Task : Active learning in drug-target space

Elham Nour Ghassemi

1 Introduction

The identification of drug-target interactions (DTIs) is a critical step in drug development. Target proteins are large bio molecules made up of amino acids. Long string of amino acids fold in a particular way determining the shape of the proteins. A drug is small molecule in comparison to a protein. When we think of drug interaction with the protein, we think in terms of molecular level interactions, for example hydrogen bond keep molecules tightly bound to proteins. Experimental methods for drug–target interaction identification remain to be

extremely costly, time-consuming and challenging even nowadays. Alternatively, virtual screening (also refers to in silico screening) is using computational techniques to analyze large chemical databases to identify possible new drug candidates. Nowadays, machine learning plays an important role in virtual screening. The main goal that follows here is cutting expenses and duration of early drug discovery phases and extending virtual screening capabilities and accuracy. To follow the context of the problem and to find solution, first we focus on some concepts.

2 Drug-target interaction

2.1 Binding affinity

Knowing exactly which drugs or compounds will bind tightly to which receptors is key to drug discovery and development. If a molecule binds tightly and specifically to a target, it is much more likely to be a safe and effective drug. So, understanding binding affinity is a key to appreciation of the intermolecular interactions driving biological processes, structural biology, and structure-function relationships.

Binding affinity is the strength of the binding interaction between a single biomolecule (e.g. protein or DNA) to its ligand/binding partner (e.g. drug or inhibitor).

Binding affinity is typically measured and reported by the equilibrium dissociation constant (KD), which is used to evaluate and rank order strengths of bimolecular interactions. The smaller the KD value, the greater the binding affinity of the ligand for its target. The larger the KD value, the more weakly the target molecule and ligand are attracted to and bind to one another.

Binding affinity is influenced by non-covalent intermolecular interactions such as hydrogen bonding, electrostatic interactions, hydrophobic and Van der Waals forces between the two molecules. In addition, binding affinity between a ligand and its target molecule may be affected by the presence of other molecules. In drug discovery, binding affinity is used to rank hits binding to the target and design drugs that bind their targets selectively. "Selectively" means the drug must have a high affinity to the selected target and the lowest possible affinities to other targets to avoid off-target binding and caused side effects.

2.2 Virtual screening

Virtual screening is a computational technique which uses computer programs to search potential hits from virtual fragment libraries. There are two main widely used approaches for virtual screening: Ligand-based, in which 3D structure of the target is unknown, and structure-based, in which 3D structure of the target is known. Molecular docking such as AutoDock, AutoDock Vina, Vina with modern scoring functions like RF [1], is a conventional structure-based virtual screening method that optimizes the orientation of a ligand and a drug target. However, when three-dimensional structure of the drug target is not available, ligand-based virtual screening should be the first choice for drug discovery. For ligand-based virtual screening, fingerprints with high dimension of features are firstly generated to describe the chemical characteristic of ligands, and then similarity searching or machine learning approaches are applied for screening in high-dimensional data [2]. Docking simulation requires the 3D structures of the target proteins that are not always readily available. Furthermore, this is an expensive process. On the other hand, the ligand-based approaches suffer from low performance when the number of known ligands of

target proteins is small.

Beside conventional machine learning model such as support vector machine (SVM) and random forest (RF), deep learning becomes a prominent tool in Virtual screening, where it requires minimum feature engineering. I chose graph neural networks as a virtual screening tool here and lets divide this complex task to small parts and see how graph neural networks are used for the featurization of molecules.

3 Deep learning virtual screening

3.1 Encoding

In this task, input is a drug-target pair, where the drug is described using the simplified molecular-input line-entry system (SMILES) string and the target uses the amino acid sequence (see figure 1). The interactions between protein and ligands are complex, and encoding the most informative bits in a computer-readable format is one of the main challenges. In the following sections, the encoding for ligand is described followed by protein and complex encodings.

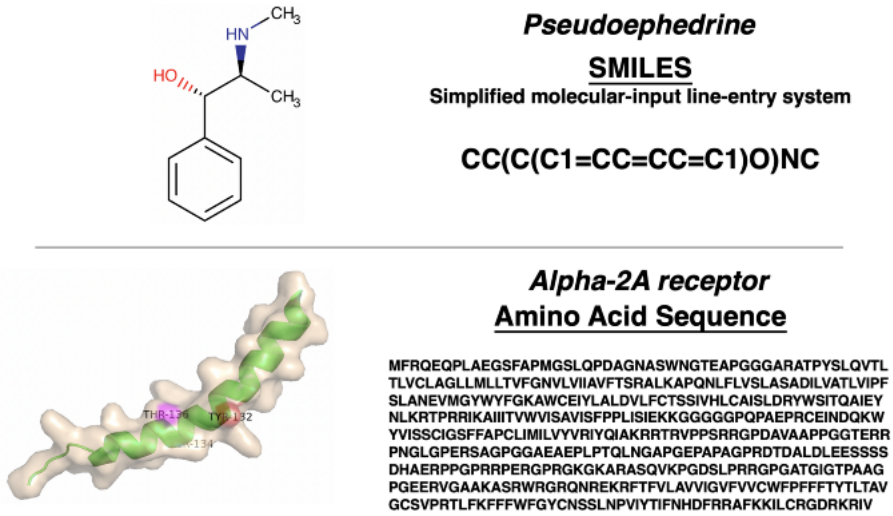

Figure 1: Drug and target protein representation [3].

3.1.1 Ligand Encodings

Chemical compounds can be described naturally in a human-readable format such as string, graph, or image in computational approaches. The most widely used string format is Simplified Molecular-Input Line-Entry System (SMILES). SMILES describes a chemical compound as a linear string. In this task, the starting point of ligand (small molecule) encodings is the molecular graph, where a chemical structure is represented as a graph. In the graph nodes and edges represent the molecular atoms and bonds, respectively. The adjacency matrix indicates the connectivity of the nodes. Each atom of a ligand is featured with

mass, total and partial charges, number of radical electrons (integers); atom type, valence, hybridization, aromaticity, chirality types (one-hot encodings), etc. Each bond of a ligand is featured with a type (single, double, triple, aromatic), ring affiliation, whether a bond is conjugated (0/1), stereo configuration (cis-/trans-, E/Z, S/R, none), direction (upright/downright) of a bond, etc. Intramolecular interactions learned by attention mechanism. The core idea of applying the attention mechanism to the graph is to obtain a context vector for the target node by focusing on its neighbors and local environment to provide expressive representations of small molecules.

3.1.2 Target Encoding

At the chemical level, a protein consists of one or more chains of smaller molecules, which we interchangeably refer to as residues for their role in the chain, or as amino acids for their chemical composition. A suitable representation for any learning task should reflect both the identity and sequence of the residues, i.e. the primary structure, and geometric information about the protein's arrangement in space. A graph-based learning can explicitly model both the sequential and geometric struc-

tures of proteins. In the simplest form, a protein can be represented as a linear graph where each amino acid in the protein structure is represented as a node, and these nodes (amino acids) are connected by edges based on the strength of noncovalent interaction between the side chains of the two amino acid residues. A node can be parameterized with such features as charge, flexibility (Smith), hydrogen bond donors/acceptors, hydrophobicity, polarity (Zimmerman), Vander-Waals volume, etc. Edges are mostly the same (peptide bonds for proteins/peptides and alternating sugar-phosphate backbone along the polynucleotide chain) and do not require parameterization.

3.1.3 Complex Encoding

The graph representation can be adapted to molecular complex. When considering a complex, the atoms from both the protein and the ligand can simply be viewed as the nodes of the graph. In this case atomic coordinates must be used. The complex interaction graph is defined as a directional graph. In this case two adjacency matrices are considered. One is constructed in such a way that it only takes into account covalent bonds. The second matrix captures bonded intramolecular

and non-bonded intermolecular interactions. It additionally considers their strength through distances. In this task I do not focus to this representation, because a protein-ligand graph is computed from the atomic coordinates in a PDB file [4].

3.1.4 Adjacency matrix in graph

Every graph (network) can be expressed mathematically in the form of an adjacency matrix. In these matrix the rows and columns are assigned to the nodes in the graph and the presence of an edge is symbolised by a numerical value. Depending on the nature of underlying edge information, different types of analysis can be performed. For this reason, it is useful to highlight the main types of edges that can be found in a graph (see the figure 2). Undirected edges : The relationship between the nodes is a simple connection and represented by a symmetric matrix containing only the values 1 and 0 to represent the presence and absence of connections, respectively. Directed edges : a set of nodes that are connected together, where all the edges are directed from one node to another, The adjacency matrix of a directed graph can be asymmetric.

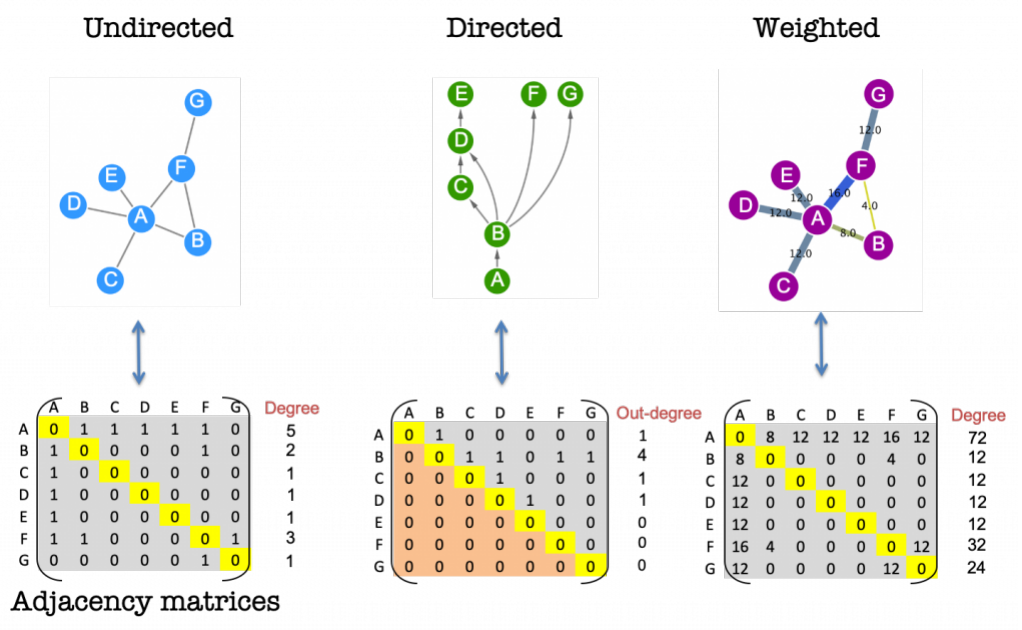

Figure 2: Graphs by edge type and their adjacency matrices

3.2 Graph neural networks

Convolutional neural networks (CNNs) are networks specialized for interacting with grid-like data, such as a 2D image. As molecules are typically not represented as 2D grids, chemists have focused on a variant of this approach: the Conv-GNN on molecular graphs. GNNs are the generalized version of CNNs where the numbers of nodes connections vary and the nodes are unordered (irregular on non-Euclidean structured data). GCN model learns the features by inspecting neighboring nodes.

3.2.1 Graph convolutional network

Graph convolutional neural (GCN) is a powerful type of neural network designed to work directly on graph and strength its structural information. GCN operates on graph to update node states. Duvenaud et al. [5] used the message passing architecture to extract valuable information from graph molecules and then transform it into a single feature vector. Neural message passing networks use a convolutional layer to exchange information between atoms or bonds within a molecule and produce a fixed-length, real-valued vector that embeds the molecular information. This layer has a weight matrix. First, these networks compute a feature vector for each atom within the molecule. These feature vectors are then collected into a matrix. Then, they generate a graph adjacency matrix that specifies the connectivity of the graph. In a forward convolutional pass, these matrices (weight, feature, and adjacency matrices) are multiplied together. This allows information to be exchanged between the feature vectors of each atom with its immediate neighbors, in accordance with the connectivity specified by the adjacency matrix. This updates each atom's feature vector to include information about its local environment. This updated feature vector

matrix is then passed through an activation function (i.e., ReLU) and can then be iteratively updated by using it as the feature matrix in another convolutional pass. This propagates information throughout the molecule. Finally, these atom feature vectors are either summed or concatenated to give a unique, learned representation of the molecule as a real-valued vector. The Conv-GNNs apply different weights at each iteration [6].

This latent node representation (feature vector that contains information about its neighborhood), often called an embedding, can be of any chosen length. It is then used as the input for a traditional fully connected Deep Neural Network(DNN) to finally make the classification or prediction. Backpropagation is once again used to train these networks by propagating gradients backward and determining how to change the convolution matrix weights and the parameters in the DNN. The figure 3 shows a GNN that the graphs of molecule and protein pass through two GCNs to get their representations. Then the affinity can be predicted after multiple fully connected layers [7].

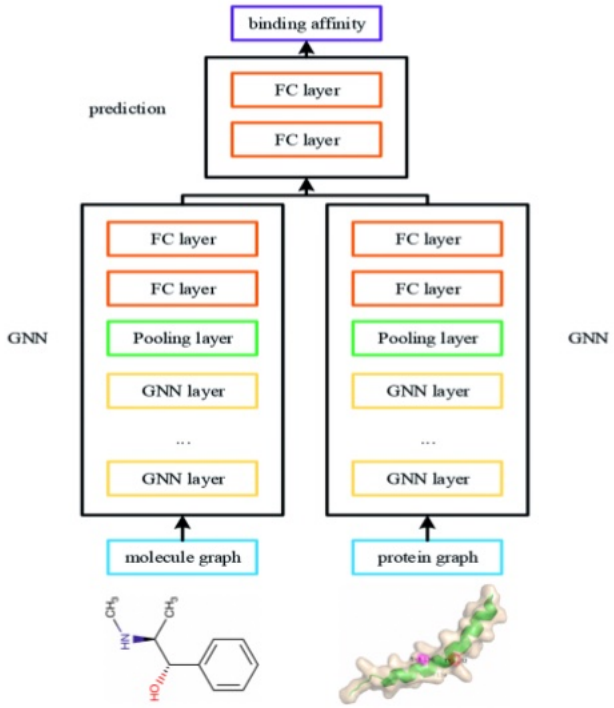

Figure 3: An encoder-decoder framework.

3.3 Workflow

First, load the dataset from a local file or load a benchmark dataset (KIBA/DAVIS), specify the compound and protein encoders. The next step is splitting the dataset into training, validation and testing sets. To evaluate the model, we can choose part of data as out of distribution of the train data set. This data set can evaluate whether our model can extrapolate to the other chemical space (for example see figure 4). Then, initialize a model using the configuration file. Train the model using train function and monitor the progress of training and performance metrics.

4 Summary

Predicting drug-target interactions is crucial for novel drug discovery,but it is challenging. Deep learning (GCN) is proposed and features are learned by the models. Graphs encoding (that describe the connectivity, bonded and non-bonded, between the atoms) seems to capture well the variety of information important for ligand-binding.

To generalize the trained model to other chemical space, we divided

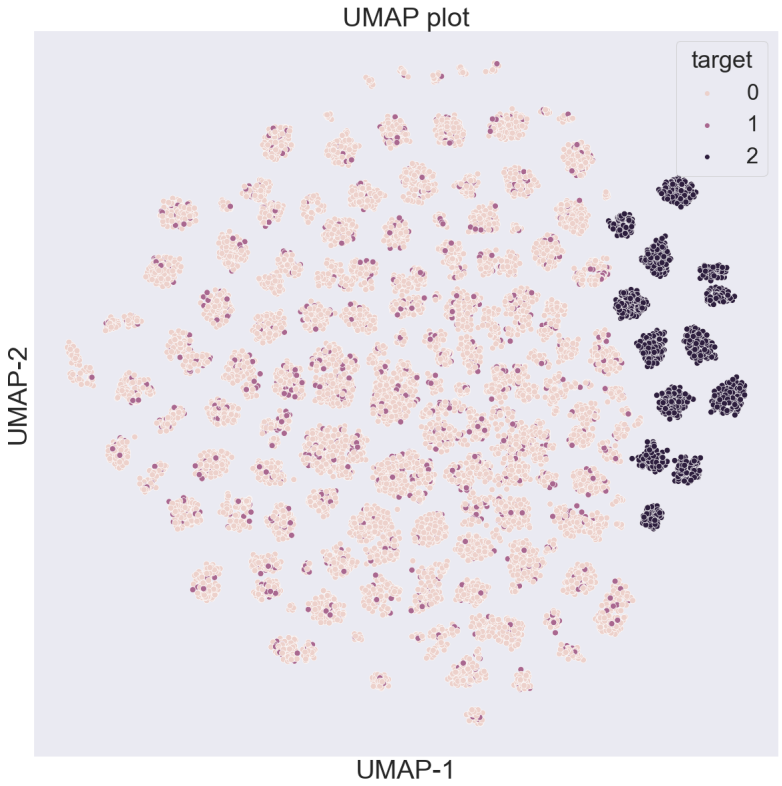

Figure 4: Chemcal space for train (0), test1 (1) and test2 (2) data sets. Test1 has the same distribution of train data set, test2 dataset is outside of train dataset chemical space.

available dataset in three parts, two of the datasets (train and test1) are in the same chemical space and test2 is in another one where train date set has never seen this space. Evaluation of trained model with test2 dataset and obtain high accuracy R^2 score show that our model can be extrapolated to the external dataset or chemical space.

Deep learning is very data greedy and usually, the bigger the training set is, the better the results. However, biochemical data are still considerably smaller than, for example, image or video data sets. DeepAtom translates and rotates the protein-ligand complex to gain more training data [8]. In QSAR predictions, using SMILES augmentation has also become popular as means to enlarge the training set [9]. Also, using ensemble methods to construct a set of models, then to aggregate them into a single model can improve the performance of models. For example, the ensemble of ML models such as RF and deep learning (GCN) model may improve the accuracy of the model for the available datasets.

References

[1] Meng X. Y., Zhang H. X., Mezei M., and Cui M., **Molecular docking: a powerful approach for structure-based drug discovery**, Curr Comput Aided Drug. 7(2):146-57 (2011).

[2] Dai W., and Guo D. A., **Ligand-Based Virtual Screening Method Using Direct Quantification of Generalization Ability** Molecules, 24(13), 2414 (2019).

[3] Huang K., Fu T., Glass L. M., Zitnik. M, Xiao C., and Sun J., **DeepPurpose: a deep learning library for drug–target interaction prediction** Bioinformatics, Vol. 36, Issue 22-23, PP5545–5547 (2020).

[4] Gomes J., Ramsundar B., Feinberg E. N., and Pande V. S., **Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity** arXiv:1703.10603, (2017).

[5] Duvenaud D., Maclaurin D., Aguilera-Iparraguirre J., Gómez-Bombarelli R., Hirzel T., Aspuru-Guzik A., and Adams R. P.,

Convolutional Networks on Graphs for Learning Molecular Fingerprints arXiv:1509.09292 (2015).

[6] Gurbych A., **Graph Neural Networks for Binding Affinity Prediction** towards data science.

[7] Jiang M., Li Z., Zhang S., Wang S., Wang X., Yuana Q. and Weia Z., **Drug–target affinity prediction using graph neural network and contact maps**, RSC Adv., 10, PP 20701-20712 (2020).

[8] Li Y., Rezaei M.A., Li C., Li X. **DeepAtom: A Framework for Protein-Ligand Binding Affinity Prediction** IEEE International Conference on Bioinformatics and Biomedicine (BIBM); San Diego, CA, USA. 18–21 November (2019).

[9] Bjerrum E.J., **SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules** arXiv. 20171703.07076.