Investigating Molecular Graph Convolution in Drug Discovery: Application to SARS-CoV-2 Main Protease Inhibitors

Markus Ferdinand Dablander

The computer-aided prediction of the properties of molecules on the basis of their chemical structure is an important field of research at the intersection of applied mathematics, computer science and chemistry. In recent years, modern artificial intelligence (AI) techniques have shown great promise for the computational prediction of important molecular attributes and subsequently for the discovery of novel molecules with desired biological or chemical functions. As a result, such techniques are of strong interest for Lhasa Limited, whose key scientific goals include the streamlining and standardisation of drug development processes, the reduction of animal testing, and the computer-based estimation of the potential toxicity of molecular compounds.

If designed carefully, then AI models can automatically extract knowledge and statistical rules from a given data set D in a process called *training*. In the case of molecular property prediction, such a training data set D usually contains hundreds or thousands of molecules which were sampled roughly from the same *chemical space* C. A chemical space is a often extremely large family of molecules which all fulfill a certain set of constraints. For example, all molecules with less than 10 atoms form a chemical space. Each molecule in D must come with a number p which represents a measurement for a target property which one wants to predict. Such target properties include elementary chemical features like water solubility, but also complex biological and pharmacological qualities like how much of an orally-taken drug is absorbed by the human body. The goal is to train an AI model on the molecular data in D in such a way that it learns to predict the target properties of previously unseen compounds from the chemical space C and ideally also from new chemical spaces.

However, the problem is that AI models are computer programs that cannot directly process man-made chemical descriptions of molecules but can only process hard numbers. The first step must therefore be to somehow transform each molecule m in the training data set D into a sequence of numbers $\boldsymbol{x} = (x_1, ..., x_n)$. The sequence of numbers $\boldsymbol{x}$ is called the *feature vector* of the molecule m and the process of turning a molecule into a feature vector is referred to as *molecular featurisation*. The full workflow for AI-based molecular property predit is depicted in Figure 1. The use of a smart molecular featurisation technique is crucial for the success of a molecular property prediction model. Featurisations should ideally capture all quantities which are relevant for the predictive tasks at hand and at the same time omit all unnecessary information in order not to confuse the AI model.

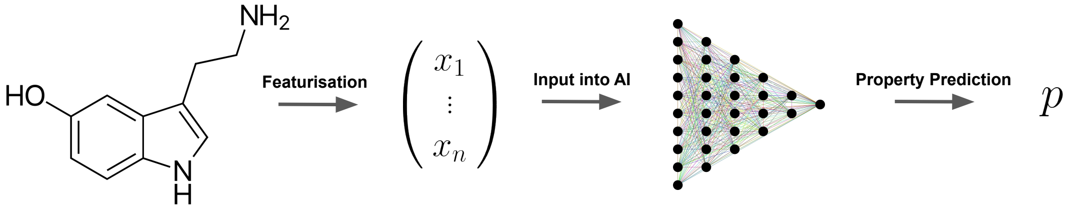

Figure 1 – Standard workflow of AI-based molecular property prediction.

Engineering and Physical Sciences Research Council

In this project, we investigated the relative performance of two competing molecular featurisation techniques: *extended-connectivity fingerprints* (ECFPs) and *molecular graph convolutions* (MGCs). ECFPs represent a well-established featurisation method which has been applied widely in the last decade. In comparison, the first MGCs only entered the scene in 2015 and are currently under research. ECFPs encode a molecule into a long fixed-size sequence of 0s and 1s; each digit symbolizes the absence (= 0) or presence (= 1) of a particular molecular substructure. MGCs, on the other hand, are based on trainable artificial neural networks which encode a molecule into a hard-to-interpret, information-dense sequence of numbers. MGCs are much more flexible than ECFPs because they can be trained to adapt to a specific prediction task, but they also take much longer to compute and are more prone to get confused by random errors in the data by falling into a trap called *overfitting*. In recent years, both ECFPs and MGCs have reached state-of-the-art results in important molecular property prediction tasks. Some researchers report that MGCs are superior to ECFPs in many applications, while others report the opposite. The question of, under which circumstances are each of the two featurisation methods superior to the other, has yet to be resolved. To shed more light on this question, we conducted a comparative computational study on two different data sets to see which featurisation method (ECFPs or MGCs) would do better. To do this, we developed our own version of an MGC which we called *rational graph convolution* (RGC).

In our first group of experiments, we used ECFPs and RGCs to build AI models for the classification of compounds which inhibit the pharmacologically important hERG potassium channel using a data set of 9748 compounds. In our second group of experiments, we again applied ECFPs and RGCs, but to try and learn to classify potential inhibitors of SARS-CoV-2 from a data set of 546 tested compounds. SARS-CoV-2 is the virus which caused the global public health emergency which started in 2019 in Wuhan, China. Finally, we experimented with pretrained MGCs: we trained RGCs on the very large hERG data set, used the pretrained RGCs to featurise the molecules in the small SARS-CoV-2 data set and then used these features to again try and identify SARS-CoV-2 inhibitors.

Overall, our results suggest that RGC-based models are slightly superior to ECFP-based models when tested on novel molecular data from the same chemical space on which they were trained on. However, when the models are tested on molecules from a different chemical space, then RGC-based models show slightly inferior performance compared to ECFP-based models and a much greater variance in their achieved accuracy. This high variance in the generalisation abilities of RGCs to new chemical spaces might be due to some underlying systematic reason, perhaps associated with the highly adaptive nature of RGCs. The investigation of such a reason could form an angle of attack for the future technical improvement of RGCs and MGCs in general. Our best model for the classification of hERG inhibitors is based on RGCs and achieves 81.54% accuracy.

Our pretraining-related results indicate that RGCs which are trained *a priori* on large molecular data sets can be used as powerful featurisation tools for small data sets from new chemical spaces. In our appliation to SARS-CoV-2, this featurisation technique was able to beat both ECFPs and non-pretrained RGCs by achieving an accuracy of 78%. These highly encouraging results provide motivation for the future exploration of the full potential of pretrained MGCs as a tool for the boosting of AI model performance during molecular property prediction when data is scarce.

Dr Thierry Hanser, Head of Molecular Informatics and AI at Lhasa Limited, said:

> *Markus was given a challenging project with many Cheminformatics and Machine Learning concepts to digest and leverage in a very brief project period. I was very impressed how Markus became quickly able to understand the technical aspects of the problem and how his perseverance allowed him to design and implement a solution in this short time. Markus became quickly operational and effective using a number of sophisticated tools and software libraries. Markus was very proactive and independent; he developed his own Graph Convolutional Network to overcome the complexity of existing solutions in order to be able to extend the approach in the future. Markus combined his model with other Neural Network methodologies to form a unique approach (Rational Convolutional Graphs) with increased performances. Markus also applied successfully transfer learning to his solution which resulted in a substantial gain in performance. The outcome of Markus' project provides a solid foundation for a more in-depth research on Molecular Graph Convolutional Networks. Markus demonstrated excellent problem solving skills along with the ability to implement sophisticated solutions.*