Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Biological insights through omics data integration

Elad Noor[1,a], Sarah Cherkaoui[1,2,a] and Uwe Sauer[1]

Abstract

The number of studies that collect and publish large-scale multiomics data is rapidly increasing. So far, we have not realized the full potential of the biological insights that can be drawn from integrating these data. In this review, we present the different existing approaches for such multiomics integration and highlight innovative articles that present new methods or biological insights. We claim that the difficulty in scaling remains one of the main reasons for the underuse of dynamic mechanistic models and discuss the potential of machine learning to disrupt this scientific field as it has done for many others.

Addresses

[1] Institute of Molecular Systems Biology, ETH Zürich, Switzerland
[2] Program in Systems Biology, Life Science Graduate School, Zürich, Switzerland

Corresponding author: Sauer, Uwe. (sauer@ethz.ch)
[a] These authors contributed equally to this review.

Current Opinion in Systems Biology 2019, **15**:39–47

This review comes from a themed issue on **Gene regulation**

Edited by **Mariko Okada** and **Shinya Kuroda**

For a complete overview see the Issue and the Editorial

Available online 21 March 2019

https://doi.org/10.1016/j.coisb.2019.03.007

Keywords

Multiomics, Data integration, Machine learning, Knowledge-based approaches, Data-driven approaches.

Introduction

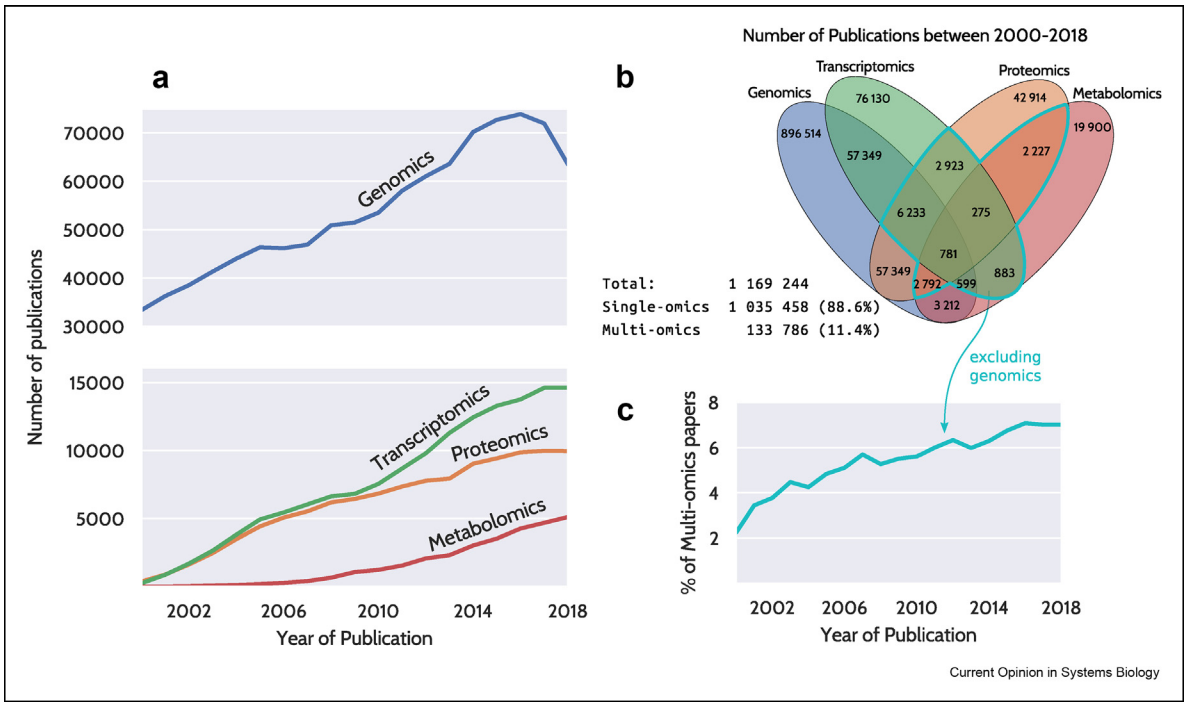
The genomics revolution marked the transition from focusing on a handful of genes at a time to simultaneously measuring as many of these cellular components as possible. The study of many other biological entities, such as mRNA transcripts, proteins, or metabolites, has since followed similar paths, carrying with them the -omics suffix. The logical next step, denoted as multiomics, was to scale up the number of different omics performed in each study, boosted by the increasing popularity of collaborative science and the availability of service providers that facilitate outsourcing some/most of these analyses (Figure 1a) [1–4]. As more and more biological data sets are being gathered at a rapidly accelerating pace, the data analysis problem is greatly exacerbated, in particular, the challenge of drawing biological insights from it. In this review, we define integration as the process of analyzing multiomics data sets in a synergistic fashion, in contrast to separate omics analysis, in the prospect of gaining more knowledge from the combination. In 2013, the U.S. National Institute of Health (NIH) started the Big Data to Knowledge program to foster development of new ways to analyze the overwhelming amount of gathered data [5]. For example, one of the program's test-case data sets, Trans-Omics for Precision Medicine, aims to collect whole-genome sequences, DNA methylation signatures, transcriptomics, metabolomics, and proteomics from 120,000 individuals.

These and other multiomics databases lay the foundation for medical prognosis tools or predicting disease trajectories. Even though the number of multiomics studies increases steadily (Figure 1b and c), surprisingly few of them integrate different omics and even fewer develop novel ways for performing the integration. Although general statistical methods exist for analyzing multiomics data [7,8], they do not consider known mechanistic or functional interactions between the measured components. Models describing these interactions on the physical level are key for a deeper understanding of the underlying mechanisms. The most common abstraction that ubiquitously appears in such models is the network [9], but more often than not, more specific descriptions that include quantitative and temporal information are necessary to capture biological phenomena. Here, we review recent work that presented novel approaches for integrating omics data or provided biological insights from such integration (Figure 2), broken down into two main categories: data-driven and knowledge-based approaches.

Data-driven approaches

Although the number of multiomic studies is rapidly increasing, the diversity of methods for integrating these data remains limited. Rather than relying on prior information such as component interaction or kinetic parameters, the most commonly used methods are purely data driven. We classify such methods into three categories: statistical approaches (e.g. correlation, enrichment analysis), unsupervised methods (clustering, factorization, dimension reduction techniques, and so on), and supervised machine learning.

Figure 1



Breakdown of PubMed publications where one or more omics were used. (a) The total number of articles published per year per omics since 2000. We included both -ome and -omic in the search of each of the four types (e.g. genome or genomic for the genomics query). The number of transcriptomics articles might be underestimated as the term transcriptomics is often times replaced by gene expression. Because of a change in naming conventions, the terms RNAseq and microarray were included in the transcriptomic query. **(b)** A Venn diagram showing the overlap of articles that include one or more omics, published between 2000 and 2017. Multiomics is defined as a publication that includes more than one type of omics. To date, the most common omics overlaps are genomics + proteomics and genomics + transcriptomics, which together amount to more than 10% of all omics articles. **(c)** Percentage of multiomics publications (not counting genomics) of the total publications with at least two of the three omics: transcriptomics, proteomics, and metabolomics. Genomics was omitted in order not to bias the results because the term genome is often used in other contexts (e.g. genome-wide or genome-scale). All PubMed queries were performed using BioPython [6], and the scripts can be found on the GitLab repository https://gitlab.com/elad.noor/cosb18-multiomics-integration.

Statistical methods

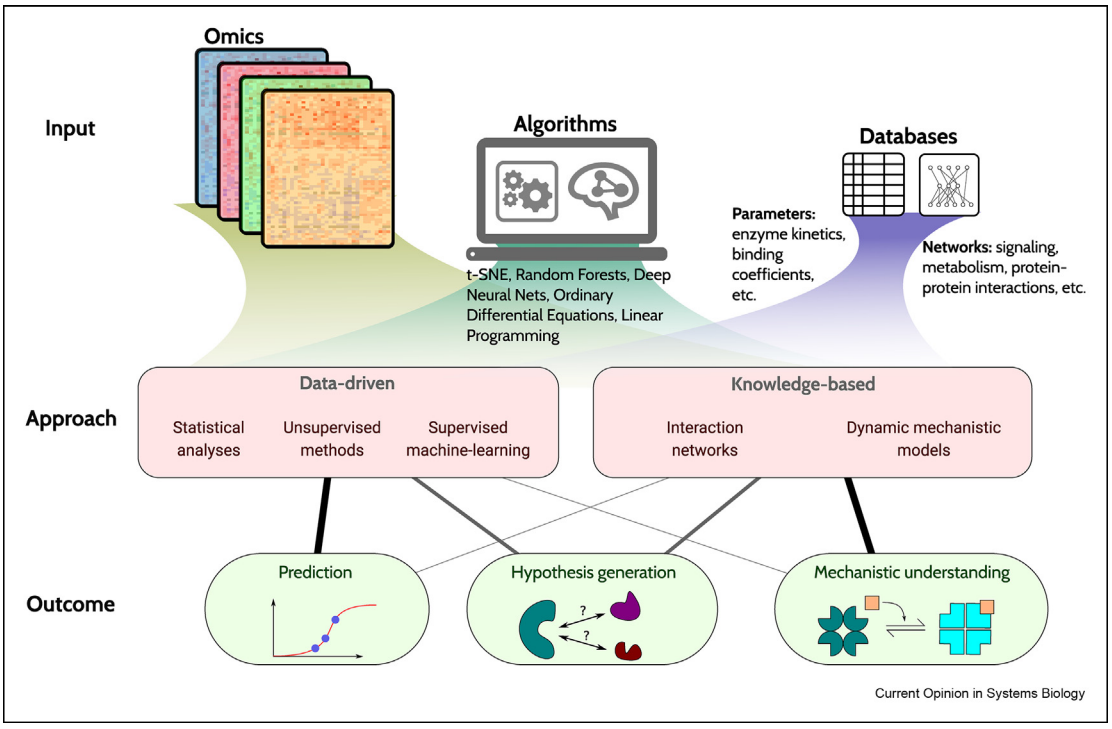
The most widely used data-driven methods for interpreting multiomics data sets are based on simple statistics, such as correlations or associations. Changes in measured entities such as proteins, mRNAs, or metabolites are compared across conditions, assuming their levels will respond similarly if they are involved in the same biological process. Such methods are commonly used in genome-wide association studies to associate genetic variants to phenotypes. As an example, a recent study of mitochondrial phenotypes collected genomic, transcriptomic, proteomic, and metabolomic data from nearly 400 mice, amounting to one of the most comprehensive multiomics data sets to date [3]. Correlating genetic loci with concentration changes through molecular quantitative trait locus[1] analysis, novel links between genes, proteins, and metabolites were identified and associated with complex metabolic phenotypes. More recently, a similar analysis of mammalian cells used correlation of proteins and metabolites to profile the temporal changes throughout the cell cycle [10].

Focusing on the small genome microbe *Mycoplasma pneumoniae*, nonparametric statistics pointed to correlations within large multiomics data sets [11]. For example, multivariate adaptive regression splines[2] allowed to infer factors that contributed most to mRNA and protein abundance, surprisingly revealing that protein half-life and the frequency of acetylation sites contributed more to protein abundance than mRNA levels. Even for well-studied model organisms such as *Escherichia coli*, very little is known about interactions between small molecules and transcription factors. By correlating metabolomics with measured promoter activities, Kochanowski et al. [13] identified the three major input signals into the transcription factor network that suffice to explain gene expression in

[1] A quantitative trait locus is a section in the genome whose sequence correlates with a quantitative trait that varies across individuals in the population.

[2] Multivariate adaptive regression splines is a nonparametric regression analysis method that is able to fit nonlinear relationships [12].

Figure 2



Scheme of two archetypal approaches to multiomics data integration. Data-driven approaches choose models that require fewer parameters or prior knowledge, therefore offering an abstract representation of the interactions between the entities. Knowledge-based approaches require more assumptions and heavily depend on literature information to populate a concrete mechanistic model but can provide more mechanistic understanding.

central metabolism under more than twenty conditions and suggested novel allosteric interactions of transcription factors. Statistical methods have the advantage of generating hypotheses rapidly by linking the different omics variables to one another; however, these links do not always reflect causality.

Unsupervised methods

Statistics methods typically evaluate single or pairs of variables at a time, rather than taking full advantage of high dimensional data. Because processing all information concurrently, or even just visualizing it, can be challenging, often unsupervised methods are used to reduce data dimensions, to highlight underlying factors within the data or to identify clusters on the basis of similarity. Unsupervised methods use machine learning algorithms to identify patterns without reference to known outcome, using unlabeled samples. For holistic analysis of high-dimensional data sets, a new technique based on matrix factorization was successfully used to identify patterns that can later be interpreted and given biological context (see review [14]). As a variant of matrix factorization, multiomics factor analysis unravels principal sources of variation and allows to infer low-dimensional factors that are representations of the original data [15]. In an application of multiomics factor analysis to single-cell multiomics data, the advantage of learning continuous factors (rather than discrete) was demonstrated for the first time and used to identify coordinated transcriptional and epigenetic changes along cell differentiation. In brief, factorization methods decompose the variation of omics data, and these factors can then be related to the studied phenotype.

Nonlinear dimension reduction techniques, such as T-distributed stochastic neighbor embedding[3] and unsupervised clustering, are powerful to characterize diverse cell types from single-cell omics data. Frequently, each omics data set is reduced using T-distributed stochastic neighbor embedding, they are then clustered, and the clustering results are compared with the cell typing results from flow cytometry (see review for all single-cell integration approaches [16]). This approach provides information on which omic is most informative for each cell type. For example, combining transcriptomics and proteomics in single-cell measurements successfully identified an unknown immune cell type that could not have been found without unsupervised methods [17]. However, one should be aware that unsupervised

[3] t-distributed stochastic neighbor embedding—a high-dimensional data visualization method based on nonlinear embedding in two or three dimensions that aims to conserve the original distances.

methods assume that the largest signal in the data is biologically meaningful. This might not always be the case if unwanted covariates, such as batch effects, are driving this variation.

Supervised machine learning

Machine learning has been hailed as a boon for the new era of data-rich biology for some time now [18–20]. In supervised learning, a set of input attributes are used to predict the value of a target. Machine learning algorithms based on linear models, such as regression, have been extensively used for such purpose. Supervised machine learning has proven useful in many applications, such as predicting synergistic drug combinations [21], microbial interactions [22], and classifying skin cancer [23]. More recently, complex machine learning algorithms such as deep neural networks have been developed to easily handle very large data sets and identify highly intricate patterns that may help in predicting biological functions, and they will increasingly be used for the integration of omics data. The challenge lies in gathering enough curated data to train the algorithm and measure its success.

A recent study of gut microbiota of 220 patients with inflammatory bowel disease has used both a statistical approach and supervised machine learning to integrate metagenomics with metabolomics and predict the disease status [24]. Using correlations, Franzosa et al. could identify metabolites associated with microbial species abundance measured in each of the two inflammatory bowel disease subtypes, and using random forest, they trained this machine learning classifier to predict a patient's disease status on the basis of metagenomics, metabolomics, and their combination. The integration of microbial species and metabolites did not, however, improve the high predictive power of using metabolites solely.

A comparative study of five unsupervised multiomics integration methods demonstrated that combining more omics can improve the accuracy of clustering but on the other hand might add noise and decrease signal strength [25]. This highlights the opportunity and the danger of adding more and more layers of information into one's model. Thus, there is an urgent need for robust, coherent, and statistically sound methods that will facilitate multiomics integration, while preventing bad practices such as p-hacking—a known phenomenon where scientists increase the significance score of their study by subtle manipulation of the data (e.g. by not including some of experimental results in the statistical analysis).

A priori, all data-driven integrations could be performed using machine learning. These methods have the potential to identify patterns from excessively large omics data to predict complex phenotypes and generate novel hypotheses. Even though machine learning holds great promise for omics integration, few applications have delivered, solely from data, mechanistic understanding of a biological system (Figure 2). Like other data-driven approaches, machine learning hardly considers the vast existing biochemical knowledge, as this is a very challenging task. It thus delivers plausible novel relationships but rarely reveals causality or evidence for direct interactions. In the following sections, we therefore review how existing knowledge-based methods have been successfully applied to multiomics data integration.

Knowledge-based approaches

To achieve mechanistic understanding, one needs to move beyond purely data-driven approaches and capitalize on the extensive prior knowledge about component interactions. In this section, we focus on studies that exemplify (in our eyes) the process of conducting knowledge-based studies. Notably, these examples do not always use knowledge-based approaches exclusively, but rather combine them with data-driven methods.

There are many ways to divide knowledge-based approaches into subcategories. Here, we chose to distinguish models on the basis of interaction networks from dynamic mechanistic ones (or topological versus kinetic). In a sense, this separation is superficial, and it might not always be obvious to which category a certain paper belongs. For example, using boolean logic to simulate signaling networks lies somewhere in between. Furthermore, constraint-based genome-scale metabolic models (which appear in many of our examples) would classically fall in the network category. In the recent years, however, more and more applications use the same models to predict dynamic metabolic behavior even without explicitly solving systems of ordinary differential equations (ODEs). Nevertheless, we believe this distinction is helpful to understand the difference in how the two categories approach multiomics integration.

Interaction networks

Sequence-based data from genome to proteome is naturally easier to enrich with literature data because the sequence information itself is transferred directly from DNA to protein. A good example is comparing transcriptome and proteome to shed light on translation efficiency, post-translational modification, and protein half-life [11,26]. Because metabolites are not directly and certainly not unambiguously connected to the other data sets [1,27,28], their integration requires some type of interaction network [29,30]. Such interaction networks are primarily metabolic models that rely on our advanced knowledge of enzymatic reactions [31] but may also include less mechanistic interaction maps, such

as those between metabolites and genes that became available recently [32] and have been used to infer the underlying genetic basis for metabolite responses [33]. Here, we focus on knowledge-based methods that integrate metabolomics and sequence-based data sets to learn about interactions between genes/proteins and metabolites such as regulation mechanisms.

One of the key questions is which mechanism regulates metabolic function, that is, flux. Gerosa et al. gathered transcriptomic, metabolomic, and fluxomic data of *E. coli* grown under different conditions. By systematically comparing the flux through each reaction in central metabolism to the different factors that affect it—namely, enzyme abundance, substrate level, and thermodynamic driving force—they were able to suggest which mechanisms regulate the flux when cells transition from one environment to another [34].

The Rabinowitz lab used a similar multiomic data set to specifically study the regulation of flux by small molecules (i.e. allosteric regulation and substrate/product inhibition). This systematic identification of meaningful metabolic enzyme regulation combined steady-state proteomic, fluxomic, and metabolomic data to suggest control mechanisms that regulate the flux in one reaction at a time [35]. By fitting a detailed metabolic rate law that uses the concentrations of an enzyme, together with all its substrates and products, and testing whether one or two allosteric regulators can improve that fit, they identified 29 yeast enzymes where regulation by a metabolite was statistically supported. The works of Hackett et al. and Gerosa et al. together are an inspiring demonstration of how multiomics data can (paradoxically) reduce the amount of computation required for modeling a system. Directly observing most (or all) of the time-dependent variables circumvents the need to simulate a kinetic model and instead fits a kinetic rate law for every single enzymatic reaction, one at a time.

The Milo lab used a similar focus on single reactions to bypass the need for a kinetic rate law altogether, by simply comparing fluxomic and proteomic data across many conditions and finding the maximal k_{eff} —that is, the ratio of flux to protein abundance for a specific reaction [36]. This way, they were able to predict *in vivo* turnover rates for 132 *E. coli* enzymes. Such effective enzyme turnover rates have become increasingly valuable as models of metabolism have grown in complexity to include protein translation and genetic regulation [37–39]. By constraining the maximal possible flux in each reaction, based on the enzyme abundance and k_{eff}, Sánchez et al. [40] showed how genome-scale metabolic models predict phenotypes such as growth rates and overflow metabolism more precisely than previous approaches. As with previous examples, using the notion of k_{eff} helps in overcoming the computational complexity that is required for solving dynamic models of enzyme kinetics, by keeping the system linear and greatly reducing the number of variables and parameters (e.g. metabolite concentrations and Michaelis–Menten constants can be ignored).

One of the first approaches that combined machine learning and biochemical knowledge came from the Ralser lab. Rather than using simplified kinetic models or focusing on single reactions at a time, they applied a hybrid approach combining a metabolic network with machine learning to construct a predictive model of metabolite abundances [41]. They showed that half of the variation in metabolite concentrations among yeast kinase knockouts can be predicted from proteomic measurements of enzymes that are 1–2 steps away in the network. The significance of this finding is that although we do not have a mechanistic description of how enzyme abundances determine steady-state metabolite levels, the information is there and can be used predictively. From the comparison of 12 different machine learning algorithms, ridge regression (a type of linear regression) was demonstrated to have the highest predictive value, although it is simpler and involves fewer parameters and more stringent assumptions than other algorithms.

The works from the Rabinowitz, Milo, and Ralser labs are promising examples of interdisciplinary science. All cases start by gathering multiomics data, but rather than applying the data-driven methods discussed earlier directly, the authors added an intermediate step, where prior knowledge (e.g. about the metabolic network or mechanisms of enzyme kinetics) was used to focus the analysis and to infer high-quality information and insights from it. Finally, statistical or machine learning methods were applied to validate the models.

Constraint-based models, such as flux balance analysis, originally focused only on metabolic networks and are now being expanded to include more cellular processes, such as translation and transcription. These algorithms, together with a growing set of tools, facilitate the integration of genomic, transcriptomic, and metabolomic data and aid in interpreting multiomic analyses [42–46]. However, while the use of prior information in the form of molecular component interactions improves biological insights, the primary value of such approaches is in generating new hypotheses. The ability of such models to simulate a given system is somewhat limited because they do not typically represent the observable entities in a quantitative fashion (unlike dynamic models which we discuss next). On the flip side, topological models are much more scalable and have proven to be very useful in countless applications, such as predicting gene essentiality, metabolic engineering, drug design, and stability analysis.

Dynamic mechanistic models

To maximize predictive power and mechanistic insights on the molecular level, ODE simulations based on physical models of binding and catalysis remain the gold standard. Combining such mechanistic models for all the different systems that exist within a cell is one of the biggest challenges in system biology [47]. A very precise model of an entire human cell, for example, could potentially make all experiments in live cell cultures obsolete. However, simulating large ODE systems is a computationally intensive task, and therefore, most contemporary computational efforts focus on specific cell compartments or processes [48], use a higher level of abstraction such as logic modeling for gene regulation networks [49], or use linear kinetic rate laws for metabolic networks [43,50,51]. Others take advantage of the separation of timescales between metabolic response (seconds) and gene expression (minutes), to construct a bilevel model where proteins are assumed to have constant concentrations on the short time scale and metabolites are assumed to reach quasi-steady-state in the long time scale [52–56]. These approaches allow modelers to avoid integrating large ODE systems and are partially solved by linear programming.

In the recent years, larger models with more concrete kinetic rate laws have been developed for *E. coli* [57,58], *Saccharomyces cerevesiae* [59], and even human cells [60]. However, increasing the scale of mechanistic models requires more than just faster computers but also depends on the combined effort of countless biochemists for collecting the relevant parameters and interaction networks [61]. So far, there are no high-throughput methods for experimentally measuring binding constants, enzyme Michaelis–Menten constants, or reaction equilibrium constants, and even if those were known, one would need to know the regulatory mechanisms that modulate them.

Owing to these challenges, dynamic models are rarely used to integrate multiomics data, and when they are applied, only small subsets of the entire omics data are actually utilized. Perhaps this explains why knowledge-based approaches are often discarded completely and replaced by data-driven methods. For instance, understanding the relevant mechanisms that cause high-level, emergent phenomena such as disease state in humans is extremely complex and would likely require an impossibly large model. One could imagine, though, that mechanistic models of disease-relevant subsystems such as a signaling pathway could be combined in hybrid models with more coarse-grained interaction networks to make use of the entire omics data set.

Conclusions

Multiomics articles will be more frequent in the near future, and promising new high-throughput methods such as limited proteolysis for unbiased, and proteome-wide profiling of protein conformational changes [62], proteomics of proteoforms [63], single-cell omics measurements [17], or small-molecule interactions [64] will be more routinely added to the repertoire. Such rich multiomics data sets will inspire the development of more complex mechanistic models that can take advantage of the new data. For example, genome-scale metabolic models have already began to routinely integrate transcriptomic and proteomic data [43], and we will likely have more layers of information added in the near future.

Hence, there will be an increased demand for data-driven and knowledge-based integration methods. In this context, machine learning holds great promise and is already used intensively in biomedical applications [65,66]. It remains presently unclear how this powerful approach can actually forward our understanding of biological systems. The difficulty of 'looking into the brain' of a learning algorithm is a known problem and is under active research by computer scientists. Notably, this difficulty goes also in the other direction—namely, we do not have effective ways for facilitating the machine learning process by letting it access the vast amount of literature data and insights. One general idea called surrogate modeling is to use a mechanistic model to simulate more data and use that as extra input for the machine learning algorithm [67]. As new methods and machine learning approaches are developed and adopted by biologists [68], they are likely to become an essential part of multiomics data analysis. Hybrid techniques that take advantage of the vast computational capabilities while maintaining the underlying mechanistic model are now beginning to emerge and will hopefully gain popularity. It remains to be seen whether a generalized algorithm that can handle all the different types of omics data and network-based information will one day become the default tool for biological data analysis.

Conflict of interest statement

Nothing declared.

Acknowledgements

The authors are grateful to Mattia Zampieri for critical comments on the manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M: **Designing and interpreting 'multi-omic' experiments that may change our understanding of biology**. *Curr Opin Struct Biol* 2017, **6**:37–45, https://doi.org/10.1016/j.coisb.2017.08.009. http://www.sciencedirect.com/science/article/pii/S2452310017300835.

2. Vilanova C, Porcar M: **Are multi-omics enough?** *Nat Microbiol* 2016, **1**:16101, https://doi.org/10.1038/nmicrobiol.2016.101. https://www.nature.com/articles/nmicrobiol2016101.

3. ** Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, Houten SM, Amariuta T, Wolski W, Zamboni N, Aebersold R, Auwerx J: **Systems proteomics of liver mitochondria function**. *Science* 2016, **352**, https://doi.org/10.1126/science.aad0189. aad0189, http://science.sciencemag.org/content/352/6291/aad0189.
Studied the metabolic function of 386 mice from a genetic reference population under various environmental conditions. On top of their genotypes, they have measured over 25′,000 transcripts, 2′500 proteins and nearly 1000 metabolites, which makes it one of the largest multi-omics study to date. They identified the genomic variants of mitochondrial enzymes that caused inborn error in metabolism and revealed two genes that appear to function in cholesterol metabolism.

4. Yugi K, Kubota H, Hatano A, Kuroda S: **Trans-omics: how to reconstruct biochemical networks across multiple 'omic' layers**. *Trends Biotechnol* 2016, **34**:276–290, https://doi.org/10.1016/j.tibtech.2015.12.013. http://www.sciencedirect.com/science/article/pii/S0167779915002735.

5. Bui AAT, Van Horn JD: **Envisioning the future of 'big data' biomedicine**. *J Biomed Inform* 2017, **69**:115–117, https://doi.org/10.1016/j.jbi.2017.03.017. http://www.sciencedirect.com/science/article/pii/S1532046417300709.

6. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**:1422–1423, https://doi.org/10.1093/bioinformatics/btp163. https://academic.oup.com/bioinformatics/article/25/11/1422/330687.

7. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L: **Methods for the integration of multi-omics data: mathematical aspects**. *BMC Bioinf* 2016, **17**:S15, https://doi.org/10.1186/s12859-015-0857-9. https://doi.org/10.1186/s12859-015-0857-9.

8. Rohart F, Gautier B, Singh A, Cao K-AL: **mixOmics: an R package for 'omics feature selection and multiple data integration**. *PLoS Comput Biol* 2017, **13**, https://doi.org/10.1371/journal.pcbi.1005752. e1005752, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752.

9. Newman M: *Networks*. Oxford University Press; 2018. google-Books-ID: YdZjDwAAQBAJ.

10. * Lee H-J, Jedrychowski MP, Vinayagam A, Wu N, Shyh-Chang N, Hu Y, Min-Wen C, Moore JK, Asara JM, Lyssiotis CA, Perrimon N, Gygi SP, Cantley LC, Kirschner MW: **Proteomic and metabolomic characterization of a mammalian cellular transition from quiescence to proliferation**. *Cell Rep* 2017, **20**:721–736, https://doi.org/10.1016/j.celrep.2017.06.074. http://www.sciencedirect.com/science/article/pii/S2211124717309051.
Collected time-course metabolomic and proteomic data from arrested pro-B cells in response to IL-3 activation and find many patterns commonly associated with cancer cells such as high uptake of methionine in G1.

11. Chen W-H, van Noort V, Lluch-Senar M, Hennrich ML, Wodke JAH, Yus E, Alibés A, Roma G, Mende DR, Pesavento C, Typas A, Gavin A-C, Serrano L, Bork P: **Integration of multi-omics data of a genome-reduced bacterium: prevalence of post-transcriptional regulation and its correlation with protein abundances**. *Nucleic Acids Res* 2016, **44**:1192–1202, https://doi.org/10.1093/nar/gkw004. https://academic.oup.com/nar/article/44/3/1192/2503077.

12. Maaten L v d, Hinton G: **Visualizing Data using t-SNE**. *J Mach Learn Res* 2008, **9**:2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html.

13. Kochanowski K, Gerosa L, Brunner SF, Christodoulou D, Nikolaev YV, Sauer U: **Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli**. *Mol Syst Biol* 2017, **13**:903.

14. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, Xu Y, Fertig EJ: **Enter the matrix: factorization uncovers knowledge from omics**. *Trends Genet* 2018, **34**:790–805, https://doi.org/10.1016/j.tig.2018.07.003. http://www.sciencedirect.com/science/article/pii/S0168952518301240.

15. ** Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O: **Multi–Omics Factor Analysis–a framework for unsupervised integration of multi–omics data sets**. *Mol Syst Biol* 2018, **14**, https://doi.org/10.15252/msb.20178124. e8124, http://msb.embopress.org/content/14/6/e8124.
Developed a method to unravel principal sources of variation in multi-omics data using factor analysis. The method disentangles axes of heterogeneity that are shared across multiple omics and those specific to individual omic. They applied their method to two datasets where they successfully identified the major drivers in chronic lymphocytic leukemia, a clinically and biologically heterogeneous disease, and recover a differentiation trajectory in single-cell multi–omics data.

16. Colomé-Tatché M, Theis FJ: **Statistical single cell multi-omics integration**. *Curr Opin Struct Biol* 2018, **7**:54–59, https://doi.org/10.1016/j.coisb.2018.01.003. http://www.sciencedirect.com/science/article/pii/S2452310018300039.

17. * Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA: **Multiplexed quantification of proteins and transcripts in single cells**. *Nat Biotechnol* 2017, **35**:936–939, https://doi.org/10.1038/nbt.3973. https://www.nature.com/articles/nbt.3973/.
Developed a tool to simultaneously measure proteins and mRNAs in single cells. Using unsupervised methods, they successfully combined these transcriptomics and proteomics datasets to identify and characterize an unknown immune cell type that could not have been found without this approach.

18. Almeida JS: **Predictive non-linear modeling of complex data by artificial neural networks**. *Curr Opin Biotechnol* 2002, **13**:72–76, https://doi.org/10.1016/S0958-1669(02)00288-4. http://www.sciencedirect.com/science/article/pii/S0958166902002884.

19. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ: **Next-Generation machine learning for biological networks**. *Cell* 2018, **173**:1581–1592, https://doi.org/10.1016/j.cell.2018.05.015. http://www.sciencedirect.com/science/article/pii/S0092867418305920.

20. Tarca AL, Carey VJ, Chen X-w, Romero R, Drăghici S: **Machine learning and its applications to biology**. *PLoS Comput Biol* 2007, **3**:e116, https://doi.org/10.1371/journal.pcbi.0030116. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030116.

21. Weinstein ZB, Bender A, Cokol M: **Prediction of synergistic drug combinations**. *Curr Opin Struct Biol* 2017, **4**:24–28, https://doi.org/10.1016/j.coisb.2017.05.005. http://www.sciencedirect.com/science/article/pii/S2452310017300197.

22. DiMucci D, Kon M, Segré D: **Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks**. *mSystems* 2018, **3**, https://doi.org/10.1128/mSystems.00181-18. e00181–18, https://msystems.asm.org/content/3/5/e00181-18.

23. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature* 2017, **542**:115–118, https://doi.org/10.1038/nature21056. https://www.nature.com/articles/nature21056.

24. ** Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, Sauk JS, Wilson RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F, Porter JA, Zhernakova A, Fu J, Weersma RK, Wijmenga C, Clish CB, Vlamakis H, Huttenhower C, Xavier RJ: **Gut microbiome structure and metabolic activity in inflammatory bowel disease**. *Nat Microbiol* 2018, https://doi.org/10.1038/s41564-018-0306-4. https://www.nature.com/articles/s41564-018-0306-4; 2018.
Collected metabolomic and metagenomic data from the gut microbiota of 220 patients with inflammatory bowel diseases. They identified many metabolites associated with the different conditions, and trained a machine learning classifier that could accurately predict a patient's disease status.

25. Tini G, Marchetti L, Priami C, Scott-Boyer M-P: **Multi-omics integration–a comparison of unsupervised clustering methodologies**. *Brief Bioinform* 2017, https://doi.org/10.1093/bib/bbx167. https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbx167/4758623; 2017.

26. Hausser J, Mayo A, Alon U: **Central dogma rates and the trade-off between precision and economy**. *bioRxiv* 2018:276139,

https://doi.org/10.1101/276139. https://www.biorxiv.org/content/early/2018/03/04/276139; 2018.

27. Donati S, Sander T, Link H: **Crosstalk between transcription and metabolism: how much enzyme is enough for a cell?** *Wiley Interdiscip Rev: Syst Biol Med* 2018, **10**, https://doi.org/10.1002/wsbm.1396. e1396, https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1396.

28. Yugi K, Kuroda S: **Metabolism as a signal generator across trans-omic networks at distinct time scales**. *Curr Opin Struct Biol* 2018, **8**:59–66, https://doi.org/10.1016/j.coisb.2017.12.002. http://www.sciencedirect.com/science/article/pii/S2452310017302044.

29. Chiappino-Pepe A, Pandey V, Ataman M, Hatzimanikatis V: **Integration of metabolic, regulatory and signaling networks towards analysis of perturbation and dynamic responses**. *Curr Opin Struct Biol* 2017, **2**:59–66, https://doi.org/10.1016/j.coisb.2017.01.007. http://www.sciencedirect.com/science/article/pii/S245231001730032X.

30. Sévin DC, Kuehne A, Zamboni N, Sauer U: **Biological insights through nontargeted metabolomics**. *Curr Opin Biotechnol* 2015, **34**:1–8, https://doi.org/10.1016/j.copbio.2014.10.001. http://www.sciencedirect.com/science/article/pii/S0958166914001694.

31. King ZA, Lloyd CJ, Feist AM, Palsson BO: **Next-generation genome-scale models for metabolic engineering**. *Curr Opin Biotechnol* 2015, **35**:23–29, https://doi.org/10.1016/j.copbio.2014.12.016. http://www.sciencedirect.com/science/article/pii/S0958166914002316.

32. Fuhrer T, Zampieri M, Sévin DC, Sauer U, Zamboni N: **Genomewide landscape of gene–metabolome associations in Escherichia coli**. *Mol Syst Biol* 2017, **13**:907.

33. Zampieri M, Zimmermann M, Claassen M, Sauer U: **Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations**. *Cell Rep* 2017, **19**:1214–1228, https://doi.org/10.1016/j.celrep.2017.04.002. https://www.cell.com/cell-reports/abstract/S2211-1247(17)30461-8.

34. Gerosa L, van Rijsewijk BBR, Christodoulou D, Kochanowski K, Schmidt BTS, Noor E, Sauer U: **Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data**. *Cell Syst* 2015, **1**.

35. * Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, Gibney PA, Botstein D, Storey JD, Rabinowitz JD: **Systems-level analysis of mechanisms regulating yeast metabolic flux**. *Science* 2016, **354**, https://doi.org/10.1126/science.aaf2786. aaf2786, http://science.sciencemag.org/content/354/6311/aaf2786.
Collected metabolomic and proteomic data from yeast cells in 25 different steady-state conditions and develop a method to identify small molecule regulations in vivo. By fitting a reversible Michaelis–Menten kinetic rate law for 56 reactions and assessing whether a potential regulator improves that fit, they found three previously unknown cross-pathway regulators.

36. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, Barenholz U, Goldenfeld M, Shlomi T, Milo R: **Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements**. *Proc Natl Acad Sci Unit States Am* 2016, **113**:3401–3406, https://doi.org/10.1073/pnas.1514240113. http://www.pnas.org/content/113/12/3401.

37. Davidi D, Milo R: **Lessons on enzyme kinetics from quantitative proteomics**. *Curr Opin Biotechnol* 2017, **46**:81–89, https://doi.org/10.1016/j.copbio.2017.02.007. http://www.sciencedirect.com/science/article/pii/S0958166916302154.

38. Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, Lerman JA, Lechner A, Sastry A, Bordbar A, Feist AM, Palsson BO: **Multi-omic data integration enables discovery of hidden biological regularities**. *Nat Commun* 2016, **7**:13091, https://doi.org/10.1038/ncomms13091. https://www.nature.com/articles/ncomms13091.

39. Ramon C, Gollub MG, Stelling J: **Integrating –omics data into genome-scale metabolic network models: principles and challenges**. *Essays Biochem* 2018, **62**:563–574, https://doi.org/10.1042/EBC20180011. http://essays.biochemistry.org/content/62/4/563.

40. Sánchez BJ, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, Nielsen J: **Improving the phenotype predictions of a yeast genome–scale metabolic model by incorporating enzymatic constraints**. *Mol Syst Biol* 2017, **13**:935, https://doi.org/10.15252/msb.20167411. http://msb.embopress.org/content/13/8/935.

41. * Zelezniak A, Vowinckel J, Capuano F, Messner CB, Demichev V, Polowsky N, Mülleder M, Kamrad S, Klaus B, Keller MA, Ralser M: **Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knockouts**. *Cell Syst* 2018, https://doi.org/10.1016/j.cels.2018.08.001. http://www.sciencedirect.com/science/article/pii/S2405471218303168; 2018.
Used a library of 97 yeast kinase knockout strains to gather a large proteomic and metabolomic dataset. Testing many different machine learning algorithms, they found that about half of the variation in a metabolite concentrations can be predicted using the levels of its neighboring enzymes.

42. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, Diener C, Dräger A, Ebert BE, Edirisinghe JN, Faria JP, Feist A, Fengos G, Fleming RMT, Garcia-Jimenez B, Hatzimanikatis V, Helvoirt W v, Henry C, Hermjakob H, Herrgard MJ, Kim HU, King Z, Koehorst JJ, Klamt S, Klipp E, Lakshmanan M, Novere NL, Lee D-Y, Lee SY, Lee S, Lewis NE, Ma H, Machado D, Mahadevan R, Maia P, Mardinoglu A, Medlock GL, Monk J, Nielsen J, Nielsen LK, Nogales J, Nookaew I, Resendis O, Palsson B, Papin JA, Patil KR, Poolman M, Price ND, Richelle A, Rocha I, Sanchez B, Schaap P, Sheriff RSM, Shoaie S, Sonnenschein N, Teusink B, Vilaca P, Vik JO, Wodke JA, Xavier JC, Yuan Q, Zakhartsev M, Zhang C: **Memote: a community-driven effort towards a standardized genome-scale metabolic model test suite**. *bioRxiv* 2018:350991, https://doi.org/10.1101/350991. https://www.biorxiv.org/content/early/2018/07/11/350991; 2018.

43. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO: **COBRAme: a computational framework for genome-scale models of metabolism and gene expression**. *PLoS Comput Biol* 2018, **14**, https://doi.org/10.1371/journal.pcbi.1006302. e1006302, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006302.

44. Machado D, Andrejev S, Tramontano M, Patil KR: **Fast automated reconstruction of genome-scale metabolic models for microbial species and communities**. *Nucleic Acids Res* 2018, **46**:7542–7553, https://doi.org/10.1093/nar/gky537. https://academic.oup.com/nar/article/46/15/7542/5042022.

45. Salvy P, Fengos G, Ataman M, Pathier T, Soh KC, Hatzimanikatis V: **pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based flux analysis**. *Bioinformatics* 2018, https://doi.org/10.1093/bioinformatics/bty499. https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty499/5047753; 2018.

46. Sergushichev AA, Loboda AA, Jha AK, Vincent EE, Driggers EM, Jones RG, Pearce EJ, Artyomov MN: **GAM: a web-service for integrated transcriptional and metabolic network analysis**. *Nucleic Acids Res* 2016, **44**:W194–W200, https://doi.org/10.1093/nar/gkw266. https://academic.oup.com/nar/article/44/W1/W194/2499310.

47. Waltemath D, Karr JR, Bergmann FT, Chelliah V, Hucka M, Krantz M, Liebermeister W, Mendes P, Myers CJ, Pir P, Alaybeyoglu B, Aranganathan NK, Baghalian K, Bittig AT, Burke PEP, Cantarelli M, Chew YH, Costa RS, Cursons J, Czauderna T, Goldberg AP, Gómez HF, Hahn J, Hameri T, Gardiol DFH, Kazakiewicz D, Kiselev I, Knight-Schrijver V, Knüpfer C, König M, Lee D, Lloret-Villas A, Mandrik N, Medley JK, Moreau B, Naderi-Meshkin H, Palaniappan SK, Priego-Espinosa D, Scharm M, Sharma M, Smallbone K, Stanford NJ, Song J, Theile T, Tokic M, Tomar N, Touré V, Uhlendorf J, Varusai TM, Watanabe LH, Wendland F, Wolfien M, Yurkovich JT, Zhu Y, Zardilis A, Zhukova A, Schreiber F: **Toward community standards and software for whole-cell modeling**. *IEEE Trans Biomed Eng* 2016, **63**, https://doi.org/10.1109/TBME.2016.2560762. 2007–2014.

48. Heiske M, Letellier T, Klipp E: **Comprehensive mathematical model of oxidative phosphorylation valid for physiological and pathological conditions**. *FEBS J* 2017, **284**:2802–2828,

https://doi.org/10.1111/febs.14151. https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/febs.14151.

49. Le Novére N: **Quantitative and logic modelling of molecular and gene networks**. *Nat Rev Genet* 2015, **16**:146–158, https://doi.org/10.1038/nrg3885. https://www.nature.com/articles/nrg3885.

50. Karr RJ, Sanghvi CJ, Macklin ND, Gutschow VM, Jacobs MJ, Bolival B, Assad-Garcia N, Glass IJ, Covert WM: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**:389–401.

51. Reimers A-M, Knoop H, Bockmayr A, Steuer R: **Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth**. *Proc Natl Acad Sci Unit States Am* 2017, **114**: E6457–E6465, https://doi.org/10.1073/pnas.1617508114. http://www.pnas.org/content/114/31/E6457.

52. Lee JM, Gianchandani EP, Eddy JA, Papin JA: **Dynamic analysis of integrated signaling, metabolic, and regulatory networks**. *PLoS Comput Biol* 2008, **4**, https://doi.org/10.1371/journal.pcbi.1000086. e1000086, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000086.

53. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophys J* 2002, **83**:1331–1340, https://doi.org/10.1016/S0006-3495(02)73903-9. http://www.sciencedirect.com/science/article/pii/S0006349502739039.

54. Richard G, Chang H, Cizelj I, Belta C, Julius AA, Amar S: **Integration of large-scale metabolic, signaling, and gene regulatory networks with application to infection responses**. In *2011 50th IEEE conference on decision and control and European control conference*; 2011:2227–2232, https://doi.org/10.1109/CDC.2011.6160954.

55. Watanabe L, König M, Myers C: **Dynamic flux balance analysis models in SBML**. *bioRxiv* 2018:245076, https://doi.org/10.1101/245076. https://www.biorxiv.org/content/early/2018/01/08/245076; 2018.

56. Yang L, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO: **DynamicME: dynamic simulation and refinement of integrated models of metabolism and protein expression**. *bioRxiv* 2018:319962, https://doi.org/10.1101/319962. https://www.biorxiv.org/content/early/2018/05/15/319962; 2018.

57. Stanford NJ, Lubitz T, Smallbone K, Klipp E, Mendes P, Liebermeister W: **Systematic construction of kinetic models from genome-scale metabolic networks**. *PLoS One* 2013, **8**, e79195, https://doi.org/10.1371/journal.pone.0079195. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079195.

58. * Khodayari A, Maranas CD: **A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains**. *Nat Commun* 2016, **7**:13806, https://doi.org/10.1038/ncomms13806. https://www.nature.com/articles/ncomms13806.
Introduce a genome scale kinetic model for *E. coli* comprising 457 reactions and 337 metabolites which is the most comprehensive to date. They parameterize it using a genetic algorithm and multiple measurements of fluxomic data.

59. Smallbone K, Mendes P: **Large-scale metabolic models: from reconstruction to differential Equations**. *Ind Biotechnol* 2013, **9**:179–184, https://doi.org/10.1089/ind.2013.0003. https://www.liebertpub.com/doi/abs/10.1089/ind.2013.0003.

60. Bordbar A, McCloskey D, Zielinski D, Sonnenschein N, Jamshidi N, Palsson B: **Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics**. *Cell Syst* 2015, **1**:283–292, https://doi.org/10.1016/j.cels.2015.10.003. http://www.sciencedirect.com/science/article/pii/S2405471215001490.

61. Tummler K, Klipp E: **The discrepancy between data for and expectations on metabolic models: how to match experiments and computational efforts to arrive at quantitative predictions?** *Curr Opin Struct Biol* 2018, **8**:1–6, https://doi.org/10.1016/j.coisb.2017.11.003. http://www.sciencedirect.com/science/article/pii/S2452310017301920.

62. Piazza I, Kochanowski K, Cappelletti V, Fuhrer T, Noor E, Sauer U, Picotti P: **A map of protein-metabolite interactions reveals principles of chemical communication**. *Cell* 2018, **172**, https://doi.org/10.1016/j.cell.2017.12.006. 358–372.e23, http://www.sciencedirect.com/science/article/pii/S0092867417314484.

63. Smith LM, Kelleher NL: **Proteoforms as the next proteomics currency**. *Science* 2018, **359**:1106–1107, https://doi.org/10.1126/science.aat1884. http://science.sciencemag.org/content/359/6380/1106.

64. Diether M, Sauer U: **Towards detecting regulatory protein–metabolite interactions**. *Curr Opin Microbiol* 2017, **39**: 16–23, https://doi.org/10.1016/j.mib.2017.07.006. http://www.sciencedirect.com/science/article/pii/S1369527417300164.

65. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: **Machine learning applications in cancer prognosis and prediction**. *Comput Struct Biotechnol J* 2015, **13**: 8–17, https://doi.org/10.1016/j.csbj.2014.11.005. http://www.sciencedirect.com/science/article/pii/S2001037014000464.

66. Sirinukunwattana K, Raza SEA, Tsang Y, Snead DRJ, Cree IA, Rajpoot NM: **Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images**. *IEEE Trans Med Imaging* 2016, **35**:1196–1206, https://doi.org/10.1109/TMI.2016.2525803.

67. Kogadeeva M, Zamboni N: **SUMOFLUX: a generalized method for targeted 13c metabolic flux ratio analysis**. *PLoS Comput Biol* 2016, **12**, https://doi.org/10.1371/journal.pcbi.1005109. e1005109, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005109.

68. Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, Desouki AA, Lercher MJ, Palsson BO: **Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models**. *Nat Commun* 2018, **9**: 5252, https://doi.org/10.1038/s41467-018-07652-6. https://www.nature.com/articles/s41467-018-07652-6.