Adapting Deep Learning Architectures for Drug Discovery

mgr Łukasz Maziarka

Abstract

The process of discovering new drugs is complex, costly and resource-intensive, with a significant portion of drug candidates failing in late-stage clinical trials. This high attrition rate highlights the need for improved predictive models and more efficient methods for optimizing chemical compounds early in the drug development pipeline. This dissertation posits that deep learning architectures can be specifically adapted and optimized for cheminformatics tasks, addressing challenges in molecular representation to enhance the efficiency and accuracy of various stages in drug discovery.

This doctoral dissertation explores the adaptation of deep neural networks to key cheminformatics challenges, with a specific focus on computer-aided drug design. The research presented in this thesis is organized around two main areas: the prediction of molecular properties and the generation of new chemical compounds. The dissertation introduces novel deep learning models that leverage graph convolutional networks (GCNs), attention mechanisms, and transformer architectures to enhance the accuracy and efficiency of molecular property prediction. Additionally, it presents innovative methods for the generation and optimization of molecules, including the use of generative adversarial networks (GANs) and flow-based models for conditional molecular generation.

The contributions of this work include the development of a *Set Aggregation Network (SAN)* for improved feature aggregation in graph-based models, the introduction of a *Spatial Graph Convolutional Network (SGCN)* that integrates 3D spatial information into the convolution process, and the creation of the *Molecule Attention Transformer (MAT)* and its variant, the *Relative Molecule Attention Transformer (R-MAT)*, both utilizing attention mechanisms to capture complex molecular interactions. Furthermore, this dissertation presents *HuggingMolecules*, an open-source library that facilitates access to pre-trained models for molecular property prediction, compares various atomic representations used in graph neural networks to identify optimal input features and introduces *EmBERT* - an adaptation of BERT architecture to multitask learning.

In the domain of molecular generation, this dissertation introduces *Mol-CycleGAN*, a method for lead optimization using GANs to modify existing molecules, and *PluGeN*, a plugin generative network that transforms latent spaces for controlled molecular generation.

These models demonstrate the potential of deep learning to not only predict the properties of chemical compounds but also to generate novel molecules with desired characteristics, thereby accelerating the drug discovery process.

The findings and models presented in this dissertation have been widely recognized in the cheminformatics community, contributing to the advancement of deep learning methods in drug design, further evidenced by the candidate's substantial citation record, exceeding 800 citations according to Google Scholar as of December 2024. The research has resulted in several high-impact publications and has established a foundation for future work in the application of artificial intelligence to drug discovery.

To summarize, this doctoral thesis focuses on the demonstration of the potential of adapting deep learning architectures for cheminformatics, with a particular focus on computer-aided drug design.

Nine papers were published based on this research. Two at CORE rank A* conferences, four at rank A conferences, one on workshop track of rank A* conference and two in a reputable cheminformatics journal. In five of them, the Ph.D. candidate was the first author. Additionally, the Ph.D. candidate was a principal investigator of the NCN Preludium grant which was strongly connected to this research.

Keywords: deep learning, cheminformatics, drug design, molecular property prediction, molecular generation.