COMPUTATIONAL MODELING IN THE LARGE DATA LIMIT

LUKE RAST

Mathematical models inspired by computing and engineering analogies play an important role in the biological sciences: they provide a means of describing the function of biological systems. As such, we might expect that, as the size of the systems that we study scales up from individual circuits to large networks, and their function becomes correspondingly more complex and difficult to describe, these models should become more important. However, this has not yet happened. Instead, such computational models have mostly remained limited to small-scale, well-studied systems, while phenomenological (e.g. machine learning) models dominate for large-scale measurements. This disconnect highlights a critical gap between models that describe function, and models that can be fit to large empirical datasets. I propose that bridging this gap, incorporating understanding of function into learning models, and vice versa, represents a promising path forward, not only for understanding physiology, but also for ensuring that our learned models capture robust and predictive features of the systems that they aim to describe. Taking this viewpoint highlights immediate questions in both the realms of data collection and model fitting.

We face two fundamental challenges related to the *sample complexity of learning*: the scaling relationship between the size of the objects we aim to model and the number of datapoints required to fit such models [1]. First, as demonstrated by classical results about finite automata [2], even simple computing systems can require sample numbers that scale exponentially with the size of the system, by virtue of the fact that complex computations can be 'hidden' so that random sampling cannot elucidate them without many samples. Second, the curse of dimensionality means that, as we observe more and more features of our system, even mechanistic models, like Markov processes, become intractible, as they require sample numbers that scale as the number of states, exponentially in dimension. This often necessitates unsupervised dimensionality reduction, which risks obscuring or ignoring features that may be physiologically important [3]. Thus, fitting biological function is both intrinsically difficult, and also requires finding signal in a large number of diffuse features. Models based on computational/functional foundations promise guidance on both fronts. The classical work on automata [2] reveals that their intrinsic complexity can be managed with active learning: by choosing the correct inputs, based on current knowledge, to use as probes, we can achieve tractable sample complexity. Moreover, recent models that explicitly incorporate functional priors [4] have shown strong performance in modeling gene expression.

Some of my previous work [5] focused on developing simple computational models for neuroscience, and analyzing how to fit the underlying computations in these models. While abstract, this work points to some ideas that motivate future directions. We considered efficient coding models, which describe the neural code (i.e. the dependence between stimuli and neural activity) using optimization problems as computational models. That is to say, these models describe the neural code as the solution to an optimization problem, meaning that the function of a code is described by the objective and constraints in this optimization. We aimed to invert this modeling approach, asking what the neural code looks like it's optimized for, by solving the optimization with general objective and constraint functions, thereby finding signatures that distinguish codes with different

objectives and different constraints. In doing so, we found that the adaptation of the code to its stimulus context is a crucial signature, essentially providing multiple samples of the underlying computation. Moreover, adaptation can also be used to fully identify both objective and constraint functions in a specific way: by finding contexts that act as fixed-points of the neural adaptation. This yields a particular active learning approach to identify these optimization models by using a fixed-point iteration to find highly informative input stimulus distributions.

This motivates one direction that arises in the context of automated sample collection. Despite much promising work, efforts still continue toward constructing AI foundation models that can capture the behavior of the cell [6, 7], with the scale of available data frequently noted as a critical bottleneck [8]. This has lead to the ideas of both model-tuned and model-in-the-loop approaches approaches to data collection. The former are based around the idea that a large volume of seemingly noisy data can be more useful for model training than less data that is apparently cleaner [9]. The latter, meanwhile generally focus on greedy information collection, choosing points that are the most informative based on current knowledge [10]. A third intriguing approach relates to ideas of active sensing [11, 12]. When designing experiments that measure the response of a system to specific conditions, rather than aiming for greedy information collection, we can instead think about *controlling* the system toward informative points, that allow features to be fit with low sample complexity. For example, mean statistics can be fit with data that scales linearly in dimension. The question, then, is whether and how we can steer the our system into a state that allows the features that we care about to be fit as mean statistics. To my knowledge, this approach is fairly open, with learning of causal structure and inverse reinforcement learning representing promising areas of application.

A second direction arises in the modeling domain related to ideas of inverse and bi-level optimization problems. Recent technical advances in fitting optimization models show great promise for applications to physiological models. These include the introduction of soft-optimization objectives [13], which use duality arguments learn on the outputs of optimization problems without having to differentiate through the optimization process itself, as well as more structured models in inverse reinforcement learning [14], and inverse optimization generally [15], which have been shown to be both tractable and performant. Meta-learning and test-time-training [16, 17] have also seen advances in the machine learning literature, and touch on many of the same ideas required for functional modelling in biological science. In particular, test-time training (TTT) updates models at inference time by training them on the incoming data, using an auxiliary objective function. In other words, this approach performs adaption by solving an optimization problem. This mirrors the physiological modelling problem, which aims to describe preserved function in highly adaptive systems. If we choose to model function using optimization problems, which we fit by observing adaptation behavior, then we have an exact parallel, but inverse, of test time training. In this view, our modelling task is equivalent to learning the auxiliary objective to be used in test-time training, so TTT systems provide both a generative model and a validation system for the analyses that we want to develop. These approaches are primed for application to biological systems, with TTT providing a reliable test-bed for emerging inverse optimization approaches.

Methods of measurement in biology are continually improving, producing more detailed and complex datasets. However, the unique properties of both these datasets and the systems that they are measured from means that it is not sufficient to merely apply machine learning approaches out of the box. In fact, by scaling up the size of our data, these methods, far from enabling ML approaches, actually produce new fundamental challenges to how we fit models. If we believe that biology, unlike the subjects of other natural sciences, does anything resembling computation, then

we must take this into account when thinking about both how we generate that data and how we fit that data. Otherwise, we risk producing brittle models that don't capture the core function of the systems we study.

References

[1] M. J. Kearns and U. Vazirani, *An introduction to computational learning theory*. MIT press, 1994.

[2] D. Angluin, "Learning regular sets from queries and counterexamples," *Information and computation*, vol. 75, no. 2, pp. 87–106, 1987.

[3] T. Hamilton, B. Sparta, S. Cooley, S. D. Aragones, J. C. J. Ray, and E. J. Deeds, "A novel metric reveals previously unrecognized distortion in the analysis of scRNA-seq data," 2022.

[4] L. Hu, H. Qin, Y. Zhang, Y. Lu, P. Qiu, Z. Guo, L. Cao, W. Jiang, Y. Shen, Q. Chen, Y. Shang, T. Xia, Z. Deng, H. Zhao, X. Xu, S. Fang, Y. Li, and Y. Zhang, "RegFormer: a single-cell foundation model powered by gene regulatory hierarchies," *Nature Communications*, 2026.

[5] L. Rast and J. Drugowitsch, "Adaptation Properties Allow Identification of Optimized Neural Codes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[6] Y. H. Roohani, T. J. Hua, P.-Y. Tung, L. R. Bounds, F. B. Yu, A. Dobin, N. Teyssier, A. Adduri, A. Woodrow, B. S. Plosky, R. Mehta, B. Hsu, J. Sullivan, C. Ricci-Tam, N. Li, J. Kazaks, L. A. Gilbert, S. Konermann, P. D. Hsu, H. Goodarzi, and D. P. Burke, "Virtual Cell Challenge: Toward a Turing test for the virtual cell," *Cell*, vol. 188, 2025.

[7] L. Kovačević, T. Gaudelet, J. Opzoomer, H. Triendl, J. Whittaker, C. Uhler, L. Edwards, and J. P. Taylor-King, "No Foundations without Foundations – Why semi-mechanistic models are essential for regulatory biology," 2025.

[8] V. M. Rao, S. Zhang, B. S. Plosky, P. D. Hsu, B. Wang, J. Zou, M. Zitnik, E. J. Topol, and P. Rajpurkar, "Generalist biological artificial intelligence in modeling the language of life," *Nature Biotechnology*, 2026.

[9] L. Naef and M. Bronstein, "Black-box data: a new paradigm for biomedicine in the AI era," *Chemical Science*, vol. 17, no. 17, 2026.

[10] J. Zhang, L. Cammarata, C. Squires, T. P. Sapsis, and C. Uhler, "Active learning for optimal intervention design in causal models," *Nature Machine Intelligence*, vol. 5, no. 10, 2023.

[11] S. C.-H. Yang, D. M. Wolpert, and M. Lengyel, "Theoretical perspectives on active sensing," *Current Opinion in Behavioral Sciences*, vol. 11, 2018.

[12] E. Tse, Y. Bar-Shalom, and L. Meier, "Wide-sense adaptive dual control for nonlinear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 18, 1973.

[13] M. Blondel, F. Llinares-Lopez, R. Dadashi, L. Hussenot, and M. Geist, "Learning Energy Networks with Generalized Fenchel-Young Losses," *Advances in Neural Information Processing Systems*, vol. 35, Dec. 2022.

[14] D. Garg, S. Chakraborty, C. Cundy, J. Song, M. Geist, and S. Ermon, "IQ-Learn: Inverse soft-Q Learning for Imitation," in *Advances in Neural Information Processing Systems*, Nov. 2022.

[15] Z. Ma, Y. Liang, and D. Li, "Behavior Learning (BL): Learning Hierarchical Optimization Structures from Data," in *ICLR*, Feb. 2026.

[16] A. Bartler, A. Bühler, F. Wiewel, M. Döbler, and B. Yang, "Mt3: Meta test-time training for self-supervised test-time adaption," in *International Conference on Artificial Intelligence and Statistics*, pp. 3080–3090, PMLR, 2022.

[17] Y. Sun, X. Li, K. Dalal, C. Hsu, S. Koyejo, C. Guestrin, X. Wang, T. Hashimoto, and X. Chen, "Learning to (Learn at Test Time)," 2024.