Guiding Data-Driven Integrative Design of Secure Biological Systems with Artificial Intelligence Techniques

Carla M. Mann[1,3]* (cmann@anl.gov), Michael Irvin,[1,3] Rebecca Weinberg,[2] Gyorgy Babnigg,[2] Christopher Henry,[1] Dionysios A. Antonopoulos,[2] and **Arvind Ramanathan**[1]

[1]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL; [2]Biosciences Division, Argonne National Laboratory, Lemont, IL; [3]These authors contributed equally

Project Goals: The long-term goal for this Project is to validate strategies for secure microbial systems that operate in dynamic environmental conditions, thus enabling systems-level and rational biological design for a range of applications. There are several key challenges to incorporating safeguard systems at the design stage including: (1) lack of knowledge for how well safeguards operate across the broad set of environmental and physiological conditions that an organism experiences; (2) a need to integrate the safeguard with other cellular components so that it can sense and recognize specific signals from the intracellular or extracellular environment, and mediate a response; and (3) a need for rapid and reliable methods to engineer and optimize the biological components for safeguard construction and functional integration. To address these challenges, we propose to use recent advances in the fields of synthetic biology, artificial intelligence (AI), and automation, which together pave the way for a paradigm shift in our understanding of the ways that cellular function can be designed at the level of bacterial communities.

Developing organismal systems to apply rational biological design to areas of need necessitates development of mechanisms to safely contain those organisms to protect the environment and public. Rational design of biological modules (e.g., biosensors, novel enzymes for biosynthesis and/or degradation, control circuits) to enable stability in engineered microbial systems also presents challenges due to the size of potential design-search spaces. For example, optimizing a 10 amino acid span of a 300 residue-length protein has 10^{20} potential combinations embedded within a highly non-linear (and potentially sparsely sampled) space. Scaling this type of optimization problem to the level of pathways and cellular systems produces an effectively infinite search space. This requires developing AI frameworks to intelligently explore that space to reduce time and effort needed to conduct Design, Build, Test, and Learn (DBTL) cycles.

As part of this optimization process, we are developing a bacterial "self-destruct" mechanism that activates under specific environmental conditions. This mechanism relies on a gRNA in the bacterium that self-targets a genomic locus such that the gRNA has low activity in the lab, but under changing cellular conditions in response to a specific environmental condition, becomes highly active. As gRNA activity can widely vary between target locations within the same gene [1, 2], and the same location targeted under different physiological conditions can lead to gRNA activity changes orders of magnitude in size, machine learning (ML) and AI strategies are needed to elucidate rules governing CRISPR/Cas gRNA activity under changing conditions. In order to optimally identify gRNAs capable of producing desired behavior, we developed a deep-learning model, CRISPRAct, that combines a natural language model with a neural network (NN) model to predict the fold change in cell population as a proxy for gRNA activity. This model utilizes Google AI's ALBERT [3] architecture by treating the 21 bp upstream and 21 bp downstream of the PAM site as "sentences" composed of seven "words" which themselves

comprise three bases. Our model is pre-trained on representative *Escherichia coli* genomes, then fine-tuned while a regression model is built on top of the pre-trained model. The NN model uses 444 features comprising physicochemical and positional features [1] along with environmental and cellular conditions (e.g., media type and growth phase). Predictions from the models are combined through polynomial regression to predict the fold change in cell population for a given gRNA. CRISPRAct produces a Mean Absolute Error (MAE) of 0.39 and Spearman Correlation Coefficient of 66.4% at different timepoints in variable physiological conditions.

Environment-responsive systems are not only useful in biocontainment, but also lay foundations for further development of self-contained biological modules that respond to cellular signals. We are thus developing biological modules that interpret and respond to complex dynamics of intracellular and extracellular signaling interactions. To accomplish this, we combine AI, ML, and dynamical modeling approaches to capture the dynamics of our biological modules and their impact on phenotypes. Our dynamical models (DMs) encode, as first-principles, complex networks of interactions between proteins and their regulatory mechanisms that underlie functions targeted in rational biological design objectives. We leveraged an ML approach that uses feature selection to identify a testable, low-dimensional representation of intracellular signaling to guide more efficient exploration of the design search space [4]. We also developed convolutional autoencoders (CAEs) that enable efficient approximation of computationally-expensive DM simulations. The AI model treats multivariate results of DM simulations as images wherein each row of pixels represents the simulated time course of a cellular component. Using a DM with 14 observable proteins and 100 simulation timepoints, we generated 10,000 1400-pixel images (each representing distinct configurations of the initial amount and activity of proteins) which were compressed via a CAE to a 196-dimensional latent space. Hyperparameters (e.g. the dimensionality of the latent space) were optimized using Optuna [5] which achieved a minimum running Mean Squared Error (MSE) loss of 0.02. This approach provides compressed latent representations of intracellular signaling dynamics, which enables them to readily support rational biological design by guiding more efficient exploration of an otherwise sparse nonlinear design-search space. The dynamical modeling and AI approach were developed in tandem and are part of a larger modeling and optimization loop that offers a synergistic impact on our model-based design objectives.

References

1. Guo, J., et al., *Improved sgRNA design in bacteria via genome-wide activity profiling*. Nucleic Acids Res, 2018. 46(14): p. 7052-7069
2. Gutierrez, B., et al., *Genome-wide CRISPR-Cas9 screen in E. coli identifies design rules for efficient targeting*. bioRxiv, 2018: p. 308148
3. Lan, Z. et al., *ALBERT: A Lite BERT for self-supervised learning of language representations*. ICLR, 2020.
4. Irvin, M., et al. *The misleading certainty of uncertain data in biological network processes*. bioRxiv 2021.
5. Akiba, T., et al. *Optuna: A next-generation hyperparameter optimization framework*. Proc 25th ACM SIGKDD 2019.

This Project is funded by the Biological Systems Science Division's Genomic Science Program, within the U.S Department of Energy, Office of Science, Biological and Environmental Research.