**Drug to target activity predication analysis using ML**

**Introduction**

In recent years, the average price-tag for getting a new drug to market has risen to about $2.5 billion, with an estimated delivery date of 10-15 years. The Merck Kaggle challenge on chemical compound activity was won by Hinton’s group with deep networks in 2012. This indicates the high potential of deep learning in drug design and attracted the attention of big pharma.

Only about 108 substances have ever been synthesized, [7] whereas the commonly reported range of potential drug-like molecules is 1023 - 1060 [8]. The opportunity is huge and the opportunity resides in initial drug screening process.

The first step of a drug design pipeline is to identify a biomolecular target upon which a potential

drug can act, e.g. a protein whose activity can be modified by a compound to achieve a beneficial

therapeutic effect. The next step is to screen tens of thousands of chemical compounds by biological

High-throughput assays for interactions with this target — typically measured via IC50 or EC50 values. Finally, a target-interacting lead compound is selected.

Virtual screening is an in silico predictions of compound-target interactions. Our research is focusing on compound to target binding prediction. The Protein binding has an extremely important role in biochemistry. Protein binding is often reversible and can be stable or unstable depend on the structures and the activities of it. Furthermore, that is also a reason why protein binding can be influence the drug's biological half life in our body.

Our goal is to identify candidate to begin testing. Our research subjects include

1. Formally, the task of target prediction presents itself as follows: given a chemical compound i, we want to predict whether the compound is active on a target t.
2. In order to identify compounds for interactions with the primary targets. We would need to compare each algorithm vs. DL algorithm.
3. Off-target analysis. Typically, chemical compounds interact with more than one protein, and most of these interactions result in unwanted side-effects. We would like to show the drug candidate interaction with other targets.
4. Identify features of drug compounds. Feature extraction is the key to DL.
5. Compare different features affect the virtual screening modeling.
6. Extraction the common features associated with the primary target.

Encouraged by the Merck Kaggle challenge on chemical compound activity won by Hinton’s

group with deep networks. Referring the Merck Kaggle Challenge to develop medicine [[1]](https://www.kaggle.com/content/kaggle/img/casestudies/Kaggle%20Case%20Study-Merck.pdf) and winner using DL networks [[2]](http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html?_r=2&)

We set down the path to learn drug to target prediction using public datasets since we don’t have in-house data of pharmaceutical companies.

1. **Datasets**

Journey starts with step-by-step tasks. First step is to extract public datasets as our base training sets on drug to target interactions.

Luckily there are publicly available drug activity database like ChEMBL which is a large-scale bioactivity database for drug discovery. ChEMBL [[3]](https://www.ebi.ac.uk/chembl/) has 1.6M distinct compounds, 11K targets, compared to the Kaggle dataset with 11k descriptors (??), 164k compounds and 15 drug targets.

[Qs] What is the descriptor inside ChEMBL database?

The ChEMBL database schema is downloadable at [[4]](ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_22_1_schema.png). Also the detailed documentation of ChEBML database schema is downloadable at [[5]](ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_22_1_schema_documentation.txt). From the database schema, we are mostly interested in “ACTIVITIES” table, which captures the drug to target activities’ results recorded in a scientific document. Each activity is described by a row. E.g. The column of “STANDARD\_TYPE” of type VARCHAR2(250) stands for Standardised version of the published\_activity\_type (e.g. IC50 rather than Ic-50/Ic50/ic50/ic-50)

Also the column “PCHEMBL\_VALUE” of type NUMBER(4,2) stands for Negative log of selected concentration-response activity values (IC50/EC50/XC50/AC50/Ki/Kd/Potency). The unit of measure is nM. In our current study, we would use all types the same, since they reflect a concentration that is associated with drug binding or activity. The lower the number the better, meaning more potent, cutoff typically 0.1 micro molar or lower considering being good.

Some potential relevant information regarding binding efficiency is yet to be explored. In table “LIGAND\_EFF”, it contains BEI (Binding Efficiency Index) and SEI (Surface Binding Efficiency Index) for each activity\_id where such data can be calculated. Ligand efficiency data is yet to be examed, such as

Binding efficiency index (BEI), Surface Efficiency index (SEI), and Ligand efficiency (LE), HEAVY\_ATOMS asNumber of heavy (non-hydrogen) atoms

[Open Discussion] How to utilize the potential relevant information? How to calculate delta G?

So far, we have extracted all activities data in ChEMBL database of compounds associated with 2 targets and generated 2 target datasets. Our target datasets are coming from ChEMBL database extraction including target\_name, target\_type, compound\_name, compound\_key, canonical\_smiles, measurement\_value, measurement\_type , ligand\_eff and heavy\_atoms\_count. Later we will use those data values in our DL analysis. The database SQL extraction is referred at [6]

|  |  |  |
| --- | --- | --- |
| Target Name | Number of Active drug compounds | Drug to Target Activity value  (IC50/EC50/Ki in nM units) |
| Dopamine D2 receptor | 7515 |  |
| Beta-secretase 1 | 7483 |  |

[Open Item 1] We have generated over 11k datasets. The datasets are yet to be analyzed.

[Open Item 2] We would need to calculate delta G as other sets of measurement value. Currently we only use EC50/IC50/Ki as measurement value published in ChEMBL database.

1. **Data Experiment using FingerPrints**

As for target prediction, given a chemical compound, we want to predict whether the compound is active on a target t. We encode this information in the binary array for each compound based on feature extraction methods. The output of data experiments produce true (or 1) if the compound is active on a target and false (or 0) otherwise. We are also interested in predicting the behavior of a compound on off targets at the same time, the off-targets range is in the thousands.

Each compound is represented using a number of binary features described later in this section.

* 1. **Types of Molecular FingerPrints**

1. ECFP and ECFC

The de facto standard circular fingerprints are the Extended-Connectivity Fingerprints (ECFPs), based on the Morgan algorithm [26], which were specifically designed for their use in structure-activity modeling [27]. They represent circular atom neighborhoods and produce fingerprints of variable length. They are most commonly used with a diameter of 4 and referred to as ECFP4. A diameter of 6 (ECFP6) is also commonly used, although some benchmarks have shown small performance differences between the two [28]. Additionally, there is a variation that keeps track of the frequency counts of the ECFP features, recording each identifier as many times as it appears in the molecule instead of only once. This variation is often denoted as ECFC. Notable software programs that provide these fingerprints are Pipeline Pilot [29], Chemaxon's JChem [30], the CDK [20] and the RDKit [31] (referred to as “Morgan fingerprints”).

The following figures illustrate the whole ECFP generation process, including the derivation of the fixed length bit string from the identifier list representation. [29]

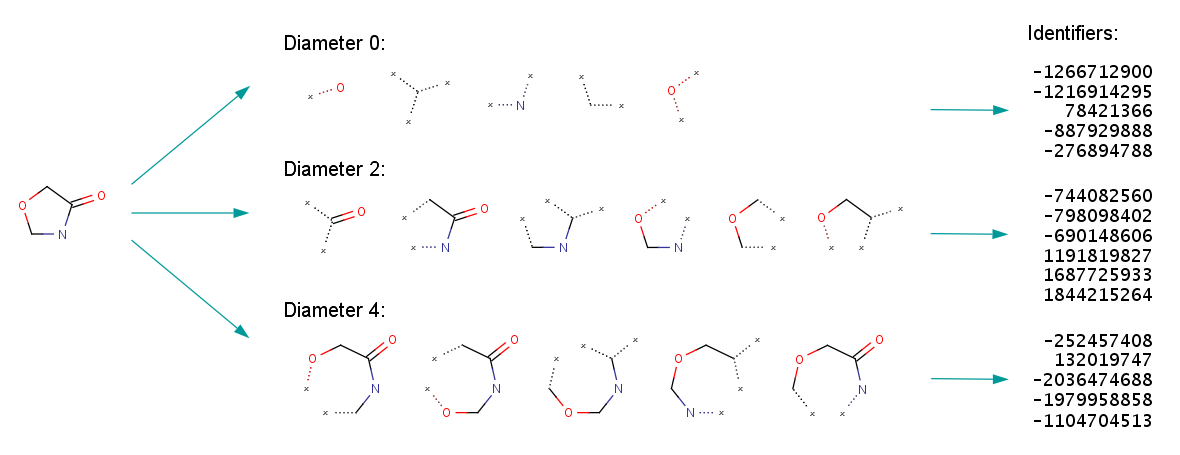


Fig. 2. ECFP generation process

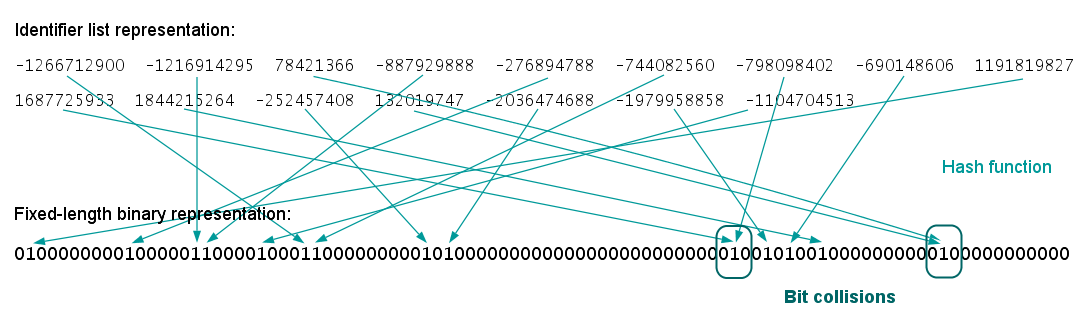


Fig. 3. Generation of the fixed-length bit string ("folding")

We are using the latest version of the popular open source [Chemical Development Kit](http://cdk.sourceforge.net/) (CDK) [20] for highly regarded ECFP of chemical structure fingerprints.

Using ECFP as compound representation is by means of lists of integer identifiers. Each identifier represents a particular substructure, more precisely, a circular atom neighborhood, which is present in the molecule. The substructures are represented in SMARTS format.

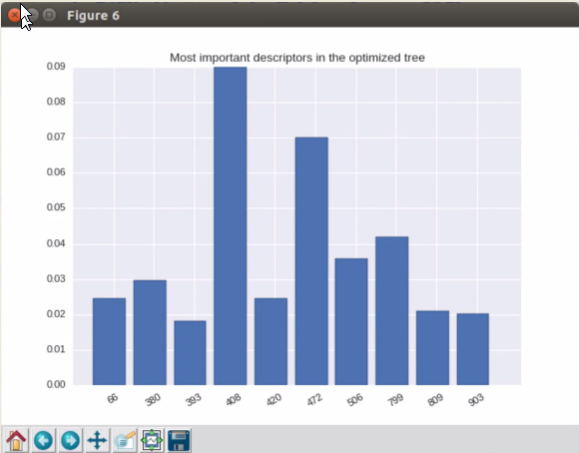
ECFP of compound for smiles format, take an example of one compound in ChEMBL database which interacts with “Dopamine D2 Receptor”.

A Smiles format of compound as "Fc1ccc(cc1)C(=O)CCCN2CCN(CC2)c3ccccn3"

The ECFP generated by CDK is as

"0000000010001010000100000000000000100000010000000000000000000000001000000000000010000000000000011000000000000000000000000000000000100100000000000000000010000000000000000000001000000000000000100000000000000000000000000000000000000000000000000000000000000100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010100000000000000100100000000000100010000000011000001000000000000000000000100000000000000000000000000000000100000000000100000001000000000000000000001010000000000000100000100000000000000000001000000000000100000000000000001000010000000000000000000100000000000000000000000000000100000000000000000000000000000000001000000001000000000000000000000010000000100000000000000001000001000000000000000000000000000000000000000000001000000000000000000000000100000000000000000000000000000000010000000010000000000000000000000000001000001000010000000000010100000000100000000000000000000001000000000000000000000000000000000000100000000000000000000000001000000000000000000000000000000"

ECFP contains the fixed-length bit string associated with identifiers. We use compounds-to-protein activity interaction datasets from ChEMBL database, then generate ECFP to start training. After the training, we get the most important descriptors as bit string’s index number as shown below. For example, we get the index of 408 472 and 799 as the most significant identifier descriptors associated with target “Dopamine D2 Receptor”.



To understand what the bit string’s index number (e.g. 408 472 and 799) means in terms of identifiers and chemical substructure. We would need to find a way to convert back from ECFP bit string’s index number to chemical substructure representation. In CDK, there is no direct conversion from ECFP bit string’s index to chemical substructure. We had to modify CDK code to accomplish the conversion.

From identified number of 408 index, we found each index is associated with multiple substructures in the form of SMARTS format. The SMARTS format is readable using [32]

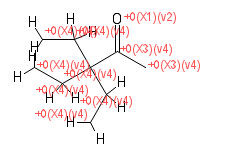
For example,

SMARTS format as [nH0v3X3+0][cH0v4X3+0]([cH1v4X3+0][cH0v4X3+0]([cH1v4X3+0])[OH0v2X2+0])[cH0v4X3+0] has 1 occurrence

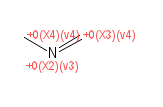
[cH1v4X3+0][cH0v4X3+0]([cH1v4X3+0][cH0v4X3+0]([cH1v4X3+0])[CH1v4X4+0])[CH0v4X2+0] has 2 occurrence

[CH1v4X4+0][cH0v4X3+0]1[cH1v4X3+0][cH0v4X3+0]([cH0v4X3+0]([cH1v4X3+0][cH0v4X3+0]1[CH2v4X4+0][CH2v4X4+0])[OH1v2X2+0])[OH1v2X2+0] has 14 occurrences

[CH2v4X4+0][CH2v4X4+0][CH0v4X4+0]([CH2v4X4+0][CH2v4X4+0])([CH2v4X4+0][CH2v4X4+0])[CH0v4X3+0]([cH0v4X3+0])=[OH0v2X1+0] has 6 occurrences



[CH0v4X3+0]=[NH0v3X2+0][CH0v4X4+0] has 963 occurrences



The atoms SMARTS chemical substructure is reference at smarts\_test3.csv

[Open Item 3] Since we don’t have enough knowledge of SMARTS format, we would need to get help with understanding the significance of chemical substructure.

1. Modified ECFP

In ECFP fingerprints’ generation, each identified index is associated with many chemical substructures SMARTS. This approach presents difficulty to identify most significant substructure associated each chemical compounds.

Therefore we modified to produce hybrid ECFP fingerprints. The process involves initial filtering process of identifying most popular substructures associated with targets. We first identify the X number of most popular substructures associated with certain targets, with X being configurable, for example, 1024.

CDK helps us break down the compounds into substructures by given target(s), then we extract the atomic information and map to the substructures from the process. We sort the substructures in descending order of number of occurrences, and cut off the number of substructures based on X configuration. Finally we could generate ECFP bit string based on the index of substructures with most popular vote.

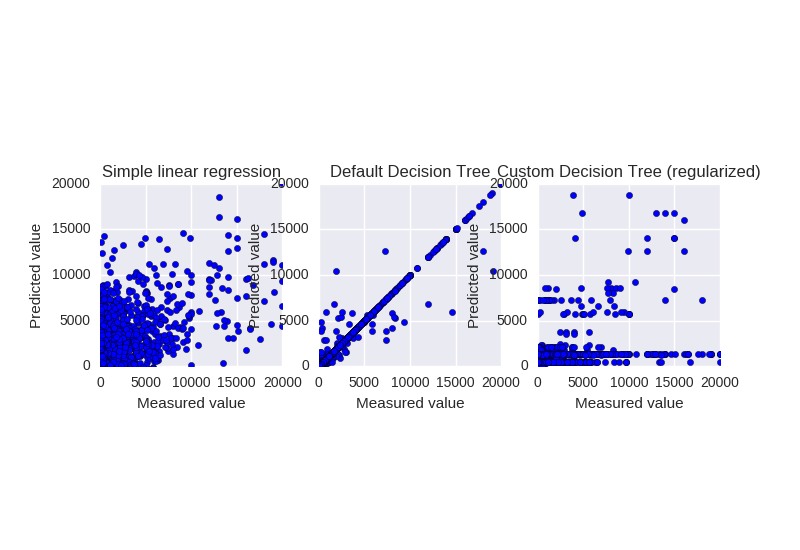
With the hybrid ECFP, we could uniquely identify the most significant substructure SMARTS format after training.

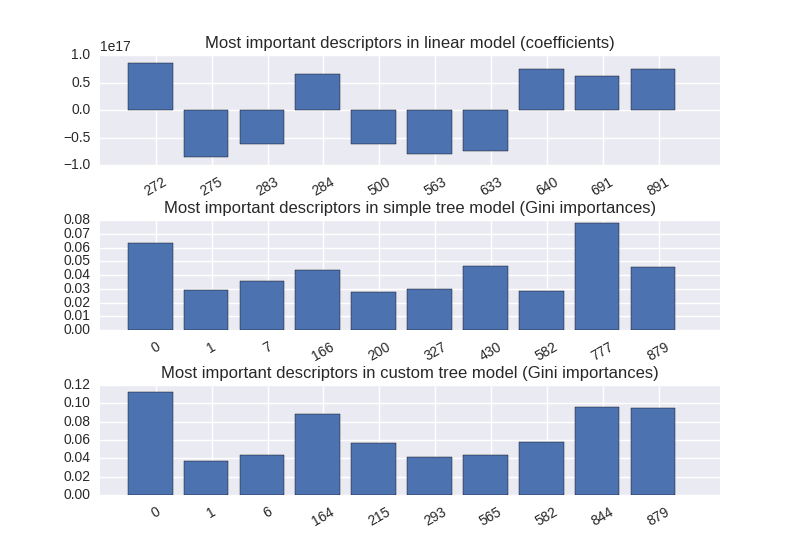
* 1. **ML Packages**

We used scikit-learn [33] as ML packages in our current study.

1. **Algorithm Experiments**

Algorithm we are using

1. Simple Linear regression
2. default decision tree regression
3. custom decision tree regression
4. Logistic Regression 



We try to predict measurement values through ECFP fingerprints. We tried several ML algorithms and found there are no linear relationships between most significant substructure SMARTS with measurement values (EC50/IC50/Ki)*.*

[Open Item 4] We are still experimenting using other algorithms. We plan to use Deep Network, SVM, BKD, k-NN, Pipeline Pilot Bayesian Classifier, Parzen-Rosenblatt, SEA other than Logistic Regression

1. **References**

[1] https://www.kaggle.com/content/kaggle/img/casestudies/Kaggle%20Case%20Study-Merck.pdf

[2] http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html?\_r=2&

[3] <https://www.ebi.ac.uk/chembl/>

[4] <ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_22_1_schema.png>

[5] <ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_22_1_schema_documentation.txt>

[6] SELECT 'target\_name', 'target\_type', 'compound\_name', 'compound\_key', 'canonical\_smiles', 'measurement\_value', 'measurement\_type', 'ligand\_eff', 'heavy\_atoms\_count' UNION ALL SELECT td.pref\_name AS target\_name, td.target\_type, cr.compound\_name, cr.compound\_key, s.canonical\_smiles, act.standard\_value, act.standard\_type, leff.le AS ligand\_eff, p.heavy\_atoms FROM target\_dictionary td, compound\_records cr, compound\_structures s, activities act, assays a, ligand\_eff leff, compound\_properties p WHERE s.molregno = cr.molregno AND cr.molregno = p.molregno AND cr.record\_id = act.record\_id AND act.assay\_id = a.assay\_id AND act.activity\_id = leff.activity\_id AND (act.standard\_type = 'IC50' OR act.standard\_type = 'EC50' OR act.standard\_type = 'ED50' OR act.standard\_type = 'Ki') AND act.standard\_units = 'nM' AND a.tid = td.tid AND td.chembl\_id = 'CHEMBL217' INTO OUTFILE '/tmp/compounds\_matching\_CHEMBL217\_limited.csv' FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n';

[7] Kim, S. et al. Pubchem substance and compound databases. Nucleic Acids Res. 44,

1202{1213 (2016).

[8] Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like

chemical space based on gdb-17 data. J. Comput.-Aided Mol. Des. 27, 675{679 (2013).

[9] https://www.researchgate.net/publication/264866258

[26] H.L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A

Technique Developed at Chemical Abstracts Service., J. Chem. Doc. 5 (1965-5) 107–113.

doi:10.1021/c160017a018.

[27] D. Rogers, M. Hahn, Extended-connectivity fingerprints., J. Chem. Inf. Model. 50 (2010-5)

742–54. doi:10.1021/ci100050t.

[28] S. Riniker, G.A. Landrum, Open-source platform to benchmark fingerprints for ligand-based

virtual screening., J. Cheminform. 5 (2013-1) 26. doi:10.1186/1758-2946-5-26.

[20] <http://cdk.sourceforge.net/>

[29] <https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP#ExtendedConnectivityFingerprintECFP-morganref>

[30] ChemAxon – cheminformatics platforms and desktop applications,

(https://www.chemaxon.com/, accessed on 06/18/2014).

[31] G. Landrum, RDKit: Open-source cheminformatics, (http://www.rdkit.org, accessed on ).

[32] <https://pubchem.ncbi.nlm.nih.gov/edit2/index.html>

[33] http://scikit-learn.org/stable/auto\_examples/index.html