Towards a self-driving lab for metabolic network discovery in biological systems

Lun Ai [*1] **Shishun Liang** [*1] **Stephen Muggleton** [2] **Geoff Baldwin** [1]

[*]Equal contribution [1]*Department of Life Sciences, Imperial College London, UK* [2]*Department of Computing, Imperial College London, UK*. Correspondence to: Lun Ai lun.ai15@imperial.ac.uk, Geoff Baldwin g.baldwin@imperial.ac.uk.

1. Introduction

Here we demonstrate the concepts required for a self-driving lab to optimise genome-scale metabolic networks in biological systems. A key application of Synthetic Biology is to genetically engineer reliable microbial systems to maximise the production of valuable compounds. This endeavour is enhanced by genome-scale metabolic network models (GEMs), which describe genome-wide mappings between genes and metabolic reactions.

To harness the knowledge in GEMs for cellular engineering, the conventional method uses flux balance analysis (FBA) [1]. FBA predicts the growth rate of an organism or the production rate of a compound by computing the metabolic fluxes in the network. However, the accuracy of FBA predictions on gene perturbations is dependent on the complexity and correctness of gene reaction associations [2]. A comprehensive and accurate understanding of gene-reaction relationships would ensure more reliable genotype-phenotype predictions for subsequent engineering.

Our work integrates abductive reasoning, active learning and high-throughput gene repression experiments to accelerate the discovery of gene-reaction relationships. We present a Design, Build, Test and Learn (DBTL) workflow called *Auto*GEM, which is built on our active abductive learning system $BMLP_{active}$ [3] and pooled screening data from CRISPR-mediated high throughput gene repression biotechnology [4]. Based on synthetic data of *Escherichia coli* (*E. coli*), we showed a 90% reduction in both experiment number and experimental resource required to learn gene-reaction mappings [3]. Our workflow presents a realistic approach for creating a self-driving lab to reliably engineer biological systems.

2. *Auto*GEM: DBTL with $BMLP_{active}$

2.1 Abduction

Abduction is the process of learning the most suitable hypothesis to explain disagreements between data and background knowledge. Inductive Logic Programming (ILP) [5] offers an automated approach to abduction, representing data, hypotheses and background knowledge through verifiable rule-like logic programs. Candidate hypotheses that might explain the contradictions between data and background knowledge are assumed to hold and then verified against additional data.

$BMLP_{active}$ uses this ILP approach and performs abduction by generating pseudo labels, which are *predictions for unlabelled data by a model that we want to train*. We predict binary pseudo labels of the producibility of essential metabolites to logically infer errors in gene-reaction associations. Based on the GEM model for *E. coli* strain MG1655, iML1515 [6], we showed $BMLP_{active}$ can produce accurate pseudo labels [3] of single gene perturbations.

2.2 Active learning

Owing to the complexity of GEMs, comprehensively perturbing all gene-reaction mappings requires enormous experimental data. We use active learning to identify the most critical gene repression experiments to perform rather than passively performing all experiments or randomly selecting experiments, thereby reducing the number of experiments and resources needed.

Our approach exploits the GEM model iML1515 [6], by using it as background knowledge. For every gene-reaction mapping hypothesis, we use the GEM model to predict a pseudo label per experiment, which includes a set of growth medium nutrients and perturbed genes. A label informs us about the effect of introducing or deleting a gene-reaction association. $BMLP_{active}$ selects experiments by approximating the minimal expected value of a user-defined cost function based on the pseudo labels, and it iteratively updates the posterior probability over the hypothesis space. $BMLP_{active}$ can actively learn single-gene and digenic functions with only 20 synthetic gene perturbation data [3].

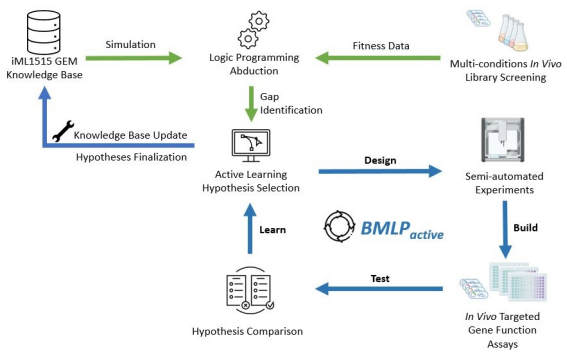

Fig. 1: DBTL with actively abductive learning, targeted multi-gene repressions and pair-wise library screening experiments.

2.3 AutoGEM

Our novel workflow *Auto*GEM is shown in Figure 1. It uses the abductive learning system $BMLP_{active}$ [3] with the input GEM model iML1515. First, we identify error sources and genes of interest by contrasting in-silico predictions and multi-condition in vivo library pooled screening data. Our screening data supplies a diverse pool of genotypes for $BMLP_{active}$ to hypothesise new gene-reaction mappings for genes of interest and predict the effect of updating the GEM.

$BMLP_{active}$ can efficiently navigate exponentially growing spaces of experimental designs and gene-reaction mapping hypotheses by actively proposing multi-gene repression trials. These targeted (as opposed to genome-wide) trials are performed experimentally, leveraging laboratory automation, thus closing the loop in the DBTL design cycle. After iterations, the hypothesis that agrees with data from all proposed experiments is added to the GEM.

$BMLP_{active}$ is the first ILP system applied to GEMs due to the improved computational efficiency from leveraging boolean matrices [7]. *Auto*GEM applies this advance to create a self-driving lab for biological discovery in the engineering of biological systems.

3. Related work

FBA [1] uses linear programming to compute fluxes in a biochemical reaction network with a matrix that contains the stoichiometric coefficients of metabolites in reactions. Ensemble machin learning approaches [8, 9, 10] can learn support vector machines, decision trees and artificial neural networks to represent constraints between genetic factors and metabolic fluxes. Recent research [11] also explores autoencoders to learn the relationship between gene expression data and metabolic fluxes from synthetic data. However, the availability of experimental data [12] is a significant limitation for learning GEMs. These systems do not identify inconsistencies between simulations and experimental data to initialise learning or guide experimentation [13].

On the other hand, mechanistic information that reflects whole-cell dynamics has been incorporated, for instance in artificial neural networks to tune their parameters [14]. The hybrid approach in [15] embeds FBA within artificial neural networks based on custom loss functions surrogating the FBA constraints. In contrast to the above, which all rely on the GEM model as the fixed metabolic ground truth, $BMLP_{active}$ uses GEM as mechanistic background knowledge and can actively guide experimentation to efficiently improve the GEM model.

In addition, our approach differs from other DBTL platforms [16, 17] since we can produce verifiable and human-interpretable hypotheses, avoiding the need for biologists to make sense of black-box AI models. Notably, the GEM iML1515 has 1515 genes and 2719 reactions, which is remarkably larger than the metabolite network model used in the related Robot Scientist project [18] . The incorporation of genome-wide CRISPR screens significantly advances our ability to test whole genome and multi-gene associations.

Acknowledgments

The first, second and fourth authors acknowledge support from the UKRI 21EBTA: EB-AI Consortium for Bioengineered Cells & Systems (AI-4-EB) award (BB/W013770/1). The second author acknowledges support from the UK's EPSRC Human-Like Computing Network (EP/R022291/1), for which he acted as Principal Investigator. ChatGPT and Grammarly have been used as language editing tools after all intellectual content has been drafted.

References

[1] B. O. Palsson. *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, Cambridge, 2015.

[2] David B Bernstein et al. Evaluating E. coli genome-scale metabolic model accuracy with high-throughput mutant fitness data. *Molecular Systems Biology*, 19(12), 2023.

[3] Lun Ai et al. Boolean matrix logic programming for active learning of gene functions in genome-scale metabolic network models. *arXiv*, 2024.

[4] Tianmin Wang et al. Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nature Communications*, 9(1):2475, 2018.

[5] S. H. Muggleton. Inductive logic programming. *New Generation Computing*, 8:295–318, 1991.

[6] J. Monk et al. iML1515, a knowledgebase that computes escherichia coli traits. *Nature Biotechnology*, 35:904–908, 10 2017.

[7] Lun Ai and Stephen H. Muggleton. Boolean Matrix Logic Programming. *arXiv*, 2024.

[8] Stephen Gang Wu et al. Rapid Prediction of Bacterial Heterotrophic Fluxomics Using Machine Learning and Constraint Programming. *PLOS Computational Biology*, 12(4), 2016.

[9] Tolutola Oyetunde et al. Machine learning framework for assessment of microbial factory performance. *PLOS ONE*, 14(1):e0210558, 2019.

[10] David Heckmann et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications*, 9(1):5252, 2018.

[11] Ankur Sahu et al. Advances in flux balance analysis by integrating machine learning and

mechanism-based models. *Computational and Structural Biotechnology Journal*, 19:4626–4640, 2021.

[12] Partho Sen et al. Deep learning meets metabolomics: a methodological perspective. *Briefings in Bioinformatics*, 22(2):1531–1542, 2021.

[13] Pratip Rana et al. Recent advances on constraint-based models by integrating machine learning. *Current Opinion in Biotechnology*, 64:85–91, 2020.

[14] Jason H. Yang et al. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell*, 177(6):1649–1661, 2019.

[15] Léon Faure et al. A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models. *Nature Communications*, 14(1):4669, 2023.

[16] Pablo Carbonell et al. An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Communications Biology*, 1(1):1–10, 2018.

[17] Mohammad HamediRad et al. Towards a fully automated algorithm driven platform for biosystems design. *Nature Communications*, 10(1):5150, 2019.

[18] R. D. King et al. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004.