Human Comprehensible Active Learning of Genome-Scale Metabolic Networks

Lun Ai[1], Shi-Shun Liang [2], Wang-Zhou Dai[3], Liam Hallett[2], Stephen H. Muggleton[1], Geoff S. Baldwin[2]

[1]Department of Computing, [2]Department of Life Science, Imperial College London, UK.
[3]School of Intelligence Science and Technology, Nanjing University, China.
{lun.ai15,shishun.liang20,l.hallett19, s.muggleton, g.baldwin}@imperial.ac.uk, daiwz@nju.edu.cn

A key application of Synthetic Biology is the engineering of organisms to produce useful compounds. This typically requires heterologous biosynthetic pathways to be introduced into the production chassis of choice. The product yield is dependent on both the efficiency of the heterologous system and the ability of the host metabolic network to produce the correct precursor in a high enough amount. Engineering the host metabolic network to improve production is the basis of metabolic engineering, but it remains a significant challenge due to the complexity of the biological system. Efficient computational approaches to both learn and navigate the biological design space would greatly enhance our ability to predictably engineer biological systems.

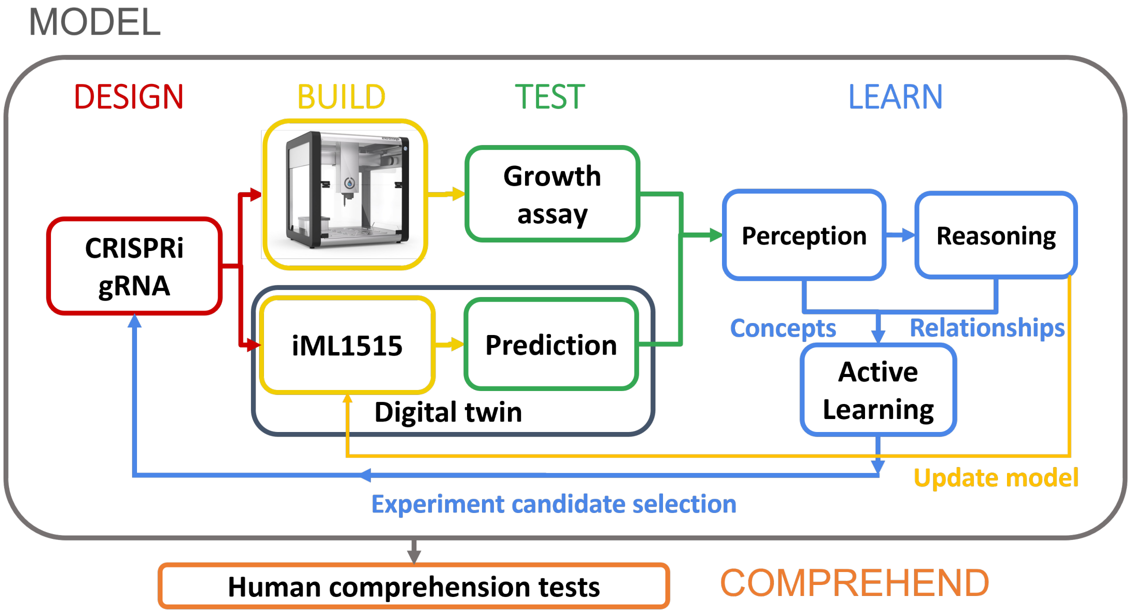

Figure 1: Model-Comprehend framework

The Robot Scientist (King *et al.* 2004) demonstrated that abductive learning could be used to learn metabolic networks faster and at less cost than random experiments. This demonstration was based on a subset of only 17 genes in aromatic amino acid metabolism. Here we expand this approach to include all 1515 genes included in the most complete E. coli genome-scale metabolic model

(GEM) (Monk *et al.* 2017). Through optimisation we demonstrate a 4000-fold improvement in computational time, making whole-genome hypothesis generation and testing machine accessible. We propose a novel bio-synthetic designer framework called Model-Comprehend which is empowered by Abductive Logical Reasoning and Active Learning together with CRISPRi technology to efficiently test theoretical predictions by experiment (Figure 1).

Gene repression is a common approach in metabolic engineering, which can force metabolite flux towards the target product, increasing yield. Predicting the phenotypic effects of gene repression is challenging due to complex interactions between genes, proteins, and reactions within cells. Models that can accurately predict the phenotypic effects of multiple gene perturbations would greatly accelerate the development of cell factories.

As one of the most extensively characterized bacteria, E. coli is the foundation of synthetic biology research and a large number of genetic tools and models have been developed for it. The iML1515 GEM was developed with the objective of being able to predict metabolic phenotypes under different conditions and with different genetic perturbations. However, even though iML1515 is incomplete with respect to core gene functions it has been demonstrated to contain errors which affect the accuracy of predictions and the reliability of the derived knowledge. A key constraint is being able to learn from experimental data and improve model accuracy based on experimental observation. The size of the design space for exploration is challenging both from the computational as well as experimental standpoint.

Our framework uses Abductive Logic Reasoning (Muggleton and Bryant 2000) with Active Learning (King *et al.* 2004) and by applying matrix encoding and parallelisation, we improve the computation time 4000-fold, making genome-scale metabolic learning a viable task. We demonstrate this by deliberately removing key nodes from the iML1515 model and asking the Model-Comprehend framework to rebuild it based on experimental observation. The utility of Active Learning is also benchmarked against random and naïve approaches. Future work aims to integrate this approach with an experimental workflow that will use CRISPRi to knock down specific genes to experimentally validate hypothesis generation. Importantly this framework will enable us to address multiple gene-loci in combination, which are not addressable by current high throughput mutagenesis approaches.

References

King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; and Oliver, S. G. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427:247–252.

Monk, J. M.; Lloyd, C. J.; Brunk, E.; Mih, N.; Sastry, A.; King, Z.; Takeuchi, R.; Nomura, W.; Zhang, Z.; Mori, H.; Feist, A. M.; and Palsson, B. O. 2017. iML1515, a knowledgebase that computes escherichia coli traits. *Nature Biotechnology* 35(10):904–908.

Muggleton, S. H., and Bryant, C. H. 2000. Theory completion using inverse entailment. In *Proceedings of the 10th International Workshop on Inductive Logic Programming (ILP-00)*, 130–146.