Predicting cellular phenotypes of unseen environmental compounds using flow matching models

Supervisors: Jess Ewald (EBI, primary), Mo Lotfollahi (Sanger)

Project summary: People are exposed to thousands of human-made compounds each day, most of which have never been evaluated for their potential to perturb biological systems. High-throughput, high-content imaging techniques are emerging as scalable methods to assess how compounds affect cells. However, chemical space is vast, and even with these advances, it remains logistically infeasible to test all compounds in all biological contexts.

In silico models of cellular responses offer a potential solution to this challenge. Deep generative models can be trained to learn joint representations of compound structure and cell morphology from large-scale imaging data of chemical perturbations such as JUMP-CP and EU-OPENSCREEN. These models can then generate predicted images of perturbed cells when given a molecular structure, even if that compound was not part of the training data. We developed IMPA[1], the first generative style-transfer model for in-silico prediction of morphological changes across genetic and chemical perturbations, enabling robust modeling of unseen perturbations and batch effects in high-content imaging screens. This approach was extended in PhenoDiff[2] to disentangle perturbation-specific phenotypic signals from confounders, and further advanced in CellFlux[3] to model temporal dynamics and directional transitions of cellular phenotypes across perturbation trajectories.

While effective for batch correction and data integration, the ability of these models to generalize to out-of-distribution compounds such as environmental chemicals remains uncertain. Most training datasets include drug-like molecules, which may not represent broader environmental exposures. Demonstrating generalization would be valuable for chemical risk assessment, while failure would highlight areas for future improvement. From a technical perspective, challenges remain in modeling unseen perturbations and in solving the inverse problem: identifying perturbations to produce a desired phenotype, a potentially powerful but underexplored application.

The specific aims of the proposed project are to:

1. Train a conditional generative model based on the flow matching framework on large-scale imaging datasets like JUMP-CP and EU-OPENSCREEN that link molecular structures to cellular morphology. Unlike standard flow models that begin from random noise, our approach conditions the initial noise on perturbation embeddings derived from molecular or genetic inputs. This enables the model to learn structured latent representations where biologically similar perturbations produce similar outputs, improving generalization to unseen compounds such as environmental chemicals.

2. Predict cell morphology after exposure to 1,000 pesticides from the EMBL pesticide screening library. The Ewald Lab is currently generating Cell Painting data for a subset of this library, providing a unique opportunity to assess generalization to a truly independent test set.

3. Evaluate predictions using multiple metrics. We will compare compound–compound similarity in generated vs. real Cell Painting images, assessing performance using mean average precision (mAP) in a top-N nearest-neighbor retrieval task. Additional metrics will

include Precision@K, Recall@K, nDCG, and visual fidelity measures like FID/KID adapted to cell morphology embeddings. Latent space alignment will be assessed using methods like Procrustes analysis or Centered Kernel Alignment.

4. Compare against simple baselines using only chemical structure or cytotoxicity (e.g. cell count), to quantify the added value of joint structure-morphology modeling.

By testing model generalization to a distinct and underexplored chemical space, this project will provide critical insight into the limits and capabilities of current generative models for phenotype prediction. It will also create a benchmark for evaluating *in silico* approaches in environmental toxicology and chemical risk assessment. Beyond this specific application, the framework developed here can be extended to other compound classes and cell systems, offering a scalable strategy for predicting cellular responses in other contexts, for example drug discovery.

Collaboration between labs: This project represents a new collaboration between the Lotfollahi Lab (est. 2024) at Sanger and the Ewald Lab (est. 2025) at EMBL-EBI, two groups with complementary expertise:

- The Lotfollahi Lab brings deep expertise in machine learning for biological systems. The group has developed approaches that learn across modalities such as gene expression, spatial data, and imaging.
- The Ewald Lab is focused on using high-throughput, *in vitro* screens to detect hazardous compounds in the environment, with expertise in high-content imaging. The group is currently conducting Cell Painting screens of pesticides that can be used to evaluate predictions.

Training plan: The Fellow will be co-located across the Lotfollahi Lab at the Sanger Institute and the Ewald Lab at EMBL-EBI, receiving tailored training in machine learning, including generative modeling with flow matching, multimodal learning, and graph-based molecular modeling, as well as in high-content imaging and morphological profiling using the Cell Painting assay. The project offers a unique opportunity to validate computational predictions experimentally, enabling rapid iteration and robust discovery. The Fellow will build expertise in both environmental toxicology and drug discovery, gaining experience that is broadly applicable across biomedical science. Working in a highly interdisciplinary AI and biology team, they will be well positioned to lead collaborative research projects. They will be supported in submitting work to top machine learning conferences such as ICLR, NeurIPS, and ICML, as well as to leading journals in computational biology, and will have access to leadership development opportunities. This training will prepare the Fellow for future roles at the interface of machine learning and life sciences, in either academic or industry settings.

References

1. Palma, A., Theis, F. J. & Lotfollahi, M. Predicting cell morphological responses to perturbations using generative modeling. *Nat. Commun.* **16**, 505 (2025).
2. Bourou, A. *et al.* PhenDiff: Revealing subtle phenotypes with Diffusion Models in real images. *arXiv [eess.IV]* (2023).
3. Zhang, Y. *et al.* CellFlux: Simulating cellular morphology changes via flow matching. *arXiv [q-bio.QM]* (2025).