LEARNING TO GENERATE MOLECULES

Using the diffusion model to drug discovery

Authors
Frédéric Charbonnier & Joël Clerc

Research Director
Alexandros Kalousis

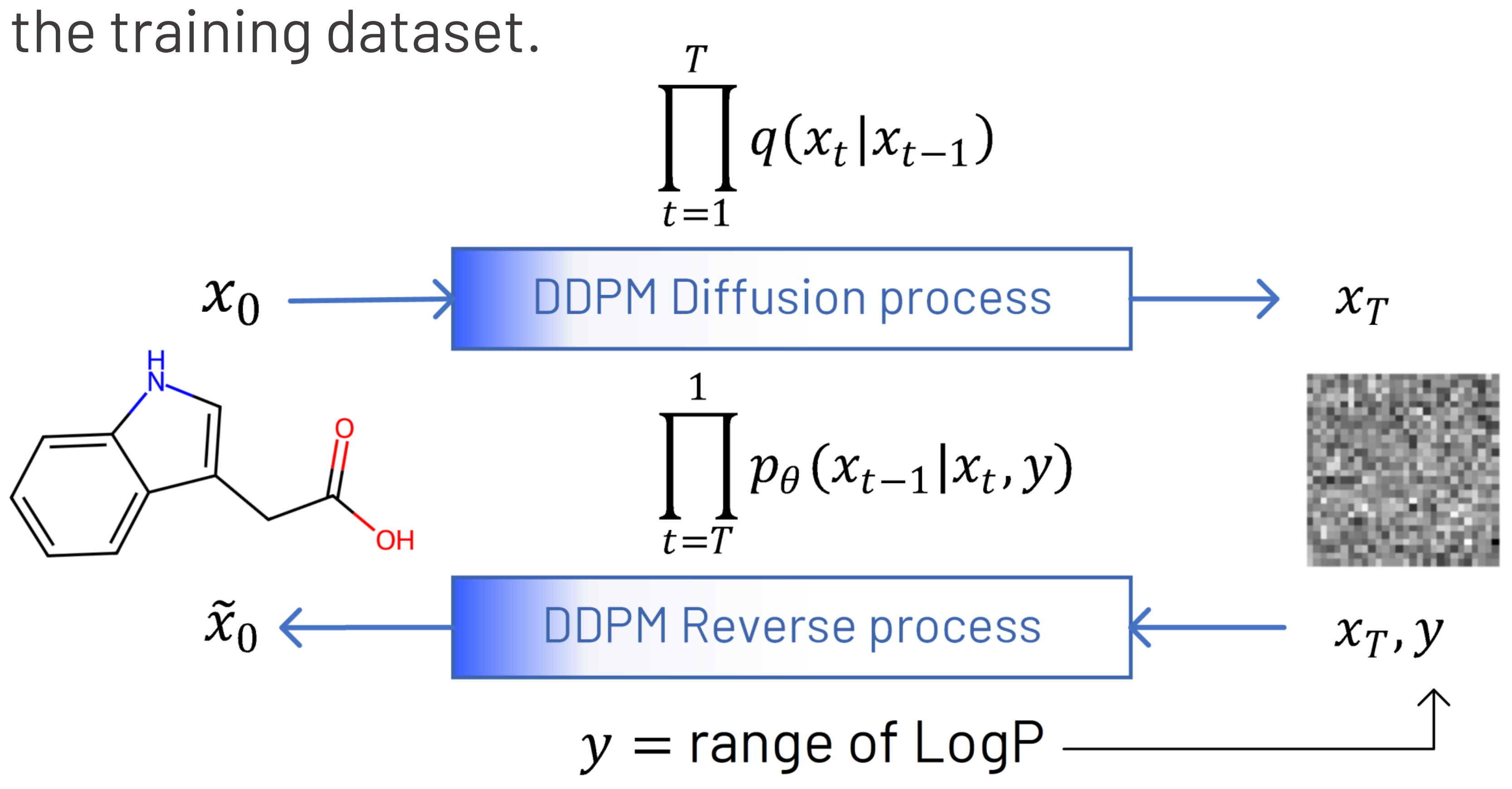
Affiliations
Haute École de Gestion de Genève
HES-SO // Genève

1 Introduction

The development of a new drug typically takes 10 to 15 years and costs around USD 2.6 billion. Machine learning generative models, like diffusion-based models, have the potential to revolutionize molecular design by generating candidate molecules with desired properties, significantly reducing time and cost in drug development and related applications.

2 Objective

Evaluate the efficiency of Denoising Diffusion Probabilistic Models[1] (DDPM) to generate molecules with a desired range of LogP (octanol/water partition coefficient), syntactically valid and novel compared to the training dataset.



$\prod_{t=1}^{T} q(x_t|x_{t-1})$

x_0 → DDPM Diffusion process → x_T

$\prod_{t=T}^{1} p_\theta(x_{t-1}|x_t, y)$

$\tilde{x}_0$ ← DDPM Reverse process ← x_T, y

y = range of LogP

3 Method

Pre-processing: Transformation of 105’625 molecules in the QM9 dataset[2] from SMILES[3] to SELFIES[4] and encoding into vectors of continuous values. Calculation of the LogP for each molecule, followed by classification into 3 classes.

Training: Training of the model on 330’000 steps of batches of 16 molecules and their associated classes (around 50 epochs, for a total duration of around 46 hours).

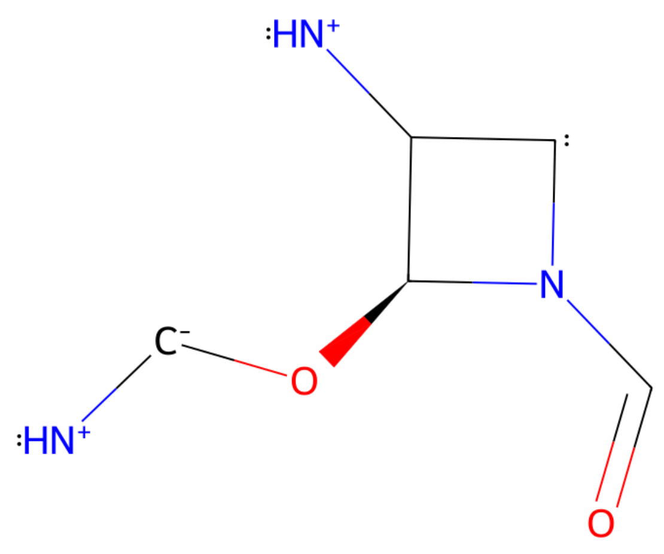
Evaluation: Every 10,000 steps, generation of 100 molecules and calculation of the corresponding class match (% desired class = real class of the molecule generated). The best model is saved based on the corresponding class match.

4 Results

To obtain these results, 1’000 molecules are generated with the best model, followed by calculation of the corresponding class match, the valid syntax rate, the rate of new molecules compared with the training set and the Wasserstein distance (similarity measure between two distributions) between the LogP values in the training set and the molecules generated.

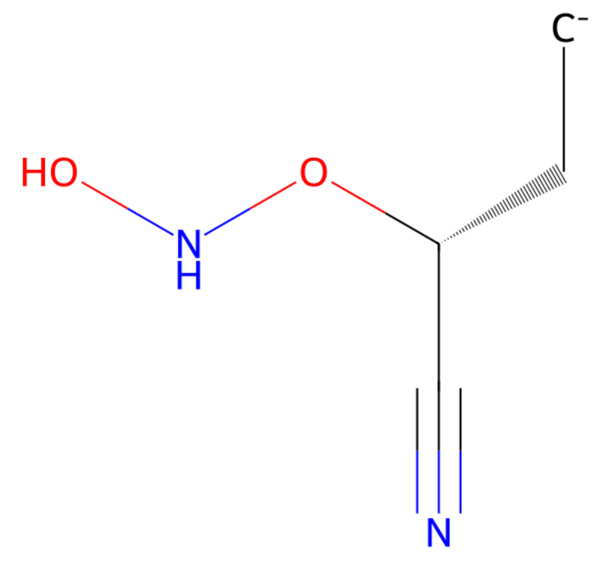
LogP increase →

Class 1



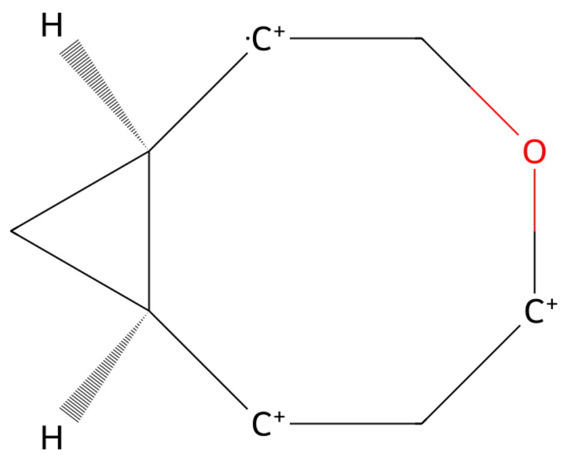
LogP -4.045

Class 2



LogP 0.013

Class 3



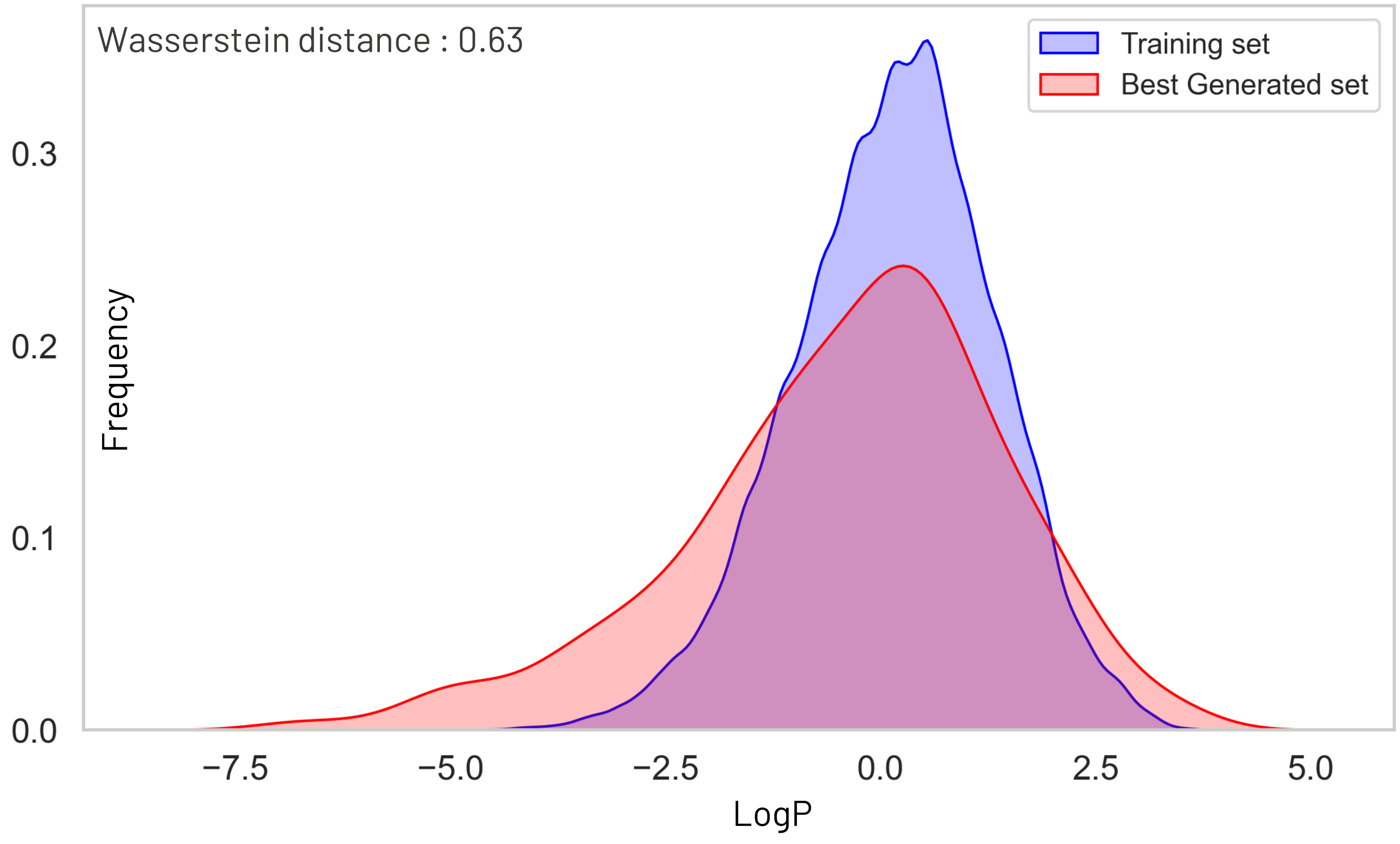
LogP 1.798

55.3% class match

100% syntax validity

94.3% molecules novelty

Distribution of LogP for the generated and training sets



5 Conclusion

The results show that DDPM can be of interest in the field of drug discovery. They can generate new syntactically valid molecules with properties close to those of the training set. Their conditional use allows the desired LogP range to be targeted, with a success rate of over 33.3% (random generation with 3 classes).

References

[1] HO, Jonathan, JAIN, Ajay and ABBEEL, Pieter, 2020. *Denoising Diffusion Probabilistic Models*. arXiv:2006.11239. arXiv. arXiv:2006.11239. DOI 10.48550/arXiv.2006.11239.

[2] RAMAKRISHNAN, Raghunathan et al., 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*. Vol. 1, no. 1, p. 140022. DOI 10.1038/sdata.2014.22.

[3] BJERRUM, Esben Jannik, 2017. *SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules*. arXiv:1703.07076. arXiv. arXiv:1703.07076. DOI 10.48550/arXiv.1703.07076.

[4] KRENN, Mario et al., 2022. SELFIES and the future of molecular string representations. *Patterns*. Vol. 3, no. 10, p. 100588. DOI 10.1016/j.patter.2022.100588.

Realized as part of the Master of Science HES-SO en Sciences de l'information.