UNIVERSITY OF CALIFORNIA, SAN DIEGO

Structural Systems Biology Perspective on the Metabolic Impact of Physicochemical Stress

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

in

Bioinformatics and Systems Biology

by

Roger Larken Chang

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Philip E. Bourne, Co-chair
Professor Adam Godzik
Professor Trey Ideker
Professor Milton H. Saier, Jr.

2012

The dissertation of Roger Larken Chang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-chair

_____
Chair

University of California, San Diego

2012

Dedication

To Michelle

For all of your sacrifices, goodness, and love.

You are my soulmate.

To Mom

For giving me my life, my mind, and my heart.

I'll always be your boy forever.

Epigraph

Ask not what you can do for a reconstruction, but what a reconstruction can do for you.

Bernhard Ø. Palsson

In some ways structure is the devil in the details of systems biology.

Philip E. Bourne

Table of Contents

List of Figures

Acknowledgements

I owe many people greatly for their support in reaching this point in my academic career, culminating in this doctoral thesis. This could not have been accomplished without many sacrifices from the people who have supported me and led me down this path.

First, I would like to thank all of my scientific mentors who have taught me, encouraged me, and otherwise contributed to furthering my academic career. I thank Dr. Aziz Aboobaker for teaching me techniques of molecular biology, for introducing me to my current field of study, and for commiserating with me during a time when we were both enduring personal loss. Thanks to Dr. Richard Scheuermann for giving me a chance to develop and prove myself as an independent bioinformatics researcher when I was a complete novice in the field. Richard helped set me on the path towards the doctorate I am now completing and has continued to be a great mentor and friend. I would like to thank my doctoral co-advisor Dr. Philip Bourne for being an open-minded and unconditionally supportive mentor, starting when he helped me during my research rotation in his group to design the first project that would eventually grow into this thesis. I have richly benefited from many conversations with Phil, not only intellectually but in determining what is most important to me in my career. I would particularly like to thank my doctoral advisor Dr. Bernhard Palsson for recognizing my personal research style and advising me accordingly, for supporting my extra-thesis activities over the years despite the burdens they placed upon my primary research, and for showing me perspectives to approaching science that have guided my career development. Thanks, Dr. Palsson, for trusting me with my sometimes seemingly crazy research ambitions. We were able to break new ground in a number of areas together.

I would also like to thank the many scientists with whom I have collaborated and from whom I have learned during my graduate career. At UCSD, I have befriended and collaborated with many people. In particular, I would like to thank those I have collaborated with most closely, including Nate Lewis, Joshua Lerman, Teddy O'Brien, Lei Xie, Li Xie, Kathy Andrews, Hojung Nam, and Donghyuk Kim. I would also like to thank Harish Nagarajan, Nitin Udpa, Jan Schellenberger, Jeff Orth, Monica

Mo, Daniel Zielinski, Pep Charusanti, and Adam Feist, who have all provided deep and fruitful discussion and collaboration over the last few years. I also thank Ines Thiele and Iman Famili, who acted as senior mentors and role models to me early in my graduate career. I would also like to thank Marc Abrams and Kathy Andrews for their wonderful support in getting grant and fellowship applications submitted, helping me to manage official business, and for their friendship.

I would also like to thank my closest collaborators in my three-year long work on algal metabolism (my original thesis plan), without whom I would not have accomplished and learned nearly as much as I have as a graduate student. Thanks to Jason Papin, Ani Manichaikul, Erik Hom, Lila Ghamsari, Kourosh Salehi-Ashtiani, Phil Lee, and Stephen Mayfield.

I also thank Autumn Ross and the students of her 2011-2012 biology and chemistry classes for their support as I developed my teaching skills at Patrick Henry High School.

I would also like to thank the members of my thesis committee collectively for their candor, for rigorously evaluating my research, and for taking their personal time to provide their evaluation and feedback.

Lastly, my success in this endeavor would not have been possible without the loving support of my family. I thank my parents, Janet Fraser and Roger A. Chang, for bringing me into this world, supporting and loving me, and nurturing my scientific curiosity as a child. I thank Mary, who will always be “Granny” to me, for supporting me in my life and education for as long as she lived. I hope I have made her proud that I am her grandson. I am grateful to my siblings Brook, Sean, and Meli for their moral support over the years. Most importantly, I am grateful to my wife Michelle Irick for her love, friendship, and support over the last nine years. If not for her sacrifices, this would not have been possible.

Chapter 1, in part, is a reprint of the material as it appears in 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 Sep 23;6(9):e1000938. I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 2, in part, is a reprint of the material as it appears in Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BØ. Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. *In preparation*. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Chapter 3, in part, is a reprint of the material as it appears in Chang RL, Bourne PE, Palsson BØ. Antibacterial mechanisms identified through structural systems pharmacology. *In preparation*. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Vita

2005 B.A., Molecular and Cell Biology, University of California, Berkeley

2012 Ph.D., Bioinformatics and Systems Biology, University of California, San Diego

Publications

11. **Chang RL**, Bourne PE, Palsson BØ. Antibacterial mechanisms identified through structural systems pharmacology. *In preparation.*
10. O'Brien ET, Lerman JA, **Chang RL**, Hyduke DR, Palsson BØ. Enhanced prediction and understanding of functional states of *Escherichia coli*. *In preparation.*
9. **Chang RL**, Andrews K, Kim D, Li Z, Godzik A, Palsson BØ. Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. *In preparation*.
8. Lewis NE, **Chang RL**, Kim D, Hefzi HH, Palsson BØ. Prokaryotes employ enzyme post-translational modification to globally regulate metabolism. *In preparation.*
7. Nam H, Lewis NE, Lerman JA, Lee D, **Chang RL**, Kim D, Palsson BO. Network context and selection in the evolution to enzyme specificity. *Science*. 2012, Aug 31;337(6098):1101-1104.
6. **Chang RL**, Ghamsari L, Manichaikul A, Hom EF, Balaji S, Fu W, Shen Y, Hao T, Palsson BØ, Salehi-Ashtiani K, Papin JA. Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol Syst Biol*. 2011, Aug 2;7:518. doi: 10.1038/msb.2011.52.
5. **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, Sep 23;6(9):e1000938.
4. **Chang RL**, Luo F, Johnson S, Scheuermann RH. Deterministic graph-theoretic algorithm for detecting modules in biological interaction networks. *Int J Bioinform Res Appl*. 2010;6(2):101-19.
3. Manichaikul A*, Ghamsari L*, Hom EFY*, Lin C*, Murray RR*, **Chang RL***, Balaji S, Hao T, Shen Y, Chavali AK, Thiele I, Yang X, Mello E, Hill DE, Vidal M, Salehi-Ashtiani K, Papin JA. Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat Methods* 2009, Aug;6(8):589-92.
2. Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V, **Chang R**, Larsen CN, Klem E, Biersack K, Scheuermann RH. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Res.* 2008, Jan;36(Database issue):D497-503.
1. Luo F, Yang Y, Chen CF, **Chang R**, Zhou J, Scheuermann RH. Modular organization of protein interaction networks. *Bioinformatics* 2007, Jan 15;23(2):207-214.

* Authors contributed equally

ABSTRACT OF THE DISSERTATION

Structural Systems Biology Perspective on the Metabolic Impact of Physicochemical Stress

by

Roger Larken Chang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2012

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

Due to the relative prevalence and centrality of the proteome, it is unsurprising that a great number of environmental stressors exert pressure on the cell via impacting proteome function. Whether it be in a global sense, as in temperature or pH destabilizing large fractions of the proteome, or in a more local sense, as in the targeting of one or a handful of proteins by an inhibitory compound, it is not possible to understand physicochemical pressures on the cellular system without considering properties of the proteome. This thesis aims to enable the analysis and simulation of such physicochemical constraints upon the cellular system through integration of systems biology with protein structural data and computational methods. This is the approach of the emerging field of structural systems biology. This work demonstrates examples of interrogating physicochemical stress imposed upon metabolic systems by exposure to exogenous chemicals and non-optimal temperatures. An extensive data resource was developed to capture biologically-relevant protein structural states to be integrated with the

genome-scale metabolic model of the bacterium *Escherichia coli*. The primary results include 1) prediction of causal drug off-targets to explain a lethal but poorly understood drug side effect in humans, 2) establishing metabolic activities as growth-rate limiting under heat shock conditions and discovering specific bottlenecks such stress creates in the metabolic system of *E. coli*, and 3) analysis of antibacterial mechanisms of both well- and poorly-understood compounds and drug design via protein targeting. Thus, through the integration of structural and systems biology, new insights are provided about the impact of physicochemical stress on complex biological systems.

Chapter 1: Drug off-target effects predicted using structural analysis in the context of a metabolic network model.

Abstract

Recent advances in structural bioinformatics have enabled the prediction of protein-drug off-targets based on their ligand binding sites. Concurrent developments in systems biology allow for prediction of the functional effects of system perturbations using large-scale network models. Integration of these two capabilities provides a framework for evaluating metabolic drug response phenotypes *in silico*. This combined approach was applied to investigate the hypertensive side effect of the cholesteryl ester transfer protein inhibitor torcetrapib in the context of human renal function. A metabolic kidney model was generated in which to simulate drug treatment. Causal drug off-targets were predicted that have previously been observed to impact renal function in gene-deficient patients and may play a role in the adverse side effects observed in clinical trials. Genetic risk factors for drug treatment were also predicted that correspond to both characterized and unknown renal metabolic disorders as well as cryptic genetic deficiencies that are not expected to exhibit a renal disorder phenotype except under drug treatment. This study represents a novel integration of structural and systems biology and a first step towards computational systems medicine. The methodology introduced herein has important implications for drug development and personalized medicine.

Introduction

Despite the advantages gained from drug therapy in medicine, drug development has historically presented an expensive and frequently perplexing challenge for researchers. Identifying useful drug targets for treating disease and matching them to chemical compounds that can elicit the desired effect through drug-target interaction has been the paradigm for the drug development process in the era of molecular medicine. However, this approach has yielded many failed drug treatments and an incomplete understanding of the consequences of treatments for human health, even with drugs that have made it to market and been prescribed for decades. Two major contributing factors that confound individual molecular target-based drug discovery are drug off-target binding and the lack of systems-level understanding of drug response[1]. Adopting a new, systems-based approach to drug development is therefore a desirable goal in the era of systems medicine.

The growing wealth of omics data offers a valuable opportunity for novel approaches in systems medicine but also presents significant challenges for data integration[2]. Increasingly sophisticated computational approaches are being developed to analyze and manipulate omics data in order to gain a greater understanding of complex biological systems. An algorithm for identifying and comparing ligand binding sites on protein structures[3] was recently employed to predict drug off-target binding sites across the proteome[4]. Such a tool offers unique capabilities for drug development by providing a comprehensive survey of uncharacterized drug targets that may participate directly in drug response, which is likely to be important as polypharmacology interactions suggest that drug promiscuity is a predominant property of existing drugs[4, 5].

Biological systems exhibit redundant pathways and synergistic effects conferring a robustness of phenotype when confronted with external stimuli. As a result, multi-target drugs are generally more clinically efficacious than single-target drugs. These facts highlight the critical importance of studying polypharmacology in a systems level context[6]. The increasing use of genome-scale metabolic network models for a variety of applications[7, 8] has established this research platform as a promising means for studying the emergent properties of complex systems. The published applications of metabolic models for drug development have thus far focused on identifying drug targets for antibacterial treatment in

such pathogens as *M. tuberculosis*[9, 10], *S. aureus*[10, 11], *H. pylori*, and *E. coli*[10]. However, the human metabolic network reconstruction (Recon 1)[12] and developed context-specific metabolic modeling algorithms[13, 14] permit human-centered *in silico* drug studies. Integrating these structural bioinformatics and human system modeling techniques for application in drug development represents a first computational step into the era of systems medicine. As an example of this integrative approach, the results of protein off-target prediction for the drug torcetrapib[4], a cholesteryl ester transfer protein (CETP) inhibitor, were evaluated in the context of a model of renal metabolism.

CETP inhibitors are intended to treat patients at risk for atherosclerosis and other cardiovascular diseases by raising high-density lipoprotein cholesterol (HDL-C) and lowering low-density lipoprotein cholesterol (LDL-C)[15]. Torcetrapib was withdrawn from phase III clinical trials after a substantial investment of labor and capital due to its observed side effect of fatal hypertension in some patients[16]. It has since been of great interest to elucidate the cause of this side effect in order to avert such failures in the future and to better define the potential of CETP inhibitors for treatment[17]. Subsequent studies have provided evidence in favor of the hypothesis that the cause of this side effect was not due directly to the mechanism of HDL-C and LDL-C regulation via CETP inhibition[18]. Instead, it has been suggested that the hypertensive side effect may result from uncharacterized drug off-target effects[17]. Two other CETP inhibitors are now under clinical trial, anacetrapib[18] and JTT-705[19]. Thus far, studies have not indicated the same risk of hypertension associated with the latter two drugs; however, these studies have been limited to relatively small patient groups lacking in diversity and over relatively short-term treatment. Even if these alternative CETP inhibitors do not carry the same adverse side effects, it is still of value to future drug development to determine the exact mechanism of torcetrapib's adverse action. It has been suggested that off-target effects of torcetrapib lead to increased activity of the renin-angiotensin-aldosterone-system (RAAS) and thereby hypertension[4, 20], but a recent review of the published CETP inhibitor clinical studies[21] concludes that the effect on blood pressure is most likely independent of the increase in aldosterone. Currently the exact cause of the hypertensive side effect of torcetrapib remains to be unambiguously identified.

The predicted torcetrapib off-targets include many metabolic enzymes and metabolite transport proteins. Although there are several mechanisms involved in regulating blood pressure that may be responsible for the hypertensive side effect, one possible mode is the renal regulation of blood pressure via metabolite reabsorption and secretion. The kidneys are the primary organs that filter the blood and therefore are strong contributors to maintaining a normotensive state even independent of RAAS function. Thus a model of renal metabolism was developed as the system context in which to analyze torcetrapib off-targets and predict drug response phenotypes. The two best-supported causal off-targets predicted in this study are prostaglandin I2 synthase (PTGIS), due to decreased capacity for renal prostaglandin I2 (PGI2) secretion, and acyl-CoA oxidase 1 (ACOX1), due to decreased capacity for renal citrate and amino acid reabsorption. Four other predicted off-targets are also predicted to impact amino acid, glucose, citrate, or bicarbonate reabsorption. As well, the model predicts no effect on renal reabsorption or secretion for a number of other predicted off-target metabolic proteins.

The goal of this study is not only to provide new insight into the torcetrapib problem but also to reveal the theoretical implications that this computational systems medicine platform has for drug development and personalized medicine. Characterizing the influence that genetic variation has in determining drug response phenotypes has been recognized as a crucial goal for the future of drug development[22]. To this end, the renal model was also used to analyze metabolic disorders resulting from genetic deficiencies and to identify those deficiencies that may pose additional risks for drug treatment in select individuals.

Although many of the predictions generated by this approach are supported by clinical and other experimental evidence that describe the impact of loss of function for predicted causal off-targets and genetic deficiencies, the full set of exact metabolic mechanisms of drug action predicted by our model remain to be completely validated. While this is seen as a limitation of this study, it also offers a number of opportunities to experimentally evaluate promising hypotheses that, if validated, will lead to significant advancements in developing CETP inhibitors for treatment and novel insight into certain renal disorders.

Results

Renal Metabolic Model

The approach for context-specific organ modeling proposed in this study (see Methods and Figure 1.1) yielded a renal metabolic model capturing functions of the kidney for reabsorption and secretion (Table 1.1). Many components of the renal objective function are factors known to be relevant determinants of blood pressure. However, there is currently incomplete knowledge about the exact role that some of these components play in blood pressure regulation. Calcium reabsorption, for example, leads to vasoconstriction in kidney glomeruli through the action of L-type and N-type calcium ion channels[23] suggesting a resulting increase in blood pressure if this mechanism applies across all vascular tissues. Calcium reabsorption also leads to an inhibition of renal sodium reabsorption in the proximal tubule[24] suggesting a blood pressure lowering effect consistent with the observation that increased dietary calcium also lowers blood pressure[25]. This highlights the complexity of the effect certain renal reasborptions have on blood pressure. Nevertheless, the many components accounted for in the renal objective function enabled explicit predictions about how system perturbations such as drug treatment and genetic deficiencies affect the kidney's ability to regulate the small molecule content of the blood.

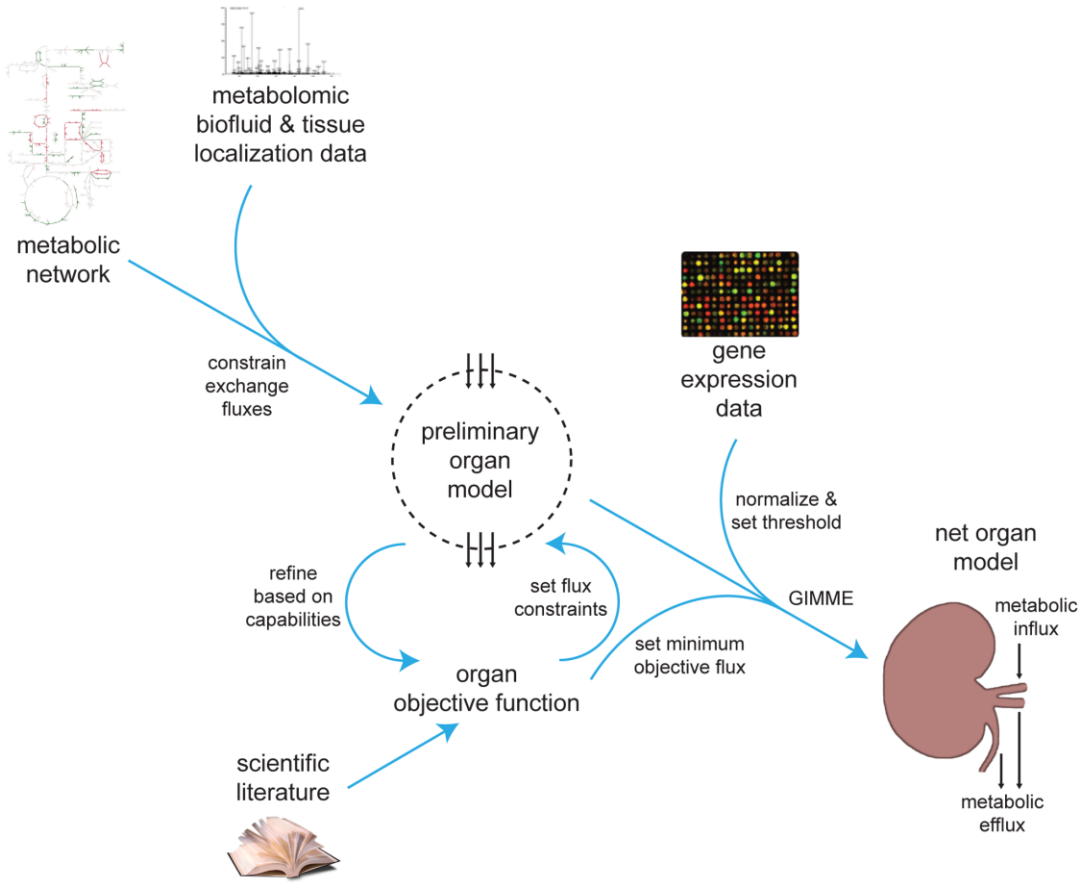

Figure 1.1. Context-specific organ metabolic modeling. Preliminary constraints were imposed upon metabolite exchange fluxes of the full metabolic network based on coordinated experimental detection of transportable metabolites both in the organ tissue and the biofluids processed by the organ. Metabolites detected in both biofluid and organ were assumed freely exchangeable in the model, and the remainder of the metabolite exchanges were tentatively constrained to zero. Organ physiology literature was reviewed to compile an objective function consisting of the metabolic functions of the organ. Each function was tested for compatibility with the preliminary model. Metabolite exchange, transport, and demand reactions required to achieve some functions were added to the network, and exchange fluxes for objective metabolites were directionally constrained in accordance with the literature. Functions not compatible with the model were removed from the overall objective function. The objective function was then integrated with gene expression data obtained from an organ tissue sample to derive a net, context-specific metabolic organ model representing the metabolic exchange between the organ and the rest of the body and the metabolic reactions that take place within the organ to achieve this exchange.

Table 1.1. Renal objective function.

Exchange	Class	Metabolite	Abbreviation	Relation to Blood Pressure
secretion	hormones	Prostaglandin I2	PGI2	vasodilation
		Prostaglandin D2	PGD2	vasodilation
		Calcitriol	-	lowers blood pressure, Ca2+ reabsorption
	urea	Urea	-	water/ion counter current system regulating osmolality
	cyclic amp	Cyclic AMP	cAMP	important for vaso-dilation/constriction
	urate	Urate	-	uknown, but secreted
	tryptamine	Tryptamine	-	uknown, but secreted
absorption	water	Water	H2O	determinant of blood pressure, ion absorption
	ions/electrolytes	Phosphate	-	determinant of blood pressure
		Sodium	Na+	determinant of blood pressure
		Calcium	Ca2+	determinant of blood pressure
		Chloride	Cl-	determinant of blood pressure
		Protium	H+	determinant of blood pressure
		Potassium	K+	determinant of blood pressure
		Bicarbonate	HCO3-	determinant of blood pressure
	carboxylates	Acetate	-	unknown, but reabsorbed
		Citrate	-	effects sodium reabsorption
		Oxalate	-	effects sodium reabsorption
	glucose	D-Glucose	-	effects sodium reabsorption
	amino acids	L-Alanine	Ala	associated reduction of hypertension/vasodilation
		L-Arginine	Arg	associated reduction of hypertension/vasodilation
		L-Asparagine	Asn	associated reduction of hypertension/vasodilation
		L-Aspartate	Asp	associated reduction of hypertension/vasodilation
		L-Cysteine	Cys	associated reduction of hypertension/vasodilation
		L-Glutamine	Gln	associated reduction of hypertension/vasodilation
		L-Glutamate	Glu	associated reduction of hypertension/vasodilation
		Glycine	Gly	associated reduction of hypertension/vasodilation
		L-Histidine	His	associated reduction of hypertension/vasodilation
		L-Isoleucine	Ile	associated reduction of hypertension/vasodilation
		L-Leucine	Leu	associated reduction of hypertension/vasodilation
		L-Lysine	Lys	associated reduction of hypertension/vasodilation
		L-Methionine	Met	associated reduction of hypertension/vasodilation
		L-Phenylalanine	Phe	associated reduction of hypertension/vasodilation
		L-Proline	Pro	associated reduction of hypertension/vasodilation
		L-Serine	Ser	associated reduction of hypertension/vasodilation
		L-Threonine	Thr	associated reduction of hypertension/vasodilation
		L-Tryptophan	Trp	associated reduction of hypertension/vasodilation
		L-Tyrosine	Tyr	associated reduction of hypertension/vasodilation
		L-Valine	Val	associated reduction of hypertension/vasodilation
	oligopeptides	L-Carnosine	-	unknown, but reabsorbed
		Glutathione	GSH	unknown, but reabsorbed

The kidney model included 336 explicitly predicted active metabolic genes (see Table S1 in [26]) that met criteria for activity as summarized in Figure 1.2. The majority, 243 genes, satisfied the gene expression significance threshold (see Methods), although the activity of 58 genes was predicted despite expression values below the threshold. These genes were activated by the GIMME algorithm[13] to optimally achieve the renal objectives while remaining minimally inconsistent with gene expression data and may represent post-transcriptionally upregulated genes. The other 35 genes were predicted to be active without penalty since no corresponding probesets existed on the microarray upon which the transcriptomic data was obtained. Since many of these genes participated in optimal pathways for

achieving renal objectives, it is projected that experimental measurement would confirm their expression if performed.

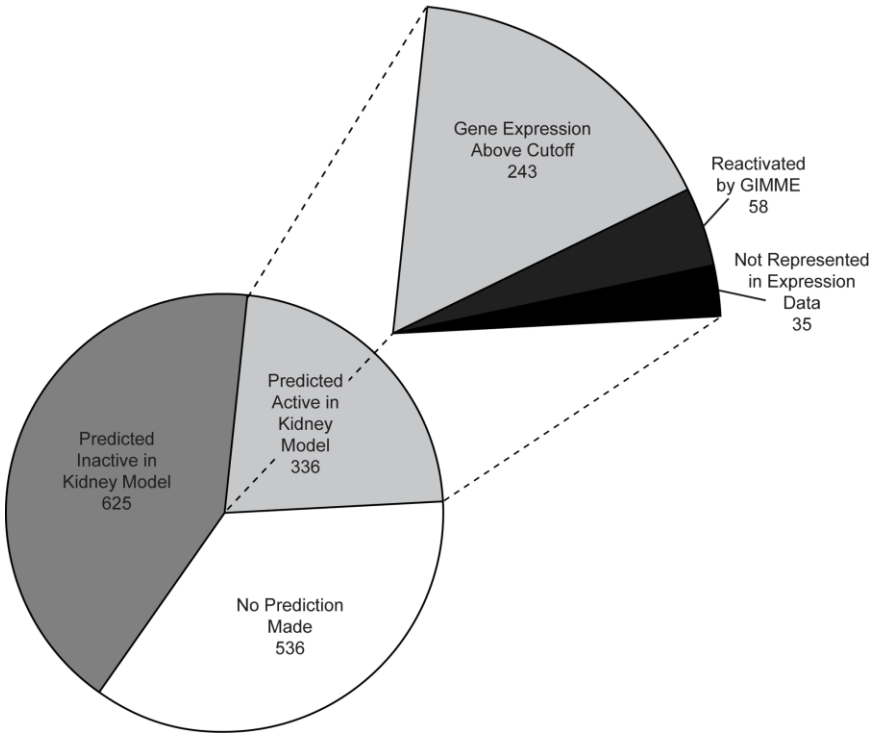

Figure 1.2. Summary of gene activity predictions in the full kidney model. The pie chart at bottom represents the Recon1 gene activity predictions resulting from deriving the kidney model. Genes predicted inactive are those genes with no associated active reaction fluxes in the kidney model. Genes for which no activity prediction was made are those associated with active reaction fluxes in the kidney model but either are not represented in the gene expression data or were not determined as the gene whose expression level is most limiting for any associated reaction through evaluation of GPR Boolean rules with respect to gene expression data. The slice at top represents genes predicted active in the kidney model.

The active reactions in the model reflect both the possible pathways by which the kidney can achieve the specified renal objectives as well as other functions supported by the gene expression data. The model included 1587 active reactions (see Table S2 in [26]), excluding model-based reactions such as objective functions, exchanges, and demands. Of these active reactions, 333 comprised a single connected sub-model accounting for all pathways which could possibly support the specified renal objectives. We refer to this sub-model as the reduced kidney model (see Table S1 and Table S2 in [26] for the contents of the reduced model and Dataset S1 in [26] for the actual model in SBML format). It should

be noted that because the reduced model included all reactions that can carry flux in support of the renal objectives, it had the exact same effective flux state solution space as the full renal model. The reduced kidney model reactions spanned a broad range of metabolic subsystems (Figure 1.3). The largest subsystem consisted of plasma membrane-spanning transport reactions, which is expected given that this model captured renal filtration and secretion functions. The second largest subsystem represented intracellular transport, signifying the importance of interaction among sub-cellular compartments in renal function including the cytosol, endoplasimic reticulum, Golgi apparatus, and mitochondria. A significant proportion of the other active subsystems in the reduced kidney model were involved directly in the metabolism of components of the renal objective function including carbohydrate, amino acid, vitamin, lipid, carboxylate, and glutathione metabolism as well as the urea cycle. These permitted the indirect reabsorption of metabolites as well as the required synthetic pathways for renal secretions.

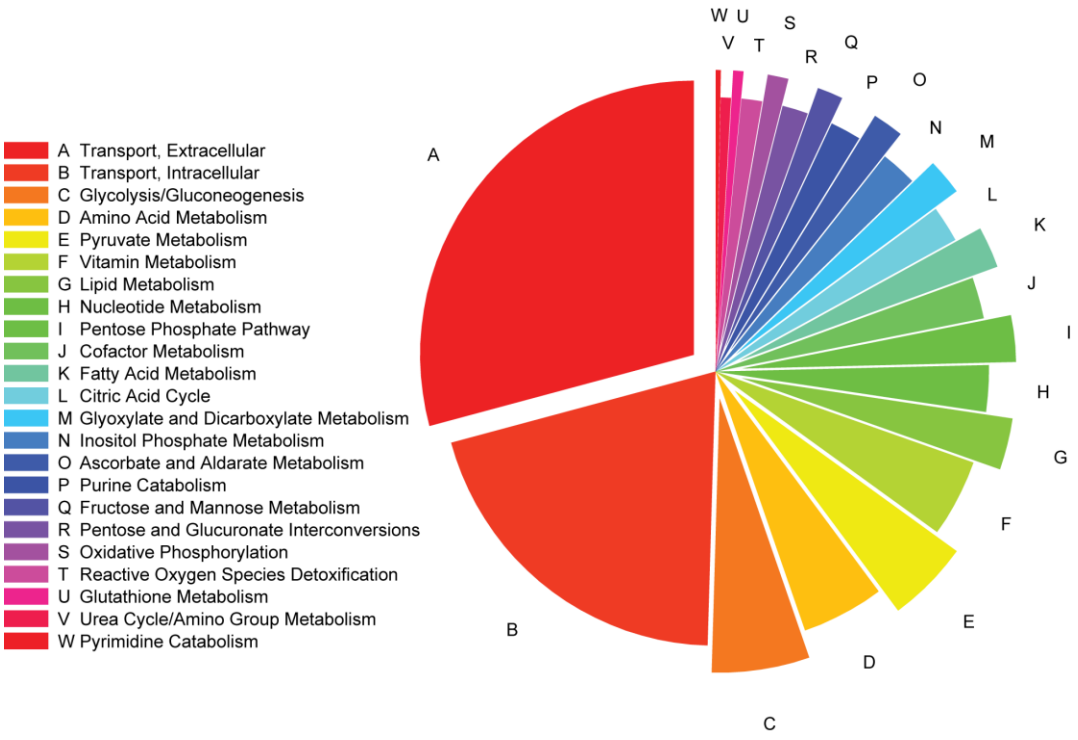

Figure 1.3. Reduced kidney model subsystem distribution. The distribution of metabolic reactions predicted to be active in the reduced kidney model with respect to broad metabolic subsystem categories is shown. The distribution excludes objective function, exchange, and demand reactions used to perform simulations in the model.

Causal Drug Off-Targets

The integrative framework adopted for predicting causal drug targets associated with response phenotypes employed both structural bioinformatics tools as well as modeling techniques of systems biology (see Methods and Figure 1.4). The workflow begins with screening of the entire human structural proteome, with each subsequent step in the process narrowing the list of proteins ultimately into a set of targets for which a response phenotype was predicted upon functional inhibition. The first step of this process identified putative off-target drug binding sites using a ligand-binding site structural alignment algorithm (see Methods). The 41 predicted metabolic protein off-targets were the focus of this study (see Table 1.2), 28 of which had predicted drug binding sites overlapping with their functional sites. Simulated inhibition of these targets in the reduced kidney model (see Methods) predicted response phenotypes for 6 of the off-target proteins with respect to renal function (Figure 1.5). The results of all analysis steps for these 6 off-targets are summarized in Table 1.3. The expression of all of these targets was determined to be the most limiting for their associated metabolic reactions included in the reduced kidney model (see Methods), providing additional evidence supporting that inhibition of these targets would be expected to have at least some deleterious impact on those reactions.

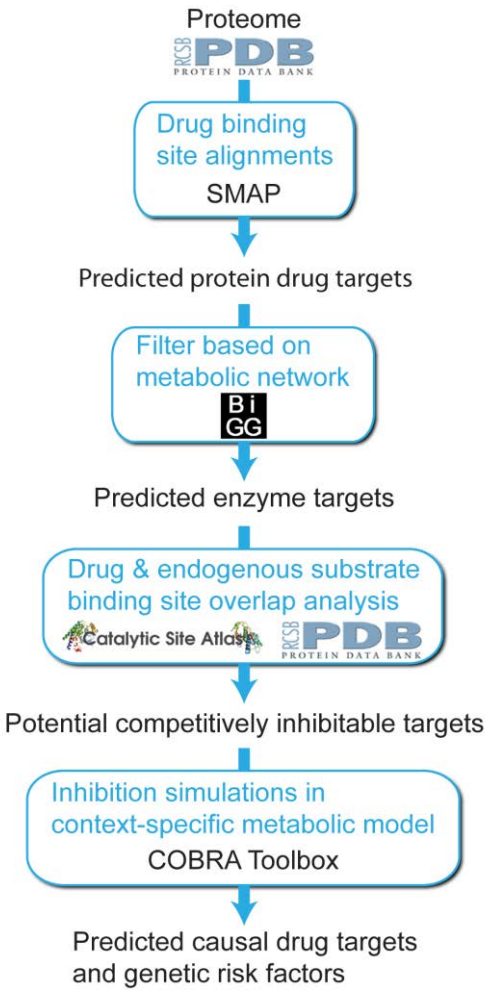

Figure 1.4. Identifying causal genes for drug response phenotypes and metabolic disorders. First, the human proteome was screened to identify off-target drug-binding sites. The resulting list of putative off-targets was filtered to focus on just metabolic proteins. Then, for each predicted metabolic off-target, the endogenous functional sites were compared to the predicted drug-binding site to identify overlap. Off-target proteins for which overlapping binding sites were identified were considered to be competitively inhibitable by the drug at the overlapping endogenous functional sites. The functional consequences of such inhibitions were then tested in an appropriate context-specific metabolic model. All possible individual gene knockouts were also simulated to predict genetic disorders that lead to functional deficiencies either alone or in combination with drug treatment. Those off-targets whose inhibition impacted model function represent causal off-targets predicted to be associated with the drug response phenotype, and the gene knockouts that impacted model function represent genetic risk factors for metabolic disorders, which may lead to amplification of the drug response phenotype.

Table 1.2. Metabolic protein drug target predictions.

Official Symbol	PDB ID	Gene ID	SMAP Prediction	Functional Site Overlap	Reduced Model Reactions Limited by Expression	Impacts Renal Function in Simulation	Stronger Drug Binding Affinity
PTGIS	2IAG	5740	x	x	x	x	x
ACOX1	1IS2[a]	51	x	x	x	x	x
AK3L1	2BBW	205	x	x	x	x	
HAO2	1LTD[a]	51179	x	x	x	x	
MT-CO1	1V54[a]	4512	x	x	x	x	
UQCRC1	1PP9[a]	7384	x	x	x	x	
SLC4A1	1HYN	6521	x		x		
HSD17B10	1U7T	3028	x	x			x
CRAT	1XMC[a]	1384	x	x			x
GLTP	1TFJ[a]	51228[b]	x	x			x
PDE10A	2OUN	10846	x	x			x
PFKFB1	1K6M	5207	x	x			x
TTPA	1OIP	7274[b]	x	x			
UGP2	2I5K[a]	7360	x				
PDE1C	1LXS	5137	x				
PCTP	1LN1	58488[b]	x	x			x
CYP2C9	1R9O	1559	x	x			x
HMOX2	2Q32	3163	x	x			x
IDO1	2D0T	3620	x	x			x
STARD3	1EM2	10948	x	x			x
PYCR1	2IZZ	5831	x	x			x
CYP19A1	3EQM	1588	x	x			x
HAL	1B8F[a]	3034	x	x			x
PFAS	1T3T[a]	5198	x	x			x
PPOX	2IVD[a]	5498	x	x			x
TDO2	2NOX[a]	6999	x	x			x
INMT	1VLM[a]	11185	x	x			x
DLAT	3B8K	1737	x	x			x
DHODH	2FPT	1723	x	x			
HAO1	1LCO[a]	54363	x	x			
HSD17B1	1I5R	3292	x	x			
AANAT	1KUX[a]	15	x				
STS	1P49	412	x				
CYP21A2	2GEG	1589	x				
HNMT	1JQE	3176	x				
HSD17B4	1IKT	3295	x				
CHP	2E30	11261	x				
COASY	2F6R[a]	80347	x				
SLC2A7	1YG7	155184	x				
SLC2A5	1YG1	6518	x				
CSAD	2JIS	51380	x				

[a] Non-human protein structures were mapped to human genes via BLAST against the human proteome and choosing the top hit only if it had E-value < 1E-50.

[b] Incomplete or incorrect gene-protein-reaction associations (GPRs) were associated with the correct gene IDs based on comparative BLAST results and gene functional annotation.

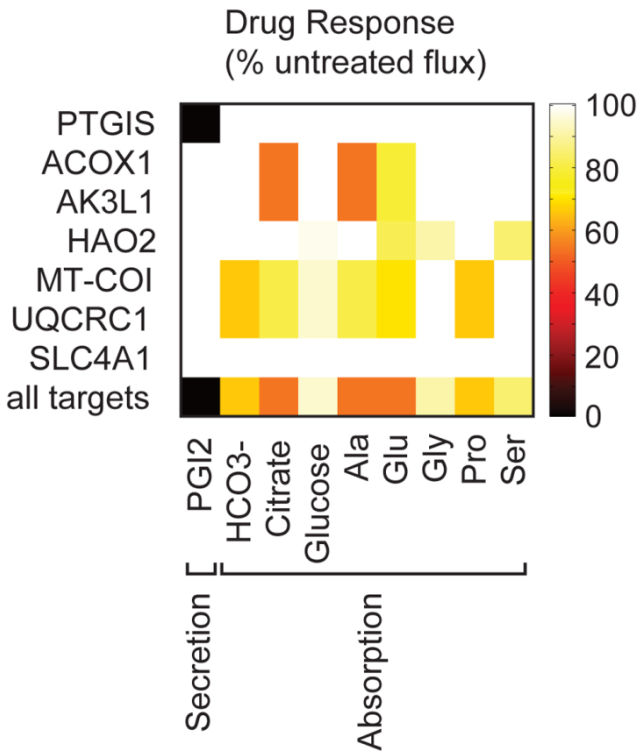

Figure 1.5. CETP inhibitor renal response phenotypes. Elements of the color matrix represent the percent of the maximum normal, untreated renal objective flux achievable by the CETP-inhibitor-treated normal kidney model. The x-axis corresponds to individual renal objective functions, and the y-axis corresponds to the predicted drug off-targets. Metabolite abbreviations are defined in Table 1.1. Only the subset of renal objective functions for which a drug response phenotype was predicted is displayed.

Table 1.3. Drug side effect causal off-targets.

Official Symbol	PDB ID	Gene ID	SMAP Prediction	Functional Site Overlap	Reduced Model Reactions Limited by Expression	Impacts Renal Function in Simulation	Stronger Drug Binding Affinity	Cryptic Genetic Risk Factors
PTGIS	2IAG	5740	×	×	×	×	×	
ACOX1	1IS2[a]	51	×	×	×	×	×	
AK3L1	2BBW	205	×	×	×	×		
HAO2	1LTD[a]	51179	×	×	×	×		SLC3A1; SLC7A9; SLC7A10; ABCC1
MT-COI	1V54[a]	4512	×	×	×	×		CYP27B1; ABCC1
UQCRC1	1PP9[a]	7384	×	×	×	×		CYP27B1; ABCC1

[a]Non-human protein structures were mapped to human genes via bi-directional BLAST against the human proteome and choosing the top hit only if it had E-value < 10^{-50}.

The renal response phenotypes for inhibition of two of the predicted drug off-targets were supported by existing scientific literature. Simulated PTGIS inhibition completely precluded PGI2 secretion. Based on the relation of renal PGI2 secretion to blood pressure (see Table 1.1), this inhibition would be expected to have a hypertensive effect. Experimental studies confirmed that PTGIS is associated with essential hypertension in humans[27] and that transgenic rats highly expressing human PTGIS exhibited decreased mean pulmonary arterial pressure despite treatment with monocrotaline to induce hypertension[28]. Inhibition of hydroxyacid oxidase 2 (HAO2) in the reduced kidney model led to reduced glutamate, glycine, and serine reabsorption suggesting a possible role for HAO2 in the hypertensive side effect following CETP inhibitor treatment based on the association of amino acid reabsorption with vasodilation and hypertension (see Table 1.1). HAO2 is highly expressed in human kidney[29] and was identified as a candidate quantitative trait locus for blood pressure in rat kidney in a study comparing normal to hypertensive rats[30].

Two predicted causal CETP inhibitor off-targets, PTGIS and ACOX1, exhibited notable binding affinity differences when comparing docking results for their endogenous substrates to those for the three CETP inhibitors (Figure 1.6). The mean predicted binding affinity of PTGIS for its endogenous substrate prostaglandin H2 was weaker than for all three CETP inhibitors (Figure 1.6A). Anacetrapib was predicted to have the strongest mean binding affinity of all four tested molecules for PTGIS and JTT-705 the weakest of the three drugs. The predicted mean binding affinity of ACOX1 for its endogenous substrate palmitoyl-CoA was weaker than for torcetrapib and anacetrapib but stronger than the affinity of the protein for JTT-705 (Figure 1.6B). These results supported potential competitive inhibition of PTGIS and ACOX1 by torcetrapib and anacetrapib, but the predictions suggested a lesser effect of JTT-705 on ACOX1.

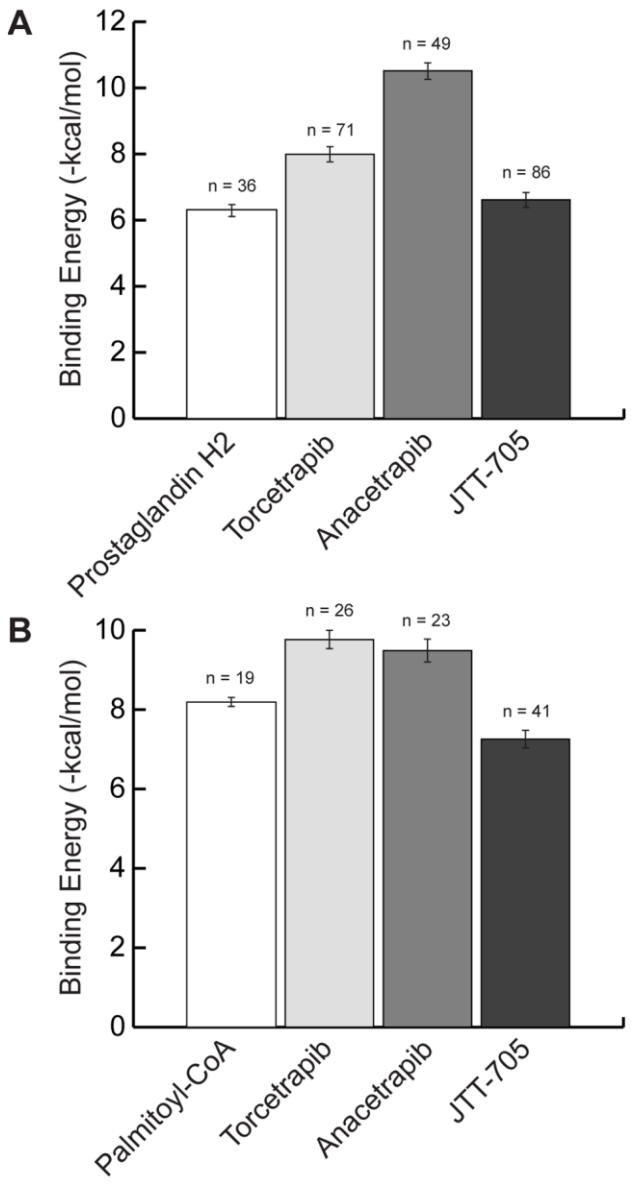

Figure 1.6. Differential causal off-target ligand and drug binding affinities. (A) Binding affinities of the prostaglandin I2 (prostacyclin) synthase protein for CETP inhibitors and prostaglandin H2, the endogenous substrate. **(B)** Binding affinities of the acyl-Coenzyme A oxidase 1, palmitoyl protein for CETP inhibitors and palmitoyl-CoA, the endogenous substrate. Each bar shows the mean binding energy predicted from docking trials. The standard error is indicated for each bar along with the number of predicted binding poses.

Renal Disorders and Drug Treatment

Similar to the use of the model to test inhibitory effects on drug targets, the model was also used to predict genetic deficiencies that lead to renal disorders and drug off-targets that act

synergistically with genetic deficiencies. Simulated gene knockouts predicted to impact renal objective functions are displayed in Figures S1 and S2 and Table S4 in [26]. The 118 deficient genes predicted to cause disorders impacted a variety of renal secretions and absorptions to varying degrees. Thirteen of these deficiencies predicted total loss of at least one renal function (see Figure S2 in [26]).

Some renal disorders were only predicted in the gene-deficient models in combination with drug treatment, not in the untreated gene-deficient models or in the normal drug-treated model, and are referred to in this study as cryptic genetic risk factors. Five such gene deficiencies were predicted (see Table S4 in [26]). A deficiency in CYP27B1, which impacted vitamin D secretion alone, also exhibited defects in proline reabsorption when combined with drug treatment in simulation. Defects in three amino acid transport proteins (SLC7A10, SLC3A1, and SLC7A9) were predicted to decrease renal glycine reabsorption in combination with drug treatment along with the disorders predicted in the absence of drug treatment. The model deficient in the ATP-binding cassette sub-family C member 1 gene (ABCC1) was predicted to exhibit a cryptic deficiency in renal phosphate reabsorption under drug treatment. These predictions are of special importance because they suggest that these renal phenotypes would only surface in gene-deficient individuals under certain conditions, such as when treated with CETP inhibitors.

Model Evaluation and Validation

Multiple evaluations were performed to analyze and validate the content of the reduced kidney model. The reduced kidney model effectively predicted activity of significantly expressed metabolic genes. The ability of our modeling approach to correctly and robustly predict activity of highly expressing genes was evaluated by a five-fold cross validation (see Methods). Our approach showed significant recall of the 20% most highly expressed metabolic genes, p-value = 4.57×10^{-22}. This observation is especially notable since the reduced kidney model was not a global model of kidney metabolism, and the result suggests the relative importance of the renal functions captured by our model within the context of total kidney gene activity.

We compared the metabolic gene activity predictions from the reduced kidney model to the set of significantly expressed genes as well as to a proteomic dataset derived from normal, healthy human

kidney glomerulus tissue[31] (Figure 1.7A). A total of 164 genes active in the reduced kidney model, 72% of the predicted activities, were supported by either significantly expressed mRNA levels, high-confidence protein detection, or both (see Table S1 in [26] for a detailed list). The remaining 64 gene activities accounted for in the model include 23 genes with no corresponding microarray probesets, and therefore not experimentally measured mRNA, and 41 genes that were determined to express more marginally below the established significance threshold. Despite a strong overlap between the transcriptomic and proteomic datasets, there were also large proportions of both which are unique. This disagreement may be due to tissue samples being taken from different kidney sub-tissues in each experiment, absent probesets on the microarray, or the propensity of mass spectrometry proteomic experiments to produce false negatives. All of the counted activities in Figure 1.7A were included in the full human metabolic network, signifying that the reduced kidney model was not a global kidney model and that there is potential for expansion to account for more metabolic functions than those of concern in this study.

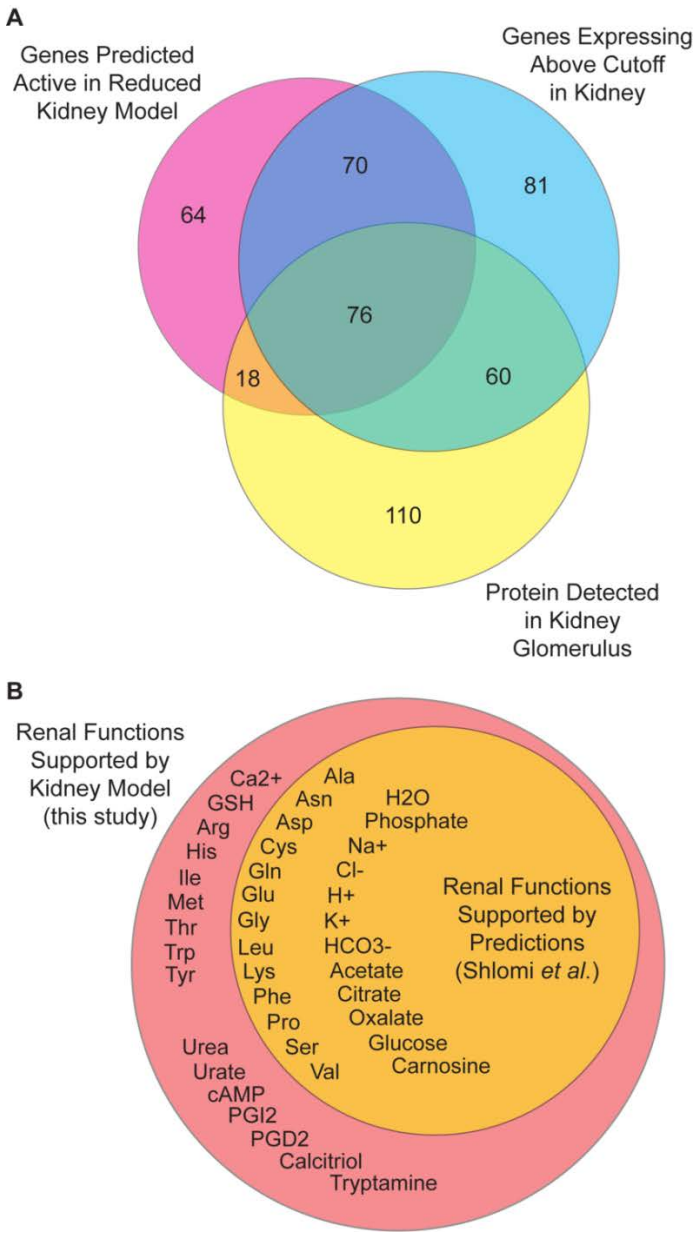

Figure 1.7. Comparative reduced kidney model evaluation. (A) Overlap of gene activity predictions with genes expressing above the significance threshold. Regions of the diagram are approximately proportional to their associated set sizes. The magenta circle represents the set of genes predicted active in the reduced kidney model. The cyan circle represents the set of Recon1-associated genes with expression levels above the significance threshold in the kidney tissue data. The yellow circle represents the set of genes encoding proteins that were detected in normal human kidney glomerulus tissue. **(B)** Renal metabolic objectives supported by predicted reaction flux states. The orange circle represents renal metabolic objectives supported both by the kidney model developed in this study and a kidney model derived from the reaction activity predictions of Shlomi *et al*. The red circle represents renal metabolic objectives supported only by the kidney model from this study. Metabolite abbreviations are defined in Table 1.1.

The literature-curated renal functions achievable by the kidney model were also compared to those achievable by a model derived from the predictions of Shlomi *et al* (Figure 1.7B). While the kidney model developed in this study was compatible with all 41 curated renal functions, the predictions of Shlomi *et al* were only compatible with 25 functions. This difference in functionality was due to false negative inactivity predictions made by Shlomi *et al* such as inactive urea transport, prostaglandin synthesis, and ATP synthesis. These results underscore the need to manually curate automatically generated metabolic network reconstructions and the advantage of integrating objective functions with context-specific modeling.

Next, the model was functionally validated by comparing the gene deficiencies predicted to cause renal disorder to disease phenotypes in the OMIM database collected from clinical studies. Twenty known gene deficiencies leading to specific disease phenotypes were accurately predicted using the model (see Table S4 in [26]). Loss of function mutations in the gene encoding 25-hydroxyvitamin D3-1-alpha hydroxylase (CYP27B1) have been linked to vitamin D-dependent rickets type I in both human patients[32] and pigs[33] consistent with the predicted inability of the gene-deficient model to secrete calcitriol. Hypouricosuria, low urinary excretion of urate, is a symptom of xanthinuria that is caused by xanthine dehydrogenase (XDH) deficiency[34], which is consistent with the deficient model's inability to excrete urate. Similarly, hypouricemia, low blood serum urate, is a consequence of nucleoside phosphorylase (NP) deficiency[35] also predicted in the model. Deficiency of aromatic L-amino acid decarboxylase (DDC) leads to increased urinary excretion of 5-hydroxytryptophan[36], which is consistent with the decreased ability to reabsorb tryptophan and secrete tryptamine predicted through simulation. Mutations in the mitochondrial cytochrome c oxidase gene (COX6B1) lead to de Toni-Fanconi-Debre renal syndrome, whose symptoms include a deficiency in the renal reabsorption of glucose, amino acids, and bicarbonate[37, 38], all of which were predicted in the model. Deficiencies in seven NADH dehydrogenase genes all lead to hypoglycemia, confirmed in simulation, and a decreased ability to oxidize citrate and glutamate[39], reactions important for indirect renal reabsorption of citrate and glutamate in the model. Proline dehydrogenase (PRODH) deficiency causes an inability to oxidize proline in kidney and other tissues leading to hyperprolinemia that includes increased urinary excretion

of proline as a symptom[40-42], which is also consistent with the predicted decrease in renal proline reabsorption. Deficiencies in two genes that take part in the ubiquinol-cytochrome c reductase complex III (UQCRQ and UQCRB) lead to proximal tubulopathy, including an inability to reabsorb amino acids[43]; the gene-deficient model exhibited reduced renal reabsorption of alanine, glutamate, and proline. Fumarate hydratase (FH) deficiency leads to defects in glutamate oxidation in kidney and other tissues[44, 45], which is also consistent with the decreased indirect renal reabsorption of glutamate predicted by the model. Renal glucosuria, recapitulated in the model, results from deficiency in a sodium-glucose transporter (SLC5A2)[46]. Dicarboxylicamino aciduria[47] exhibits impaired renal glutamate and aspartate reabsorption and hypoglycemia resulting from a deficient glutamate transporter (SLC1A1), all symptoms predicted by the model. Severe dehydration is one symptom resulting from another deficient transporter (SLC5A1)[48], confirmed through decreased reabsorption of water in the model. These results qualitatively describe the ability of our modeling approach to predict perturbed phenotypic states.

To more rigorously quantify the predictive ability of our model simulation approach, we performed area under receiver operating characteristic (AROC) analysis based on not only the abovementioned clinical validations of our gene-deficient phenotype predictions but based on the entire set of such known clinical phenotypes that could potentially have been investigated using our model (see Figure 1.8 and Methods). The sharp declines in rates with increasingly stringent classifier ratio thresholds (see Figure 1.8) reflect the likely low coverage of actual disorder phenotypes by existing clinical studies. Nevertheless, our approach performed very well based on this analysis, with an AROC of 0.7565. Permutation trials resulted in a mean AROC of 0.5112, in close agreement with the expected theoretical randomly achievable AROC of 0.5. Our approach achieved a significantly greater AROC than could be expected by chance, p-value = 8.71×10^{-70}. Given the relatively low number of actual clinical negatives available (see Table S5 in [26]), we also assessed the significance of our prediction results based purely on the true positive rates determined through the AROC analysis. The mean true positive rate of our results in this analysis was 0.2859, significantly greater than the 0.0215 mean true positive rate obtained randomly, p-value = 3.29×10^{-127}. These analysis results illustrate that our

approach for predicting perturbation phenotypes exhibits both favorable sensitivity and specificity based on actual clinical data and should hold not only for predicting genetic deficiency phenotypes but also enzyme inhibition by drugs, which exhibits a similarly deleterious phenotypic effect.

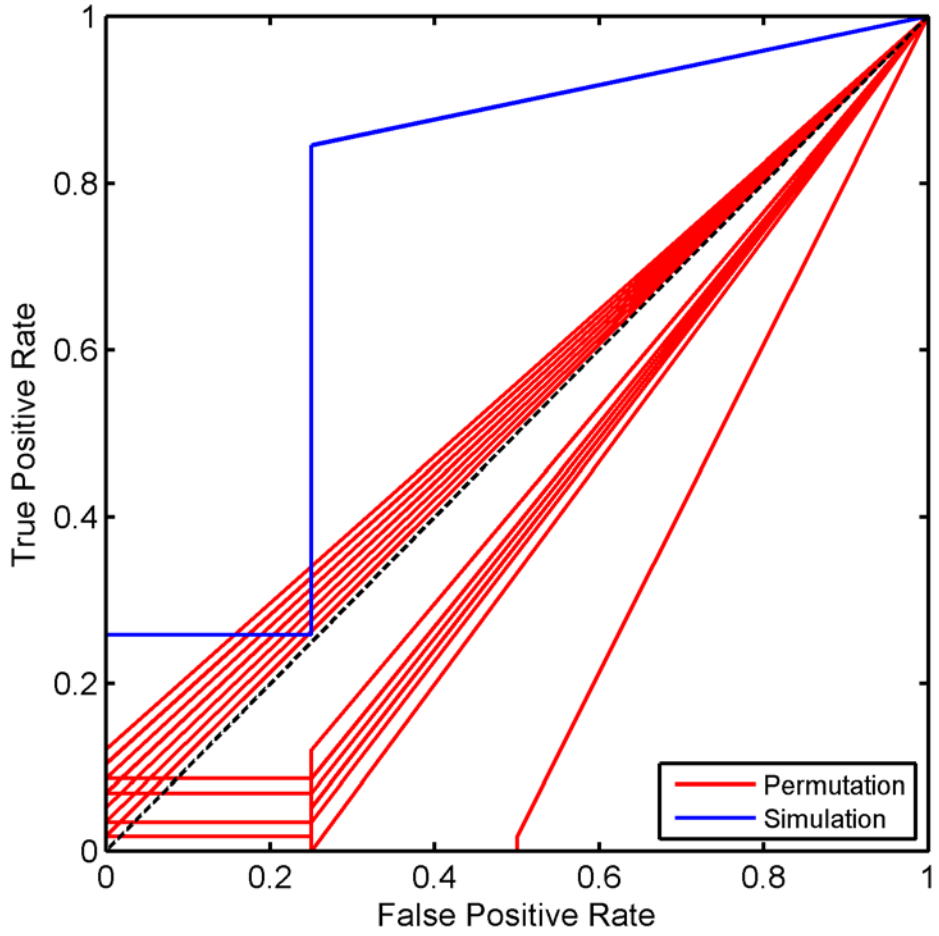

Figure 1.8. ROC curves for gene-deficient phenotype prediction. The blue line represents the analysis of the predictions of the model simulations presented in this study. The red lines represent the analysis of 100 different permutation trials. The dashed black line is the line y=x.

Parameter Sensitivity Analysis

In order to assess the effects of some of the critical assumptions made in the model development and simulation procedures, we performed sensitivity analysis with respect to the predicted renal disorder phenotypes.

First, we compared the predictive capability of our reduced kidney model to that of the original, unconstrained human Recon1 metabolic network. The same approach to simulating renal

disorder states was employed using both models (see Methods). We simulated all single gene knock outs in both models and assessed the renal disorder phenotypes with respect to each individual component of the renal objective function based on the ratio of maximum objective flux in the perturbed state to maximum objective flux in the unperturbed state. Comparing the results achieved by each model (Figure 1.9), it is apparent that although there are a few cases where both models predict an equal degree of renal disorder given the same genetic perturbation, the vast majority of disorder phenotypes are more apparent in the reduced kidney model than in Recon1 alone. In fact, 427 out of the 608 (71%) disorder phenotypes predicted by the reduced kidney model showed no degree of disorder relative to the unperturbed state in Recon1, including 36 of the most severe phenotypes for which a total loss of renal function was predicted by the reduced kidney model. These observations display the predictive ability gained through integration of the gene expression data via the GIMME algorithm, incorporating metabolomics data to set exchange constraints, and the addition of six key membrane transport reactions during the limited function-enabling manual curation of the model. These reactions involve the transport of prostaglandins I2 and H2, calcitriol, and carnosine. It should be noted that the 7 disorders for which Recon1 predicted a more severe phenotype than the kidney model result directly from the addition of these transporters in that these transporters have enabled additional pathways in the kidney model that are absent in Recon1. All but one of the predictions concerning CETP inhibitors showed a clearer phenotype in the kidney model as well; this off-target is PTGIS for which both models predict a complete loss of function when fully inhibited. Finally, 28 out of the 33 clinically validated phenotypes are predicted more noticeably by the kidney model, 17 of these showing no disorder phenotype in Recon1. Overall, this comparison establishes the relative contribution of context-specific modeling in studying disorder and drug response phenotypes.

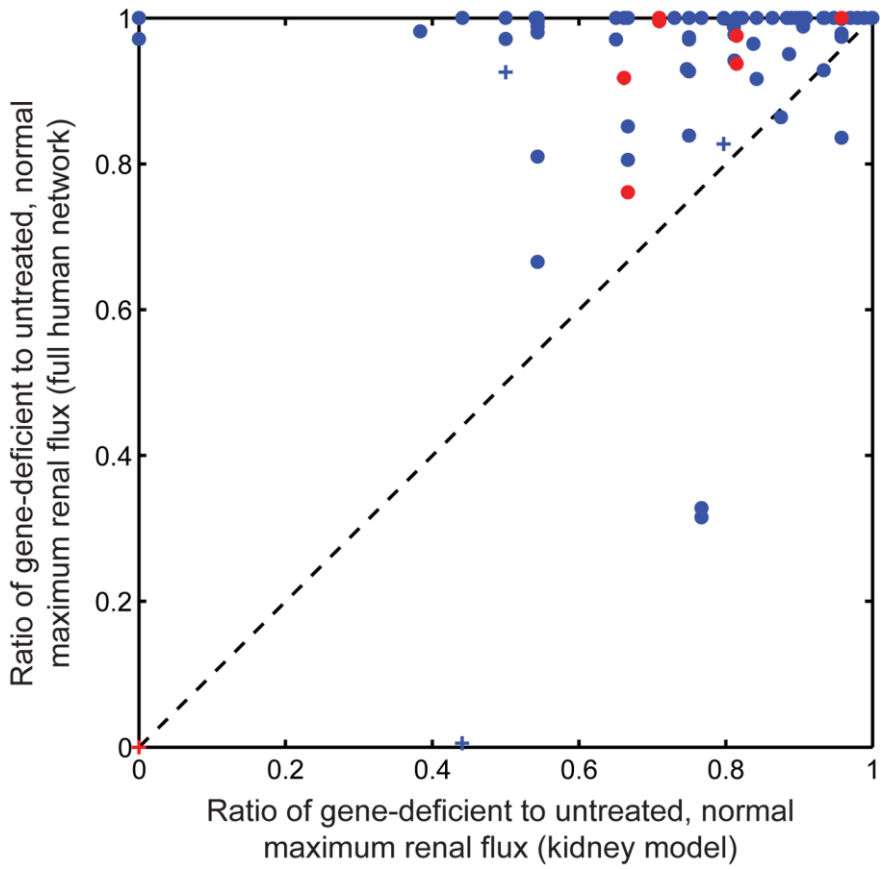

Figure 1.9. Predictive ability gained by modeling. The dotted black line is the line y=x for ease of visual comparison. Red marks represent predictions resulting from inhibition of a predicted CETP inhibitor off-target. Blue marks represent predictions resulting from non-drug-target gene inhibition. Pluses represent predictions validated in the OMIM database. There are 608 marks in total plotted and exact and partial overlap of some marks precludes complete visual resolution.

Second, we investigated the sensitivity of our drug off-target response phenotype predictions to the variability of two important parameters used in our simulations, the system boundary flux constraint, set as equal fractions of the upper bound on renal objective fluxes (see Methods), and the degree of enzymatic activity inhibition assumed to result from drug treatment.

The system boundary flux constraint was imposed upon demand and exchange reactions other than those optimized during a given simulation. By default we set this constraint assuming that all allowed boundary fluxes can carry an equal fraction of the potential maximum renal objective flux. This assumption was made to allow all pathways that could possibly contribute to the objective to be used simultaneously in the optimal flux state, providing the most flexible state while maintaining maximum

sensitivity of our model to additional system perturbations such as gene deficiencies or drug effects. This approach was unbiased in that it did not favor any possible pathway over another in achieving a set objective without imposing additional constraints, which may not always reflect biological reality but was the most conservative assumption in the absence of additional experimental data required to more precisely set these flux constraints. In our sensitivity analysis, we varied this parameter between 0 and 1000 flux units, the absolute lower and upper magnitudes possible in our model, and repeated the simulations of drug off-target effects. The result of this analysis (Figure 1.10) was captured in the normalized sensitivity coefficient computed for each simulation (see Methods). The coefficient can vary between negative and positive unity and displays the deviation from a base result, the primary predictions we have presented in this study. The base result is indicated by a black star in Figure 1.10, and the parameter value in this case equals 13.5 flux units.

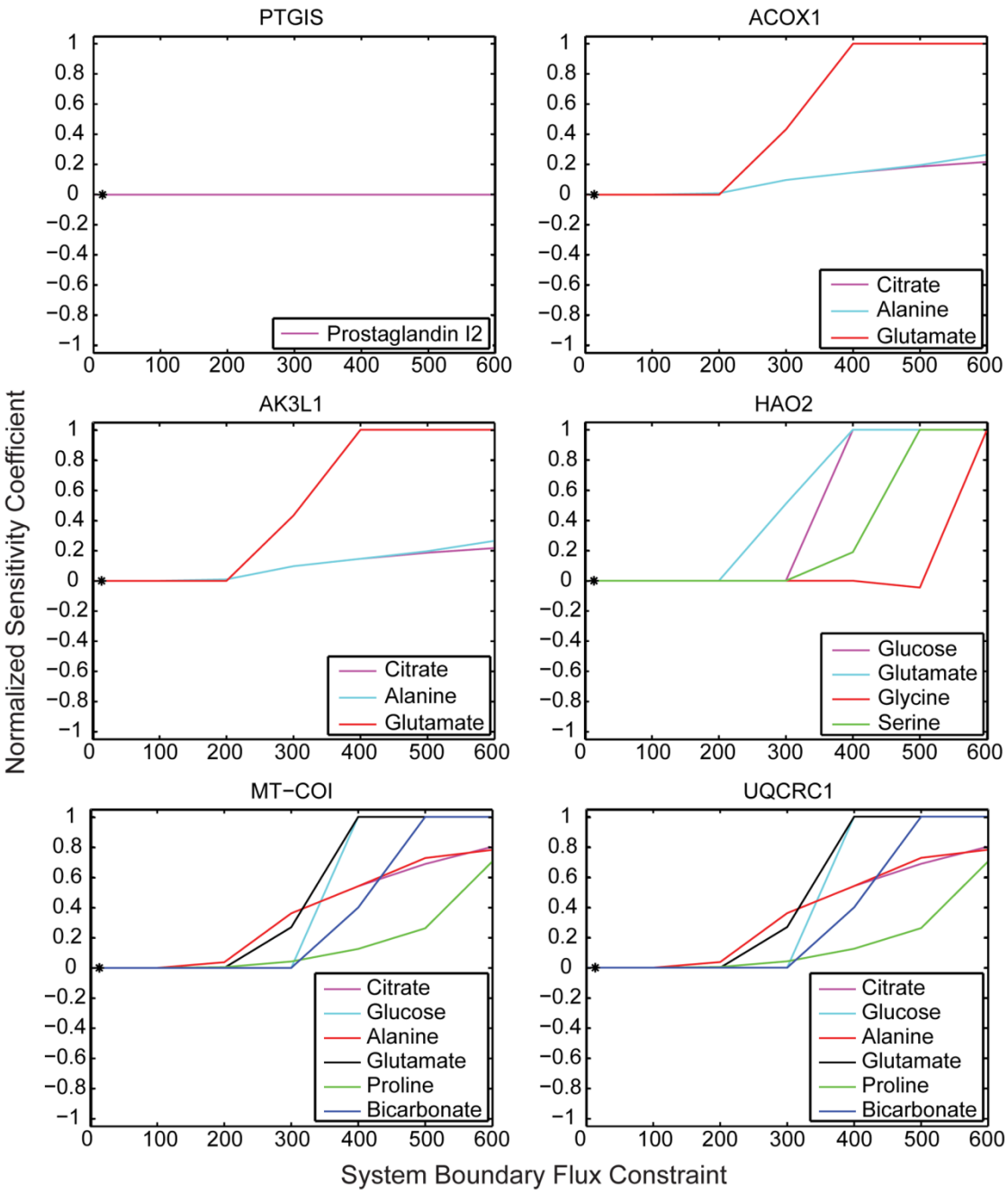

Figure 1.10. System boundary flux constraint sensitivity. Only those drug targets and renal functions are shown for which a deficient phenotype was predicted. The x-axis is in units of flux. The black star represents the base case which is presented as our primary result.

It is clear from Figure 1.10 that PTGIS inhibition resulted in the same renal disorder phenotype regardless of the value of the system boundary flux constraint parameter. This was because there was only one pathway in the model by which prostaglandin I2 could be secreted. Most other disorder phenotype predictions begin to diverge from the base result around a parameter value of 200 flux units, a fairly permissive value, which shows that the predictions were fairly robust to variability of this parameter. The closer to 1000 flux units this parameter was set, the more completely alternative pathways could compensate for a loss of function in the simulations. If alternative pathways existed to achieve a renal function, it was guaranteed that the ability to predict a disorder phenotype with respect to that function would be completely lost at the maximum possible parameter value of 1000.

We similarly analyzed the sensitivity of our predictions to changes in the degree of enzyme inhibition assumed to follow from drug treatment (Figure 1.11). For the primary results presented in this study, we assumed complete inhibition of activity by the drug, corresponding to a fraction of maximum enzymatic reaction flux equal to 0 in Figure 1.11. Similar to the default setting of our system boundary flux constraint, this default of complete inhibition was chosen in order to maximize the sensitivity of our model in detecting disorder phenotypes. Most of the phenotypes were still detectable to varying degrees with as much as 25% of the maximum activity of drug targets. The predicted phenotypes associated with PTGIS, ACOX1, and AK3L1 were especially robust to variation in degree of inhibition, still exhibiting a phenotype near 50% of maximum activity. Decreased glucose and bicarbonate reabsorption under drug-induced MT-COI and UQCRC1 inhibition exhibited the most sensitivity to variability in this parameter, although none of the predicted phenotypes required complete inhibition of the drug target in order to be detected.

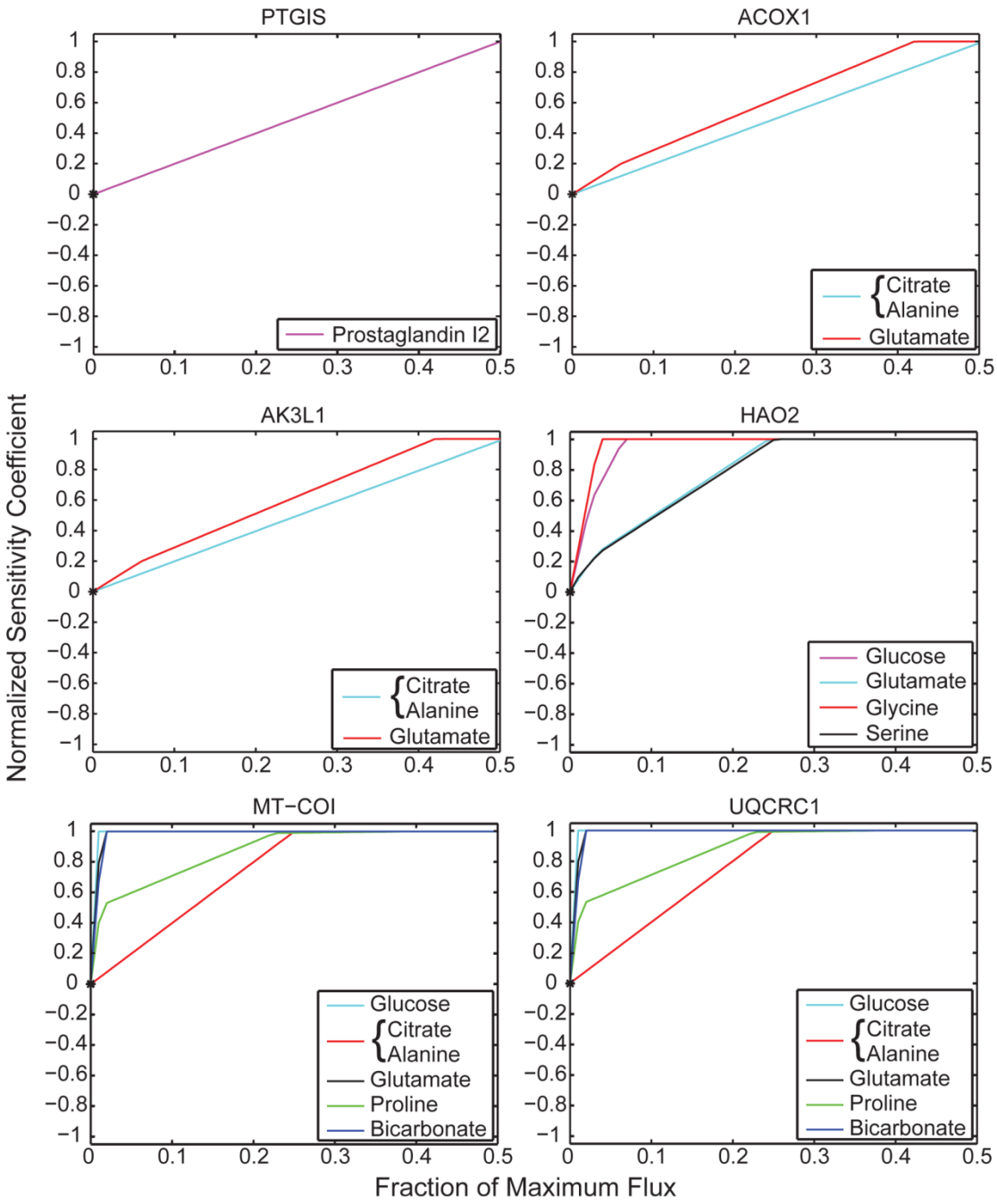

Figure 1.11. Degree of drug-induced inhibition sensitivity. Only those drug targets and renal functions are shown for which a deficient phenotype was predicted. The x-axis values correspond to the fraction of maximal enzymatic flux achievable in the untreated simulation, which represents the constraint placed on associated reactions for each simulation. The black star represents the base case which is presented as our primary result.

Discussion

A novel approach for making functional predictions of drug response phenotypes has been introduced that integrates techniques of both structural bioinformatics and systems biology. Although the current study focused on a specific metabolic system, the general methodology excluding techniques particular to metabolic modeling are extensible to other systems such as signaling or transcriptional regulation. Non-metabolic protein drug off-targets are predictable using the same structural analysis tools, and many such off-targets have indeed been predicted as well for CETP inhibitors[4].

The context-specific organ metabolic modeling strategy employed in this study represents an improvement upon previous efforts in this realm. Model development algorithms such as GIMME[13] or that developed by Shlomi *et al*, when integrated with multiple omics datasets, can lead to more biologically realistic models. It is also of critical importance to include context-specific metabolic objective functions in the model development process in order to yield a fully functional and predictive model, as is evident from the functional comparisons of models performed in this study.

As an early effort at modeling such a context-specific metabolic system it is important to discuss the limitations of our model. Although the functional validations presented here are compelling, currently available clinical data only permits the assessment of a subset of the predictions possible in the model. Also, the functional portion of the model, the reduced kidney model, does not and is not intended to represent a global model of kidney metabolism but only the specific renal functions studied in this work. As such, our model does not fully resolve of complexity of the human kidney. The human kidney fulfills a number of functions not studied here and is a spatially distributed system across multiple distinct tissue types. Here we have summarily replaced the various kidney sub-tissues with a single, net system model. Because we integrated expression data with curated renal functions that operate across multiple kidney tissues, it is likely that our model approximates a superset of the metabolic pathways supporting these functions. Although we have made several simplifying assumptions in the model development process, even the current level of model validation suggests that

the gene and reaction content of the model is fairly accurate and that simulations in this model indeed hold predictive capability.

The simulation approach taken, optimization of a linear objective function, does not fully capture the full physiological role of the kidney. The goal of these simulations was to determine drug-target effects that may limit the capacity of the kidney to move towards a homeostatic nominal state from a state of high blood pressure, thereby decreasing the capacity of the kidney to lower blood pressure. This strategy is appropriate for the goals of the current study but would not be appropriate to simulate all physiological states of interest in the kidney. On a related note, the choice to define a disorder state based on the ratio of perturbed to unperturbed maximum achievable renal objective flux demonstrates a difference in the capacity of the renal function and not necessarily a precise flux state. Therefore this strategy too is not appropriate for modeling all physiological states.

The predictions made for CETP inhibitors in this study serve as illustrative examples of many important implications that this approach has for drug development and personalized medicine. Predicted causal off-targets for renal metabolic disorders related to blood pressure may be responsible in part or full for the clinically observed hypertensive side effect of torcetrapib. The evidence resulting from this study suggests that PTGIS and ACOX1 are both potential causal torcetrapib off-targets, the inhibition of which may explain the side effect of hypertension. In addition, AK3L1, HAO2, MT-COI, and UQCRC1 may also play a role in this side effect as we have predicted, although our docking trials did not suggest that they are bound as strongly by torcetrapib. The specific predicted deficiencies in renal function associated with the drug off-targets can serve as biomarkers to be measured in patients participating in clinical trials. A positive correlation of these biomarkers with side effects would lend support to the predictions of this study and confirm these biomarkers as risk indicators in future patient treatment. It is important to note that although these predictions comprise the basis for a renal filtration and secretion-based hypothesis explaining the hypertensive side effect of torcetrapib, these results do not refute the hypothesis based on a RAAS-mediated mechanism. These two hypotheses are not mutually exclusive and could potentially contribute alternatively or synergistically to the clinically

observed side effects. This possibility illustrates the major tenet for systems biology: studying a single protein or even a single pathway is not necessarily sufficient to explain complex biological phenomena.

Aside from the confirmation that some of our predicted off-targets are known to be involved in renal disorders, we do not currently present direct experimental verification that torcetrapib binds and inhibits the predicted targets and that this inhibition leads to the predicted response phenotypes. Although this would be the obvious next step, a retrospective validation is currently hampered by the availability of the drug and the nature of the phenotypes both predicted and known. Ideally, relevant physiological studies would be carried out during actual clinical trials, when a method such as ours would be most useful, in preclinical and clinical phases of drug development.

The extended structural analysis of causal drug off-targets to identify differential binding affinities for endogenous substrates and drug molecules suggests possible differences in drug response phenotypes across the CETP inhibitors tested. The results suggest that anacetrapib may potentially lead to a similar response phenotype to that of torcetrapib, while JTT-705 may not carry the same adverse effect, at least with respect to the off-targets detailed in this study. This particular type of analysis may aid in differentiating between likely response phenotypes expected for chemically and functionally similar drugs. Results of the computational pipeline for interaction prediction between proteins and CETP inhibitors employed in this study, SMAP and docking, have yet to be confirmed experimentally. Although we are currently unable to provide direct experimental evidence for the off-target interaction predictions for this class of drugs, multiple recent studies have shown experimental support for the general efficacy of this approach for interaction prediction[49, 50].

The predicted renal metabolic disorders with a genetic basis suggest classes of individuals in which treatment with CETP inhibitors may pose a higher risk for adverse side effects. These predictions suggest a likely relationship between participants in torcetrapib clinical trials exhibiting symptoms of these disorders and the observed adverse side effects. The concept of cryptic genetic risk factors for drug treatment introduced in this study suggests a novel approach to personalized medicine. Should polymorphisms within these genes be clinically linked to side effects of drug treatment, the result would comprise a basis for genetic screening to assess the risk of drug treatment for future patients. Given that

these cryptic risk factors are not expected to elicit the predicted abnormal phenotypes in the absence of drug treatment, identification of causal polymorphisms through association studies could only occur during clinical phase when a sufficient number of patients could be observed to gain the statistical power needed to draw significant correlations.

As illustrated above, this approach for *in silico* drug testing could become an indispensible tool during the pre-clinical and clinical phases of new drug development for studying the nature of adverse side effects. In addition, this platform holds obvious potential for analyzing drug efficacy in general and identification of potential beneficent drug side effects that may be useful for drug repositioning and could also be easily adapted for studying combinatorial drug treatment. For a failed drug like torcetrapib, results from this approach could reinitiate the drug development process, providing new insight to help target patients who could benefit from the treatment without the risk of serious adverse side effects.

Methods

Prediction of CETP Inhibitor Drug Off-Targets

The binding site for CETP inhibitors on the CETP structure and the predicted off-target binding sites for this class of drug across the proteome were assumed to be as previously predicted using the SMAP program[4], which implements the Sequence Order Independent Profile-Profile Alignment (SOIPPA) algorithm to identify significant structural similarity to a given ligand-binding site[3]. The results contained proteins from all organisms represented in the PDB, not just human structures.

Mapping Off-Target Proteins to the Metabolic Network

In order to integrate the result of drug off-target predictions with the metabolic network, it was necessary to first map all PDB structures (http://www.pdb.org) corresponding to human metabolic proteins included in Recon1, downloaded from the BiGG database, to their respective gene identifiers as represented in Recon1. The BiGG database requires registration and a password, which can be requested by visiting (http://bigg.ucsd.edu/bigg/home.pl). The UniProt ID mapping tool

(http://www.uniprot.org/) was used to map PDB structures corresponding to human proteins to gene identifiers linked to metabolic reactions in Recon1 accounting for all predicted human metabolic protein drug off-targets. All non-human predicted metabolic protein drug off-targets were mapped to their human orthologs using the Basic Local Alignment Search Tool (BLAST)[51] to perform a bi-directional BLAST with a mutual best hit criterion. BLAST was also used to resolve inconsistencies in functional annotation between Recon1 gene-protein-reaction associations (GPRs) and gene annotations from the Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) with respect to predicted drug targets, leading to the reannotation of three Recon 1 GPRs. The overall result of this mapping was that 97 metabolic reactions in Recon1 were linked to 41 predicted CETP inhibitor off-targets.

Enzyme Inhibition Analysis

The metabolic enzymes predicted as CETP inhibitor off-targets using SMAP were evaluated to determine potential enzymatic inhibition by the drug. The predicted drug-binding sites of the putative off-targets were compared to endogenous ligand-binding sites from existing PDB protein-ligand complex structures (http://www.pdb.org) and catalytic sites from the Catalytic Site Atlas (http://www.ebi.ac.uk/thornton-srv/databases/CSA/). Ligand-binding sites were defined as amino acid residues lying within 4.5 Å from atoms of the ligand. Drug-binding sites were defined as residues aligned with the cholesteryl ester binding sites on the CETP structure using SMAP. Overlap between endogenous ligand-binding sites and drug-binding sites was defined by a sharing of any amino acid residues between the sites. The function of predicted drug targets present in Recon1 with at least a partial such overlap was considered to be competitively inhibitable by the drug.

Protein-Ligand Docking

Enzyme substrates were identified from Recon1 reaction formulas. Certain molecules (H^+, H_2O, O_2, phosphate, ferricytochrome C, and ferrocytochrome C) were excluded from docking trials due to size or structural challenges prohibiting a useful docking result for the purposes of binding affinity predictions. All protein structures used in this study were downloaded from the PDB (http://www.pdb.org). Three-dimensional structures for endogenous enzyme substrates were

downloaded directly from the PDB if available. If the PDB ligand structure did not exist or was non-functional for docking, the structure was searched for in PubChem (http://pubchem.ncbi.nlm.nih.gov/). The subsequently downloaded SDF file was converted to PDB format using the ChemAxon web applet available at the PDB website (http://www.rcsb.org/pdb/ligand/chemAdvSearch.do). If the three-dimensional ligand structure could not be found in PubChem, the two-dimensional structure was derived from the canonical SMILES[52] representation of the compound available in PubChem and then converted to a three-dimensional structure in PDB format using the Clean3D Fine Build tool available through the Marvin web applet (http://www.chemaxon.com/marvin/sketch/index.jsp). The three-dimensional structures for glycolipids were derived from their KEGG glycan structures (http://www.genome.jp/kegg/glycan/) using SWEET-II (http://www.glycosciences.de/spec/sweet2/doc/index.php).

Protein structures were pre-processed for docking using AutoDockTools (ADT) version 1.5.2[53] by adding polar hydrogen atoms, removing all non-protein molecules from the PDB structure including water, detergents, and ligands, adding Kollman charges to the protein and converting it to PDBQT format. Ligand structures were also prepared using ADT, using the default method for preparing ligands for docking that adds hydrogens and charges. The default rotatable bonds were accepted as well, and the structure was converted to PDBQT format. The search space for docking was determined visually by centering the Grid Box in ADT central to the experimentally determined binding site of the ligand and expanding the dimensions of the cubic search space to just completely encompass the site.

Docking was performed using AutoDock Vina[54] with default parameter settings other than the search space specification described above, and the mean predicted binding affinity from the set of predicted binding poses was accepted as the true binding affinity for each docking run. The predicted binding affinities for endogenous substrates were compared to the affinity of the same site for the CETP inhibitor drugs in order to make predictions about differential responses with respect to each of the drugs.

Renal Objective Function

As the preliminary step in modeling human renal function, the scientific literature was reviewed to compile a list of specific metabolic functions of the kidney, with a focus on functions implicated as determinants of blood pressure. This list includes a number of renal reabsorptions and secretions. Each function in this list was tested for compatibility with Recon1, downloaded from the BiGG database (http://bigg.ucsd.edu/bigg/home.pl), by performing flux balance analysis (FBA) on the fully unconstrained network optimizing for the given function. Those functions compatible with Recon1 were those that could achieve a positive flux and are summarized in Table 1.1. These metabolic functions were combined with a basic ATP maintenance function to form a single model reaction that represents the kidney's ability to filter the metabolic content of blood with preference for lowering blood pressure. This model reaction was used as the objective function in developing the metabolic kidney model and is referred to as the renal objective function in this study. All stoichiometric coefficients in this reaction were set equal to one, which is a safe assumption for the model development step as this only significantly impacts the magnitude of fluxes through pathways that support each individual renal objective and not generally whether or not certain fluxes will be active in the resulting model. For the full renal objective function reaction to be seen as useful in performing simulations, more careful balancing of these coefficients based on experimental evidence would be required. As such, the full renal objective function was not used in any subsequent simulations with the model, instead being substituted as an objective by the reactions representing individual reabsorptions or secretions.

Metabolite exchange and transport reactions needed to achieve some of the renal functions were also added to the network. It was observed that Recon1 as a base model could not achieve flux through certain key renal metabolite reabsorptions: sodium, calcium, chloride, potassium, and oxalate. These deficiencies were corrected for by simply adding demand fluxes for these metabolites in the cytosol model compartment. Demand fluxes were also added for the remaining kidney reabsorptions and secretions as well to enable an array of simulations involving individual components of the renal objective function to be tested. In the case of reabsorption, this allows for direct reabsorption of

metabolites in addition to indirect reabsorption in which the absorbed metabolite is first metabolized into other compounds and then reabsorbed into the blood, as is the primary mechanism of reabsorption for some metabolites, such as reduced glutathione (GSH)[55].

Kidney Metabolite Exchange Flux Constraints

A preliminary model was created by imposing kidney-specific exchange flux constraints representing the metabolic exchanges the kidney carries out with the blood and urine. The preliminary model was initialized by loading Recon1 into the COBRA Toolbox[56] and, by default, unbounding all reaction fluxes by setting them to the default maximum magnitude of 1000 flux units. Next, the renal objective function was added to the network as a single reaction. Exchange fluxes for kidney secretion objectives were constrained to preclude uptake of those metabolites to achieve the renal objective, forcing the model to synthesize those metabolites in order to secrete them. The resulting preliminary model included 407 exchange fluxes, only 49 of which were explicitly unconstrained based on literature-curated kidney functions and the most basic of metabolic precursor requirements. The basic metabolic exchanges assumed to take place include ions and other inorganic compounds.

The Human Metabolomics Database (HMDB) (http://www.hmdb.ca/) was queried to derive further evidence in support of allowable exchange fluxes for the kidney. All 407 exchange metabolites in the preliminary model were searched in HMDB for experimental detection in specific biofluids and tissues. Those metabolites detected both in the blood and kidney tissue were assumed to be freely exchangeable in the kidney model, leading to 78 more explicitly unconstrained exchanges beyond what was derived from basic and curated kidney-specific metabolic functions. This assumption is based on the kidney's physiological role of filtering the blood and the observation that if both the blood and kidney contain a metabolite, it must either be exchanged between the two or synthesized separately in both. In the former case, this data provides evidence of that exchange. In the latter case, although the model might allow a biologically unrealistic exchange, because the metabolite exists in both blood and kidney, the impact on simulations using the resulting model should be merely quantitative in terms of the maximum renal objective fluxes achievable by the unperturbed model. The integration of gene expression data in the model development process described below should reduce the propensity for

biologically unsound metabolic pathway activation that could follow from precursors introduced by any biologically unsound exchanges. Those metabolites detected both in the urine and kidney were assumed to be possible excretions, and exchange constraints were set accordingly. Excretions determined utilizing the urine metabolomics data mostly showed redundancy in determining exchange constraints with exchanges determined using blood data or literature curation with the exception of 4 additional metabolites. The remaining 276 exchange fluxes for which no evidence was found to support were tentatively constrained to 0 flux units.

The resulting preliminary model was again tested for the ability to achieve all kidney-specific metabolic functions. It was found that this model could not absorb and metabolize GSH, without also absorbing oxidized glutathione, the exchange of which was subsequently unconstrained. Also, L-threonine and L-methionine could not be absorbed and metabolized in this model without exchange of 2-hydroxybutyrate and 2-methylcitrate, the exchanges of which were similarly unconstrained as a corrective measure. The resulting preliminary model could still achieve all the same renal objectives as the fully unconstrained model. As a final preliminary constraining measure, all system effluxes were bound to equal fractions of the default upper bound on influxes of 1000 flux units; we term this parameter the system boundary flux constraint. This was done so that any available direct or indirect reabsorption pathways could possibly be used to achieve metabolite reabsorption without biasing the model towards use of any particular pathways without further evidence. This represents the state of the model just prior to final processing using the GIMME algorithm. The fitting of the allowable fluxes to the gene expression data by GIMME ultimately determined the usable reabsorption and secretion pathways in accordance with gene expression.

Gene Expression Microarray Data Processing

Two gene expression microarray dataset for normal, healthy kidney tissue[57] were obtained from the GEO database (http://www.ncbi.nlm.nih.gov/geo/), accession GSE803. The two background-subtracted datasets were first normalized using a global normalization factor equal to the sum of probe intensities from the first dataset divided by the sum of probe intensities from the second dataset to account for any systematic differences in procedure between the two experiments. The resulting data

were then normalized using the Lowess method[58] to reduce random noise. The resulting normalized datasets were then weighted equally as replicates in determining the final data for integration with the human metabolic network by taking the mean of the two normalized datasets.

The gene-protein-reaction associations (GPRs) in Recon1 use Entrez Gene IDs to annotate reactions in the network. To map the data from the AffyHG-U95 chips to Recon1, all genes included in Recon1 were mapped to corresponding AffyHG-U95 probesets using Bioconductor[59] and the most recent chip annotations[60]. A single expression value was then assigned to each gene in Recon1 based on the maximum normalized data value associated with any of the probesets mapped to a given gene. Next, a single expression value was assigned to each reaction in Recon1 by evaluating the Boolean rules in the GPRs with respect to the normalized expression data. The minimum data point was chosen for genes linked by an AND operator in a GPR, and the maximum data point was chosen for genes linked by an OR operator in a GPR.

Finally, a significant expression threshold was established for subsequent use in the GIMME algorithm. This was done by fitting the normalized gene expression data to a Gaussian distribution, estimating the mean and standard deviation of this distribution, and calculating p-values associated with each data point by subtracting the cumulative distribution function from one. The normalized data value corresponding to the p-value closest to but not exceeding 0.05 was chosen as the significance threshold; this resulted in a threshold of 991.3698 for the normalized expression data.

Implementation of GIMME to Obtain Metabolic Kidney Model

To integrate the renal objective function and kidney gene expression data with the preliminary model to derive a functional kidney model, the GIMME algorithm[13] was implemented. The GIMME algorithm takes a metabolic network model, a gene expression dataset, and specified required metabolic functions as input and solves a linear programming optimization to yield the network flux activity state that maximizes the specified functions while remaining as consistent as possible with the gene expression data. The complete renal objective function, the combination of all functions presented in Table 1.1, was set as the metabolic objective with a minimum requirement of 90% of the maximum possible flux set as a parameter for GIMME in determining the final kidney model. The reaction

expression threshold parameter was set as described above. GIMME was run with these parameters and the normalized expression data and preliminary model as inputs. The resulting reaction activity predictions were used to constrain metabolic reactions yielding the full kidney model. Subsequently, the connected sub-graph of this full kidney model, which includes all functioning reactions possible for achieving the renal objectives, was isolated and is this portion of the model we focused on for validation and simulation. We refer to this sub-model as the reduced kidney model (available in SBML format as Dataset S1 in [26]).

Validation of Kidney Model Content

Gene activity predictions made when deriving the metabolic kidney model were compared to the set of expressed genes with normalized expression values above the significance threshold described above. Activity predictions were also validated against a comprehensive proteomics dataset from normal human kidney glomerulus tissue[31] for overlap with network-associated proteins detected with high confidence, that is, identified through detection of two or more peptides.

To evaluate the modeling approach used in this study, a five-fold cross validation was performed in which the data corresponding to the most highly expressed 20% of network-associated genes were excluded before deriving the kidney model. The ability of each approach to correctly predict the activity of these most highly expressed 20% of genes was measured from the overlap of predictions with the highly expressed gene set assuming a hypergeometric distribution, and the resulting probability was Bonferroni-adjusted.

Simulating Drug Target Effects and Renal Metabolic Disorders

All predicted metabolic protein drug off-targets were tested in the kidney model to assess the drug response phenotype caused by inhibitory effects in this system. Inhibition of metabolic proteins by the drug was modeled by constraining corresponding reactions catalyzed by drug targets to 0 flux units. Simulations of the consequences of these drug effects were performed using FBA as implemented in the COBRA Toolbox[56] in the MATLAB programming environment. Each drug target was evaluated with respect to its impact on each individual renal function to determine if target inhibition by the drug leads

to a renal deficiency relative to the untreated normal kidney model. This was done by optimizing single exchange or demand fluxes at a time, representing reabsorptions and secretions respectively, out of the full set listed in Table 1.1. The cumulative effect of all predicted drug targets being simultaneously inhibited was also tested against each individual renal function. Renal secretion fluxes were maximized in simulation. Renal reabsorption fluxes were set as unbounded and then maximized while the remainder of allowable uptakes were constrained to equal fractions of the default maximum magnitude of 1000 flux units. The additional constraints were imposed for reabsorption simulations in order to allow the resulting network flux state to include concurrently active alternative optimal direct and indirect reabsorption pathways rather than having to identify alternative optimal pathways by performing multiple simulations.

Single gene deficiencies were also simulated in the kidney model to assess their effects on renal function and their potential as risk factors for treatment with CETP inhibitors. Each of the genes annotated to reactions in Recon1 was knocked-out of the kidney model and simulations were run using the gene-deficient kidney model both with and without drug treatment to assess effects on each individual renal reabsorption and secretion.

Drug response and metabolic disorder phenotypes were assessed by taking the ratio of maximum gene-deficient, untreated renal function flux to maximum normal, untreated renal function flux. A ratio of less than unity indicates a deleterious phenotype. Predicted metabolic disorder phenotypes were validated against previous clinical studies as represented in the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim/).

Cryptic genetic risk factors for drug treatment were also predicted for which the maximum gene-deficient, untreated renal objective flux equals the maximum normal, untreated renal objective flux but the ratio of maximum gene-deficient, drug-treated renal objective flux to maximum normal, drug-treated renal objective flux is less than unity.

Parameter Sensitivity Analysis

Sensitivity of our prediction approach to variability in parameters was performed through repeated simulation in which we varied the parameter value across the full range of possible values. We

investigated sensitivity with respect to each parameter independently. A normalized sensitivity coefficient was calculated as the result of each of these simulations. This coefficient was calculated by first taking the percent difference in the predicted outcome relative to a base case, our primary results, and then dividing it by the maximum possible percent difference.

Area under Receiver Operating Characteristic (AROC) Analysis

Benchmark data was collected from the OMIM database (http://www.ncbi.nlm.nih.gov/omim/) by searching for all metabolic disorders related to renal reabsorptions or secretions that are associated with deficiencies in genes included in the reduced kidney model. The resulting list of disorders was manually curated using literature references to identify precisely which metabolic renal reabsorptions and secretions were impacted. These included not only those renal functions captured in Table 1.1, but also other renal exchanges. All resulting reabsorptions and secretions that can have corresponding non-zero fluxes under unperturbed conditions in the reduced kidney model were included in our benchmark data set (see Table S5 in [26]). Every phenotype in the benchmark data was investigated through our model as described for simulating drug target effects and renal metabolic disorders, taking the ratio of perturbed to unperturbed flux capacities as a measure of phenotype, where a ratio of one signifies no disorder phenotype and a ratio of less than one signifies some degree of disorder. Next, the ratio threshold for classifying normal versus disorder phenotype was iteratively set to assess the sensitivity and specificity of our approach for predicting true and false positives across the full range from zero to one. Note that a threshold of one was used by default for the main results presented in this study. The true positive rate was plotted against the false positive rate (see Figure 1.8), the ROC curve, and the AROC was computed using the trapezoidal rule for approximating definite integrals. The statistical significance of our result was determined by comparison to 100 permutation trials in which all reaction flux ratios, perturbed to unperturbed, were randomly shuffled for each simulated gene deficiency and AROC-analyzed. The permutation trials exhibited true positive and false negative rates expected for random disorder phenotype classification (see Figure 1.8), and thus comprised an appropriate assessment of the predictive ability of our model simulation approach relative to chance. One-sample left-tailed student t-tests were performed using an alpha value of 0.05 to assess the statistical

significance of the AROC and mean true positive rate achieved by our model simulation approach relative to the permutation results.

Chapter 1 is a modified version of material in 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 Sep 23;6(9):e1000938. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Chapter 2: Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*

Abstract

Systems biology of metabolism has been developed using genome scale network reconstruction. Traditionally, protein structural information has not been represented in such reconstructions. Experimental and computed protein structures were used to achieve 93% coverage of enzymes in the *Escherichia coli* K-12 MG1655 metabolic network. This expanded reconstruction enabled the analysis of protein thermostability in a network context, leading to prediction of protein activities that limit network function at super-optimal temperature (i.e. network hot spots) and mechanistic interpretations of mutations found in strains adapted to heat. Predicted growth-limiting factors for thermotolerance were validated through supplementation experiments, leading to the discovery of previously unknown metabolic sensitivities to heat stress and providing new evidence that metabolic enzyme thermostability is rate limiting at super-optimal temperature, as represented by specific enzymes. This study thus notably expands the content and predictive capability of genome-scale metabolic networks enabling structural systems biology of metabolism.

Introduction

The dependence of cellular thermosensitivity on proteome stability has long been known, first highlighted by the presence of many chaperones and proteases among well-characterized heat shock

proteins (HSPs)[61] and supported by evidence that chemical chaperones improve survival at super-optimal temperatures[62]. Both proper folding and structural stability required for function are disrupted at sufficiently high temperature. Many individual proteins and their mutant variants have been studied to identify structural hot spots[63], loci within a protein that are destabilized at sufficiently high temperatures leading to thermal denaturation. Replacing such structural hot spots with more stabilizing residues has proven a successful strategy to engineering thermostable proteins[64]. Just as structural hot spots provide a molecular basis for understanding individual protein thermostability, identification of the proteins that comprise analogous hot spots at the level of the cellular system, referred to as network hot spots in this report, is critical to uncovering the molecular mechanisms for cellular thermosensitivity. Thus far, strategies for increasing thermotolerance have been limited to indirect measures like introducing chemical chaperones, overexpressing known HSPs, pretreatment with more moderate heat, or random mutagenesis to evolve desirable stress tolerance[65], never directly identifying the actual points of thermosensitivity in the system, the network hot spots.

The emerging discipline of structural systems biology[66] has enabled new insights into a variety of scientific topics including the structure-function relationship in the metabolism in a hyperthermophile[67], identifying causal drug off-targets for an adverse side effect[26], identification of protein-protein interactions[68, 69], and determining causal missense mutations for disease susceptibility[69, 70]. In this study, a structural systems biology approach was taken to uncover network hot spots in the metabolism of the mesophilic bacterium *Escherichia coli* K-12 MG1655, assessing metabolic thermosensitivity as a function of protein thermostabilities as they affect enzyme activity in a genome-scale model (GEM), providing mechanistic explanations for recently reported mutations in evolved thermotolerant strains[71, 72] and leading to the discovery of previously unknown metabolic limitations to thermotolerance.

Results

In order to assess protein thermostability, a genome-scale model integrated with protein structures (GEM-PRO) (see Dataset S1 in [73]), exemplified for one reaction in Figure 2.1A, was

generated defined by the scope of single-peptide chains in the *E. coli* metabolic network (*i*JO1366)[74]. The four main objectives of this reconstruction effort were to: 1) best-represent the experimentally-measured native structure of each wild type (WT) protein, 2) maximally cover amino acid sequence length, 3) best-represent the functional conformation or induced fit caused by each protein-substrate interaction, and 4) map existing amino-acid-residue-resolution functional annotation to all structures[75-78]. Thus, in this GEM-PRO a protein may be represented by no, one, or multiple separate structures. Extensive curation of experimentally-measured structures[79] and homology modeling to compute structures were carried out (Figure 2.2) to achieve 93% coverage of every protein by at least 1 structure (Figure 2.1B) and between 24% and 33% coverage of protein-substrate-pair induced fit (Figure 2.1C). Notably, the majority of coverage was enabled by established structure modeling techniques[67], without which such a reconstruction would not currently be possible.

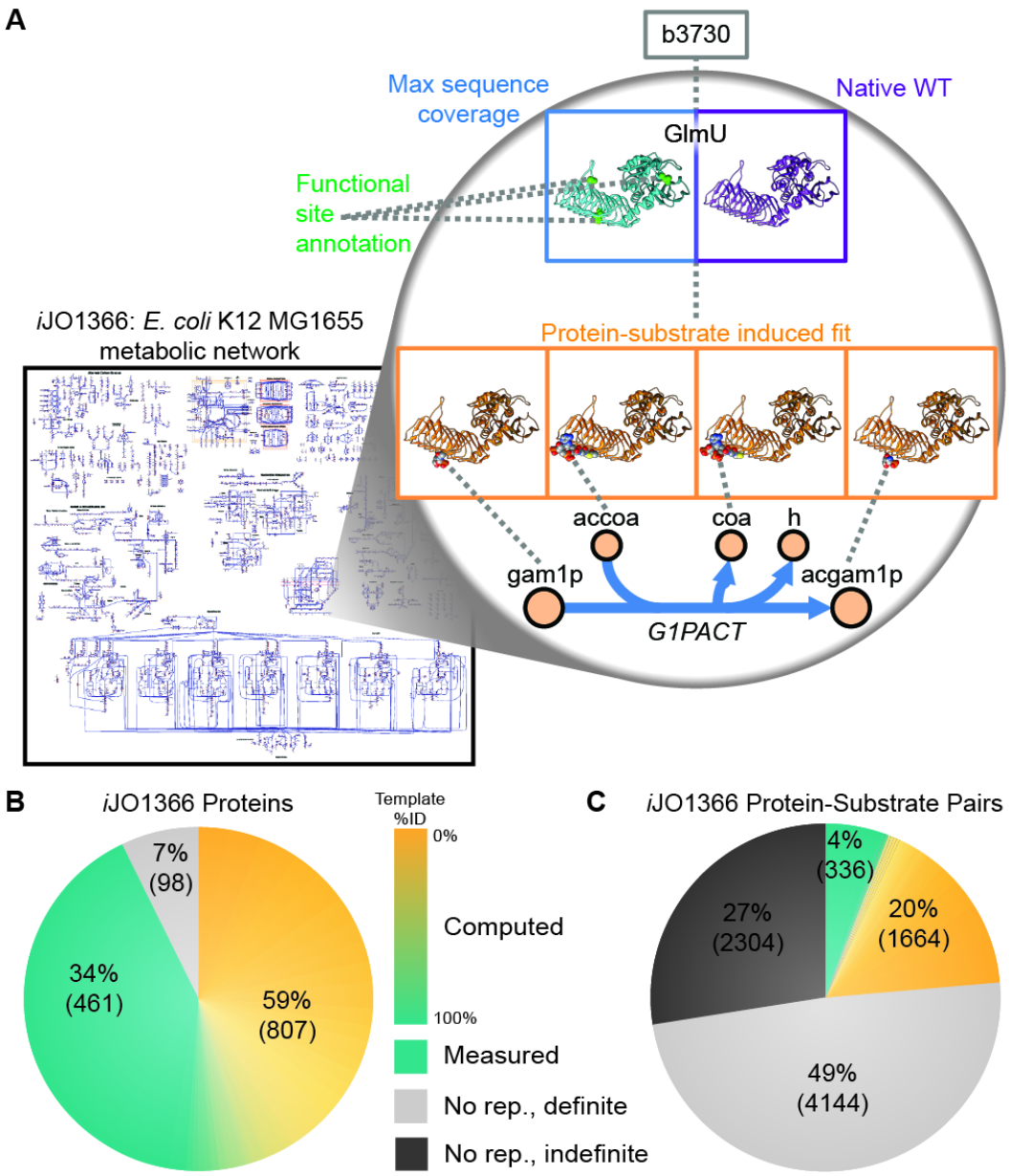

Figure 2.1. The *E. coli* GEM-PRO. (A) The *E. coli* GEM-PRO provides native WT structures, maximal sequence length coverage, protein-substrate induced fit, and functional annotation for proteins included in *i*JO1366. A conceptual illustration is depicted for the GlmU protein catalyzing the "G1PACT" reaction. **(B)** Coverage of each protein by at least one structure is categorized whether measured experimentally or computed by homology modeling. If a protein is covered by more than one structure, the structure with greatest sequence identity to the reference sequence is represented. The color gradient represents the percent sequence identity between the protein and the template structure used in homology modeling. **(C)** Coverage of induced fit counted as protein-substrate pairs is similarly categorized. The metabolites H^+, H_2O, and H_2 are excluded from the pair count. Proteins and induced fits with no structural representative but definitely known to exist are shown in light gray, and those for which knowledge is indefinite are shown in dark gray.

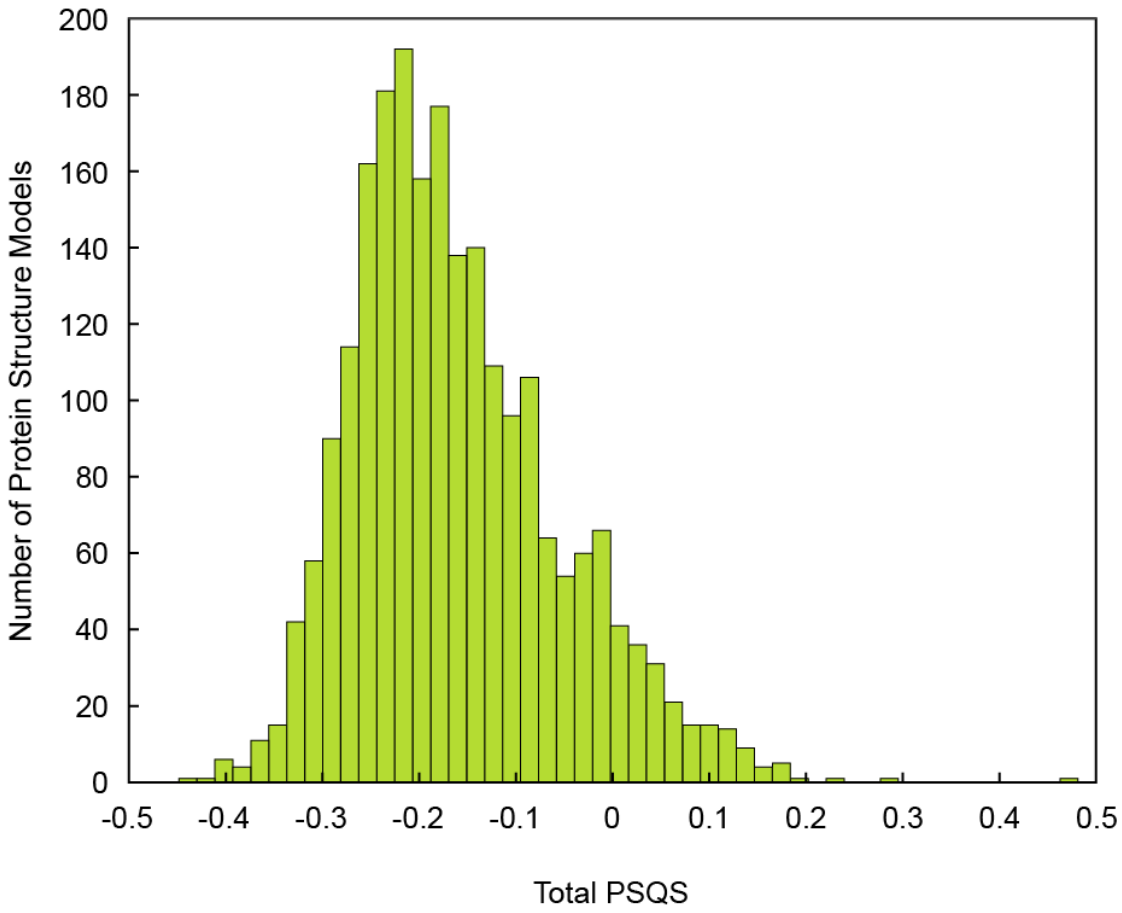

Figure 2.2. PSQS scores for protein structures generated through homology modeling. The distribution of the sum of local, burial, and contact PSQS scores for modeled protein structures is depicted. A negative PSQS score indicates higher quality of a model. The mean of total PSQS scores across all structures is -0.16.

Experimentally-measured critical temperatures[80, 81] for *E. coli* protein activities accounting for optimal, half-maximum, and total loss of activity were supplemented with bioinformatic predictions of protein melting temperatures[82-85] (Figure 2.3 and see Table S1 in [73]) using the GEM-PRO. Protein activity functions were defined by these critical temperatures to impose temperature-dependent constraints on the metabolic model (Figure 2.4), comprising a basis for genome-scale metabolic simulation with growth temperature as a parameter affecting protein function for the first time to our knowledge. Simulated temperature-dependent growth showed good qualitative agreement with experimental growth data using three different nutrient media (Figure 2.5A), especially in the range from 32°C to 43°C where growth is above 50% of the maximum. This result provides new evidence

that the thermostability of metabolic proteins alone could suffice to explain thermosensitivity within this temperature range.

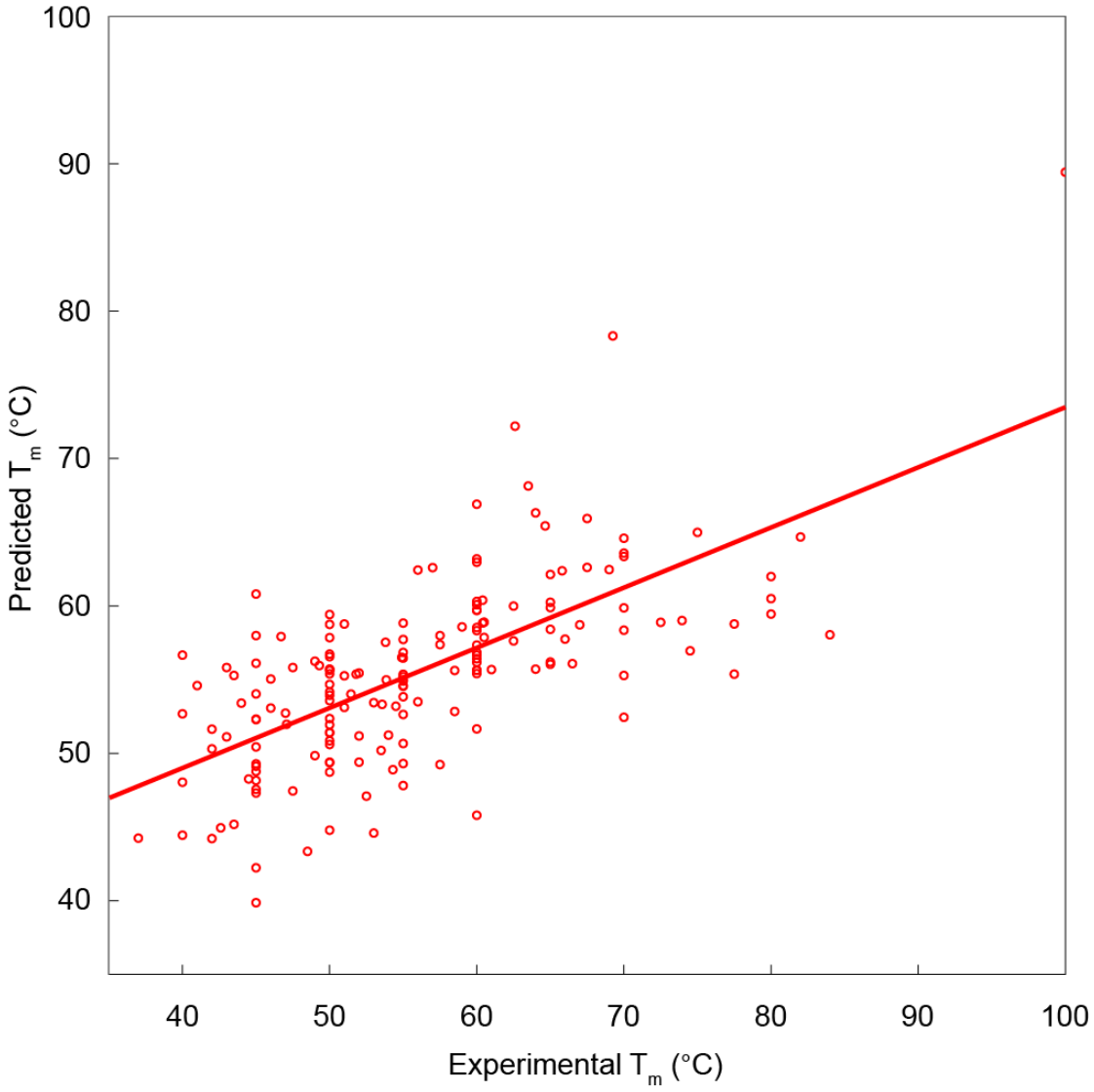

Figure 2.3. Correlation between experimentally-measured and composite predicted T_m values. Correlation between experimentally-measured and composite predicted T_m values. Data are shown for the 172 proteins in the training set for the composite prediction pipeline. The Pearson correlation coefficient is $\rho = 0.69$ (p-value = 6.55×10^{24}), and the red line represents the best fit line with a slope equal to the correlation coefficient. The axes have equal limits and scale.

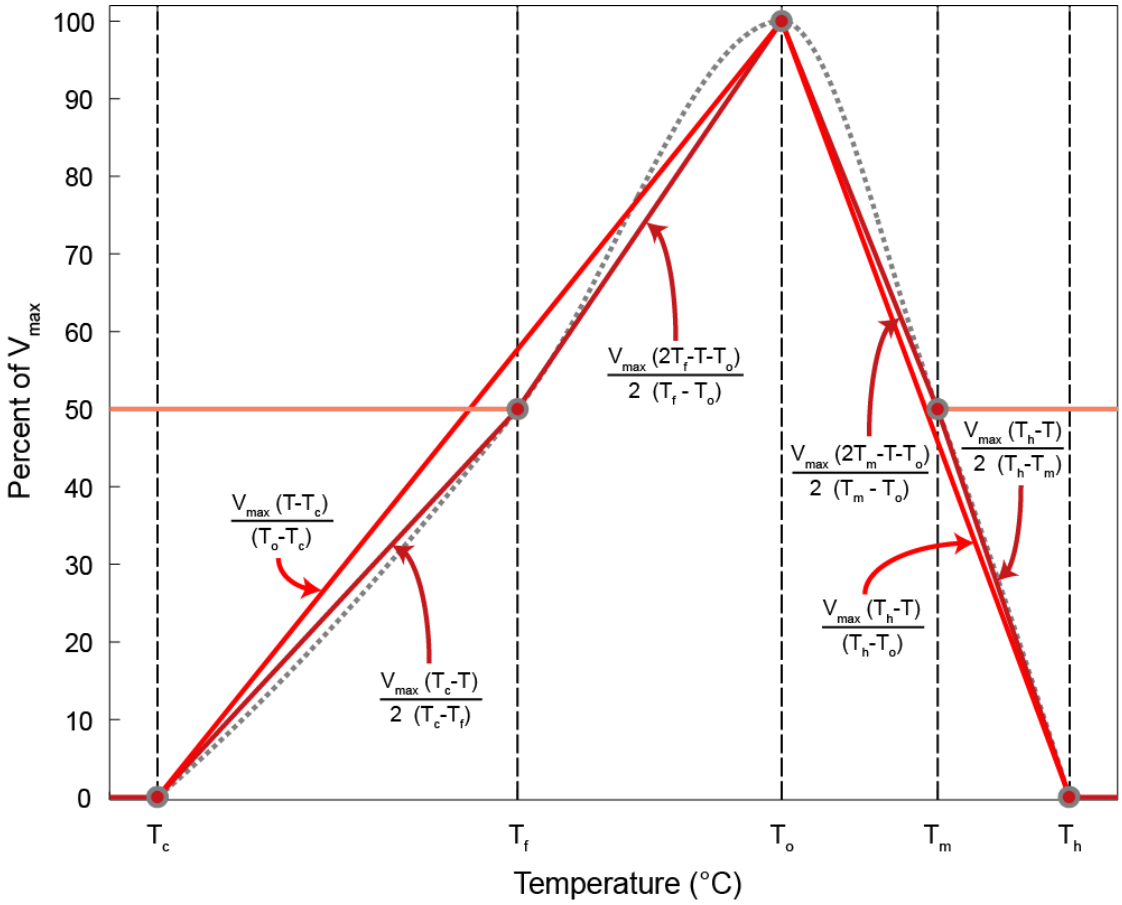

Figure 2.4. Conceptual graph of critical temperatures and the temperature-dependent protein activity constraint function for a generic protein. Conceptual graph of critical temperatures and the temperature-dependent protein activity constraint function for a generic protein. The gray dotted line represents typical ideal protein activity and is shown to illustrate how the piecewise linear functions approximate this ideal. The equations for the piecewise linear functions are displayed. Points indicate the position of critical temperatures labeled on the x-axis. The priority of constraint function usage depends on the availability of critical temperatures and is indicated by the darkness of the red colored lines: dark red lines have the highest priority when the most critical temperature information is available, medium red lines have second priority for when T_m or T_f are unavailable, and light red lines have third priority for when T_h or T_c are unavailable. For conditions not depicted, a default activity equal to V_{max} is assumed.

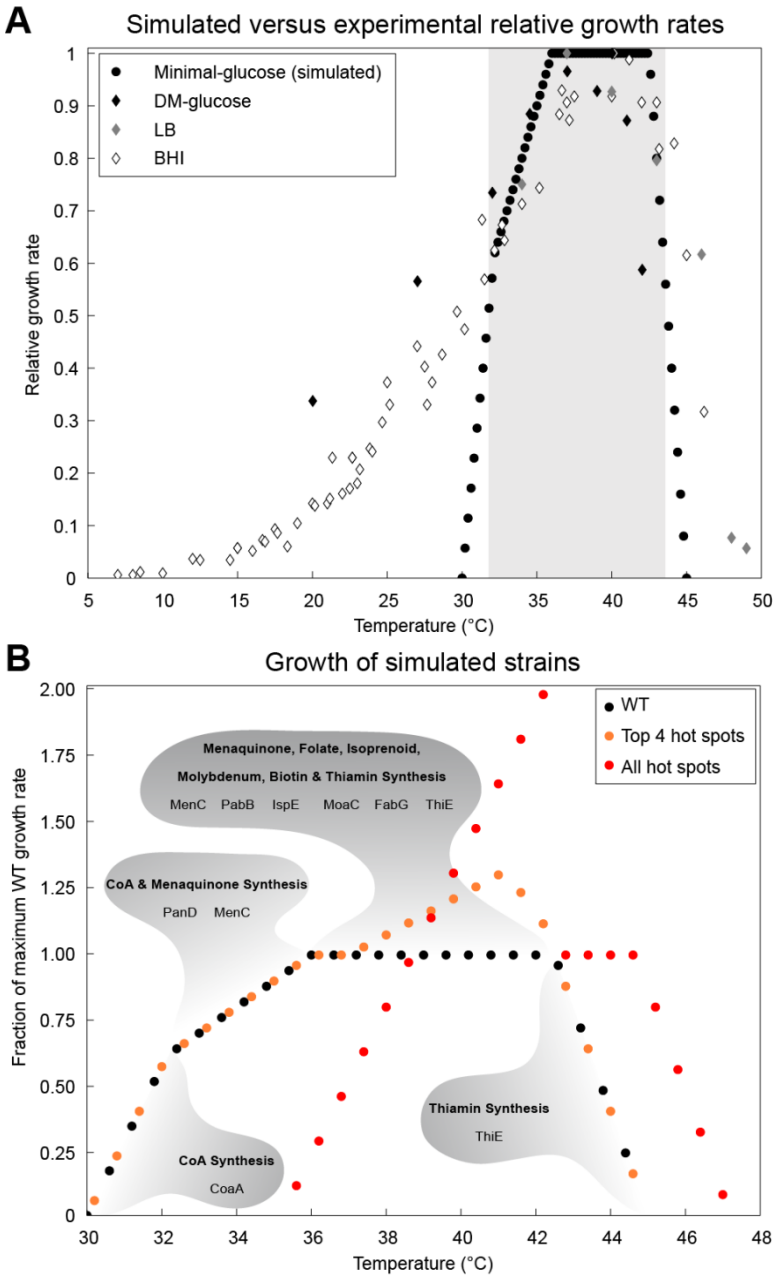

Figure 2.5. Growth rates as a function of temperature. (A) Growth rates as a function of temperature are depicted relative to maximum growth rates under each condition. Circles are growth on minimal media with glucose simulated in this study, and diamonds are experimentally-measured growth on Davis minimal medium (DM) with glucose[86], lysogeny broth (LB)[72], and brain heart infusion (BHI) broth[87]. The shaded region highlights the temperature range for which the model best predicts relative growth rates. **(B)** Simulated growth rates relative to maximum WT growth rate are shown for the WT strain, a strain with the four predicted most growth-limiting network hot spots at 42.2°C completely unconstrained, and a strain with all network hot spots at 42.2°C completely unconstrained. Each phase of WT growth is labeled with the predicted most temperature-limited protein activities and pathways.

The novel modeling framework developed also enabled precise prediction of metabolic network hot spots (Figure 2.6 and see Table S2 in [73]). The most temperature-limited protein activities for each phase of simulated growth are reported in Figure 2.5B, highlighting the trend that cofactor synthesis pathways tend to be the most temperature-limited metabolic processes in the model. Alleviating the temperature-dependent activity constraints on all network hot spots predicted at 42.2°C by optimally shifting their activity functions for growth at that temperature produces a 2-fold increase in maximum growth rate, shifts the optimal temperature to 42.2°C, but narrows the range of growth temperatures due to the incompatibility of these more thermophilic activity functions at mesophilic growth temperatures (Figure 2.5B). Adjustment of activity functions for just the four most-growth-limiting hot spots has similar but dampened effects on temperature-dependent growth (Figure 2.5B). Such activity function changes most directly parallel the engineering of proteins with increased thermostability.

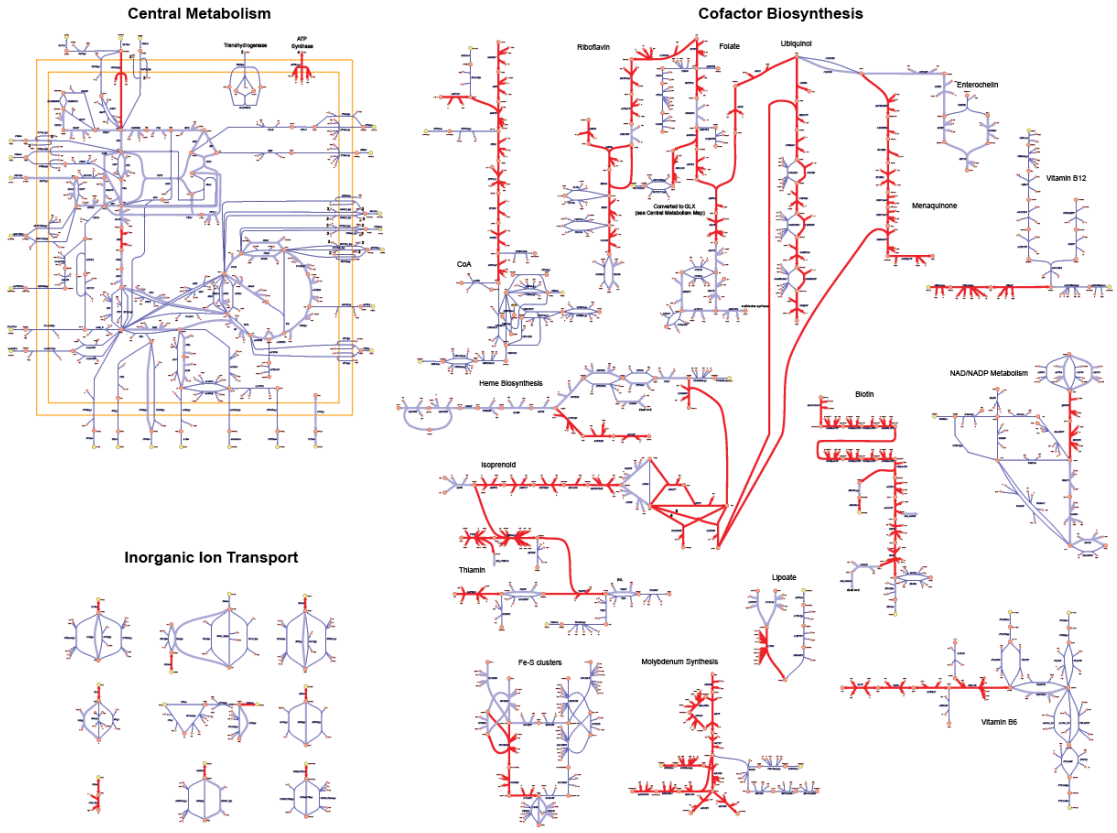

Figure 2.6. Network hot spots at 42.2°C in *i*JO1366 subsystems. All predicted network hot spots are shown as red reactions. Reaction fluxes that increase upon relief of network hot spots are shown as thick blue lines. Only subsystems containing at least one network hot spot are shown. These subsystems include central carbon metabolism, inorganic ion transport, and a number of cofactor synthesis pathways.

Recent adaptive laboratory evolution experiments have yielded 119 thermotolerant *E. coli* mutants[71, 72]. Re-sequencing these mutant genomes revealed many point mutations in coding and non-coding regions of hundreds of genes, and the original studies sought to assign causality to these mutations via frequency of occurrence and gene annotation. Retrospectively investigating the mutations occurring in metabolic genes and their regulators[88] using our novel modeling framework to generate best-validating models of each evolved strain yielded not only classification of potential causal mutations for 51 strains (see Table S3 in [73]) but also mechanistic explanations for their functionality in thermotolerance through compensating for heat-limited growth factors (Figure 2.7). Statistical analysis of the range of model capabilities (see Table S3 in [73]) established that predicted causal mutant gene combinations were significant outliers in conferring thermotolerance at 42.2°C and have extremely low

probability of being identified by chance, signifying the predictive accuracy of heat-affected metabolic activities.

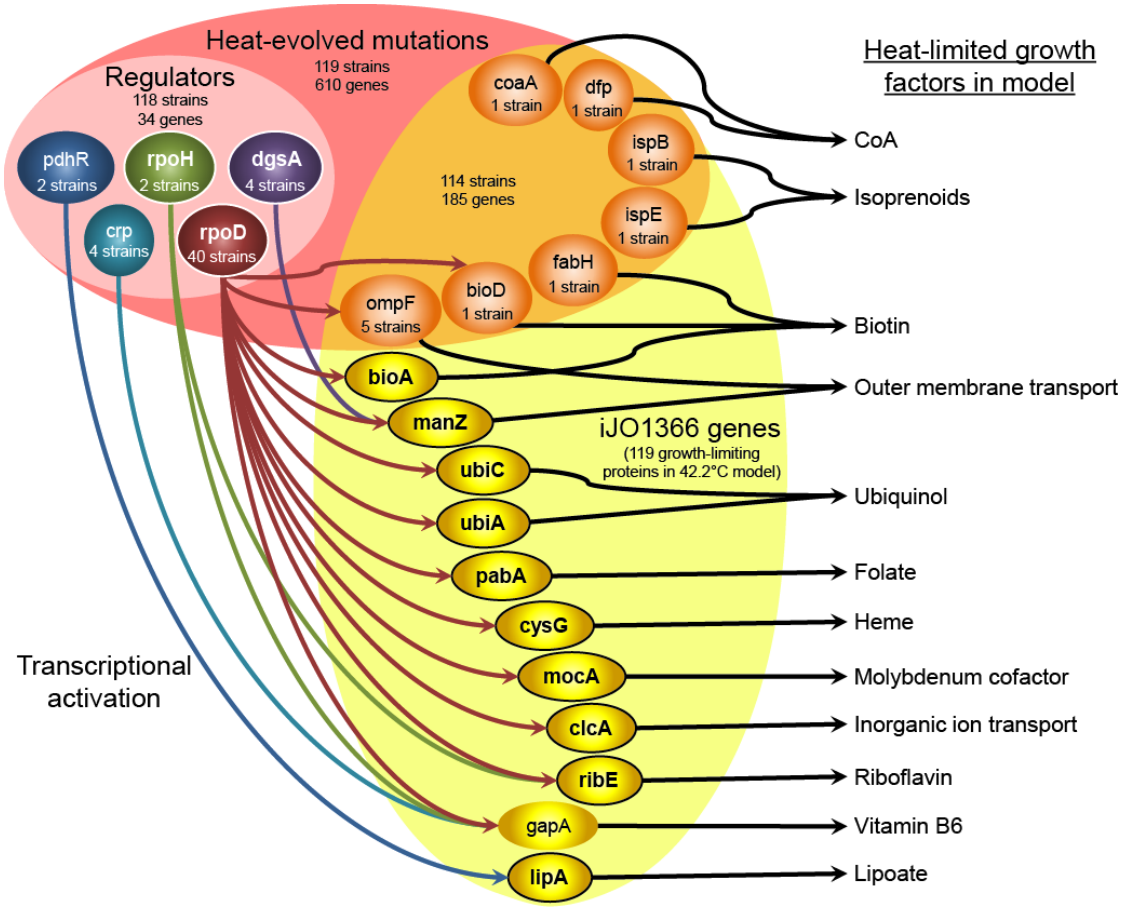

Figure 2.7. Explanatory mechanisms predicted to confer thermotolerance. Explanatory mechanisms predicted to confer thermotolerance are summarized for heat-evolved *E. coli* strains. The total number of heat-evolved strains and mutated genes is given and also noted for the regulatory and metabolic subsets of mutated genes. Only regulators acting upon metabolic genes both predicted to lead to thermosensitivity and with heat-induced transcription in WT are depicted, except for crp. Only metabolic genes predicted to lead to thermosensitivity and either mutated in the set of evolved strains or both activated by depicted regulators and with heat-induced transcription in WT are depicted, except for gapA. Encircled, bolded genes show heat-induced transcription in WT. The predicted metabolic factors limited by heat-dependent decreases in protein activity are indicated at right.

Mutations decreasing thermosensitivity of metabolic activities could be protein-thermostabilizing or otherwise increase protein activity, as through increased gene expression. To further support that increased expression of predicted causal genes that evolved non-coding mutations led to thermotolerance, we gene-expression profiled WT *E. coli* at 37°C and 42°C (see Table S4 in [73])

to identify genes with significantly heat-induced transcriptional activity. Such genes participate in native heat-shock response and therefore offer probable mechanisms for adaptive evolution to cope with elevated temperature.

Of the 119 metabolic genes that can limit growth rate at 42.2°C in the model, 67 are targeted by 12 mutated transcriptional regulators (see Table S3 in [73]), suggesting that the mutations in these regulators may lead to increased expression of these metabolic genes, contributing to thermotolerance in the heat-evolved strains. Ten of these 67 metabolic genes regulated by rpoD, rpoH, dgsA, and pdhR also show significant heat-induced transcription in WT (Figure 2.7 and see Table S4 in [73]). Multiple strains were isolated with mutations in each of these regulators; most notably, 40 of the 119 thermotolerant strains contained mutations in rpoD, corresponding to its apparent centrality in heat shock response through regulation of 64 genes predicted in the model to limit growth rates at 42.2°C, accounting for 9 of the 10 metabolic genes that also exhibit heat-induced transcription (Figure 2.7). Furthermore, rpoD, rpoH, and dgsA also show heat-induced transcription in WT, indicating their native roles in thermotolerance.

The known heat-inducible sigma factors rpoD and rpoH are shown to contribute to thermotolerance in every best-validating model of evolved strains in which they were mutated. Further support that rpoH was likely instrumental in conferring thermotolerance in these strains comes from an independent long-term laboratory evolution experiment at 41.5°C showing significant increase in rpoH expression in multiple evolved K-12 lines[89]. A loss-of-function rpoH mutant also showed increased thermosensitivity[90].

Interestingly an rpoS loss-of-function mutant showed no increased thermosensitivity over WT[90], contrary to the hypothesis that the 3 evolved strains with rpoS mutations therein gained thermotolerance[71]. The best-validating models for 2 of these 3 evolved strains failed to uncover any growth rate dependence on the 77 metabolic rpoS regulatory targets at elevated temperature, the one exception consisting of activity of exactly one rpoS target, folK.

The gene gapA is a known regulatory target of rpoD, rpoH, and crp, which were shown cumulatively to contribute to thermotolerance in best-validating models of 45 evolved strains. Notably, gapA was shown consistently to exhibit the greatest increase in expression out of 35 induced genes following long-term evolution at 41.5°C[89], although there was no observed heat-induced expression in WT.

The WT heat-induced gene lipA, loss-of-function of which leads to thermosensitivity[91], was found to contribute significantly to thermotolerance in best-validating models of both evolved strains in which its regulator pdhR accumulated mutations.

The preceding examples highlight that within the scope of mutations that may impact metabolism, this prediction method is capable of discerning causal from non-causal mutations. Another intriguing observation is that the 7 *i*JO1366 genes with mutations for which the model predicts a mechanistic explanation for thermotolerance (Figure 2.7) showed no heat-induced expression in WT. This result suggests that if these mutations indeed confer thermotolerance, they may lead to increased gene expression via non-native mechanisms, whether constitutive or regulatory, or lead to increased thermostability of protein activity through a non-synonymous coding mutation, as may be the case for the A71V mutation in dfp. The rarer occurrence of mutations in these genes compared to mutations in regulatory genes participating in native heat shock response mechanisms supports this novel mechanism hypothesis, since the probability of evolving a mechanism *de novo* is lower than adapting a pre-existing one. Unfortunately, hypotheses concerning these 7 mutations cannot currently be further evaluated because no expression data is available from the heat-evolved mutants[71, 72], and the mutation locus for dfp is not covered in the GEM-PRO presented in this work or in any available published *E. coli* structures.

With specific network hot spots identified, they may be directly addressed using several strategies: replacement with more thermostable proteins, increasing gene expression to compensate for decreased activity as in heat shock response[89, 92], or bypassing the network hot spots via supplementation. The observed trend that predicted network hot spots limit cofactor synthesis pathways

suggested a relatively simple supplementation approach to test these predictions. The metabolites most immediately downstream of network hot spots (Figure 2.6) and for which transport mechanisms are known to exist in *E. coli* were chosen as supplements, yielding a set of 9 compounds meant to supplement 6 heat-limited growth factors (Table 2.1). Each individual compound and a cocktail combining all compounds were tested for heat-dependent supplementation conferring increased thermotolerance at 42°C relative to 37°C. The supplement cocktail showed significant increase, about 13%, in log phase growth rate (Figure 2.8A) and decrease in the time spent in lag phase (Figure 2.8B) at 42°C but yielded little to no benefit at 37°C.

Table 2.1. Compounds used to test predicted network hot spots in supplementation experiments.

Abbrev.	Name	Conc.	Melting Point (°C)	Growth Factor Supplemented	Uptake Mechanism
pnto-R	(R)-Pantothenate	2 µM	184	CoA	facilitated
btn	Biotin	2 µM	233	Biotin	facilitated
ribflv	Riboflavin	2 µM	280	Riboflavin	passive
lipoate	Lipoate	2 µM	61	Lipoate	facilitated
4abz	4-Aminobenzoate	2 µM	189	Folate	facilitated
dhna	1,4-Dihydroxy-2-naphthoate	2 µM	191	Menaquinone	unknown
cit	Citrate	2 µM	153	Carbon metabolism	facilitated
g6p	D-Glucose 6-phosphate	2 µM	204	Carbon metabolism	facilitated
glyclt	Glycolate	2 µM	75	Carbon metabolism	facilitated

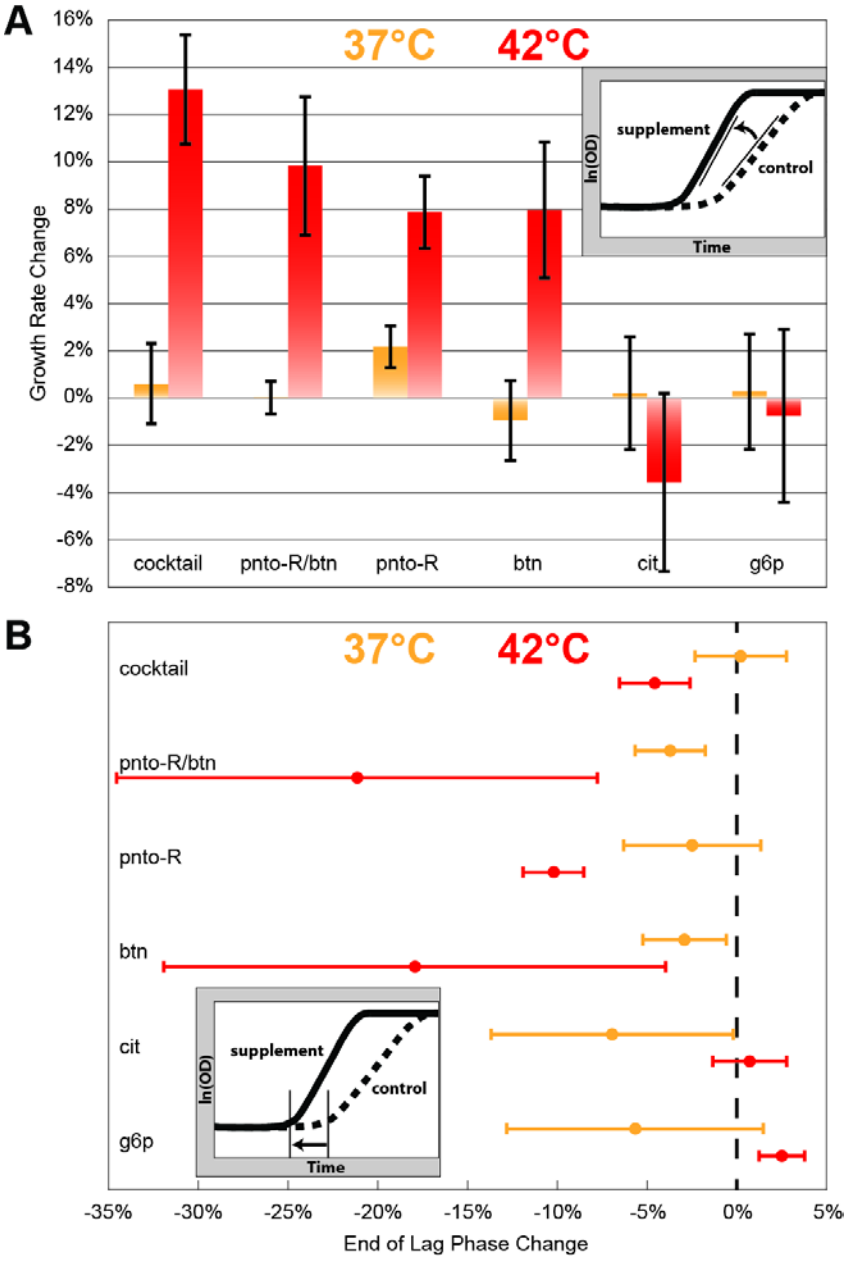

Figure 2.8. Heat-dependent supplementation increases thermotolerance. (A) Changes in specific growth rate upon supplementation relative to a no supplement control are depicted in orange for 37°C and red for 42°C. Error bars indicate the standard deviations with n = 3 for each condition. The inset graph illustrates how growth rate changes were computed by comparing the maximum slopes of growth curves between the control and supplement condition. **(B)** The change in lag phase time under each supplementation condition is plotted in orange for 37°C and red for 42°C. Error bars indicate the standard deviations with n = 3 for each condition. The inset graph illustrates how the end of lag phase changes were computed by comparing the times of log phase initiation between the control and supplement condition. Cocktail: combination of all 9 supplements, pnto-R: pantothenate, btn: biotin, cit: citrate, g6p: glucose-6-phosphate.

Triplicate experiments at both temperatures for subsets of the cocktail compounds (see Table S6 in [73]) were prioritized for the 4 compounds resulting in the highest growth rates at 42°C in single experiments (Figure 2.9). The individual supplements pantothenate and biotin provided a significant although lesser degree of heat-dependent supplementation than the cocktail (Figure 2.8A). Notably, pantothenate production in WT at 37°C has been found to be in excess of the required amount for growth by as much as 15-fold, leading to excretion[93]. A pathway with excess activity at 37°C being successfully supplemented at 42°C highlights the significance of the prediction that this pathway suffers lowered activity due to thermal deactivation of just four proteins, PanB, PanC, PanD, and IlvC. The heat dependency of this supplementation precludes the scenario where the supplements simply alleviate the burden of synthesizing cofactors from the nutrient carbon source; it indicates that the model-based predictions of thermosensitive metabolic activities were accurate and, due to the location of supplement entry into the network, supports that the precise proteins predicted to be network hot spots are limiting at 42°C. The alternative carbon source supplements citrate and glucose-6-phosphate did not individually yield significant benefits to growth rate or log phase initiation, suggesting that central carbon metabolism is not rate-limiting under heat shock. However, this also establishes both compounds as valid negative controls for heat-dependent supplementation in WT *E. coli*.

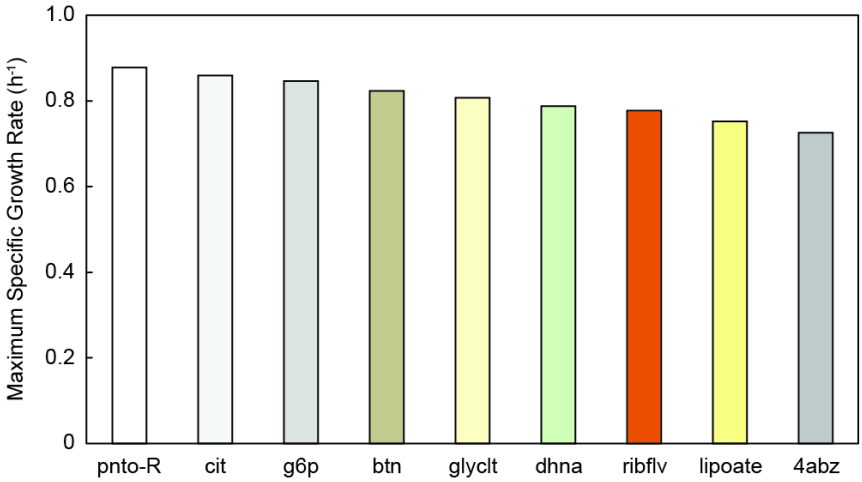

Figure 2.9. Screen of individual supplement conditions at 42°C. The data is sorted from left to right in decreasing order of growth rate to show how subsequent triplicate experiments at both 42°C and 37°C were prioritized. No error bars are displayed because each of these results was derived from a single experiment.

The relatively greater benefit of the supplement cocktail suggested a combined effect beyond what was observed for any of the individual compounds. The combination of pantothenate and biotin provides a partial explanation for this combined effect, as illustrated in Figure 2.8. However, the precise combination of cocktail components responsible for the full effect remains incompletely determined, suggesting that perhaps additions to pantothenate and biotin may compensate for less rate-limiting network hot spots in a manner not observable individually. This result supports the existence of at least a rough rate-limiting rank order to network hot spots, similar to that predicted using our modeling framework.

Discussion

In this study, we have provided new evidence establishing that metabolic processes are among the most growth-limiting under heat stress, in particular CoA and biotin synthesis and perhaps other cofactor pathways as well. The hypothesis that cofactor pathway enzymes are most rate-limiting under heat shock has been raised previously[94], but no mechanistic evidence for these dependencies has been substantiated to date. In this study, we show that these dependencies arise directly from the systemic constraints that proteome thermostability imposes upon growth, network hot spots, which can be relieved through increased individual protein thermostability, through compensatory transcriptional regulatory responses, or through exogenous supplementation of the most limiting processes. Understanding and controlling thermotolerance in microbes has important implications in developing industrial microbial biocatalysts[95], probiotics[96], and bacterial vaccines[97]. The most efficient producers of compounds of interest are rarely also naturally thermotolerant, but the absence of a genetic system limits the usefulness of native thermophiles in industrial processes. Therefore, strategies for increasing thermotolerance of production strains are of great interest.

Incorporating molecular properties of proteins into a metabolic model through development of the GEM-PRO enabled discovery of these thermosensitive processes and comprised a framework through which disparate data types were reconciled to explain fundamental properties of heat shock response. Because metabolic proteome thermostability is such a major evolutionary pressure, this

framework was also able to provide mechanistic explanations to distinguish causal mutations conferring microbial thermotolerance, a result illustrating the importance of systems biology in interpreting complex mutation data and perhaps even personal genomics data. Our result provides a solid argument for the necessity of systems biology in understanding complex stress responses and mechanisms through which tolerance to these stresses evolve and could be controlled. Our discoveries of previously unknown supplements to heat stress exemplify such control. Furthermore, these discoveries would not have been possible using either the protein structure data or the metabolic network in isolation, illustrating the potential of the emerging field of structural systems biology.

Methods

E. coli GEM-PRO Reconstruction

Reference amino acid sequences for all 1366 proteins included in *i*JO1366 were collected from the UniProt database[78]. The entire set of *E. coli* proteins represented by structures in the PDB[79] was searched by sequence alignment for all structures corresponding to each protein (sequence identity cutoff >95%). For each protein, the set of corresponding structures was manually checked for exact correspondence to the reference protein by name to ensure that close homologs were not counted as corresponding to the reference protein. Metabolite compounds included in *i*JO1366 were mapped to PDB ligands first via their KEGG entries[98], which often contain direct links to the corresponding PDB ligand. For those metabolites not yet mapped to a PDB ligand, their canonical SMILES[52] were obtained from the PubChem database[99] and then searched for similar compounds (similarity cutoff >80%) in the PDB through the chemical component search. Significant hits for the compound were manually inspected for exact correspondence to the query metabolite (Dataset S1); if an exact matching PDB ligand was not found but a similar compound was found with exactly one functional group different from the query metabolite, then this ligand was mapped to the metabolite as an analog (Dataset S1) to be used in discriminating among PDB structures for the below objectives. For each objective below, the corresponding PDB structures were manually curated as described, and, when indicated, a structure

template was chosen for homology modeling from the entire PDB database. Thus 2784 protein structures were generated.

Objective 1: Native WT Structure

For each protein, from the corresponding set of PDB structures a single best-representative structure was chosen that accounted for, in order of priority, the minimal number of point mutations, maximum coverage of the length of the reference amino acid sequence, binding as many of the native metabolic substrates as possible (with priority for primary metabolites or chemical analogs of primary metabolites), and structural resolution. If the resulting best-representative structure contained one or more point mutations deviating from the reference sequence, the protein was slated for homology modeling to replace mutated residues with their analogs from the reference sequence.

Objective 2: Maximum Sequence Coverage

For each protein, from the corresponding set of PDB structures the single structure with maximum coverage of the reference amino acid sequence was identified. If the structure had 100% coverage of the reference sequence and 100% sequence identity, it was accepted as the maximum-coverage structure. Multiple perfect-matching structures were chosen from based on binding of native metabolic substrates or analogs and structural resolution. Otherwise, FFAS[100] was run on the protein (with default parameters) to search for the best homology modeling template based on FFAS score, reference sequence coverage, and reference sequence identity as described previously[67]. If the template had greater reference sequence coverage than any available *E. coli* PDB structure, the protein was slated for homology modeling using the template. If the best available *E. coli* PDB structure had greater reference sequence coverage than the template but was not 100% identical to the reference, it was slated for homology modeling using the *E. coli* structure as a template.

Objective 3: Protein-Substrate Induced Fit

Protein-substrate pairs were taken from *i*JO1366 by enumerating all pairwise combinations of every protein taking part in a reaction's gene-protein-reaction association (GPR) and every metabolite compound taking part in the associated reaction, discounting H^+, H_2, and H_2O. For each protein-

substrate pair, the corresponding set of PDB structures for the protein were searched to find those containing preferably the exact substrate as a ligand or secondarily a chemical analog of the substrate. If multiple such structures existed, priority was given for the structure with higher sequence identity and coverage, inclusion of a higher number of metabolic substrate molecules, and structural resolution, in that order. If no corresponding protein structure also contained the metabolite compound or a chemical analog, FFAS was run on the protein to identify candidate templates for homology modeling. Because the structure database used by FFAS is a set of sequence-clustered PDB structures (99% identity cutoff), each FFAS hit was mapped back to its structure cluster, and each of these cluster members was investigated for presence of the metabolite or a chemical analog included in the structure. The template with lowest FFAS score that contained either the exact metabolite compound as a ligand or a chemical analog was chosen for homology modeling of the protein. In the event of ties for best candidate template, structural resolution was used to choose the best structure.

Homology Modeling Using ProtMod

Proteins slated in the above steps for homology modeling were modeled using their chosen template using the ProtMod server (http://ffas.burnham.org/protmod-cgi/protModHome.pl) with default parameters. The three homology modeling algorithms implemented in ProtMod were SCWRL[101], Jackal[102], and MODELLER[103], generating three structures per protein. Every structure was evaluated for quality using PSQS[104], and the modeled structure with the lowest total PSQS score (Figure 2.2) was chosen to represent the protein. File names for chosen representative PDB structures and modeled structures are presented in Dataset S1.

Objective 4: Residue-Resolution Functional Annotation

In order to enable use of the GEM-PRO to study molecular functions of proteins, known functional sites and other structural features were mapped to the structure files. Protein residue-resolution annotation was collected for catalytic sites, allosteric regulatory sites, secondary structure, and other structural features[75-78]. The locations of these features within the reconstructed structures was determined by aligning the sequences from the annotation sources to the sequences as contained in the

PDB format files of the GEM-PRO. Thus, annotated features have start and stop residue positions corresponding to the residue numbering contained in the GEM-PRO files, and when available, specific amino acids are specified (Dataset S1).

Modeling Temperature-Dependent Protein Activity Constraints

The two-state model for enzyme activity as a function of temperature[105] was assumed. Although this classical model has been criticized as incompletely capturing the relationship between protein activity and the transition between native folded and thermally denatured states[106], assuming the two-state model for this study enables approximation of activity functions using the limited data and bioinformatic techniques currently available for temperature-dependent protein activity and thermostability. The two-state model essentially treats the process of global denaturation as directly causal and co-occurring with loss of protein activity, where activity is directly proportional to the molar fraction of native folded protein and all protein molecules are either in the native folded state or in the denatured state. The potential error in the two-state model is that partial denaturation of active sites may occur at more intermediate temperatures than the temperature at which global denaturation is complete. Thus, assuming the two-state model may underestimate the extent of activity loss in the range between the optimal temperature and the temperature of complete global denaturation for some proteins. Nevertheless, evidence exists that the two-state model may be the most appropriate explanation of this relationship in at least some well-characterized proteins[107, 108].

Critical Temperatures

The temperature at which maximal protein activity (V_{max}) occurs is called the optimal temperature (T_o) for a protein. Given the two-state model, the definition of protein melting temperature (T_m) is the temperature at which 50% of the molar fraction of protein is in the native folded state as opposed to the denatured state and therefore 50% of V_{max} activity is present. The melting temperature is also sometimes referred to as the transition temperature in the literature. The temperature at which the molar fraction of native folded protein falls to 0%, leading to a total loss of activity, we refer to in this study as the temperature of complete heat denaturation (T_h). Although the nature of protein inactivation

is somewhat different at sub-optimal temperatures, for our conceptual framework, we again assume a two-state model, with a cold transition temperature we call a freezing temperature (T_f) and a temperature of complete cold denaturation (T_c). Together, T_c, T_f, T_o, T_m, and T_h define the critical temperatures for protein activity utilized in this study.

Processing Experimentally-Measured Protein Critical Temperatures

Protein activity and stability data at various temperatures was collected from the BRENDA database[80] and melting temperatures from the ProTherm database[81]. Data for mutant proteins from either database was disregarded. Because BRENDA data is assigned using Enzyme Commission (EC) numbers instead of specific protein identifiers, as ProTherm does, it was necessary to map the BRENDA data from EC numbers to proteins. If an EC number corresponded to exactly one *E. coli* protein in the KEGG database[98], the BRENDA data was mapped to that protein; otherwise, the linked literature references in BRENDA were reviewed to determine precisely which protein the data corresponded to. The data from BRENDA specifically denotes T_o and T_m but also often includes percentages of maximum activity, which could be anywhere between 100% and 0%. Any activity data point not precisely labeled as T_o or T_m was used to estimate the critical temperatures, first by rounding to the nearest 50% of activity, then by classifying as T_o (if rounded to 100%), T_m (if rounded to 50%), or T_h (if rounded to 0%). In the absence of experimentally-measured T_o of a protein, a default T_o of 37°C was assumed based on the optimal growth temperature of *E. coli* K-12 unless experimentally-measured and precisely-labeled T_m values fell below 37°C, in which case T_o was left null. Estimated T_m and T_h temperatures were then reclassified, if they were less than T_o, as T_f and T_c, respectively. This procedure often resulted in multiple distinct values for a given critical temperature. Whether precisely labeled in the databases or classified based on rounding, the critical temperatures were then processed further to arrive at just one value for each by taking the median of all temperatures of the same class for a protein. If at this point the order of critical temperatures was not $T_c < T_f < T_o < T_m < T_h$, adjustments were made by taking the minimum of T_c or T_f classified temperatures and maximum of T_h or T_m

classified temperatures as necessary to create this order. This final step was only necessary for 7 proteins out of the 376 proteins with at least one experimentally-measured critical temperature.

Bioinformatic Prediction of Protein Melting Temperatures

Due to the relatively low coverage of critical temperatures by experimental data, it would not currently be possible to assess whole metabolic proteome activity under temperature stress without supplementing the experimental data with bioinformatic predictions of thermostability. Existing bioinformatics methods focus on predicting the transition temperature for heat denaturation, the melting point (T_m). Because this is the most moderate critical temperature in the super-optimal temperature range, identification of T_m should suffice to capture activity of most proteins within the moderate heat-shock range for *E. coli*. Different structural properties of proteins are evaluated by existing methods to predict T_m. Four such methods were implemented in this study that focus on distinct structural properties: occurrence of dipeptides significantly enriched in known highly thermostable proteins[82], thermodynamic contribution of chemical groups to total energetic difference between native and unfolded states[83], polar and non-polar buried surface area and configurational entropy[84], and length of primary structure[85].

Ku *et al* method: The frequency of occurrence of specific dipeptides has been shown to correlate strongly with protein melting temperatures[82]. The scoring matrix for dipeptides published with that method was used in this study as described previously[82] yielding T_m indices for each protein. These indices were previously shown to linearly correlate with melting temperatures such that an index of 0 corresponds to a melting temperature prediction of 55°C, an index of 1 corresponds to 65°C, and a linear scale derived from this correspondence can yield precise T_m predictions for an index < 0 and > 1.

Oobatake *et al* method: Formulation of energy functions from first principles based on native folded and unfolded structures[83] comprise another basis for T_m prediction. In that study, chemical group contribution to thermodynamic terms in these functions was determined based on free forms of all 20 standard amino acids. Given a constant temperature, the only additional input necessary to implement this prediction method was the solvent accessible surface area of each chemical group in the native and

unfolded conformations of the protein. The structure file for each protein provided the native structure and a basis for computing these terms using the Chimera analysis software[109] to compute solvent accessible surface areas. A simplifying assumption was made to compute solvent accessible surfaces for the unfolded protein; when computing the accessible surface of chemical groups on a given amino acid residue, the only atoms considered were those contained in the tripeptide defined with the given amino acid located in the center. This approach estimates the solvent exposure of atoms on the central residue as would result from complete unfolding of the protein. Using this approach to compute energy terms for native and unfolded states allowed for explicit calculation of T_m by finding the temperature at which the ΔG term converged on 0. For proteins with multiple corresponding structures, the median prediction value was taken.

Dill *et al* method: More recently, a method for T_m prediction has been developed based on the assumption that primary structure length alone is a valid predictor of thermostability[85]. An energy function in that study was formulated through reduction of the classical thermodynamic terms to functions of length. Using the published formulation, explicit calculation of T_m resulted from finding the temperature at which the ΔG term converged on 0.

Murphy *et al* method: Another energy function formulation was developed previously[84] based on configurational entropy changes of amino acid residues upon unfolding and polar versus non-polar buried surface area. The Chimera analysis software[109] was used to compute the polar and non-polar buried surface areas of protein structures, providing the necessary parameters for implementation of this method when combined with the estimated amino acid configurational entropy changes and constants published with the original study[84]. Melting temperatures were predicted from this method by explicit calculation of the temperature at which the ΔG term converged on 0. For proteins with multiple corresponding structures, the median prediction value was taken.

Composite method: Given the divergent bases for the four implemented T_m prediction methods and the observation that different methods outperformed others at distinct temperature ranges, we chose to train a composite prediction pipeline on the 172 proteins for which experimentally-measured T_m

values were available and for which all four methods were able to make a prediction. We first computed the root-mean-square deviation (RMSD) with respect to the experimental T_m for each predicted T_m. For each method, we sorted the data by predicted T_m value and identified the temperature ranges for which the total RMSD was smallest for the given method; these roughly correspond to the temperature ranges for which the given method is the best T_m predictor. Because the rank order of predicted T_m values differed across methods, there were overlaps between the temperature ranges where RMSD was smallest with respect to each method. It was found that, at least with respect to the training set, overlaps could always be resolved by using a consistent priority list of the four methods in choosing which T_m prediction to accept. The priority list in order of decreasing priority was Ku *et al*, Oobatake *et al*, Dill *et al*, and Murphy *et al*. Using this priority list to resolve all overlapping temperature ranges and to make predictions outside of the composite temperature ranges of best performance for each method, T_m predictions were generated for use in subsequent applications of critical temperatures in this study (see Table S1 in [73]). The Pearson correlation coefficient of the composite T_m predictions with experimentally-measured T_m values is $\rho = 0.69$ (p-value = 6.55×10^{24}). Figure 2.3 shows the direct comparison of experimental and predicted T_m values. Notably, regardless of what priority list of the four methods is used, the minimum correlation is $\rho = 0.64$ for the composite predictions. The performance of the composite pipeline was evaluated further by 5-fold cross validation where total RMSD was computed with respect to the experimentally-measured T_m data and statistical significance was evaluated for each fold based on 1000 trials of randomly choosing a T_m prediction method for each of the proteins in the test set. This 5-fold cross validation showed that the composite prediction method performed significantly better than expected by chance (average p-value = 0.0032).

Approximating Protein Activity V_{max} by Simulation

In order for temperature to affect protein activity in a metabolic model, a term V_{max} must be defined. For the purpose of this study, V_{max} accounts for the maximum possible activity of the set of all molecules of a particular metabolic protein under a reference nutrient condition in the absence of temperature stress. This approach is based on the assumption that all proteins in the cell are operating at

approximately V_{max} under steady-state with constant nutrient conditions. The V_{max} of a given protein is assessed through flux variability analysis (FVA)[110], determining the maximum magnitude flux possible through any reaction associated with that protein under the constraint that the model must achieve ≥ 90% of the maximum possible biomass flux. This biomass flux cutoff is a tunable parameter for affecting the eventual protein activity functions. The lower the cutoff, the higher the potential V_{max} values will be and the less constraining the temperature parameter will be on individual proteins for which V_{max} increases. Although a biomass flux cutoff of 90% maximum was ultimately chosen, relative temperature-dependent growth rate predictions were robust within the tested range of parameter values from 50% to 100%. This analysis provides a V_{max} value for each protein in the network.

Formulating Protein Activity Functions

Protein activity as a function of temperature was formulated as piecewise linear functions sequentially connecting the points defined by V_{max} and the critical temperatures (Figure 2.4). The formulation of these linear functions depends on the availability of specific critical temperatures. For example, if T_o and T_m are both available, the line segment connecting points (T_o, V_{max}) and (T_m, $V_{max}/2$) is chosen for the activity function over this range of temperatures; alternatively, if T_o and T_h are both available but T_m is not, the line segment connecting points (T_o, V_{max}) and (T_h, 0) is chosen for the activity function over that range of temperatures. If T_m is available and T_h is not, for temperatures > T_m, the activity function is conservatively set equal to $V_{max}/2$. For any temperatures not covered by linear functions in Figure 2.4 due to a lack of critical temperature availability, a conservative default level of activity equal to V_{max} is assumed.

Constraining the Model Using Protein Activity Functions

With the temperature-dependent protein activity functions defined, constraining the metabolic model to account for temperature is as simple as inputting the desired environmental temperature for growth. This temperature is used to compute the magnitude of both upper and lower bound (if the reaction is reversible) reaction flux constraints in the model with respect to each protein in the corresponding GPR. If this results in multiple constraints for a single reaction, the most limiting protein

activity constraint is chosen to constrain the reaction by choosing the maximum protein activity to resolve "OR" relationships and the minimum protein activity to resolve "AND" relationships. The most limiting protein for each reaction is tracked for subsequent analysis whenever temperature-dependent constraints are imposed. With the temperature-dependent constraints imposed upon network reactions, standard steady-state constraint based simulation methods can be implemented to analyze network fluxes, including flux balance analysis (FBA)[111] and a variety of other established methods[112]. In this way, temperature is established as a model parameter simulating the effect of thermal denaturation on proteome function.

Network Hot Spot Prediction

Basic Model Constraints

The *E. coli* metabolic network *i*JO1366[74] was loaded into the COBRA toolbox[112] from the published SBML model using Matlab. Since the time of publication of *i*JO1366 a thermodynamic constraint error was discovered in the published model; as a result, the malate oxidase, "MOX," reaction was set as irreversible. The superoxide dismutase, "SPODM," reaction was set with an initial upper bound of 1000 as well. The objective function was set as the complete wild type biomass reaction "Ec_biomass_*i*JO1366_WT_53p95M." Default exchange reaction constraints were used, except for a glucose uptake lower bound of -8 mmol/gDW/h and an oxygen uptake lower bound of -18.5 mmol/gDW/h, representing aerobic growth on glucose. These basic constraints were used for all reported simulations in this study. Using the temperature-dependent protein activity functions described above, the environmental temperature parameter was set as reported in the results of this study. For identification of network hot spots, the temperature parameter was set equal to 42.2°C.

Identifying Hot Spots

The following steps describe the sequential identification of network hot spots (Figure 2.6 and see Table S2 in [73]) in the constrained model:

1. All reactions constrained by temperature-dependent protein activity in the model are identified.

2. All correlated reaction sets[113], those that are fully-coupled and therefore always carry flux if any other reaction in the same set does[114], are identified that include at least one of the reactions identified in step 1.

3. Identify the minimum set of proteins responsible for the greatest increase in biomass flux upon relieving temperature constraints, the hot spot(s) for this iteration. For each correlated reaction set identified in step 2:

    a. Identify if completely relieving the temperature-dependent constraints on these reactions leads to an increase in biomass flux. If so, proceed to step 3b.

    b. From the correlated reaction set identified in step 3a, determine the minimum set of associated temperature-dependent protein activity constraints that must be relieved to achieve any increase in biomass flux. If multiple protein sets with the same minimum number of members exist, choose the one leading to the greatest increase in biomass flux.

    c. If the number of relieved protein constraints from step 3b is smaller than any previously found set of network hot spots in this identification iteration, replace the most limiting set of hot spots with the set from step 3b. If the number of reactions equals that of the previously found most limiting set of hot spots, replace with the set from step 3b only if the increase in biomass flux is greater upon relief of constraints.

4. If no network hot spots were identified in step 3, terminate operation. Otherwise, record the most limiting network hot spot(s) identified in step 3 for this iteration. Update the metabolic model by permanently relieving the constraints on these hot spots first by setting T_o for these proteins equal to the environmental temperature parameter value and second by re-deriving the temperature-constrained model using this updated set of critical temperatures. Return to step 1 using the updated model.

The method described above may identify a single protein as the most limiting network hot spot or multiple proteins together as the most limiting network hot spots in a given iteration. The range of number of proteins in these most limiting hot spots was from 1 to 4 in this study. It should also be noted that the method formulated above guarantees eventually achieving the maximum theoretical biomass flux that would be possible in the complete absence of temperature-dependent constraints on the model. For the simulated strains reported in Figure 2.5B, the referenced sets of network hot spots were relieved by shifting all critical temperatures for an individual protein by the minimum amount required to achieve maximal growth and re-deriving the metabolic model based on that set of critical temperatures.

Generating Best-Validating Models of Evolved Strains

In order to study the mechanistic causality of mutated genes reported in heat-evolved *E. coli* strains[71, 72], a metabolic modeling strategy was devised to find the maximal contribution to thermotolerance that the combined mutations of a single strain could achieve and the minimal set of these mutations necessary to achieve this thermotolerance. In brief, thermotolerance was defined by the simulated growth rate at 42.2°C relative to the wild type model, the effect of mutated genes was simulated by relieving temperature-constraints on encoded proteins, and the non-mutant proteins were left as constrained for the wild type model. The detailed method follows for generating the best-validating model for a given evolved strain:

1. Beginning with the wild type model constrained at the set environmental temperature, all correlated reaction sets are identified that include at least one reaction associated with a mutated gene or a regulatory target of a mutated regulator in the evolved strain.

2. All proteins associated with correlated reaction sets identified in step 1 are found.

3. Relieve the temperature-dependent constraints on all proteins identified in step 2. If the maximum achievable biomass flux increases relative to the wild type model, proceed to step 4. Otherwise, the metabolic model is incapable of predicting causal mutations for this strain.

4. Iteratively reintroduce temperature-dependent protein activity constraints on proteins not directly mutated or targeted by mutated regulators in the evolved strain. If reintroducing each protein activity constraint does not diminish the increase in biomass flux observed in step 3, retain the reintroduced temperature-dependent constraint. Otherwise, maintain the relief of the constraint tested in this step.

5. Iteratively reintroduce temperature-dependent protein activity constraints on proteins that are targeted by mutated regulators in the evolved strain. If reintroducing each protein activity constraint does not diminish the increase in biomass flux observed in step 3, retain the reintroduced temperature-dependent constraint. Otherwise, maintain the relief of the constraint tested in this step.

6. Iteratively reintroduce temperature-dependent protein activity constraints on proteins encoded by mutated genes from the evolved strain. If reintroducing each protein activity constraint does not diminish the increase in biomass flux observed in step 3, retain the reintroduced temperature-dependent constraint. Otherwise, maintain the relief of the constraint tested in this step.

7. If any correlated reaction set from step 1 is no longer represented by at least one protein encoded by a mutated gene or target of a mutated regulator through relieved constraints, remove all proteins only associated with that reaction set from further analysis and permanently reintroduce the corresponding protein activity constraints. This step prevents usage of reactions outside of correlated reactions sets directly associated with predicted causal mutations in achieving thermotolerance in the model.

8. Repeat steps 4 – 7 until no change in a single protein activity constraint can be made without diminishing the biomass flux from step 3. The protein activities that remain relieved after this step comprise the utilized proteins required to achieve the maximal thermotolerant growth rate at the environmental temperature.

9. Generate the best-validating model by shifting all critical temperatures for each utilized protein from step 8 by the difference between the environmental temperature parameter value and the T_o for that protein and re-deriving the metabolic model based on that set of critical temperatures. The overlap between utilized proteins in step 8 and the mutated genes or mutated regulators in the evolved strain that target the genes encoding these proteins are the predicted causal mutations for this strain.

Of the 119 evolved strains, 51 yielded a best-validating model using the method described above (see Table S3 in [73]). The number of utilized mutated genes and targets of mutated regulators in a best-validating model ranged from 1 to 65. Every best-validating model included relieved activity constraints on proteins not implicated in the corresponding evolved strain through mutation or in association with mutated regulators, ranging from 3 to 93 such proteins in a single model. All such proteins are associated via correlated reaction sets with predicted causal mutations. These proteins may be encoded by genes that represent unknown targets of mutated regulators, or they may result from inaccurate relative T_m predictions or missing model constraints. Because these proteins participate in the same metabolic pathways as predicted causal mutated genes, they do not change the prediction of temperature-limited metabolic pathways. The relative impact of including such proteins in the best-validating models upon the prediction of thermotolerance was evaluated in the analysis described in the section below.

Statistical Analysis of Best-Validating Models

To assess the significance of causal mutations predicted for evolved strains, a random sampling approach was taken to statistically evaluate the best-validating models within the broader flux space of metabolic network. This approach also addresses the impact of allowing increased activity of proteins not known to be associated with measured mutations in achieving thermotolerance in the best-validating models. The steps to this approach were as follows:

1. Identify the number of genes predicted by the best-validating model to represent causal mutations in the evolved strain. This is the sum of mutated genes and targets of mutated regulators from the evolved strain that are utilized in the best-validating model.

2. Randomly select the number of genes identified in step 1 from the full set of 1366 genes contained in the *i*JO1366 network.

3. Identify all genes participating in the same correlated reaction sets as one or more of the genes from step 2.

4. Starting from the wild type model with temperature-dependent protein activity constraints at the set temperature, relieve the temperature-dependent constraints on proteins encoded by genes from step 3 by shifting their critical temperatures by the difference between the environmental temperature parameter value and T_o. Then derive a metabolic model based on that new set of critical temperatures.

5. Maximize the biomass flux through the model from step 4 using FBA.

6. Repeat steps 1 – 5 to randomly sample models and their maximum biomass fluxes 1000 times, keeping track of the number of random models that achieved equal or greater biomass fluxes than the maximum biomass flux for the best-validating model currently under evaluation.

7. Compute an empirical p-value for the best-validating model equal to the number of equivalently or more thermotolerant models from step 8 divided by 1000.

The empirical p-values resulting from the method described above represent how significantly thermotolerant the best-validating models are relative to chance predictions from the network flux space. These p-values are reported in Table S3 in [73]. The predicted thermotolerance of 45 of the 51 best-validating models was never randomly matched or outperformed using the same number of genes and all genes associated via correlated reaction sets. The other 6 best-validating models had empirical p-values ranging from $0.001 \leq \text{p-value} \leq 0.051$. These results indicate that the precise combinations of

causal mutations for thermotolerance predicted in the best-validating models are all highly significant and could not have been predicted by chance.

Gene Expression Profiling

Bacterial Strains, Media, and Growth Conditions

Escherichia coli K-12 MG1655 was grown in glucose (2 g/L) minimal M9 medium containing 2 mL/L 1 M $MgSO_4$, 50 mL/L 1 M $CaCl_2$, 12.8 g/L $Na_2HPO_4.7H_2O$, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl and 1 mL trace element solution (100X) containing 1 g EDTA, 29 mg $ZnSO_4.7H_2O$, 198 mg $MnCl_2.4H_2O$, 254 mg $CoCl_2.6H_2O$, 13.4 mg $CuCl_2$, and 147 mg $CaCl_2$. Glycerol stock of the *E. coli* strain was inoculated into the minimal medium supplemented with glucose and cultured at 37°C with constant agitation overnight. The culture was diluted 1:100 into 50 mL of the fresh minimal medium and then cultured at 37°C to mid-exponential phase (OD_{600} ~ 0.6). For heat-shock experiments, cells were grown to mid-exponential phase at 37°C and half of the culture was sampled as a control. The other half culture was transferred into pre-warmed (50°C) medium to get the media temperature to 42°C and then incubated for 10 min.

Total RNA Isolation

Three milliliters of cells from mid-exponential phase culture were mixed with 6 mL RNAprotect Bacteria Reagent (Qiagen). Samples were mixed immediately by vortexing for 5 seconds, incubated for 5 minutes at room temperature, and then centrifuged at 5000×g for 10 minutes. The supernatant was decanted and any residual supernatant was removed by inverting the tube once onto a paper towel. Total RNA samples were then isolated using RNeasy Plus Mini kit (Qiagen) in accordance with the manufacturer's instruction. Samples were then quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific) and quality of the isolated RNA was checked by visualization on agarose gels and by measuring the sample's A_{260}/A_{280} ratio.

Transcriptome Analysis

From the total RNA sample, 20 μg was reverse transcribed with 1,500 U SuperScript II reverse transcriptase (Invitrogen), 30 U SUPERase•In (Ambion), 750 ng random primer, 10 mM dNTP mixture

containing 4 mM amino-allyl dUTP, 10 mM DTT and 8 μg/mL actinomycin D. QIAquick PCR purification columns (Qiagen) were used to purify the amino-allyl labeled cDNAs. To protect amino-allyl residues, phosphate wash (5 mM KPO_4 and 80% ethanol) and elution buffer (4 mM KPO_4) were used instead of PE and PB buffers, respectively. The amino-allyl-labeled cDNAs were subsequently incubated with Cy5 monoreactive dyes (Amersham) to obtain Cy5 labeled cDNAs. The cDNA samples were fragmented by 0.3 U RNase-free DNaseI (Epicentre) per μg cDNA, and were then purified and hybridized onto the high-density oligonucleotide tiling microarrays. After hybridization and washing steps, the arrays were scanned on an Axon GenePix 4000B scanner and features were extracted by using NimbleScan software. Normalization of raw expression data and determination of differential gene expression between 37°C and 42°C was performed using the ANAIS web tool and the methods presented in the associated publication[115]. Fold changes and FDR-adjusted ANOVA p-values are reported in Table S4 in [73].

Supplementation Experiments

Bacterial Strains, Media, and Growth Conditions

Escherichia coli (Migula) Castellani and Chalmers MG1655 (ATCC 700926) was grown in glucose (4 g/L) minimal M9 medium containing 1X M9 salts (47.9 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.6 mM NaCl, and 18.7 mM NH_4Cl), 0.1 mM $CaCl_2$, 2 mM $MgSO_4$, and Trace Elements pH 7.1 (15 μM $FeCl_3$, 0.16 μM $ZnSO_4$, 0.18 μM $CuCl_2 \cdot 2H_2O$, 0.18 μM $MnSO_4$, 0.19 μM $CoCl_2$, and 4.4 μM Na_2EDTA). Volumes of 100 mL media were inoculated at 0.01 OD from washed cells of 37°C LB 5-7 h growth precultures. Supplementation for each condition was at 2 μM using various combinations of 4abz (4-aminobenzoic acid, CAS 150-13-0), btn (biotin, CAS 58-85-8), cit (citric acid, CAS 77-72-9), dhna (1,4-dihydrozy-2-napthoic acid, CAS 31519-22-9), g6p (D-glucose 6-phosphate dipotassium salt hydrate, EC 227-837-6), glyclt (glycolic acid, CAS 79-14-1), lipoate ((+/-)-α-lipoic acid, CAS 1077-28-7), pnto-R (D-pantothenic acid calcium salt, CAS 137-08-9), and ribflv ((-)-riboflavin, CAS 83-88-5). See Table 2.1 for additional details regarding these supplement compounds. Cocktail supplementation consisted of 4abz, btn, cit, dhna, g6p, glyclt, lipoate, pnto-R, and ribflv. Verification of MG1655 strain

was confirmed with forward primer AATGCCTGGAAATGGTTCAC and reverse primer AATAGGACGATTTGCGTTGC for 316 bp product.

Growth Curve Experimental Setup

Equipment for heat-dependent supplementation experiments consisted of Bellco Multi Stir-9 set at 650 rpm, as verified using MI Calibration strobe light. Water bath circulation and heating was performed at 42.0°C or 37.0°C with Lauda E100 control unit and temperature verification using certified Traceable Lollipop thermometer (certification number 4371-4382019). Cultures were grown in 500 mL Erlenmeyer flasks with 2" stir bars, aluminum foil wrapped around sides, and aluminum foil loosely-sealing the caps. Sampling was performed with 2 mL serological pipets. OD_{600} measurements were taken (see Table S6 in [73]) with a Thermo Spectronic Biomate 3 with readings up to 0.35 OD_{600} before dilution at 1X, 3X, 5X, 11X, and 17X made from 500 μL sample aliquots.

Growth Parameter Analysis

In order to compute growth parameters from the raw OD_{600} data at 37°C and 42°C, the natural log-transformed data was analyzed using the DMFit web tool (http://modelling.combase.cc/DMFit.aspx). The growth curve model of Baranyi and Roberts[116] was selected for all experimental data, and maximum growth rates, lag times, and R^2 values for the fitted model were collected (see Table S6 in [73]). Means and standard deviations for these growth parameters were then computed. Growth rate changes (Figure 2.8A) and lag time changes (Figure 2.8B) reported in this study were computed by taking the ratio of the mean supplementation experiment value to the mean no supplement control value and then subtracting one. Standard deviations for these changes (as reported in Figure 2.8) were computed based on the set of pairwise comparisons of all supplementation to all control replicates within an experiment.

Chapter 2 is a modified version of material in Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BØ. Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. *In*

preparation. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Chapter 3: Antibacterial mechanisms identified through structural systems pharmacology

Abstract

The continuously growing discipline of structural systems biology is applied prospectively in this study to predict pharmacological outcomes of antibacterial compounds in *Escherichia coli* K12. This work builds upon previously established methods for structural prediction of ligand binding pockets on protein molecules and utilizes and expands upon the previously developed genome scale model integrated with protein structures (GEM-PRO) for *E. coli*, structurally accounting for protein complexes. Carefully selected case studies are demonstrated to display the potential for this structural systems pharmacology framework in discovery and development of antibacterial compounds.

Introduction

A previously developed local structure homology-based approach to predicting ligand binding pockets[3, 4, 117] has been applied efficaciously in multiple contexts to study pharmacological phenomena[26, 49, 118, 119]. The recent development of a structural biology resource with which to study physiological stresses upon the proteome of *Escherichia coli* K12 MG1655 metabolism[73] enables a diversity of potential applications. Thus, we applied the SMAP methodology and the *E. coli* metabolic genome-scale model integrated with protein structures (GEM-PRO), to analyze and predict antibacterial effects of chemical compounds. *E. coli* K12, although not pathogenic under normal circumstances, is a

well-characterized laboratory model for enteropathogenic bacteria that infect humans. Thus methods, and perhaps even some specific predictions of antibacterial properties made in this study, are extensible to pathogenic *E. coli* and other bacterial pathogens. In addition to the integrative framework presented in this study for structural systems pharmacology, this effort also included significant expansion of the previously developed GEM-PRO to account for physiological assemblies of protein complex structures with activities accounted for in the *E. coli* K12 metabolic network *i*JO1366[74]. Results from this study show promising proof of principle for such an analysis framework and raise specific molecular and systemic hypothesis about antibacterials that are amenable to experimental testing.

Results

Many proteins do not act as monomers in the cell but as part of multimeric protein complexes that may include proteins encoded by one or several distinct genes. The previously constructed *Escherichia coli* genome-scale model integrated with protein structures (GEM-PRO)[73] considered proteins solely as single-peptide chains. As a result, we sought to expand the scope of this GEM-PRO to account for protein complexes. The structures of protein complexes are complementary to the existing single-peptide chain structures already included in the *E. coli* GEM-PRO. The objective was to best represent the physiological assemblies of metabolic enzyme complexes, that is, the best structural representation of the active form of enzyme complexes *in vivo*. A conceptual representation of this expansion with respect to one metabolic reaction is displayed in Figure 3.1A. The overall coverage of complexes in this reconstruction was 519 of 1106 known complexes (Figure 3.1B) included in the metabolic network *i*JO1366; 39% of complexes are completely represented by a single structure in the expanded GEM-PRO. Another 8% of complexes are partially represented by structures. This effort yielded 527 individual protein structure files, 149 of which were redundant with structures contained in the previously developed GEM-PRO. As is clear from Figure 3.1B, a slight majority of known complexes are not represented at all in the complex expansion to the GEM-PRO. A combination of the EcoCyc database[76], PDB structure curation[79], computational assessment of symmetry operations on the asymmetric unit of protein crystals[120], and literature review were used to identify the consensus most

physiologically accurate assemblies currently possible (approach summarized in Figure 3.2). These assemblies were distributed among different classes of oligomeric states: monomers, homomultimers, and heteromultimers (Figure 3.1B). The monomers directly overlap with contents previously reconstructed[73].

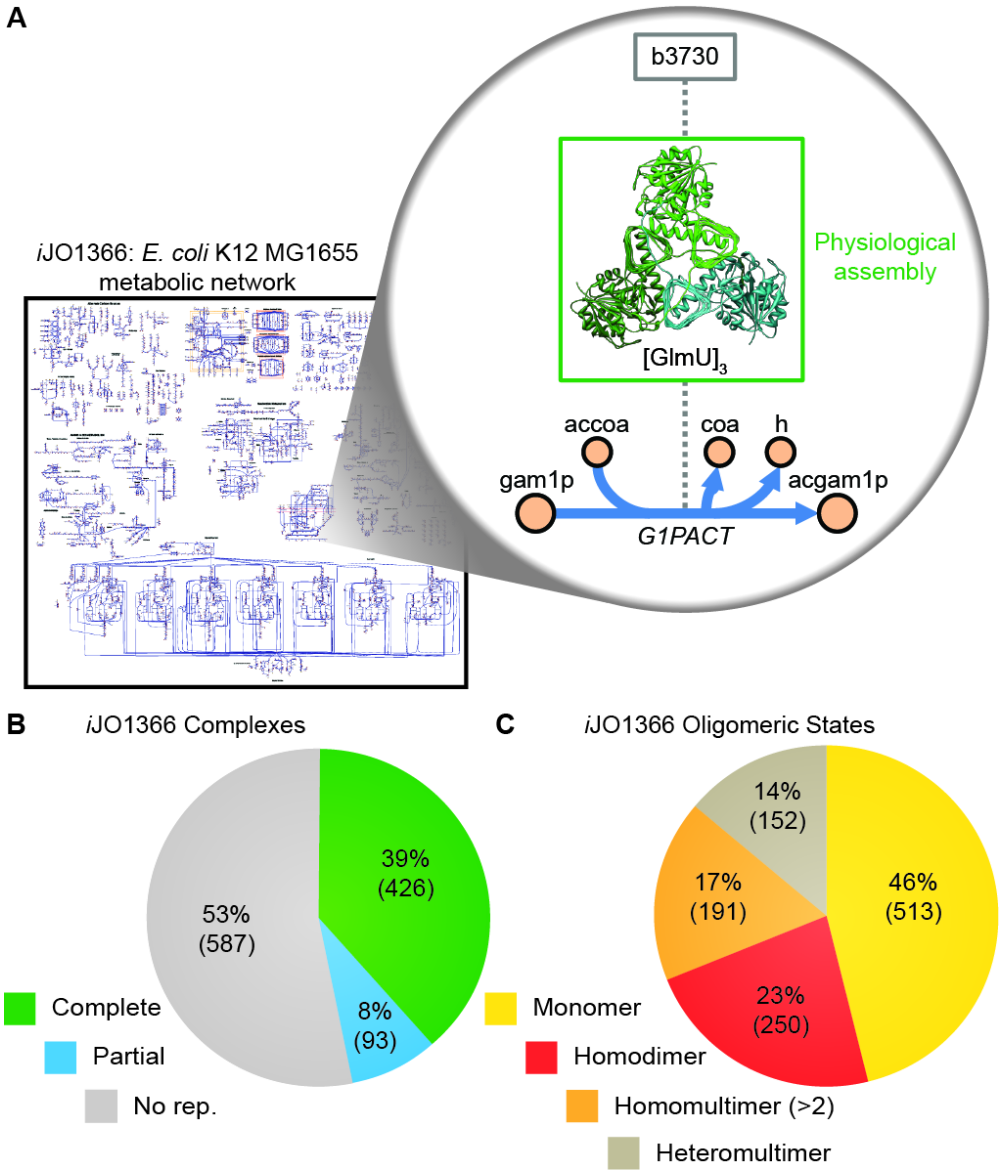

Figure 3.1. Complex expansion of *E. coli* GEM-PRO. (A) This expansion of the *E. coli* GEM-PRO provides structural coverage of protein complexes included in *i*JO1366. A conceptual illustration is depicted for the GlmU protein catalyzing the "G1PACT" reaction. **(B)** Complete and partial coverage of each protein complex by at least one structure is categorized. **(C)** The oligomeric states of complexes included in this expansion are distributed across monomers, homomultimers, and heteromultimers.

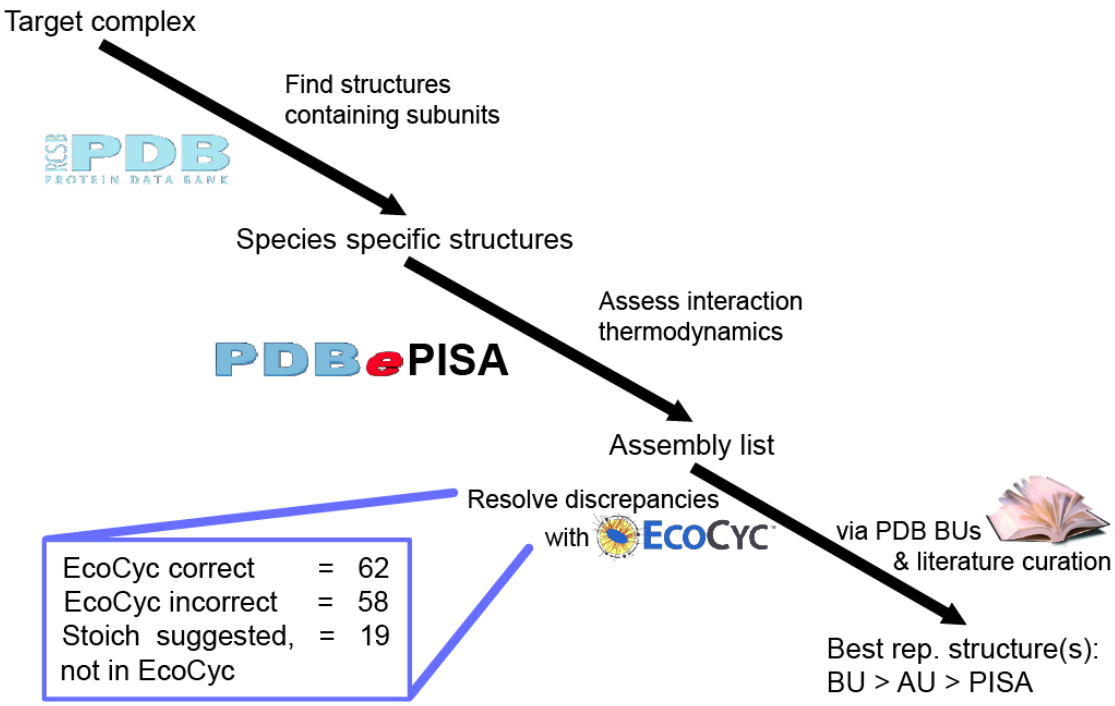

Figure 3.2. Complex physiological assembly reconstruction pipeline. The step-wise process of reconstructing physiological assemblies of protein complexes is summarized. The box at lower left notes the number of discrepancies and the outcome of their resolution.

The expanded *E. coli* GEM-PRO was then employed prospectively to explore possible currently unknown antibacterial properties. In particular, protein targets for orphan antibacterials, compounds known to have antibacterial effects but without known molecular targets, were predicted, and anti-metabolite compounds were also predicted as novel antibacterials to target promising metabolic protein targets without known inhibitors. The structural systems pharmacology approach taken here consisted of structure-based screens for protein-ligand targeting using the previously developed SMAP method[117] and metabolic model simulation to test the potential effects of predicted protein-ligand interactions. Some negative and positive controls were also screened, for which there is existing data on antibacterial capacity and at least some proven mechanisms of action within metabolism.

The results of these screens are summarized in Table 3.1. In the negative control screen for glucose (BGC) SMAP predicted that glucose binds to 7 individual metabolic *E. coli* proteins and 2

protein complexes significantly, 1 of which is a known target (MglB). Limiting the significance criteria to SMAP p-value showed that SMAP predicted a second known target (Glk) as well. Some of these targets are expected because glucose is a known substrate of these proteins, but the meaning of some of these remains unclear and could represent unknown alternative substrates of the proteins, unknown allosteric sites, or just false positives. Of the positive antibacterial controls, the top SMAP hit for the sulfonamide 4-amino-N-(1,3-thiazol-2-yl)benzenesulfonamide (YTZ) is the known primary target, dihydropteroate synthase (FolP). Two other positive controls, fosfomycin (FCN) and trimethoprim (TOP), were predicted by SMAP to bind significantly to a number of proteins (Table 3.1), none of which were known targets, leaving these predictions as uncertain validations or perhaps possible unknown side mechanisms leading to an antibacterial effect, which will be described further below. The positive control 2,2'-methanediylbis(3,4,6-trichlorophenol) (H3P) was not predicted to significantly bind any proteins; although the known primary target (FabI) was ranked 133rd out of 3303 protein structures, this result does constitute a false negative.

Table 3.1. Summary of *in silico* antibacterial screens.

Screen	Ligand ID	Target Name	SMAP prediction (significant)	Antibacterial Simulation	Functional Site Overlap
Negative control	BGC	-	-	-	-
Positive control: PEP analogue	FCN	BtuC	x	x	-
Positive control: sulfonamide	YTZ	FolP	x	x	-
Positive control: trimethoprim	TOP	RibD	x	x	x
		IspU	x	x	x
		EntA	x	x	x
		FabG	x	x	x
		KdtA	x	x	-
		MurJ	x	x	-
		WaaB	x	x	-
		MenH	x	x	-
		WaaQ	x	x	-
		MoeA	x	x	-
		TyrA	x	x	-
Positive control: chlorophenol	H3P	-	-	-	-
Orphan antibacterial	028	IspA	x	x	x
		IspB	x	x	x
	4AZ	-	-	-	-
	2OB	PheA	x	x	x
		AcpP	x	x	x
		EntA	x	x	x
		AtpB	x	x	x
		CyoB	x	x	x
		Cytochrome *bo* terminal oxidase	x	x	x
		Succinate dehydrogenase	x	x	x
		MurJ	x	x	-
		ProC	x	x	-
		ArgA	x	x	-
		IspU	x	x	-
		NuoB	x	x	-
		CyoC	x	x	-
		GdhA	x	x	-
		Ppk	x	x	-
		FadE	x	x	-
	TMM	-	-	-	-
Novel target: b1261	F6F	TrpB	x	x	x
	PLT	TrpB	x	x	x
	7MN	TrpB	x	x	x
	IDM	TrpB	x	x	x
	PLS	TrpB	x	x	x
Novel target: b2320	-	PdxB	-	x	-
Novel target: b3642	-	PyrE	-	x	-

The two orphan antibacterials 4-(aminomethyl)benzoic acid (4AZ) and 1,3,5-benzenetricarboxylic acid (TMM) were not predicted to significantly bind to any metabolic *E. coli*

proteins, providing no evidence to support a metabolic mechanism for these compounds. Since their mechanisms are currently not known in *E. coli*, this result may at least suggest searching elsewhere in the cellular system than metabolism for the mechanism. Intriguingly, the two other orphan antibacterials screened in this study, (1-hydroxyheptane-1,1-diyl)bis(phosphonic acid) (028) and cholesteryl oleate (2OB), were both predicted as significant binders by SMAP to multiple metabolic *E. coli* proteins (Table 3.1), suggesting possible mechanisms for their antibacterial activity.

Of the three screens aiming to identify anti-metabolite inhibitors of known essential genes in *E. coli*, SMAP predicted 5 candidate inhibitors for the tryptophan synthase β subunit (TrpB). The potential inhibitors of TrpB are listed in Table 3.1. SMAP screens for inhibitors of erythronate-4-phosphate dehydrogenase (PdxB) and orotate phosphoribosyltransferase (PyrE) failed to predict any significant candidate inhibitors.

Several other known metabolic targets of the control compounds were not predicted to be significant by SMAP. In our preliminary control screens, it was hypothesized that there may exist distinct binding pocket motifs for an individual compound such that using a single protein template to search for other targets may not identify all known targets of a compound. Expanding the number of search templates for a single compound, as was done for BGC, FCN, and TOP, indeed identified more significant targets, supporting this hypothesis. We were interested in seeing the relative accuracy of SMAP in predicting true positive protein-ligand interactions; thus we performed statistical analysis of the entire set of SMAP results including non-significant calls. Mann Whitney U-tests were run on the ranked lists of SMAP predictions with respect to each template protein structure, yielding inconsistently statistically significant p-values for some compounds (Figure 3.3). This result too supports that different binding motifs may exist for an individual compound, as is most apparent for BGC and TOP, which show the widest range of p-values. To highlight the overall efficacy of SMAP in predicting true positives, the results from all screens for a particular compound were combined by considering only the top rank number for each protein structure, whether a known target or not. It is apparent from Figure 3.3 that the examples BGC, FCN, TOP, and H3P all show a noticeable bias in favor of SMAP's predictive

accuracy; however, the stringency of significance criteria used may obscure this ability for many protein-ligand interactions. Because there is no obvious a priori approach to choosing a single structural template for screening a compound that may bind to multiple distinct motifs, our results suggest that using as wide array of diverse templates as appropriate should be considered when running SMAP screens. This phenomenon may explain some of the false negative SMAP predictions for controls in this study.

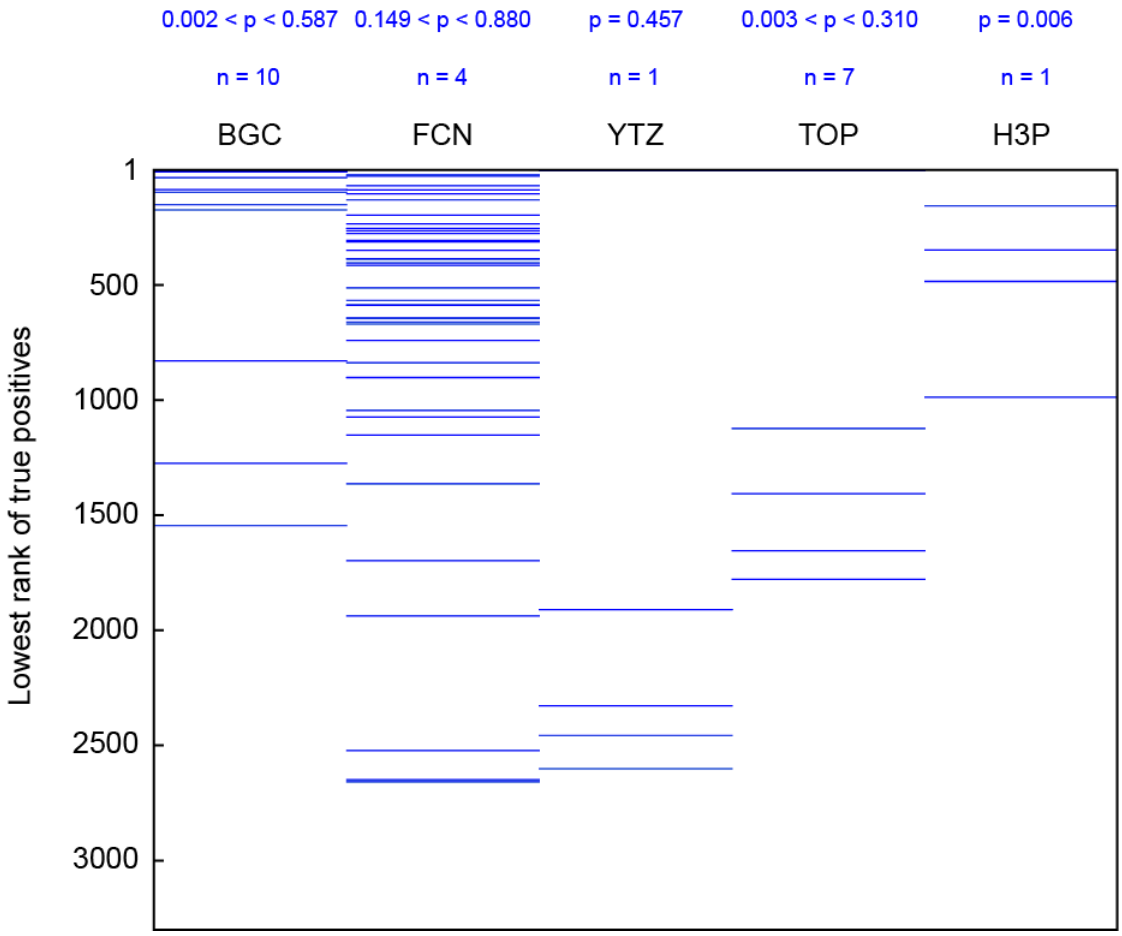

Figure 3.3. SMAP performance in recalling true positives. The lowest rank for each protein structure predicted as an SMAP hit is displayed for the set of known protein targets for the five control compounds. Blue lines indicate the rank position (out of 3303) of a known target for a given compound. n = the number of screens using different protein structure templates performed for each compound. p = the p-value resulting from Mann Whitney statistical tests for individual SMAP results with respect to an individual template screen. BGC: beta-D-glucose; FCN: fosfomycin; YTZ: 4-amino-N-(1,3-thiazol-2-yl)benzenesulfonamide ; TOP: trimethoprim; H3P: 2,2'-methanediylbis(3,4,6-trichlorophenol).

Next, we turned to the metabolic network portion of the *E. coli* GEM-PRO, *i*JO1366 [REF], to simulate the outcomes of known and predicted binding events leading to inhibition of protein activity and determine whether or not these events may be detrimental to growth. First, we found that inhibition of all known targets of all positive controls did lead to no growth or reduced growth rates in the model. In combination, the collective inhibition of all known targets for each positive control compound led to complete growth inhibition, but remarkably, most of these targets individually also led to complete loss of growth if completely inhibited. We also tested if inhibition of the individual protein targets predicted by gene-knockout phenotypes to be effective antibacterial targets leads to growth deficits in the model and found that all three individual inhibitions lead to no growth in the model. The effects of inhibition of SMAP-predicted targets were then evaluated in the model. Each of the individual predicted protein targets reported in Table 3.1 exhibited decreased or no growth upon full inhibition in simulation. These predictions helped to pare down the list of significant SMAP predictions to focus on those that satisfy both lines of evidence for antibacterial effects. With the exception of the FolP-YTZ binding interaction, all of the interactions reported in Table 3.1 are previously unknown, which suggests that in the case of positive control compounds, we may have uncovered previously unknown antibacterial targets. For the orphan antibacterial compounds, we predicted that inhibition of IspA and IspB by 028 leads to decreased growth rate and that inhibition of 14 individual proteins and 2 protein complexes by 2OB leads to decreased growth rate.

For the predicted protein-ligand interactions that also showed antibacterial effects in the metabolic model, we next utilized the residue-resolution functional annotation presented in the previously generated *E. coli* GEM-PRO to identify whether the SMAP-predicted ligand binding sites overlapped with known functional sites, such as catalytic and substrate binding sites. Such interactions could be expected to exhibit competitive inhibitory effects. For cases where an SMAP prediction was made on the basis of a protein complex structure, we also identified predicted ligand binding sites at the interface between subunits, which may lead to disruption or prevention of protein complex formation in vivo and therefore have a deleterious impact on enzyme function. Overlap between predicted TOP binding sites and native nucleotide and substrate binding sites occurred on RibD, partial overlap with

the catalytic site of IspU, and almost complete overlap with the catalytic sites of both EntA and FabG. The predicted binding sites for 028 completely overlapped with the catalytic site of IspA and overlapped with the substrate binding site and a Mg^{2+} ion binding site of IspB. In the case of 2OB, predicted binding sites showed at least partial overlap with the catalytic sites of PheA, CyoB, EntA, AtpB, and AcpP. Predicted 2OB binding sites also had implications with respect to two protein complexes not exhibited with respect to the complex subunits in isolation. The predicted 2OB binding site on the cytochrome *bo* terminal oxidase appears at the interaction site between CyoB and CyoC. The 2OB binding site also overlapped with the heme binding sites of the SdhC and SdhD subunits of the succinate dehydrogenase complex as well as the protein-protein interaction region between these subunits. These last few predictions speak to the importance of the complex expansion of the GEM-PRO, without which such molecular predictions involving multiple subunit interfaces would not have been possible.

Discussion

In this study, we have demonstrated the first structural systems pharmacology antibacterial screens for the model bacterium *E. coli*. This effort was enabled in part through the complex expansion of the *E. coli* GEM-PRO. In the first attempt at this reconstruction, we chose to utilize solely structures supported by strong experimental; however, this could be further expanded through modeling of protein complex structures as has been attempted by others recently[68]. Our previous and current efforts at reconstructing the *E. coli* metabolic GEM-PRO have enabled *in silico* exploration of diverse physicochemical stress, but given the relative novelty of this resource, much broader expansions are likely to emerge and enable still more diverse avenues of analysis.

This study illustrates another example of how structural and systems biology combine to an effect greater than they are capable of in isolation. For example, some of the SMAP hits of lesser quantitative significance showed promise as antibacterial targets in simulation, sometimes accounting for known antibacterial targets that otherwise would have been called as false negatives by SMAP alone. Conversely, although metabolic model predictions have previously been shown to accurately

predict the effects of many targeted gene knockouts[74] and have been applied to select individual and multiple antibacterial targets[11, 121], these metabolic models have not yet been capable of pairing these targets with compounds. Not only does the expansion from the GEM to GEM-PRO framework enable prediction of candidate compounds, it enables prediction of specific molecular mechanisms (e.g. competitive inhibition or complex disruption) that explain how the candidate compounds may affect the function of their targets.

In addition to providing a promising proof of principle that such a structural systems biology strategy can be used to understand antibacterial mechanisms, we have made specific predictions of novel candidate antibacterial compounds that target a protein currently unutilized for antibacterial applications (TrpB) and previously unknown mechanisms of existing antibacterial compounds, both those with and without established mechanisms. Future work will aim to experimentally validate some of the predictions presented in this study. These experiments would first assess the phenotypic predictions of compound treatment through measurement of growth rates in wild type (WT) *E. coli*. WT screens can show whether or not a compound has antibacterial properties. Comparison of WT and gene-knockout mutant sensitivities to varying concentrations will provide a basis for validating the predicted targets as well; a lesser dosage-dependent effect in a gene-knockout strain supports targeting of the knocked-out protein by the candidate antibacterial. Milder concentrations of growth-inhibiting compounds would serve as a condition under which metabolomics measurements can be made to confirm the prediction of targeted metabolic pathways, providing direct evidence of specific response phenotypes and an additional line of indirect evidence for antibacterial target validation. Finally, molecular predictions of protein-ligand interactions may be directly assessed through ligand-binding assays. Seeding the relatively simple experimental workflow described with the structural systems pharmacology framework we present here should permit rapid discovery in the area of antibacterials.

Methods

Complex expansion of the *E. coli* GEM-PRO

Enzyme complexes included in the metabolic network *i*JO1366[74] were reviewed as annotated in EcoCyc[76]. The annotation from EcoCyc includes protein subunit compositions, which served as a starting point for this reconstruction. The EcoCyc subunit compositions were evaluated from a structural perspective based on biological units of crystal structures in the PDB[79] and through thermodynamic analysis of possible physiological assemblies using the PDBePISA software[120]. The most thermodynamically feasible PISA assembly for each complex, based on computed ΔG of dissociation, was compared to PDB biological units and EcoCyc composition annotation for each complex. In many cases, these three sources were in perfect agreement, in which case the PDB biological unit was chosen as the structure to represent the physiological assembly of the complex. However, many discrepancies were also found among the compositions assigned by these sources, including protein membership in complexes but missing stoichiometries in EcoCyc. To reconcile these discrepancies, the scientific literature was reviewed to find experimental evidence supporting the correct physiological assembly for a complex. These references reported data from a variety of experiments including: X-ray crystallography, gel filtration, size-exclusion chromatography, ultracentrifugation, functional assays, substrate binding assays, cooperative analysis, and mutant studies. A few studies also provided evidence from bioinformatic analysis such as kinetic assembly, molecular docking, and inference based on knowledge about orthologous structures. The consensus of these experimental results and the three preliminary sources was taken to determine the most likely physiological assembly. If the PDB biological unit agreed with the consensus, that structure was taken as the physiological assembly structure. If not, then the PISA structure that best agreed with the consensus was taken as the physiological assembly. In some cases, no PDB structure or PISA assembly completely accounted for the consensus complex assembly. In such cases, multiple structures were taken to represent as many sub-parts of the physiological complex assembly as possible. This resulted in some overlap with single-peptide chain structures included in the previously developed *E. coli* GEM-PRO.

Protein-ligand interaction predictions

Different types of SMAP screens were run to answer three different types of questions: 1) positive and negative controls for antibacterials with known effective mechanisms in wild type *E. coli* K12 through known metabolic protein targets; 2) orphan antibacterials known to be effective against *E. coli* K12 but with unknown mechanisms, seeking to answer the question of whether those compounds may target metabolic functions; 3) searches for potential novel antibacterials that are competitive inhibitors of metabolic proteins known to hinder growth of *E. coli* K12 if subjected to gene knockout. These are all open-ended questions, and candidate compounds and protein targets to be selected for these purposes are not immediately obvious. Also because SMAP is a method requiring substantial computational resources, the number of screens that could be performed was limited. For these reasons, filtering the wealth of candidate compounds and targets to choose candidates for the screens was necessary. Therefore, large data sources were filtered to pick most promising candidates to test these three types of questions.

Selecting antibacterial controls for screen

At the time of this writing, there are 12,785 chemically distinct ligand molecules represented in at least one PDB structure. Given that SMAP performs best when starting with a well-defined ligand binding site for the search template, we chose only to use experimentally-determined binding sites for this type of screen. The collection of all known antibacterials and their known targets was collected from KEGG[98], EcoCyc[76], DrugBank[122], and ChEMBL[123], and the overlapping set of these and the PDB ligands was found. Antibiotic classifications were derived from KEGG, EcoCyc, and DrugBank. All PDB ligands were clustered by their chemical similarity using their canonical SMILES[52] and the EI-Clustering software[124]. The distance matrix output by EI-Clustering was used to form the clusters by hierarchical clustering and cutoff of 1.15 was determined such that the classified antibiotics were clustered together and not in the same clusters with antibiotics of other classes. Thus functionally and chemically distinct groups of antibacterials were identified from which to choose positive controls. Positive controls were chosen from these groupings such that they represented a breadth of diversity and only if they had at least one known metabolic protein target in *E. coli*. Glucose was chosen as a

negative control since it does not have inhibitory effects on growth, and its cellular binding partner proteins mostly are not negatively impacted through its interaction.

Selecting orphan antibacterials for screen

The ChEMBL database[123] was reviewed to find biological assays in which antibacterial activity of compounds was identified in *E. coli*. Within this set of compounds, we search for those with no known binding partners encoded by *E. coli* in KEGG, EcoCyc, DrugBank, ChEMBL, or the PDB. We then prioritized for those compounds that are ligands in PDB structures of only non-bacterial proteins. Small compounds consisting only of C, H, N, O, P, and S elements were chosen from this set as the orphan antibacterials of interest for this study.

Selecting antibacterial protein targets for screen

Previously published essentiality screens of the *E. coli* K12 single-gene knockout library[74] were analyzed to choose novel antibacterial protein targets to search for anti-metabolites to inhibit. Phenotypes that showed a small measure of growth but dramatically hindered (minimum $OD_{600} < 0.26$) were selected from, prioritizing for high PDB ligand cluster size, low number of native metabolic substrates, and extent of structural coverage in the GEM-PRO.

Prediction of antibacterial mechanism

In searching for possible metabolic protein targets for known antibacterial compounds, template structures were chosen from PDB crystal structures that included the compound bound to a protein. These structures were used with SMAP to search for potential binding pockets for these antibacterial compounds within both the previously published *E. coli* GEM-PRO and also the newly-generated physiological complex assemblies. The entire set of PDB proteins was clustered using a 50% sequence identity cutoff. The best resolution structure from each cluster that contained the ligand of interest was chosen as an alternative template for SMAP screens. SMAP was run with default numerical parameters screening each template in turn across the database of proteins comprising the GEM-PRO structures. SMAP hits were considered significant if p-value $< 1.0 \times 10^{-4}$ and Tanimoto coefficient > 0.5. A secondary tier of lesser significance was determined using just the aforementioned p-value criterion.

Prediction of anti-metabolite protein inhibitors

Searching for possible inhibitors of predicted antibacterial metabolic protein targets was performed by taking the structure of the protein target of interest from the *E. coli* GEM-PRO, docking[125] the primary native metabolic substrate into the known catalytic site (as annotated in the GEM-PRO), and using the resulting structure as a template for SMAP screens. SMAP was then used to search across all ligand-bound protein structures in the PDB, excluding structures that only bind metal ions or metabolites included in iJO1366, to find ligands that bind to structurally similar sites. The query database contained 51,608 PDB structures. SMAP was run with default numerical parameters and specifying that only ligand binding sites be considered. SMAP hits were considered significant if p-value $< 1.0\times10^{-4}$ and Tanimoto coefficient > 0.5. A secondary tier of lesser significance was determined using just the aforementioned p-value criterion.

Simulating protein inhibitory effects

The *E. coli* metabolic network *i*JO1366[74] was loaded into the COBRA toolbox[112] from the published SBML model using Matlab. Since the time of publication of *i*JO1366 a thermodynamic constraint error was discovered in the published model; as a result, the malate oxidase, "MOX," reaction was set as irreversible. The superoxide dismutase, "SPODM," reaction was set with an initial upper bound of 1000 as well. The objective function was set as the complete wild type biomass reaction "Ec_biomass_*i*JO1366_WT_53p95M." Default exchange reaction constraints were used, except for a glucose uptake lower bound of -8 mmol/gDW/h and an oxygen uptake lower bound of -18.5 mmol/gDW/h, representing aerobic growth on glucose. These basic constraints were used for all reported simulations in this study.

The combined sets of known targets and predicted targets were first tested for antibacterial effects by constraining all associated reactions to 0 flux and then maximizing biomass using flux balance analysis (FBA)[111]. Individual targets were tested in the same manner to determine causal targets from the broader sets. Resulting biomass fluxes were compared to a simulated untreated condition where just the basic constraints were imposed and biomass was maximized; any decrease in biomass

flux relative to the untreated condition was treated as a prediction of antibacterial effect by degree of decrease.

Analysis of impact of protein-ligand binding on molecular function

The specific amino acid residues comprising the ligand binding sites predicted by SMAP were compared to residue-resolution functional annotation contained in the original GEM-PRO[73]. If precise residues overlapped between these sets, we flagged these proteins as having predicted binding sites for the given ligand that should be seen as competitively inhibitory since they would bind to the same location as substrates required for normal function. Functional features included in this analysis consisted of catalytic sites and substrate binding sites. For SMAP query structures that were protein complexes containing multiple subunits, if the predicted ligand binding site included residues from distinct subunits, we flagged these as possible ligand binding events that could prevent or disrupt complex formation and therefore function.

Chapter 3 is a modified version of material in Chang RL, Bourne PE, Palsson BØ. Antibacterial mechanisms identified through structural systems pharmacology. *In preparation*. I was the primary author, while the co-authors provided support in the research that served as the basis for this study.

Conclusion: Structural systems biology, present and future

Biological systems are composed of both the "visible" and the "invisible," the "visible" being the molecular components that make up cells, the "invisible" being the governing constraints that drive the movement, interaction, and transformation of the molecular components. The traditional structural biologist focuses largely on the study of proteins, the main "visible" actors in the cell, and the form-to-function relationship so intricately specified by the subtleties of structure. Systems biology, still a relatively young field in itself, grasps at the "invisible," often forced to seek explanations for biological phenomena through mathematical mimicry. Seemingly distant fields, but perhaps there is a rich land of discovery somewhere in-between. It is this notion that inspired the research in this dissertation, to explore this relatively untouched frontier.

Armed with the benefit of molecular detail conferred by protein structures and the theoretical scaffolds that are genome-scale models, new ground was broken in the analysis of response to two distinct kinds of physicochemical stress, that of exogenous chemicals and that of exposure to super-optimal temperature. In the course of this work, I moved from studying a human system to studying a bacterial system, both using the framework of structural systems biology. A new structural systems biology data source was created in the form of the genome-scale model integrated with protein structures (GEM-PRO). This resource was designed with the motivation of integrating computational tools of both structural bioinformatics and systems biology, and it was demonstrated to be effective in

predicting systemic response to physicochemical stresses through the impact that these stresses exert differentially upon the collection of distinct proteins that make up the proteome. Furthermore, the effectiveness of this merging of fields was distinctly shown to only work with both components; the same predictions could not have been made using either set of tools in isolation.

The applications of such an analysis platform are very far reaching. In the course of this thesis alone, we have explored issues of drug side effects in treating human disease, gained new biological understanding of the response of a microbe to high temperature environments, and provided a basis for predicting antibacterial activity. In part, these applications make up the scheme displayed in Figure 4.1. This figure also hints at the utility that a structural systems biology approach may have in understanding native mechanisms for regulation, such as post-translational modification of proteins, but these applications most likely barely scratch the surface of what can be done with this integrative framework. One can easily imagine empowering a protein engineering, or even strain engineering, study by adopting such a research perspective or adding new dimensionality to evolutionary studies.

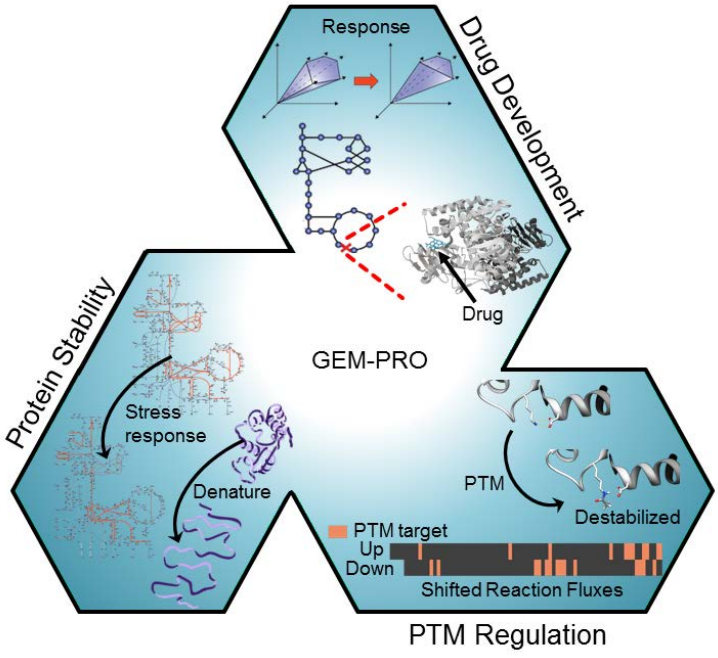

Figure 4.1. Established and prospective applications of structural systems biology. Drug development in the form of human-targeted pharmaceuticals (Chapter 1) and antibacterials (Chapter 3) was presented in this work. The impact of environment stress upon protein stability was exemplified in Chapter 2. Post-translational protein modification (PTM) represents another frontier for structural systems biology.

Where is this field heading? When I first took interest in systems biology, I said that I dreamt of dynamically simulating cellular systems in three-dimensions. We are getting closer with these early steps into structural systems biology, but I am not watching molecules diffuse across my screen in an accurate representation of the cellular milieu right now. As computing technology continues to become more powerful and as new, faster, higher-throughput experimental techniques are developed, one can start to more realistically expect that we will be simulating and visualizing the cellular milieu soon enough. Meanwhile, we will focus on the more tangible but equally remarkable: interpretation of genomic variation leading to human disease and personalized medicine, finding the genetic linchpins limiting industrial microbe output, managing invasive pathogens without disturbing the native microflora, and a plethora of other open scientific problems I've seen hints towards while completing this thesis.

The invisible mirror reflects with greater clarity if it is not blind to the visible stage.

References

1. M. K. Hellerstein. Exploiting complexity and the robustness of network architecture for drug discovery. *J Pharmacol Exp Ther*. **325**, 1 (2008).

2. D. B. Searls. Data integration: challenges for drug discovery. *Nat Rev Drug Discov*. **4**, 45 (2005).

3. L. Xie, P. E. Bourne. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A*. **105**, 5441 (2008).

4. L. Xie, J. Li, P. E. Bourne. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput Biol*. **5**, e1000387 (2009).

5. G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason, A. L. Hopkins. Global mapping of pharmacological space. *Nat Biotechnol*. **24**, 805 (2006).

6. A. L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*. **4**, 682 (2008).

7. A. M. Feist, B. O. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol*. **26**, 659 (2008).

8. M. A. Oberhardt, B. O. Palsson, J. A. Papin. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*. **5**, 320 (2009).

9. N. Jamshidi, B. O. Palsson. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol*. **1**, 26 (2007).

10. T. Y. Kim, H. U. Kim, S. Y. Lee. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metab Eng*. **12**, 105 (2010).

11. D. S. Lee, H. Burd, J. Liu, E. Almaas, O. Wiest, A. L. Barabasi, Z. N. Oltvai, V. Kapatral. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J Bacteriol*. **191**, 4015 (2009).

12. N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, B. O. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*. **104**, 1777 (2007).

13. S. A. Becker, B. O. Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*. **4**, e1000082 (2008).

14. T. Shlomi, M. N. Cabili, M. J. Herrgard, B. O. Palsson, E. Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*. **26**, 1003 (2008).

15. G. J. de Grooth, J. A. Kuivenhoven, A. F. Stalenhoef, J. de Graaf, A. H. Zwinderman, J. L. Posma, A. van Tol, J. J. Kastelein. Efficacy and safety of a novel cholesteryl ester transfer

protein inhibitor, JTT-705, in humans: a randomized phase II dose-response study. *Circulation*. **105**, 2159 (2002).

16. P. J. Barter, M. Caulfield, M. Eriksson, S. M. Grundy, J. J. Kastelein, M. Komajda, J. Lopez-Sendon, L. Mosca, J. C. Tardif, D. D. Waters, C. L. Shear, J. H. Revkin, K. A. Buhr, M. R. Fisher, A. R. Tall, B. Brewer. Effects of torcetrapib in patients at high risk for coronary events. *N Engl J Med*. **357**, 2109 (2007).

17. A. R. Tall, L. Yvan-Charvet, N. Wang. The failure of torcetrapib: was it the molecule or the mechanism? *Arterioscler Thromb Vasc Biol*. **27**, 257 (2007).

18. R. Krishna, M. S. Anderson, A. J. Bergman, B. Jin, M. Fallon, J. Cote, K. Rosko, C. Chavez-Eng, R. Lutz, D. M. Bloomfield, M. Gutierrez, J. Doherty, F. Bieberdorf, J. Chodakewitz, K. M. Gottesdiener, J. A. Wagner. Effect of the cholesteryl ester transfer protein inhibitor, anacetrapib, on lipoproteins in patients with dyslipidaemia and on 24-h ambulatory blood pressure in healthy individuals: two double-blind, randomised placebo-controlled phase I studies. *Lancet*. **370**, 1907 (2007).

19. J. A. Kuivenhoven, G. J. de Grooth, H. Kawamura, A. H. Klerkx, F. Wilhelm, M. D. Trip, J. J. Kastelein. Effectiveness of inhibition of cholesteryl ester transfer protein by JTT-705 in combination with pravastatin in type II dyslipidemia. *Am J Cardiol*. **95**, 1085 (2005).

20. M. J. Forrest, D. Bloomfield, R. J. Briscoe, P. N. Brown, A. M. Cumiskey, J. Ehrhart, J. C. Hershey, W. J. Keller, X. Ma, H. E. McPherson, E. Messina, L. B. Peterson, W. Sharif-Rodriguez, P. K. Siegl, P. J. Sinclair, C. P. Sparrow, A. S. Stevenson, S. Y. Sun, C. Tsai, H. Vargas, M. Walker, 3rd, S. H. West, V. White, R. F. Woltmann. Torcetrapib-induced blood pressure elevation is independent of CETP inhibition and is accompanied by increased circulating levels of aldosterone. *Br J Pharmacol*. **154**, 1465 (2008).

21. M. Hermann, F. T. Ruschitzka. The hypertension peril: lessons from CETP inhibitors. *Curr Hypertens Rep*. **11**, 76 (2009).

22. K. Sangkuhl, D. S. Berlin, R. B. Altman, T. E. Klein. PharmGKB: understanding the effects of individual genetic variants. *Drug Metab Rev*. **40**, 539 (2008).

23. Y. Konno, K. Kimura. Vasodilatory effect of cilnidipine, an L-type and N-type calcium channel blocker, on rat kidney glomerular arterioles. *Int Heart J*. **49**, 723 (2008).

24. D. C. Hatton, D. A. McCarron. Dietary calcium and blood pressure in experimental models of hypertension. A review. *Hypertension*. **23**, 513 (1994).

25. L. J. Appel, T. J. Moore, E. Obarzanek, W. M. Vollmer, L. P. Svetkey, F. M. Sacks, G. A. Bray, T. M. Vogt, J. A. Cutler, M. M. Windhauser, P. H. Lin, N. Karanja. A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. *N Engl J Med*. **336**, 1117 (1997).

26. R. L. Chang, L. Xie, P. E. Bourne, B. O. Palsson. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol*. **6**, e1000938 (2010).

27. T. Nakayama, M. Soma, Y. Watanabe, B. Hasimu, M. Sato, N. Aoi, K. Kosuge, K. Kanmatsuse, S. Kokubun, J. D. Marrow, J. A. Oates. Splicing mutation of the prostacyclin synthase gene in a family associated with hypertension. *Biochem Biophys Res Commun*. **297**, 1135 (2002).

28. T. Ito, T. Okada, J. Mimuro, H. Miyashita, R. Uchibori, M. Urabe, H. Mizukami, A. Kume, M. Takahashi, U. Ikeda, Y. Sakata, K. Shimada, K. Ozawa. Adenoassociated virus-mediated prostacyclin synthase expression prevents pulmonary arterial hypertension in rats. *Hypertension*. **50**, 531 (2007).

29. J. M. Jones, J. C. Morrell, S. J. Gould. Identification and characterization of HAOX1, HAOX2, and HAOX3, three human peroxisomal 2-hydroxy acid oxidases. *J Biol Chem*. **275**, 12590 (2000).

30. S. J. Lee, J. Liu, N. Qi, R. A. Guarnera, S. Y. Lee, G. T. Cicila. Use of a panel of congenic strains to evaluate differentially expressed genes as candidate genes for blood pressure quantitative trait loci. *Hypertens Res*. **26**, 75 (2003).

31. M. Miyamoto, Y. Yoshida, I. Taguchi, Y. Nagasaka, M. Tasaki, Y. Zhang, B. Xu, M. Nameta, H. Sezaki, L. M. Cuellar, T. Osawa, H. Morishita, S. Sekiyama, E. Yaoita, K. Kimura, T. Yamamoto. In-depth proteomic profiling of the normal human kidney glomerulus using two-dimensional protein prefractionation in combination with liquid chromatography-tandem mass spectrometry. *J Proteome Res*. **6**, 3680 (2007).

32. S. Kitanaka, K. Takeyama, A. Murayama, T. Sato, K. Okumura, M. Nogami, Y. Hasegawa, H. Niimi, J. Yanagisawa, T. Tanaka, S. Kato. Inactivating mutations in the 25-hydroxyvitamin D3 1alpha-hydroxylase gene in patients with pseudovitamin D-deficiency rickets. *N Engl J Med*. **338**, 653 (1998).

33. J. T. Wang, C. J. Lin, S. M. Burridge, G. K. Fu, M. Labuda, A. A. Portale, W. L. Miller. Genetics of vitamin D 1alpha-hydroxylase deficiency in 17 families. *Am J Hum Genet*. **63**, 1694 (1998).

34. C. J. Dickinson, J. M. Smellie. Xanthinuria. *Br Med J*. **2**, 1217 (1959).

35. Y. Sasaki, M. Iseki, S. Yamaguchi, Y. Kurosawa, T. Yamamoto, Y. Moriwaki, T. Kenri, T. Sasaki, R. Yamashita. Direct evidence of autosomal recessive inheritance of Arg24 to termination codon in purine nucleoside phosphorylase gene in a family with a severe combined immunodeficiency patient. *Hum Genet*. **103**, 81 (1998).

36. K. Hyland, P. T. Clayton. Aromatic amino acid decarboxylase deficiency in twins. *J Inherit Metab Dis*. **13**, 301 (1990).

37. G. Eshel, E. Lahat, K. Fried, J. Barr, V. Barash, A. Gutman, S. DiMauro, M. Aladjem. Autosomal recessive lethal infantile cytochrome C oxidase deficiency. *Am J Dis Child*. **145**, 661 (1991).

38. M. Zeviani, I. Nonaka, E. Bonilla, E. Okino, M. Moggio, S. Jones, S. DiMauro. Fatal infantile mitochondrial myopathy and renal dysfunction caused by cytochrome c oxidase deficiency: immunological studies in a new patient. *Ann Neurol*. **17**, 414 (1985).

39. R. W. Moreadith, M. L. Batshaw, T. Ohnishi, D. Kerr, B. Knox, D. Jackson, R. Hruban, J. Olson, B. Reynafarje, A. L. Lehninger. Deficiency of the iron-sulfur clusters of mitochondrial reduced nicotinamide-adenine dinucleotide-ubiquinone oxidoreductase (complex I) in an infant with congenital lactic acidosis. *J Clin Invest*. **74**, 685 (1984).

40. H. Jacquet, J. Berthelot, C. Bonnemains, G. Simard, P. Saugier-Veber, G. Raux, D. Campion, D. Bonneau, T. Frebourg. The severe form of type I hyperprolinaemia results from homozygous inactivation of the PRODH gene. *J Med Genet*. **40**, e7 (2003).

41. V. Humbertclaude, F. Rivier, A. Roubertie, B. Echenne, H. Bellet, C. Vallat, D. Morin. Is hyperprolinemia type I actually a benign trait? Report of a case with severe neurologic involvement and vigabatrin intolerance. *J Child Neurol*. **16**, 622 (2001).

42. T. L. Perry, D. F. Hardwick, R. B. Lowry, S. Hansen. Hyperprolinaemia in two successive generations of a North American Indian family. *Ann Hum Genet*. **31**, 401 (1968).

43. P. de Lonlay, I. Valnot, A. Barrientos, M. Gorbatyuk, A. Tzagoloff, J. W. Taanman, E. Benayoun, D. Chretien, N. Kadhom, A. Lombes, H. O. de Baulny, P. Niaudet, A. Munnich, P. Rustin, A. Rotig. A mutant mitochondrial respiratory chain assembly protein causes complex III deficiency in patients with tubulopathy, encephalopathy and liver failure. *Nat Genet*. **29**, 57 (2001).

44. A. B. Zinn, D. S. Kerr, C. L. Hoppel. Fumarase deficiency: a new cause of mitochondrial encephalomyopathy. *N Engl J Med*. **315**, 469 (1986).

45. C. Gellera, G. Uziel, M. Rimoldi, M. Zeviani, A. Laverda, F. Carrara, S. DiDonato. Fumarase deficiency is an autosomal recessive encephalopathy affecting both the mitochondrial and the cytosolic enzymes. *Neurology*. **40**, 495 (1990).

46. L. P. van den Heuvel, K. Assink, M. Willemsen, L. Monnens. Autosomal recessive renal glucosuria attributable to a mutation in the sodium glucose cotransporter (SGLT2). *Hum Genet*. **111**, 544 (2002).

47. H. L. Teijema, H. H. van Gelderen, M. A. Giesberts, M. S. Laurent de Angulo. Dicarboxylic aminoaciduria: an inborn error of glutamate and aspartate transport with metabolic implications, in combination with a hyperprolinemia. *Metabolism*. **23**, 115 (1974).

48. L. J. Elsas, R. E. Hillman, J. H. Patterson, L. E. Rosenberg. Renal and intestinal hexose transport in familial glucose-galactose malabsorption. *J Clin Invest*. **49**, 576 (1970).

49. S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, P. E. Bourne. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput Biol*. **5**, e1000423 (2009).

50. J. D. Durrant, R. E. Amaro, L. Xie, M. D. Urbaniak, M. A. Ferguson, A. Haapalainen, Z. Chen, A. M. Di Guilmi, F. Wunder, P. E. Bourne, J. A. McCammon. A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput Biol*. **6**, e1000648 (2010).

51. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. Basic local alignment search tool. *J Mol Biol*. **215**, 403 (1990).

52. D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*. **28**, 31 (1988).

53. M. F. Sanner. Python: a programming language for software integration and development. *J Mol Graph Model*. **17**, 57 (1999).

54. O. Trott, A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. **31**, 455 (2010).

55. I. M. Frey, I. Rubio-Aliaga, A. Siewert, D. Sailer, A. Drobyshev, J. Beckers, M. H. de Angelis, J. Aubert, A. Bar Hen, O. Fiehn, H. M. Eichinger, H. Daniel. Profiling at mRNA, protein, and metabolite levels reveals alterations in renal amino acid handling and glutathione metabolism in kidney tissue of Pept2-/- mice. *Physiol Genomics*. **28**, 301 (2007).

56. S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. O. Palsson, M. J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*. **2**, 727 (2007).

57. O. Shmueli, S. Horn-Saban, V. Chalifa-Caspi, M. Shmoish, R. Ophir, H. Benjamin-Rodrig, M. Safran, E. Domany, D. Lancet. GeneNote: whole genome expression profiles in normal human tissues. *C R Biol*. **326**, 1067 (2003).

58. W. S. Cleveland, S. J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*. **83**, 596 (1988).

59. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. **5**, R80 (2004).

60. F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G. A. Danieli, S. Bicciato. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*. **8**, 446 (2007).

61. M. J. Schlesinger. Heat shock proteins. *J Biol Chem*. **265**, 12111 (1990).

62. E. Van Derlinden, K. Bernaerts, J. F. Van Impe. Dynamics of Escherichia coli at elevated temperatures: effect of temperature history and medium. *J Appl Microbiol*. **104**, 438 (2008).

63. V. Potapov, M. Cohen, G. Schreiber. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel*. **22**, 553 (2009).

64. A. Korkegian, M. E. Black, D. Baker, B. L. Stoddard. Computational thermostabilization of an enzyme. *Science*. **308**, 857 (2005).

65. C. R. Fischer, D. Klein-Marcuschamer, G. Stephanopoulos. Selection and optimization of microbial hosts for biofuels production. *Metab Eng*. **10**, 295 (2008).

66. P. Beltrao, C. Kiel, L. Serrano. Structures in systems biology. *Curr Opin Struct Biol*. **17**, 378 (2007).

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

68. Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, B. Honig. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*. **490**, 556 (2012).

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

70. T. M. Cheng, L. Goehring, L. Jeffery, Y. E. Lu, J. Hayles, B. Novak, P. A. Bates. A structural systems biology approach for quantifying the systemic consequences of missense mutations in proteins. *PLoS Comput Biol*. **8**, e1002738 (2012).

71. O. Tenaillon, A. Rodriguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, B. S. Gaut. The molecular diversity of adaptive convergence. *Science*. **335**, 457 (2012).

72. I. K. Blaby, B. J. Lyons, E. Wroclawska-Hughes, G. C. Phillips, T. P. Pyle, S. G. Chamberlin, S. A. Benner, T. J. Lyons, V. Crecy-Lagard, E. Crecy. Experimental evolution of a facultative thermophile from a mesophilic ancestor. *Appl Environ Microbiol*. **78**, 144 (2012).

73. R. L. Chang, K. Andrews, D. Kim, Z. Li, A. Godzik, B. Ø. Palsson. Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. *In preparation*.

74. J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, B. O. Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol*. **7**, 535 (2011).

75. C. T. Porter, G. J. Bartlett, J. M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*. **32**, D129 (2004).

76. I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muniz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, P. D. Karp. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res*. **39**, D583 (2011).

77. Z. Huang, L. Zhu, Y. Cao, G. Wu, X. Liu, Y. Chen, Q. Wang, T. Shi, Y. Zhao, Y. Wang, W. Li, Y. Li, H. Chen, G. Chen, J. Zhang. ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res.* **39**, D663 (2011).

78. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40**, D71 (2012).

79. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*. **28**, 235 (2000).

80. M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Sohngen, M. Stelzer, J. Thiele, D. Schomburg. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res*. **39**, D670 (2011).

81. M. D. Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, A. Sarai. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res*. **34**, D204 (2006).

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

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

84. K. P. Murphy, E. Freire. Structural energetics of protein stability and folding cooperativity. *Pure and applied chemistry*. **65**, 1939 (1993).

85. K. A. Dill, K. Ghosh, J. D. Schmit. Physical limits of cells and proteomes. *Proc Natl Acad Sci U S A*. **108**, 17876 (2011).

86. V. S. Cooper, A. F. Bennett, R. E. Lenski. Evolution of thermal dependence of growth rate of Escherichia coli populations during 20,000 generations in a constant environment. *Evolution*. **55**, 889 (2001).

87. E. Van Derlinden, J. F. Van Impe. Modeling growth rates as a function of temperature: model performance evaluation with focus on the suboptimal temperature range. *Int J Food Microbiol*. **158**, 73 (2012).

88. S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, J. Collado-Vides. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*. **39**, D98 (2011).

89. M. M. Riehle, A. F. Bennett, R. E. Lenski, A. D. Long. Evolutionary changes in heat-inducible gene expression in lines of Escherichia coli adapted to high temperature. *Physiol Genomics*. **14**, 47 (2003).

90. A. Diaz-Acosta, M. L. Sandoval, L. Delgado-Olivares, J. Membrillo-Hernandez. Effect of anaerobic and stationary phase growth conditions on the heat shock and oxidative stress responses in Escherichia coli K-12. *Arch Microbiol*. **185**, 429 (2006).

91. M. Murata, H. Fujimoto, K. Nishimura, K. Charoensuk, H. Nagamitsu, S. Raina, T. Kosaka, T. Oshima, N. Ogasawara, M. Yamada. Molecular strategy for survival at a critical high temperature in Eschierichia coli. *PLoS One*. **6**, e20063 (2011).

92. T. S. Gunasekera, L. N. Csonka, O. Paliy. Genome-wide transcriptional responses of Escherichia coli K-12 to continuous osmotic and heat stresses. *J Bacteriol*. **190**, 3712 (2008).

93. S. Jackowski, C. O. Rock. Regulation of coenzyme A biosynthesis. *J Bacteriol*. **148**, 926 (1981).

94. G. Nonaka, M. Blankschien, C. Herman, C. A. Gross, V. A. Rhodius. Regulon and promoter analysis of the E. coli heat-shock factor, sigma32, reveals a multifaceted cellular response to heat stress. *Genes Dev*. **20**, 1776 (2006).

95. P. Turner, G. Mamo, E. N. Karlsson. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microb Cell Fact*. **6**, 9 (2007).

96. C. Stanton, C. Desmond, M. Coakley, J. K. Collins, G. Fitzgerald, R. P. Ross. Challenges facing development of probiotic-containing functional foods. *Handbook of fermented functional foods*. 27 (2003).

97. B. N. Duplantis, M. Osusky, C. L. Schmerk, D. R. Ross, C. M. Bosio, F. E. Nano. Essential genes from Arctic bacteria used to construct stable, temperature-sensitive bacterial vaccines. *Proc Natl Acad Sci U S A*. **107**, 13456 (2010).

98. M. Kanehisa, S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28**, 27 (2000).

99. E. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant. PubChem: integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*. **4**, 217 (2008).

100. L. Jaroszewski, Z. Li, X. H. Cai, C. Weber, A. Godzik. FFAS server: novel features and applications. *Nucleic Acids Res*. **39**, W38 (2011).

101. A. A. Canutescu, A. A. Shelenkov, R. L. Dunbrack, Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*. **12**, 2001 (2003).

102. D. Petrey, Z. Xiang, C. L. Tang, L. Xie, M. Gimpelev, T. Mitros, C. S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I. Y. Koh, E. Alexov, B. Honig. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*. **53 Suppl 6**, 430 (2003).

103. A. Sali, T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. **234**, 779 (1993).

104. L. Jaroszewski, K. Pawlowski, A. Godzik. Multiple model approach: exploring the limits of comparative modeling. *Journal of Molecular Modeling*. **4**, 294 (1998).

105. J. Bandekar. A simple model for protein thermal denaturation. *Int J Pept Protein Res*. **11**, 191 (1978).

106. R. M. Daniel, M. E. Peterson, M. J. Danson, N. C. Price, S. M. Kelly, C. R. Monk, C. S. Weinberg, M. L. Oudshoorn, C. K. Lee. The molecular basis of the effect of temperature on enzyme activity. *Biochem J*. **425**, 353 (2010).

107. M. P. Byrne, W. E. Stites. Thermal denaturations of staphylococcal nuclease wild-type and mutants monitored by fluorescence and circular dichroism are similar: lack of evidence for other than a two state thermal denaturation. *Biophys Chem*. **125**, 490 (2007).

108. J. Ramprakash, V. Doseeva, A. Galkin, W. Krajewski, L. Muthukumar, S. Pullalarevu, E. Demirkan, O. Herzberg, J. Moult, F. P. Schwarz. Comparison of the chemical and thermal denaturation of proteins by a two-state transition model. *Anal Biochem*. **374**, 221 (2008).

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

110. R. Mahadevan, C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*. **5**, 264 (2003).

111. J. D. Orth, I. Thiele, B. O. Palsson. What is flux balance analysis? *Nat Biotechnol*. **28**, 245 (2010).

112. J. Schellenberger, R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, B. O. Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc*. **6**, 1290 (2011).

113. J. L. Reed, B. O. Palsson. Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res*. **14**, 1797 (2004).

114. C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, B. O. Palsson. Genome-scale metabolic model of Helicobacter pylori 26695. *J Bacteriol*. **184**, 4582 (2002).

115. A. Simon, E. Biot. ANAIS: analysis of NimbleGen arrays interface. *Bioinformatics*. **26**, 2468 (2010).

116. J. Baranyi, T. A. Roberts. A dynamic approach to predicting bacterial growth in food. *Int J Food Microbiol*. **23**, 277 (1994).

117. J. Ren, L. Xie, W. W. Li, P. E. Bourne. SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. *Nucleic Acids Res*. **38**, W441 (2010).

118. S. L. Kinnings, L. Xie, K. H. Fung, R. M. Jackson, P. E. Bourne. The Mycobacterium tuberculosis drugome and its polypharmacological implications. *PLoS Comput Biol*. **6**, e1000976 (2010).

119. S. J. Ho Sui, R. Lo, A. R. Fernandes, M. D. Caulfield, J. A. Lerman, L. Xie, P. E. Bourne, D. L. Baillie, F. S. Brinkman. Raloxifene attenuates Pseudomonas aeruginosa pyocyanin production and virulence. *Int J Antimicrob Agents*. **40**, 246 (2012).

120. E. Krissinel, K. Henrick. Inference of macromolecular assemblies from crystalline state. *J Mol Biol*. **372**, 774 (2007).

121. D. Perumal, A. Samal, K. R. Sakharkar, M. K. Sakharkar. Targeting multiple targets in Pseudomonas aeruginosa PAO1 using flux balance analysis of a reconstructed genome-scale metabolic network. *J Drug Target*. **19**, 1 (2011).

122. C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, D. S. Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. **39**, D1035 (2011).

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

124. Y. Cao, T. Jiang, T. Girke. Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. *Bioinformatics*. **26**, 953 (2010).

125. A. Grosdidier, V. Zoete, O. Michielin. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res*. **39**, W270 (2011).