [47].

Figure 2.5. Systems modeling framework and modeling results.
a) Systems modeling framework used in this study. Inputs used for constraint-based and kinetic modeling are derived from molecular modeling calculations and experimental data when available. In order to understand how small-scale changes from enzyme variants affect the entire system, we look at the internal system changes (in reaction flux and metabolite concentration), differences in metabolite import & export, and how the cell handles an increase in oxidative or energy loads. Oxidative load is defined as the conversion of NADPH to NADP+, whose rate of reaction is increased under states of oxidative stress. Energy load is defined as the use of ATP. For all panels, the change in metabolic flux is colored by a difference from the wild-type flux state, red being a decreased flux in the mutant state and blue being an increased flux. b) Constraint-based modeling for the mutant COMT enzyme. The SNP is predicted to decrease the binding affinity of the enzyme in norepinephrine and dopamine metabolism. Increasing the $K_M$ (predicted) of COMT for the respective reactions leads to decreased flux and as a result decreased export of their methylated counterparts. Inhibitors tolcapone (TCW) and entacapone (ENT) are also predicted to have a lowered binding affinity to COMT, leading to similar effects. c) Kinetic modeling for the mutant G6PD enzyme. Decreases of the $K_M$ (predicted and experimental) and of the $k_{cat}$ (experimental) lead to major systemic changes of the pentose phosphate pathway and glycolysis. The ratio of NADPH to NADP+ greatly decreases and subsequently the oxidative load able to be handled also decreases. d) Kinetic modeling for the mutant GAPDH enzyme. The cell is unable to handle the predicted increase in $K_M$ (predicted) and results in an infeasible state of the model, corresponding to cell lysis.


A)
System inputs
Enzyme variant $K_M$ & $K_{cat}$
Oxidative/energy load requirements
System outputs
Internal system changes
Metabolite import & export
Maximum oxidative/energy load handled
Oxidative load
NADPH
NADP+
H+
Energy load
ATP
ADP
Pi
B)
COMT
$K_{M(LNR)}$ 58x
Inhibitors
ENT
TCW
COMT
$K_{M(LDP)}$ 9.4x
LNR
SALMCOM
SAM
SAH
Normetanephrine
LDP
DOPAMT
SAM
SAH
Methoxytyramine
Decreased metabolite export
C)
G6PD
$K_{M(G6P)}$ 5.4x, $K_{cat(G6P)}$ 6.9x
HEX1
G6P
G6PDH2r
6PGL
PGL
6PGC
GND
RU5P
F6P
RPE
RPI
XU5P
R5P
HCO3E
H2O
HCO3
KCCt
K
Cl
Pentose Phosphate Pathway
PGM
2PG
ENO
PEP
PYK
PYR
LAC
LDH
Glycolysis
NADPH/NADP+ ratio decreased 70-fold
Decreased ability to handle oxidative load
D)
GADPH
$K_{M(G3P)}$ 7x
G3P
NAD+
Pi
GAPD
NADH
H+
1,3-BPG
Cell lysis
Metabolic flux of mutant enzyme reaction
Lower flux
Higher flux
No change from WT

*Catechol-O-methyltransferase (COMT)*

As COMT is not present in the core erythrocyte kinetic model [119], we therefore turned to constraint-based modeling techniques, utilizing the entire genome-scale model of the erythrocyte. We used the established Markov Chain Monte Carlo-based (MCMC) sampling approach [122] to calculate distributions of all feasible flux states for both wild-type and SNP systems. Ligand binding differences between wild-type and SNP variant (computed from the molecular simulations) were integrated into the erythrocyte model by altering the reaction flux bounds, which represent the rates that metabolites flow through a reaction. We used the relative ratio of the binding affinity differences to effectively constrain the quantitative relationship between the wild-type and mutant metabolic state as well as the difference in behavior of the enzyme under a drug load.

Our findings suggest significant changes in the uptake of dopamine and norepinephrine, and the secretion of their methylated counterparts (Figure 5(b)) as a result of the sequence variant. In contrast, for epinephrine, the computed binding free energy difference in wild-type and SNP protein was positive (i.e. it binds more strongly to the mutant), which did not influence network analysis of the uptake or secretion of its associated metabolites. Furthermore, the mutant COMT decreases the effectiveness of the drugs entacapone and tolcapone, which is again reflected by an increase in the secretion of the methylated metabolites when compared to the wild-type cell inhibited by these drugs. These findings are consistent with previous studies related to entacapone, which report decreased efficacy of entacapone in individuals with the SNP (Met108) [92], though it may be dependent on different human-specific characteristics [123] and tolcapone, which has a reported increased efficacy in individuals with the wild-type (Val108) [124]. These findings oppose the previous claim that the genotype did not contribute to the clinical response [125].

*Glucose-6-phosphate dehydrogenase (G6PD)*

In patients with G6PD deficiency, the most common symptom is hemolytic anemia resulting from the erythrocyte's loss of ability to respond to oxidative stress. This ability can be measured by simulating an increase in the oxidative load within a kinetic modeling framework. The predicted increase in binding affinity of G6P to the mutant enzyme corresponds to the experimentally calculated binding affinity reported in [107]. If we assume the same catalytic rate ($k_{cat}$) of the reaction carried out by this enzyme, the cell's ability to respond to an oxidative load does not decline. Though the ratio of NADPH to $NADP^+$ increases, it does not lead to a significant increase in the oxidative load tolerated when compared to the baseline, wild-type model. Integrating the experimentally measured $k_{cat}$ from [107], however, drastically reduces the ratio of NADPH to $NADP^+$ and subsequently lowers the maximum tolerable oxidative load of the cell under stress conditions (Figure 5(c)). Incorporating these kinetic parameters within the erythrocyte model, we find specific systemic effects that correspond to the classification of this SNP as a "severe deficiency with intermittent hemolysis" by the WHO [126]. This behavior is also observed from constraint-based modeling after decreasing the flux through the corresponding reaction and induces significant changes (defined as <40% the original, wild-type flux span, see Methods) in the glutathione reductase pathway, which utilizes NADPH to combat oxidative stress.

*Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)*

GAPDH deficiencies in humans are rare, and mostly cause mild hemolytic anemia [127]. By integrating the relative change in $K_M$ of the mutant, based on binding free energy computations, into the erythrocyte kinetic model, we observed that this change led to lethality (Figure 5(d)). Smaller relative changes in the $K_M$ or $k_{cat}$ (compared to wild-type) were not lethal and did not impact the ability of the erythrocyte to respond to an increase of oxidative or energy load,

suggesting that only a small degree of change in protein structure and reactivity may be tolerated. This finding is consistent with studies in mice where those with lower activity mutants or those heterozygous for a lethal mutant rarely showed symptoms, while those homozygous for a mutant encountered mortality at the development stage [128]. The human subject with this annotated variant (within the HapMap dataset) was noted as having a heterozygous form of this specific mutant, which would explain the non-lethality observed. It is important to note that GAPDH is involved in several non-metabolic processes [129], and while variation of the enzyme sequence may be tolerated to a certain extent, these additional processes may be impacted due to the causal effect of this mutation.

## 2.4 Conclusion

Here, we propose a framework for mapping protein structural information to genome-scale models of human erythrocyte metabolism for the characterization SNP-drug associations. Three case studies presented in this contribution point to the complexity of pharmacogenomic associations and being able to conduct integrated *in silico* simulations that extend from the molecular scale to the systems level. Using parameters from molecular simulations to guide genome-scale modeling, we are able to study how changes in protein structure and binding affinity influence the phenotypic states of an entire metabolic network. We find that the union of genome-scale modeling and molecular, physics-based methods, presents, to the best of our knowledge, the first workflow capable of systematically integrating data from pharmacogenomics research, in conjunction with 3D high resolution protein structural information, to model changes on both the pathway (i.e. metabolic network) and molecular (i.e. protein) scales. The information gained through molecular modeling simulations can be utilized to supply parameters to both kinetic

models and constraint-based modeling approaches and has been found to be amenable to the study of other enzymopathies [47, 130]. Our findings indicate that there is consistency between experimental and computational trends in substrate and drug compound binding in wild-type versus mutant proteins.

Currently, most systems biology approaches lack the ability to utilize insights from structure-based analyses related to metabolite and/or drug binding. Fortunately, atomistic molecular simulations have evolved to become powerful tools for the characterization of binding mechanisms and as such constitute valuable assets for systems modeling. Extending analysis beyond crystallographic structures through the use of ensemble confirmations substantially enhances the predictive scope of docking methods by identifying alternative binding modes for a drug molecule [96-100]. Ensembles of the thermodynamically accessible states of a protein, generated from molecular dynamics, allows for the mechanistic characterization of how sequence and structural variation may influence metabolite or drug binding [131].

The scalability of this workflow is mainly limited (i) to the documentation and experimental analysis of exonic SNVs/SNPs, and (ii) by the execution of molecular dynamics simulations, which takes a significant manual effort and requires high performance computing resources. For the second point, certain efforts have already shown that high-throughput simulations using classical MD can be performed on large numbers of proteins [132, 133]. However, performing high accuracy computations on a systems scale is currently intractable, due to the intense computational and time requirements of quantum-based simulations or free energy calculations. Therefore, a trade-off between accuracy and cost must be considered (see Fig B in Supplemental File 2.1 and recent reviews on the subject [134-136]). In light of these limitations, we find that the additional information gained from protein structure greatly contribute to our

understanding of causal mutations and can assist in selecting protein targets for more detailed molecular studies. Thus, when combined with other developing frameworks [46] and experiments [137], the contributed workflow provides a first step in the translation of Big Data in the pharmaceutical industry to practical therapeutic applications and is expected to have a positive transformative impact on the fields of systems medicine, population studies and drug discovery efforts.

## 2.5 Methods

### 2.5.1 GEM-PRO construction

The techniques used here are a consolidation of 4 previous methods to add protein structural information to genome-scale models [31, 32, 64, 138], and described in detail in [27]. To do so, the SBML model of the erythrocyte genome-scale model was first obtained from the BiGG Models website (http://bigg.ucsd.edu/models/iAB_RBC_283) [139], and all gene IDs were mapped to their corresponding amino acid sequences (UniProt and RefSeq entries). This model differed from the construction of previous GEM-PROs due to the appearance of protein isoforms, and required additional manual mapping to ensure correctness. Gene isoforms led to inconsistencies between database entries and additional difficulty linking to available homology models (discussed in the section “Homology Modeling”). Additional QC/QA steps were taken in order to ensure the correct sequence was being retrieved, as described below.

*Mapping to UniProt accession numbers*

For a given gene in *i*AB-RBC-283, there are a number of associated isoforms, annotated as the gene name and a isoform number, separated by a decimal (eg. "Aldoa.1"). We take the gene name, which is taken from the corresponding gene in Recon 2, obtain the Entrez gene ID [140],

and directly map this to its corresponding UniProt accession code (UAC). Then, we directly map isoform numbers to available isoforms in the UAC entry (Fig A in Supplemental File 2.1, top panel). These are annotated with reviewed isoform-specific sequences, allowing us to filter for the correct experimental PDB structure in later stages.

*Mapping to RefSeq and Ensembl identifiers*

In some cases, the number of isoform sequences annotated in *i*AB-RBC-283 does not match the number of isoforms available in UniProt. For these, we generated a separate mapping pipeline to the RefSeq and Ensembl databases [141]. The Bioservices Python package [42] and Ensembl Biomart tables [142] were used in order to first map the gene IDs without their isoform identifier to their corresponding entries, and then back to isoform IDs according to the transcript name as listed in Ensembl (see Fig A in Supplemental File 2.1, bottom panel). The information here was also utilized in order to cross-reference what was successfully mapped with the UniProt mapping service. Once the correct isoform entry was found, available PDB mappings were found using the entry ID (RefSeq or Ensembl Protein), or by sequence alignment to the PDB. We note that the difficulty in mapping isoforms and inconsistencies between databases points to a larger need of consistency and standardization for this biological property.

*Homology modeling*

We have filled in the gaps where there are no experimental structures by querying previously generated databases of I-TASSER homology models for *H. sapiens* [84], and manually generating homology models for genes that were not part of these databases [14]. The I-TASSER Suite version 4.4 was utilized for the construction of missing structures, and provides an especially useful method in modeling splice isoforms, which are specialized in the erythrocyte [143]. In the final GEM-PRO data frame, we note where available homology models have been mapped to their

respective genes. We also include additional information in the data frame that explains the type of computational prediction method used to model the protein structure (e.g. template-based versus ab initio), the corresponding URL (for downloading the homology file from the source), the label (i.e. the identifier of the model given by the homology model database), and information related to the confidence of the homology model (e.g. C-score), the native (homologous) template used for the model, etc. All columns added to the master data frame from this stage are preceded by a 'i' for I-TASSER. It is important to note that certain PDB structures with unresolved residues or gaps in the structure can also be homology modeled to enhance the structural coverage of the amino acid sequence. Quality scores for each model are included as PSQS and PROCHECK scores [144, 145].

*QC/QA procedure*

For the purpose of molecular modeling, it is important to select high-quality starting structures when conducting docking or molecular dynamics simulations. On a genome-scale, an automatic ranking system becomes a requirement if there are multiple structures or homology models that represent one gene. The main objective of this section is to discuss the quality assessment and quality control of the data that has been thus far mapped to the metabolic network reconstruction. In previous versions of GEM-PROs, experimental structures were additionally classified and ranked according to whether a protein was bound to a native metabolite or ligand, in order to ensure proper binding predictions. While the updated version of the GEM-PRO modeling framework does not include the bound state of a protein as a target characteristic in the quality control pipeline, this data is accessible in the knowledge base. Instead, we are mainly interested in quantifying the general quality attributes of the experimental structure of the protein.

An ideal starting platform for higher-level modeling methods are experimental protein structures without missing residues (especially at the interior of the protein) and 100% sequence identity compared to wild-type. To determine which experimental structures required further modeling or modification, we devised a scoring metric that ranks each PDB based on 1) the coverage of the wild-type amino acid sequence, 2) the resolution, and 3) the similarity of secondary structural features between the PDB structure and its corresponding homology model. The final outcome of the quality assessment is the classification of experimental structures into three groups: (i) high quality structures requiring no modification; (ii) high quality structures requiring minimal (site-directed) modification and (iii) low quality structures requiring homology modeling. For more information on the ranking scheme, please see Supplemental File 2.1 and [27]. All protein structure files following ranking and quality control are included within Supplemental File 2.3.

### 2.5.2 Genetic variation, drug-target interactions, and essential genes

Previous work was done to map data from the Online Mendelian Inheritance in Man (OMIM) database in order to find disease causing mutations that could map to erythrocyte proteins [50]. We also collected all known SNPs from dbSNP, and filtered them down to variations in exons that could be studied utilizing protein structure information. Information was additionally cross-referenced with UniProt variant annotations [146].

There are a number of drug target databases that were queried for this study. DrugBank was used in a previous study to gather drug targets based on sequence [50]. In order to be as comprehensive as possible, we also obtained data from ChEMBL [147] and MATADOR [83], with MATADOR providing annotations for indirect interactions. With this, we were able to verify targets that appeared in all 3 databases. Drug adverse effects due to variation were mainly gathered from the PharmGKB, a pharmacogenomics database with information from clinical studies,

research articles, and individual cases [148]. The PharmGKB further annotates for the significance of an association, as well as details of the clinical trial or GWAS study carried out. Finally, the DrugBank contains a simple list of SNP-drug associations in their SNP-ADR and SNP-FX sub-databases [82], which was cross-referenced with all information found in the PharmGKB.

As a final source of parameters for validation of our model, experimentally determined kinetic values for binding of a drug or inhibitor to a target (wild-type as well as mutant) were obtained from BRENDA and the BindingDB [149, 150]. As expected, information for this step was much sparser than the previous information, which indicates the need for experimental assays if we are to validate the predictions made from this model. For the targets in this study, we also manually searched for additional information from published biochemical studies.

Finally, for the selection of interesting targets to study with molecular and systems modeling techniques, we also wanted to understand the essentiality of each gene within the erythrocyte model. Gene knockouts were performed for each gene contained within *i*AB-RBC-283, as per [50]. A gene was marked as interesting to study within the context of the erythrocyte if there were significant changes within the reaction fluxes of metabolite import and export through the membrane using flux variability analysis (FVA) simulations [151]. In order to detect these significant differences, all reaction fluxes were compared to the normal "wild-type" state of the cell. Specifically, similar procedures to Shlomi et al. and Bordbar et al. were followed [50, 121]. Changes in exchange fluxes were categorized into i) activation/inactivation, ii) shift to a fixed direction, iii) a change in magnitude of flux, or iv) no change (refer to [50], Figure 5). For changes in magnitude of flux, if the new flux span (defined as maximum flux - minimum flux) was less than 40% of the original flux span, it was considered to be a significant change.

### 2.5.3 Molecular modeling and docking

Experimental PDB structures or homology models representing the genes of interest in this study were taken from the GEM-PRO data frame following ranking and QC/QA. Mutant forms of the enzymes were either taken directly from the PDB, if available, or modeled by point mutations of the structure. Next, the general approach for each target was to first understand the binding position and energetics of either the native metabolite or a drug of interest to a wild-type protein structure and its corresponding mutant. Flexible docking simulations using DOCK6 were carried out with default parameters and binding sites defined when known [152]. Furthermore, simulations were conducted with and without cofactors, to account for competitive binding drugs or cases where the order of substrate binding was not known. To compare flexible docking results to ensemble docking, simulations were repeated under different random seeds for a total of 500 docking runs.

### 2.5.4 Molecular dynamics simulations and ensemble docking

Molecular dynamics simulations were run utilizing the PMEMD module of the AMBER14 toolkit [153]. Initial parameterization of ligands and cofactors were carried out utilizing the Gaussian 09 software [154] or obtained from previously published data sets (see Supplemental File 2.1 for protein-specific methods and Supplemental File 2.3 for parameter sets). For generating topologies as input to AMBER, 99SB force field charges and atom types were then used and then solvated in a periodically repeated TIP3P 12 Å water box with counterions being added as needed ($Na^+$ or $Cl^-$). Minimization was carried out under constant volume conditions at while being heated to 300 K. Structures were then equilibrated under constant temperature and pressure conditions with restraints being released. Finally, the structures were run in production phase of 75 ns or more under a Langevin thermostat and Particle Mesh Ewald (PME) cutoff of 12 Å.

At least 4 separate MD simulations (representing WT and SNP structures in cofactor unbound and bound states, more for additional cofactor bound states) were carried out on each enzyme (see Tables D-F in Supplemental File 2.1 for all simulation information). Every 100 frames from these trajectories were utilized as input for ensemble docking of the substrate of interest.

All docked positions were clustered into 5 representative poses based on the distances from known binding residues. Specifically, distances from 3 known binding or interacting residues to the atoms of the drug or metabolite were calculated for each extracted frame, and k-means clustering of the Euclidean distance separated these frames into 5 distinct binding modes for use in further simulation. These docked positions were subject to additional MD production runs of 10 ns each, in order to examine the stability of the bound position and if they would converge into one distinct pose. We conducted free energy calculations for each of the ligands in the cofactor bound state of the WT and SNP enzymes. MM-PBSA calculations were carried out to predict the difference in free energies of binding ($\Delta\Delta G$). The binding energies of all 5 representative conformations were averaged per ligand, and the resulting value indicates if the ligand is more favorable to bind to WT (negative $\Delta\Delta G$) or SNP (positive $\Delta\Delta G$) structures.

### 2.5.5 Binding energy calculations

MM-GBSA/MM-PBSA calculations utilizing the MMPBSA.py script available in the AMBER14 toolkit were carried out on the 10 ns simulated receptor-ligand complexes [101]. The first nanosecond of simulations were discarded before running calculations to account for initial stabilization of the docked ligand. Thermodynamic integration (TI) calculations were calculated utilizing the Simulated Annealing with NMR-derived Energy Restraints (SANDER) module within AMBER14 [155]. The dual topology paradigm was utilized with a three step alchemical

transformation, with state 0 representing a wild-type enzyme and state 1 the mutant form. Step 1 carried out the decharging of the WT utilizing 10 λ points and simulations of 1 ns each. Step 2 transformed the residue atoms of the WT to the SNP again utilizing 10 λ points and simulations of 1 ns each. Step 3 carried out the recharging of the mutant residue atoms with the same number of λ points and simulation time. This was run for both ligand bound and unbound states. Finally, the change in potential energy of the system with ligand bound was calculated by integration over the λ points and subtracted from the ligand unbound state. For full information on docking, MD, MM-PBSA, and TI parameters, please refer to the section entitled "Molecular modeling simulations" in Supplemental File 2.1.

### 2.5.6 Systems modeling

The constraint-based modeling approach was carried out for all enzymes in this study by simulating a normal (wild-type) and perturbed (mutant) erythrocyte condition utilizing FVA followed by a Markov chain Monte Carlo (MCMC) based sampling approach [122, 130, 156]. Previous simulations for identifying biomarkers have simulated perturbed states by setting the upper and lower bounds of flux through affected enzymes of the cell to 0, effectively mirroring a full gene inhibition, and then analyzing the exchange conditions [50, 121]. For the purposes of this study, we are now able to understand the relative differences in native metabolite catalysis utilizing the ratio of differences in the binding affinity between wild-type and mutant forms of the enzymes. This ratio was then converted into a ratio of flux in wild-type to mutant enzymes, assuming equal concentration of substrate and enzyme (see Equation S3). From this, the determined normal wild-type minimum and maximum fluxes through the corresponding reaction were adjusted to a perturbed mutant state, and both FVA and MCMC simulations were then run with the goal of analyzing 1) the flux differences through the exchange reactions (import/export of metabolites) of

the erythrocyte (as described above in the section "Genetic variation, drug-target interactions, and essential genes") and 2) significant flux shifts within the internal network. In this way, hypotheses for the altered phenotypic state of the erythrocyte and its impact on the body could be deduced based on the differences of uptake or secretion of metabolites or large-scale internal network changes. For MCMC simulations, significant shifts in the distribution of fluxes were considered (p-value < 0.05). Additional information on MCMC sampling is included in the section entitled "Systems modeling" in Supplemental File 2.1.

With the kinetic rate law model, we are able to directly integrate the predicted $K_m$ and experimental $k_{cat}$ values as well as simulate the cell under oxidative or energy load conditions. This detailed model was utilized for the simulations of normal and perturbed G6PD and GAPDH enzymes. Simulation of COMT within the kinetic model was not available due to the current model being limited to core metabolic enzymes. We utilize the model to also understand the erythrocyte's capability to withstand oxidative stress or increased energy needs and compare wild-type to mutant states. Oxidative stress is simulated as an increase in the rate of NADPH usage, to mirror the fact that a cell under stress requires NADPH to neutralize reactive oxygen species. Energy load is simulated as an increase in the rate of ATP usage. The normal, wild-type cell was first simulated and the maximum oxidative and energy loads were determined for comparison to the mutant state. Integration of the predicted $K_M$ without any change in $k_{cat}$ was then simulated for the mutant state, to understand if only changes in binding affinity led to a change in maximum tolerable oxidative or energetic load. Finally, changes from predicted $K_M$, experimental $K_M$, and experimental $k_{cat}$ were fully integrated to investigate the model's accuracy to the known phenotype.

## 2.6 Supplemental Files

Supplemental File 2.1. Expanded methods and results text and figures detailing GEM-PRO construction, molecular modeling simulations, and systems modeling.

Supplemental File 2.2. GEM-PRO of the human erythrocyte and related pharmacogenomics files.

Supplemental File 2.3. Parameters used in molecular modeling simulations.

## 2.7 Acknowledgements

This work was funded by the Swiss National Science Foundation (grant p2elp2_148961 to E.B.), and by the National Institutes of Health (grant GM057089 to B.O.P.). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors would like to thank Prof. Ross Walker and the members of his laboratory, Prof. Philip E. Bourne, Dr. Roger Chang, Dr. Daniel Zielinski, Prof. Nathan Lewis, Dr. Ke Chen, and Chia-An Yen. We are also grateful to Gilbert Omenn and the two anonymous reviewers for helpful review comments.

Chapter 2, in full, is a reprint of the material as it appears in Mih N*, Brunk E*, Bordbar A, Palsson BO. 2016. A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism. PLoS Comput. Biol. 12:e1005039. http://dx.doi.org/10.1371/journal.pcbi.1005039. The dissertation author was the primary author (equally contributing with Elizabeth Brunk) of this paper.

# Chapter 3.

# Systems Biology of the Structural Proteome

## 3.1 Abstract

The success of genome-scale models (GEMs) can be attributed to the high-quality, bottom-up reconstructions of metabolic, protein synthesis, and transcriptional regulatory networks on an organism-specific basis. Such reconstructions are biochemically, genetically, and genomically structured knowledge bases that can be converted into a mathematical format to enable a myriad of computational biological studies. In recent years, genome-scale reconstructions have been extended to include protein structural information, which has opened up new vistas in systems biology research and empowered applications in structural systems biology and systems pharmacology.

Here, we present the generation, application, and dissemination of a genome-scale model with protein structures (GEM-PRO) for *Escherichia coli* and *Thermotoga maritima*. We show the utility of integrating molecular scale analyses with systems biology approaches by discussing several comparative analyses on the temperature dependence of growth, the distribution of protein fold families, substrate specificity, and characteristic features of whole cell proteomes. Finally, to aid in the grand challenge of big data to knowledge, we provide several explicit tutorials of how protein-related information can be linked to genome-scale models in a public GitHub repository (https://github.com/SBRG/GEMPro/tree/master/GEMPro_recon/).

Translating genome-scale, protein-related information to structured data in the format of a GEM provides a direct mapping of gene to gene-product to protein structure to biochemical

reaction to network states to phenotypic function. Integration of molecular-level details of individual proteins, such as their physical, chemical, and structural properties, further expands the description of biochemical network-level properties, and can ultimately influence how to model and predict whole cell phenotypes as well as perform comparative systems biology approaches to study differences between organisms. GEM-PRO offers insight into the physical embodiment of an organism's genotype, and its use in this comparative framework enables exploration of adaptive strategies for these organisms and opens the door to many new lines of research. With these provided tools, tutorials, and background, the reader will be in a position to run GEM-PRO for their own purposes.

## 3.2 Background

The success of genome-scale modeling can be attributed to high-quality, bottom-up reconstructions of metabolic, protein synthesis, and transcriptional regulatory networks on an organism-specific basis [157-160]. Such network reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledge bases (or k-bases) [161] that can be used for discovery purposes (such as model-driven discovery of unidentified metabolic reactions [162], studies of evolutionary processes [28], and analysis of biological network properties), as well as practical applications (such as metabolic engineering, prediction of cellular phenotypes [163], and interspecies similarities and differences). Others have explored host/pathogen interactions [164], cocultures and microbial communities [165-168], ecology [169], and chemotaxis [170]. Numerous recent developments have broadened the predictive scope of genome-scale models by incorporating other sources of biological data, such as protein structural data, into reconstructions [28, 31, 32].

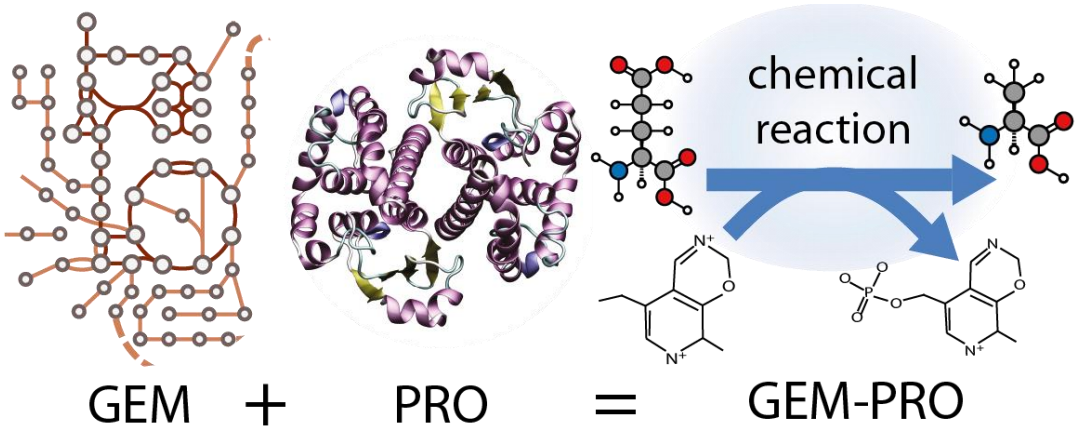



Figure 3.1. Structural systems biology emerges from the integration of networks and structural biology. Genome-scale models incorporate multi-omic data and large-scale curation from databases such as KEGG and UniProt. Molecular-level analyses enable atomic-level characterizations of secondary structure, substrate binding, and comparisons of similar catalytic sites among proteins in the metabolic network.

The complementarity of molecular-level and systems-level data types has led to the integration of protein structurally-derived data into genome-scale models. Using genome-scale models of metabolism (GEMs), we link metabolic enzyme activities to characteristics of observed phenotypes, whereas using structural biology, we link molecular interaction details (e.g., protein-ligand binding) to the activities of enzymes. The genome-scale models with protein structures (GEM-PRO) framework, therefore, gives a direct mapping of gene to transcript, to protein structure, to biochemical reaction, to network states, and finally to phenotype (Figure 1). Understanding the structural properties of proteins as well as their respective ligand binding events (e.g., metabolite, drug or oncometabolite) enables the characterization of molecular-level events that trigger changes in states of an entire network. Such a multi-scale approach acts as bridge between systems biology and structural biology, two scientific disciplines that, when combined, become the emerging field of structural systems biology [1, 171-174]. This union has brought about exciting advances, which would have otherwise been out of reach: the evolution of fold

families in metabolism [28], identification of causal off target actions of drugs [31], identification of protein-protein interactions [175, 176], and determination of causal mutations for disease susceptibility [176, 177].

In recent years, the number of publicly available biological macromolecules has grown to more than 110,000 entries, and continues to increase yearly by roughly 10% [178]. The increasing availability of protein structural data brings about a number of implications for GEM-PRO models. First, to keep pace with the deluge of protein data coming from experiments, there is a developing need for pipelines that use systematic mapping and quality assurance processes to read, filter, and process all newly deposited structures, ultimately managing all relevant data in an easy-to-use knowledgebase. Second, increasingly accessible protein structural data enhances the predictive scope of systems biology research; the more description we have of the biological components involved in complex systems, the more we can understand cellular processes that span a wide range of biological, chemical, and structural detail. Expanding these models would allow for the progressive description from a 1 to a 2- to a 3-D view of biology. Finally, to aid in the dissemination and further development of these resources, growing datasets and pipelines should be developed together with in silico tools that increase data accessibility and training.

Here, we address each of the above implications and demonstrate how linking protein structural data to GEMs enables the generation, dissemination, and application of GEM-PRO for studying two contemporary organisms, *T. maritima* and *E. coli*. For the generation and updating of GEM-PRO, we present a novel pipeline that systematically maps genes in a metabolic model to their respective high-quality structural data. We present four novel applications areas which demonstrate the utility of modeling at the intersection of systems and structural biology: (i) metabolic protein specificity; (ii) the relationship between protein complex stoichiometry and in

vivo protein abundance; (iii) the diversity of bacterial proteomes; (iv) protein properties of growth rate-limiting reactions at high temperatures. Finally, for dissemination and training purposes, we distribute the GEM-PRO knowledgebase together with tutorials, which explicitly describe how GEM-PRO can address the following questions: (i) How are protein fold families distributed over metabolism? (ii) How does temperature, and hence protein instability, determine growth rate?

## 3.3 Results and Discussion

### 3.3.1 Generation and updating of GEM-PRO using a systematic pipeline

As with metabolic network reconstructions [157], structural proteome reconstructions require constant curation and updating to incorporate newly deposited experimental protein structures. For example, over the course of two years, the number of available experimentally determined protein structures for *E. coli* has increased substantially (since 2013, 356 additional experimental *E. coli* protein structures can be linked to genes in the metabolic network model, *i*JO1366 [179]) and the structural coverage of genes in the model has increased by 10% (133 genes). In this section, we describe the construction of a quality assessment pipeline which enables newly deposited crystallographic or NMR structures to be searched, assessed, and managed within a structured k-base. In total, 2 person-hours are required by this workflow, once all homology models have been constructed for proteins without available crystallographic structures. Time and computational requirements for homology modeling are discussed in the I-TASSER pipeline [14]. The workflow discussed here can be carried out with no specific hardware requirements, and software requirements are outlined within the tutorial notebooks.

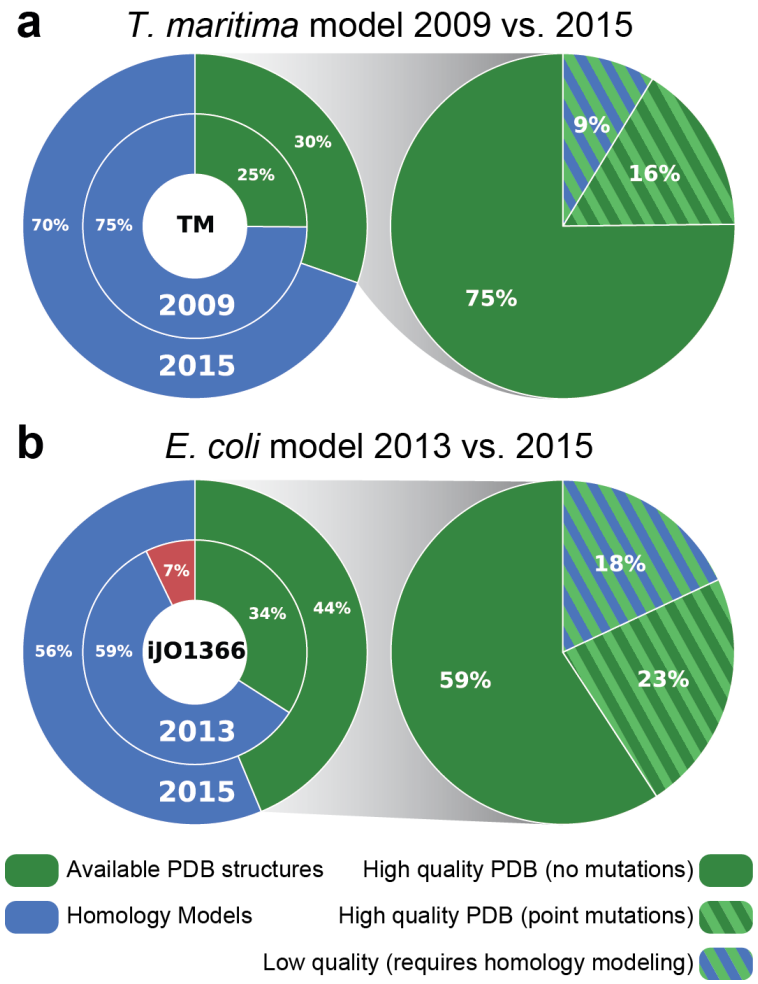



Figure 3.2. New GEM-PRO models for *T. maritima* and *E. coli*.
(a) The new GEM-PRO model for T. maritima (TM). Displayed in the pie chart on the left is the coverage of genes by a PDB structure or homology model, and a comparison of those structures available in 2009 versus 2015. In the pie chart on the right, the available PDB structures are further classified into three groups based on the overall quality of the structure: (i) high quality structures that have no mutations in the interior of the protein (112 genes involved in 210 reactions; in green); (ii) high quality structures that have some mutations and require minimal modification to revert back to wildtype sequence (24 genes involved in 49 reactions; in dashed green) and low quality structures (13 genes involved in 20 reactions; dashed green/blue) that may have large gaps of unresolved sections of the protein or a large number of mutations at the interior of the protein and require further homology modeling (in blue). Determining the quality of a PDB is explained in detail in the section entitled Quality control and quality assessment of all structures. The same quality assessment evaluations were carried out for *E. coli* in (b).

*Coverage of protein structures in metabolism*

We find that the coverage of all experimental (X-ray crystallography and NMR) protein structures (PDB) for genes in *T. maritima* and *E. coli* is between 30-45%, which is 6-10% higher compared to the original GEM-PRO reconstructions (Figure 2). The updated GEM-PROs for *T.*

*maritima* (*i*BM478-GP) and *E. coli* (*i*BM1366-GP) include 336 and 3425 PDB structures, respectively, an additional 5-10% of newly deposited protein structures compared to the original versions (see inner versus outer nested pie chart in Figure 2). Of the newly deposited protein structures, the majority are linked to subsystems in metabolism with a higher coverage of protein structures compared to others (e.g., alanine and aspartate metabolism, see Figure 3).

As shown in Figure 2, nearly 56-69% of genes in the GEMs cannot be mapped to available experimental protein structural information. To a large extent, the 3D structure of a protein can be estimated from homology modeling, which predicts structure based on experimental templates of proteins that are homologous in sequence to the protein of interest. Here, we selected the I-TASSER (iterative threading assembly refinement) suite of programs [180, 181], which has been the highest ranking program for automated protein structure prediction for the the past two CASP experiments [181-184]. Mapping the *E. coli* model to available I-TASSER homology models [176, 185, 186], we find that the coverage is nearly complete for its metabolic proteome (1343 genes have available template-based homology models and 23 have *ab initio* models [185]). For *T. maritima*, we have manually performed homology modeling using the I-TASSER protocol to generate models for a total of 333 genes lacking experimental protein structure information. We find that the updated GEM-PRO models make use of over 100 recently deposited (and higher quality) experimental structures compared to the previous models.

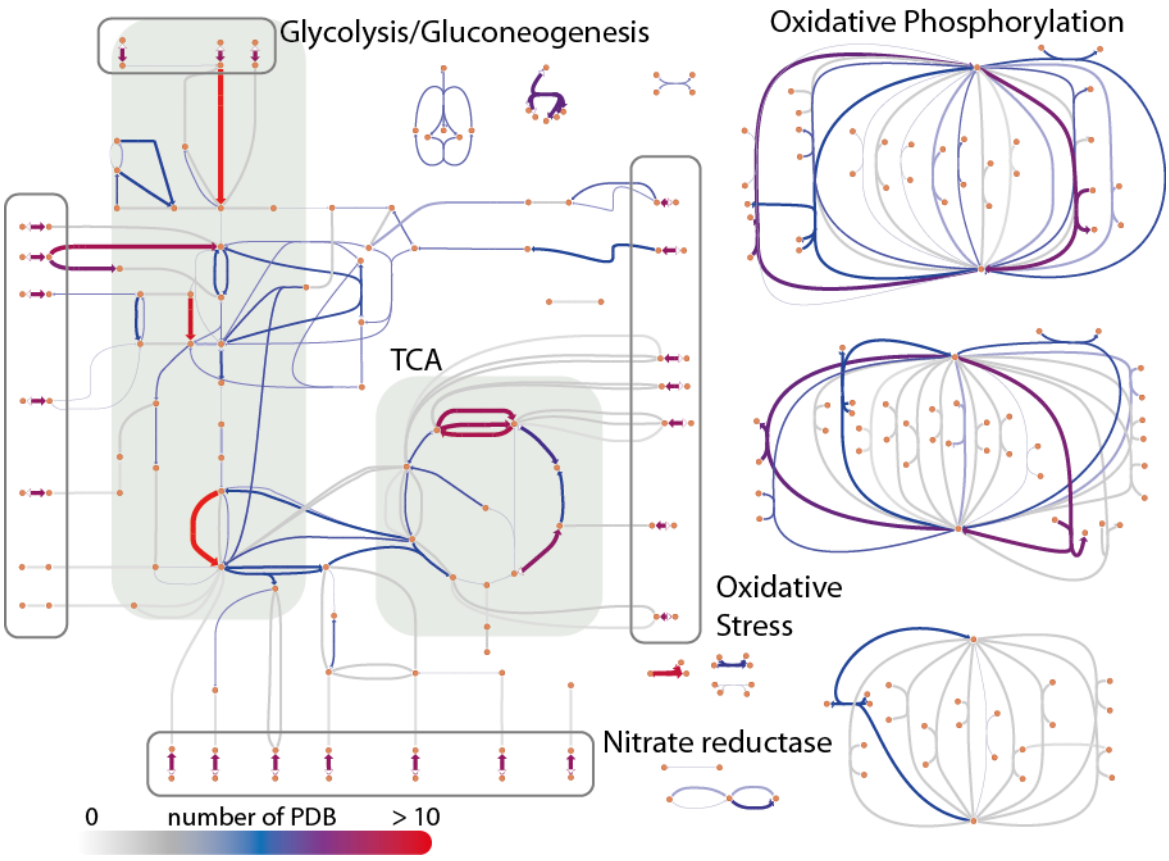



Figure 3.3. All available PDB structures mapped to the network of *E. coli* metabolism (iJO1366 model).
The heat map indicates an increase in the number of available experimental protein structures that map to a given reaction in the pathway (grey to blue to red transitions represents 0 to more than 10 PDB structures). Subsystems such as glycolysis and TCA are highlighted by the colored grey squares and transporters by transparent rectangles with grey borders. The largest increase in coverage in subsystems involved in alanine and aspartate metabolism, glycolysis and gluconeogenesis, folate metabolism, cysteine metabolism, the citric acid cycle, arginine and proline metabolism, tRNA charging, and nitrogen metabolism.

*Quality of experimental and homology-based structures*

In many cases, experimental protein structures may contain unresolved fragments of the protein or mutations in the sequence (often as artifacts or the result of a crystallization protocol or due to natural disorder). Small variations in sequence can have large-scale effects on the structure and function of proteins. Thus, we perform a rigorous assessment of the quality of all structural data for each model organism. To determine which experimental structures require further modeling (e.g., group iii proteins, displayed in Figure 2 (a) and (b)) or minimal modification (e.g.,

group ii proteins, displayed in Figure 2 (a) and (b)), we devised a scoring metric that ranks each PDB structure based on a set of criteria: the maximum coverage of the wild-type amino acid sequence, PDB resolution, and minimal number of missing or unresolved parts of the structure (Figure 4(b)).

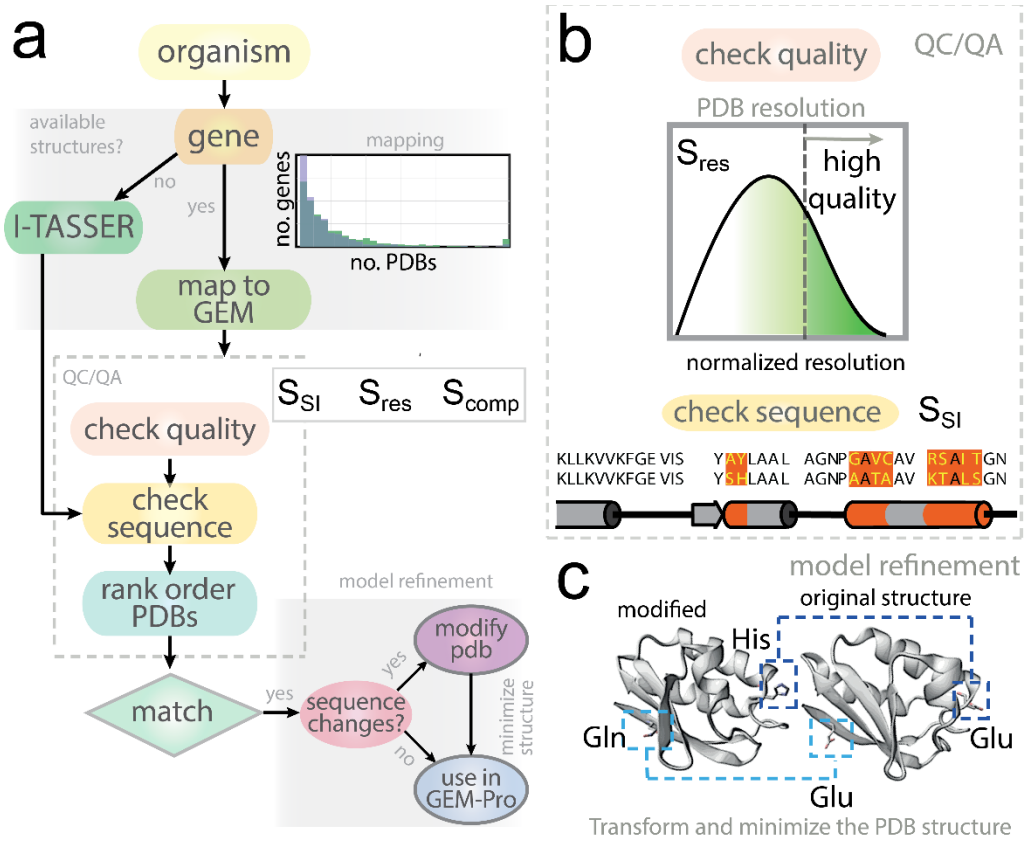



Figure 3.4. Workflow for generating simulation-ready models of all proteins in metabolism.
(a) The first stage involves mapping the genes of the organism to available crystallographic and NMR protein structures, found in the Protein Data Bank (PDB). The second stage performs homology modeling for genes without available structures. The third stage performs ranking and filtering of structures and homology models for each gene based on set selection criteria (e.g., $S_{SI}$, $S_{res}$ and $S_{comp}$). These criteria refer to a scoring metric that ranks a PDB structure based on sequence identity ($S_{SI}$), resolution ($S_{res}$), or homology model based on the similarity in secondary structure composition ($S_{comp}$) compared to the structure. As shown in (b), evaluation of the sequence identity between the protein structure sequence and that of the wild-type sequence and PDB resolution (in Å) allows filtering of low-quality structures. In the final stage, all high quality PDB files that require minimal modification (e.g., reversion of the sequence to match that of the wild-type) are further refined, as depicted in (c)

In the previous *E. coli* GEM-PRO, 43% of all proteins contained unresolved fragments. After carrying out the QC/QA pipeline, we correct for all cases and provide 100% complete (gap-less) and sequence identical structures of proteins. To further assess the quality of protein structures in the updated GEM-PRO, we have evaluated all structures using PROCHECK [187], which assesses the stereochemical quality of a protein structure, and PSQS (based on statistical potentials of the mean force between residue pairs and between solvent and residue) [188]. While the average quality scores for all protein structures in the updated versions of GEM-PRO are similar to those of previous versions, the completeness of all structural models in the updated GEM-PROs substantially enhances the quality of the structures in the model and their capacity for future applications.

*Structural and sequence refinement of structures*

The final step in the workflow (Figure 4 (c)) carries out minimal sequence modifications of nearly perfect, high-quality experimental structures (e.g., group ii proteins, displayed in Figure 2 (a) and (b)). Modifications of this set of structures are mainly needed to fix: (i) a minimal number of single-residue mutations (i.e., not more than two sequential mutations); or (ii) a minimal number of deletions or missing residues in the interior of the protein. This final step enables one of the most considerable improvements in the updated GEM-PRO framework, providing a complete set of minimally modified experimental structures that have 100% sequence identity to wild-type sequence. Using our PDB refinement pipeline (Figure 5), we find that 16% (24/136) and 23% (136/490) of experimental protein structures in the GEM-PRO of *T. maritima* and *E. coli*, respectively, require minimal modifications to revert the PDB sequence to the wild-type sequence. See Table 1 for details on average sequence identity and completeness.

Table 3.1. Quality statistics of all available protein structures in GEM-PRO models.
Mean sequence identity, completeness, and resolution refers to the average of the three metrics over all experimental protein structures in GEM-PRO. The standard deviation is given for each metric. Mean sequence identity refers to exact amino acid matches between sequence and structure, while mean completeness disregards exact matches.

| **Property** | ***T. maritima*** | ***E. coli*** |
|---|---|---|
| Mean Sequence Identity | 92.1 ± 15.8% | 91.8 ± 16.4% |
| Mean Completeness | 92.3 ± 15.5% | 91.9 ± 16.1% |
| Mean Resolution | 2.2 ± 0.5 Å | 2.3 ± 1.0 Å |

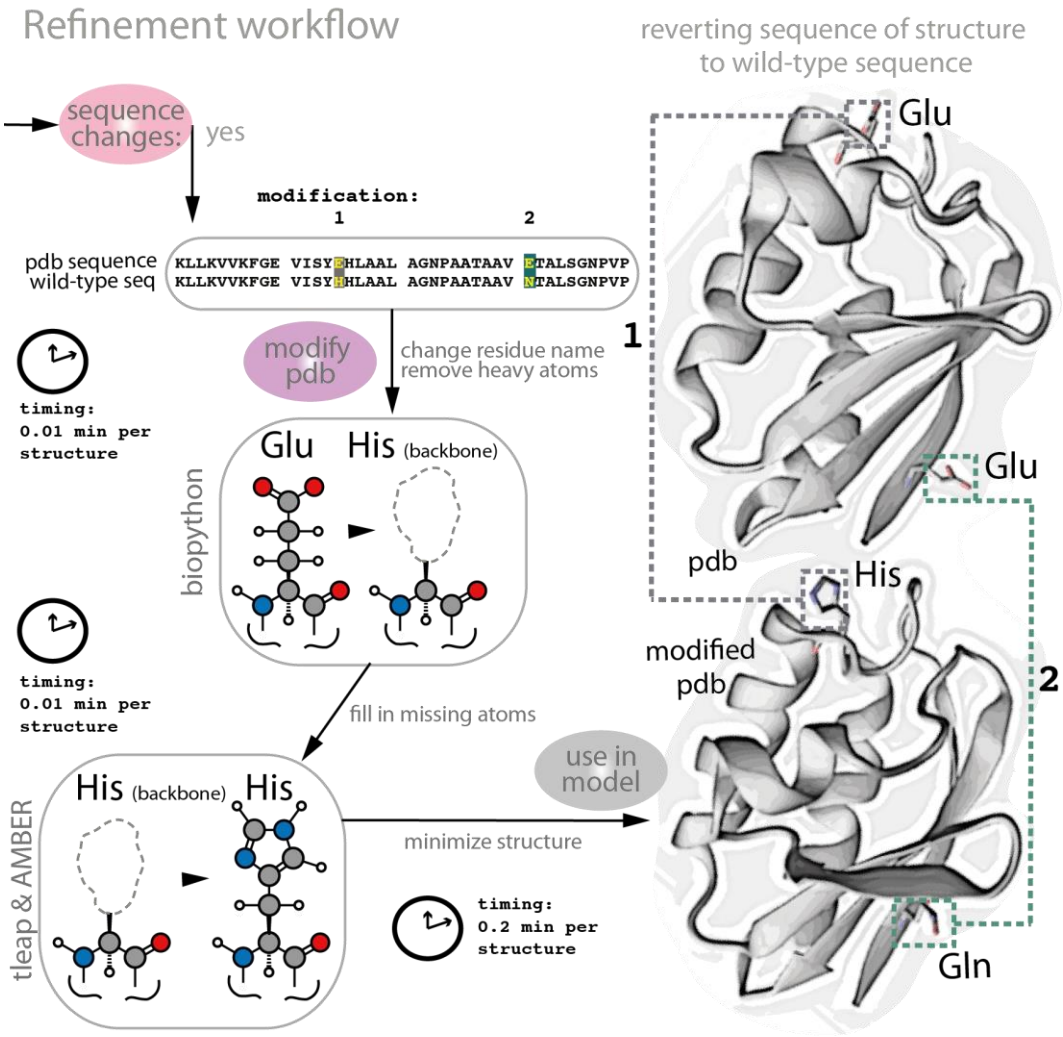



Figure 3.5. Final stage of refinement in the GEM-PRO pipeline.
This workflow demonstrates the final stage of refinement for PDB structures, performed to replace atomic coordinates of atoms in a mutated residue with atomic coordinates corresponding to the wild-type residue. Using a combination of Biopython modules and the AMBER suite of programs, each PDB structure is modified and the final structure is minimized. For example, an original crystal structure and its wild-type sequence differ by two residues (Glu115His and Glu131Gln). The modified structure is reverted back to the original wild-type sequence in three stages: (i) all atoms in the Rgroup of the target amino acid (except for the peptide backbone atoms) are stripped from the file; (ii) new atoms with their respective 3D atomic coordinates are placed in the “empty” amino acid ‘site’ (e.g., the R-group atoms of Glu); (iii) the modified structure undergoes energy minimization using a steepest descent algorithm to relieve any bad contacts (i.e., steric hindrance) that may be caused by the addition of new atoms

*Final outcome of mapping protein structures to genome-scale data*

The overall coverage and quality of the selected experimental and homology-based structures for each organism is detailed in Table 2. This database increases the scope and capacity of genome-scale models when applied within a model and data-driven workflow. As shown in Figure 6 (a), the combination of protein data (e.g. melting temperature) and a genome-scale model of metabolism can be used to predict the effect of temperature on the growth rate of a model organism. These *in silico* findings can then be tested with experiments to provide input into the next round of this iterative workflow.

Table 3.2. Quality statistics of GEM-PRO models.
[a]number of total genes with PDB structures (includes minimally modified) after QC/QA; [b]number of total genes with homology models. Note that there may be overlap between PDB and homology model coverage; [c]mean quality score of PDB structures in the GEM-PRO model for all available PDB structures. In parentheses are the subset of "best representative structures" for all metabolic gene (as ranked by the QC/QA pipeline), scaled (0, 1] where 0 is low quality and 1 is the highest quality; [d]mean quality score of the homology models taken from the I-TASSER TM-score metric, is the range [0,1] with a value >0.5 implying correct topology of the model.

| **Model** | **PDB coverage[a]** | **Homology model coverage[b]** | **PDB quality score[c]** | **Homology model quality (TM-score)[d]** |
|---|---|---|---|---|
| *T. maritima* | 136/478 | 342 | 0.82 (0.86) | 0.79 |
| *E. coli* | 490/1366 | 1366 | 0.77 (0.95) | 0.82 |

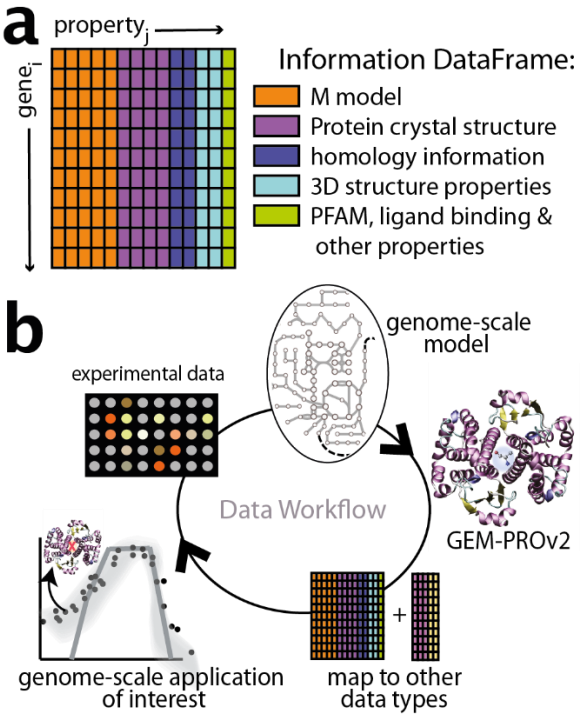



Figure 3.6. Storing GEM-PRO in a computable format.
(a) The master GEM-PRO data frame which stores various protein-related properties for a specified organism. (b) A proposed data workflow, in which a genome-scale model is integrated with protein structural information, thus forming a GEM-PRO which can then be mapped to other data types, such as melting point temperatures, and can subsequently be applied to genome-scale applications, such as predicting growth rate of *E. coli* at different temperatures. Finally, these in silico predictions are compared to experiments for validation

### 3.3.2 Modeling at the intersection of systems and structural biology

Once a GEM-PRO database has been constructed, it can be queried and used in conjunction with experimental data and genome-scale modeling approaches to understand the nature of the underlying biology. Here, we present four novel case studies which demonstrate how properties derived from the structures of proteins determine systems-level behavior.

*Characterizing the degree of diversity in substrate specificity of metabolic proteins*

Evaluating protein structural properties together with their binding capacities provides insight into structure-function relationships of isozymes and proteins that catalyze similar reactions. We are interested in using GEM-PRO to formulate hypotheses about which proteins are

most likely to act promiscuously on substrates other than their native one (i.e. substrate ambiguity). Assessing the degree of substrate ambiguity with EC numbers has been explored through evaluation of fourth digit of the enzyme commission number (e.g. 2.6.1.**X**) [189]. Here, we a different approach, we apply GEM-PRO to evaluate the degree of diversity in the substrates/ligands bound to crystallized proteins within various EC families.

Many enzymes in the transaminase family are known to be capable of dual substrate recognition [190, 191]. Querying GEM-PRO, we find that aspartate aminotransferase, aspC (2.6.1.1), and tyrosine aminotransferase, tyrB (2.6.1.57), are both pyridoxal 5' phosphate (PLP)-dependent enzymes, share a common Pfam (PF00155; Figure 7 (b)) and structurally align to give a high overlap of the substrate and cofactor binding sites. Structural properties such as these have been used to generate hypotheses about possible "underground" activities of enzymes, and some have been recently validated *in vivo* using an isozyme discovery workflow [162]. Extending the above analysis to the entire proteome, we are interested in addressing the question: "What is the degree of substrate specificity of proteins in a metabolic network?" Using the metabolic network models of *E. coli* and *T. maritima*, we find that both organisms have a subset of multi-functionality genes (i.e. genes that can catalyze more than one reaction); in *E. coli*, 4.4% (60) of metabolic genes are involved in multiple enzymatic complexes and in *T. maritima*, over 19% (90) are multi-functional. Although *T. maritima* has a higher degree of multifunctional peptides, the number of reactions with isozymes is consistent with that of *E. coli* (~30%).

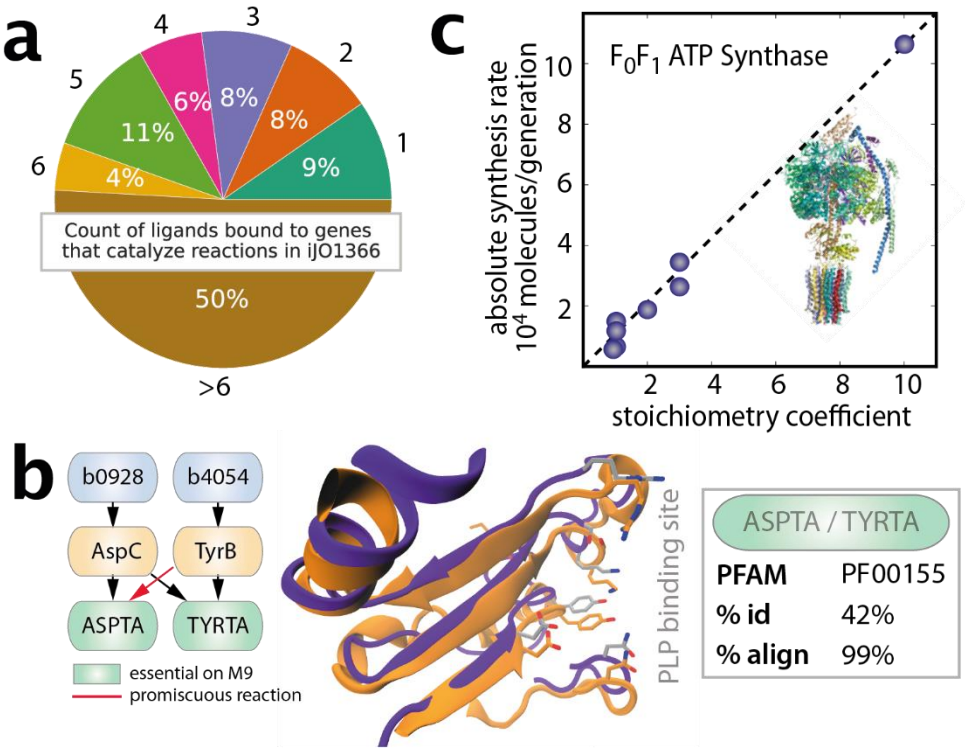



Figure 3.7. New structural systems biology applications using GEM-PRO.
(a) The counts of different ligands from the Ligand Expo database (PDB) that are bound to holoenzyme protein structures in the *E. coli* GEM-PRO model and are linked to catalytic metabolic reactions. (b) An example of a highly promiscuous family of enzymes, transaminases, which have been shown to rescue the activity of another protein when its respective gene has been knocked out. Pfam refers to shared protein fold family, '% id' refers to percent sequence identity, and '% align' refers to the 3D structural alignment of the two proteins. The plot in (c) demonstrates how the GEM-PRO model can be combined with experimental data, such as ribosomal profiling, to predict the in vivo abundance of proteins and their complex stoichiometry. The example shown here is that of ATP synthase, which indicates a high overlap between the complex stoichiometry stored in GEM-PRO and an experimental measurement

Protein structures of holoenzymes (i.e. proteins co-crystallized with cofactors or substrates/analogs) also provide a wealth of information on different protein-ligand interactions, as they can be directly compared to existing enzyme-substrate interactions in the metabolic network. We analyzed proteins bound to a representative set of compounds present in metabolism (e.g., not bound to glycerol, non-catalytic water molecules, or other types of detergents). To filter the large majority of these cases from the dataset, we classified the types of ligands bound to protein structures, which clusters ligands using a fast heuristic graph-matching algorithm [192, 193]. The type of ligand bound to a protein structure is grouped into different superclasses (e.g.

lipids, amino acids, sugars, antibiotics), by comparing discriminating factors, such as the atom element, chirality, valence, and/or bond order [192]. After filtering the ligands into metabolic (and non-metabolic) superclasses, we find 39% of the total genes in the *E. coli* GEM-PRO model are representative holoenzymes (26% of *T. maritima* genes). Surprisingly, we observe a large amount of metabolite binding versatility in *E. coli*, as 50% of holoenzymes are experimentally shown to bind more than six different ligands (i.e., in different crystallographic structures of the same protein, see Figure 7 (a)). Each metabolite was described (according to its metabolite fingerprint similarity using Tanimoto coefficients [194]) and these coefficients were compared across the set of ligands bound to a given protein to determine the degree of variation in substrate specificity. We find that certain classes of enzymes, such as transferases (EC 2.-.-.-.), are only bound to very similar metabolites (which consistent between *E. coli* and *T. maritima*), whereas lyases (EC 4.-.-.-.), are bound to the most structurally diverse set of substrates.

*Protein complex stoichiometry predicts in vivo enzyme abundances*

Does protein complex stoichiometry determine *in vivo* enzyme abundance? Previous work using ribosome profiling techniques revealed that multiprotein complexes have proportional synthesis rates [195]. This is both interesting and important because catalysis or activation of proteins is dependent on the proper complex formation of a specific number of homo- or hetero-subunits. Here, we apply a complementary approach, using genome-scale modeling of metabolism in conjunction with ribosome profiling data to identify which protein abundances are constrained by complex stoichiometry and which have higher free protein abundances.

Information about the stoichiometry (or ratio) of genes in the respective enzyme complex (and its functional properties) is found in organismal [196, 197] or protein databases [198] and can be directly incorporated into GEMs (e.g., in the annotated gene-protein-reactions, or GPRs). GPRs

link a set of genes to the metabolic enzyme to the catalyzed reaction, providing a starting point for the reconstruction of enzyme complex stoichiometry. In this section, we discuss how to predict enzyme abundances, identify peptides that are not expressed stoichiometrically, and predict the partitioning of peptides across the multiple complexes to which it belongs. To associate the metabolic reactions with structures of their catalyzing enzymes, we integrated GEM-PRO together with the genome-scale models of metabolism and expression (ME-model) for *E. coli* [199]. The coverage of complex stoichiometry is relatively complete (95%). We find that the majority of metabolic enzymes are homomers (90.3%), for which, we see a strong preference for even stoichiometry. This is consistent with general trends among homomeric complexes towards even stoichiometry, and has been explained based on the ability of complexes with even stoichiometry to form complexes with dihedral symmetry as well as rotational symmetry [200]. Furthermore, we find that 4.4% (60) of metabolic genes are involved in multiple enzymatic complexes and 30% of reactions are catalyzed by isozymes.

Coupling information from genome-scale reconstructions, known enzyme complex stoichiometry, and ribosomal profiling data, we can predict *in vivo* protein abundance in *E. coli.* As depicted in Figure 7 (c), this novel framework can be applied to identify and predict protein complex stoichiometry [195, 201]. As illustrated in the Supplementary IPython notebook, "Complex_Stoichiometry.ipynb" (available in the online resource), protein complex stoichiometry can be converted into a computable (mathematical) format for validation with experimental ribosomal profiling data [195, 202]. A protein stoichiometric matrix is assembled in which the rows represent proteins, the columns represent enzymes, and the entries indicate the stoichiometry of the protein within the enzyme (akin to a stoichiometric matrix of metabolism, used in GEMs [161]). This matrix, combined with quantitative data on protein expression [203, 204], can then be

used to determine feasible enzyme (and free peptide) abundances using constraint-based modeling methods [117] and available software [205, 206]. We find that the maximal and minimal enzyme abundances, computed using flux variability analysis (assuming free peptide abundance are minimized) indicate that enzyme abundances are quite constrained by stoichiometry alone. Interestingly, we find that many of the proteins with the largest free abundances are periplasmic substrate binding proteins. These proteins are not always in complex with the transporter protein itself and, therefore, are not produced stoichiometrically with the rest of the transporter complex, making their abundances less constrained.

*Comparative systems biology of different bacterial proteomes*

To date, there has been a great deal of attention placed on understanding the genetic differences between *T. maritima* and other Eubacteria [207-213]. Whole-genome similarity comparisons indicate that *T. maritima* is the most Archaea-like organism compared to other eubacterial species [207-213], with 24% of genes appearing to be more closely related to archaeal genes [213, 214]. Less attention, however, has been focused on characterizing the differences between proteomes of species. Of the studies that evaluate protein-level differences, many have focused on families of proteins [215, 216], and few have focused on comparing proteins that span across entire metabolic networks. The novelty of using GEM-PRO for comparative studies is the ability to map genes to their gene products (proteins) to the reactions they catalyze within a single database. Such a mapping allows for high-level structural comparisons of *functionally* relevant sets of genes: homologous genes, genes that catalyze more than one reaction (i.e. promiscuous), genes that catalyze similar reactions (i.e. isozymes) and genes with high sequence or structural similarity. Here, we apply GEM-PRO to address the question, “How different are bacterial proteomes and what are the main properties that distinguish them?”

The first notable difference, when comparing GEM-PROs of *E. coli* and *T. maritima*, is the spread of molecular motifs across metabolic proteins, which greatly distinguishes the two proteomes from one another. We used the Flexible structure AlignmenT by Chaining AFPs (Aligned Fragment Pairs) with Twists (FATCAT) [25] algorithm to detect all of the aligned fragment pairs (AFPs), based on previous PDB-wide alignment of representative protein domains [217]. The observed AFPs are regions of a protein that cluster based on similarities in local geometry and take into consideration protein flexibility by clustering regions of the protein that can undergo different geometric transformations. Considering all proteins in both the *E. coli* and *T. maritima* GEM-PRO models, we found a total of 874 and 197 unique domains (according to SCOP or PDB-based annotations), respectively, which span the whole of metabolism (i.e. 1819 total protein structures). We find that 36 domains are shared between *T. maritima* and *E. coli*. Furthermore, comparing the distribution of complex stoichiometry between *E. coli* and *T. maritima*, we find that for both organisms, the majority of metabolic enzymes are homomers (90.3% and 71.1%, respectively).

To understand whether the properties of entire proteomes are distinguishable between organisms, we carried out PCA on 29 computed secondary structural properties (see Figure 8 (a)). The projections of the first two principal components explain 60% of the normalized property distribution. Using K-means clustering, we find that protein properties separate into four discrete clusters (based on the percent variance within clusters). The main difference between the clusters of proteins is the percent composition of secondary structural elements, such as $\alpha$-helical and $\beta$-extended strand, solvent-accessible surface area and percentage of charged residues (Figure 8 (b)). For example, in one cluster (‘1’), 64.7% of amino acids are found in $\alpha$-helices. A correlation matrix derived from the properties of proteins in this cluster indicates that the majority of residues

found in $\alpha$-helices also have higher percentages of hydrophobic content while other residues found in $\beta$ strands are highly charged. The majority (155 out of 247) of this cluster of proteins are membrane-bound proteins, which are known to have distinguishing exterior domains [218, 219], and correlate based on a preference for $\alpha$-helices and a neutral surface charge, compared to those proteins in other clusters.

As illustrated in Figure 8 (c), the percentage of the proteome in each of the four clusters differs between organisms; certain clusters are present (or enriched) in only one of the organisms (such as cluster 0 for *E. coli* and cluster 2 for *T. maritima*). Comparing the unique aspects of proteins within each of the clusters, we find that certain characteristic features distinguish proteins based on their metabolic roles as well as based on which organism they belong to. For most clusters, proteins belong to a single (or a select few) subsystem(s), which suggests that these features may play a role in self assembly and cellular localization. For example, comparing the second and third clusters (1 and 2), many of the members (over 70%) function as transport proteins versus alternative carbon metabolism and cofactor biosynthesis (33%). For differences between proteomes, we find that the first cluster (0) consists of only *E. coli* proteins (Figure 8 (c)), which are enriched in surface-exposed residues and tend to be polar or positively charged (Figure 8 (b)). However, in the third cluster (2), we find an increased number of thermophilic proteins compared to the number of mesophilic proteins with a higher degree of buried, nonpolar residues, and are less polar and solvent accessible. This is consistent with what is generally known about protein stability [220], such as those dominated by forces that drive protein folding (e.g. the burial of nonpolar groups, increased number of hydrophobic interactions and decreased solvent accessibility).

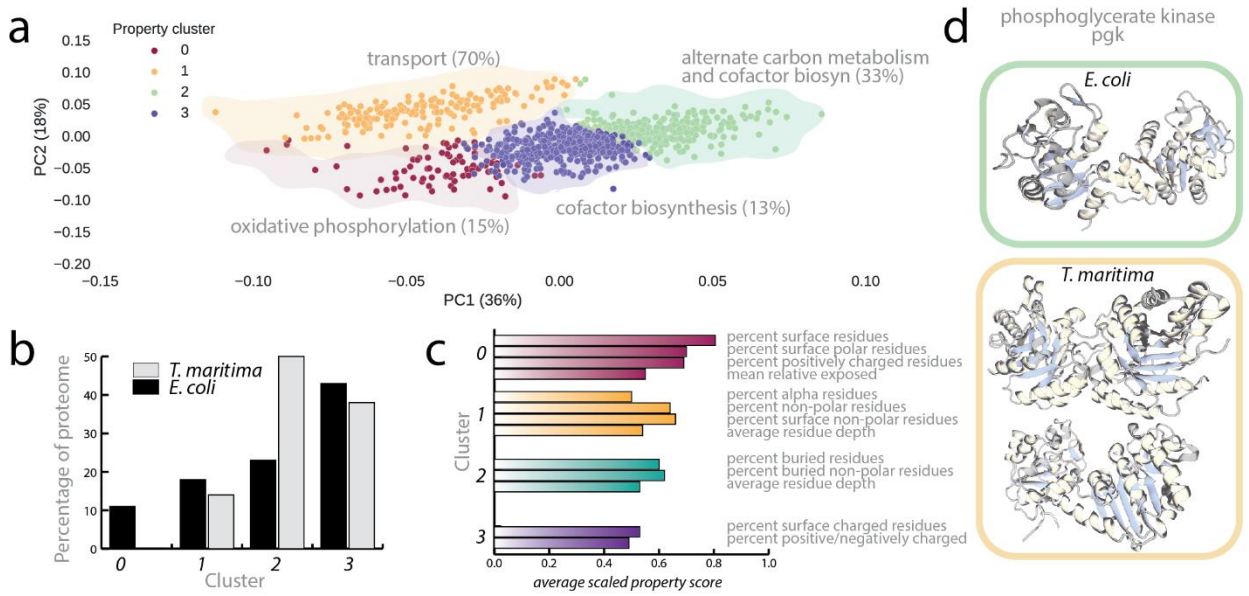



Figure 3.8. Distinguishing properties of entire proteomes.
In (a), K-means clustering of all *E. coli* and *T. maritima* protein structural properties (29 features, including SASA, percent polar, nonpolar, buried, surface, charged residues and others). The K-means clustering algorithm clusters all proteins into four distinct clusters (based on the percent variance explained per cluster using the elbow method, see Additional file 1). Interestingly, metabolic subsystems in *E. coli* show distinct structural characteristics in their respective proteins. The subsystem with the most proteins in a given cluster is reported. In (b), we report the main structural characteristics that distinguish proteins across clusters. The numbers represent averaged scaled property values across all proteins within a given cluster (see Additional file 1). The property values generally represent the percentage of the protein that is described by a given property (e.g., percentage of the protein which is nonpolar). In (c), the percentage of *E. coli* and *T. maritima* proteomes within each cluster are shown. Surprisingly, certain clusters are enriched in *E. coli* proteins (cluster 0) and certain in *T. maritima* proteins (cluster 2). Total numbers of proteins in each cluster are 154, 318, 592, and 763 for cluster 0–4, respectively. In (d), an example of a homolog (pgk) which is present in entirely different clusters (cluster 2 for *E. coli* and cluster 1 for *T. maritima*). The structural differences can mainly be explained by the fact that in *T. maritima*, pgk (PDB 1VPE) is fused with tpi (PDB 1B9B), creating a protein which is triple in length to that of its *E. coli* counterpart (PDB entry 1ZMR).

*Characterization of proteins with growth rate-limiting reactions at high temperatures*

High temperatures impose a heavy burden on organisms with respect to the functioning of cellular metabolism. Understanding the molecular basis for stability is necessary to grasp the the fundamental nature of protein structure as well as to engineer high-temperature industrial processes [221]. In general, structure-based analyses have been used to discover properties of thermostability [15, 16, 220, 222], however, there remains a significant challenge to pinpoint which characteristic features of proteins lead to detectable differences between thermophiles and mesophiles [223, 224]. Using an entirely different approach, genome-scale models of metabolism point to specific proteins that limit the ability of the cell to grow and function at a given temperature [32]. For example, specific *E. coli* proteins, identified as "hotspots," are linked to reactions in the metabolic network that limit or diminish the cellular growth rate at higher temperatures (e.g., due to protein unfolding/degradation). The novelty of this approach is that we can hypothesize that these "hotspot" proteins are under selective pressure (on the basis of how important their function is to the entire metabolic network) and require adaptation to function at higher temperatures. Here, we are interested in understanding what characteristic molecular properties of homologs in *T. maritima* are different, relative to their *E. coli* counterparts, to ensure functional proteins at higher temperatures.

For the identification of homologs, we took advantage of the extensive database of both GEMs to effectively map between *E. coli* and *T. maritima* genes that have a similar sequence and metabolic function (a total of 219 homologs). In this case, we clustered alignments of *E. coli* with *T. maritima* PDB templates into three classes (high, medium, and low-medium overlap) based on the root-mean-squared-deviation (RMSD) of the protein backbones (less than 5 Å, 5-7 and 7-10 Å, respectively) and an alignment coverage of greater than 70% of the total length of the protein.

Surprisingly, we find that, out of 219 homologs, only 10% (19) of *E. coli* genes share a structurally similar domain with their *T. maritima* homologs (all cases align with RMSD < 5 Å). Of the 10% that are structurally similar, we linked their respective metabolic functions to amino acid biosynthesis, cofactor biosynthesis, or cell envelope biosynthesis. A few cases related to tRNA and methionine metabolism also show a high degree of structural similarity, despite low nucleotide sequence identity (e.g. b3559/TM_0216 have 30.1% sequence identity and b4019/TM_0269 have 27.8% sequence identity).

Particularly interesting cases pulled out from this analysis are those of 3-phosphoglycerate kinase (pgk, EC 5.3.1.1) and the *b* subunit of atp synthase (atpB, EC 3.6.3.14). Comparing the extremely stable thermophilic pgk with its less stable, mesophilic homolog reveals that this peptide correlates to proteins in cluster 2, whereas the thermophilic pgk correlates to proteins in cluster 1. The crystallographic structure of the thermophilic pgk shows increased rigidity from the many intramolecular contacts, alpha helices, and loop regions [225] consistent with cluster 1 properties. Furthermore, the size of the *T. maritima* pgk is three times that of its *E. coli* counterpart (280 kDa versus 43 kDa), as it is a tetrameric fusion protein (pgkfus) of two enzymes, namely pgk and triosephosphate isomerase (tpi, 2.7.2.23), illustrated in Figure 8 (d), bottom. Despite a difference in relative enzyme efficiency, the fusion protein is active when previously cloned and expressed in *E. coli*, confirming the authenticity of the two separable proteins and enzyme activities resulting from this gene in the mesophilic host [226]. In this context, covalent fusion of two proteins to complexes or assemblies might represent an additional stabilization strategy, particularly for "hotspot" enzymes that become unstable at higher temperatures, like pgk.

Structural comparison of the β subunit of ATP synthase polypeptides indicates that the *T. maritima* protein has a higher degree of buried, nonpolar residues that, on average, are less solvent

exposed (i.e., a larger average residue depth of the alpha carbon atoms in the protein). In contrast, the *E. coli* peptide is much more solvent exposed and its residues are, on average, more polar or positively charged. A previous study, which characterized the chimeric soluble β polypeptides *in vitro* showed that the *T. maritima* protein melted cooperatively with a midpoint more than 20°C higher than that of the *E. coli* sequence [227]. The study revealed the effects of substituting different sequences in the *E. coli* peptide, showing which parts of the peptide tolerated the most change without a loss of function and which changes led to an increased thermostability. The structural differences brought out by this pairwise comparison are consistent with the fact that the average relative contact order (which correlates to solvent accessibility) of *T. maritima* proteins is significantly different than their close mesophilic homologs [224].

### 3.3.3 Dissemination of GEM-PRO and development of new training resources

Equally important to providing higher quality models is providing the community with complete knowledge bases, tools, and training examples for the continuous development of genome-scale modeling approaches. Historically, advances in genome-scale modeling have been accelerated by the wide dissemination of network reconstructions, modeling methods, and their continual curation and updating to incorporate new information. Furthermore, as GEM-PRO enables modeling of cellular processes that span a wide range of biological, chemical, and structural detail, input from different scientific disciplines could vastly enhance the capabilities of current methods and approaches used in systems biology. To make GEM-PRO accessible to a wide-range of scientific backgrounds, we present GEM-PROs workflows for these two contemporary organisms, *E. coli* and *T. maritima*.

As supplementary information, we describe how various protein-related data types are paired with GEMs (Figure 6). We provide bioinformatics scripts together with tutorials (in the

form of IPython notebooks) as Supplementary Information (available in the online resource) to explicitly describe how protein-related information can be linked to genome-scale models to study: (i) the evolution of protein fold families in metabolism; (ii) temperature-dependent growth rate predictions; (iii) the diversity in protein-ligand interactions in a metabolic network; (iv) the organization of protein complex stoichiometry and how it can be paired with ribosomal profiling data to describe *in vivo* protein abundance.

## 3.4 Conclusion

Protein structures and their molecular assemblies offer a wide range of possibilities to further enhance the predictive scope of genome-scale modeling by providing information on the sequence of molecular events in a pathway, how to interfere with a pathway to treat a pathology, or the evolutionary history of contemporary organisms. The further integration of protein-related data into metabolic network reconstructions will rely on clear mapping protocols and the development of bioinformatics tools that will aid in this process. This contribution, the bioinformatics tools, and the accompanying tutorials, which are based on constraint-based modeling methods through COBRApy [206], describe the generation and application of GEM-PRO models. Here, we have shown the utility of integrating molecular scale analyses with systems biology approaches by discussing several comparative analyses on the temperature dependence of growth, the distribution of protein fold families, substrate specificity, and characteristic features of whole cell proteomes.

The dissemination of the GEM-PRO modeling framework is likely to broadly impact work in a wide array of disciplines, including structural biology, computational chemistry, systems biology, and biotechnology. The ability to characterize the structural, chemical, and binding

characteristics of metabolic proteins in different organisms also enables the further development of *in silico* tools capable of identifying isozyme activity on a genome-scale. Recently, a number of studies have emerged [162, 228, 229] that have used genome-scale models together with complementary bioinformatics techniques to characterize the versatility of enzymes on a systems level. Such studies can easily be extended to include the assessment of protein structural data and can be used to complement current "gap-filling" methods [230, 231] for model improvement. Current "gap-filling" methods typically use amino acid sequence identity as a measure for predicting enzyme similarity. However, some candidates are likely to be overlooked, since proteins with low sequence identities (e.g., <15% in the globin family) have also been shown to share similar folds and functions [215, 232, 233]. Evaluating the capacity of a protein to catalyze more than one reaction is also especially important to applications in metabolic engineering [234-236], where such proteins serve as an ideal starting platform for engineering novel capabilities as well as increasing substrate specificity.

Finally, GEM-PRO offers insight into the physical embodiment of an organism's genotype and provides a new way to compare genomes by linking genes to their encoded gene product, to the protein's structure, and finally, to the reaction catalyzed by that protein (or its molecular assembly). The use of GEM-PRO as a comparative systems biology approach demonstrates that important aspects of the functional differences between organisms (e.g., due to lifestyle changes) are not only derived from differences in their genetic components but also from the physical interactions of their molecular components. Together with previous applications on the phylogenomic analysis of protein structure [237], global motifs on protein fold and domain architecture [238, 239], and evolution of modern metabolism [28, 240], mapping the properties of proteins to their respective genes offers a novel perspective of the molecular, biochemical, and

phenotypic features of contemporary organisms. This comparative framework enables exploration of adaptive strategies for these organisms and opens the door to many new lines of research, including metabolic engineering and the design of thermostable enzymes.

## 3.5 Methods

### 3.5.1 Data retrieval and manipulation

Incorporating protein-related information into a GEM involves four stages of semi-automated curation: (i) map the genes of the organism to available experimental protein structures, found in publicly available databases, such as the Protein Data Bank (PDB); (ii) determine genes with and without available protein structures and perform homology modeling using the I-TASSER suite of programs [181] to fill in gaps where crystallographic or NMR structures are not available; (iii) perform ranking and filtering of PDB structures for each gene based on a set selection criteria (e.g., resolution, number of mutations, completeness); (iv) map GEM genes to other databases (e.g., BRENDA [241, 242], SwissProt [243], Pfam [244], SCOP [245]) for complementary protein-structure derived data. The quality of the reconstruction expansion process to include high confidence protein structures is considered by carrying out a series of QC/QA verification steps during the ranking and filtering stage. The GEM annotation of the organism of interest is stored in SBML and Matlab formats and many organisms can be found in the BiGG database [161]. Amino acid sequence of the proteins of interest are stored in FASTA format. To map protein structural data to a GEM, we make use of Python modules, ProDy [38, 246] and Biopython [19] to parse information in the PDB files. The molecular visualization software VMD [247] was used for viewing the 3D structure of the modeled protein and the predicted functional sites and the creation of images. Installation of PfamScan and HMMER3 algorithms are required

for generating protein fold families for certain proteins [248, 249]. Open source software for protein structural predictions are available and are used in conjunction with the IPython framework.

### 3.5.2 Data organization into IPython notebooks

In the Supporting Information (available in the online resource), we provide discrete examples of how to use the expanded metabolic network reconstructions with protein information to predict cellular phenotypes, which include (i) the discovery of multimeric properties of metabolic enzymes; (ii) the predicted growth of *E. coli* at different temperatures; (iii) predicting the effects of antibacterial drugs in *E. coli*; (iv) the discovery of patterns in fold families distributed across the metabolic network in *E. coli* and (v) the discovery of ligand similarity and potential for promiscuity in the metabolism *E. coli*. The tutorials provided in Supporting Information are designed in such a way that aids the user to properly access information in the GEM-PRO database, easily reproduce previously reported findings and organize information into meaningful representations. The main objective of the designed framework is to assist in (i) mapping between useful and unique identifiers; (ii) locate and query various data sources and (iii) identify fruitful and meaningful associations between the disparate datasets. We provide tutorial-like IPython notebooks as a means to organize the output of the database into easily manageable and understandable modules. Such a framework is the first of its kind for constraint-based modeling and provides full details that can be reproduced and updated as new data becomes available.

### 3.5.3 Homology modeling framework

The I-TASSER protocol is described by the following steps: (i) for each protein of interest, homologous templates are identified and used to assemble the queried protein; (ii) modified Monte

Carlo based replica exchange simulations are performed to cluster the lowest-free energy states of the assembled structure; (iii) the fragment-based assembly simulation is performed a second time to further refine the model and remove steric clashes; (iv) the function of the query protein is inferred by structurally matching the predicted 3D models against the proteins of known structure and function in the PDB. In order to assess the quality of the predicted structure, the accuracy is predicted from a confidence score (C-score or TM-score), which is defined based on the quality of the threading alignments and the convergence of the assembly refinement simulation used in steps ii and iii. I-TASSER is capable of generating multiple model predictions with a rank-ordered C-score. For more details about I-TASSER, please refer to the published literature [14].

### 3.5.4 Prediction of Pfam family folds (HMMER)

The database currently maintains 14,831 manually curated entries in the current release and is accessible via web servers (http://pfam.sanger.ac.uk/ and http://pfam.janelia.org/). This information allows for the classification of proteins via amino acid sequence into distinct protein families who share domain architecture through the HMMER suite of programs [250]. The challenges of predicting protein families using HMMER3 are discussed elsewhere [251]. For the genes in our models without Pfam annotations, we have run the freely available HMMER source code [248, 249] to fill in the "gaps" in the Pfam knowledgebase.

### 3.5.5 Temperature-based predictions in the *E. coli* metabolic network

Temperature-related properties of proteins (e.g., melting point temperature or $T_M$) were determined using both experimental and predicted values for the melting temperatures of proteins. The two main sources of this experimental data were taken from ProTherm [252] and BRENDA [242] online data services. By querying ProTherm and BRENDA temperatures of specific

metabolic proteins were linked to metabolic genes via their respective EC number. In the Supplementary Information , we have provided a script that performs the direct mapping between Blattner number and EC for querying both ProTherm and BRENDA databases (see the Supplementary IPython notebook titled, "Predicting Growth Rate at Various Temperatures", available in the online resource). For the *i*JO1366 model of *E. coli*, we find low coverage of temperature related data (only 29 out of 1366 genes with automated querying and 193 genes with semi-automated and manual curation). Thus, the experimentally determined $T_M$ values were supplemented with predicted $T_M$ using a previously published method [253]. We provide an example of one out of the four bioinformatics-based computational prediction of $T_M$ which derived from the amino acid sequence.

### 3.5.6 Reconstruction of protein complex stoichiometry

We updated the reconstruction of complex stoichiometry of enzyme complexes that catalyze metabolic reactions to include over 500 new complexes. We have included the list of added reactions together with the nearly complete mapping to complex stoichiometry in the Supplementary Information (available in the online resource). Metabolic models contain gene-protein-reaction relationships (GPRs), which are boolean statements on the requirements of genes for catalysis. However, more detailed reconstructions that include protein structures and models of metabolism and protein expression (ME-Models) benefit [199, 254] from information on enzyme stoichiometry. While the previous versions of GEM-PRO [28, 255] included information on single protein chains and protein complexes (using information both experimentally determined and putative PISA predictions [256]), the updated GEM-PRO extends the coverage to include additional data derived from experimentally determined enzyme complex stoichiometry. There are several additional sources of data on the stoichiometry of proteins in complexes, including PDB

structures and protein gels; much of this data is already compiled in databases such as Ecocyc [196, 197] or UniProt [198]. Experimentally determined structures and structures from homology modeling were used to achieve 93% structural coverage of proteins in the *i*JO1366 network and between 24% and 33% coverage of protein-substrate binding conformations. Manual curation for enzymes and metabolic reactions that do not perfectly match between the M-Model and databases is necessary. This procedure was performed by O'Brien *et al. [199]* starting from the *i*JO1366 metabolic model and mapping to the enzyme annotation in EcoCyc [196].

### 3.5.7 Calculation of protein 3D structural properties

We calculate 29 physical properties of the protein to construct a multidimensional data matrix, including solvent-accessible surface area (SASA), number of total contacts, disulfide bond distance (SS-bond), percent of the protein that is buried, percent of the protein that is on the surface, secondary structure composition ($\alpha$-helical content, $\beta$-strand content, $3^{10}$ helix content, $\pi$-helix content, hydrogen bonded turn content, bend content, disordered content), ovality ($SASA/N_{res}^{2/3}$), residue depth (distance of the C$\alpha$ atom from the protein surface), percent of the total structure that is nonpolar, polar, positively charged, or negatively charged, and percentage of the surface/buried residues that are nonpolar, polar, positively charged, or negatively charged. SASA was calculated according to the algorithm of Lee and Richards [257, 258] with a probe radius of 1.4 Å. Residues with a SASA measurement greater than 3 $Å^2$ are assigned as surface residues. The residue depth has been calculated for all atoms in the entire protein based on Michel Sanner's Molecular Surface (MSMS) method [12] and is evaluated from the average distance of all atoms to the surface of the protein. The number of disulfide-bonds is calculated from the 3D coordinates of sulfur atoms (using a 5 Å bonding distance cutoff).

## 3.6 Author Contributions

All authors have read and approved the final version of the manuscript. Conceptualization, E.B, N.M., and B.O.P.; Methodology, E.B., N.M., J.M., R.L.C., P.E.B.; Investigation, E.B., N.M., J.M., S.E.B., E.J.O., Z.Z. and K.C.; Writing – Original Draft, E.B., and B.O.P; Writing – Review & Editing, E.B., N.M., J.M., E.J.O, B.O.P., P.E.B.; Funding Acquisition, B.O.P.; Resources, B.O.P.; Supervision, B.O.P.

## 3.7 Acknowledgements

The authors acknowledge support from the Swiss National Science Foundation (grant p2elp2_148961 to E.C.B), the Gordon and Betty Moore Foundation GBMF 2550.04 Life Sciences Research Foundation postdoctoral fellowship (to R.L.C.). We also acknowledge funding from the National Institutes of Health (grant GM057089). This research was supported in part by the Intramural Research Program of the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (support to S.B. and P.B.). The authors also gratefully acknowledge NERSC supercomputer facilities and Ali Ebrahim for technical support.

Chapter 3, in full, is a reprint of the material as it appears in: Brunk E*, Mih N*, Monk J, Zhang Z, O'Brien EJ, Bliven SE, Chen K, Chang RL, Bourne PE, Palsson BO. 2016. Systems biology of the structural proteome. BMC Syst. Biol. 10:26. http://dx.doi.org/10.1186/s12918-016-0271-6. The dissertation author was the primary author (equally contributing with Elizabeth Brunk) of this paper.

# Chapter 4.

# Adaptations of *Escherichia coli* strains to oxidative stress are reflected in properties of their structural proteomes

## 4.1 Abstract

The growing availability of the three-dimensional structures of proteins has enabled the reconstruction of a reference structural proteome for *Escherichia coli* K-12 MG1655. This reference enables us to generate a comprehensive map of genetic changes in 1,764 strains of *E. coli*, by mapping their DNA sequence variations to 4,118 proteins with available 3D models. Metabolic model simulations enable predictions of their basal reactive oxygen species (ROS) production levels (ROStype) for 695 strains, and surprisingly, strains predicted to have higher as well as lower basal levels share antioxidative properties within their structural proteomes. We show that it is possible to assess a strain's sensitivity to an oxidative environment, based on known chemical mechanisms of oxidative damage to groups of proteins defined by their localization and functionality. Two general protein classes - metalloproteins and periplasmic proteins - show the clearest enrichment of antioxidative properties between the 695 strains with a predicted ROStype as well as 116 strains with a defined pathotype. Specifically, proteins that a) utilize a molybdenum ion as a cofactor and b) are involved in the biogenesis of fimbriae show striking differences in these antioxidative properties. We thus demonstrate that structural systems biology enables a

proteome-wide, computational assessment of changes to atomic-level physicochemical properties and of oxidative damage mechanisms for multiple strains in a species. This integrative approach opens new avenues to study adaptation to a particular environment based on physiological properties predicted from sequence alone.

## 4.2 Introduction

Reactive oxygen species (ROS) can cause severe oxidative damage to cellular proteins, which often results in chain reactions that spread the damage to neighboring macromolecules and leads to systems-level changes of cellular function [259]. While ROS can be beneficial - and even necessary in some contexts [260-263] - at a high concentration, oxidative damage must be responded to and repaired. As a result, cells are equipped with a number of mechanisms to quench ROS, directly repair the damaged components, or manage ROS indirectly [264]. Certain amino acids that comprise the structures of proteins are principal sites of oxidative damage [265]. This damage can be divided into two groups - reversible or irreversible modifications. The sulfur-containing residues methionine and cysteine incur reversible modifications, while irreversible damage impacts histidine, arginine, lysine, proline, threonine (RKPT), and tyrosine (with reactive nitrogen species) [266]. The damage to the RKPT amino acids is a post-translational modification known as carbonylation, and is commonly used as an experimental measurement of oxidative damage with mass spectrometric methods [267-269].

A number of studies have explored the adaptations of an organism's proteome to deal with an oxidative environment. These studies have examined how the amino acid usage of their proteomes differs between anaerobes versus aerobes [270], or between short- and long-living organisms [271-273]. The latter case has been of interest due to the proposed oxidative stress

theory of aging [274] which states that the accumulation of oxidative damage to macromolecules is a major reason why organisms age. As a result of these studies, there are a number of proposed hypotheses regarding how aerobic organisms have evolved to live in oxygen-rich environments or why certain species have increased longevity. In the context of the structural proteome, these proposals include: 1) the use of "amino acid sponges" that can absorb oxidative damage by enriching cytosolic protein surfaces with methionine [275-278] or cysteine [279], and additionally tyrosine or tryptophan in transmembrane proteins [280]; 2) the avoidance of cysteines [281] or carbonylation-susceptible residues [268, 282] on the surface of a protein; 3) the avoidance of charged or disorder-prone residues [283-286] to favor a more stable folded state; 4) the protection of reactive cofactors such as transition metal ions or flavins with extended domains [287] or by altering global or local structural characteristics [288-290]; and 5) a number of other novel mechanisms [291-295]. The true impact of these adaptations remains unclear as patterns observed could be attributed to other factors [270], but it is clear from these studies that antioxidative protein properties have manifested themselves within the genetic code over time.

In this study, we use a combination of both structural and systems biology approaches to evaluate if the differential manifestation of these antioxidative properties in the proteomes of multiple *E. coli* strains reflects the varying oxidative environments they may encounter. We extend a pipeline to construct genome-scale models with protein structures (GEM-PROs) to the entire reference proteome of *E. coli* K-12 MG1655, with additional methods to select representative cofactor-bound structures and protein complexes. We use this information along with a set of 1,764 sequenced *E. coli* strains to map DNA sequence variation to the reference proteome, with the goal of characterizing physicochemical changes in groups of proteins defined by their localization and functionality (Figure 1). We additionally create strain-specific genome-scale metabolic models

capable of predicting basal ROS production levels (henceforth referred to as "ROStypes") [29, 296] to understand how adaptation may arise not only due to levels of ROS encountered exogenously, but also produced endogenously. With this information, we are able to pinpoint shared changes in relation to a strain's phenotype and the antioxidative properties of its proteome. We unexpectedly find that strains with predicted higher levels of endogenous ROS relative to MG1655 ($ROS^{hi}$) share antioxidative properties in their proteomes to those with predicted lower levels ($ROS^{lo}$). We find that generally, metalloproteins and periplasmic proteins differ in these antioxidative properties, and detail two specific examples of molybdenum-binding enzymes and proteins involved in the biogenesis of fimbriae. This work demonstrates a structural systems biology approach to explore patterns of protein sequence variation in relation to predicted or known phenotypes, specifically in the context of adaptation to an oxidative environment.

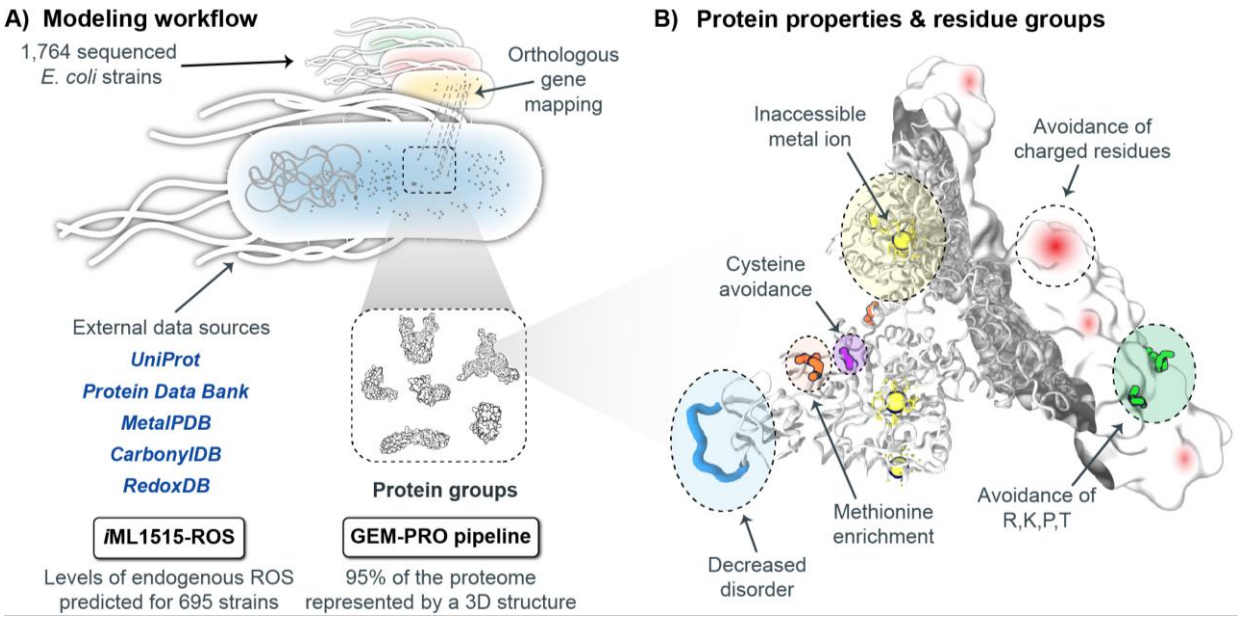



Figure 4.1. Schematic of modeling workflow and the hypothetical antioxidative properties of a protein.
A) The genomes of 1,764 strains of *E. coli* were gathered and orthologous genes were mapped to the reference *E. coli* K-12 MG1655 strain. External data sources were integrated to gather protein sequence and structure annotations with regards to susceptibility of oxidative damage, such as the locations of metal-binding sites [297], known carbonylation sites [298], and known cysteine damage sites [299]. The structural proteome is further categorized into protein groups by their annotated localization and functionality. We conducted gene deletions upon the genome-scale metabolic model of MG1655 integrated with ROS generating reactions (*i*ML1515-ROS [29, 296]) for strain-specific predictions of basal ROS production levels, defining a strain's "ROStype". We utilized the GEM-PRO pipeline [300, 301] to select representative protein structures for 95% of the MG1655 proteome. B) A hypothetical protein resistant to oxidative damage. Protein sequence and structure properties are highlighted based on previous studies finding enrichment of these properties in aerobes or long-living organisms. Structural properties defined by locations in 3D space, such as surface-exposed residues or those in a specified radius within a metal-binding site, are used to further divide a single protein into residue groups.

# 4.3 Results

### 4.3.1 Reconstructing the reference structural proteome of *E. coli* K-12 MG1655

The construction of a reference structural proteome for the 4,313 proteins of the *E. coli* K-12 MG1655 strain resulted in 1,457 proteins that could be represented by an experimentally determined 3D structure, an additional 2,661 proteins with a homology model, and 195 with no available structure. Proteins were segregated into functionally similar groups, first based on their localization within the cell and further using clusters of orthologous group (COG) categories and

metabolic subsystem groupings, resulting in over 200 groups of proteins (henceforth referred to as *protein groups*) analyzed for differences in their structural properties that could potentially contribute to oxidative stress resistance (henceforth referred to as *antioxidative properties*) (Figure 1). These antioxidative properties were selected on the basis of previous studies that characterized hypothesized resistance patterns seen in aerobes or in long-living organisms.

Improvements to the GEM-PRO pipeline [300, 301] allowed for the refined selection of experimental protein structures for 42 iron and iron-sulfur binding enzymes. For these, selection of a representative structure was extended beyond sequence identity and structural resolution by considering cofactor-bound states if experimental structures were available in both apo (cofactor-unbound) and holo (cofactor-bound) forms. Additionally, the amino acid residues that create interfaces between subunits in enzyme complexes were also integrated within the metabolic network reconstruction. The integration of external data sources enabled the assessment of changes to experimentally determined damage points on proteins for carbonylation and cysteine damage. These improvements result in a more rigorous reconstruction of the structural proteome and also enable future analyses to consider important protein subsequences (Figure 1B), such as for changes within a certain radius of a metal-binding site or within the residues that make up an enzyme complex interface.

### 4.3.2 Strains with significant variance in ROStypes display enrichment of antioxidative properties in their proteomes

The protein sequences of 1,764 strains of *E. coli* were gathered from public databases (see Methods). Simulations to predict the ROStype of available *E. coli* strains were successful for 695 strains, and resulted in a set of 16 strains that had a significantly higher basal ROS production rate ($ROS^{hi}$, >5% measured K-12 MG1655 production rate [302]) and 26 that had a lower basal rate

($ROS^{lo}$, <5% measured rate) (Figure 2A). This cutoff was selected from a previous study that validated a number of gene deletions and their impact on measured endogenous ROS levels [296]. The production rate of $H_2O_2$ and $O_2^-$ was highly correlated ($r$=0.98) and thus, classification of a strain based on their ROStype refers to rates of both $H_2O_2$ and $O_2^-$ production. A large number of strains (653) displayed non-varying levels (within 5%) compared to K-12 MG1655 ($ROS^{K\text{-}12}$) while the remaining strains (1,069) failed to simulate growth under minimal media conditions, most often due to missing amino acid synthesis pathways that were not mapped from orthologous genes. Gap-filling of the strain-specific metabolic models was not carried out.

From the set of $ROS^{hi}$ and $ROS^{lo}$ strains, we inspected the gene deletions that resulted in the predicted phenotypes. In both cases, the major class of reactions missing due to gene deletions were involved in alternate carbon metabolism and lipopolysaccharide biosynthesis, consistent with other studies of the core and pan genome [303, 304]. Both strain sets shared missing reactions in methionine metabolism, nitrogen metabolism, and inner membrane transport pathways. $ROS^{hi}$ strains were missing additional lipopolysaccharide (LPS) biosynthesis reactions, and the only non-LPS reaction unique to these strains was a deletion of UDP-galactopyranose mutase. $ROS^{lo}$ strains on the other hand, had a variety of missing reactions in different subsystems (Figure 2B). Interestingly, in most protein groups, the computed antioxidative properties did not significantly cluster apart the $ROS^{hi}$ and $ROS^{lo}$ strains in principal component analyses (PCA) (top panels of Figure 3A, 4A, 5A). It was observed that these strains clustered together, but apart from strains with non-varying levels of endogenous ROS.

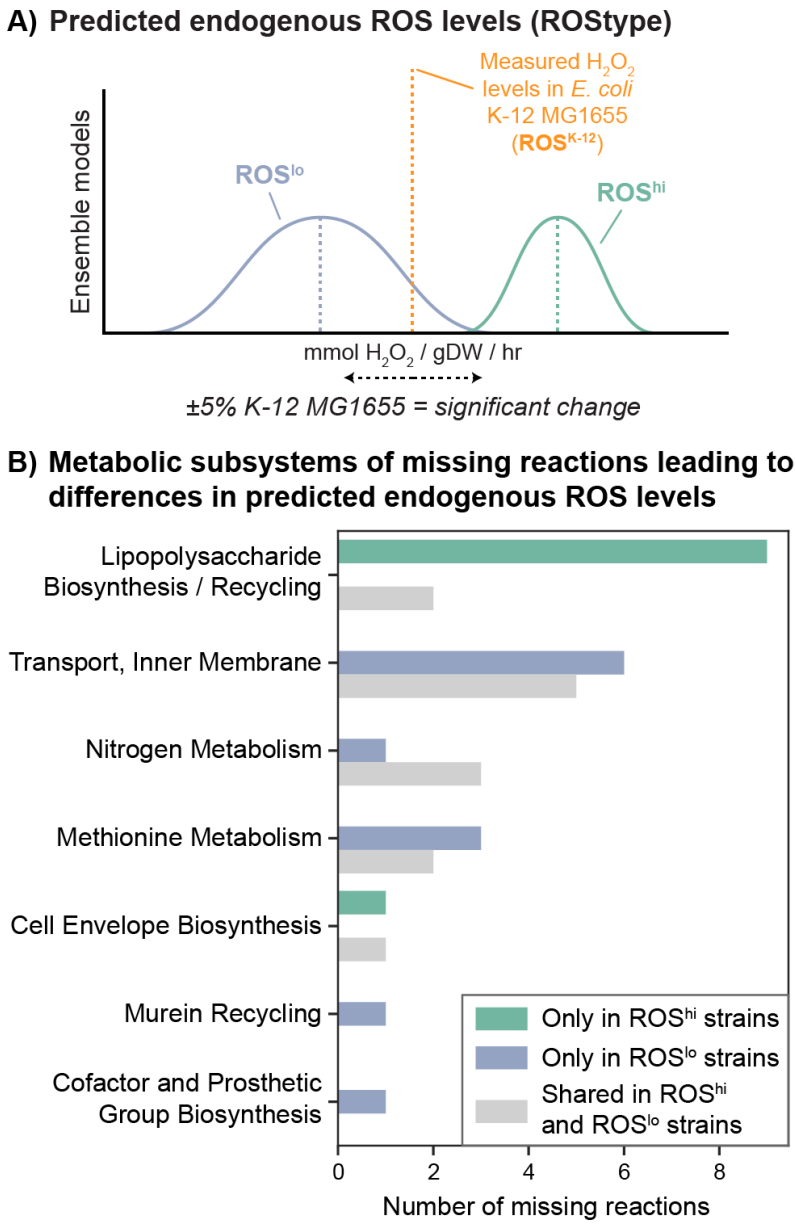



Figure 4.2. Classifying strains by predicted endogenous ROS levels and missing reactions that contribute to the predicted phenotype.
A) Simulations of strain-specific metabolic models enable the prediction of endogenous ROS levels, or ROStype. A defined ROStype results from changes to pathway usage due to the deletion of certain reactions from missing genes. Strains are classified by their predicted endogenous ROS levels as "high" (ROS$^{hi}$) or "low" (ROS$^{lo}$) ROStypes if their predicted rates of production of hydrogen peroxide ($H_2O_2$) or superoxide ($O_2^-$) differ by more than 5% from the measured production rate in the K-12 MG1655 strain (orange dotted line). If predicted endogenous ROS levels do not differ by more than 5%, a strain is classified as similar to MG1655 (ROS$^{K-12}$). B) Histogram of the metabolic subsystems of missing reactions that contribute to the ROStype. Reactions that are shared between strains with ROS$^{hi}$ and ROS$^{lo}$ predictions are denoted as shared missing reactions in gray.

### 4.3.3 Known chemical mechanisms of oxidative damage to proteins enable the assessment of a strain's oxidative environment

Pathotype classifications were available for a subset of 116 from the overall set of 1,764 strains, while growth/no growth phenotypes in a variety of media conditions were available for up to 650 strains depending on the condition [305]. PCA of the selected antioxidative properties successfully segregated strains with predicted ROStypes (i.e., $ROS^{hi}$ and $ROS^{lo}$ strains vs. $ROS^{K-12}$) and those with defined pathotypes (i.e., ExPEC/AIEC/APEC vs. other pathotypes) for a number of protein groups (Table 1). These groups were ranked by cluster homogeneity following Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Examples of proteins showing significant cluster homogeneity include the general protein groups of all metal-binding enzymes as well as all proteins localized to the periplasm (Figure 3). Specifically, in these groups, the metal-binding enzymes that utilize a molybdenum cofactor and periplasmic enzymes involved in the assembly of fimbriae showed the highest homogeneity. These specific groups are expanded upon below. Other protein groups with high homogeneity included other extracellular and motility proteins, murein biosynthesis enzymes, and proteins involved in inorganic ion transport and metabolism (Table 1).

We observed very little segregation of the panel of strains with available growth/no growth phenotypes in various media conditions [305], with our original hypothesis being that strains with growth phenotypes in oxidative environments would show enrichment of antioxidative properties in proteins important for metabolic function. A major reason for this observation was due to a much lower statistical power as many conditions contained a very small number of strains with "no growth" phenotypes compared to those with a "growth" phenotype.

Table 4.1. Homogenous clusters formed using DBSCAN on the first two principal components, following labeling with ROStype or pathotype.
Both V-measures and Adjusted Rand Index (ARI) are reported for generated clusters to describe cluster homogeneity. For ROStypes, both high and low predictions were considered the same in homogeneity measurements. If a ROStype or pathotype is unavailable for a strain, they were excluded from the homogeneity measurements.

| ***ROStype: $ROS^{hi}$ & $ROS^{lo}$ vs. $ROS^{K-12}$*** | | | | |
|---|---|---|---|---|
| **Protein group** | **Localization** | **# clusters** | **V-measure** | **ARI** |
| COG I (Lipid transport and metabolism) | Cytosol | 2 | 0.3 | 0.44 |
| COG P (Inorganic ion transport and metabolism) | Cytosol | 5 | 0.27 | 0.43 |
| Molybdenum-binding enzymes | Periplasm | 4 | 0.27 | 0.28 |
| Metabolism - Murein Biosynthesis | Periplasm | 2 | 0.25 | 0.32 |
| Fimbriae assembly proteins | Periplasm | 4 | 0.23 | 0.31 |
| All metal-binding proteins | All | 7 | 0.19 | 0.32 |
| All periplasmic proteins | Periplasm | 3 | 0.09 | -0.02 |
| | | | | |
| ***Pathotype: ExPEC/AIEC/APEC vs. all others*** | | | | |
| **Protein group** | **Localization** | **# clusters** | **V-measure** | **ARI** |
| Fimbriae assembly proteins | Periplasm | 3 | 0.45 | 0.55 |
| COG Q (Secondary metabolites biosynthesis, transport and catabolism) | Cytosol | 4 | 0.44 | 0.57 |
| COG N (Cell motility) | Cytosol | 3 | 0.41 | 0.46 |
| COG W (Extracellular structures) | Periplasm | 3 | 0.41 | 0.5 |
| Molybdenum-binding enzymes | Cytosol | 3 | 0.4 | 0.44 |
| All metal-binding proteins | All | 3 | 0.08 | 0.02 |
| All periplasmic proteins | Periplasm | 2 | 0.16 | 0.02 |

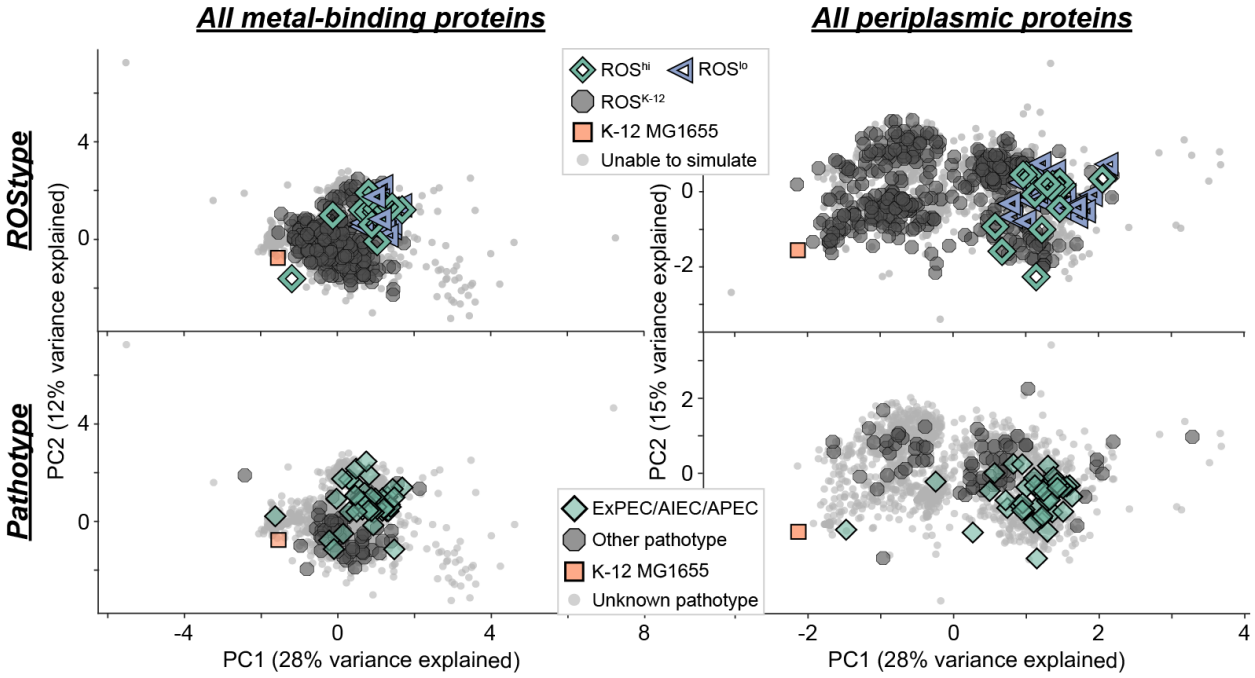



Figure 4.3. Principal components analysis (PCA) of antioxidative properties in all metal-binding proteins as well as all periplasmic proteins.
Antioxidative properties are computed for all metal-binding proteins and averaged for every strain. A feature matrix containing these properties is used as input to PCA, and subsequently, strains with predicted ROStypes (top left) and strains with pathotype annotations (bottom left) are highlighted. PCA for all proteins localized to the periplasm is also shown (right). These two general groups of proteins show clusters with the highest homogeneity in regards to predicted endogenous ROS or pathotype labels.

### 4.3.4 Molybdoenzymes with promiscuous activity are enriched with ROS-resistant structural properties

Enzymes that utilize molybdenum as an inorganic cofactor in catalysis carry out a variety of redox reactions in *E. coli* [306] and in all eukaryotic organisms [307, 308]. The molybdoenzyme group created through the structural proteome reconstruction was composed of 14 enzymes containing molybdenum-binding sites in their associated UniProt entries. PCA of their antioxidative properties displayed clear clustering of ROStypes and pathotypes (Figure 4A).

However, analysis of the antioxidative properties contributing to the first principal component revealed conflicting properties among the proteins contained within the group (see XdhD, below). Thus, we proceeded to inspect the properties of each of the 14 molybdoenzymes individually.

The subset of these enzymes that show enrichment of antioxidative properties in the pathotype and endogenous ROS clusters display dual functionalities in many cases. For example, the main function of biotin sulfoxide reductase (BisC) is in biotin salvage (i.e., to reduce the oxidized form of biotin), but it has also been shown to reduce oxidized free methionine [309]. BisC most significantly shows an avoidance of surface-exposed cysteines and negatively-charged residues ($p<0.0001$, Mann-Whitney U test) (Figure 4B). An avoidance of surface-exposed methionines was also observed. Trimethylamine-N-oxide (TMAO) reductase 2 (TorZ) reduces the compound TMAO, an alternative electron acceptor for anaerobic growth [310, 311]. Additionally, TorZ carries out the same biotin sulfoxide reductase reaction as BisC [312], although at a lower catalytic rate. TorZ displayed properties in the same strain sets trending towards more ordered residues, and less surface-exposed carbonylatable and charged residues. Xanthine dehydrogenase subunit A (XdhA), which showed similar property enrichments as TorZ (Figure 4C), is involved in purine catabolism and reduces NAD+ to NADH in the process [313]. Increases in xanthine levels are potentially indicative of higher rates of DNA damage, such as from ROS [314]. Out of the set of molybdoenzymes which avoided antioxidative properties, most are either known to only have a single function, or do not have their functions well characterized as of yet. As an example, a hypothesized xanthine oxidase (XdhD), displayed changes that shifted it to be more susceptible to damage. These changes may be due to the fact that XdhD cannot carry out the dehydrogenase reaction that XdhA does, since it lacks the FAD-binding domain to catalyze it [313]. As such,

being an oxidase, XdhD is involved in the generation of ROS and may not be desirable for use in high ROS environments.

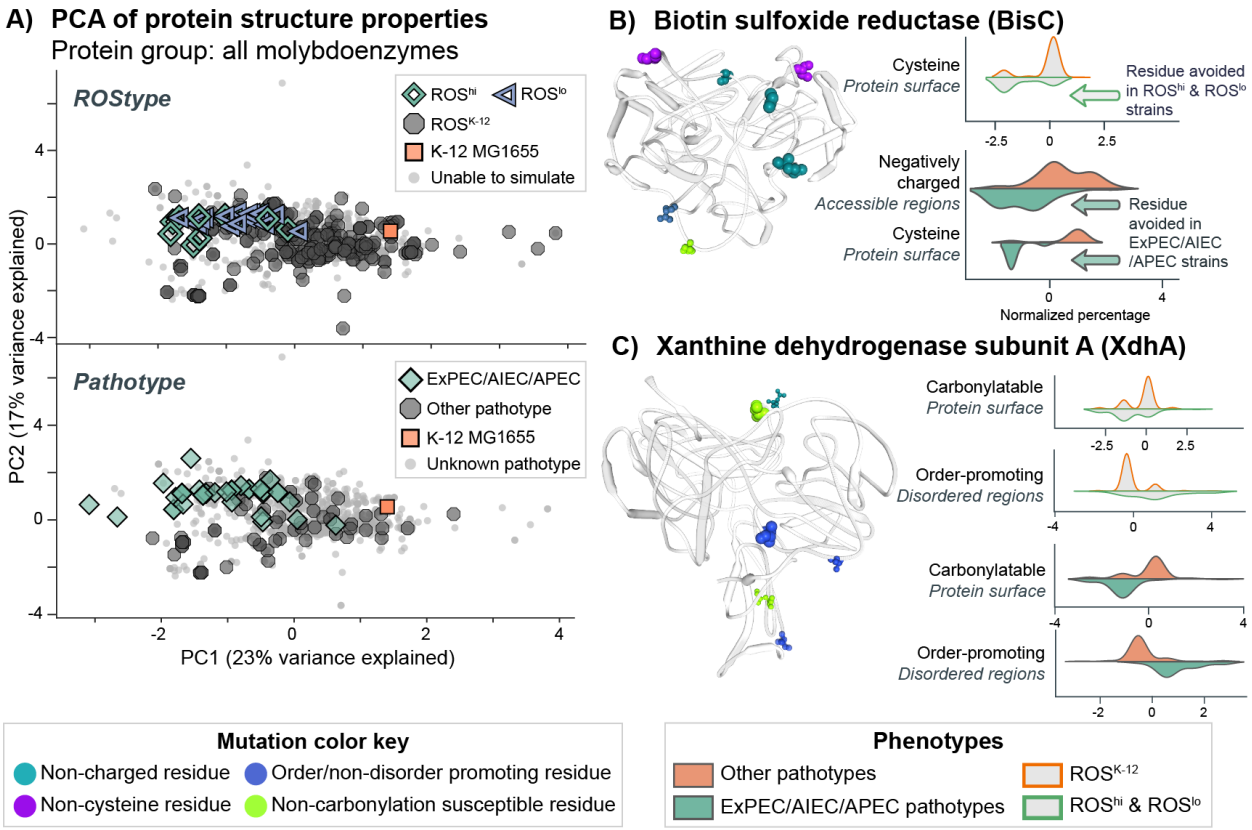



Figure 4.4. Molybdenum-binding proteins are enriched in antioxidative properties in strains with both high and low predicted levels of endogenous ROS as well as strain pathotypes likely encountering oxidative environments.
A) PCA of antioxidative properties for molybdenum-binding proteins of strains, in relation to their predicted ROStype and annotated pathotype. Proteins were inspected individually for changes in antioxidative properties, since analysis of the component contributions showed both enrichment and avoidance of certain properties. ExPEC/AIEC/APEC strains are grouped together as they likely encounter oxidative environments more frequently [315]. B) Biotin sulfoxide reductase shows avoidance of surface-exposed cysteine residues in both $ROS^{hi}$/$ROS^{lo}$ and ExPEC/AIEC/APEC strains. The residues highlighted on the protein structure indicate common mutations in strains of ExPEC/AIEC/APEC pathotypes. The size of the highlighted residue corresponds to the number of strains that mutation appears in. Note that all mutations do not co-occur together in all strains. The distribution plots for strain phenotypes to the right of the protein figure show the normalized percentage of the residue in relation to the protein subsequence, i.e. percentage of cysteines on the protein surface. C) Xanthine dehydrogenase subunit A similarly shows enrichment of antioxidative properties, by avoiding carbonylatable and charged residues on the protein surface, along with an increase of a total percentage of order-promoting residues. Structural models shown here are all homology models from the SWISS-MODEL database [316].

### 4.3.5 Operons containing type 1 fimbriae biogenesis proteins are enriched in antioxidative properties

Fimbriae are special pili that are synthesized and transferred to the outer membrane of *E. coli*. They are involved in the attachment of the bacteria to their host environments [317]. The *fim* operon is the most well characterized system that assembles type 1 fimbriae [318]. A number of other similar operons are encoded within the K-12 MG1655 genome, but require specific environmental stimuli to be expressed [319]. PCA and subsequent DBSCAN clustering identified this set of proteins as creating highly homogenous clusters, again for both strains with predicted ROStypes and defined pathotypes (Figure 5A).

One cluster contained a majority of the annotated ExPEC/AIEC/APEC strains (38 of 40), along with two asymptomatic (ABU) 83972 strains and 6 strains with a variety of other pathotypes. Another cluster was comprised by a majority of EHEC strains (37 of 69). The antioxidative properties contributing to the first principal component were largely consistent with many of the previously outlined properties hypothesized to contribute to oxidative stress resistance. Specifically, the rightmost strains on PC1 are enriched in order-promoting residues in disordered regions of their proteins, along with solvent accessible methionines. The leftmost strains on PC1 are characterized by amino acid features opposite to providing oxidative stress resistance, such as an increase in disorder-promoting residues in disordered regions, more solvent accessible cysteines, charged residues, and residues susceptible to irreversible carbonylation (Figure 5B). The proteins that diverged in sequence the most from the orthologous K12 sequence were those in the *yeh* and *yfc* operons (Figure 5C). The function of these operons is relatively unknown but hypothesized to assemble other type 1 fimbriae due to their homology to *fim* components [319].

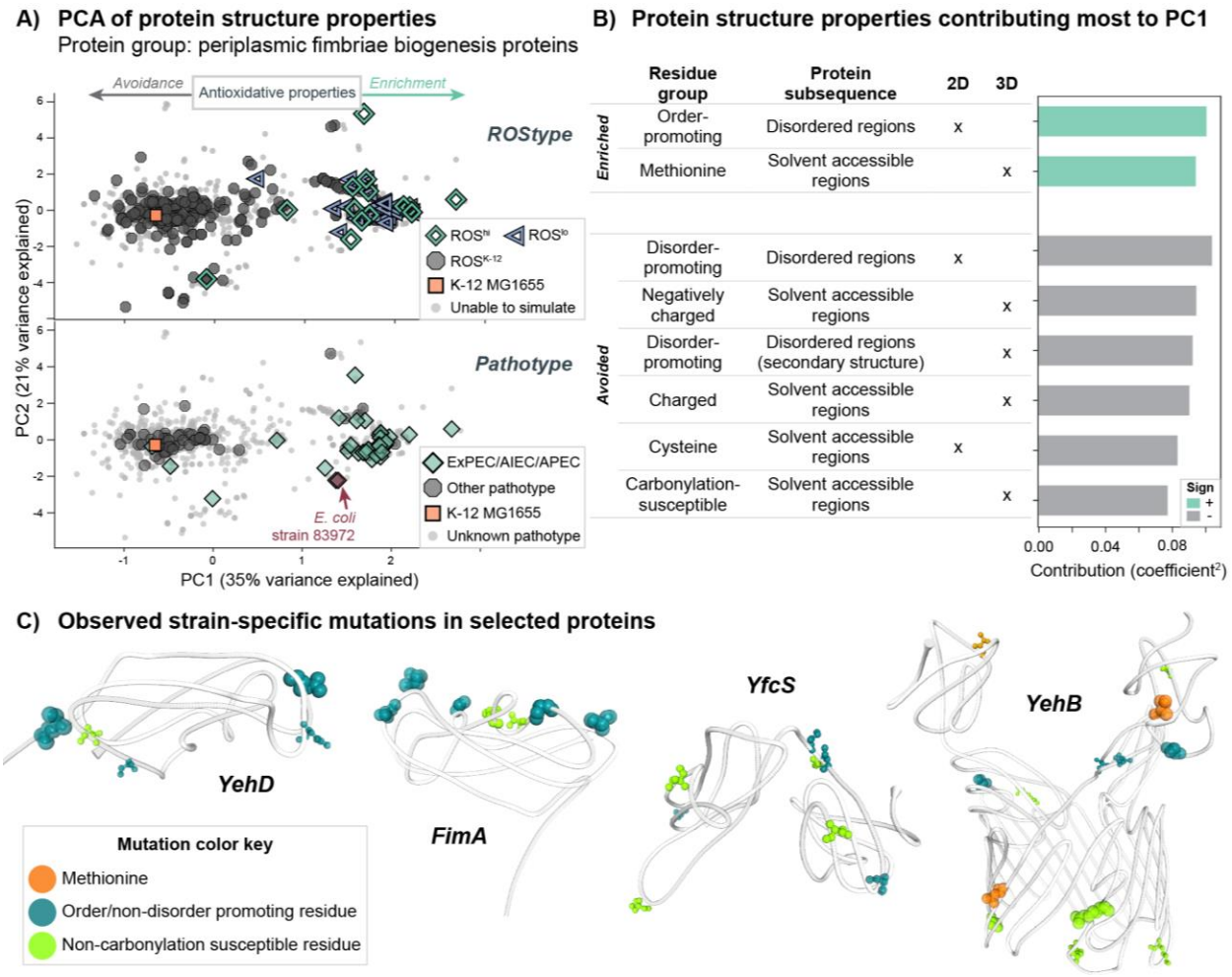



Figure 4.5. Proteins in the periplasm involved in the assembly of fimbriae are enriched in antioxidative properties. A) PCA of antioxidative properties for periplasmic fimbriae assembly proteins, in relation to their simulated ROStype and annotated pathotype. PC1 separates the observed ExPEC/AIEC/APEC strains and DBSCAN was applied to cluster the results into 3 (for ROStype) or 4 (for pathotype) clusters (Table 1). The location of two asymptomatic strains (83972) is specified as they have been found to outcompete UPEC strains in adhesion of the bladder wall [320]. B) Specific antioxidative properties contributing to PC1. The 2D and 3D columns indicate if the protein subsequences were determined by predictions from sequence (2D) or calculations from structure (3D). C) Selected examples from the proteins involved in fimbriae assembly that show enrichment of antioxidative properties. YehB, YehD, and YfcS are relatively unknown components of operons similar to the *fim* operon [319]. Up to 3000 copies of FimA form the structural pilus of the fimbriae [321]. Highlighted residues indicate mutations which are seen in the cluster with the most ExPEC/AIEC/APEC pathotypes. The size of the highlighted residue corresponds to the number of strains that mutation appears in. Note that all mutations do not co-occur together in all strains. The structures shown here are from a collection of homology models from the SWISS-MODEL and I-TASSER modeling pipelines [14, 316].

## 4.4 Discussion

### 4.4.1 Endogenous ROS levels suggest convergent evolution in oxidative stress resistance

The principal component analyses of the antioxidative properties of protein groups revealed that similar adaptive features are displayed by strains with both $ROS^{hi}$ and $ROS^{lo}$ ROStypes. This result was surprising as we had initially expected to see features in opposition to each other, such that $ROS^{hi}$ strains would have adapted their proteomes to deal with this constant source of oxidative stress while $ROS^{lo}$ strains would perhaps vary in other ways unique to their environment. However, it has been found that other organisms adapt to environments of high oxidative stress by lowering their endogenous ROS levels [322], but conversely, high endogenous ROS can allow for natural adaptive mutations to occur lending to a general increased tolerance to oxidative environments [323]. Thus, the relationship between genetic variation, endogenous ROS, and exogenous ROS is complex, but trends towards similar genetic adaptations as seen in our simulation results. Unfortunately, metadata for the gathered *E. coli* strain sequences was sparse with the only consistent annotation being the isolation site, which does not show any correlation to oxidative environments. The analysis of strains and their associated genome sequences could benefit greatly from richer and standardized annotations of observed phenotypes for large-scale studies.

### 4.4.2 Dual functionalities of molybdoenzymes potentially contribute to the oxidative damage repair response

The molybdoenzymes within *E. coli* generally catalyze unique redox reactions under aerobic conditions, such as biotin salvage, and also enable the usage of alternative electron acceptors in anaerobic conditions, such as nitrate [306]. Due to their redox capabilities, some

molybdoenzymes can also act promiscuously to reverse oxidation events to other metabolites such as methionine [309]. Interestingly, the standard repair system for methionine sulfoxide - specifically in the stressful oxidative environment of the periplasm (MsrPQ) - utilizes molybdenum as its cofactor of choice and is able to reduce both stereoisomers of oxidized methionine [266]. The proposed stability of these molybdoenzymes under oxidative conditions suggests two factors: 1) that these enzymes would be upregulated in response to oxidative stress to reduce their main binding partners that are being oxidized by ROS, and 2) that they may be called upon to carry out their promiscuous repair functions on other oxidized metabolites.

Although the function of molybdoenzymes in anaerobic respiration seems unrelated to oxidative stress, there may be reasons for their use in aerobic conditions. Interestingly, a previous study indicated the metabolite trimethylamine-N-oxide (TMAO) to confer protein structural stability in vitro, by stabilizing charged residues, disordered regions, and preventing protein aggregation [324]. The identified TorZ enzyme in this analysis may then have a role in maintaining reduced TMAO in conditions of oxidative stress. The usage of nitrate as an electron acceptor in anaerobic respiration generates reactive nitrogen species, which, similarly to ROS, damage cysteine and additionally tyrosine residues [325]. Manual analysis of TorZ and BisC structures displayed avoidance of surface-exposed tyrosines, suggesting that similar structural analysis could be carried out for other sites of protein damage.

### 4.4.3 Type 1 fimbriae biogenesis operons and their use in oxidative environments

The type 1 fimbriae biogenesis components are interesting to approach from the standpoint of oxidative damage resistance due to three reasons. First, the assembly of the fimbriae depends on an oxidation event to create a single disulfide bond on the components that make up the tip of the fimbriae, which ends up adhering to the environmental surface [326]. Second, there are many

similar operons with homologous genes that encode for other fimbriae, supposedly for different environmental conditions [319]. Third, the expression of these components is sensitive to oxygen levels, being inactive under anaerobic conditions [327]. We grouped together ExPEC, AIEC, and APEC strains since their encountered environments are associated with high oxidative stress and because of their similarity in both phylogenetic origin and virulence factors [328-330].

The assembly of the fimbriae begins when a subunit is translocated into the periplasm and bound to a chaperone, which accompanies the subunit to the outer membrane "usher" (collectively known as the chaperone-usher pathway). This binding event depends on the subunit being oxidized by DsbA, a periplasmic enzyme that creates and maintains disulfide bonds [326]. The formation of disulfide bonds would be accelerated by an environment with higher levels of ROS, but would additionally demand the chaperone to have antioxidative properties if the fimbriae were to properly assemble. Additionally, previous studies have shown that most sequence variation on the extracellular fimbriae subunits do not decrease the specificity of adhesion to carbohydrates [331]. However, specific point mutations on the FimH adhesin do provide higher adhesion capabilities [332]. The results presented here suggest that variations are likely implicated in greater stability during translocation and when being presented on the exposed regions of the fimbriae. The existence of the other type 1 fimbriae operons points to a highly adaptable set of adhesins available to an *E. coli* cell. Finally, the finding that the expression of these operons responds to changes in oxygen levels [327] points to a likely link between their antioxidative properties and survival within an oxidative environment.

The two *E. coli* ABU 83972 strains in the cluster of ExPEC/AIEC/APEC strains have been shown to outcompete UPEC strains in colonization of the bladder [333]. The similar properties of their type I fimbriae biogenesis proteins may point towards a similar resistance to oxidative

damage in the urinary tract, which is a stressful environment during urinary tract infection [334, 335]. We hypothesize that the greater stability of the biogenesis enzymes under these conditions potentially allow these nonpathogenic strains to assemble their fimbriae and colonize the bladder, outcompeting the UPEC strains. This finding suggests that further experimental work to characterize these relatively unknown fimbriae operon components may elucidate adherence properties of *E. coli* strains.

## 4.5 Conclusion

In this study, we have applied two protocols in a structural systems biology approach to develop an understanding of the genotype-phenotype relationship. At the systems-level, we utilized strain-specific genome-scale metabolic models to predict levels of endogenous ROS, while also guiding orthologous gene mapping of strains for sequence-level variation analysis. With structural information, we mapped this variation to specific groups of proteins and their location within three-dimensional space. Specific protein groups were identified as differentiating in regards to their antioxidative properties. The analysis of 1,764 sequenced strains confers strong evidence pointing towards further experimental study of the contributions of molybdenum-binding enzymes as well as fimbriae assembly proteins in the oxidative stress response. The workflow presented here also demonstrates a method for understanding important features of the structural proteome to enable modeling of different stress responses *in silico*. Looking forward, the exploration of natural variation in regards to enzymatic capabilities to confer a specific environmental advantage could be applied in the creation of fine-grained metabolic models taking into account the properties of the structural proteome. In the lab, this approach could also guide

drug development pipelines by identifying susceptible targets as well as library design for directed evolution experiments in protein engineering.

## 4.6 Methods

### 4.6.1 Construction of the E. coli structural proteome

We applied an existing pipeline to create genome-scale models of metabolism with protein structures (GEM-PRO models) [32, 300] for the entire proteome of *E. coli* str. K-12 substr. MG1655 (reference proteome downloaded from UniProt [69] on March 20, 2018), using the current implementation contained in the ssbio Python package [301]. This pipeline results in the selection of a single representative three-dimensional tertiary protein structure, either from experimental data in the Protein Data Bank [70] or from homology models generated from the I-TASSER pipeline [14] and the SWISS-MODEL repository [316]. A number of improvements were implemented for this study and are slated for incorporation into the publicly available pipeline in ssbio. These improvements include 1) the selection of representative experimental structures with the consideration of bound cofactors, as described in the methods of [336]; 2) the definition of membrane spanning domains in transmembrane proteins by consolidating information as predicted by TMHMM [17] or OPM [337] and as annotated in UniProt; 3) the selection of quaternary structures of protein complexes by a breadth-first search matching algorithm to determine the structure with the highest quality and coverage of annotated subunits, either from biological assemblies in the PDB or from predicted complexes in the SWISS-MODEL repository.

### 4.6.2 Division of the proteome into groups

To delineate search spaces within the structural proteome we created groups of proteins first defined by their localization, and then defined by their functional assignment (Figure 1A). Localization within a cell is defined by the categories: outer membrane, periplasm, inner membrane, and cytosol. This information was taken from a consensus of a previously generated genome-scale model of proteome synthesis [338], proteomics information from [339], the EchoLOCATION database [340], the UniProt database, and finally predictions from TMHMM [17] if no other information was available. Secondary functional groups were either created with 1) the clusters of orthologous group (COG) categories [341]; 2) metabolic network subsystem as defined in *i*ML1515; 3) metal-binding enzymes as annotated in UniProt; and 4) manual assignment in relation to ROS generation or repair. The manually curated list of proteins were collected based on their functional relation to oxidative stress levels in *E. coli*. These include proteins that are known generators of reactive oxygen species, are involved in the repair of oxidative damage to macromolecules, are involved in regulation of metal ion transport, etc.

### 4.6.3 Protein property calculations and division into subsequences

The physicochemical properties of each representative protein's sequence and structure were calculated using a variety of tools and stored as per-residue annotations using ssbio. The properties used in this study include: 1) definitions of protein "sites" (i.e. binding, catalytic, or active sites) as annotated in UniProt, the Catalytic Site Atlas [342], and MetalPDB [297]; 2) regions of protein flexibility and disorder as retrieved from PDBFlex [343] or predicted with DisEMBL [344] and IUPred [345]; 3) parameters of solvent accessibility as calculated using FreeSASA [24] or predicted using SCRATCH [13]; 4) parameters of residue depth as calculated using Biopython [3] and MSMS [12]; and 5) definitions of secondary structure using DSSP [23] or predicted using SCRATCH. We additionally incorporated the locations of known oxidizable

cysteines from RedoxDB [299], known carbonylatable amino acids (R, K, P, T) from CarbonylDB [298], and predicted disulfide bridges using a distance-based metric (3 Å cutoff) in an extension of the PDB module in Biopython (http://biopython.org/wiki/Struct).

The aforementioned properties were then used to delineate search spaces within single proteins into subsequences (Figure 1B). For example, if a 3D structure was available for a protein, a "surface" residue was defined as one with a depth of less than 2.5 Å and a relative solvent accessibility of above 25%. These cutoffs were then applied to all residues in a protein sequence and those that meet the cutoff then form a defined surface subsequence. The main properties that formed these definitions are: depth, solvent accessibility, disordered regions, transmembrane domains, solvent-exposed residues surrounding a metal-binding site or a catalytic site, residues within a DNA-binding site or any annotated "site", and residues within sensitive sites as found in RedoxDB or CarbonylDB.

### 4.6.4 Gathering strains, phenotypes, and pathotypes

The proteomes of *E. coli* strains in this study were gathered from three separate sources: 1) the Ecoref strain panel [305]; 2) the *i*ML1515 metabolic network reconstruction resource [29]; and 3) a manually curated set of adherent-invasive *E. coli* (AIEC) strains. The Ecoref strain panel is a published panel of 696 strains that were tested for growth/no growth under a number of media conditions. The sequenced genomes, resulting protein sequences, and pathotype annotations were downloaded from the publicly accessible database (https://evocellnet.github.io/ecoref/download/). Out of the 214 media conditions, 24 of which include a chemical that induces oxidative stress in some manner were selected for this study. Out of the 696 strains, 676 were selected due to the presence of phenotypic data under the selected media conditions. From the *i*ML1515 resource, the proteomes and pathotype information of 1045 strains of *E. coli* were downloaded from the

PATRIC database [346]. The number differs from the original reported number of strains in the resource due to database updates and removal of poorly annotated genomes. The strains obtained from Ecoref and PATRIC also contained pathotype information for 116 of the strains. Finally, the manually curated set of 23 AIEC strains was obtained through literature review and downloading of the individual genomes from various publications [347-353]. Two pathotype groups was created by 1) extra-intestinal pathogenic (ExPEC), adherent-invasive (AIEC), and avian pathogenic (APEC) strains (which have been shown to be similar to ExPEC strains, see [354, 355]) and 2) all other pathotypes. This grouping was chosen to discriminate pathotypes by the oxidative environments they may encounter.

### 4.6.5 Simulation of strain-specific endogenous ROS levels

The construction of strain-specific metabolic models follows a previously established protocol [356]. Briefly, this involves creating a presence/absence orthology matrix of proteins in a strain following orthology detection through bidirectional-best BLAST hits (BBH) at an 80% sequence identity cutoff. The orthology matrix is then used to trim reactions in a metabolic model given a protein's usage in a reaction. Instead of trimming the original metabolic model of *E. coli* K-12 MG1655, the *i*ML1515-ROS model was used as the base model. Based on a previously developed model [296], *i*ML1515-ROS includes 298 reactions that have the potential to produce $H_2O_2$ and $O_2^-$ [29]. Thus, the strain-specific ROS models predicts changes in endogenous ROS production levels based on the deviation from the measured MG1655 ROS production rates.

Ensemble models (sampling different stoichiometric coefficients of the ROS generating reactions) of each strain were then simulated to sample total endogenous production of ROS. Simulations were conducted under glucose minimal media per [296]. The mean $H_2O_2$ and $O_2^-$ production rates were normalized to the growth rate, thus assigning per-strain predictions of mmol

$H_2O_2$ $gDW^{-1}$ and $O_2^-$ $gDW^{-1}$. The deviations of endogenous ROS production from the "wild-type" MG1655 strain were classified as "high" or "low" if they deviated above or below 5% of the measured ROS production levels, and "non-varying" if otherwise [302].

### 4.6.6 Characterization of strain-specific changes

The orthology matrix used to generate strain-specific models was used to align orthologous strain protein sequences to the K12 proteins, using default parameters in the Needleman-Wunsch pairwise alignment tool in the EMBOSS package [357]. All orthologous sequences were loaded into their related Protein objects using the ssbio Python package, and alignments were executed in parallel using the Apache Spark Python API (PySpark, https://spark.apache.org). From this, strain-by-feature matrices describing the proteomic features of all strains and the averaged antioxidative properties were generated. To deal with missing data, such as proteins that are absent from certain strains or portions of protein sequences that have been truncated (either due to technical reasons such as genome sequence or annotation errors, or true deletions compared to MG1655), we tested multiple filtering and imputation methods and compared consistency between them.

The strain-by-feature matrices are created for principal component analysis (PCA) as follows. For all combinations of protein groups and subsequences (i.e. a protein group of cytoplasmic metal-binding enzymes and subsequences of their metal-binding sites), amino acid ratios were calculated relative to the subsequence length. Ratios of certain groupings of amino acids were also calculated, such as from the number of positively charged, bulky, or disorder promoting residues. These groupings were again chosen due to previous hypotheses that enrichment or avoidance of these changes may be associated with resistance to oxidative damage [265, 267, 284, 289, 358].

### 4.6.7 Statistical analysis of protein groups

For each protein group, subsequence, and associated strains, features were first l2-normalized to unit norm using the Python package scikit-learn [359]. Features were then filtered for PCA following the hierarchical clustering by rank correlation coefficient method as presented in [284, 360]. Briefly, this clusters the features together based on Spearman rank correlation, and cuts off similar features over a coefficient of 0.9, keeping the feature closest to the center of the cluster as representative. Since the availability of features varied from prediction from sequence and calculation from structure, this allowed for the filtering out of redundant features when both predictions and calculations were available. PCA was then carried out to understand if specific features contributed to the differentiation of the following: 1) growth/no growth phenotypes in different media conditions; 2) strains with predicted higher or lower than "wild-type" (K12) endogenous ROS levels; and 3) annotated strain pathotypes. To do so, observation labels associated with these phenotypes were assigned back to the transformed data. Next, we wanted to identify protein groups which showed the largest separation between phenotypes. DBSCAN was utilized to cluster the strain phenotypes into dense clusters, and cluster homogeneity was ranked by a variety of measures such as the Rand index.

## 4.7 Acknowledgements

Chapter 4, in full, is a reprint of the material as it appears in: Mih N, Monk JM, Fang X, Catoiu E, Heckmann D, Yang L, Palsson BO. Adaptations of *Escherichia coli* strains to oxidative stress are reflected in properties of their structural proteomes. Submitted. The dissertation author was the primary author of this paper.

# Chapter 5.

# Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms

## 5.1 Abstract

Catalysis using iron-sulfur clusters and transition metals can be traced back to the last universal common ancestor. The damage to metalloproteins caused by reactive oxygen species (ROS) can completely inhibit cell growth when unmanaged and thus elicits an essential stress response that is universal and fundamental in biology. We develop a computable multi-scale description of the ROS stress response in *Escherichia coli*. We show that this quantitative framework allows for the understanding and prediction of ROS stress responses at three levels: 1) pathways: amino acid auxotrophies, 2) networks: the systemic response to ROS stress, and 3) genetic basis: adaptation to ROS stress during laboratory evolution. These results show that we can now develop fundamental and quantitative genotype-phenotype relationships for stress responses on a genome-wide basis.

## 5.2 Main text

All aerobic life requires management of corrosive oxidative stress. Oxygen toxicity is manifested in damage to cellular components by reactive oxygen species (ROS). Cells generate

ROS endogenously when flavin, quinol, or iron cofactors are autoxidized [332]. ROS species are known to damage DNA, certain iron-containing metalloproteins, and other cellular processes [311]. In addition to inherent endogenous ROS production, eukaryotes and microbes produce exogenous ROS as $H_2O_2$, superoxide, or redox-cycling compounds to inflict oxidative stress on competitors [328]. In spite of the fundamental importance of ROS damage on cellular functions, we lack a framework that connects known and hypothesized individual molecular targets of ROS to systemic physiological responses. Here, we address this gap using a genome-scale computational systems biology approach focused on the processes that determine homeostasis of iron, which is essential for *E. coli*'s growth, yet is vulnerable to ROS.

We developed a computable multi-scale description of ROS damage to metalloproteins in the context of metabolic function and macromolecular expression in *Escherichia coli* MG1655 [287, 288, 321, 325]. To mechanistically account for metalloprotein inactivation by ROS, we integrated metal cofactor damage and repair pathways for 43 protein complexes involving mononuclear iron or iron-sulfur clusters (Fig. 1). While all of these enzymes are potentially inactivated, the extent of ROS damage has not been experimentally determined for most of them. Certain iron metalloproteins have been shown to avoid inactivation by ROS due to structural properties including full incorporation or solvent accessibility [282]. To account for these cases, we integrated protein structural property calculations [295, 361] into a Bayesian network to compute the probability of cofactor inactivation by ROS (Fig. S1). The resulting description is a computable model, referred to as OxidizeME, that accounts for 1,581 proteins that collectively represent up to 85% of the proteome by mass. OxidizeME makes clear interpretations and predictions of ROS responses, three of which are addressed below; 1) ROS and amino acid

auxotrophies, 2) the systemic response to ROS stress, and 3) adaptation to ROS stress during laboratory evolution.

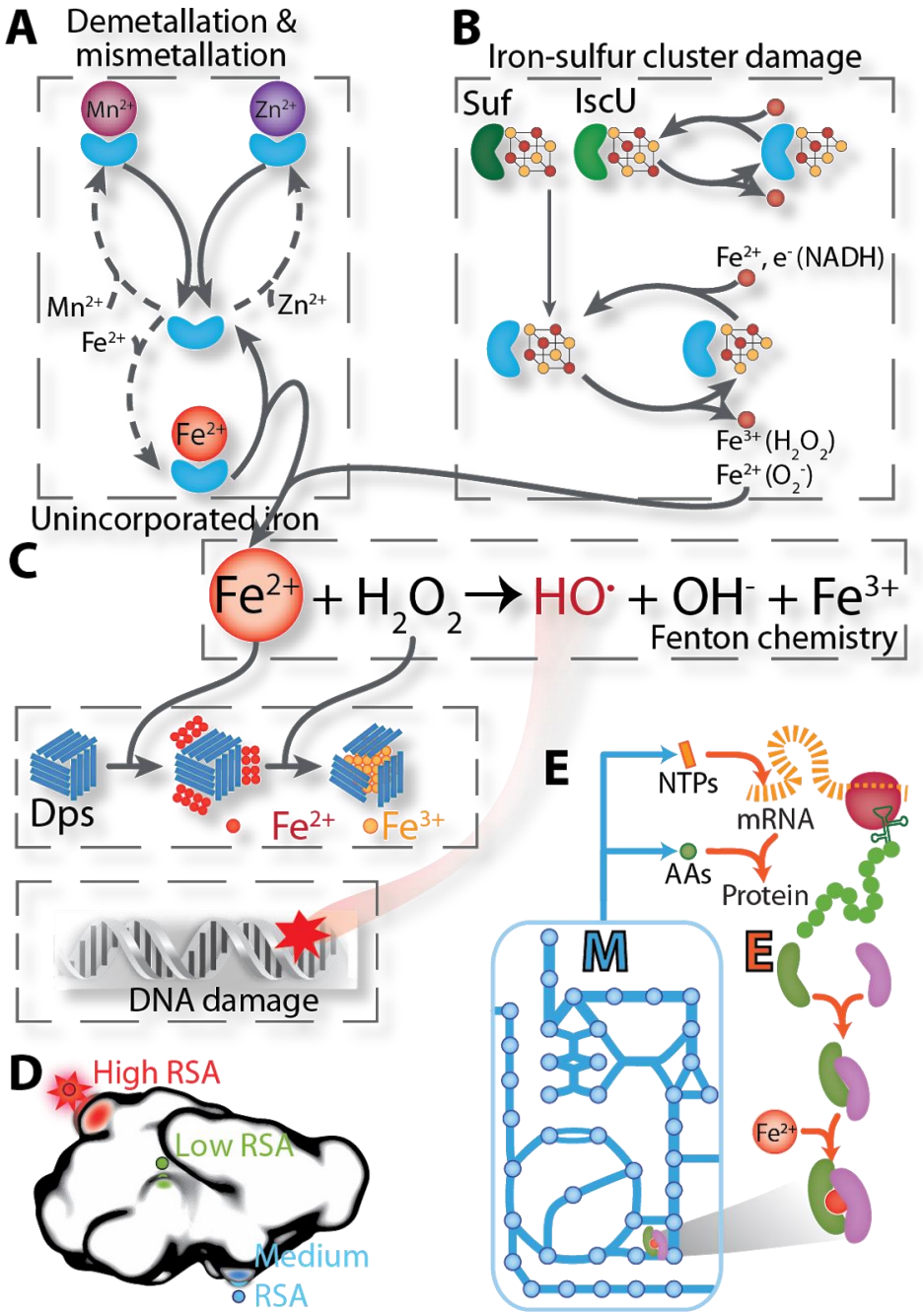



Figure 5.1. OxidizeME: a multi-scale description of metabolism and macromolecular expression that accounts for damage by reactive oxygen species (ROS) to macromolecules.
(A) Mononuclear Fe(II) proteins are demetallated by ROS and mismetallated with alternative divalent metal ions. (B) Iron-sulfur clusters are oxidized and repaired. (C) Unincorporated Fe(II) spontaneously reacts with H2O2 via Fenton chemistry, generating hydroxyl radicals that damage DNA, while the Dps protein stores unincorporated iron and protects DNA from damage. (D) Protein structural properties are computed to estimate the probability of metal cofactor damage by ROS (RSA: relative solvent accessibility). (E) Processes in A-D are integrated into a multiscale oxidative model, named OxidizeME. OxidizeME is used to compute the scope of macromolecular damage and the cellular response for varying intracellular concentrations of superoxide, hydrogen peroxide, and divalent metal ions (Fe(II), Mn(II), Co(II), Zn(II)), See SI for details.

A hallmark response to ROS damage for *E. coli* is the deactivation of branched-chain and aromatic amino acid biosynthesis pathways, which is alleviated by supplementing these amino acids [311]. Compared to supplementing all 20 amino acids, OxidizeME correctly predicted that excluding Ile and Val had a greater impact on growth rate than did excluding Phe, Trp, and Tyr (Fig. 2A). The reason that *E. coli* cannot grow under ROS stress without supplementation of branched-chain amino acids is that the iron-sulfur clusters of dihydroxy-acid dehydratase and isopropylmalate isomerase are inactivated by ROS, thus debilitating the branched-chain amino acid biosynthetic pathway [263]. The auxotrophy for aromatic amino acids was originally attributed to inactivation of the transketolase reaction [303], but was recently traced to the mismetallation of the mononuclear iron cofactor in 3-deoxy-D-arabinheptulosonate 7-phosphate (DAHP) synthase [320]. OxidizeME correctly predicted these molecular mechanisms and their phenotypic consequences (Fig. 2).

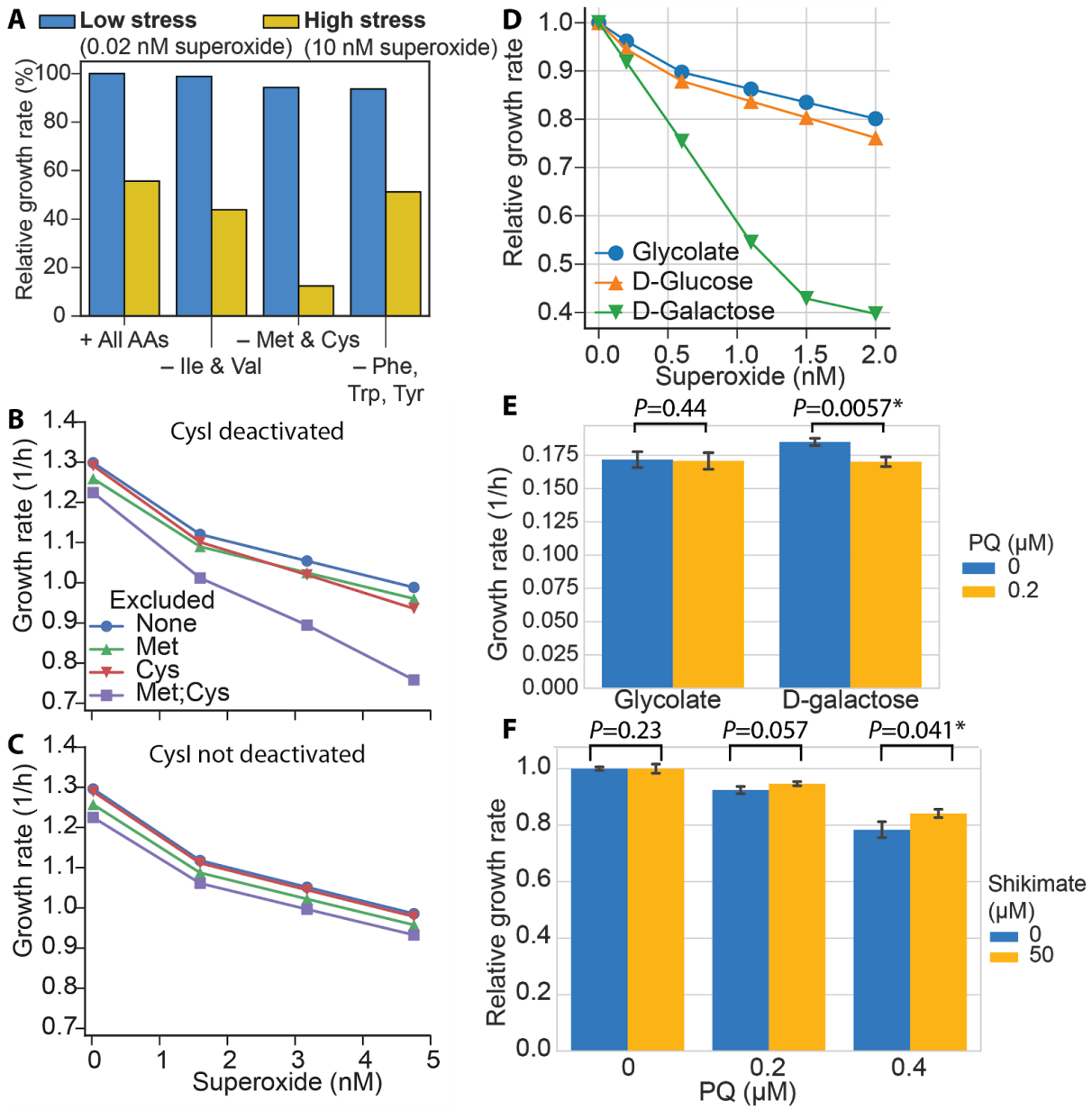



Figure 5.2. Systemic consequences of ROS stress.
(A) Predicted growth rate under low and high superoxide concentrations with different supplementation of amino acids (AAs). All AAs refers to all 20 common amino acids, and "- Ile & Val" means all amino acids except Ile and Val were supplemented. (B) Predicted growth rate versus superoxide concentration in various sulfurous amino acid supplementation media. (C) Same as (B) but simulated without damage to CysI by ROS. (D) Simulated growth rate in varying superoxide concentrations relative to simulations with zero intracellular superoxide. (E) Growth rate of MG1655 with and without 0.2 μM PQ on glycolate, D-galactose. * denotes that the growth rate without PQ is significantly greater than that with PQ (one-tailed Welch's t-test, P<0.01). (F) Relative growth rate of MG1655 grown on D-galactose, relative to that at 0 μM PQ with or without supplementation of 50 μM shikimate. * denotes that the relative growth rate without shikimate is significantly less than that with shikimate (one-tailed Welch's t-test, P<0.05).

Meanwhile, the basis of sulfurous amino acid auxotrophy in *E. coli* remains inconclusive despite multiple investigations [294, 329]. OxidizeME correctly predicted auxotrophy for sulfurous amino acids (cysteine and methionine) under ROS stress (Fig. 2A). We traced a plausible mechanism to damage of the iron-sulfur cluster in CysI, which catalyzes the sulfite reductase step of Cys biosynthesis. Sulfite reductase binds four cofactors: iron-sulfur, FAD, FMN, and siroheme. Consistent with prior studies [277], our structural model estimated the siroheme group of sulfite reductase to be difficult to reach by ROS, mainly due to the depth of the cofactor binding residue. Previous studies showed that the iron-sulfur cluster is likely not autoxidized with molecular oxygen because it is not solvent-exposed [277]. However, our structural model predicted that the iron-sulfur cluster is reached by ROS when considering both solvent exposure and depth of the cluster-binding residue from the solvent accessible surface. Simulations confirmed that alleviating damage to sulfite reductase was sufficient to reverse the observed growth rate defect and enable growth at higher ROS concentrations in the absence of Cys and Met (Fig. 2B,C). Our hypothesis that sulfite reductase is deactivated by ROS is consistent with studies in *Salmonella enterica* showing that the activity of this enzyme is indeed reduced by elevated superoxide [362]. Furthermore, the deactivation of sulfite reductase is consistent with accumulation of its substrate, sulfite, and explains the previously-observed accumulation of sulfite [294]. We note that CysI inactivation does not exclude the possibility that superoxide additionally leads to cell envelope damage, facilitating leakage of small molecules [278]. Thus, OxidizeME can be used to understand and predict the basis for amino acid auxotrophies as a systemic response to specific macromolecular vulnerabilities to ROS.

To investigate how environmental context affects ROS tolerance, we simulated growth under superoxide stress in 180 carbon sources (Fig. S2). OxidizeME predicted that glycolate and

D-galactose conferred enhanced and reduced ROS tolerance, respectively (Fig. 2D). To validate this prediction, we measured growth of *E. coli* MG1655 on these two carbon sources with 0.2 μM paraquat (PQ) and without PQ. PQ is a divalent cation that is taken up opportunistically, typically by polyamine transmembrane transporters, and then undergoes reduction and autoxidation cycles catalyzed by any of three *E. coli* PQ diaphorases to generate superoxide [330]. Consistent with OxidizeME's predictions, PQ treatment significantly decreased the growth rate by 8% on D-galactose (one-tailed Welch's t-test, $P$=0.0057), and insignificantly (0.58% decrease) on glycolate (one-tailed Welch's t-test, $P$=0.44) (Fig. 2E). OxidizeME suggested that biosynthesis of aromatic amino acids was strongly impacted by ROS in D-galactose but not in glycolate, due to a ROS-induced bottleneck in shikimate metabolism through inactivation of shikimate kinase II (AroL). To validate this hypothesis, we supplemented shikimate for *E. coli* growing on D-galactose at different PQ concentrations. The relative growth rate (relative to 0 μM PQ) was significantly greater by 7.3% with shikimate supplementation than without shikimate at 0.4 μM PQ (one-tailed Welch's t-test, $P$=0.041) (Fig. 2F). This result confirms our hypothesis that shikimate biosynthesis becomes a bottleneck under ROS stress and is consistent with our proposed mechanism that shikimate kinase II is inactivated.

Next, we assessed the systemic response of *E. coli* to ROS stress. We measured the transcriptome of *E. coli* under superoxide stress using PQ treatment and identified 914 differentially expressed genes (DEGs), of which 501 were accounted for in OxidizeME (Fig. 3, Fig. S3). In particular, 87 genes were up-regulated. Using OxidizeME, we determined that these 87 genes were more likely activated due to damage that is specific to iron metalloproteins than to any other protein ($P$<0.001). Furthermore, of the DEGs that were correctly predicted, a large fraction (84%) of the repressed genes changed due to decreased growth rate from PQ treatment,

while 95% of the activated genes were specific responses to stress (Fig. 3). Gene expression is expected to respond to ROS stress directly--e.g., by up-regulating ROS detoxification genes--and indirectly--in response to decreased metabolic rates caused by ROS damage. The responses we identified as being specific to ROS, not growth rate, spanned eight cellular processes (Fig. 3): ROS detoxification, central metabolism, anaerobic respiration, amino acid biosynthesis, cofactor synthesis and repair, translation, iron homeostasis, and transcriptional regulation by the *rpoS* sigma factor.

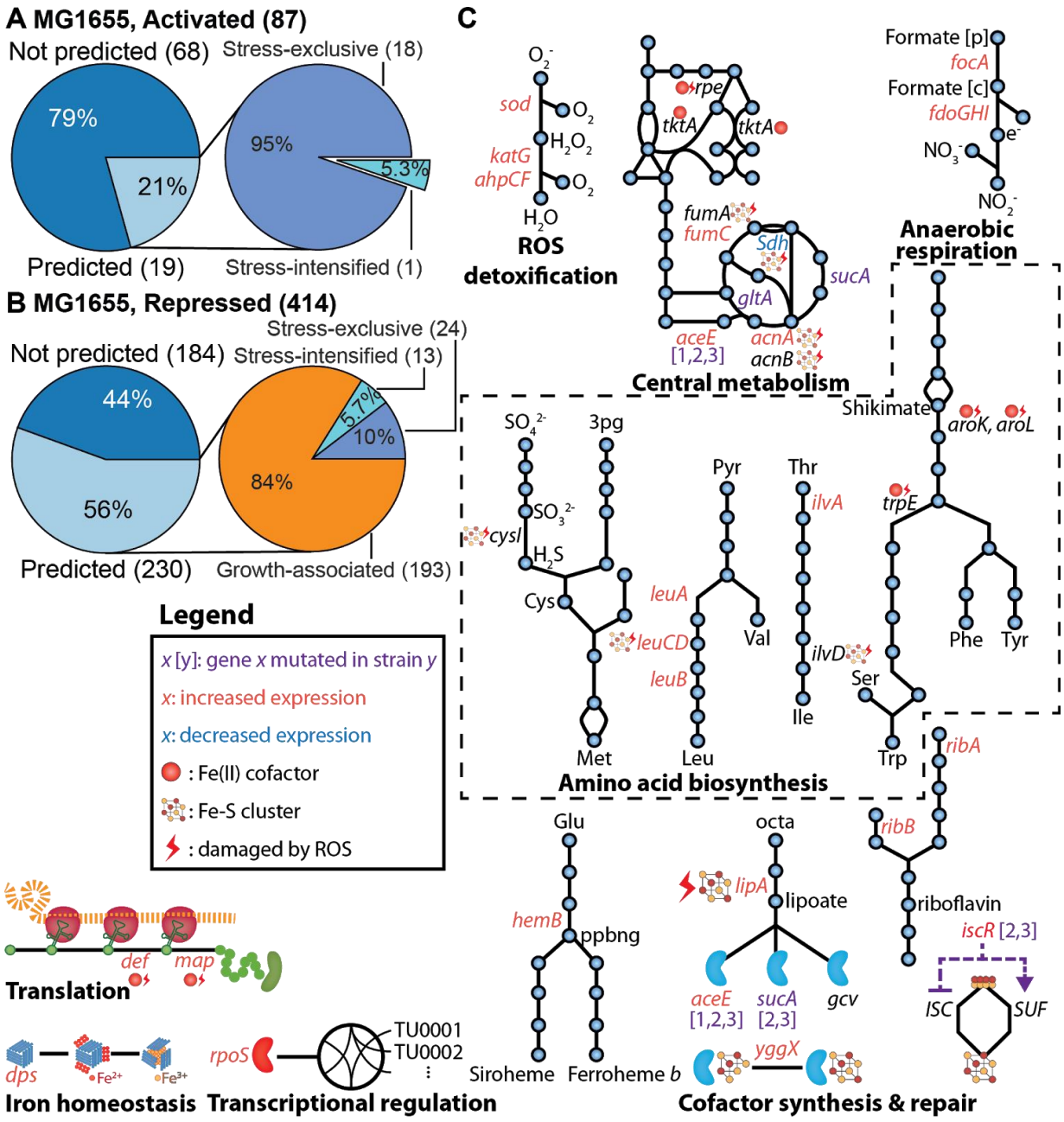



Figure 5.3. Validation of the consequences and responses to ROS stress.
Differentially expressed genes (DEGs) (|log2(fold-change)|>0.9|, FDR<0.01) that are activated (A) and repressed (B). Correctly predicted DEGs are distinguished from global growth-associated regulation using OxidizeME. (C) Cellular processes involved in a systemic response to iron metalloprotein damage by ROS.

To investigate the genetic basis for ROS tolerization, we deployed laboratory evolution to investigate how microbes adapt to tolerate high ROS stress. We adaptively evolved a glucose-optimized strain of *E. coli* [323], named GLU, in parallel cultures to tolerate three different concentrations of PQ (400, 600, and 800 μM). Re-sequencing identified a common set of genetic changes consisting of mutations in *ygfZ* (folate-binding protein), *aceE* (pyruvate dehydrogenase), and at least one gene in the citric acid cycle: *sucA* (alpha-ketoglutarate dehydrogenase), *gltA* (citrate synthase), or *icd* (isocitrate dehydrogenase) (Table S1). The two most ROS-tolerant strains additionally had mutations in *iscR* (iron-sulfur cluster regulator), while *arnA* (bifunctional, UDP-4-amino-4-deoxy-L-arabinose formyltransferase/UDP-glucuronate dehydrogenase) or *pitA* (metal phosphate:$H^+$ symporter) were mutated in each of these two strains. These results revealed that relatively few mutations confer microbial ROS tolerance, unlike adaptation to thermal stress, which can involve dozens to hundreds of mutations [307].

We performed RNA-Seq measurements on the PQ-adapted strains and found cellular responses that contrasted with those of the wild-type (MG1655) and with GLU. These differences also contrast with the conventional understanding of the ROS response. ROS damages components of the iron-sulfur cluster (ISC) assembly pathways through mismetallation of labile iron-sulfur clusters on the scaffold proteins IscU and SufA [314]. Under ROS stress, conventional understanding would suggest that the ISC assembly pathway is repressed and the sulfur assimilation (SUF) system is up-regulated [270, 308, 311, 322].

IscR regulates the transcription of both ISC and SUF based on coordination of 2Fe-2S at its Cys92, Cys98, and Cys104 residues [306, 308]. In the two most ROS-tolerant strains, we observed the mutation C104S in *iscR*. This mutation may hinder IscR's ability to incorporate 2Fe-2S and therefore to regulate expression of the ISC and SUF systems under ROS stress. Consistent

with deactivation of IscR, we observed expression of ISC and SUF in PQ-adapted strains that was opposite to the unadapted response, where ISC genes are repressed and SUF genes are up-regulated. When treated with 0.6 mM PQ, the most ROS-tolerant strain (PQ3) down-regulated the entire *sufABCDSE* transcription unit. Furthermore, compared to GLU, the PQ3 strain showed higher expression of the ISC pathway by up-regulating the *iscRSUA* and *hscBA-fdx-iscX* transcription units in 0.2 mM PQ.

This evolved response directly opposes that of wild-type MG1655 and GLU, which both down-regulated ISC genes and up-regulated *sufA*. These results show that *E. coli* does not necessarily rely on SUF under ROS stress, and that up-regulation of ISC may be integral for adaptation to higher ROS concentrations. OxidizeME correctly predicted up-regulation of the ISC system and repression of *sufABCDSE* under high ROS stress based on the relative cost-efficiency of Fe-S cluster biosynthesis from each respective system.

The use of iron-sulfur clusters and transition metals to catalyze biological processes can be traced back to the last universal common ancestor [274] and ROS stress has a profound impact on all aerobic life forms. Thus, OxidizeME represents a fundamental advance in our understanding of stress-response mechanisms as it provides a genome-wide description of metabolism, protein expression, prosthetic group engraftment, and ROS protecting mechanisms, that all together account for up to 85% of the proteome by mass. Finally, the ability to quantitatively and mechanistically describe responses to ROS damage of ancient conserved molecular targets has broad implications for organisms across the tree of life.

## 5.3 Materials and Methods

### 5.3.1 Demetallation and mismetallation of mononuclear iron enzymes

We assume that metallation occurs rapidly and is close to equilibrium [70]. Therefore, the concentrations of metal and complexes will depend on stability constants. Demetallation of mononuclear iron enzymes occurs with rate constants ($k_{cat}/K_M$) of $10^3$ to $10^4$ $M^{-1}s^{-1}$ *in vitro [310]*. We added reactions for mismetallation of mononuclear iron enzymes by alternate metals. The $k^{eff}$ was scaled relative to $Fe^{2+}$ for enzymes mismetallated with $Mn^{2+}$, $Co^{2+}$, and $Zn^{2+}$, based on published data [271, 359]. We know the total metal concentrations and some free concentrations [70]. We then fix the metal concentrations within these ranges. We then use proteomics to constrain the total enzyme concentrations. We assume that metallation proceeds rapidly and sufficiently close to equilibrium [70] that apoenzyme concentrations and dilution are negligible.

### 5.3.2 Limits of ROS damage rates

The $K_M$ of catalase for $H_2O_2$ is $8.65\times10^7$ nM [363], which represents a lower limit $K_M$ for reaction of $H_2O_2$ with the mononuclear metal or Fe-S cluster enzymes. Under glucose aerobic conditions, $H_2O_2$ concentration is around 50 nM. At 200 nM, OxyR is activated, and at 400 nM, growth defects are observed [310]. Therefore, for a viable cell, the $H_2O_2$ concentration is far below $K_M$ and we can approximate the damage rate as $v^{dmg} \approx k_{cat}/K_M ES$, where $E$ and $S$ are enzyme and $H_2O_2$ concentrations, respectively. Values for $k_{cat}/K_M$ for dehydratase and mononuclear iron enzymes range from $10^3$ to $10^4$ $M^{-1}s^{-1}$ [310]. We thus used a lower limit of $10^3$ $M^{-1}s^{-1}$ and created an ensemble of damage scenarios where damage rate constants varied within this range. Thus, at 50 nM $H_2O_2$, effective rate constants (i.e., $k^{eff} = k_{cat}/K_M S$) were between $5 \times 10^{-5}$ to $5 \times 10^{-4}$ $s^{-1}$ and at 400 nM $H_2O_2$, $k^{eff}$ was between $4 \times 10^{-4}$ to $4 \times 10^{-3}$ $s^{-1}$.

Similarly, rate constants for damage by $O^-_2$ exceed $10^6$ $M^{-1}s^{-1}$ [310]. Thus, at a basal $O^-_2$ concentration of 0.2 nM, $k^{eff}$ of damage by $O^-_2$ is about $2 \times 10^{-4}$ $s^{-1}$.

### 5.3.3 Modeling damage rates as functions of ROS concentration

We simulate ROS damage and repair by specifying intracellular ROS concentrations, which affect effective rate constants of damage and detoxification enzymes. Assuming ROS concentrations are much smaller than $K_M$ of damage reactions, individual damage fluxes are coupled to enzyme abundance as follows:

$$v^{dmg} = k_{cat}/K_M[ROS]E = k_{eff}^{dmg}\frac{v^{dil}}{\mu}$$

Thus, damage fluxes are proportional to ROS concentration and abundance of the damageable protein. Generated ROS is either detoxified or damages macromolecules:

$$\sum_{j\in Damaged} s_j v_j + \sum_{j\in Detox} s_j v_j = \sum_{j\in Generated} s_j v_j$$

$$[ROS] = [ROS]^{detox} + [ROS]^{damage}$$

$$= \sum_{j\in Detox} \frac{v_j \mu}{k_{eff,j} v_j^{dil}} + \sum_{j\in Damage} \frac{v_j \mu}{k_{eff,j} v_j^{dil}}$$

Thus, increased detoxification decreases the ROS left to damage macromolecules. The concentration of ROS that is not detoxified determines the rate of macromolecule damage. For a fixed [ROS], OxidizeME determines the optimal distribution of ROS between detoxification and damage. Typically, detoxification will be maximized until its capacity is exceeded. We also assume [ROS] concentrations do not change at steady-state:

$$\frac{d[ROS]}{dt} = \sum_{j\in Production} s_j v_j - \sum_{j\in Consumption} s_j v_j = 0$$

### 5.3.4 Solving the OxidizeME model as an optimization problem

Combining all new variables and constraints, the final OxidizeME optimization problem for maximizing growth rate subject to oxidative stress is the following:

$$\max_{\mu,v} \ \mu$$

$$s.t. \ \ Sv = 0$$

$$v^{dmg} = \frac{k_{cat}}{K_M} \cdot [ROS] \cdot \frac{v^{dil}}{\mu}$$

$$v_j^{dmg} - v_j^{repair} - v_j^{dil} = 0$$

$$v_j^{dil} = \frac{\mu}{k_{repair,j}} v_j^{repair}, j \in DamagedComplex$$

$$v_i^{met} = \frac{\beta^i [Metal\ i]}{\beta^{Fe}[Fe(II)]} \cdot \left( \sum_j v_{E:Fe}^{demet} + v_{E:Fe}^{dil} \right),$$

$$i \in AltMetals$$

where $W$ is the cell specific weight (gDW/L) $= 278 * \frac{10^{-15} gDW}{6.8*10^{-16} L}$, $\mu$ is the specific growth rate (1/h), [$E$], [$E{:}i$], [$i$] are concentrations ($M$) of apoenzyme, holoenzyme, and metal ions, respectively, $AltMetals = \{Zn(II), Mn(II), \dots\}$, and $\beta^k$ are metal-protein stability constants (M).

### 5.3.5 Protein structural property computation

To enable a structural systems biology approach of computing ROS damage probabilities for each metal-binding enzyme in the OxidizeME model, we utilized the GEM-PRO pipeline [273] implemented in the ssbio Python package [309] to gather all available protein structures for model enzymes from the Protein Data Bank [285]. We set a single representative structure for each enzyme based on the following four factors: 1) sequence identity and coverage of the wild-type amino acid sequence; 2) presence of the annotated metal binding site 3) presence of the model-annotated metal ion; and 4) presence or absence of any model-annotated cofactors other than the metal ion. The MetalPDB database [312] was used to expedite this analysis. We used homology

models from the I-TASSER pipeline if no experimental structure was available [313]. We used these models to compute structural properties that would help determine if a metal cofactor was damaged by ROS. Specifically, we built a Bayesian network (Fig. S1) using the pomegranate Python package [280]. The conditional nodes of the Bayesian network are: A) the probability of ROS reaching the metal-binding site which depends on 1) the maximum solvent accessibility of each of the metal-binding residues computed by FreeSASA [278] as well as 2) its associated depth computed by MSMS [286]; and B) the probability of ROS damaging the metal-binding site which depends on A), as well as if the site uses a cysteine residue for binding or is in a 5 Å radius around the binding site as computed using Biopython PDB [279] and MetalPDB [312]. The damage probability is then used to scale damage rates in the OxidizeME model.

### 5.3.6 Differentiating stress–specific from growth-associated responses

We simulated transcriptomes over lowered growth rates using two models—one accounting for stress and the other not. The growth rate was determined by the stress model over increasing superoxide concentrations, leading to lowered growth rate, and this growth rate was used for the non-stress model. DEGs that were correctly predicted by the stressed model but not the unstressed model were considered to be stress-associated. DEGs that were correctly predicted by the unstressed model, and were not considerably more differentially expressed by the stressed model were considered growth-associated.

### 5.3.7 Classifying differentially expressed genes under oxidative stress using OxidizeME

We used OxidizeME to classify genes as activated, repressed, or unchanged in expression in response to ROS stress. To do so, we computed the change in transcription rates of all transcription units in the model for intracellular superoxide concentrations ranging from 2e-10 to

2e-9, which is approximately 10 to 100 times the concentration under aerobic growth without PQ treatment. We then used the average of these computed transcription rate changes. Because certain genes changed expression nonlinearly with superoxide concentration, including exhibiting an inverse response, and because the transcription rate change depends on the reference state simulation, we also computed the Spearman rank correlation of transcription rate with superoxide concentration. Thus, if a gene had a lower relative rate than the reference but its rate was positively and significant ($P < 0.05$) correlated with superoxide, we used this Spearman rank value as the predictor of expression change.

We then classified genes as activated or repressed based on a threshold transcription rate change, which we varied and investigated using a precision-recall curve. The best classifier could then be chosen using the threshold maximizing $F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$ (Fig. S3).

### 5.3.8 OxidizeME software

All code used to build and use OxidizeME is available at https://github.com/SBRG/oxidizeme.

### 5.3.9 RNA sequencing

RNA-sequencing data were generated under conditions of exponential-phase, aerobic growth in glucose M9 minimal media with appropriate concentration paraquat. Cells were washed with Qiagen RNAprotect Bacteria Reagent and pelleted for storage at -80 ◦C prior to RNA extraction. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% sodium dodecyl sulfate for 20min at 37 ◦C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns, following vendor procedures. An on-column DNase-treatment was performed for 30 min at room temperature. RNA was quantified using a

Nano drop and quality assessed by running an RNAnano chip on a bioanalyzer. Paired-end, strand-specific RNA-seq was performed following a modified dUTP method [269]. The rRNA was isolated using Epicentres Ribo-Zero rRNA removal kit for Gram Negative Bacteria. Sequences were run on an Illumina HiSeq using a KAPA Stranded RNA-seq kit.

Reads were mapped to the *E. coli* K12 MG1655 Genome (NC 000913.2) with bowtie2 [307]. Expression levels in fragments per kilobase per million mapped (FPKM), and differentially expressed genes (DEGs) were found using DESeq2 [293]. A fold-change of 1.5 and FDR-adjusted p-value cutoff of 0.05 were used to call significant differential expression.

### 5.3.10 Adaptive laboratory evolution and phenotypic profiling

The starting strain for ROS tolerization adaptive laboratory evolution was taken from previous adaptive laboratory evolution study (insert reference) where *E. coli* strain MG1655 (ATCC47076) was evolved under 4 g/l glucose M9 media. Throughout the experiment, all the strinas were cultured in flasks filled with 15 ml of 4 g/l glucose M9 media with varying concentrations of paraquat. Each flask was aerated by a magnetic stirrer at 1,100 rpm. The temperature of the cultures were kept at 37 ◦C by being placed in a heat block. The starting strains were inoculated into 15 ml of 4 g/l glucose M9 media with 0.01 mM of paraquat and cultured overnight to reach stationary phase. 800 µl of each culture was then inoculated into three independent flasks containing 15 ml of 4 g/l glucose M9 media with 0.1 mM of paraquat. For each flask, OD at 600 nm was measured four times while the OD of the culture was within 0.05-0.3 ranges. Once the OD reaches 0.3, 800 µl of the culture was passed into a new flask. Based on the growth rate of the previous flask, the sampling time points were calculated so that they would evenly spread out to capture accurate growth rate. The slope of the least-squares linear regression line that fits the logarithm of the OD measurements was taken as the growth rate.

While doing the tolerization adaptive laboratory evolution experiment, the concentration of paraquat was increased by 0.1 mM whenever the culture has been passed at least three times without increasing the paraquat concentration and the most recent growth rate was above 0.5. The concentration of paraquat is also increased by 0.1 mM whenever the culture has been passed at least six times without increasing the paraquat concentration and the most recent growth rate was in between 0.15 and 0.5 $h^{-1}$.

Periodically, 800 $\mu$l of culture samples were mixed with 800 $\mu$l of autoclaved 50% glycerol solution and stored at -80°C. Glucose M9 media consisted of 4 g/l glucose, 0.1 mM $CaCl_2$, 2.0 mM $MgSO_4$, 1x trace element solution and 1x M9 salts. 4000x trace element solution consisted of 27 g/l $FeCl_3*6H_2O$, 2 g/l $ZnCl_2*4H_2O$, 2 g/l $CoCl_2*6H_2O$, 2 g/l $NaMoO_4*2H_2O$, 1 g/l $CaCl_2*H_2O$, 1.3 g/l $CuCl_2*6H_2O$, 0.5 g/l $H_3BO_3$, and concentrated HCl dissolved in $ddH_2O$ and sterile filtered. 10x M9 salts solution consisted of 68 g/l $Na_2HPO_4$ anhydrous, 30 g/l $KH_2PO_4$, 5 g/l NaCl, and 10 g/l $NH_4Cl$ dissolved in $ddH_2O$ and autoclaved.

### 5.3.11 DNA sequencing

Selected cultures were streaked on to Lysogeny Broth (LB) agar plates and grown overnight at 37 °C. The genomic DNA of the isolated colonies was extracted using Promegas Wizard DNA Purification Kit. The quality of DNA was assessed with ultraviolet absorbance ratios using a Nano drop. DNA was quantified using Qubit dsDNA High Sensitivity assay. Paired-end resequencing libraries were generated using a Kapa Hyper Plus kit from Kapa Biosystems (Wilmington, MA). Sequences were obtained using an Illumina NextSeq with a NextSeq 500/550 High Output Kit v2 (300 cycles). The breseq pipeline [324] version 0.30.2 with bowtie2 [307] was used to map sequencing reads and identify mutations relative to the *E. Coli* K12 MG1655 genome

(NCBI accession number NC 000913.2). All samples had an average mapped coverage of at least 25x

## 5.4 Acknowledgements

Chapter 5, in full, is a reprint of the material as it appears in: Yang L, Mih N, Park JH, Tan J, Seo S, Kim D, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Feist AM, Palsson BO. 2018. Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms. Submitted. The dissertation author was a contributing author of this paper.

# Chapter 6.

# Conclusions and Future Directions

## 6.1 Conclusions

Much of the motivation for this dissertation was originally driven by the promise of discovery in the relatively unexplored realm of structural systems biology. It is easier said than done to attempt to combine the theories and ideologies of distinctly different fields in computational biology. The initial challenge presented itself as somewhat of a mundane one, with logistical challenges of grunt work and reproducibility. However, therein lied opportunity for software to be written and educational outreach to be conducted. One of my personal goals with the work presented here was to provide robust tools and examples for the merging of these two fields to continue to find its footing.

In this dissertation, each application was enabled by this integrative approach. What was originally a mish-mash of scripts and custom workflows matured into a Python package which sparked new collaborations with my colleagues and beyond. Taking an analogy to different magnification levels on a microscope, the structural systems biology approach allows one to start exploring biological questions from a zoomed-out view of cellular networks, while retaining the ability to zoom in to the atomic detail provided by protein structures. With that in mind, we have attempted here to zoom in almost as far as molecular modeling tools allowed us to in understanding the impact of mutational changes on structures and systems. Zooming out, we then utilized genome-scale models as a knowledgebase to compare features of the structural proteome in different strains and species. Finally, we have informed systems-level simulations with structural

parameters to enable new capabilities in metabolic modeling (Figure 1). These studies have further extended the GEM-PRO methodology, with an overarching goal of reaching out to both the systems and structural biology communities.

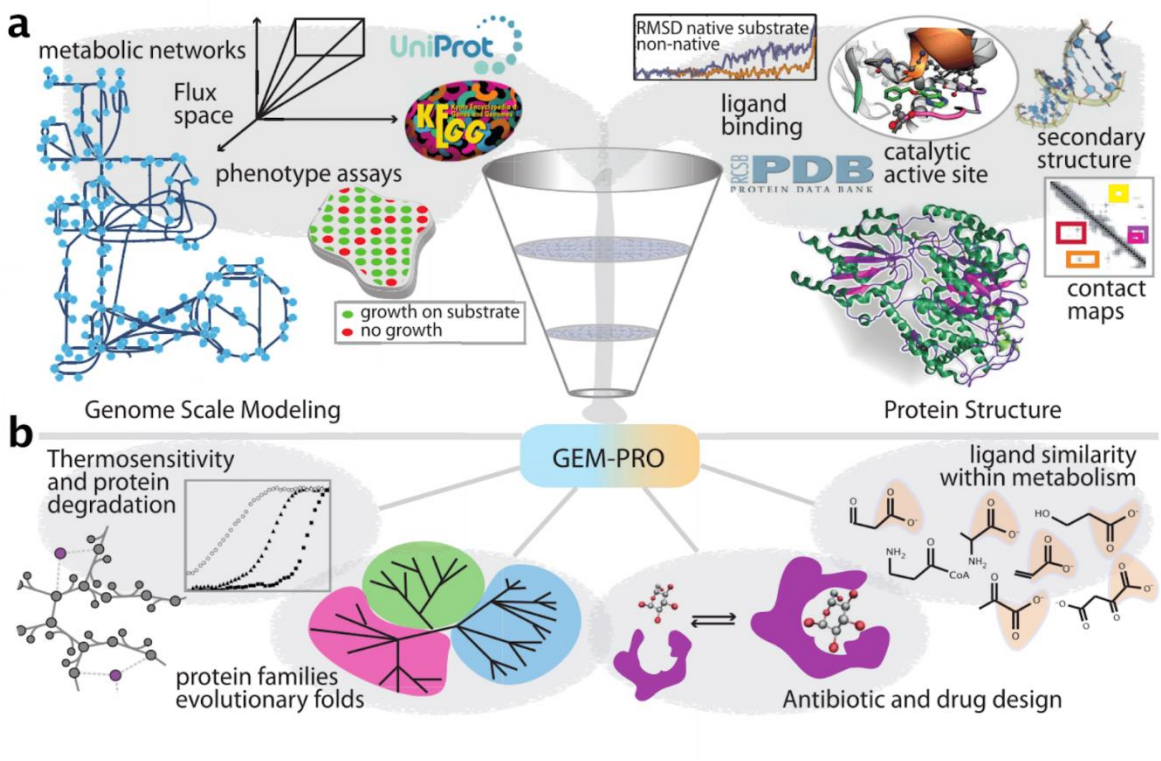



Figure 6.1. Integrating metabolic systems biology with protein structures brings the promise of structural systems biology (a) and applications within multiple areas of study (b).

## 6.2 Future Directions

There remains much to be gained from the integration of molecular-level details into systems approaches to modeling. One obvious, perhaps currently intractable model that could be built, is a detailed representation of the protein folding and repair machinery in response to all types of cellular stress. We have attempted to do so for the context of thermal stress [34], and these types of models provide the groundwork for true multi-scale modeling in wide-ranging applications such in understanding the complicated repair response to oxidative stress or creating models of aging processes. In the biotechnology space, metabolic models alone have proved their

worth in engineering yeast and bacteria to synthesize high-value chemicals [275, 330, 364]. However, a common challenge lies in the optimization of certain enzymes in synthesis pathways, and structural approaches to enhance low-performing components could be the ultimate step in the quest for efficient biosynthetic processes.

Technical issues and computational bottlenecks also must be addressed. The prediction of kinetic rate constants across an entire metabolic network with quantum mechanics/molecular mechanics (QM/MM) methods is most definitely decades away. However, there has been much work done in scaling up and coarse-graining molecular modeling simulations [69], and in one example even enabled the dynamic simulation of an entire cytoplasm [321, 342]. The interest of the molecular modeling community in bringing their methods to the systems-scale is absolutely required to integrate the best of both worlds.

In what can be considered another branch of systems biology, a deep understanding of the interactome - including macromolecular interactions, signaling networks, and regulatory networks - still evades us. Modeling protein-protein and protein-DNA interactions within a cell may be one of the biggest challenges to overcome since so many of these interactions are transient, difficult to measure experimentally, and unreliably predicted computationally [23, 277, 336]. Inspecting and measuring condition-dependent states of cells provides us with small but incremental steps towards elucidating the interactome.

Lastly, what I consider to be the "final frontier" of structural systems biology would be an accurate visual model of a cell, incorporating all known interactions, localizations, enzymes complexes, etc. (Figure 2). The construction of this model would serve as a proof-of-concept that we actually do, in fact, understand a large portion of a model organism enough to visualize it computationally. This could serve as a starting point for large-scale molecular modeling

simulations, for mathematical systems models incorporating principles of diffusion and cell size, and for contextualization of interactome data. On the flip side, a visual model would be a rich entry point for students and non-scientists to begin to learn about the basics of life right down to the atomic level.

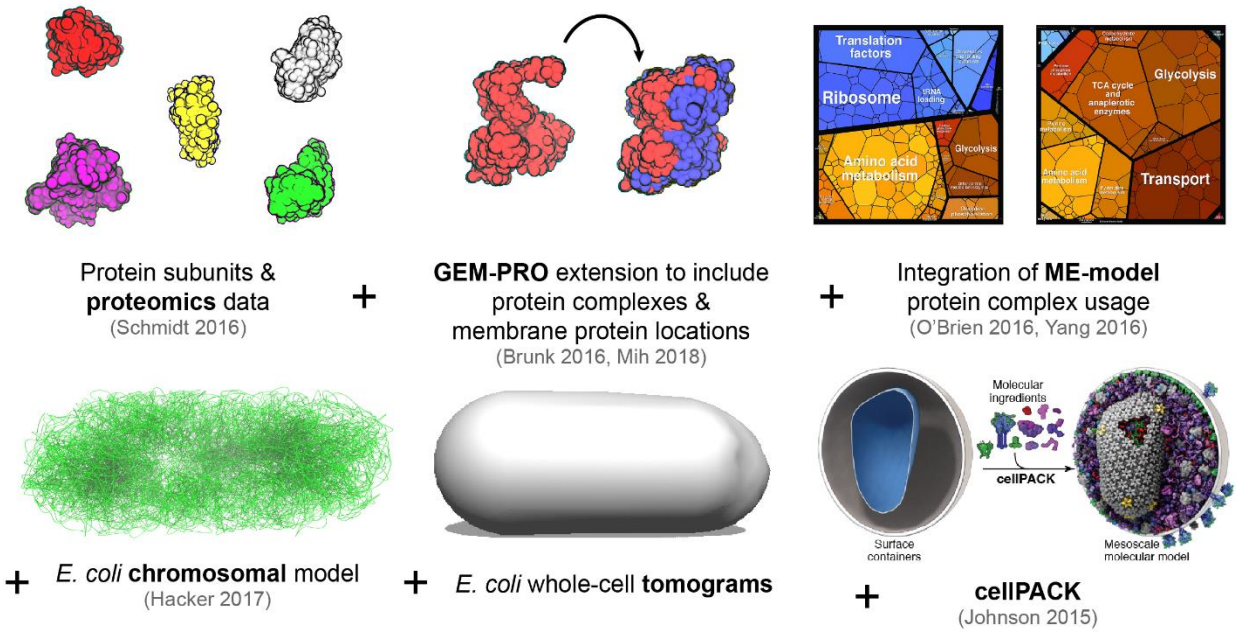



Figure 6.2. A proposed workflow to generate a visual model of an *E. coli* cell.
To build a high-level visual model of a cell, certain experimental data and computational techniques are required. This proposed model integrates proteomics data measuring copy numbers of individual protein subunits [365], established pipelines for the annotation of structural information [27, 301], simulations of protein complex usage within a metabolic network [357, 362], a model of the compacted *E. coli* chromosome [305], publicly available three-dimensional tomograms, and finally, an algorithm to "pack" the components into the cell [329].

# References


1. Beltrao P, Kiel C, Serrano L: **Structures in systems biology**. *Curr Opin Struct Biol* 2007, **17**(3):378-384.

2. Gu J, Bourne PE: **Structural Bioinformatics**: John Wiley & Sons; 2009.

3. Hamelryck T, Manderick B: **PDB file parser and structure class implemented in Python**. *Bioinformatics* 2003, **19**(17):2308-2310.

4. Biasini M, Mariani V, Haas J, Scheuber S, Schenk AD, Schwede T, Philippsen A: **OpenStructure: a flexible software framework for computational structural biology**. *Bioinformatics* 2010, **26**(20):2626-2628.

5. O'Donoghue SI, Sabir KS, Kalemanov M, Stolte C, Wellmann B, Ho V, Roos M, Perdigão N, Buske FA, Heinrich J, Rost B, Schafferhans A: **Aquaria: simplifying discovery and insight from protein structures**. *Nat Methods* 2015, **12**(2):98-99.

6. Grünberg R, Nilges M, Leckner J: **Biskit--a software platform for structural bioinformatics**. *Bioinformatics* 2007, **23**(6):769-770.

7. Mizianty MJ, Fan X, Yan J, Chalmers E, Woloschuk C, Joachimiak A, Kurgan L: **Covering complete proteomes with X-ray structures: a current snapshot**. *Acta Crystallogr D Biol Crystallogr* 2014, **70**(Pt 11):2781-2793.

8. Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H: **Software for systems biology: from tools to integrated platforms**. *Nat Rev Genet* 2011, **12**(12):821-832.

9. O'Brien EJ, Monk JM, Palsson BO: **Using Genome-scale Models to Predict Biological Capabilities**. *Cell* 2015, **161**(5):971-987.

10. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR: **COBRApy: COnstraints-Based Reconstruction and Analysis for Python**. *BMC Syst Biol* 2013, **7**:74.

11. Kabsch W, Sander C: **DSSP: definition of secondary structure of proteins given a set of 3D coordinates**. *Biopolymers* 1983, **22**:2577-2637.

12. Sanner MF, Olson AJ, Spehner J-C: **Reduced surface: an efficient way to compute molecular surfaces**. *Biopolymers* 1996, **38**(3):305-320.

13. Cheng J, Randall AZ, Sweredoski MJ, Baldi P: **SCRATCH: a protein structure and structural feature prediction server**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W72-76.

14. Roy A, Kucukural A, Zhang Y: **I-TASSER: a unified platform for automated protein structure and function prediction**. *Nat Protoc* 2010, **5**(4):725-738.

15. Oobatake M, Ooi T: **Hydration and heat stability effects on protein unfolding**. *Prog Biophys Mol Biol* 1993, **59**(3):237-284.

16. Dill KA, Ghosh K, Schmit JD: **Physical limits of cells and proteomes**. *Proc Natl Acad Sci U S A* 2011, **108**(44):17876-17882.

17. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**(3):567-580.

18. Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ: **A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins**. *PLoS One* 2013, **8**(1):e54175.

19. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, Others: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11):1422-1423.

20. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.

21. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL: **OPM database and PPM web server: resources for positioning of proteins in membranes**. *Nucleic Acids Res* 2012, **40**(Database issue):D370-376.

22. Gromiha MM, Thangakani AM, Selvaraj S: **FOLD-RATE: prediction of protein folding rates from amino acid sequence**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W70-74.

23. Frishman D, Argos P: **Knowledge-based protein secondary structure assignment**. *Proteins* 1995, **23**(4):566-579.

24. Mitternacht S: **FreeSASA: An open source C library for solvent accessible surface area calculations**. *F1000Res* 2016, **5**:189.

25. Ye Y, Godzik A: **Flexible structure alignment by chaining aligned fragment pairs allowing twists**. *Bioinformatics* 2003, **19 Suppl 2**:ii246-255.

26. Duke RE, Giese TJ, Gohlke H, Goetz AW, Homeyer N, Izadi S, Janowski P, Kaus J, Kovalenko A, Lee TS, Others: **AmberTools 16**. 2016.

27. Brunk E, Mih N, Monk J, Zhang Z, O'Brien EJ, Bliven SE, Chen K, Chang RL, Bourne PE, Palsson BO: **Systems biology of the structural proteome**. *BMC Syst Biol* 2016, **10**(1):26.

28. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, al E: **Three-dimensional structural view of the central metabolic network of Thermotoga maritima**. *Science* 2009, **325**(5947):1544-1549.

29. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, Feist AM, Palsson BO: **iML1515, a knowledgebase that computes Escherichia coli traits**. *Nat Biotechnol* 2017, **35**(10):904-908.

30. Broddrick JT, Rubin BE, Welkie DG, Du N, Mih N, Diamond S, Lee JJ, Golden SS, Palsson BO: **Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis**. *Proc Natl Acad Sci U S A* 2016, **113**(51):E8344-E8353.

31. Chang RL, Xie L, Xie L, Bourne PE, Palsson BØ: **Drug off-target effects predicted using structural analysis in the context of a metabolic network model**. *PLoS Comput Biol* 2010, **6**(9):e1000938.

32. Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BO: **Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli**. *Science* 2013, **340**(6137):1220-1223.

33. Mih N, Brunk E, Bordbar A, Palsson BO: **A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism**. *PLoS Comput Biol* 2016, **12**(7):e1005039.

34. Chen K, Gao Y, Mih N, O'Brien EJ, Yang L, Palsson BO: **Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation**. *Proceedings of the National Academy of Sciences* 2017, **114**(43):11548-11553.

35. Ong WK, Vu TT, Lovendahl KN, Llull JM, Serres MH, Romine MF, Reed JL: **Comparisons of Shewanella strains based on genome annotations, modeling, and experiments**. *BMC Syst Biol* 2014, **8**:31.

36. Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, Palsson BO, Herrgård MJ, Feist AM: **Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes**. *Cell Syst* 2016, **3**(3):238-251.e212.

37. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ: **Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity**. *Proc Natl Acad Sci U S A* 2016, **113**(26):E3801-3809.

38. McKinney W: **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython**: " O'Reilly Media, Inc."; 2012.

39. Rose AS, Hildebrand PW: **NGL Viewer: a web application for molecular visualization**. *Nucleic Acids Res* 2015, **43**(W1):W576-579.

40. Nguyen H, Case DA, Rose AS: **NGLview - Interactive molecular graphics for Jupyter notebooks**. *Bioinformatics* 2017.

41. King ZA, Dräger A, Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO: **Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways**. *PLoS Comput Biol* 2015, **11**(8):e1004321.

42. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J: **BioServices: a common Python package to access biological Web Services programmatically**. *Bioinformatics* 2013, **29**(24):3241-3242.

43. Ingelman-Sundberg M: **Pharmacogenetics: an opportunity for a safer and more efficient pharmacotherapy**. *J Intern Med* 2001, **250**(3):186-200.

44. Frueh FW, Amur S, Mummaneni P, Epstein RS, Aubert RE, DeLuca TM, Verbrugge RR, Burckart GJ, Lesko LJ: **Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use**. *Pharmacotherapy* 2008, **28**(8):992-998.

45. Wedlund PJ, de Leon J: **Pharmacogenomic testing: the cost factor**. *Pharmacogenomics J* 2001, **1**(3):171-174.

46. Zielinski DC, Filipp FV, Bordbar A, Jensen K, Smith JW, Herrgard MJ, Mo ML, Palsson BO: **Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis**. *Nat Commun* 2015, **6**:7101.

47. Jamshidi N, Wiback SJ, Palsson B BØ: **In silico model-driven assessment of the effects of single nucleotide polymorphisms (SNPs) on human red blood cell metabolism**. *Genome Res* 2002, **12**(11):1687-1692.

48. Rajith B, C GPD: **Path to Facilitate the Prediction of Functional Amino Acid Substitutions in Red Blood Cell Disorders – A Computational Approach**. *PLoS One* 2011, **6**(9):e24607.

49. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ: **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proceedings of the National Academy of Sciences* 2007, **104**(6):1777-1782.

50. Bordbar A, Jamshidi N, Palsson BO: **iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states**. *BMC Syst Biol* 2011, **5**:110.

51. Mardinoglu A, Gatto F, Nielsen J: **Genome-scale modeling of human metabolism - a systems biology approach**. *Biotechnol J* 2013, **8**(9):985-996.

52. Hecht M, Bromberg Y, Rost B: **Better prediction of functional effects for sequence variants**. *BMC Genomics* 2015, **16 Suppl 8**:S1.

53. Jorgensen WL: **The many roles of computation in drug discovery**. *Science* 2004, **303**(5665):1813-1818.

54. Wist AD, Berger SI, Iyengar R: **Systems pharmacology and genome medicine: a future perspective**. *Genome Med* 2009, **1**(1):11.

55. Yang R, Niepel M, Mitchison TK, Sorger PK: **Dissecting variability in responses to cancer chemotherapy through systems pharmacology**. *Clin Pharmacol Ther* 2010, **88**(1):34-38.

56. Duran-Frigola M, Mosca R, Aloy P: **Structural systems pharmacology: the role of 3D structures in next-generation drug development**. *Chem Biol* 2013, **20**(5):674-684.

57. Xie L, Ge X, Tan H, Xie L, Zhang Y, Hart T, Yang X, Bourne PE: **Towards structural systems pharmacology to study complex diseases and personalized medicine**. *PLoS Comput Biol* 2014, **10**(5):e1003554.

58. Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J: **Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling**. *Mol Syst Biol* 2014, **10**:721.

59. Turner RM, Park BK, Pirmohamed M: **Parsing interindividual drug variability: an emerging role for systems pharmacology**. *Wiley Interdiscip Rev Syst Biol Med* 2015, **7**(4):221-241.

60. Tan H, Ge X, Xie L: **Structural systems pharmacology: a new frontier in discovering novel drug targets**. *Curr Drug Targets* 2013, **14**(9):952-958.

61. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R: **Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review**. *Pharmacol Ther* 2013, **138**(3):333-408.

62. Hart T, Xie L: **Providing data science support for systems pharmacology and its implications to drug discovery**. *Expert Opin Drug Discov* 2016, **11**(3):241-256.

63. Xie L, Xie L, Bourne PE: **Structure-based systems biology for analyzing off-target binding**. *Curr Opin Struct Biol* 2011, **21**(2):189-199.

64. Chang RL, Xie L, Bourne PE, Palsson BO: **Antibacterial mechanisms identified through structural systems pharmacology**. *BMC Syst Biol* 2013, **7**:102.

65. Sorger PK, Allerheiligen SRB, Abernethy DR, Altman RB, Brouwer KLR, Califano A, D'Argenio DZ, Iyengar R, Jusko WJ, Lalonde R, Others: **Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms**. In: *An NIH white paper by the QSP workshop group: 2011*. 2011: 1-48.

66. Zhao S, Iyengar R: **Systems pharmacology: network analysis to identify multiscale mechanisms of drug action**. *Annu Rev Pharmacol Toxicol* 2012, **52**:505-521.

67. Gottlieb A, Altman RB: **Integrating systems biology sources illuminates drug action**. *Clin Pharmacol Ther* 2014, **95**(6):663-669.

68. Iyengar R, Zhao S, Chung S-W, Mager DE, Gallo JM: **Merging systems biology with pharmacodynamics**. *Sci Transl Med* 2012, **4**(126):126ps127.

69. The UniProt Consortium: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2017, **45**(D1):D158-D169.

70. Berman HM: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

71. Thorn CF, Klein TE, Altman RB: **Pharmacogenomics and bioinformatics: PharmGKB**. *Pharmacogenomics* 2010, **11**(4):501-505.

72. Cossum PA: **Role of the red blood cell in drug metabolism**. *Biopharm Drug Dispos* 1988, **9**(4):321-336.

73. Hinderling PH: **Red blood cells: a neglected compartment in pharmacokinetics and pharmacodynamics**. *Pharmacol Rev* 1997, **49**(3):279-295.

74. Boudíková B, Szumlanski C, Maidak B, Weinshilboum R: **Human liver catechol-O-methyltransferase pharmacogenetics**. *Clin Pharmacol Ther* 1990, **48**(4):381-389.

75. Fujii H, Miwa S: **Red blood cell enzymes and their clinical application**. *Adv Clin Chem* 1998, **33**:1-54.

76. Zimmerman HJ: **Hepatotoxicity: the adverse effects of drugs and other chemicals on the liver**: Lippincott Williams & Wilkins; 1999.

77. Fujii H, Miwa S: **Other erythrocyte enzyme deficiencies associated with non-haematological symptoms: phosphoglycerate kinase and phosphofructokinase deficiency**. *Baillieres Best Pract Res Clin Haematol* 2000, **13**(1):141-148.

78. Sender R, Fuchs S, Milo R: **Revised estimates for the number of human and bacteria cells in the body**. *bioRxiv* 2016.

79. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.

80. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A: **OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders**. *Nucleic Acids Res* 2015, **43**(D1):D789-D798.

81. Famiglietti ML, Estreicher A, Gos A, Bolleman J, Géhant S, Breuza L, Bridge A, Poux S, Redaschi N, Bougueleret L, Xenarios I, UniProt C: **Genetic variations and diseases in**

**UniProtKB/Swiss-Prot: the ins and outs of expert manual curation**. *Hum Mutat* 2014, **35**(8):927-935.

82. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS: **DrugBank 4.0: shedding new light on drug metabolism**. *Nucleic Acids Res* 2014, **42**(Database issue):D1091-1097.

83. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R: **SuperTarget and Matador: resources for exploring drug-target relationships**. *Nucleic Acids Res* 2008, **36**(Database issue):D919-922.

84. Zhou H, Skolnick J: **Template-based protein structure modeling using TASSER(VMT.)**. *Proteins* 2012, **80**(2):352-361.

85. Shen Y, Liu J, Estiu G, Isin B, Ahn YY, Lee DS, Barabási AL, Kapatral V, Wiest O, Oltvai ZN: **Blueprint for antimicrobial hit discovery targeting metabolic networks**. *Proc Natl Acad Sci U S A* 2010, **107**(3):1082-1087.

86. Kazakiewicz D, Karr JR, Langner KM, Plewczynski D: **A combined systems and structural modeling approach repositions antibiotics for Mycoplasma genitalium**. *Comput Biol Chem* 2015.

87. Weinshilboum RM, Raymond FA: **Inheritance of low erythrocyte catechol-o-methyltransferase activity in man**. *Am J Hum Genet* 1977, **29**(2):125-135.

88. McLeod HL, Fang L, Luo X, Scott EP, Evans WE: **Ethnic differences in erythrocyte catechol-O-methyltransferase activity in black and white Americans**. *J Pharmacol Exp Ther* 1994, **270**(1):26-29.

89. Maltête D, Cottard AM, Mihout B, Costentin J: **Erythrocytes catechol-o-methyl transferase activity is up-regulated after a 3-month treatment by entacapone in parkinsonian patients**. *Clin Neuropharmacol* 2011, **34**(1):21-23.

90. Männistö PT, Kaakkola S: **Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors**. *Pharmacol Rev* 1999, **51**(4):593-628.

91. Arnold MA, Bartholini G, Black IB, Bloom FE, Brownstein MJ, Conolly ME, Jonakait GM, Koob GF, Kopin IJ, Martin JB, Others: **Catecholamines II**, vol. 90: Springer Science & Business Media; 2012.

92. Corvol J-C, Bonnet C, Charbonnier-Beaupel F, Bonnet A-M, Fiévet M-H, Bellanger A, Roze E, Meliksetyan G, Ben Djebara M, Hartmann A, Lacomblez L, Vrignaud C, Zahr N, Agid Y, Costentin J, Hulot J-S, Vidailhet M: **The COMT Val158Met polymorphism affects the response to entacapone in Parkinson's disease: a randomized crossover clinical trial**. *Ann Neurol* 2011, **69**(1):111-118.

93. Rutherford K, Le Trong I, Stenkamp RE, Parson WW: **Crystal structures of human 108V and 108M catechol O-methyltransferase**. *J Mol Biol* 2008, **380**(1):120-130.

94. Palma PN, Bonifácio MJ, Loureiro AI, Wright LC, Learmonth DA, Soares-da-Silva P: **Molecular modeling and metabolic studies of the interaction of catechol-O-methyltransferase and a new nitrocatechol inhibitor**. *Drug Metab Dispos* 2003, **31**(3):250-258.

95. Rutherford K, Alphandéry E, McMillan A, Daggett V, Parson WW: **The V108M mutation decreases the structural stability of catechol O-methyltransferase**. *Biochim Biophys Acta* 2008, **1784**(7-8):1098-1105.

96. Wong CF, Kua J, Zhang Y, Straatsma TP, McCammon JA: **Molecular docking of balanol to dynamics snapshots of protein kinase A**. *Proteins* 2005, **61**(4):850-858.

97. Bolstad ESD, Anderson AC: **In pursuit of virtual lead optimization: pruning ensembles of receptor structures for increased efficiency and accuracy during docking**. *Proteins* 2009, **75**(1):62-74.

98. Paulsen JL, Anderson AC: **ChemInform Abstract: Scoring Ensembles of Docked Protein: Ligand Interactions for Virtual Lead Optimization**. *ChemInform* 2010, **41**(13):no-no.

99. Cheng LS, Amaro RE, Xu D, Li WW, Arzberger PW, McCammon JA: **Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase**. *J Med Chem* 2008, **51**(13):3878-3894.

100. Yoon S, Welsh WJ: **Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring**. *J Chem Inf Comput Sci* 2004, **44**(1):88-96.

101. Miller BR, III, McGee TD, Jr., Swails JM, Homeyer N, Gohlke H, Roitberg AE: **MMPBSA. py: an efficient program for end-state free energy calculations**. *J Chem Theory Comput* 2012, **8**(9):3314-3321.

102. Lotta T, Vidgren J, Tilgmann C, Ulmanen I, Melén K, Julkunen I, Taskinen J: **Kinetics of human soluble and membrane-bound catechol O-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme**. *Biochemistry* 1995, **34**(13):4202-4210.

103. Chen J, Lipska BK, Halim N, Ma QD, Matsumoto M, Melhem S, Kolachana BS, Hyde TM, Herman MM, Apud J, Egan MF, Kleinman JE, Weinberger DR: **Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain**. *Am J Hum Genet* 2004, **75**(5):807-821.

104. Kirkman HN, Gaetani GD, Clemons EH, Mareni C: **Red cell NADP+ and NADPH in glucose-6-phosphate dehydrogenase deficiency**. *J Clin Invest* 1975, **55**(4):875-878.

105. Beutler E: **The molecular biology of G6PD variants and other red cell enzyme defects**. *Annu Rev Med* 1992, **43**:47-59.

106. Mason PJ, Bautista JM, Gilsanz F: **G6PD deficiency: the genotype-phenotype association**. *Blood Rev* 2007, **21**(5):267-283.

107. Wang X-T, Lam V, Engel PC: **Marked decrease in specific activity contributes to disease phenotype in two human glucose 6-phosphate dehydrogenase mutants, G6PDUnion and G6PDAndalus**. *Hum Mutat* 2005, **26**(3):284-284.

108. Notaro R, Afolayan A, Luzzatto L: **Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history**. *FASEB J* 2000, **14**(3):485-494.

109. Kotaka M, Gover S, Vandeputte-Rutten L, Au SWN, Lam VMS, Adams MJ: **Structural studies of glucose-6-phosphate and NADP+ binding to human glucose-6-phosphate dehydrogenase**. *Acta Crystallogr D Biol Crystallogr* 2005, **61**(Pt 5):495-504.

110. Giles NM, Giles GI, Jacob C: **Multiple roles of cysteine in biocatalysis**. *Biochem Biophys Res Commun* 2003, **300**(1):1-4.

111. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protoc* 2009, **4**(7):1073-1081.

112. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.

113. Kisters-Woike B, Vangierdegom C, Müller-Hill B: **On the conservation of protein sequences in evolution**. *Trends Biochem Sci* 2000, **25**(9):419-421.

114. Sirover MA: **New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase**. *Biochim Biophys Acta* 1999, **1432**(2):159-184.

115. Soukri A, Mougin A, Corbier C, Wonacott A, Branlant C, Branlant G: **Role of the histidine 176 residue in glyceraldehyde-3-phosphate dehydrogenase as probed by site-directed mutagenesis**. *Biochemistry* 1989, **28**(6):2586-2592.

116. Cook WJ, Senkovich O, Chattopadhyay D: **An unexpected phosphate binding site in glyceraldehyde 3-phosphate dehydrogenase: crystal structures of apo, holo and ternary complex of Cryptosporidium parvum enzyme**. *BMC Struct Biol* 2009, **9**:9.

117. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**(3):245-248.

118. Bordbar A, Monk JM, King ZA, Palsson BO: **Constraint-based models predict metabolic and associated cellular functions**. *Nat Rev Genet* 2014, **15**(2):107-120.

119. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO: **Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics**. *Cell Systems* 2015, **1**(4):283-292.

120. Jamshidi N, Palsson BØ: **Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models**. *Biophys J* 2010, **98**(2):175-185.

121. Shlomi T, Cabili MN, Ruppin E: **Predicting metabolic biomarkers of human inborn errors of metabolism**. *Mol Syst Biol* 2009, **5**:263.

122. Schellenberger J, Palsson BØ: **Use of randomized sampling for analysis of metabolic networks**. *J Biol Chem* 2009, **284**(9):5457-5461.

123. Kim JS, Kim J-Y, Kim J-M, Kim JW, Chung SJ, Kim SR, Kim MJ, Kim H-T, Choi K-G, Shin D-I, Others: **No correlation between COMT genotype and entacapone benefits in Parkinson's disease**. *Neurology Asia* 2011, **16**(3).

124. Apud JA, Mattay V, Chen J, Kolachana BS, Callicott JH, Rasetti R, Alce G, Iudicello JE, Akbar N, Egan MF, Goldberg TE, Weinberger DR: **Tolcapone improves cognition and cortical information processing in normal human subjects**. *Neuropsychopharmacology* 2007, **32**(5):1011-1020.

125. Chong DJ, Suchowersky O, Szumlanski C, Weinshilboum RM, Brant R, Campbell NR: **The relationship between COMT genotype and the clinical effectiveness of tolcapone, a COMT inhibitor, in patients with Parkinson's disease**. *Clin Neuropharmacol* 2000, **23**(3):143-148.

126. **Glucose-6-phosphate dehydrogenase deficiency. WHO Working Group**. *Bull World Health Organ* 1989, **67**(6):601-611.

127. Lee B, Scaglia F: **Inborn Errors of Metabolism: From Neonatal Screening to Metabolic Pathways**: Oxford University Press; 2014.

128. Pretsch W, Favor J: **Genetic, biochemical, and molecular characterization of nine glyceraldehyde-3-phosphate dehydrogenase mutants with reduced enzyme activity in Mus musculus**. *Mamm Genome* 2007, **18**(10):686-692.

129. Seidler NW: **GAPDH: Biological Properties and Diversity: Biological Properties and Diversity**, vol. 985: Springer Science & Business Media; 2012.

130. Price ND, Schellenberger J, Palsson BO: **Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies**. *Biophys J* 2004, **87**(4):2172-2186.

131. Rutherford K, Daggett V: **Polymorphisms and disease: hotspots of inactivation in methyltransferases**. *Trends Biochem Sci* 2010, **35**(10):531-538.

132. Beck DAC, Jonsson AL, Schaeffer RD, Scott KA, Day R, Toofanny RD, Alonso DOV, Daggett V: **Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations**. *Protein Eng Des Sel* 2008, **21**(6):353-368.

133. van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, Benson NC, Anderson PC, Merkley ED, Rysavy S, Bromley D, Beck DAC, Daggett V: **Dynameomics: a comprehensive database of protein dynamics**. *Structure* 2010, **18**(4):423-435.

134. Zwier MC, Chong LT: **Reaching biological timescales with all-atom molecular dynamics simulations**. *Curr Opin Pharmacol* 2010, **10**(6):745-752.

135. Lavecchia A, Di Giovanni C: **Virtual screening strategies in drug discovery: a critical review**. *Curr Med Chem* 2013, **20**(23):2839-2860.

136. Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, Murcko M, Frye L, Farid R, Lin T, Mobley DL, Jorgensen WL, Berne BJ, Friesner RA, Abel R: **Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field**. *J Am Chem Soc* 2015, **137**(7):2695-2703.

137. Bordbar A, Palsson BØ: **Moving Toward Genome-Scale Kinetic Models: The Mass Action Stoichiometric Simulation Approach**. In: *Functional Coherence of Molecular Networks in Bioinformatics.* Springer New York; 2012: 201-220.

138. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A, Godzik A: **Three-dimensional structural view of the central metabolic network of Thermotoga maritima**. *Science* 2009, **325**(5947):1544-1549.

139. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE: **BiGG Models: A platform for integrating, standardizing and sharing genome-scale models**. *Nucleic Acids Res* 2016, **44**(D1):D515-522.

140. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2005, **33**(Database issue):D54-58.

141. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: **RefSeq: an update on mammalian reference sequences**. *Nucleic Acids Res* 2014, **42**(Database issue):D756-763.

142. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P: **Ensembl BioMarts: a hub for data retrieval across taxonomic space**. *Database* 2011, **2011**:bar030.

143. Menon R, Roy A, Mukherjee S, Belkin S, Zhang Y, Omenn GS: **Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers**. *J Proteome Res* 2011, **10**(12):5503-5511.

144. Jaroszewski L, Pawlowski K, Godzik A: **Multiple Model Approach: Exploring the Limits of Comparative Modeling**. *J Mol Med*, **4**(10):294-309.

145. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **ıt PROCHECK: a program to check the stereochemical quality of protein structures**. *J Appl Crystallogr* 1993, **26**(2):283-291.

146. Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, Bairoch A: **Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase**. *Hum Mutat* 2008, **29**(3):361-366.

147. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res* 2012, **40**(Database issue):D1100-1107.

148. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE: **Pharmacogenomics knowledge for personalized medicine**. *Clinical Pharmacology & Therapeutics* 2012, **92**(4):414-417.

149. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009**. *Nucleic Acids Res* 2009, **37**(Database issue):D588-592.

150. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK: **BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities**. *Nucleic Acids Res* 2007, **35**(suppl 1):D198-D201.

151. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**(4):264-276.

152. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, Rizzo RC, Case DA, James TL, Kuntz ID: **DOCK 6: Combining techniques to model RNA–small molecule complexes**. *RNA* 2009, **15**(6):1219-1230.

153. Case DA, Babin V, Berryman J, Betz RM, Cai Q, Cerutti DS, Cheatham TEI, Darden TA, Duke RE, Gohlke H, Others: **Amber 14**. 2014.

154. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Others: **01; Gaussian, Inc**. *Wallingford, CT* 2004.

155. Martins SA, Sousa SF, Ramos MJ, Fernandes PA: **Prediction of Solvation Free Energies with Thermodynamic Integration Using the General Amber Force Field**. *J Chem Theory Comput* 2014, **10**(8):3570-3577.

156. Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási AL: **Global organization of metabolic fluxes in the bacterium Escherichia coli**. *Nature* 2004, **427**(6977):839-843.

157. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**(1):93-121.

158. Thiele I, Jamshidi N, Fleming RMT, Palsson BØ: **Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization**. *PLoS Comput Biol* 2009, **5**(3):e1000312.

159. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ: **Reconstruction of biochemical networks in microorganisms**. *Nat Rev Microbiol* 2008, **7**(2):129-143.

160. Barrett CL, Herring CD, Reed JL, Palsson BO: **The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states**. *Proc Natl Acad Sci U S A* 2005, **102**(52):19103-19108.

161. Schellenberger J, Park JO, Conrad TM, Palsson BØ: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions**. *BMC Bioinformatics* 2010, **11**(1):213.

162. Guzmán GI, Utrilla J, Nurk S, Brunk E, Monk JM, Ebrahim A, Palsson BO, Feist AM: **Model-driven discovery of underground metabolic functions in Escherichia coli**. *Proceedings of the National Academy of Sciences* 2015, **112**(3):929-934.

163. Monk J, Palsson BO: **Predicting microbial growth**. *Science* 2014, **344**(6191):1448-1449.

164. Jain R, Srivastava R: **Metabolic investigation of host/pathogen interaction using MS2-infected Escherichia coli**. *BMC Syst Biol* 2009, **3**(1):121.

165. Hanly TJ, Henson MA: **Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures**. *Biotechnol Bioeng* 2011, **108**(2):376-385.

166. Tzamali E, Poirazi P, Tollis IG, Reczko M: **A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities**. *BMC Syst Biol* 2011, **5**(1):167.

167. Wintermute EH, Silver PA: **Emergent cooperation in microbial metabolism**. *Mol Syst Biol* 2010, **6**(1).

168. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, Bonilla G, Kar A, Leiby N, Mehta P, Others: **Metabolic resource allocation in individual microbes**

**determines ecosystem interactions and spatial dynamics**. *Cell Rep* 2014, **7**(4):1104-1115.

169. Klitgord N, Segrè D: **Environments that induce synthetic microbial ecosystems**. *PLoS Comput Biol* 2010, **6**(11):e1001002.

170. Kugler H, Larjo A, Harel D: **Biocharts: a visual formalism for complex biological systems**. *J R Soc Interface* 2010, **7**(48):1015-1024.

171. Aloy P, Russell RB: **Structural systems biology: modelling protein interactions**. *Nat Rev Mol Cell Biol* 2006, **7**(3):188-197.

172. Betts MJ, Russell RB: **The hard cell: from proteomics to a whole cell model**. *FEBS Lett* 2007, **581**(15):2870-2876.

173. Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Others: **Proteome organization in a genome-reduced bacterium**. *Science* 2009, **326**(5957):1235-1240.

174. Kortemme T, Baker D: **Computational design of protein--protein interactions**. *Curr Opin Chem Biol* 2004, **8**(1):91-97.

175. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Others: **Structure-based prediction of protein-protein interactions on a genome-wide scale**. *Nature* 2012, **490**(7421):556-560.

176. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease**. *Nat Biotechnol* 2012, **30**(2):159-164.

177. Cheng TMK, Goehring L, Jeffery L, Lu Y-E, Hayles J, Novák B, Bates PA: **A structural systems biology approach for quantifying the systemic consequences of missense mutations in proteins**. *PLoS Comput Biol* 2012, **8**(10):e1002738.

178. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M, Quinn GB, Ramos AG, Westbrook JD, Young J, Zardecki C, Berman HM, Bourne PE: **The RCSB Protein Data Bank: new resources for research and education**. *Nucleic Acids Res* 2013, **41**(Database issue):D475-482.

179. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism?2011**. *Mol Syst Biol* 2011, **7**(1).

180. Wu S, Skolnick J, Zhang Y: **Ab initio modeling of small proteins by iterative TASSER simulations**. *BMC Biol* 2007, **5**(1):17.

181. Zhang Y: **I-TASSER: Fully automated protein structure prediction in CASP8**. *Proteins: Struct Funct Bioinf* 2009, **77**(S9):100-113.

182. Battey JND, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T: **Automated server predictions in CASP7**. *Proteins: Struct Funct Bioinf* 2007, **69**(S8):68-82.

183. Zhang Y: **Template-based modeling and free modeling by I-TASSER in CASP7**. *Proteins: Struct Funct Bioinf* 2007, **69**(S8):108-117.

184. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: **Evaluation of template-based models in CASP8 with standard measures**. *Proteins: Struct Funct Bioinf* 2009, **77**(S9):18-28.

185. Xu D, Zhang Y: **Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment**. *Sci Rep* 2013, **3**.

186. Zhou H, Gao M, Kumar N, Skolnick J: **SUNPRO: Structure and function predictions of proteins from representative organisms**. 2012.

187. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures**. *J Appl Crystallogr* 1993, **26**(2):283-291.

188. Godzik A, Koliński A, Skolnick J: **Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets**. *Protein Sci* 1995, **4**(10):2107-2117.

189. Mander L, Liu H-W: **Comprehensive Natural Products II: Chemistry and Biology**: Elsevier; 2010.

190. Hirotsu K, Goto M, Okamoto A, Miyahara I: **Dual substrate recognition of aminotransferases**. *The Chemical Record* 2005, **5**(3):160-172.

191. Steffen-Munsberg F, Vickers C, Thontowi A, Schätzle S, Meinhardt T, Svedendahl Humble M, Land H, Berglund P, Bornscheuer UT, Höhne M: **Revealing the structural basis of promiscuous amine transaminase activity**. *ChemCatChem* 2013, **5**(1):154-157.

192. Saito M, Takemura N, Shirai T: **Classification of ligand molecules in PDB with fast heuristic graph match algorithm COMPLIG**. *J Mol Biol* 2012, **424**(5):379-390.

193. **RCSB PDB - Drug To PDB IDs Mappings** [http://www.pdb.org/pdb/ligand/drugMapping.do]

194. Godden JW, Xue L, Bajorath J: **Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients**. *J Chem Inf Comput Sci* 2000, **40**(1):163-166.

195. Li G-W, Burkhardt D, Gross C, Weissman JS: **Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources**. *Cell* 2014, **157**(3):624-635.

196. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli**. *Nucleic Acids Res* 2005, **33**(suppl 1):D334-D337.

197. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, López-Bigas N: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes**. *Nucleic Acids Res* 2005, **33**(19):6083-6089.

198. Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, Chen Y, Wu CH: **A comprehensive protein-centric ID mapping service for molecular data integration**. *Bioinformatics* 2011, **27**(8):1190-1191.

199. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction**. *Mol Syst Biol* 2013, **9**(1).

200. Levy ED, Teichmann SA: **Structural, Evolutionary, and Assembly Principles of Protein Oligomerization**. In: *Oligomerization in Health and Disease.* vol. 117: Elsevier; 2013: 25-51.

201. Latif H, Szubin R, Tan J, Brunk E, Lechner A, Zengler K, Palsson BO: **A streamlined ribosome profiling protocol for the characterization of microorganisms**. *Biotechniques* 2015, **accepted**(-):-.

202. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling**. *Science* 2009, **324**(5924):218-223.

203. Blackstock WP, Weir MP: **Proteomics: quantitative and physical mapping of cellular proteins**. *Trends Biotechnol* 1999, **17**(3):121-127.

204. Cox J, Mann M: **Quantitative, high-resolution proteomics for data-driven systems biology**. *Annu Rev Biochem* 2011, **80**:273-299.

205. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**(3):727-738.

206. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR: **COBRApy: COnstraints-Based Reconstruction and Analysis for Python**. *BMC Syst Biol* 2013, **7**(1):74.

207. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science* 1995, **269**(5223):496-512.

208. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae**. *Nucleic Acids Res* 1996, **24**(22):4420-4449.

209. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of Escherichia coli K-12**. *Science* 1997, **277**(5331):1453-1462.

210. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessières P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Cordani JJ, Connerton IF, Cummings NJ, Daniel RA, Denziot F, Devine KM, Düsterhöft A, Ehrlich SD, Emmerson PT, Entian KD, Errington J, Fabret C, Ferrari E, Foulger D, Fritz C, Fujita M, Fujita Y, Fuma S, Galizzi A, Galleron N, Ghim SY, Glaser P, Goffeau A, Golightly EJ, Grandi G, Guiseppi G, Guy BJ, Haga K, Haiech J, Harwood CR, Hènaut A, Hilbert H, Holsappel S, Hosono S, Hullo MF, Itaya M, Jones L, Joris B, Karamata D, Kasahara Y, Klaerr-Blanchard M, Klein C, Kobayashi Y, Koetter P, Koningstein G, Krogh S, Kumano M, Kurita K, Lapidus A, Lardinois S, Lauber J, Lazarevic V, Lee SM, Levine A, Liu H, Masuda S, Mauël C, Médigue C, Medina N, Mellado RP, Mizuno M, Moestl D, Nakai S, Noback M, Noone D, O'Reilly M, Ogawa K, Ogiwara A, Oudega B, Park SH, Parro V, Pohl TM, Portelle D, Porwollik S, Prescott AM, Presecan E, Pujic P, Purnelle B, Rapoport G, Rey M, Reynolds S, Rieger M, Rivolta C, Rocha E, Roche B, Rose M, Sadaie Y, Sato T, Scanlan E, Schleich S, Schroeter R, Scoffone F, Sekiguchi J, Sekowska A, Seror SJ, Serror P, Shin BS, Soldo B, Sorokin A, Tacconi E, Takagi T, Takahashi H, Takemaru K, Takeuchi M, Tamakoshi A, Tanaka T, Terpstra P, Togoni A, Tosato V, Uchiyama S, Vandebol M, Vannier F, Vassarotti A, Viari A, Wambutt R, Wedler H, Weitzenegger T, Winters P, Wipat A, Yamamoto H, Yamane K, Yasumoto K, Yata K, Yoshida K, Yoshikawa HF, Zumstein E, Yoshikawa H, Danchin A: **The complete genome sequence of the gram-positive bacterium Bacillus subtilis**. *Nature* 1997, **390**(6657):249-256.

211. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV: **The complete genome of the hyperthermophilic bacterium Aquifex aeolicus**. *Nature* 1998, **392**(6674):353-358.

212. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C, Cotton MD, Fujii C, Garland S, Hatch B, Horst K, Roberts K, Sandusky M, Weidman J, Smith HO, Venter JC: **Complete genome sequence of Treponema pallidum, the syphilis spirochete**. *Science* 1998, **281**(5375):375-388.

213. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima**. *Nature* 1999, **399**(6734):323-329.

214. Logsdon JM, Jr., Faguy DM: **Evolutionary genomics: Thermotoga heats up lateral gene transfer**. *Curr Biol* 1999, **9**(19):R747-R751.

215. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G: **A database of protein structure families with common folding motifs**. *Protein Sci* 1992, **1**(12):1691-1698.

216. Nasir A, Kim KM, Caetano-Anollés G: **Global patterns of protein domain gain and loss in superkingdoms**. *PLoS Comput Biol* 2014, **10**(1):e1003452.

217. Prlic A, Bliven S, Rose PW, Bluhm WF, Bizon C, Godzik A, Bourne PE: **Pre-calculated protein structure alignments at the RCSB PDB website**. *Bioinformatics* 2010, **26**(23):2983-2985.

218. von Heijne G: **Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule**. *J Mol Biol* 1992, **225**(2):487-494.

219. Jones DT, Taylor WR, Thornton JM: **A model recognition approach to the prediction of all-helical membrane protein structure and topology**. *Biochemistry* 1994, **33**(10):3038-3049.

220. Murphy KP, Freire E: **Structural energetics of protein stability and folding cooperativity**. *J Macromol Sci Part A Pure Appl Chem* 1993, **65**(9):1939-1946.

221. Wu I, Arnold FH: **Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures**. *Biotechnol Bioeng* 2013, **110**(7):1874-1883.

222. Sawle L, Ghosh K: **How do thermophilic proteins and proteomes withstand high temperature?** *Biophys J* 2011, **101**(1):217-227.

223. Das R, Gerstein M: **The stability of thermophilic proteins: a study based on comprehensive genome comparison**. *Funct Integr Genomics* 2000, **1**(1):76-88.

224. Robinson-Rechavi M, Godzik A: **Structural genomics of thermotoga maritima proteins shows that contact order is a major determinant of protein thermostability**. *Structure* 2005, **13**(6):857-860.

225. Auerbach G, Huber R, Grättinger M, Zaiss K, Schurig H, Jaenicke R, Jacob U: **Closed structure of phosphoglycerate kinase from Thermotoga maritima reveals the catalytic mechanism and determinants of thermal stability**. *Structure* 1997, **5**(11):1475-1483.

226. Beaucamp N, Ostendorp R, Schurig H, Jaenicke R: **Cloning, sequencing, expression and characterization of the gene encoding the 3-phosphoglycerate kinase- triosephosphate isomerase fusion protein from Thermotoga maritima**. *Protein Pept Lett* 1995, **2**(1):281-286.

227. Bi Y, Watts JC, Bamford PK, Briere L-AK, Dunn SD: **Probing the functional tolerance of the b subunit of Escherichia coli ATP synthase for sequence manipulation through a chimera approach**. *Biochim Biophys Acta* 2008, **1777**(7-8):583-591.

228. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, Lázár V, Spohn R, Csörg\Ho B, Wagner A, Others: **Network-level architecture and the evolutionary potential of underground metabolism**. *Proceedings of the National Academy of Sciences* 2014, **111**(32):11762-11767.

229. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO: **Network context and selection in the evolution to enzyme specificity**. *Science* 2012, **337**(6098):1101-1104.

230. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: **Systems approach to refining genome annotation**. *Proc Natl Acad Sci U S A* 2006, **103**(46):17480-17484.

231. Orth JD, Palsson B: **Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions**. *BMC Syst Biol* 2012, **6**:30.

232. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds**. *Nature* 1994, **372**(6507):631-634.

233. Orengo CA, Flores TP, Jones DT, Taylor WR, Thornton JM: **Recurring structural motifs in proteins with different functions**. *Curr Biol* 1993, **3**(3):131-139.

234. Yoshikuni Y, Ferrin TE, Keasling JD: **Designed divergent evolution of enzyme function**. *Nature* 2006, **440**(7087):1078-1082.

235. Lee S-M, Jellison T, Alper HS: **Directed evolution of xylose isomerase for improved xylose catabolism and fermentation in the yeast Saccharomyces cerevisiae**. *Appl Environ Microbiol* 2012, **78**(16):5708-5716.

236. Bar-Even A, Tawfik DS: **Engineering specialized metabolic pathways-is there a room for enzyme improvements?** *Curr Opin Biotechnol* 2013, **24**(2):310-319.

237. Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anollés G: **History of biological metal utilization inferred through phylogenomic analysis of protein structures**. *Proceedings of the National Academy of Sciences* 2010, **107**(23):10567-10572.

238. Caetano-Anollés G, Caetano-Anollés D: **An evolutionarily structured universe of protein architecture**. *Genome Res* 2003, **13**(7):1563-1571.

239. Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal J: **The origin, evolution and structure of the protein world**. 2009, **417**:621-637.

240. Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE: **The origin and evolution of modern metabolism**. *Int J Biochem Cell Biol* 2009, **41**(2):285-297.

241. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009**. *Nucleic Acids Res* 2009, **37**(suppl 1):D588-D592.

242. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D: **BRENDA, the enzyme database: updates and major new developments**. *Nucleic Acids Res* 2004, **32**(suppl 1):D431-D433.

243. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Others: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.

244. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Others: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**(suppl 1):D138-D141.

245. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**(4):536-540.

246. Bakan A, Meireles LM, Bahar I: **ProDy: protein dynamics inferred from theory and experiments**. *Bioinformatics* 2011, **27**(11):1575-1577.

247. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics**. *J Mol Graph* 1996, **14**(1):33-38.

248. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Others: **The Pfam protein families database**. *Nucleic Acids Res* 2009:gkp985.

249. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Others: **Pfam: the protein families database**. *Nucleic Acids Res* 2013:gkt1223.

250. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. In: *Genome Inform: 2009*. 2009: 205-211.

251. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions**. *Nucleic Acids Res* 2013, **41**(12):e121-e121.

252. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A: **ProTherm, version 4.0: thermodynamic database for proteins and mutants**. *Nucleic Acids Res* 2004, **32**(suppl 1):D120-D121.

253. Ku T, Lu P, Chan C, Wang T, Lai S, Lyu P, Hsiao N: **Predicting melting temperature directly from protein sequences**. *Comput Biol Chem* 2009, **33**(6):445-450.

254. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, Others: **In silico method for modelling metabolism and gene product expression at genome scale**. *Nat Commun* 2012, **3**:929.

255. Chang RL, Xie L, Bourne PE, Palsson BO: **Antibacterial mechanisms identified through structural systems pharmacology**. *BMC Syst Biol* 2013, **7**(1):102.

256. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state**. *J Mol Biol* 2007, **372**(3):774-797.

257. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility**. *J Mol Biol* 1971, **55**(3):379-IN374.

258. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**(12):2577-2637.

259. Cabiscol E, Tamarit J, Ros J: **Oxidative stress in bacteria and protein damage by reactive oxygen species**. *Int Microbiol* 2000, **3**(1):3-8.

260. Hernández-García D, Wood CD, Castro-Obregón S, Covarrubias L: **Reactive oxygen species: A radical role in development?** *Free Radic Biol Med* 2010, **49**(2):130-143.

261. Naviaux RK: **Oxidative shielding or oxidative stress?** *J Pharmacol Exp Ther* 2012, **342**(3):608-618.

262. Jones RM, Luo L, Ardita CS, Richardson AN, Kwon YM, Mercante JW, Alam A, Gates CL, Wu H, Swanson PA, Lambeth JD, Denning PW, Neish AS: **Symbiotic lactobacilli stimulate gut epithelial proliferation via Nox-mediated generation of reactive oxygen species**. *EMBO J* 2013, **32**(23):3017-3028.

263. Rea WJ, Patel KD: **Reversibility of Chronic Disease and Hypersensitivity, Volume 4: The Environmental Aspects of Chemical Sensitivity**: CRC Press; 2017.

264. **Oxidative Stress and Redox Regulation**; 2013.

265. Ghosh K, de Graff AMR, Sawle L, Dill KA: **Role of Proteome Physical Chemistry in Cell Behavior**. *J Phys Chem B* 2016, **120**(36):9549-9563.

266. Ezraty B, Gennaris A, Barras F, Collet J-F: **Oxidative stress, protein damage and repair in bacteria**. *Nat Rev Microbiol* 2017, **15**(7):385-396.

267. Rao RSP, Møller IM: **Pattern of occurrence and occupancy of carbonylation sites in proteins**. *Proteomics* 2011, **11**(21):4166-4173.

268. Maisonneuve E, Ducret A, Khoueiry P, Lignon S, Longhi S, Talla E, Dukan S: **Rules governing selective protein carbonylation**. *PLoS One* 2009, **4**(10):e7269.

269. Grimsrud PA, Xie H, Griffin TJ, Bernlohr DA: **Oxidative stress and covalent modification of protein with bioactive aldehydes**. *J Biol Chem* 2008, **283**(32):21837-21841.

270. Vieira-Silva S, Rocha EPC: **An assessment of the impacts of molecular oxygen on the evolution of proteomes**. *Mol Biol Evol* 2008, **25**(9):1931-1942.

271. Oliver CN, Ahn BW, Moerman EJ, Goldstein S, Stadtman ER: **Age-related changes in oxidized proteins**. *J Biol Chem* 1987, **262**(12):5488-5491.

272. Andziak B, O'Connor TP, Qi W, DeWaal EM, Pierce A, Chaudhuri AR, Van Remmen H, Buffenstein R: **High oxidative damage levels in the longest-living rodent, the naked mole-rat**. *Aging Cell* 2006, **5**(6):463-471.

273. Pérez VI, Buffenstein R, Masamsetti V, Leonard S, Salmon AB, Mele J, Andziak B, Yang T, Edrey Y, Friguet B, Ward W, Richardson A, Chaudhuri A: **Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat**. *Proc Natl Acad Sci U S A* 2009, **106**(9):3059-3064.

274. Bokov A, Chaudhuri A, Richardson A: **The role of oxidative damage and stress in aging**. *Mech Ageing Dev* 2004, **125**(10-11):811-826.

275. Schindeldecker M, Moosmann B: **Protein-borne methionine residues as structural antioxidants in mitochondria**. *Amino Acids* 2015, **47**(7):1421-1432.

276. Levine RL, Mosoni L, Berlett BS, Stadtman ER: **Methionine residues as endogenous antioxidants in proteins**. *Proc Natl Acad Sci U S A* 1996, **93**(26):15036-15040.

277. Luo S, Levine RL: **Methionine in proteins defends against oxidative stress**. *FASEB J* 2009, **23**(2):464-472.

278. St John G, Brot N, Ruan J, Erdjument-Bromage H, Tempst P, Weissbach H, Nathan C: **Peptide methionine sulfoxide reductase from Escherichia coli and Mycobacterium tuberculosis protects bacteria against oxidative damage from reactive nitrogen intermediates**. *Proc Natl Acad Sci U S A* 2001, **98**(17):9901-9906.

279. Requejo R, Hurd TR, Costa NJ, Murphy MP: **Cysteine residues exposed on protein surfaces are the dominant intramitochondrial thiol and may protect against oxidative damage**. *FEBS J* 2010, **277**(6):1465-1480.

280. Moosmann B, Behl C: **Cytoprotective antioxidant function of tyrosine and tryptophan residues in transmembrane proteins**. *Eur J Biochem* 2000, **267**(18):5687-5692.

281. Moosmann B: **Respiratory chain cysteine and methionine usage indicate a causal role for thiyl radicals in aging**. *Exp Gerontol* 2011, **46**(2-3):164-169.

282. Krisko A, Radman M: **Phenotypic and genetic consequences of protein damage**. *PLoS Genet* 2013, **9**(9):e1003810.

283. de Graff AMR, Hazoglou MJ, Dill KA: **Highly Charged Proteins: The Achilles' Heel of Aging Proteomes**. *Structure* 2016, **24**(2):329-336.

284. Vidovic A, Supek F, Nikolic A, Krisko A: **Signatures of conformational stability and oxidation resistance in proteomes of pathogenic bacteria**. *Cell Rep* 2014, **7**(5):1393-1400.

285. Panda A, Ghosh TC: **Prevalent structural disorder carries signature of prokaryotic adaptation to oxic atmosphere**. *Gene* 2014, **548**(1):134-141.

286. Girod M, Enjalbert Q, Brunet C, Antoine R, Lemoine J, Lukac I, Radman M, Krisko A, Dugourd P: **Structural basis of protein oxidation resistance: a lysozyme study**. *PLoS One* 2014, **9**(7):e101642.

287. Pieulle L, Magro V, Hatchikian EC: **Isolation and analysis of the gene encoding the pyruvate-ferredoxin oxidoreductase of Desulfovibrio africanus, production of the recombinant enzyme in Escherichia coli, and effect of carboxy-terminal deletions on its stability**. *J Bacteriol* 1997, **179**(18):5684-5692.

288. Lu Z, Imlay JA: **The Fumarate Reductase of Bacteroides thetaiotaomicron, unlike That of Escherichia coli, Is Configured so that It Does Not Generate Reactive Oxygen Species**. *MBio* 2017, **8**(1).

289. Stocker R, Keaney JF, Jr.: **New insights on oxidative stress in the artery wall**. *J Thromb Haemost* 2005, **3**(8):1825-1834.

290. Valasatava Y, Rosato A, Furnham N, Thornton JM, Andreini C: **To what extent do structural changes in catalytic metal sites affect enzyme function?** *J Inorg Biochem* 2018, **179**:40-53.

291. Jang HH, Lee KO, Chi YH, Jung BG, Park SK, Park JH, Lee JR, Lee SS, Moon JC, Yun JW, Choi YO, Kim WY, Kang JS, Cheong G-W, Yun D-J, Rhee SG, Cho MJ, Lee SY: **Two enzymes in one; two yeast peroxiredoxins display oxidative stress-dependent switching from a peroxidase to a molecular chaperone function**. *Cell* 2004, **117**(5):625-635.

292. Ritz D, Lim J, Reynolds CM, Poole LB, Beckwith J: **Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion**. *Science* 2001, **294**(5540):158-160.

293. Chabrière E, Charon MH, Volbeda A, Pieulle L, Hatchikian EC, Fontecilla-Camps JC: **Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate**. *Nat Struct Biol* 1999, **6**(2):182-190.

294. Wan B, Zhang Q, Ni J, Li S, Wen D, Li J, Xiao H, He P, Ou H-Y, Tao J, Teng Q, Lu J, Wu W, Yao Y-F: **Type VI secretion system contributes to Enterohemorrhagic**

**Escherichia coli virulence by secreting catalase against host reactive oxygen species (ROS)**. *PLoS Pathog* 2017, **13**(3):e1006246.

295. Tsuchiya Y, Peak-Chew SY, Newell C, Miller-Aidoo S, Mangal S, Zhyvoloup A, Bakovic J, Malanchuk O, Pereira GC, Kotiadis V, Szabadkai G, Duchen MR, Campbell M, Cuenca SR, Vidal-Puig A, James AM, Murphy MP, Filonenko V, Skehel M, Gout I: **Protein CoAlation: a redox-regulated protein modification by coenzyme A in mammalian cells**. *Biochem J* 2017, **474**(14):2489-2508.

296. Brynildsen MP, Winkler JA, Spina CS, MacDonald IC, Collins JJ: **Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production**. *Nat Biotechnol* 2013, **31**(2):160-165.

297. Andreini C, Cavallaro G, Lorenzini S, Rosato A: **MetalPDB: a database of metal sites in biological macromolecular structures**. *Nucleic Acids Res* 2013, **41**(Database issue):D312-319.

298. Rao RSP, Zhang N, Xu D, Møller IM: **CarbonylDB: A curated data-resource of protein carbonylation sites**. *Bioinformatics* 2018.

299. Sun M-A, Wang Y, Cheng H, Zhang Q, Ge W, Guo D: **RedoxDB—a curated database for experimentally verified protein oxidative modification**. *Bioinformatics* 2012, **28**(19):2551-2552.

300. Brunk E, Mih N, Monk J, Zhang Z, O'Brien EJ, Bliven SE, Chen K, Chang RL, Bourne PE, Palsson BO: **Systems biology of the structural proteome**. *BMC Syst Biol* 2016, **10**:26.

301. Mih N, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, Monk JM, Zhang Z, Palsson BO: **ssbio: A Python Framework for Structural Systems Biology**. *Bioinformatics* 2018.

302. Imlay JA, Fridovich I: **Assay of metabolic superoxide production in Escherichia coli**. *J Biol Chem* 1991, **266**(11):6957-6965.

303. Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT: **The evolution of metabolic networks of E. coli**. *BMC Syst Biol* 2011, **5**:182.

304. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BØ: **Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments**. *Proc Natl Acad Sci U S A* 2013, **110**(50):20338-20343.

305. Galardini M, Koumoutsi A, Herrera-Dominguez L, Cordero Varela JA, Telzerow A, Wagih O, Wartel M, Clermont O, Denamur E, Typas A, Beltrao P: **Phenotype inference in anEscherichia colistrain panel**. *Elife* 2017, **6**.

306. Iobbi-Nivol C, Leimkühler S: **Molybdenum enzymes, their maturation and molybdenum cofactor biosynthesis in Escherichia coli**. *Biochim Biophys Acta* 2013, **1827**(8-9):1086-1101.

307. Hille R: **The Mononuclear Molybdenum Enzymes**. *Chem Rev* 1996, **96**(7):2757-2816.

308. Fischer M, Thöny B, Leimkühler S: **7.17 - The Biosynthesis of Folate and Pterins and Their Enzymology**. In: *Comprehensive Natural Products II*. Oxford: Elsevier; 2010: 599-648.

309. Ezraty B, Bos J, Barras F, Aussel L: **Methionine sulfoxide reduction and assimilation in Escherichia coli: new role for the biotin sulfoxide reductase BisC**. *J Bacteriol* 2005, **187**(1):231-237.

310. Ansaldi M, Théraulaz L, Baraquet C, Panis G, Méjean V: **Aerobic TMAO respiration in Escherichia coli**. *Mol Microbiol* 2007, **66**(2):484-494.

311. Méjean V, Iobbi-Nivol C, Lepelletier M, Giordano G, Chippaux M, Pascal MC: **TMAO anaerobic respiration in Escherichia coli: involvement of the tor operon**. *Mol Microbiol* 1994, **11**(6):1169-1179.

312. del Campillo Campbell A, Campbell A: **Alternative gene for biotin sulfoxide reduction inEscherichia coli K-12**. *J Mol Evol* 1996, **42**(2):85-90.

313. Xi H, Schneider BL, Reitzer L: **Purine catabolism in Escherichia coli and function of xanthine dehydrogenase in purine salvage**. *J Bacteriol* 2000, **182**(19):5332-5341.

314. Belenky P, Ye JD, Porter CBM, Cohen NR, Lobritz MA, Ferrante T, Jain S, Korry BJ, Schwarz EG, Walker GC, Collins JJ: **Bactericidal Antibiotics Induce Toxic Metabolic Perturbations that Lead to Cellular Damage**. *Cell Rep* 2015, **13**(5):968-980.

315. Conover MS, Hadjifrangiskou M, Palermo JJ, Hibbing ME, Dodson KW, Hultgren SJ: **Metabolic Requirements of Escherichia coli in Intracellular Bacterial Communities during Urinary Tract Infection Pathogenesis**. *MBio* 2016, **7**(2):e00104-00116.

316. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T: **SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information**. *Nucleic Acids Res* 2014, **42**(Web Server issue):W252-258.

317. Cookson AL, Cooley WA, Woodward MJ: **The role of type 1 and curli fimbriae of Shiga toxin-producing Escherichia coli in adherence to abiotic surfaces**. *Int J Med Microbiol* 2002, **292**(3-4):195-205.

318. Shaikh N, Holt NJ, Johnson JR, Tarr PI: **Fim operon variation in the emergence of Enterohemorrhagic Escherichia coli: an evolutionary and functional analysis**. *FEMS Microbiol Lett* 2007, **273**(1):58-63.

319. Korea C-G, Badouraly R, Prevost M-C, Ghigo J-M, Beloin C: **Escherichia coli K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities**. *Environ Microbiol* 2010, **12**(7):1957-1977.

320. Hull R, Rudy D, Donovan W, Svanborg C, Wieser I, Stewart C, Darouiche R: **Urinary tract infection prophylaxis using Escherichia coli 83972 in spinal cord injured patients**. *J Urol* 2000, **163**(3):872-877.

321. Puorger C, Vetsch M, Wider G, Glockshuber R: **Structure, folding and stability of FimA, the main structural subunit of type 1 pili from uropathogenic Escherichia coli strains**. *J Mol Biol* 2011, **412**(3):520-535.

322. Smith SW, Latta LCt, Denver DR, Estes S: **Endogenous ROS levels in C. elegans under exogenous stress support revision of oxidative stress theory of life-history tradeoffs**. *BMC Evol Biol* 2014, **14**:161.

323. Aubron C, Glodt J, Matar C, Huet O, Borderie D, Dobrindt U, Duranteau J, Denamur E, Conti M, Bouvet O: **Variation in endogenous oxidative stress in Escherichia coli natural isolates during growth in urine**. *BMC Microbiol* 2012, **12**:120.

324. Levine ZA, Larini L, LaPointe NE, Feinstein SC, Shea J-E: **Regulation and aggregation of intrinsically disordered peptides**. *Proc Natl Acad Sci U S A* 2015, **112**(9):2758-2763.

325. Souza JM, Chen Q, Blanchard-Fillion B, Lorch SA, Hertkorn C, Lightfoot R, Weisse M, Friel T, Paxinou E, Themistocleous M, Chov S, Ischiropoulos H: **Reactive Nitrogen Species and Proteins: Biological Significance and Clinical Relevance**. In: *Biological Reactive Intermediates VI: Chemical and Biological Mechanisms in Susceptibility to and Prevention of Environmental Diseases*. Edited by Dansette PM, Snyder R, Delaforge M, Gibson GG, Greim H, Jollow DJ, Monks TJ, Sipes IG. Boston, MA: Springer US; 2001: 169-174.

326. Crespo MD, Puorger C, Schärer MA, Eidam O, Grütter MG, Capitani G, Glockshuber R: **Quality control of disulfide bond formation in pilus subunits by the chaperone FimC**. *Nat Chem Biol* 2012, **8**(8):707-713.

327. Floyd KA, Moore JL, Eberly AR, Good JAD, Shaffer CL, Zaver H, Almqvist F, Skaar EP, Caprioli RM, Hadjifrangiskou M: **Adhesive fiber stratification in uropathogenic Escherichia coli biofilms unveils oxygen-mediated control of type 1 pili**. *PLoS Pathog* 2015, **11**(3):e1004697.

328. Palmela C, Chevarin C, Xu Z, Torres J, Sevrin G, Hirten R, Barnich N, Ng SC, Colombel J-F: **Adherent-invasive Escherichia coli in inflammatory bowel disease**. *Gut* 2017.

329. Bringer M-A, Glasser A-L, Tung C-H, Méresse S, Darfeuille-Michaud A: **The Crohn's disease-associated adherent-invasive Escherichia coli strain LF82 replicates in mature phagolysosomes within J774 macrophages**. *Cell Microbiol* 2006, **8**(3):471-484.

330. Martinez-Medina M, Mora A, Blanco M, López C, Alonso MP, Bonacorsi S, Nicolas-Chanoine M-H, Darfeuille-Michaud A, Garcia-Gil J, Blanco J: **Similarity and divergence among adherent-invasive Escherichia coli and extraintestinal pathogenic E. coli strains**. *J Clin Microbiol* 2009, **47**(12):3968-3979.

331. Bouckaert J, Mackenzie J, de Paz JL, Chipwaza B, Choudhury D, Zavialov A, Mannerstedt K, Anderson J, Piérard D, Wyns L, Seeberger PH, Oscarson S, De Greve H, Knight SD: **The affinity of the FimH fimbrial adhesin is receptor-driven and quasi-independent of Escherichia coli pathotypes**. *Mol Microbiol* 2006, **61**(6):1556-1568.

332. Dreux N, Denizot J, Martinez-Medina M, Mellmann A, Billig M, Kisiela D, Chattopadhyay S, Sokurenko E, Neut C, Gower-Rousseau C, Colombel J-F, Bonnet R, Darfeuille-Michaud A, Barnich N: **Point mutations in FimH adhesin of Crohn's disease-associated adherent-invasive Escherichia coli enhance intestinal inflammatory response**. *PLoS Pathog* 2013, **9**(1):e1003141.

333. Roos V, Ulett GC, Schembri MA, Klemm P: **The asymptomatic bacteriuria Escherichia coli strain 83972 outcompetes uropathogenic E. coli strains in human urine**. *Infect Immun* 2006, **74**(1):615-624.

334. Kurutas EB, Ciragil P, Gul M, Kilinc M: **The effects of oxidative stress in urinary tract infection**. *Mediators Inflamm* 2005, **2005**(4):242-244.

335. Hryckowian AJ, Welch RA: **RpoS contributes to phagocyte oxidase-mediated stress resistance during urinary tract infection by Escherichia coli CFT073**. *MBio* 2013, **4**(1):e00023-00013.

336. Yang L, Mih N, Yurkovich JT, Park JH, Seo S, Kim D, Monk JM, Lloyd CJ, Tan J, Gao Y, Broddrick JT, Chen K, Heckmann D, Feist AM, Palsson BO: **Multi-scale model of the proteomic and metabolic consequences of reactive oxygen species**. *bioRxiv* 2017.

337. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI: **OPM: orientations of proteins in membranes database**. *Bioinformatics* 2006, **22**(5):623-625.

338. Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM: **Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale**. *BMC Syst Biol* 2014, **8**:110.

339. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M: **The quantitative and condition-dependent Escherichia coli proteome**. *Nat Biotechnol* 2016, **34**(1):104-110.

340. Horler RSP, Butcher A, Papangelopoulos N, Ashton PD, Thomas GH: **EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations**. *Bioinformatics* 2009, **25**(2):163-166.

341. Galperin MY, Makarova KS, Wolf YI, Koonin EV: **Expanded microbial genome coverage and improved protein family annotation in the COG database**. *Nucleic Acids Res* 2015, **43**(Database issue):D261-269.

342. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data**. *Nucleic Acids Res* 2004, **32**(suppl 1):D129-D133.

343. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A: **PDBFlex: exploring flexibility in protein structures**. *Nucleic Acids Res* 2016, **44**(D1):D423-428.

344. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure* 2003, **11**(11):1453-1459.

345. Dosztányi Z, Csizmok V, Tompa P, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Bioinformatics* 2005, **21**(16):3433-3434.

346. Antonopoulos DA, Assaf R, Aziz RK, Brettin T, Bun C, Conrad N, Davis JJ, Dietrich EM, Disz T, Gerdes S, Kenyon RW, Machi D, Mao C, Murphy-Olson DE, Nordberg EK, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Santerre J, Shukla M, Stevens RL, VanOeffelen M, Vonstein V, Warren AS, Wattam AR, Xia F, Yoo H: **PATRIC as a unique resource for studying antimicrobial resistance**. *Brief Bioinform* 2017.

347. Desilets M, Deng X, Deng X, Rao C, Ensminger AW, Krause DO, Sherman PM, Gray-Owen SD: **Genome-based Definition of an Inflammatory Bowel Disease-associated Adherent-Invasive Escherichia coli Pathovar**. *Inflamm Bowel Dis* 2016, **22**(1):1-12.

348. O'Brien CL, Bringer M-A, Holt KE, Gordon DM, Dubois AL, Barnich N, Darfeuille-Michaud A, Pavli P: **Comparative genomics of Crohn's disease-associated adherent-invasive Escherichia coli**. *Gut* 2016.

349. Krause DO, Little AC, Dowd SE, Bernstein CN: **Complete genome sequence of adherent invasive Escherichia coli UM146 isolated from Ileal Crohn's disease biopsy tissue**. *J Bacteriol* 2011, **193**(2):583.

350. Clarke DJ, Chaudhuri RR, Martin HM, Campbell BJ, Rhodes JM, Constantinidou C, Pallen MJ, Loman NJ, Cunningham AF, Browning DF, Henderson IR: **Complete genome sequence of the Crohn's disease-associated adherent-invasive Escherichia coli strain HM605**. *J Bacteriol* 2011, **193**(17):4540.

351. Nash JH, Villegas A, Kropinski AM, Aguilar-Valenzuela R, Konczy P, Mascarenhas M, Ziebell K, Torres AG, Karmali MA, Coombes BK: **Genome sequence of adherent-invasive Escherichia coli and comparative genomic analysis with other E. coli pathotypes**. *BMC Genomics* 2010, **11**:667.

352. Zhang Y, Rowehl L, Krumsiek JM, Orner EP, Shaikh N, Tarr PI, Sodergren E, Weinstock GM, Boedeker EC, Xiong X, Parkinson J, Frank DN, Li E, Gathungu G: **Identification of**

**Candidate Adherent-Invasive E. coli Signature Transcripts by Genomic/Transcriptomic Analysis**. *PLoS One* 2015, **10**(6):e0130902.

353. Dogan B, Suzuki H, Herlekar D, Sartor RB, Campbell BJ, Roberts CL, Stewart K, Scherl EJ, Araz Y, Bitar PP, Lefébure T, Chandler B, Schukken YH, Stanhope MJ, Simpson KW: **Inflammation-associated adherent-invasive Escherichia coli are enriched in pathways for use of propanediol and iron and M-cell translocation**. *Inflamm Bowel Dis* 2014, **20**(11):1919-1932.

354. Moulin-Schouleur M, Répérant M, Laurent S, Brée A, Mignon-Grasteau S, Germon P, Rasschaert D, Schouler C: **Extraintestinal pathogenic Escherichia coli strains of avian and human origin: link between phylogenetic relationships and common virulence patterns**. *J Clin Microbiol* 2007, **45**(10):3366-3376.

355. Ron EZ: **Host specificity of septicemic Escherichia coli: human and avian pathogens**. *Curr Opin Microbiol* 2006, **9**(1):28-32.

356. Monk J, Bosi E: **Integration of Comparative Genomics with Genome-Scale Metabolic Modeling to Investigate Strain-Specific Phenotypical Differences**. In: *Metabolic Network Reconstruction and Modeling: Methods and Protocols.* Edited by Fondi M. New York, NY: Springer New York; 2018: 151-175.

357. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, **48**(3):443-453.

358. Kitazoe Y, Kishino H, Hasegawa M, Nakajima N, Thorne JL, Tanaka M: **Adaptive threonine increase in transmembrane regions of mitochondrial proteins in higher primates**. *PLoS One* 2008, **3**(10):e3343.

359. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É: **Scikit-learn: Machine Learning in Python**. *J Mach Learn Res* 2011, **12**(Oct):2825-2830.

360. Smole Z, Nikolic N, Supek F, Šmuc T, Sbalzarini IF, Krisko A: **Proteome sequence features carry signatures of the environmental niche of prokaryotes**. *BMC Evol Biol* 2011, **11**:26.

361. Acquisti C, Kleffe J, Collins S: **Oxygen content of transmembrane proteins over macroevolutionary time scales**. *Nature* 2007, **445**(7123):47-52.

362. Semeiks J, Grishin NV: **A method to find longevity-selected positions in the mammalian proteome**. *PLoS One* 2012, **7**(6):e38595.

363. Moosmann B, Behl C: **Mitochondrially encoded cysteine predicts animal lifespan**. *Aging Cell* 2008, **7**(1):32-46.

364. Lv H, Han J, Liu J, Zheng J, Liu R, Zhong D: **CarSPred: a computational tool for predicting carbonylation sites of human proteins**. *PLoS One* 2014, **9**(10):e111478.

365. Hagberg A, Schult D, Swart P, Conway D, Séguin-Charbonneau L, Ellison C, Edwards B, Torrents J: **Networkx. High productivity software for complex networks**. *Webová strá nka https://networkx lanl gov/wiki* 2013.