2026 Helmholtz – OCPC – Programme

for the involvement of postdocs in bilateral collaboration projects

PART A

Title of the project:

Scaling Generative AI to Predict Cellular Perturbation Responses Across Organoids, Tissues, and Drug Screens

Helmholtz Centre and/or institute:

Helmholtz Munich

Project leader:

Prof. Dr. Dr. Fabian J. Theis

Contact Information of Project Supervisor: (Email, telephone)

fabian.theis@helmholtz-munich.de
+4989 3187 43260

Web-address:

https://www.helmholtz-munich.de/en/
https://www.helmholtz-munich.de/en/computational-health-center
https://www.helmholtz-munich.de/en/icb
https://www.helmholtz-munich.de/en/icb/research-groups/theis-lab

Department: (at the Helmholtz centre or Institute)

Computational Health Center (CHC)
Institute of Computational Biology (ICB)

Programme Coordinator (Email, telephone and telefax)

Name: Kathrin Zahr
Address: Ingolstädter Landstr. 1, 85764 Neuherberg
Phone: +49-8931873149
E-mail: kathrin.zahr@helmholtz-munich.de

Description of the project (max. 1 page)**:**

Single-cell RNA sequencing (scRNA-seq) has enabled high-resolution mapping of cellular states, revealing heterogeneity, disease mechanisms, and developmental trajectories. Large-scale perturbation screens now profile cellular responses to drugs and genetic interventions across millions of cells. Yet the combinatorial space of perturbations and contexts far exceeds experimental capacity. The key challenge is therefore to simulate cell biology: predicting responses to unseen perturbations in silico to accelerate therapeutic discovery. Computational approaches for perturbation prediction from our lab and others have progressed from autoencoder-based methods to foundation and/or optimal transport models (Klein et al, *Nature* 2025), leveraging large-scale pretraining to improve generalization across cellular contexts. Generative frameworks based on flow matching, such as our CellFlow (Klein et al., biorxiv 2025), have advanced distributional modeling of perturbation outcomes. Despite these advances, current models typically operate on a single readout modality (gene expression), a narrow perturbation type (chemical or genetic), or a limited number of datasets, leaving substantial room for improvement in generalization, scalability, and practical utility.

We propose to develop a unified large-scale perturbation model that bridges these gaps by jointly training across gene expression profiles and functional assay readouts (e.g., cell viability, morphology, maturation indices), across chemical and genetic perturbations, including combinatorial treatments, and across cell lines, primary tissues, and complex organoid systems. Our approach combines scalable flow matching with learnable perturbation embeddings that encode chemical structure and gene identity into a shared latent space, enabling knowledge transfer: the effect of an untested compound can be inferred by traversing the learned space of related perturbations.

We are uniquely positioned to realize this vision with some of the world's largest and most diverse single-cell perturbation datasets through ongoing strategic collaborations. These include:
(i) comprehensive brain, gut, and lung organoid atlases developed together with Roche's Institute of Human Biology and ETH Zurich, providing standardized single-cell references across protocols and labs (He et al, ***Nature*** 2024; Xu et al., ***Nat. Genet.*** 2025);
(ii) a Cancer Plasticity Atlas in collaboration with the Wellcome Sanger Institute and Parse Biosciences, combining organoid perturbation screens at unprecedented scale to model drug response and resistance mechanisms;
(iii) one of the largest human lung tissue perturbation atlases in collaboration with the Parse Biosciences, profiling cellular responses to 900 pharmacological interventions in ex vivo tissue slices from healthy and diseased donors.
Together, these resources encompass a comprehensive spectrum of cell states and thousands of perturbation conditions, providing an unmatched training corpus. The resulting model will not only be benchmarked against state-of-the-art methods on established datasets, but used to propose new experiments prospectively to maximize desired cellular phenotypes. By unifying diverse perturbation data under a single generative framework, this project aims to lay the computational foundation for perturbation modeling in drug discovery and personalized medicine.

Description of existing or sought Chinese collaboration partner institute (max. half page)**:**

The project seeks collaboration with leading Chinese research institutions with strong capabilities in biomedical AI, computational biology, and medical imaging. Primary partners include **Fudan University**, which has established strengths in cancer research, clinical data science, and AI-assisted diagnostics. Additional collaboration is planned with the **Biomedical Pioneering Innovation Center (BIOPIC) at Peking University**, an interdisciplinary platform integrating genomics, advanced imaging, and systems biology for data-intensive biomedical research. The project will also engage **the MOE Key Laboratory of Bioinformatics and the Bioinformatics Division of BNRIST at Tsinghua University,** which has strong expertise in computational biology, large-scale biomedical data analysis, and machine learning for life sciences. In addition, collaboration is considered with **Westlake University**, which has developed focused strengths in quantitative biology and virtual cell research, supporting data-driven modeling and simulation of cellular systems and complex biological processes.

Required qualification of the postdoc:

- Recently obtained a PhD in Computer Science, Bioinformatics, Computational Biology, Biology, or a related field.
- Strong programming skills in data science workflows (e.g. Python, GitHub) and motivation to learn and apply modern deep learning frameworks (e.g. PyTorch).
- Solid understanding of molecular biology and a strong interest in complex disease models; experience with experimental techniques is welcome but not required.
- High scientific curiosity and motivation to contribute to AI-augmented disease atlases and perturbation models, combined with excellent communication skills and the ability to collaborate in interdisciplinary teams.

Desirable qualifications:

- Experience in identifying, curating, and integrating large-scale, multimodal biological or clinical datasets.
- Familiarity with single-cell data analysis, foundation models, multi-modal integration, perturbation experiments, organoid models, or translational target discovery.
- Exposure to active learning, iterative screening, perturbation data, or target identification steps in drug development.