Flow-based model for generative model for small molecules
Mentor: Dr. Halima Bensmail
Co-mentor:
Dr. Michael Aupetit

Flow-based generative models have been employed for Boltzmann sampling tasks, but their application to high-dimensional systems is hindered by the significant computational cost of obtaining the Jacobian of the flow.
Normalizing flows enable exact likelihood computation through invertible transformations, yet global coupling structures or rigid architectural assumptions frequently constrain their expressive capacity.
This work introduces a novel class of piecewise-affine, non-volume-preserving transformations defined by half-space conditioned mappings that operate locally while preserving tractable inverses and explicit Jacobian determinants. Each transformation applies affine scaling and shearing along a data-dependent hyperplane, yielding a flexible geometric primitive amenable to composition without sacrificing analytical tractability.
Building upon this construction, we propose a deep neural mixture architecture that composes these transformations hierarchically while preserving exact likelihood evaluation. The resulting model admits closed-form inverses, explicit Jacobian determinants, and stable layer-wise training, thereby enabling expressive density estimation in high-dimensional settings without recourse to global autoregressive or residual parameterizations.
Unlike the Hutchinson estimator, our approach is inherently unbiased in Boltzmann sampling. Notably, this method significantly accelerates Boltzmann sampling of a Chignolin mutant with all atomic Cartesian coordinates explicitly represented, while delivering more accurate results than the Hutchinson estimator.

This project focuses on using **flow-based generative models** to generate and sample **small molecules**, which is an important and rapidly growing area in drug discovery. Traditional approaches such as normalizing flows allow us to model complex data distributions and compute exact probabilities, but they often become **computationally expensive** and **less flexible** when applied to high-dimensional data, such as molecular structures.

In this work, we introduce a new type of transformation that is **simple, efficient, and flexible**. Instead of relying on global transformations, our method applies **local, piecewise-affine transformations** that operate on different regions of the data. These transformations are designed to remain **mathematically tractable**, meaning we can still compute probabilities and inverses exactly, which is essential for reliable sampling. By combining many of these transformations in a deep neural network, we build a model that can **learn complex molecular distributions** while remaining efficient and stable to train.

This approach is particularly useful for **Boltzmann sampling**, which is widely used in molecular modeling to generate realistic molecular configurations. Compared to existing methods, our model provides **faster and more accurate sampling**, making it suitable for applications such as **small molecule generation and optimization**.

Specific Aim

The main goal of this project is to: **Develop and implement a flow-based generative model capable of learning molecular distributions and generating novel small molecules for drug discovery applications.**

What is available:

1) We already developed the model and theory of DeepMix Flow
2) We already developed a basic code in Python (using libraries such as TensorFlow or PyTorch) which implement forward and inverse mappings and compute Jacobian determinants

Tasks

Students participating in this project will:

1. **Understand the model**
   - Learn the basics of generative models and normalizing flows
   - Study how the proposed piecewise-affine transformations work
2. **Implement the model**
   - Adapt the code to generate small molecule
3. **Train the model**
   - Train the model on molecular datasets (e.g., small molecule structures)
   - Monitor training stability and performance
4. **Apply to molecule generation**
   - Use the trained model to **generate new small molecules**
   - Visualize and analyze generated molecules
5. **Evaluate results**
   - Compare generated molecules with real datasets
   - Assess quality, diversity, and chemical validity

Expected Outcome

By the end of the project, students will:

- understand **modern generative AI methods in drug discovery**
- gain hands-on experience in **deep learning and scientific modeling**
- produce a **working model capable of generating small molecules**

references:

1) Github where the code is from Salem Lahlou:

https://github.com/GFNOrg/torchgfn

Paper 1: Lahlou, Salem, et al. "A theory of continuous generative flow networks." *International Conference on Machine Learning*. PMLR, 2023.

Paper 2: Jain, M., Bengio, E., Hernandez-Garcia, A., Rector-Brooks, J., Dossou, B.F., Ekbote, C.A., Fu, J., Zhang, T., Kilgour, M., Zhang, D. and Simine, L., 2022, June. Biological sequence design with gflownets. In *International conference on machine learning* (pp. 9786-9801). PMLR.

2) Github for perturbed Normalizing Flow:

Data: Modeling the Chignolin mutant using the CHARMM22 force field with the implicit OBC2 solvent model. We used simulation data from Frank Noé's research group (http://ftp.mi.fu-berlin.de/pub/cmb-data/bgmol/datasets/chignolin/ChignolinOBC2PT.tgz)

Code availability: The code used in this study for training models, generating samples, and analyzing data is openly available on GitHub at:
https://github.com/XinPeng76/Flow_Perturbation
and has been archived on Zenodo with the the https://doi.org/10.5281/zenodo.15744467

Paper 3: Peng, Xin, and Ang Gao. "Flow perturbation to accelerate Boltzmann sampling." *Nature Communications* 16, no. 1 (2025): 6604. https://doi.org/10.1038/s41467-025-62039-8