AI-Driven Discovery: Exploring Generative Models for Novel DYRK1A Protein Inhibitors

Eduardo González García [1] Pablo Varas Pardo [1] Nuria Campillo [1,2]
Simón Rodríguez Santana [1,4] Pedro González Naranjo [3] Juan A. Páez [3]

[1]Institute of Mathematical Sciences (ICMAT-CSIC)
[2]Centro de Investigaciones Biológicas Margarita Salas
[3]Institute of Medical Chemistry (IQM-CSIC)
[4]ICAI Escuela Técnica Superior de Ingeniería (U.Comillas)

Abstract

We explore the applications of two distinct generative AI models in the search for new inhibitors of the DYRK1A protein, a promising therapeutic target for addressing specific aspects of Alzheimer's disease and other conditions. A key aspect of this work involves training predictive models on properties of interest, such as molecular affinity to the protein or toxicity. These models were instrumental in guiding the generative process and screening potential candidates. Furthermore, some proposed candidates are being experimentally tested achieving promising inhibition percentages.

Data

Toxicity data:

- 11765 molecules from the Tox21 dataset.
- Variables: Uniquely identifying SMILES codes for each molecule.
- Labels: 12 binary classes that represent the outcome of 12 different toxicological experiments.

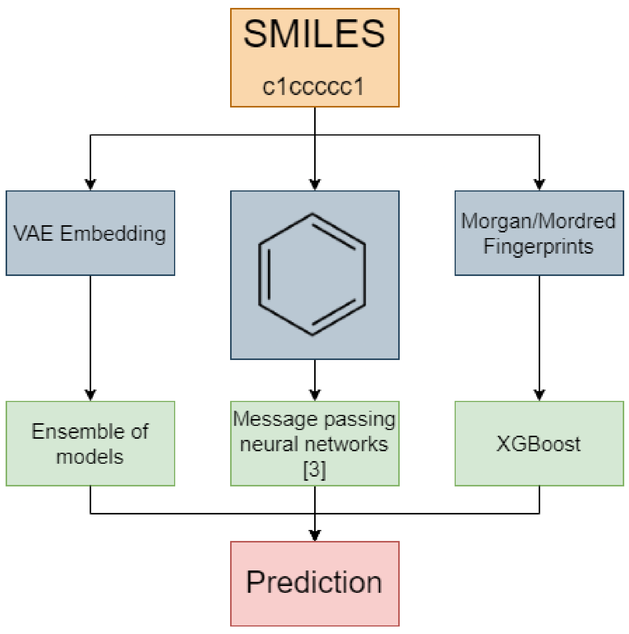
DYRK1A affinity data:

- 1783 molecules from a custom dataset.
- Variables: Uniquely identifying SMILES codes for each molecule.
- Labels: pChEMBL affinity value calculated from IC_{50}, Ki, Kd or EC_{50} measurements.

Molecule data representation:

- Text strings given by their SMILES code.
- Graphs representing the molecule structure.
- Binary vectors of their molecular fingerprints.

Predictive models



We use the graph representation of molecules to predict toxicity. For affinity predictions, we incorporate all available data representations to minimize prediction variance considering our limited data paradigm.

Generative models

Hierarchical Generation of Molecular Graphs using Structural Motifs:

As described in [1], this model aims to generate molecules by piecing together structural motifs, which work as basic building blocks.

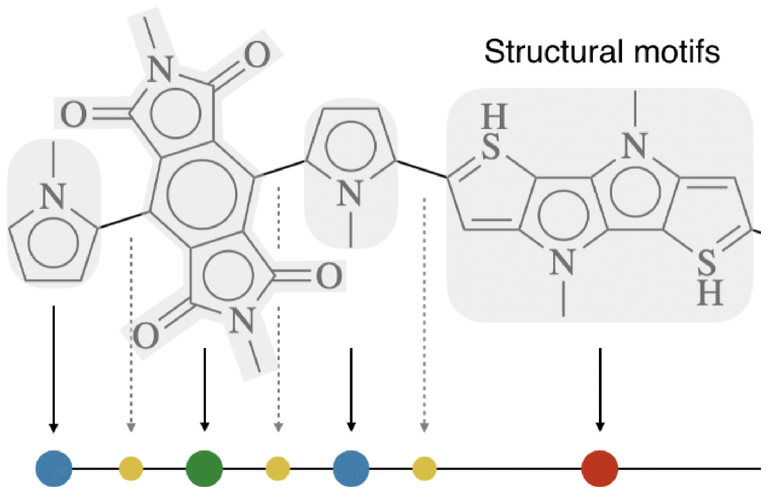


Figure 1. Structural motif representation.

We take the model pretrained on the ChEMBL database, and finetune it to generate DYRK1A inhibitors following an iterative process:

1. Train the model with an unsupervised task on our DYRK1A dataset.
2. Generate new molecules randomly and filter them out based on similarity, toxicity and affinity predictions made by the predictive models.
3. Include the new molecules in the training dataset and return to step 1.

Pocket2Mol Generator:

As described in [2], this model aims to generate molecules considering the structure of the protein pockets where the desired binding is going to occur. The algorithm follows four steps as presented in the figure.

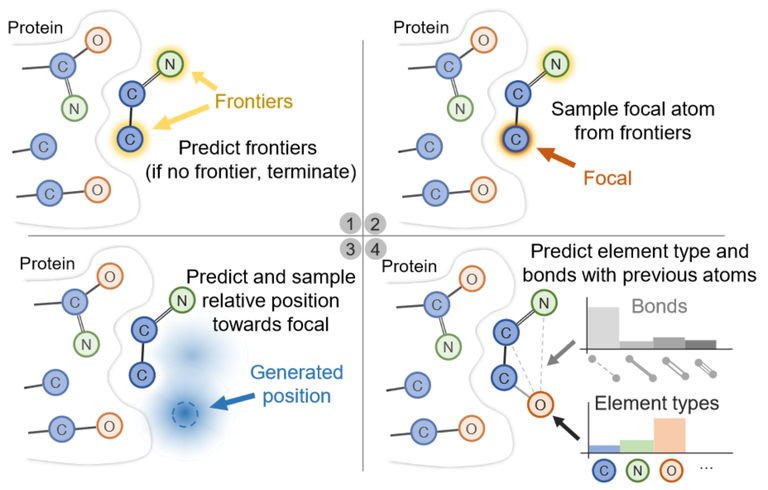
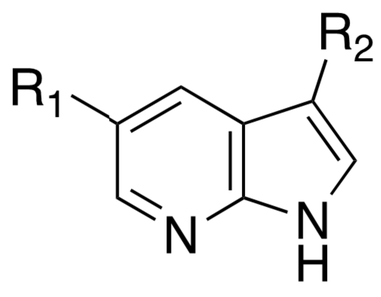


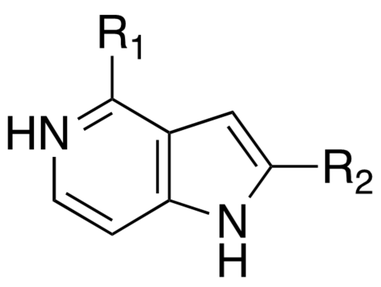
Figure 2. Generation process of Pocket2Mol.

Having the XYZ coordinates of the protein pocket of interest, we employ this method to generate candidate molecules. Subsequently, we filter and rank these molecules based on the predicted toxicity and affinity.

Results



Docking score: -13.99 kcal/mol



Docking score: -13.91 kcal/mol

Several proposed molecules present higher simulated docking scores than the crystallized reference ligand (at -8.42 kcal/mol). Furthermore, the top candidate has been synthesized achieving an experimental result of 99% inhibition at 10 μM.

Conclusions

In this work, we demonstrate the effectiveness of employing state-of-the-art generative models in conjunction with predictive models to generate potential inhibitors of the DYRK1A protein. It is crucial to highlight that these methods have shown remarkable utility, even in a low-data paradigm where limited information about protein inhibitors is available.

Future work

- Include deep learning docking estimators to the generative process in order to improve the quality of the proposed molecules.
- Expand the current models for the generation of multi target inhibitors.

Bibliography

[1] Jin W, Barzilay R, Jaakkola T. Hierarchical Generation of Molecular Graphs using Structural Motifs. ICML. **2020**;119:4839-4848.

[2] Peng X *et al*. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. ICML. **2022**;162:17644-17655.

[3] Yang K *et al*. Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model*. **2019**;59(8):3370-3388.

Acknowledgements: This work was funded by grants from Spanish Ministry of Science and Innovation. Development of an artificial intelligence-based framework to accelerate drug development (environmental and digital transition projects TED2021-129970B-C21).