Image-conditioned Molecule Diffusion Generation

Maxime Sanchez[0,1], Thomas Boyer[0]

[0]Computational Bioimaging and Bioinformatics, Institut de Biologie, ENS Ulm | 45 rue d'Ulm, Paris, France.
[1]iktos | 65 rue de Prony, 75017 Paris, France.

Internship details

Level	Master 2 or equivalent
Duration	6 months, starting February - April
Location	Computational Bioimaging and Bioinformatics, IBENS, 46 rue d'Ulm
Supervisors	Thomas Boyer, Maxime Sanchez, Auguste Genovesio
Contacts	tboyer@bio.ens.psl.eu, msanchez@bio.ens.psl.eu

Context

Cell painting assays of chemically or genetically perturbed cells

Cell Painting is a high-throughput imaging assay that captures rich morphological profiles of human cells by "painting" various cellular components with fluorescent dyes. This technique allows for the systematic observation of cellular responses to different chemical or genetic perturbations.

The Broad Institute's JUMP Cell Painting Consortium has recently released extensive datasets containing millions of images of cells subjected to a wide array of chemical and genetic perturbations ([1], [2]). These datasets offer unparalleled opportunities to study the impact of these perturbations on human cells. Analyzing those images can lead to the discovery of new therapeutics by identifying molecules that induce desired phenotypic outcomes.

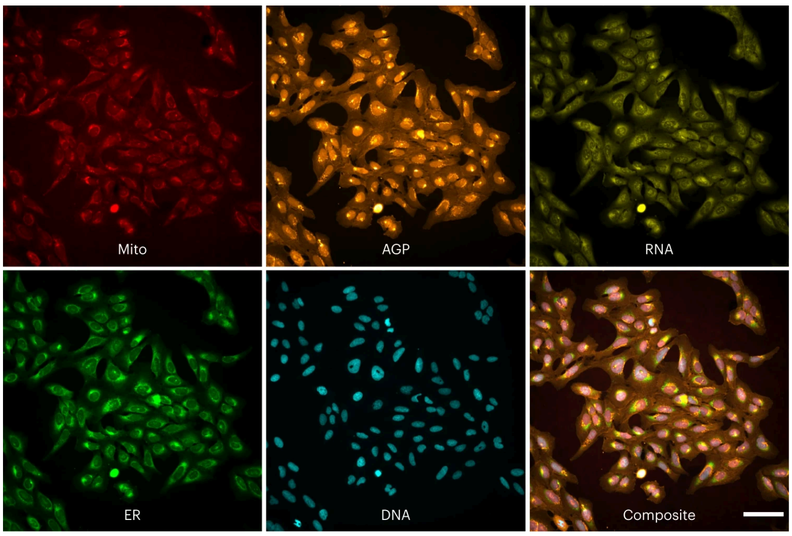

Figure 1: Image of perturbed cell culture from [1].

Conditional Molecule Diffusion Models

Diffusion models have emerged as powerful generative models in machine learning, particularly for tasks involving continuous high-dimensional data like image synthesis [3]. Recent advancements have adapted diffusion models for molecular generation, allowing for the generation *de novo* molecular structures, often with specified properties or constraints ([4], [5]).

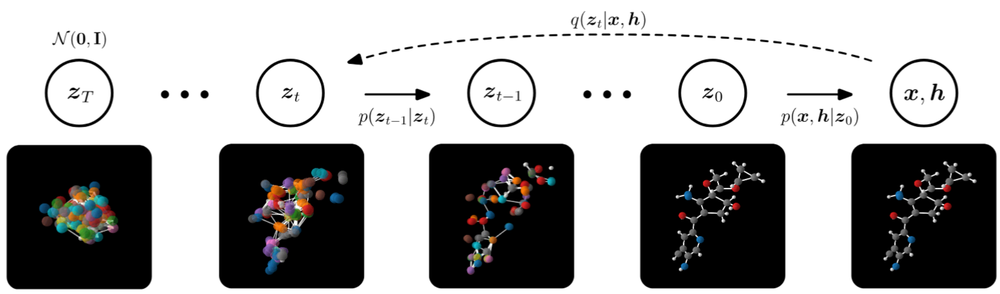

Figure 2: Figure 2 from [6]

Subject

The broad subject of the internship is to generate molecular candidates inducing an observed cellular perturbation. Leveraging recent and massive cell painting datasets released by the Broad Institute ([1], [2]), we intend to train diffusion models to generate 3D molecular representations while conditioning them on a microscopy image of cell culture.

Training an image → molecule conditional diffusion model is, to the best of our knowledge, an untackled challenge, thus very open to exploration:

- **Which image embedding to use (if any?)** We already have in-house (or potentially external: [7]) pretrained autoencoders, but since the image conditioning is of paramount importance, an extensive exploration is expected.
- **What conditioning scheme to leverage?** Cross-attention is the de-facto standard since [8], but conditioning the generation of a diffusion model on an image is not a very common task (in a non image-to-image setting), so exploration is again anticipated.
- **What specialization is needed for the generative model to output realistic molecular candidates?** The rich literature on such generative methods ([4], [5]) will give us domain-specific baselines to reuse, although we will need to adapt them them to our specific needs.
- **How to account for the diversity of 3D conformations?** Linking the compound descriptions found in the JUMP consortium datasets with the potentially *many* 3D structures described in other existing databases ([9], [10]) will be a major task.

Potential follow-ups include using [2] to link genetic and molecular perturbations, itching closer to therapeutical applications.

Pre-requisites

- Strong proficiency in Python and experience with deep learning frameworks (preferably PyTorch).
- Very solid understanding of machine learning concepts, particularly in generative modelling.
- Ability to communicate effectively in English.

Nice to have

- Experience with diffusion models.
- Experience with molecular chemistry or molecular representations.

Application Process

To apply for this internship, please submit your **resume** and a brief **motivation letter**. Do not hesitate to contact us for more details about the project or the application process.

Bibliography

[1] S. N. Chandrasekaran *et al.*, "JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations," *bioRxiv*, 2023, doi: 10.1101/2023.03.23.534023.

[2] S. N. Chandrasekaran *et al.*, "Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations," *Nature Methods*, vol. 21, no. 6, pp. 1114–1121, Jun. 2024, doi: 10.1038/s41592-024-02241-6.

[3] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *NeurIPS*, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[4] M. Xu, A. S. Powers, R. O. Dror, S. Ermon, and J. Leskovec, "Geometric Latent Diffusion Models for 3D Molecule Generation," in *ICLR*, 2023. [Online]. Available: https://proceedings.mlr.press/v202/xu23n.html

[5] A. Morehead and J. Cheng, "Geometry-complete diffusion for 3D molecule generation and optimization," *Communications Chemistry*, vol. 7, no. 1, p. 150–151, Jul. 2024, doi: 10.1038/s42004-024-01233-z.

[6] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant Diffusion for Molecule Generation in 3D," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 8867–8887. [Online]. Available: https://proceedings.mlr.press/v162/hoogeboom22a.html

[7] O. Kraus *et al.*, "Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 11757–11768.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10684–10695.

[9] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Scientific Data*, vol. 1, no. 1, p. 140022–140023, Aug. 2014, doi: 10.1038/sdata.2014.22.

[10] S. Axelrod and R. Gómez-Bombarelli, "GEOM, energy-annotated molecular conformations for property prediction and molecular generation," *Scientific Data*, vol. 9, no. 1, p. 185–186, Apr. 2022, doi: 10.1038/s41597-022-01288-4.