Data Factory

In developing a data factory for cell engineering, two primary observations guide our approach. First, sequencing technologies combined with controlled environmental conditions can be used to effectively estimate global cellular states. Second, targeted analytical techniques that focus on individual biomolecules are essential for practical applications in metabolic engineering, synthetic biology tool design, systems biology discovery, biocatalysis, and natural product discovery[4]. Early deep learning efforts in genotype-phenotype mapping have confirmed the deep learning hypothesis that more high-quality data can be leveraged for improved genotype-phenotype mapping, but also that certain technologies, like RNA-seq and high-throughput mass spectrometry (HTMS), benefit from high data density per sample which can reduce the number of equivalent experiments needed for predicting the same phenotype[5,6]. Applications of large language models (LLMs) have proven useful in helping predict phenotypes related to sequence data including DNA, and protein, but these models often neglect important biological relationships[7–9]. Analogous ideas have demonstrated that LLMs trained on single-cell RNA-seq datasets can be used to construct foundation models useful in phenotype prediction tasks in systems that depend on multiple different interacting biological entities, such as gene expression perturbation responses, gene regulatory network inference, and multiomics integration[10–12]. Perturbation to response mapping is largely possible due to Perturb-Seq which links a genetic library of possible perturbations to a genome-wide mRNA expression profiles but this method has been underutilized in microbial engineering[13–15]. This method can be used to assess whether genetic manipulations lead to desired cell states and enables screening combinatorial designs by effectively combining thousands of traditional benchtop experiments into a single library experiment. Gene expression is often just a correlate to some desired phenotype, which can be more precisely quantified by HTMS. To this end we plan to use high-throughput Echo MS+ System with the ZenoTOF 7600 System for more targeted phenotypes such as protein-protein interaction (PPI)identification via affinity purification mass spectrometry (AP-MS) and compound-protein interaction (CPI) via affinity selection mass spectrometry (AS-MS)[16]. When compared to sequencing, HTMS has a lower throughput per biological sample, but it is similar in that it can be used to generate of dense data in the order of hundreds of compounds detected per sample. In a natural product campaign such HTMS could be used in a data independent acquisition (DIA) manner to discover natural products that a bioactivty assay would not normally detect[17]. With this in mind, we believe that a modern data foundry should be focused on sample construction and preparation for downstream high-throughput sequencing technologies that are supported by sequencing core facilities, and the rest of the data foundry should be focused on efficient phenotype quantification primarily by mass spectrometry due to high data density, with some space dedicated to a few low-density phenotype detection techniques such as growth/inhibition assays and activity assays.

Notes:

- Purchase: RoboLector (approximatley $230,000). This integrates with the BiolectorXT which IGB already has and can be integrated with LIMS system. This will allow for time point sampling, followed by mass spec analysis. This type of data can used for developing systems like Dynamic Metabolic Control used by DMC biotechnologies. This can get around the issue you mentioned in last group meeting how compounds are often toxic during growth, but show less toxicity during max growth.
- There is no standard technique for sample preparation in proteomics. Depending on the application it is likely that a lot of effort will be necessary to setup the mass spec. I imagine switching applicaitons a lot will defeat the high-throughput goals of automation.
- At metabolomics center on campus cost per sample of DIA and DDA are around $75. We can do an entire price breakdown if necessary. This means take a prototypical example application and compute cost per biological sample and corresponding data density for sequence and metabolite data.

References

1. Zelezniak, A. *et al.* Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cell Systems* **7**, 269–283.e6 (2018).
2. Zhang, J. *et al.* Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat Commun* **11**, 4880 (2020).
3. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38**, 56–65 (2020).
4. Yuan, Y., Huang, C., Singh, N., Xun, G. & Zhao, H. Self-resistance-gene-guided, high-throughput automated genome mining of bioactive natural products from Streptomyces. *cels* **16**, (2025).

5. Ma, J. *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* **15**, 290–298 (2018).

6. Culley, C., Vijayakumar, S., Zampieri, G. & Angione, C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc Natl Acad Sci USA* **117**, 18869–18879 (2020).

7. Rosen, Y. *et al.* Universal Cell Embeddings: A Foundation Model for Cell Biology. 2023.11.28.568918 https://www.biorxiv.org/content/10.1101/2023.11.28.568918v1 (2023) doi:10.1101/2023.11.28.568918.

8. Dalla-Torre, H. *et al.* The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. 2023.01.11.523679 https://www.biorxiv.org/content/10.1101/2023.01.11.523679v1 (2023) doi:10.1101/2023.01.11.523679.

9. Karollus, A. *et al.* Species-aware DNA language models capture regulatory elements and their evolution. 2023.01.26.525670 https://www.biorxiv.org/content/10.1101/2023.01.26.525670v2 (2023) doi:10.1101/2023.01.26.525670.

10. Singh, A. H. *et al.* An Automated Scientist to Design and Optimize Microbial Strains for the Industrial Production of Small Molecules. 2023.01.03.521657 https://www.biorxiv.org/content/10.1101/2023.01.03.521657v1 (2023) doi:10.1101/2023.01.03.521657.

11. Cui, H. *et al.* scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

12. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS | Nature Biotechnology. *Nature Biotechnology* (2023) doi:10.1038/s41587-023-01905-6.

13. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).

14. Yao, D. *et al.* Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nat Biotechnol* **42**, 1282–1295 (2024).

15. Nadal-Ribelles, M., Solé, C., de Nadal, E. & Posas, F. The rise of single-cell transcriptomics in yeast. *Yeast* **41**, 158–170 (2024).

16. Stella, A., McCabe, J. W. & Bhalkikar, A. Automated, rapid method optimization and buffer screening using the Echo® MS+ system with ZenoTOF 7600 system.

17. Covington, B. C. & Seyedsayamdost, M. R. Unlocking hidden treasures: The evolution of high-throughput mass spectrometry in screening for cryptic natural products. *Nat. Prod. Rep.* (2025) doi:10.1039/D4NP00026A.