Novel AI Algorithms for Bridging Protein and Gene Regulation Modeling in Cellular Engineering

Zachary R Fox[1], Belinda Akpa[2], Ania-Ariadna Baetica[3]

[1]Oak Ridge National Laboratory, Advanced Computing for Health
[2]Chemical and Biological Engineering, University of Tennessee Knoxville
[3]Drexel University, Department of Mechanical Engineering and Mechanics

Topic Areas: *Novel Algorithms, Trustworthy AI, Foundation Models.*
Challenge: *Interpretable AI models for mechanism discovery in natural and synthetic gene regulatory systems. This aligns with DOE's mission to uncover principles of biological function and the energy-related pathways.*
Opportunity: *Improve interpretability of AI by bridging gene regulation models and protein models. Improve AI approaches in learning symbolic functions from sparse, noisy dynamic gene regulation data.*
Timeliness: *Advances in trustworthy AI algorithms and AI knowledge representation, Improved access to data, Access to HPC resources.*

Challenge: The modeling of gene regulation is critical for understanding how to design and reprogram microbes and plants for DOE BER's objectives in energy independence and environmental resilience. We identify two key areas in which AI can improve the accuracy, generalizability, and trustworthiness of gene regulation models. First, such models have yet to benefit from recent AI/ML breakthroughs at the molecular scale, such as AlphaFold [1] for protein structure prediction and RFDiffusion [2] for protein design. These models provide rich vector representations of biomolecules that could be used to enhance dynamical models of interacting genes. Second, dynamical gene regulatory networks have been traditionally modeled using *purely domain-specific knowledge* (e.g. through sets of known interactions) or with *purely data-driven methods* (e.g. through LSTMs or neural ODEs). Traditional modeling is challenging because we often do not know *a priori* what molecules interact (i.e. inferring the interaction graph) or the complex nonlinearities that describe how they interact. On the other hand, purely data-driven methods typically lack interpretability, making them less useful for designing perturbative experiments and *de novo* synthetic gene regulatory circuits. These applications are critical to understanding and designing behavior in the rhizobiome and plant engineering application areas. Furthermore, current tools such as AI-Feynman [3], SINDy [4], and other symbolic regression systems [5, 6] are built for standard physical systems, and not well suited for the non-Gaussian stochastic dynamics common to gene regulatory networks [7]. Our position is that the integration of advanced molecular machine learning and novel biologically-informed dynamical system discovery could enable trustworthy workflows for understanding and controlling gene regulation of plants and microbes.

Opportunity: We envision two clear opportunities to address these challenges. First, to bridge molecular scale AI/ML models [1, 8, 2] for the design of synthetic gene regulatory circuits, we hypothesize that vector-embeddings extracted from these foundational neural networks can be fine-tuned to predict the relevant parameters (e.g. Hill-function coefficients and binding affinites) that are required for simulating gene regulatory networks. Such a fine-tuning is known to be computationally challenging in terms of data access (limited DNA-TF affinity measurements) and computation (large, complex models requiring HPC solutions). DOE-BER infrastructure KBase [9] will be a key information source for plant and microbe 'omic information. The Protein Databank and publicly available data from the MITOMI assay to ascertain affinity measurements for promoter sites that bind multiple TFs [10] will be crucial to fine-tuning such models. The computational needs can be addressed through exascale computing at OLCF.

Second, we envision extending the current approaches data-driven model discovery [4, 3] to stochastic dynamics which are quintessential to understanding how genes are regulated. The loss function by which these learning systems train often assume symmetric, Gaussian noise; however we have previously shown that simple Gaussian noise characteristics in the loss function can lead to spurious and misinterpreted biological consequences [7]. Thus, there is an enormous opportunity to generalize existing deep learning based model discovery approaches to include non-Gaussian dynamics. However, such non-standard loss functions and stochastic models impose costly computations, either through Finite State Projections or Gillespie Simulations [11]. These challenges require HPC solutions available at DOE leadership computing facilities. Data for such approaches include publicly available KBase data and other single-cell microbial RNA-seq data, along with potential existing collaborations for time-series fluorescence microscopy data.

Finally, these two opportunities may be brought together to build more modular workflows for BER applications in engineering microbes for bioenergy feedstock productivity and sustainability; for example the entire end-to-end methodology could backpropagate from stochastic dynamics at the system-level (desired gene regulatory activity) to the design of transcription factors. These modeling efforts could be validated through both in-distribution and out-of-distribution datasets. For example, we envision validating promoter activity curves both through existing publicly available datasets (e.g. PDB) and with novel datasets from DOE collaborators.

Timeliness: Integrating computational modeling of proteins and stochastic dynamics of gene regulatory processes with interpretable AI can advance the discovery and engineering of complex biological systems. Modern HPC systems now enable efficient, large-scale simulations of complex systems. Tools like molecular dynamics for protein modeling and stochastic simulation algorithms for gene expression such as the Gillespie algorithm have significantly improved in scalability and performance. Furthermore, the availability of high-throughput experimental data, such as single-cell RNA sequencing and time-lapse microscopy offers detailed inputs that could enhance the accuracy of the discovered gene regulatory models. Lastly, research in systems biology and synthetic biology has matured over the past two decades, highlighting the importance of integrating molecular-scale dynamics such as protein structure and dynamics with cellular processes such as gene regulation. Multi-scale, interpretable modeling is necessary to engineer such systems. By integrating machine learning frameworks and mechanistic models will bridge these large research gaps. This aligns well with the DOE's mission to uncover principles underlying biological function and energy-relevant pathways.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[2] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

[3] Nazanin Ahmadi Daryakenari, Mario De Florio, Khemraj Shukla, and George Em Karniadakis. Ai-aristotle: A physics-informed framework for systems biology gray-box identification. *PLOS Computational Biology*, 20(3):e1011916, 2024.

[4] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[5] Alejandro Carderera, Sebastian Pokutta, Christof Schütte, and Martin Weiser. Cindy: Conditional gradient-based identification of non-linear dynamics–noise-robust recovery. *arXiv preprint arXiv:2101.02630*, 2021.

[6] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

[7] Brian Munsky, Guoliang Li, Zachary R Fox, Douglas P Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, 2018.

[8] Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

[9] Adam P Arkin, Robert W Cottingham, Christopher S Henry, Nomi L Harris, Rick L Stevens, Sergei Maslov, Paramvir Dehal, Doreen Ware, Fernando Perez, Shane Canon, et al. Kbase: the united states department of energy systems biology knowledgebase. *Nature biotechnology*, 36(7):566–569, 2018.

[10] Amir Shahein, Maria López-Malo, Ivan Istomin, Evan J Olson, Shiyu Cheng, and Sebastian J Maerkl. Systematic analysis of low-affinity transcription factor binding site clusters in vitro and in vivo establishes their functional relevance. *Nature communications*, 13(1):5273, 2022.

[11] Zachary Fox and Brian Munsky. Stochasticity or noise in biochemical reactions. *arXiv preprint arXiv:1708.09264*, 2017.