CELERIS

Identification of Drug-Target Interactions using Xanthos Match Maker™

In this whitepaper we highlight the potential of using graph representations for the identification of drug-target interactions (DTI).

TECHNOLOGY:
Xanthos Match Maker™

RESULTS:
We refer to our case study on drug-target interactions.

INTRODUCTION:

Virtual screening has become increasingly important for the discovery of novel drugs and is nowadays considered as one of the essential components for the early-stage drug development pipeline.
Given a target protein that is related to a disease, the goal is to identify compounds that interact with a target of interest, desirably with a high binding affinity. This is enabled by screening the target against libraries with millions up to billions of compounds.
Upscaling a virtual screening to a scale of billions of compounds is associated with expensive and time-consuming computations. With the third wave of AI, deep learning approaches continuously yield surprising results considering what computers can achieve running effectively on graphic processor units (GPUs).
Virtual screening can be framed as a supervised learning problem (receiving inputs and predicting real-valued outputs), which makes it an ideal application for deep learning. This makes deep learning one of the most promising *in silico* approaches for conducting these screenings. It introduces an encouraging solution for the ultra-fast identification of novel drugs.

Starting from a library of compounds and a target protein, we predict binding affinities that resemble the strength of binding between a library of compounds and a target.

X: Target protein
Y: Real-value scalar
Our goal is to identify a generalized function that can map the input to the output space.
F: X → y

It is particularly important to emphasize generalization aspects, as algorithms need to predict reliable outputs on both compounds as well as targets that are coming from real world-data.

In this whitepaper, we describe how we are predicting drug-target interactions (DTIs) based on deep learning. We propose methods featuring high generalization capacity.
The exact details of the technology Celeris Therapeutics is employing are kept confidential. Here, the general aspects of our architecture and employed methods are discussed.

METHOD:

Xanthos Match Maker™ is the screening tool in Celeris One, the *in-silico* drug discovery platform of Celeris Therapeutics. It is a generalized tool for predicting interactions between compounds and targets of interest.

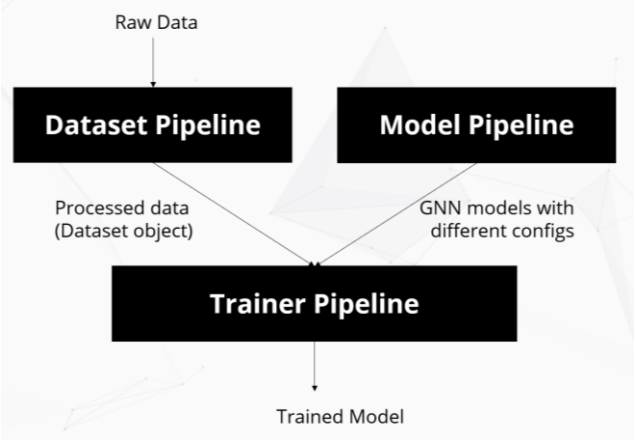

Figure 1: Three pipelines for solving drug-target interaction predictions

In figure 1, we depict the three pipelines that we implemented for DTI predictions:

- **Dataset pipeline**: Conversion of raw data into an object that can be processed by machine learning algorithms.
- **Model pipeline**: Implementation of architectures for predicting DTI.
- **Trainer pipeline**: Implementation for handling training and finding the best hyperparameters.

Dataset Pipeline

As the input for virtual screenings consists of compounds and target proteins, it is necessary to convert the raw data of these molecules into suitable representations that are fed into our deep learning model. As these molecules are inherently different objects, the Celeris Therapeutics dataset pipeline is consisting of different methods for these conversions.

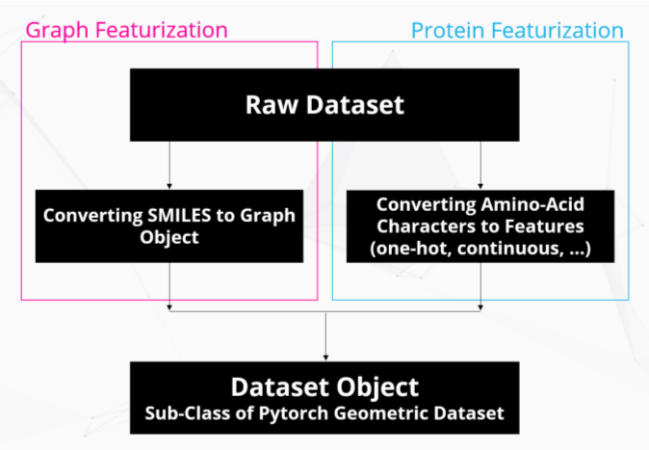

Figure 2: Conversion of input compounds and target proteins into representation that are further processed.

1.) Conversion of compound raw data:

Several representations for compounds are used. Most importantly:

- SMILES/SELFIES
- InChI
- Fingerprints (e.g., Morgan fingerprints)
- Two-dimensional graphs

Due to the breakthrough of graph representation learning (GRL; through graph neural networks) on processing graph-structured data (e.g., molecules), Celeris Therapeutics identified GRL as an ideal methodology for processing multidimensional data as graphs. In this representation, the nodes are atoms, and the edges are bonds between atoms. The Celeris One platform considers different features such as the total number of hydrogens, implicit valence, and aromaticity, as the features of each node (atom) in the graph.

We can featurize edges (bonds) and exploit message passing neural networks on top of that. Based on the problem and distribution of the data, Celeris One can automatically choose whether adding edge features can increase the performance or not.

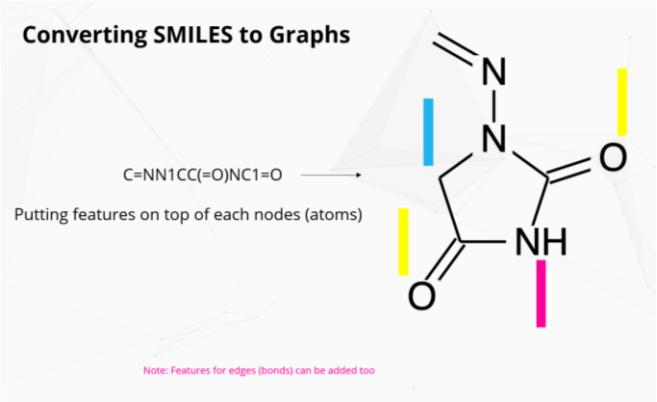

Figure 3: Graph featurization schema. This picture shows how we are converting the smiles representation into the graph protein featurization.

In traditional machine learning methods, proteins are represented in amino-acid characters and transformed into vectors via one-hot encoding. The emergence of transformer architectures such as BERT and Roberta enables a more promising representation of proteins. This leads to learning powerful protein representations that boost the performance of downstream predictions.

We have trained some of these architectures on vast amounts of unlabeled protein data and used them to extract useful representations.

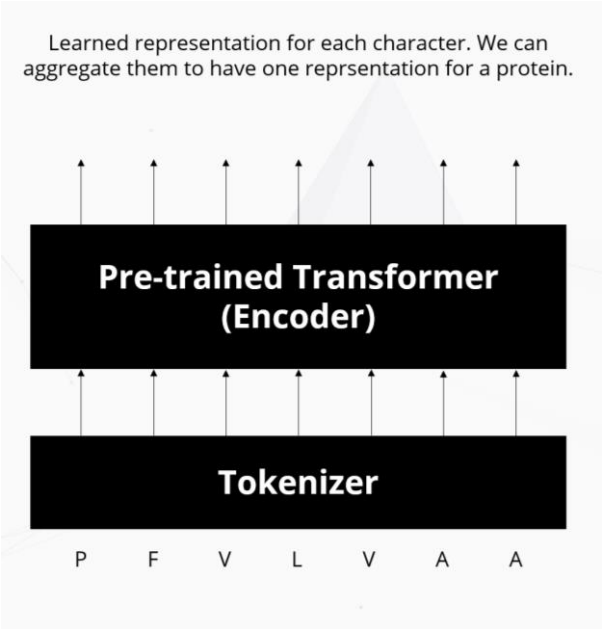

Figure 4: Protein featurization scheme.

Model Pipeline

Celeris One exploits architectures that can work on input representation and produce a real-valued output (binding affinity). A scheme of the model pipeline is highlighted in figure 5.

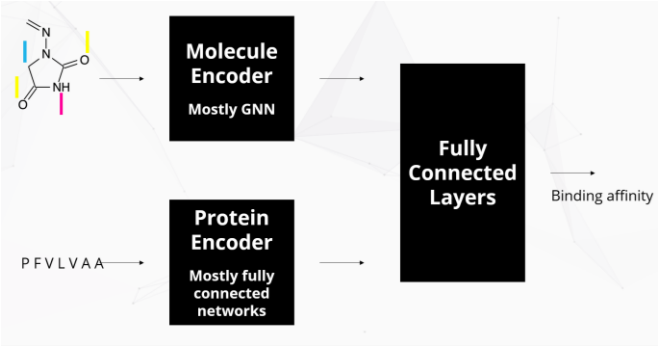

Figure 5: Overall schema of model pipeline. The architecture contains one encoder for molecules, and one for protein. The merged representation is fed into an MLP.

The architecture contains the following components: One encoder for proteins, another encoder for small molecules. Their output is fed to a multilayer perceptron (MLP) network, which takes the merged representations and yields the binding affinity. For the molecule encoder, we are using graph neural networks and choose the most promising automatically.

We have two stages for discovering the best architectures: First, we are fixing the architectures (different GNNs), and we are trying to find the best hyper-parameters through our hyperparameter optimization method. The second method is related to the active field of neural architecture search, i.e., identifying the architectures featuring highest performance automatically.

In this regard, we have been able to achieve promising results on the common benchmark data for DTI prediction. But Celeris Therapeutics goal is not just to have significant performance on common benchmarks; the goal is a reliable and strong prediction performance for the real world and unknown compounds (or target proteins) that are not necessarily coming from the input distribution. So, we have also tried to find the solution for this problem within our Celeris One platform.

CAVEATS & SOLUTIONS

First caveat: Out of distribution generalization

We can train a machine learning/deep learning algorithm to reach significant results on different test sets that have the same distribution as the training data, but we cannot guarantee the same performance on data sets with different distribution. This problem is known as out of distribution (OOD) generalization. Our end goal is to achieve significant results in the currently available chemical space and go beyond that. Celeris Therapeutics, therefore, came up with some of the most sophisticated solutions.

First solution: Data augmentation

Data augmentation is a method, well-known from computer vision, to improve the performance and generalization of prediction algorithms. Data augmentation can expand the input space, and in that way, it results in more diverse data from different regions of the input space. Hence, it boosts the generalization performance. Expanding this concept into the chemical space is a timely topic that has gained attention just recently[1] and brings measurable value for drug discovery. Consequently, we have implemented a unique feature for data augmentation to OOD generalization in the binding affinity prediction. Automatically expanding the input chemical space can make the prediction on unknown input more reliable.

Second caveat: Low-data regime affinity prediction

For most of the target proteins, we do not have enough training data for training expressive deep function approximators.

These sparse data situations need to be used efficiently. The solution is a method for finding compounds with high affinity just by having access to a small amount of training data.

Second solution: Bayesian optimization

Celeris Therapeutics exploits Bayesian optimization for finding the compounds with high affinity with the least possible steps. Bayesian optimization is a black-box (zero-order) optimization method that can be used for finding the minimum/maximum of unknown functions. We are employing flexible proxy functions (instead of actual functions) for reaching the optimum from sparse data.

SUMMARY

In this whitepaper, we have explained our Xanthos Match Maker™ tool for DTI prediction and virtual screening on an abstract level. Not only is our solution capable of achieving significant performance on a benchmark dataset, it also provides a strategy for reasonable predictions on out-of-distribution data. Moreover, by extracting the maximum amount of information, the Celeris One platform copes with sparse input data in a systematic manner.

1 Scantlebury, J.; Brown, N.; von Delft, F.; Deane, C. M. Dataset Augmentation Allows Deep Learning-Based Virtual Screening To Better Generalize To Unseen Target Classes