Automated de-novo molecule design based on Deep Neural Networks

Magdalena Wiercioch [1] and Sabina Podlewska [2]

[1] Jagiellonian University, Department of Information Technologies, Lojasiewicza 11, 30-348 Krakow, Poland

[2] Institute of Pharmacology, PAS, Department of Medicinal Chemistry, Smetna 12, 31-343 Krakow, Poland

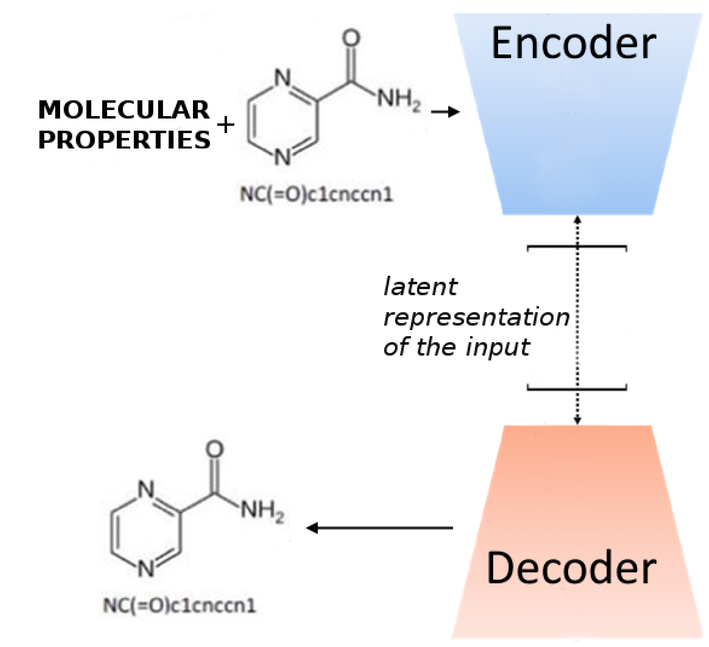
Correspondence: magdalena.wiercioch@uj.edu.pl

INTRODUCTION

Searching for the novel molecules which have the desired properties is the primary goal of material and drug design. The challenge is posed by the fact the search space is extremely immense and discrete. For this reason, the task of generating and testing new potential candidates is either costly or takes a lot of time. Recent advances in machine learning have raised increasing interests in powerful probabilistic generative models which after their training on real samples are able to yield realistic synthetic examples. We propose a molecular generative model called FGVAE that uses the grammar variational autoencoder (GVAE) [1].

MODEL

As a molecular generator we selected GVAE [1]. The molecules were represented by their SMILES codes. In our model, the molecular properties we want to consider were added as the extra production rules that can be used for constructing a molecule. As a result, the FGVAE can generate valid molecules with the desired properties imposed by the rules.



DATASET, PARAMETERS

- The total dataset is made of 1,000,000 molecules randomly selected from the PubChem dataset [2].
- To calcule the four target properties of the molecules we used RDKit [3].
- The learning rate was set to 0.0005.
- The model was trained until convergence.

REFERENCES

[1] Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1945–1954, 2017.

[2] Kim S, Thiessen PA, Bolton EE, Fu G Chen J, Gindulyte A, Han L, He J, He S, Shoemaker BA, Yu B Wang J, Zhang J, and Bryant SH. Pubchem substance and compound databases. *Nucleic Acids Res*, 44(D1):202–213, 2016.

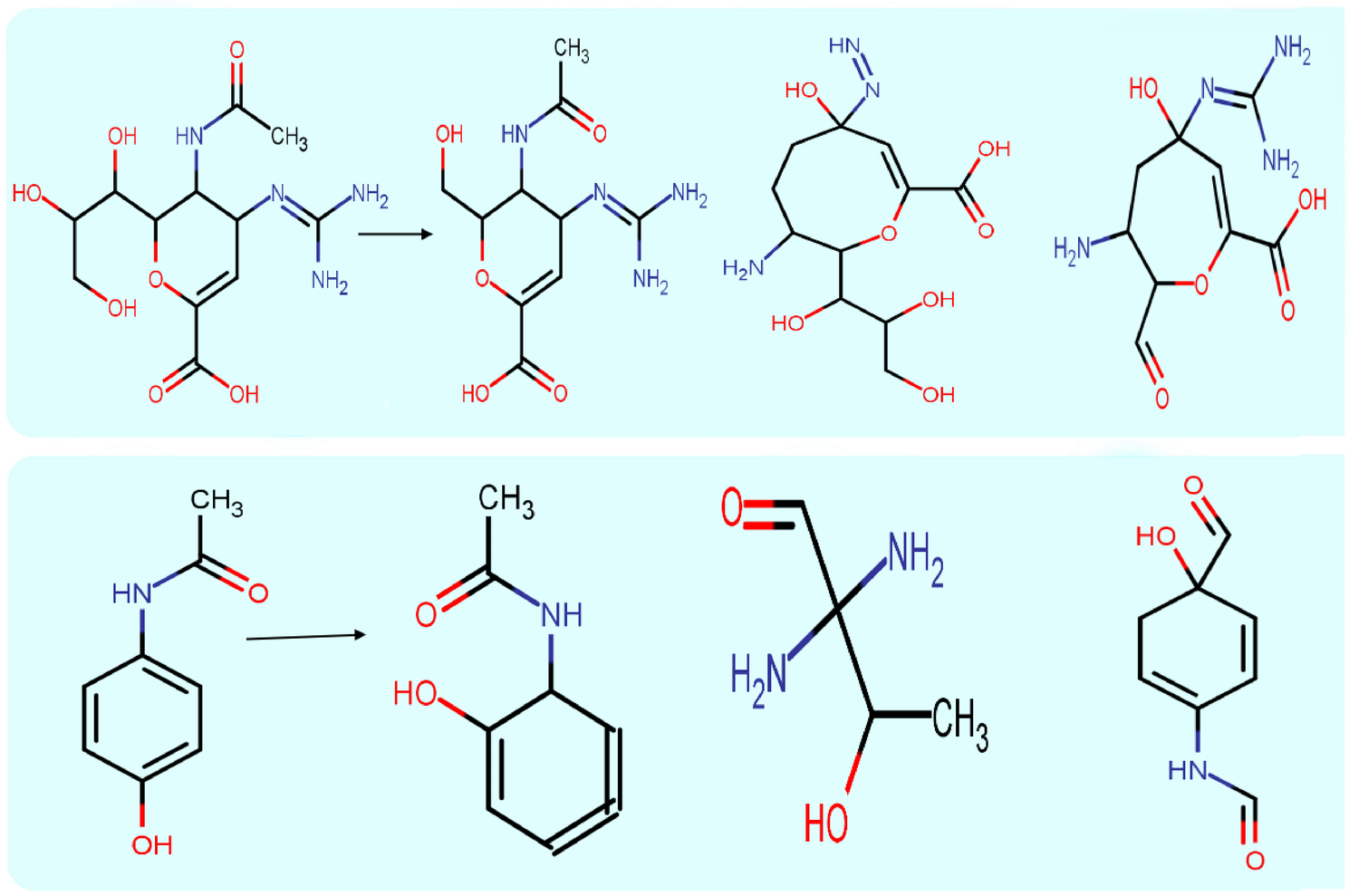
[3] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

ACKNOWLEDGEMENTS

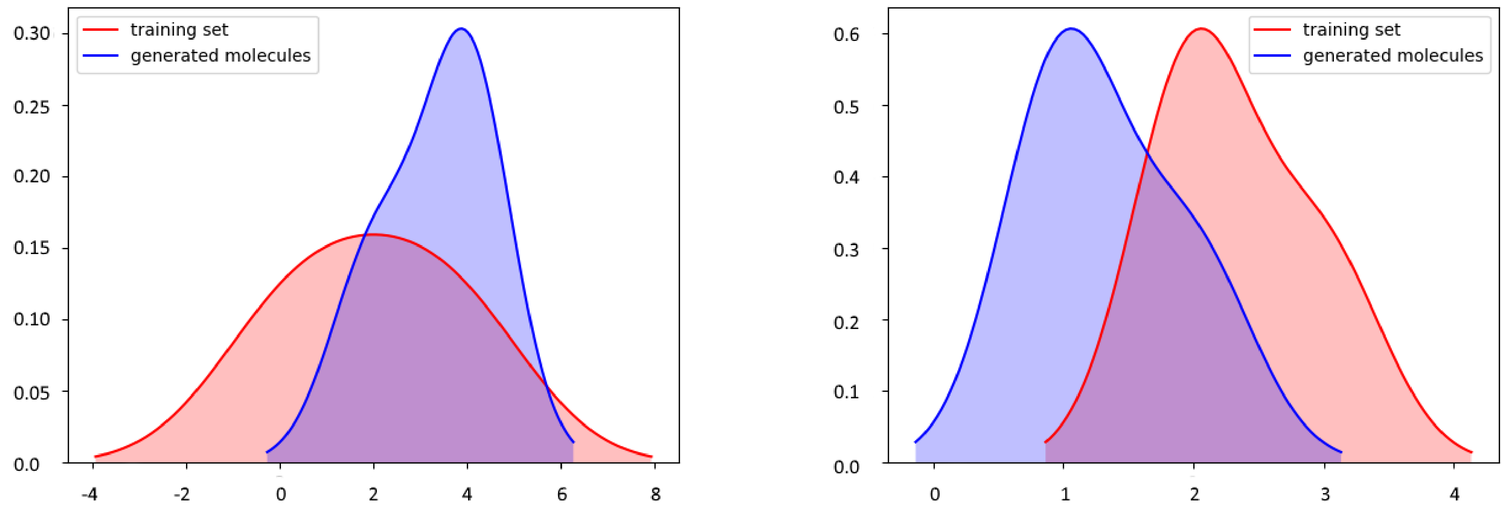
This research was partially supported by National Centre of Science (Poland) Grants No. 2016/21/N/ST6/01019.

EXPERIMENTS

As the first application, we demonstrated that the FGVAE approach can generate molecules with specific values for the four target properties (partition coefficient - logP, number of hydrogen bond donor - HBD, number of hydrogen acceptor - HBA, topological polar surface area - TPSA) by applying it to Relenza and Paracetamol. The values of the logP, HBD, HBA, and TPSA for Relenza and Paracetamol are (-3, 7, 8, and 201) and (0.46, 2, 2, and 49.3), respectively. The extra set of production rules for each molecule was made by those values. The following figures present three molecules produced with the set of production rules for Relenza and Paracetamol, respectively. All of them had similar properties to those of Relenza and Paracetamol.



We also compared the distribution of the logP and HBA for 500 randomly selected molecules from the training set and 500 generated molecules with property values outside of the range of the dataset. Comparison results values are shown below.



CONCLUSION AND DISCUSSION

We proposed a new molecular design method based on the grammar variational autoencoder which is a promising technique to construct virtual libraries and discover new drugs.

1. The method directly produces molecules with desirable target properties which are valid.
2. The model enables to controll multiple target properties simultaneously.
3. We have proved that it was possible to generate druglike molecules with specific values for the four target properties (logP, HBD, HBA, and TPSA) within an error range of 25%.