Machine learning alternative to systems biology should not solely depend on data

Hock Chuan Yeo and Kumar Selvarajoo (iD)

Corresponding author. Kumar Selvarajoo. Tel.: 64788396; E-mail: kumar_selvarajoo@bii.a-star.edu.sg

Abstract

In recent years, artificial intelligence (AI)/machine learning has emerged as a plausible alternative to systems biology for the elucidation of biological phenomena and in attaining specified design objective in synthetic biology. Although considered highly disruptive with numerous notable successes so far, we seek to bring attention to both the fundamental and practical pitfalls of their usage, especially in illuminating emergent behaviors from chaotic or stochastic systems in biology. Without deliberating on their suitability and the required data qualities and pre-processing approaches beforehand, the research and development community could experience similar 'AI winters' that had plagued other fields. Instead, we anticipate the integration or combination of the two approaches, where appropriate, moving forward.

Keywords: mechanistic modeling, machine learning, AI, systems biology, synthetic biology, metabolic engineering

Introduction

Human history was marked by transitions, whereby machines had been innovatively built to replace, or make it easier for mankind, to carry out laborious, time-consuming and/or mundane tasks. The current development of artificial intelligence (AI) and machine learning (ML) techniques and tools in the biological and biotechnological domains can be viewed similarly for bringing about smarter and more automated analysis and decision-making to these fields [1]. AI/ML applications in these domains particularly benefitted from the advent of high-throughput multi-omics profiling [2], which provides the large amount of data needed at each regulatory level (genomics, transcriptomics, proteomics, metabolomics, etc.) [3, 4] for the reliable and accurate prediction of biological behaviors. However, black-box AI/ML methods that are solely dependent on data, by nature, unable to provide the mechanistic basis for justifying the explanation of complex behaviors, and in making reliable predictions.

Conversely, systems biology [5] underlines the principle that complex behaviors can emerge holistically from the mechanistic interactions among the components of a biological system (e.g. cell fate transitions [6]), and models them mathematically to predict experimental observations. However, it often requires detailed description of the molecular mechanisms involved. For example, modeling the temporal dynamics of biomolecules in a metabolic system will entail the prescription of a set of quantitative rate laws describing the material physico-chemical interactions among enzymes, substrates and regulator molecules, taking into account various intracellular effects (compartmental, membrane, channeling effects [7], etc.) and regulation (post-translational [8] and allosteric [9] regulations, among others).

In doing so, systems biology further provides a fundamental basis for the rigorous integration of metabolomics and proteomics (enzymes) datasets, while remaining amenable to the productive analytical frameworks of other fields, such as those of chaos, control and ergodic theories. However, such intricate mechanistic details and their parameter values are mostly unavailable [10, 11] and challenging to obtain experimentally [12], requiring considerable time and effort to attain in each context.

Considering the challenges of systems biology, AI/ML remains an alternative for investigating system behaviors [13] and improving the output of interest in synthetic biology applications [14]. Nonetheless, there remains considerable challenges for an AI/ML alternative to systems biology, especially if it is solely dependent on (omics) data, despite its successes in specific categories of biomedical applications [15]. Here, we collate the practical pitfalls found in the literature and reiterate the fundamental challenges to do so. We then conclude with key considerations for its usage.

Challenges associated with requiring huge quantity of well-designed data

Firstly, the strength of AI/ML methods in handling enormous number of samples is also where its weakness lies: it needs them, but also of the right design to work well (Figure 1**A**, see [14], for example of design guidelines in synthetic biology that are also relevant to an AI/ML alternative to systems biology). With inadequate data, they perform poorly, but the processes and consequences of providing such data will give rise to difficult challenges, further compounded by the nature of both biological and system studies, as outlined in the rest of the article. The stringent

data requirement has played an important part in their uneven applications across biomedical fields, in line with the available amount of suitable data [16]. As a case in point, since its conception 4 years ago [13], there has not been any report of the usage of AI/ML to predict the temporal concentration of interacting biomolecules solely based on their data. This is despite the clear utility of such a technique for understanding complex system behaviors, as well as, for enhancing the production titer/yield in synthetic biology applications. The lack of application is likely to be due to the huge amount of data required, which we shall outline in the context of the 'Design-Build-Test-Learn' (DBTL) cycle. The latter is an iterative improvement framework, used for both refining the understanding of the biological phenomenon under study [17] and for attaining specified design objective in synthetic biology [18]. As its name implies, the cycle is named after the sequential steps involved.

For illustration, we take the scenario of enhancing the production of a metabolite via microbial strain design, whereby the synthetic pathway involved consists of nine enzymatic steps, each with three possible levels of expression (different promoter strengths). As a result, there is a total combination of 19 683 (3^9) candidate strains, a number which is considered too large for exhaustive testing. To guide the strain design more effectively, the DBTL cycle emphasizes the need to learn from each cycle (step 'L'), the candidate strains (i.e. combination of expression levels) that could produce higher output, and, thus, should be designed ('D'), built ('B') and tested ('T') in the next cycle. To do so, the learning process requires the building of a predictive model embodying the relationship between the input and output variables, which may be achieved in an automated fashion via ML [19]. Through such a systematic and informed approach for directing strain design, the framework aims to reduce the overall number of tests required. A minimum of five DBTL cycles is recommended [14], but this will still require 960 strains to be investigated (<5% of all possibilities). The number of tested strains consists of three 96-well plate instances (strains) for the first cycle, and two 96-instances for the second cycle, followed by one each for the last three cycles. Another two 96-instances are kept in reserve, in anticipation of a high rate of design failures, such as those arising from the toxicity of the pathway metabolites and cellular stresses due to the presence of exogenous plasmids [20]. Triplicates are assumed with half-hourly samples over 2 days, resulting in a total of 276 480 time-series data points (10×96-well plate instances×3 replicates×2 days×24 h×2 half-hourly timepoints) for each metabolite and enzyme under study, which is considered a huge sample number for biological studies. Note a similar sample size is required, if an AI/ML alternative to systems biology is to be used for more fundamental studies, such as in predicting dynamic and emergent cellular behaviors.

Although transfer learning can greatly reduce the data requirement by learning from prior data on different but related systems, the low temporal resolution of available data is likely to have limited such opportunity till now. This, in part, is attributed to a different research focus other than dynamic modeling, a lack of understanding of the resolution required, and exercising prudence by minimizing the cost of data generation. The challenge of insufficient data is further aggravated by the fact that the amount required is unknown for new systems [14], and it is unclear if it is knowable in advance. Currently, the practical and sensible way to gauge the difficulty of a new learning task is to observe the ease of improvement in the general ability of AI/ML models to predict the test dataset, with increasing sample size [14]. However, it is also important and suffice to be aware of the existence of relationships in nature that may be well defined mathematically

but are so complex, data requirement-wise, that no amount is sufficient to learn their form [21, 22]. In this regard, the inclusion of a mathematical description of the phenomenon grounded in first principles, if available, may be more desirable and useful by providing a qualitative structure to the predictive model, thereby markedly reducing data requirement.

Because of the various reasons mentioned above, the costs and timeline of a new project, and even their uncertainty, may be hard to estimate. This presents significant risk in correctly assessing its feasibility, as well as its amenability to be completed on budget and on time. Despite considerable advancements in the last decade [23–27], the various phases of the DBTL cycle can still benefit from exigently needed innovations to improve their automation, throughput, reliability and cost (Figure 1**A**), to mitigate these challenges. Compared to systems biology approaches, AI/ML methods are also less suitable for analyzing system failures, given its stringent demand on data quality and quantity.

Pernicious data leakage via preprocessing, temporal and technical artifacts

As AI/ML algorithms are designed to build high-scoring models, they are also naturally inclined to leverage on any correlation present in the data, some of which may be irrelevant to the fundamental working of the system being studied. These circumstantial associations can be technical in nature (temperature, reagent batches, technicians and machine performance), due to data pre-processing [28], or even come from irrelevant biological relationships (homology [16], co-regulation, etc.) and noise for a purely data-driven approach. Moreover, the sharing of replicates and fundamentally similar samples (which may not be obvious) among training and validation/test datasets may result in their superficially easier prediction [14]. Such false/artificial relationships are even more easily picked up with inappropriate training processes [29–32], but the high performance of the resulting models could not be reproduced during application. The phenomenon is termed 'data leakage', as it ascribes the cause to such misleading information that is 'leaked across' (present in both) training and validation datasets. As a result, such associations uncovered during the training process may be 'validated' by the high score achieved in predicting the same associations in the validation dataset. In all, data leakage can be defined as the phenomenon whereby circumstantial and artifactual correlations, irrelevant to the fundamental working of the system being studied, are leveraged by AI/ML to build high-scoring models. To underscore the deep-seated nature of the problem, data leakage may be said to be in the 'DNA' of AI/ML algorithms, as they are designed to build high-scoring models. In this light, the 'random' sorting of samples into training and validation datasets will not prevent the exploitation of irrelevant associations for building high-scoring models [33]. The effect may also be compounded with the variables under study to exaggerate the latter's impact in a hidden manner. In worse case scenarios, there is no real learning of the fundamental relationships under study but is simply reflecting the arbitrary input–output correlation in the data (i.e. overfitting or 'memorizing'). Conversely, realistic models may be confounded by the external factors and thus become disregarded inconspicuously. Despite the seriousness of the problem, there are no straightforward way to exhaustively detect and identify all the factors before model deployment.

While a sound experimental design can ensure the equal exposure of studied conditions to external factors (differences in operator, cell culture media preparation, etc.) [34] that can be corrected for eventually, it is however challenging to do so for all factors, due to the hidden nature of some of them, as well as experimental

Figure 1. Challenges associated with AI/ML alternative to systems biology for fundamental understanding of complex systems as well as synthetic biology applications. Three categories of issues (**A**–**C**) compound on each other from the perspective of 'Design-Build-Test-Learn' improvement cycle to make an AI/ML alternative to systems biology solely dependent on data challenging. Note that data leakage (**B**) may happen via the cumulative effects of multiple external factors. These factors may affect the same or different samples but is depicted in one sample for illustration purpose. (**B**) Top: data generation artifacts may together confound the effect of variables driving the phenomena. Although the resulting overestimation of effect is depicted, the converse is also possible. (**B**) Bottom: a high scoring but spurious model can ensue from combination of issues related to inappropriate sample selection, data pre-processing and model training process. Although the depictions may appear to be exaggerated and unlikely, poor data generation and usage strategy will increasingly elevate the chance of producing compromised models, as more external factors come into play and augment the effect of each other in a subtle manner. As AI/ML algorithms are designed to build high scoring models, they are also inclined to leverage

and resource constraints. Even without such limitations, it is still not possible to totally eliminate all known effects, as it is difficult and tedious to find the context-specific, optimal balance between their removal and preservation of biological signals [35]. As such, the results of batch effect correction can vary greatly in a poorly understood manner, depending on the algorithm (assumptions) and parameters used, and the order of the batches being merged for correction. There is also the risk of random and biological variabilities being suppressed during the process, thus exaggerating the confidence interval of biological effects [34]. In addition, the computational intractability of correcting massive datasets will become more of a hindrance to an AI/ML method that is solely dependent on data [36].

We anticipate data leakage to arise more perniciously in an AI/ML alternative to systems biology fully dependent on data. This is because the cost of samples and data preparation will, to certain degree, incentivizes the minimal generation of time-series data points, and the widespread application of data augmentation and curve smoothening techniques to synthesize replacement data [13]. However, excessive and ill-considered application for complex dynamics can result in simplistic, false relationships between the input candidates and the output, which are picked up during training, and 'validated' in validation/test datasets with similar artificial profiles (Figure 1**B**). An erroneously high model score can thus arise, which is otherwise not indicative of its predictive performance. Notably, such a practice has been further fed by a lack of understanding of the temporal resolution required for profiling dynamical systems. In addition, with increasing time, time-series measurements are more likely to be affected by technical artifacts [37] and rare chance events, unknowingly and uncontrollably by the researchers, to bias the learning process. The sheer sample size required by AI/ML will further aggravate the challenge [38]. Although AI/ML models may be trained to learn such temporal effects, it may still not be possible to fully account for them [28].

Currently, the over-reliance on using a sole metric, such as accuracy, and narrow approach, such as cross validation [38], for judging potential predictive capability is another cause for concern, as it provides ample room for AI/ML to exploit artifacts for attaining high training score, and further accentuated by researchers' search for such models, via repeated trial-and-error attempts. In this way, any AI/ML approach that is solely data dependent will be predisposed to pick up biased models, which is thus a more serious and prevalent problem compared to experimental and systems biology studies. In all, the difficulty in managing multi-faceted data artifacts will pose a serious challenge [38] to the aspiration of replacing systems biology with an AI/ML alternative that is fully reliant on data.

Unsuitability of AI/ML models solely reliant on data for systems biology endeavors

Another major challenge is that, without mechanistic underpinning, AI/ML generally could not be expected to be able to extrapolate out of data context [14] (Figure 1**C**), especially if the underlying mechanism is complex and exhibits contextual emergent behaviors. Also, without the fundamental constraints, the ability to predict is also more easily impaired by technical and preprocessing artifacts, as well as the hallmark of biological systems,

which is heterogeneity, whereby differing mechanism(s) at various regulatory levels, e.g. transcriptomics, proteomics, metabolomics, etc., often result in similar outcomes [39].

We also believe that an AI/ML approach solely reliant on data is unsuitable for the predictive modeling of chaotic biological systems and phenomena (e.g. yeast glycolysis [40], mitochondria metabolism [41], synthetic biochemical network [42], cell cycle transitions [43, 44], cyclic AMP signaling [45], etc.), since some imperceptible error of measuring the initial state will result in its nonlinear propagation, and hence, diverging predictions. Although the limit for predicting the chaotic dynamics of simulated physical systems has been recently extended with the usage of more advanced AI/ML algorithm [46], it should also be noted that the said prediction was not restricted by the quantity and quality of the data, an idealistic condition found under simulation, but not in biological experiments that are further plagued by heterogeneity. Importantly, the reported achievement of accurately predicting dynamics up to 8 'Lyapunov time' (time required for the divergence in prediction to be considered large by some measure of separation) still translates to under an hour for biomolecular systems (8×4.6 min) [47], which is still not sufficiently long to be useful for systems biology endeavors, even if it can be similarly accomplished. Nonetheless, we believe the boundary of the AI/ML predictions should continue to be 'pushed', but a more accurate and reliable approach may benefit from the incorporation of mechanistic elements into the model [48, 49].

AI/ML is similarly inappropriate for stochastic processes (e.g. at the single-cell level), since the actual outcome of random events cannot be foretold by any model. While 'training' can arguably still be conducted for chaotic and stochastic systems to achieve the more limited scope of inferring an explanatory model, it is unclear if such models can minimally recover the qualitative dynamics, given the issues of data quantity, quality and heterogeneity.

In addition, the lack of mechanistic interpretability for AI/ML models fully reliant on data continues to be hotly debated and requires further scrutiny. Although arguments have been put forth to downplay its relevance if the paramount purpose is to predict [13], we believe the ability to justify predictions based on mechanism is still the key to detecting model quality issues (e.g. data leakage), and for researchers to trust the model [50]. More fundamentally, hypothesizing of mechanisms, and hence inference of new biology, is a core endeavor of systems biology, and the capacity must be preserved by any compelling alternatives. Furthermore, without rudimentary knowledge of the activation and inhibition network topology, bioengineers cannot easily analyze and understand the stability of the system under design [51].

Concluding remarks

In recent years, we observe the rise of AI/ML modeling as an alternative to systems biology for elucidating the behaviors of biological systems. There are, however, multifaceted interlocking challenges as laid out in this article. In this light, a systems biology approach can still be used for small systems (order of tens of reactions) if the governing mechanistic rate laws are available, with good predictive performances. In this case, AI/ML methods

---

maximally on circumstantial inflating factors to generate overly optimistic models more prevalently than perceived. Another possibly negative but unobvious outcomes are the undermining of realistic models by the external factors; these are not unreported due to their poor training score. Cross: measured data; open circle: augmented data; continuous line through both crosses and open circles: fitted curve; continuous line through crosses only: the underlying phenomenon.

fully reliant on data will provide no clear benefits for the high tradeoffs in costs and risks, compared to the efforts involved in curating and reconstructing rate laws.

However, for small systems whose biochemical mechanisms are either unavailable or the models are not working well, one option we believe is to treat the broad families of rate laws (generalized mass actions, Michaelis–Menten, Hill-type cooperativity, etc.) or heuristic ones as candidate models to be similarly selected and trained using AI/ML, such as automated ML [19], alike conventional black-box models. This has the significant advantage of grounding the modeled system in more realistic mechanisms, thereby (1) providing interpretation and justification, as well as (2) the hypothesizing of obscure biochemical mechanisms and new biology. Such an approach will also (3) provide meaningful mechanistic constraints that will greatly reduce the required data for training, and thus the challenges associated with its provision. In addition, a hybrid approach (4) may enhance the accuracy and robustness of prediction due to the realism of the system model, compared to its black-box counterparts. As a technique for learning context-specific rate laws and parameter values as part of the training process, the method will also help advance the core endeavors of biochemistry and systems biology fields. In contrast, most other science-informed AI/ML approaches to date still necessitate the fundamental laws to be specifically provided [48, 49, 52]. Although we are currently working on such an implementation, this area is generally under-studied and requires further progress.

One caveat to the hybrid approach is its limited applicability to small systems; as the number of rate laws to be hypothesized increases, there will be more potential combinations of candidate rate laws and parameter values that can collectively fit the data relatively well. As such, due to the higher degree of freedom, there may be more uncertainty in the inferred laws. Consequently, for a large system with mostly unavailable rate laws, the hybrid approach may not be superior to black-box AI/ML methods in its interpretability (in terms of mechanism), and prediction accuracy and robustness. Instead, conventional black-box AI/ML may be preferred in practice, because of the greater availability of their software packages. The approach is however unsatisfactory, given the exponentially compounding effects of our discussed challenges, as more models for reactions are being trained, requiring more data. Also, kinetic simulation will still be increasingly intractable with growing system size, regardless of the approach used for building the required rate laws (i.e. systems biology, black-box AI/ML or hybrid). For large genome-scale metabolic systems, one computationally efficient solution is to use flux balance analysis (FBA) sequentially for each time point (i.e. dynamic FBA), assuming pseudo steady state for each simulation [53]. Nevertheless, the AI/ML parameters to be trained for such models remain to be studied.

On a positive note, there have been growing efforts to integrate/combine systems biology approaches with AI/ML methods, due to their specific synergies in various situations [54–56]. Nevertheless, systems biology is still likely to remain relevant to the biological and biotechnological domains for reasons given above.

Key Points

- We discuss the challenges in using AI/ML models, solely dependent on data, as an alternative to systems biology. These are associated with the required quantity and quality of data, pervasive data leakage, and the irreplaceability of a deep understanding of fundamental laws.
- Systems biology remains relevant to the elucidation of complex system behavior as well as the realization of design objective in many contexts.
- A hybrid AI/ML and systems biology approach could reduce the data-related challenges and enhance biological predictability in the future.

Acknowledgements

Funding

Data availability

There are no new data associated with this article.

References

1. Lopez R, Gayoso A, Yosef N. Enhancing scientific discoveries in molecular biology with deep generative models. *Mol Syst Biol* 2020;**16**:e9198.
2. Mirza B, Wang W, Wang J, *et al.* Machine learning and integrative analysis of biomedical big data. *Genes (Basel)* 2019;**10**:87.
3. Chen Y, Guenther JM, Gin JW, *et al.* Automated "cells-to-peptides" sample preparation workflow for high-throughput, quantitative proteomic assays of microbes. *J Proteome Res* 2019;**18**:3752–61.
4. Fuhrer T, Zamboni N. High-throughput discovery metabolomics. *Curr Opin Biotechnol* 2015;**31**:73–8.
5. Kitano H. Systems biology: a brief overview. *Science* 2002;**295**: 1662–4.
6. Torregrosa G, Garcia-Ojalvo J. Mechanistic models of cell-fate transitions from single-cell data. *Curr Opin Syst Biol* 2021;**26**: 79–86.
7. Abernathy MH, He L, Tang YJ. Channeling in native microbial pathways: implications and challenges for metabolic engineering. *Biotechnol Adv* 2017;**35**:805–14.
8. Daran-Lapujade P, Rossell S, van Gulik WM, *et al.* The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels. *Proc Natl Acad Sci USA* 2007;**104**:15753–8.
9. Hackett SR, Zanotelli VR, Xu W, *et al.* Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science* 2016; **354**:aaf2786.
10. Kyriakopoulos S, Ang KS, Lakshmanan M, *et al.* Kinetic modeling of mammalian cell culture bioprocessing: the quest to advance biomanufacturing. *Biotechnol J* 2018;**13**:e1700229.
11. Costa RS, Machado D, Rocha I, *et al.* Hybrid dynamic modeling of Escherichia coli central metabolic network combining Michaelis-Menten and approximate kinetic equations. *Biosystems* 2010;**100**:150–7.
12. Helmy M, Smith D, Selvarajoo K. Systems biology approaches integrated with artificial intelligence for optimized metabolic engineering. *Metab Eng Commun* 2020;**11**:e00149.

13. Costello Z, Martin HG. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Syst Biol Appl* 2018;**4**:19.

14. Radivojević T, Costello Z, Workman K, *et al.* A machine learning Automated Recommendation Tool for synthetic biology. *Nat Commun* 2020;**11**:4879.

15. Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;**15**:20170387.

16. Greener JG, Kandathil SM, Moffat L, *et al.* A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 2022;**23**:40–55.

17. Coutant A, Roper K, Trejo-Banos D, *et al.* Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proc Natl Acad Sci USA* 2019;**116**:18142–7.

18. Nielsen J, Keasling JD. Engineering cellular metabolism. *Cell* 2016;**164**:1185–97.

19. Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* 2019;**36**:250–6.

20. Opgenorth P, Costello Z, Okada T, *et al.* Lessons from two Design-Build-Test-Learn cycles of dodecanol production in Escherichia coli aided by machine learning. *ACS Synth Biol* 2019;**8**:1337–51.

21. Linial N, Mansour Y, Rivest RL. Results on learnability and the Vapnik-Chervonenkis dimension. *Inform Comput* 1991;**90**:33–49.

22. Blumer A, Ehrenfeucht A, Haussler D, *et al.* Learnability and the Vapnik-Chervonenkis dimension. *J ACM* 1989;**36**:929–65.

23. Carbonell P, Jervis AJ, Robinson CJ, *et al.* An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Commun Biol* 2018;**1**:66.

24. Janjic A, Wange LE, Bagnoli JW, *et al.* Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biol* 2022;**23**:88.

25. Jervis AJ, Carbonell P, Taylor S, *et al.* SelProm: a queryable and predictive expression vector selection tool for Escherichia coli. *ACS Synth Biol* 2019;**8**:1478–83.

26. Kim HK, Min S, Song M, *et al.* Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat Biotechnol* 2018;**36**:239–41.

27. Mellor J, Grigoras I, Carbonell P, *et al.* Semisupervised gaussian process for automated enzyme search. *ACS Synth Biol* 2016;**5**:518–28.

28. Kaufman S, Rosset S, Perlich C. *Leakage in data mining: formulation, detection, and avoidance. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* San Diego, California, USA: Association for Computing Machinery, 2011, 556–63.

29. Ding J, Tarokh V, Yang YJISPM. Model selection techniques: an overview. 2018;**35**:16–34.

30. Ghosh S, Stephenson W, Nguyen TD, *et al.* Approximate cross-validation for structured models. *NeurIPS* 2020;**33**:8741–52.

31. Bates S, Hastie T, RJapa T. Cross-validation: What does it estimate and how well does it do it? 2021. arXiv:2104.00673.

32. Kohavi R. *A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2.* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, 1137–43.

33. Riley P. Three pitfalls to avoid in machine learning. *Nature* 2019;**572**:27–9.

34. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 2016;**17**:29–39.

35. Goh WWB, Yong CH, Wong L. Are batch effects still relevant in the age of big data? *Trends Biotechnol.*

36. Tran HTN, Ang KS, Chevrier M, *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**:12.

37. Leek JT, Scharpf RB, Bravo HC, *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.

38. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;**35**:498–507.

39. Gough A, Stern AM, Maier J, *et al.* Biologically relevant heterogeneity: metrics and practical insights. *SLAS Discov* 2017;**22**:213–37.

40. Nielsen K, Sorensen PG, Hynne F. Chaos in glycolysis. *J Theor Biol* 1997;**186**:303–6.

41. Kembro JM, Cortassa S, Lloyd D, *et al.* Mitochondrial chaotic dynamics: redox-energetic behavior at the edge of stability. *Sci Rep* 2018;**8**:15422.

42. Yamaguchi HQ, Ode KL, Ueda HR. A design principle for post-translational chaotic oscillators. *iScience* 2021;**24**:101946.

43. Gerard C, Goldbeter A. A skeleton model for the network of cyclin-dependent kinases driving the mammalian cell cycle. *Interface Focus* 2011;**1**:24–35.

44. Gérard C, Goldbeter A. Entrainment of the mammalian cell cycle by the circadian clock: modeling two coupled cellular rhythms. *PLoS Comput Biol* 2012;**8**:e1002516.

45. Martiel JL, Goldbeter A. Autonomous chaotic behaviour of the slime mould Dictyostelium discoideum predicted by a model for cyclic AMP signalling. *Nature* 1985;**313**:590–2.

46. Pathak J, Hunt B, Girvan M, *et al.* Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. *Phys Rev Lett* 2018;**120**:024102.

47. Gaspard P, Chaos. *Scattering and Statistical Mechanics.* Cambridge University Press, 2005.

48. Sharma N, Liu YA. A hybrid science-guided machine learning approach for modeling chemical processes: a review. 2022;**68**:e17609.

49. Yazdani A, Lu L, Raissi M, *et al.* Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLoS Comput Biol* 2020;**16**:e1007575.

50. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intell* 2020;**2**:573–84.

51. Pucccia C, Levins R. *Qualitative Modeling of Complex Systems.* Harvard University Press, 1985.

52. Lee D, Jayaraman A, Kwon JS. Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling. *PLoS Comput Biol* 2020;**16**:e1008472.

53. Karr Jonathan R, Sanghvi Jayodita C, Macklin Derek N, *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* 2012;**150**:389–401.

54. Zampieri G, Vijayakumar S, Yaneske E, *et al.* Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput Biol* 2019;**15**:e1007084.

55. Ma J, Yu MK, Fong S, *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* 2018;**15**:290–8.

56. Culley C, Vijayakumar S, Zampieri G, *et al.* A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. 2020;**117**:18869–79.