HHS Public Access

Author manuscript

Curr Opin Biotechnol. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Curr Opin Biotechnol. 2015 August ; 34: 125–134. doi:10.1016/j.copbio.2014.12.017.

Computing the functional proteome: recent progress and future prospects for genome-scale models

Edward J. O'Brien[1,2] and **Bernhard O. Palsson**[1,2,3,4]

[1]Bioinformatics and Systems Biology Program, University of California, San Diego

[2]Department of Bioengineering, University of California, San Diego

[3]Department of Pediatrics, University of California, San Diego

[4]Novo Nordisk Center for Biosustainability, The Danish Technical University

Abstract

Constraint-based models enable the computation of feasible, optimal, and realized biological phenotypes from reaction network reconstructions and constraints on their operation. To date, stoichiometric reconstructions have largely focused on metabolism, resulting in genome-scale metabolic models (M-Models). Recent expansions in network content to encompass proteome synthesis have resulted in models of metabolism and protein expression (ME-Models). ME-Models advance the predictions possible with constraint-based models from network flux states to the spatially resolved molecular composition of a cell. Specifically, ME-Models enable the prediction of transcriptome and proteome allocation and limitations, and basal expression states and regulatory needs. Continued expansion in reconstruction content and constraints will result in an increasingly refined representation of cellular composition and behavior.

Keywords

systems biology; constraint-based model; genome-scale model; metabolism; gene expression; proteome

Introduction

Building computational whole cell models has been a long-standing goal of theoretical biology. In the 1980s, serious attempts to build large-scale models of a whole bacterium were undertaken [1]. A few years later, an attempt to build whole cell models for the human red cell represented a culmination of decades of work [2–6]. Perhaps the most comprehensive whole organism model appeared in the mid 1990s for the lambda-bacteriophage [7,8]. Time scale decomposition of these early models showed that their effective dynamic order was low [9] and that their dynamic structure was relatively

condition invariant [10], motivating the development of constraint-based models that minimized the need for kinetic information [11].

After the first full genome sequences appeared, constraint-based models could be scaled up to the genome-scale [12]. As the first genome-scale models (GEMs) proved their ability to predict biological functions [12,13], a vision was laid out in 2003 [14] that all cellular functions could be reconstructed in biochemical terms and seamlessly integrated. A decade later, some of this vision has been realized [15–19]. With these achievements, we can now assess what might be ahead with genome-scale models over the coming decade. We lay out some of our thoughts in this commentary.

The expanding scope of reconstructions: synthesis and function of the proteome

To date, stoichiometric reconstructions have largely focused on metabolism, resulting in genome-scale metabolic models, M-Models. The processes of enzyme synthesis including transcription, translation, protein folding, complex formation, and prosthetic group integration were formalized in a gene expression reconstruction [20]. Protein translocation and localization pathways [18,21] and DNA replication, repair, and cell division have also been reconstructed [19]. These networks have been merged with metabolic reconstructions to create integrated reaction networks [15–19] that formalize the primary chemical transformations that occur in cell (Figure 1A). Models integrating metabolism with protein expression are called ME-models.

To enable prediction of biological phenotypes, stoichiometric reconstructions are combined with constraints on their operation. Stoichiometric networks are subject to (dynamic or steady-state) mass balance constraints on the production and consumption of molecules. For ME-Models, enzyme catalytic constraints are also necessary. In contrast to the typical use of kinetic equations to simulate system dynamics, the catalytic constraints in ME-Models are approximate stoichiometric relationships between enzyme abundance and catalyzed flux (Figure 1B). Adding these catalytic 'coupling' constraints [20] enables the computation of feasible and optimal proteome and transcriptome states.

Most models encompassing gene expression have used measured expression states as a prerequisite for simulations. Often, gene expression measured under a particular condition is used to predict other molecular and physiological phenotypes [19,22]. Alternatively, some approaches utilize gene expression data under environmental or genetic perturbations to build regulatory models [23]. These two approaches can be combined to predict molecular and physiological phenotypes subject to a transcriptional regulatory model [24,25]. These are undoubtedly invaluable types of models and predictions; similar methods will likely be applied to ME-Models (Figure 1C).

ME-Models can predict gene expression with no previous input expression measurements: they can compute protein abundances that are required to (optimally) achieve integrated physiological functions (Figure 1C). Enzymes have an optimal expression level subject to their (biosynthetic) cost and (physiological) benefit [26]. The ME-Model solves this cost-

benefit optimization to compute genome-scale proteome states. Thus, compared to other methods for prediction and analysis of gene expression, predictions of gene expression in the ME-Model are based on fundamental constraints and optimality principles (as are the predictions of flux states in M-Models). ME-models can therefore be used to predict optimal expression and regulatory states.

Prediction of the molecular composition of a cell

An important distinction between M- and ME-Models is the prediction of cellular biomass composition. Instead of having protein and RNA biomass composition as an input (in the form of a biomass objective function [27]), biomass composition is an output that is predicted by ME-models. The expressed molecular machinery, such as the proteome, must support the integrated physiological functions of the cell. Rather than processes being coupled stoichiometrically through the biomass function, demands for vitamins and cofactors, chaperones, amino acids, nucleotides, tRNAs, etc. are derived directly from the computed proteome state.

Furthermore, a recent expansion of the ME-Model to include protein translocation enables predictions of a cell's coarse-grained spatial organization [18]. Protein complexes are localized in cellular compartments required for enzyme function. With this expansion in scope, aspects of compartmentalized proteome abundance and molecular crowding can be assessed [18].

Thus, ME-Models advance the predictions possible with constraint-based models from network flux states to the spatially resolved molecular composition of a cell (Figure 2A).

Phenotypic effects of proteome allocation constraints

In addition to satisfying flux balance constraints, ME-Models are subject to proteome allocation constraints. While M-Models account for the 'operating expenses' (i.e., metabolic requirements) to carry flux through pathways, ME-Models also account for the 'capital expenses' (i.e., enzyme machinery) needed to catalyze all network reactions. Therefore, in addition to cellular functions being limited by nutrients, they can also be limited by properties of the proteome (i.e., due to limited protein synthesis capacity and enzyme catalytic rates). Proteome allocation constraints govern integrated cell functions and, combined with growth-optimality assumptions, can explain several aspects of cell behavior not encompassed by previous models (Figure 2B).

First, the change in ribosomal protein abundance can be explained by growth-optimization subject to proteome allocation constraints [15,17,20,28,29]. At faster growth rates, more ribosomes are required to sustain the faster dilution of protein to daughter cells. Previous models have taken this growth rate dependent relationship as an observed (and subsequently assumed and fixed) phenomenological relationship [30], rather than a prediction.

Second, specific pathway shifts in central carbon metabolism from carbon-limited to carbon-excess environments (i.e., batch culture or non-carbon limitations), can be explained as a consequence of proteome allocation constraints. Specific pathway shifts from carbon-limited

to proteome-limited growth are consistent with pathway shifts observed between chemostat and batch cultures [17]. Whereas previous approaches required the invocation of multiple competing objectives [31], now a single objective of maximal growth rate subject to proteome allocation constraints can explain the same phenomenon.

Third, the constraints limiting absolute growth rates and substrate uptake rates have remained elusive. Proteome-limitations in the ME-Model result in a maximal growth rate and optimal substrate uptake rate that is consistent with experimental data when nutrients are available in excess [17]. The limitations placed on substrate uptake by the proteome significantly expand the scope of environments that can be simulated with constraint-based models to include nutrient-excess and complex media conditions.

Fourth, spatial limitations on the membrane proteome further refine predictions of pathway shifts and substrate uptake rates in nutrient-excess environments [18]. Limitations on protein synthesis and protein space result in similar phenotypic responses, but have some differences in enzyme utilization; membrane proteomics data and experimental evolution can help to illuminate which constraints are dominant.

The phenotypic effects of proteome allocation constraints are just beginning to be uncovered and will likely change our conception of optimal behavior and pathway use [32].

Gene expression states and molecular phenotypes can now be computed

The prediction of proteome composition is an ambitious endeavor. To date, a few predictions of absolute gene expression have been validated. The ME-Model accurately predicts ribosomal [15,17] and translocase [18] protein abundances, which have well-known catalytic rates. In general, however, catalytic rates of specific enzymes *in vivo* are not known. Nonetheless, the relatively accurate prediction of overall proteome abundance of different cellular compartments and functional subsystems is possible [18]. Furthermore, genome-scale predicted and measured mRNA abundance correlate significantly [15]. These early predictions provide support for the genome-scale prediction of absolute gene expression from evolutionary optimality principles (Figure 2B).

As with false predictions from M-Models [33], discrepancies between predicted and observed expression levels have led to discovery (Figure 3A). First, the quantitative difference between predicted and measured gross RNA and protein biomass composition has led to the realization that translation rate is a hyperbolic function of growth rate [17], which has been independently validated [34]. Second, comparing predicted and measured abundance of functional subsystems identified the processes of protein folding and metal ion and prosthetic group integration as under-predicted [18]. These under-predictions are consistent with known knowledge gaps of chaperone targets [35] and metal ion usage by proteins [36] and prioritizes these processes for further reconstruction.

Given the discordance between measured RNA and protein abundances [37], the moderate correlation between genome-scale predicted and measured gene product abundance is unsurprising. The factors contributing to the discrepancy between RNA and protein abundances are beginning to be uncovered [38,39], aided by diverse data types on the

various steps of gene expression, including promoter activity [40,41], RNA abundance, RNA degradation rates, ribosome occupancy [42], and protein abundance [43]. These data types and gene-specific rates on the steps of gene expression can readily be integrated into ME-Models. Parameterizing the steps of gene expression with these data types, biophysical models [44,45] or synthetic 'parts characterization' [46–49] will help understand the gap between RNA and protein abundance as well as *in silico* and *in vivo* gene expression levels.

Precise prediction of protein abundance is also limited by knowledge of enzyme catalytic rates. However, even though data on individual enzyme rates is noisy and sparse [50,51], statistics on the distributions of catalytic rates are robust [52], enabling confident distributions of expression levels to be computed. Furthermore, model-driven approaches can be used to infer catalytic rates that are consistent with *in vivo* data [53–55]. These efforts will iteratively result in more precise predictions of protein expression.

Bottom-up prediction of gene expression will truly test our understanding of the biological demand and activity of enzymes. We believe that quantitative proteome levels will not fully be understood until we are able to predict them from the bottom-up with GEMs. Given the early successful uses of ME-models it seems clear that there is much more discovery that lies ahead. Just as the community has become accustomed to flux balances, and thus the uses of metabolic networks, ME-models are likely to help us understand how the composition of the proteome is optimally balanced.

Defining and understanding regulatory needs

Prediction of regulatory needs during shifts in homeostatic states is another important challenge for ME-Models. Differential expression data is more abundantly available than absolute expression data, and will aide in ME-Model validation and model-driven discovery. We anticipate that this comparison and, more generally, a physiological needs perspective on gene expression will help reveal the principles underlying transcriptional regulation.

Recent examples of bottom-up prediction of differential expression include the use of an M-Model to predict transcriptional changes after redox shifts [56] and the use of a ME-Model to predict differential expression after a shift in carbon sources [15]. Furthermore, the principle of simplest pathway structure can predict gene co-expression and transcriptional regulatory relationships [57]. These examples provide evidence that transcriptional regulation is somewhat predictable based on optimality principles.

There are various reasons as to why transcriptional regulation may seem non-optimal [58]. First, there could be errors in the reaction network reconstruction, which can be rectified by systematic comparison of computational predictions and experimental data [33]. Second, discrepancies could be due to constraints or optimality principles that are not yet modeled or understood (such as proteome constraints added in moving from M- to ME-Models). Third, the environmental history of the organism may have coupled seemingly unrelated biological processes [59], or be optimized for fluctuating rather than static environments [60–63]. Finally, it is likely that transcriptional regulation is 'moderately efficient' rather than perfectly optimal. Enumerating and classifying these discrepancies can drive biological discovery as has occurred through classifying the false predictions of gene essentiality

[33,64]. Identified discrepancies can provide insight into organismal physiology and prioritize the development of explicit transcriptional regulatory models.

Parallel to the prediction of transcriptional regulation with constraint-based models, a physiological perspective has revealed striking simplicity and optimality in transcriptional regulation [56,65]. In several studies, the mass fractions of large protein subsystems and the activity of transcription factors have been shown to change linearly with growth rate or specific metabolic fluxes [66–69]. Furthermore, linear models covering several genes have been shown to capture the variation in their expression with relative accuracy [41,70]. The simplicity of these regulatory relationships (despite the complex topology and biophysical relationships [71] underlying regulatory networks) provides promise for accurate genome-scale regulatory models. However, the cross-talk and competition between transcription factors is still not generally understood; in the meantime, top-down approaches [23] may be necessary to capture the essence of these more complex relationships.

Importantly, the explicit representation of transcription in the ME-Model allows for the molecular details of transcription factor targets to be combined with the physiological principles underlying transcription factor activity. This will enable new approaches to model transcriptional regulation that move beyond binary representations of transcription factor activity [25]. Regulatory model development can be prioritized by the physiological importance of regulatory shifts and failure modes identified through comparison of predicted and measured differential expression.

Though we have focused on transcriptional regulation here, optimality and physiological principles will likely apply to translational (e.g., by sRNAs) and post-translational regulation (e.g., by post-translational modifications and allosteric interactions) as well. These regulatory networks have received less attention, partially due to the difficulty in identifying the underlying interactions networks (compared to transcriptional regulatory networks [72]). However, new computational [73,74] and experimental [45,75] methods are emerging, and optimality principles are being uncovered [76,77] to elucidate these regulatory networks. Like transcriptional regulation, we anticipate that the explicit representation of enzyme abundance and activity in ME-Models will aid in the genome-scale modeling of post-transcriptional regulation.

Seeking a comprehensive biophysical representation of cellular composition

The conceptual change in GEMs to enable the prediction of proteome abundance, localization, and limitations affords numerous opportunities for model application and expansion. The *E. coli* ME-Model currently encompasses ~80% of the proteome and transcriptome by mass in environments of exponential growth [17]; this equates to ~60% of the cell's entire mass. While the requirements for biosynthesis of a whole cell are encompassed by the model, not all molecular abundances are predicted; gaps include: 1) the non-ME proteome, 2) the cell envelope, 3) metabolite concentrations, 4) DNA replication and gene copy number, and 5) glycogen. We briefly cover factors that may enable prediction of these molecular abundances deemed most important (Figure 3B).

Metabolite concentrations are beginning to be predicted with genome-scale models using thermodynamic considerations. By extending a method to ensure metabolic fluxes are thermodynamically feasible [78], thermodynamic constraints were used to predict steady-state metabolite concentrations that are consistent with a given flux state [79]. Later, an objective to minimize metabolite concentrations over the thermodynamically feasible space was shown to increase prediction accuracy [80]. Additional constraints on osmolarity, metabolite toxicity, and correlations observed between metabolite concentrations and their enzyme affinities [81] and chemical properties [82] may improve predictions further. Then, the effects of these concentrations on enzyme and transcription factor activity may be accounted for in future genome-scale models.

A first requirement for the prediction of cell envelope composition is the prediction of cell size and shape, which determines the surface area of the cell that must be covered. The consistent growth rate dependence of cell size [83] suggests that simple principles may underlie the determination of cell size (for example, the balance between cytosolic and membrane proteome abundance [84]). However, the constraints underlying the exact composition of membrane lipid, glycans, LPS, and murein (together accounting for 15% of cell dry weight) are not well understood. Perhaps data on cell envelope composition will aide in understanding when and how it varies across environments and strains (an important characteristic of particular *E. coli* serotypes).

Expanding the proteome coverage to the remaining 20% of the proteome not encompassed by the ME-Model will require expansion of reconstruction content. Proteome abundance can be used to prioritize model expansion [85]. Many of these non-ME proteins can be broadly categorized as non-growth and stress response genes (e.g., biofilm and flagella formation and pH, osmolarity, and temperature responses). Therefore, modeling of stress responses is important to increase the coverage of the proteome and environments that can be simulated. Importantly, the ME-Model already accounts for the biosynthetic costs of synthesizing stress response proteins; however, the constraints imposed by environmental stresses will be needed to understand the protein's physiological benefit.

Protein structures will aide in formalizing the constraints on the proteome. Genome-scale models integrated with protein structures (GEM-PRO) enable simulation of the effects of temperature: structure-based predictions of protein thermostability and the subsequent limitations on metabolic fluxes result in accurate predictions of growth and nutrient supplementations at high temperatures [86]. As other cellular stresses (e.g., pH) also affect protein catalytic capacity, protein structures may enable the simulation of other physiochemical stresses as well. Protein structures combined with ME-Models will also approach a more detailed biophysical representation of a cell. Spatial resolution can be refined further with protein-protein interaction data [87] (or prediction of protein-protein interactions with protein structures themselves [88]). Spatial considerations may be important for understanding co-localization of sequential catalytic steps [89,90] and the effects of molecular crowding [91,92] in the cytosol and membranes.

Conclusion

Building whole cell computational models has been a long-standing goal. Genome-scale metabolic models, M-Models, have become widely used due to the numerous actionable predictions they can make [93], and the ease of draft model construction from readily available genome sequences and annotations [94]. Here we reviewed recent advancements expanding the scope of whole cell computational models to encompass the synthesis and localization of the proteome. The constraint-based philosophy underlying ME-Models parallels that of M-Models. However, the expanded scope of components and constraints enables the prediction of enzyme abundance and activity. Already, ME-Models have revealed how constraints on proteome allocation explain aspects of cell behavior that have remained elusive or require invocation of phenomenological relationships. Furthermore, several cases demonstrate that ME-Models enable accurate prediction of protein abundance and differential expression. As the basic capabilities of M-Models to predict flux states have led to numerous applications, future work will capitalize on the new capabilities of GEMs to compute proteome allocation and limitations. Optimality-based predictions will no doubt be imperfect, but they form a strong conceptual base to drive biological discovery, bioengineering, and further model development.

Acknowledgements

We thank Ali Ebrahim, Daniel Zielinski, Zak King, and Joshua Lerman for insightful discussions and critical feedback on the manuscript. EJO was supported by National Institute of Health R01 GM057089.

References

1. Domach MM, Leung SK, Cahn RE, Cocks GG, Shuler ML. Computer model for glucose-limited growth of a single cell of escherichia coli b/r-a. Biotechnol Bioeng. 1984; 26(3):203–216. [PubMed: 18551728]
2. Joshi A, Palsson BO. Metabolic dynamics in the human red cell. Part i--a comprehensive kinetic model. J Theor Biol. 1989; 141(4):515–528. [PubMed: 2630803]
3. Joshi A, Palsson BO. Metabolic dynamics in the human red cell. Part ii--interactions with the environment. J Theor Biol. 1989; 141(4):529–545. [PubMed: 2630804]
4. Joshi A, Palsson BO. Metabolic dynamics in the human red cell. Part iii--metabolic reaction rates. J Theor Biol. 1990; 142(1):41–68. [PubMed: 2141093]
5. Joshi A, Palsson BO. Metabolic dynamics in the human red cell. Part iv--data prediction and some model computations. J Theor Biol. 1990; 142(1):69–85. [PubMed: 2141094]
6. Heinrich R, Rapoport SM, Rapoport TA. Metabolic regulation and mathematical models. Prog Biophys Mol Biol. 1977; 32(1):1–82. [PubMed: 343173]
7. McAdams HH, Shapiro L. Circuit simulation of genetic networks. Science. 1995; 269(5224):650–656. [PubMed: 7624793]
8. McAdams HH, Arkin A. Simulation of prokaryotic genetic circuits. Annu Rev Biophys Biomol Struct. 1998; 27(199–224)
9. Joshi A, Palsson BO. Escherichia coil growth dynamics: A three-pool biochemically based description. Biotechnol Bioeng. 1988; 31(2):102–116. [PubMed: 18581570]
10. Palsson BO, Joshi A. On the dynamic order of structured escherichia coli growth models. Biotechnol Bioeng. 1987; 29(6):789–792. [PubMed: 18576521]
11. Varma A, Palsson BO. Metabolic flux balancing - basic concepts, scientific and practical use. Bio-Technol. 1994; 12(10):994–998.

12. Edwards JS, Palsson BO. The escherichia coli mg1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97(10):5528–5533. [PubMed: 10805808]
13. Edwards JS, Ibarra RU, Palsson BO. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. Nature biotechnology. 2001; 19(2):125–130.
14. Reed JL, Palsson BO. Thirteen years of building constraint-based in silico models of escherichia coli. J Bacteriol. 2003; 185(9):2692–2699. [PubMed: 12700248]
15. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, Palsson BO. In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun. 2012; 3(929) ** This study shows that ME-Models can be used to predict differential gene expression and reduce the metabolic solution space compared to M-Models.
16. Thiele I, Fleming RM, Que R, Bordbar A, Diep D, Palsson BO. Multiscale modeling of metabolism and macromolecular synthesis in e. Coli and its application to the evolution of codon usage. PloS one. 2012; 7(9):e45635. [PubMed: 23029152]
17. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular systems biology. 2013; 9(693) ** An ME-Model for *E. coli* accounts for ~80% of the proteome mass and is used to study the effects of nutrient versus proteome limitations.
18. Liu JK, EJ OB, Lerman JA, Zengler K, Palsson BO, Feist AM. Reconstruction and modeling protein translocation and compartmentalization in escherichia coli at the genome-scale. BMC Syst Biol. 2014; 8(1):110. [PubMed: 25227965] ** An ME-Model for *E. coli* that includes protein translocation and compartmentalization accurately predicts translocase and protein subsystem abundance.
19. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr, Assad-Garcia N, Glass JI, Covert MW. A whole-cell computational model predicts phenotype from genotype. Cell. 2012; 150(2):389–401. [PubMed: 22817898]
20. Thiele I, Fleming RM, Bordbar A, Schellenberger J, Palsson BO. Functional characterization of alternate optimal solutions of escherichia coli's transcriptional and translational machinery. Biophys J. 2010; 98(10):2072–2081. [PubMed: 20483314]
21. Feizi A, Osterlund T, Petranovic D, Bordel S, Nielsen J. Genome-scale modeling of the protein secretory machinery in yeast. PloS one. 2013; 8(5):e63284. [PubMed: 23667601]
22. Machado D, Herrgard M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS computational biology. 2014; 10(4):e1003580. [PubMed: 24762745]
23. Brooks AN, Reiss DJ, Allard A, Wu WJ, Salvanha DM, Plaisier CL, Chandrasekaran S, Pan M, Kaur A, Baliga NS. A system-level model for the microbial regulatory genome. Molecular systems biology. 2014; 10(7):740. [PubMed: 25028489]
24. Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of escherichia coli. Molecular systems biology. 2014; 10(7):735. [PubMed: 24987114]
25. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(41):17845–17850. [PubMed: 20876091]
26. Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005; 436(7050):588–592. [PubMed: 16049495]
27. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010; 13(3):344–349. [PubMed: 20430689]
28. Thiele I, Jamshidi N, Fleming RM, Palsson BO. Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization. PLoS computational biology. 2009; 5(3):e1000312. [PubMed: 19282977]

29. Scott M, Klumpp S, Mateescu EM, Hwa T. Emergence of robust growth laws from optimal regulation of ribosome synthesis. Molecular systems biology. 2014; 10(747) *Growth optimality principles are used to derive the growth-rate dependent change in ribosome abundance.

30. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: Origins and consequences. Science. 2010; 330(6007):1099–1102. [PubMed: 21097934]

31. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. Multidimensional optimality of microbial metabolism. Science. 2012; 336(6081):601–604. [PubMed: 22556256]

32. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R. Glycolytic strategy as a tradeoff between energy yield and protein cost. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(24):10039–10044. [PubMed: 23630264] **Thermodynamic analysis reveals a tradeoff between energy yield and protein cost across alternative pathways.

33. Orth JD, Palsson BO. Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng. 2010; 107(3):403–412. [PubMed: 20589842]

34. Klumpp S, Scott M, Pedersen S, Hwa T. Molecular crowding limits translation and cell growth. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(42): 16754–16759. [PubMed: 24082144]

35. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, Nichols RJ, Typas A, Gross CA, Kramer G, Weissman JS, et al. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell. 2011; 147(6):1295–1308. [PubMed: 22153074]

36. Cvetkovic A, Menon AL, Thorgersen MP, Scott JW, Poole FL 2nd, Jenney FE Jr, Lancaster WA, Praissman JL, Shanmukh S, Vaccaro BJ, Trauger SA, et al. Microbial metalloproteomes are largely uncharacterized. Nature. 2010; 466(7307):779–782. [PubMed: 20639861]

37. Maier T, Guell M, Serrano L. Correlation of mrna and protein in complex biological samples. FEBS Lett. 2009; 583(24):3966–3973. [PubMed: 19850042]

38. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature reviews Genetics. 2012; 13(4):227–232.

39. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011; 473(7347):337–342. [PubMed: 21593866]

40. Zaslaver A, Kaplan S, Bren A, Jinich A, Mayo A, Dekel E, Alon U, Itzkovitz S. Invariant distribution of promoter activities in escherichia coli. PLoS computational biology. 2009; 5(10):e1000545. [PubMed: 19851443]

41. Keren L, Zackay O, Lotan-Pompan M, Barenholz U, Dekel E, Sasson V, Aidelberg G, Bren A, Zeevi D, Weinberger A, Alon U, et al. Promoters maintain their relative activity levels under different growth conditions. Molecular systems biology. 2013; 9(701) *The relative activity of promoters is generally conserved across environments and can be accurately described by a handful environment-specific scaling factors.

42. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014; 157(3):624–635. [PubMed: 24766808] *Translation rate data reveals that proteins are translated stoichiometrically and much of the proteome is devoted to cell growth.

43. Valgepea K, Adamberg K, Seiman A, Vilu R. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. Mol Biosyst. 2013; 9(9):2344–2358. [PubMed: 23824091]

44. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-limiting steps in yeast protein translation. Cell. 2013; 153(7):1589–1601. [PubMed: 23791185]

45. Ingolia NT. Ribosome profiling: New views of translation, from single codons to genome scale. Nature reviews Genetics. 2014; 15(3):205–213.

46. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM. Composability of regulatory sequences controlling transcription and translation in escherichia coli. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(34): 14024–14029. [PubMed: 23924614]

47. Mutalik VK, Guimaraes JC, Cambray G, Mai QA, Christoffersen MJ, Martin L, Yu A, Lam C, Rodriguez C, Bennett G, Keasling JD, et al. Quantitative estimation of activity and quality for collections of functional genetic elements. Nat Methods. 2013; 10(4):347–353. [PubMed: 23474467]

48. Cambray G, Guimaraes JC, Mutalik VK, Lam C, Mai QA, Thimmaiah T, Carothers JM, Arkin AP, Endy D. Measurement and modeling of intrinsic transcription terminators. Nucleic Acids Res. 2013; 41(9):5139–5148. [PubMed: 23511967]

49. Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat Methods. 2013; 10(7):659–664. [PubMed: 23727987]

50. van Eunen K, Kiewiet JA, Westerhoff HV, Bakker BM. Testing biochemistry revisited: How in vivo metabolism can be understood from in vitro enzyme kinetics. PLoS computational biology. 2012; 8(4):e1002483. [PubMed: 22570597]

51. Garcia-Contreras R, Vos P, Westerhoff HV, Boogerd FC. Why in vivo may not equal in vitro - new effectors revealed by measurement of enzymatic activities under the same in vivo-like assay conditions. The FEBS journal. 2012; 279(22):4145–4159. [PubMed: 22978366]

52. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011; 50(21):4402–4410. [PubMed: 21506553]

53. Sanghvi JC, Regot S, Carrasco S, Karr JR, Gutschow MV, Bolival B Jr, Covert MW. Accelerated discovery via a whole-cell model. Nat Methods. 2013; 10(12):1192–1195. [PubMed: 24185838]

54. Cotten C, Reed JL. Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. BMC bioinformatics. 2013; 14(32)

55. Khodayari A, Zomorrodi AR, Liao JC, Maranas CD. A kinetic model of escherichia coli core metabolism satisfying multiple sets of mutant flux data. Metabolic engineering. 2014

56. Federowicz S, Kim D, Ebrahim A, Lerman J, Nagarajan H, Cho BK, Zengler K, Palsson B. Determining the control circuitry of redox metabolism at the genome-scale. PLoS Genet. 2014; 10(4):e1004264. [PubMed: 24699140]

57. Bordbar A, Nagarajan H, Lewis NE, Latif H, Ebrahim A, Federowicz S, Schellenberger J, Palsson BO. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. Molecular systems biology. 2014; 10(7):737. [PubMed: 24987116] *The principle of shortest metabolic pathways better matches biomolecular interaction data and enables prediction of new transcription factor interactions.

58. Price MN, Deutschbauer AM, Skerker JM, Wetmore KM, Ruths T, Mar JS, Kuehl JV, Shao W, Arkin AP. Indirect and suboptimal control of gene expression is widespread in bacteria. Molecular systems biology. 2013; 9(660)

59. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, Kupiec M, Dahan O, Pilpel Y. Adaptive prediction of environmental changes by microorganisms. Nature. 2009; 460(7252):220–224. [PubMed: 19536156]

60. New AM, Cerulus B, Govers SK, Perez-Samper G, Zhu B, Boogmans S, Xavier JB, Verstrepen KJ. Different levels of catabolite repression optimize growth in stable and variable environments. PLoS Biol. 2014; 12(1):e1001764. [PubMed: 24453942]

61. de Hijas-Liste GM, Klipp E, Balsa-Canto E, Banga JR. Global dynamic optimization approach to predict activation in metabolic pathways. BMC Syst Biol. 2014; 8(1)

62. Pavlov MY, Ehrenberg M. Optimal control of gene expression for fast proteome adaptation to environmental change. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(51):20527–20532. [PubMed: 24297927]

63. Bartl M, Kotzing M, Schuster S, Li P, Kaleta C. Dynamic optimization identifies optimal programmes for pathway regulation in prokaryotes. Nat Commun. 2013; 4(2243)

64. Monk J, Palsson BO. Genetics. Predicting microbial growth. Science. 2014; 344(6191):1448–1449. [PubMed: 24970063]

65. Chubukov V, Gerosa L, Kochanowski K, Sauer U. Coordination of microbial metabolism. Nat Rev Microbiol. 2014; 12(5):327–340. [PubMed: 24658329]

66. Klumpp S, Hwa T. Growth-rate-dependent partitioning of rna polymerases in bacteria. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(51):20245–20250. [PubMed: 19073937]

67. Klumpp S, Zhang ZG, Hwa T. Growth rate-dependent global effects on gene expression in bacteria. Cell. 2009; 139(7):1366–1375. [PubMed: 20064380]

68. You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang YP, Lenz P, Yan D, Hwa T. Coordination of bacterial proteome with metabolism by cyclic amp signalling. Nature. 2013; 500(7462):301–306. [PubMed: 23925119] **Proteome allocation to large protein subsystems is linearly related to growth rate depending on the limiting nutrient.

69. Kochanowski K, Volkmer B, Gerosa L, Haverkorn van Rijsewijk BR, Schmidt A, Heinemann M. Functioning of a metabolic flux sensor in escherichia coli. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(3):1130–1135. [PubMed: 23277571] *The activity of a transcription factor is linearly related to specific metabolic fluxes.

70. Rothschild D, Dekel E, Hausser J, Bren A, Aidelberg G, Szekely P, Alon U. Linear superposition and prediction of bacterial promoter activity dynamics in complex conditions. PLoS computational biology. 2014; 10(5):e1003602. [PubMed: 24809350]

71. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R. Transcriptional regulation by the numbers: Models. Curr Opin Genet Dev. 2005; 15(2):116–124. [PubMed: 15797194]

72. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Osorio G, et al. Regulondb v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 2013; 41(Database issue):D203–D213. [PubMed: 23203884]

73. Link H, Kochanowski K, Sauer U. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. Nature biotechnology. 2013; 31(4):357–361.

74. Modi SR, Camacho DM, Kohanski MA, Walker GC, Collins JJ. Functional characterization of bacterial srnas using a network biology approach. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(37):15522–15527. [PubMed: 21876160]

75. Li X, Gianoulis TA, Yip KY, Gerstein M, Snyder M. Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. Cell. 2010; 143(4):639–650. [PubMed: 21035178]

76. Goyal S, Yuan J, Chen T, Rabinowitz JD, Wingreen NS. Achieving optimal growth through product feedback inhibition in metabolism. PLoS computational biology. 2010; 6(6)

77. Chubukov V, Zuleta IA, Li H. Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(13):5127–5132. [PubMed: 22416120]

78. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. Biophys J. 2007; 92(5):1792–1805. [PubMed: 17172310]

79. Tan Y, Rivera JG, Contador CA, Asenjo JA, Liao JC. Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. Metabolic engineering. 2011; 13(1):60–75. [PubMed: 21075211]

80. Tepper N, Noor E, Amador-Noguez D, Haraldsdottir HS, Milo R, Rabinowitz J, Liebermeister W, Shlomi T. Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. PloS one. 2013; 8(9):e75370. [PubMed: 24086517]

81. Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD. Absolute metabolite concentrations and implied enzyme active site occupancy in escherichia coli. Nat Chem Biol. 2009; 5(8):593–599. [PubMed: 19561621]

82. Bar-Even A, Noor E, Flamholz A, Buescher JM, Milo R. Hydrophobicity and charge shape cellular metabolite concentrations. PLoS computational biology. 2011; 7(10):e1002166. [PubMed: 21998563]

83. Donachie, WD.; Robinson, AC. Cell division: Parameter values and the process. In: Neidhardt, FCIJ.; Low, KB.; Magasanik, B.; Schaechter, M.; Umbarger, HE., editors. Escherichia coli and

salmonella typhimurium cellular and molecular biology. Washington, D.C: American Society for Microbiology; 1987. p. 1578-1593.

84. Molenaar D, van Berlo R, de Ridder D, Teusink B. Shifts in growth strategies reflect tradeoffs in cellular economics. Molecular systems biology. 2009; 5(323)
85. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(23):8488–8493. [PubMed: 24889604]
86. Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BO. Structural systems biology evaluation of metabolic thermotolerance in escherichia coli. Science. 2013; 340(6137):1220–1223. [PubMed: 23744946] **Protein structures coupled with a metabolic model enable prediction of the growth response and metabolic bottlenecks at high temperature.
87. Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A, Hauser R, et al. The binary protein-protein interaction landscape of escherichia coli. Nature biotechnology. 2014; 32(3):285–290.
88. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012; 490(7421):556–560. [PubMed: 23023127]
89. Huang X, Holden HM, Raushel FM. Channeling of substrates and intermediates in enzyme-catalyzed reactions. Annu Rev Biochem. 2001; 70(149–180)
90. Agapakis CM, Boyle PM, Silver PA. Natural strategies for the spatial optimization of metabolism in synthetic biology. Nat Chem Biol. 2012; 8(6):527–535. [PubMed: 22596204]
91. Ellis RJ. Macromolecular crowding: Obvious but underappreciated. Trends Biochem Sci. 2001; 26(10):597–604. [PubMed: 11590012]
92. Dill KA, Ghosh K, Schmit JD. Physical limits of cells and proteomes. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(44):17876–17882. [PubMed: 22006304]
93. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nature reviews Genetics. 2014; 15(2):107–120.
94. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology. 2010; 28(9):977–982.
95. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011; 147(6):1408–1419. [PubMed: 22153082]
96. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004; 429(6987):92–96. [PubMed: 15129285]
97. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. Molecular systems biology. 2007; 3(119)

- Models of metabolism and gene expression (ME-Models) now exist for model organisms.
- ME-Models account for ~80% of the proteome mass.
- ME-Models enable prediction of proteome allocation and limitations.
- ME-Models enable prediction of basal expression states and regulatory needs.
- Increased scope and resolution will be achieved with further model expansion.

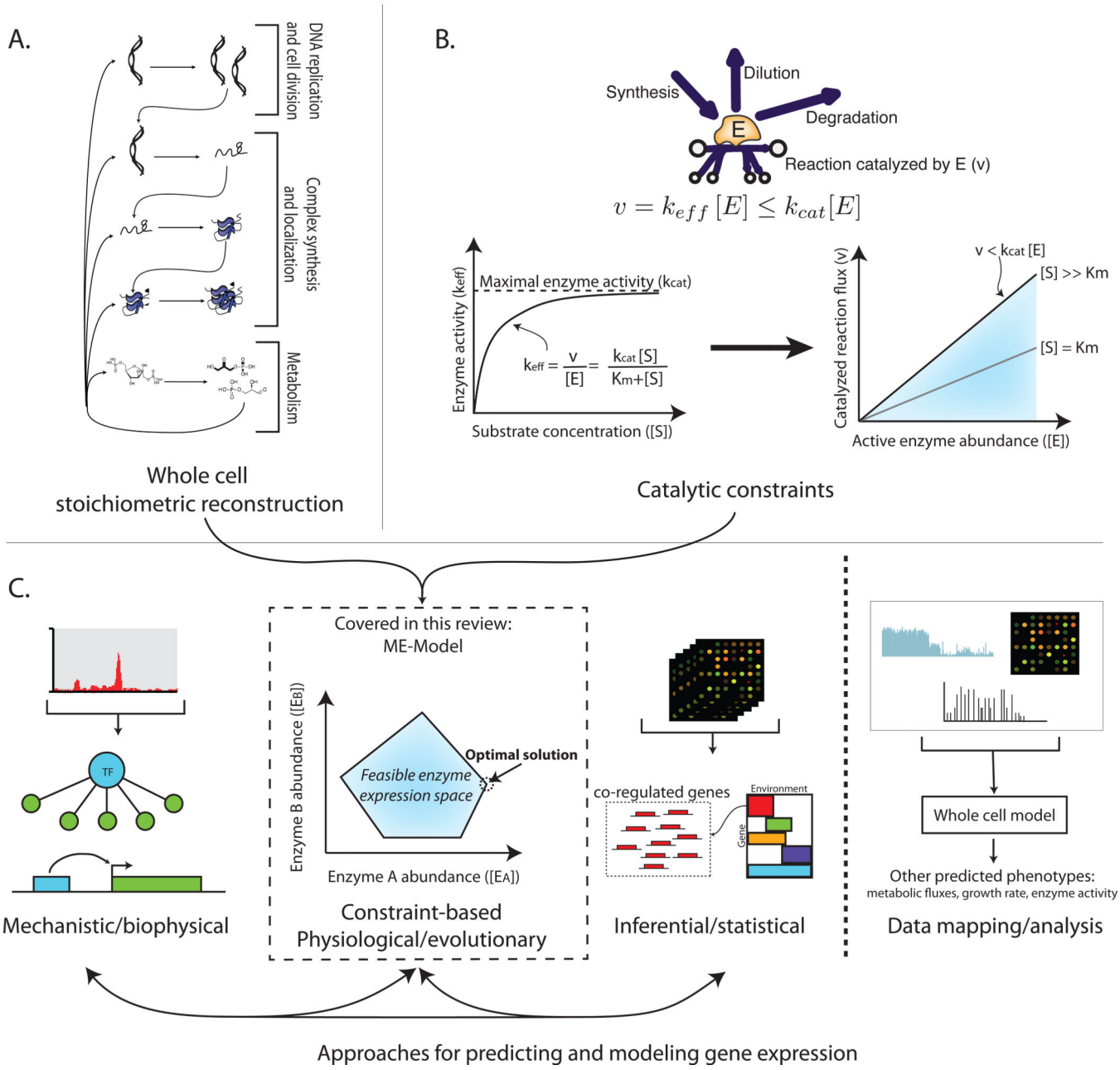

Figure 1. The expanding scope of reconstructions: synthesis and function of the proteome

A) Stoichiometric reconstructions represent the chemical transformations that can occur in a cell and form the base of whole cell models. Recent reconstructions represent all of the major steps in the central dogma of molecular biology in biochemical detail [15–19].

B) Constraints on network operation are utilized to predict functional states. For reconstructions that encompass enzyme synthesis and function, catalytic constraints are necessary [15,20]. Catalytic constraints relate enzyme abundance to its dilution (to daughter cells), degradation, and catalyzed flux with kinetic and/or thermodynamic relationships.

C) Several general approaches exist to predict and model gene expression. Mechanistic/biophysical approaches first start from bottom-up reconstructions of transcription factor

interactions and promoter architectures [56], aided by high-throughput data types [95]; models of regulatory logic can then be reconstructed and imposed [71,96]. Constraint-based models are built from reconstructions of biochemical networks and constraints on their operation and can then be used to compute feasible and optimal physiological states [15–18]. Importantly, the constraint-based approach enables prediction of gene expression states without any previous gene expression measurements. Inferential/statistical approaches are based on large gene expression datasets across environmental and genetic perturbations to identify co-regulated gene sets and their expression under novel perturbations. These general approaches can also be combined into hybrid models [24,25]. Finally, we distinguish approaches that predict gene expression from those that use gene expression data from a particular state to predict other phenotypes [19,22]—another important capability of genome-scale models.

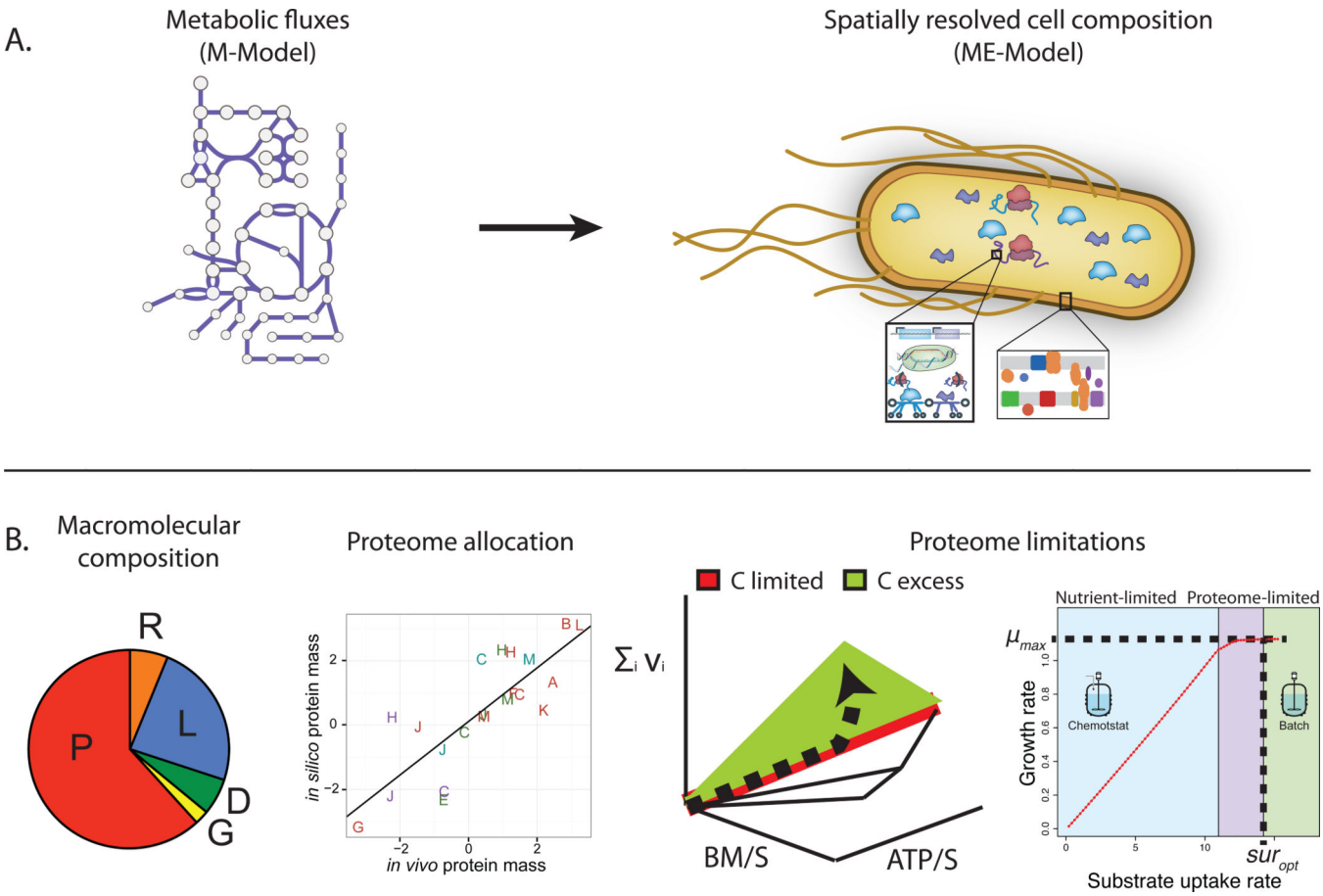

Figure 2. Prediction of spatially resolved proteome allocation and limitations

A) The expanded scope of reconstruction content advances the predictions possible with constraint-based models from network flux states (with M-Models) to the spatially resolved molecular composition of a cell (with ME-Models).

B) ME-Models predict the gross macromolecular composition of the cell and the detailed allocation of the proteome. Additionally, the effects of proteome limitations can be accounted for, including the prediction of optimal substrate uptake rates and specific pathway shifts from carbon-limited to carbon-excess environments.

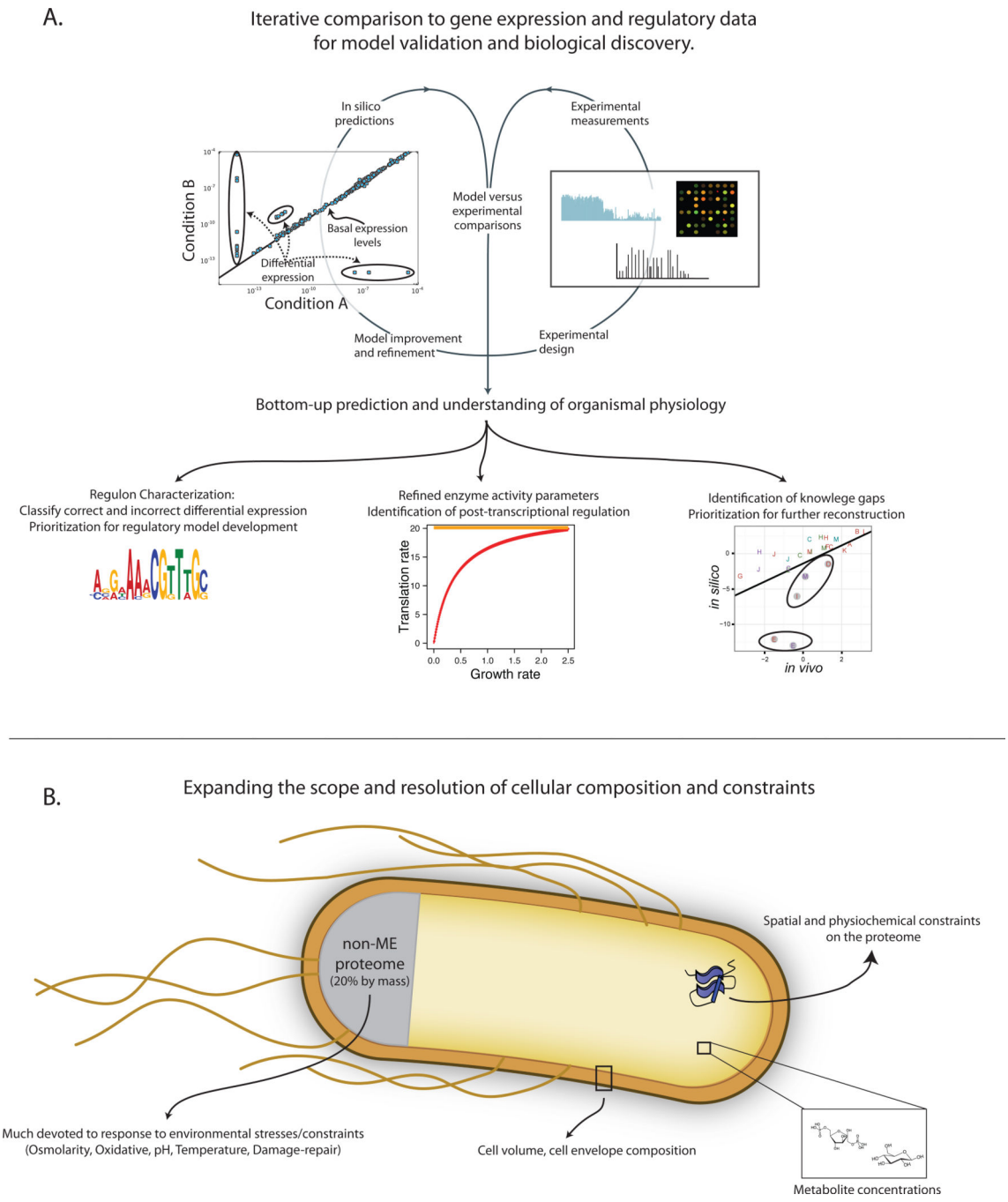

Figure 3. Iterative model validation and biological discovery enabled by expanded scope

A) The prediction of basal gene expression states and regulatory needs upon environmental shifts enables comparisons to gene expression datasets. Like M-Model predictions of gene essentiality [64] and metabolic flux [97], comparison of *in silico* and *in vivo* gene expression states will enable model validation and biological discovery.

B) To increase the scope and resolution of predicted cellular composition and organization, there are several prioritized areas for model expansion. These include metabolite

concentrations, the non-ME proteome, cell envelope composition, and the spatial organization and physiochemical constraints on the proteome (aided by protein structures).