Drug Target protein interaction prediction by PSSM and LOOP method

**Team Details:**

C. Vijayabaskar 2017103048

G. Velmurugan 2017103047

K. Sathish Kumar 2017103588

**Guided by**

G. Sudhakaran

Teaching Fellow

Department of Computer Science and Engineering

College Of Engineering Guindy

Anna University

**Problem Statement:**

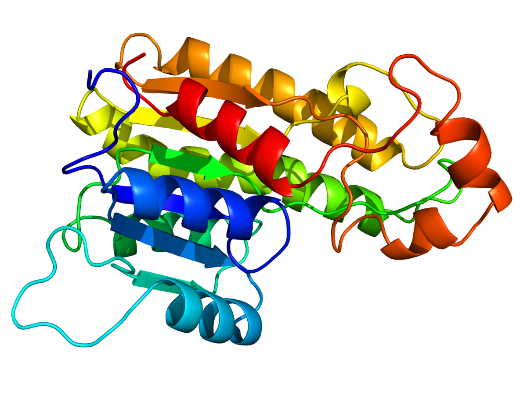
Predict the interaction between the drug and target protein with the sequence information of the protein and drug fingerprint.

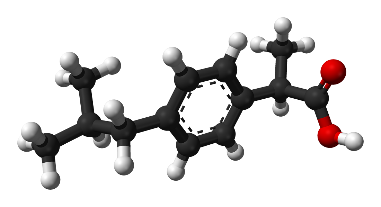
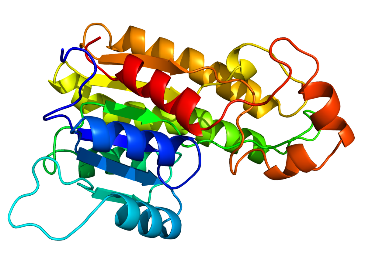
**Introduction:**

Drug target interaction is a prominent research area in the field of drug discovery. It refers to the recognition of interactions between chemical compounds and the protein targets. Wet lab experiments to identify these interactions are expensive as well as time consuming. Drug-Target Interaction Prediction (DTI) is an important application of machine learning in medicine industry, the importance is coming from the fact that we need to save the time and cost of the drugs development. The prediction of interactions between drugs and target proteins is a critical part of drug discovery pipeline as it can help and a novel drug candidate and understand side effects. Although modern medicine is aligned with antibiotic treatment, the discovery of new and potential drugs is declining, as there is an increase of the misuse of the current available medicine, causing a resistance effect to these kinds of agents. Large numbers of DTIs have been uncovered in databases such as DrugBank, Matador, and CTD, but many DTIs remain to be discovered. These public databases store a number of known drug-target interactions which validate through experimental. This also provides a good basis for researchers to develop novel computational methods to predict DTIs.

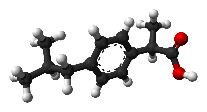
**Drug** is a chemical substance, typically of known structure, which, when administered to a living organism, produces a biological effect. A pharmaceutical drug, also called a medication or medicine, is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being. The most widely used drugs in the world include caffeine, nicotine and alcohol, etc. It interacts with protein (targets) in the body to change their physiological function this can often be noticed through a protein’s phenotype (outward expression of protein). Each drug can have many targets some of which can produce unwanted changes (side effects) and each protein can be targeted by many different drugs (many to many relationship).

**Target proteins** are functional biomolecules that are addressed and controlled by biologically active compounds. They are used in the processes of transduction, transformation and conjugation. They are used by the body to build and repair the body e.g. enzymes which are used to speed up chemical reactions such as insulin or antibodies which are used to fight off foreign particles in the body such as viruses. Proteins are made up of chained combinations of amino acids and do not have a fixed length, the amino acids come from the amino acid alphabet of length 20 but proteins can have any length e.g. 678, each individual protein has a unique 3D structure. Not all proteins will interact with drugs, a proteins affinity to binding with a drug is known as it’s druggability, it is largely believed that if a protein is druggable then the proteins that belong to the same protein family as the protein are also druggable, it largely believed that only around 5% of all the proteins in our bodies are druggable, hence why there is currently an upward trend in biotech drugs and genome editing. Proteins always interact with other proteins so even though a protein may not be druggable, a protein that is druggable could have far reaching consequences by interacting with proteins that are not directly druggable.





+ =



Drug Target DTIs

Fig 1: Visual representation of Drug Target Interaction

Computational methods for DTI prediction are divided into 3 main approaches:

1. ligand based
2. docking simulation
3. chemogenomic

Ligand based approaches are built upon the concept that similar molecules have similar properties and therefore should bind to the same group of proteins. Docking Simulation approaches are used for structure based drug design, where the interaction between a protein and a drug is simulated and scored, according to the intermolecular interaction energy, using 3D structures. Chemogenomic approaches are based on the chemical space of compounds, genomic space of target proteins and/or the pharmacological space (interactions between proteins and drugs) to predict new potential interactions.

Computational prediction of drug-target interactions (DTI) is vital for drug discovery. The experimental identification of interactions between drugs and target proteins is very onerous. Modern technologies have mitigated the problem, leveraging the development of new drugs. However, drug development remains extremely expensive and time consuming. Many machine learning approaches have been proposed over the years for DTI prediction. Nevertheless, prediction accuracy and efficiency are persisting problems that still need to be tackled. Here, we propose a new learning method which addresses DTI prediction which uses chemogenomic way of predicting.

**Summary of related work:**

Drug-target networks are receiving a lot of attention in late years, given its relevance for pharmaceutical innovation and drug lead discovery. Nidhi, Meir Glick, John W. Davies, and Jeremy L. Jenkins in 2006 proposed a Multiple-Category Bayesian Models using chemogenomics database. Useful for improving knowledge in chemogenomics databases and for predicting new targets for orphan compounds. The automated nature of multiple-category naive Bayesian models trained on the knowledge in chemogenomics databases lends itself to the creation of very large predicted chemogenomics databases that would enable additional data mining.

Yong Liu ,Min Wu,Chunyan Miao,Peilin Zhao,Xiao-Li Li proposed Neighbourhood Regularized Logistic Matrix Factorization. It is based on the drug similarities and target similarities is utilized to further improve the prediction ability of the model. But it lacks couple logistic matrix factorization with the multiple kernel learning techniques leads to low accuracy. Wang’s method of computational model is developed from (PSSM) Post Specific Scoring Matrix gives variable importance for drug testing results which helps in determining the variable which impacts positively in prediction. But it lacks the knowledge about the unknown drug target interactions hence no improvement in inferring unknown drug-target interaction. Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, Jie Zheng gives a Intuitive solution to the new candidate problem of (BLM) Bipartite Local Model by integrating a Neighbour-based Interaction-profile Inferring (NII) procedure. In this, the influence of neighbours based on their distances can be used in finding new drug/target. The major drawback is the deterioration in the prediction performance and limited exploration in local and global information in model learning.

Hailin Chen,Zuping Zhang in 2013 developed a Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks. The labeled and unlabeled interaction information of drugs and targets and developed a semi supervised model. In this, predicting interactions for new drug compounds and predicting unknown interactions of the given network are the advantages. Data incompleteness is another big issue for such prediction problem. Thus, the performance of our method could be further improved by integrating more verified drug-target interactions. Zhu-Hong You,Li-Ping Li Xin Yan in 2019 proposed a new silico approach, named DTIRF, to predict the DTI combine feature weighted Rotation Forest (FwRF) classifier with protein amino acids information. In this, the feature extraction is very effective, and the designed classifier has high recognition performance than other classification algorithms. But it executed less number of feature extraction algorithm to predict DTI. WenZhangab,Yanlin,Chenc,Dingfang,LicXiang,Yue developed a novel computational method named “Manifold Regularized Matrix Factorization” (MRMF) to predict potential drug-drug interactions (DDIs). MRMF model achieves high-accuracy performance on the benchmark dataset, and outperforms other state-of-the-art methods. But there is no accurate methods on how to combine diverse features in a manifold regulation. Wen Zhang, Yanlin Chen and Dingfang Li proposed a label Propagation method with Linear Neighbourhood Information (LPLNI) for predicting unobserved drug-target interactions. Experimental results demonstrate that integrated information (LPLNI-II) can produce improved performances, better than other state-of-the-art methods. This experiment doesn’t show how to utilize the unknown data for drug-target interaction.

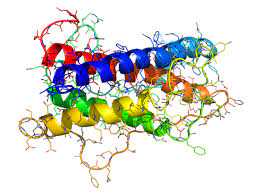
In our proposed system a novel computational approach is proposed by combining: Drug molecular Substructure Fingerprint, Local optimal oriented pattern (LOOP), Position specific scoring matrix (PSSM), Rotation forest (RF) classifier for predicting potential drug-target interactions. **Drug molecular Substructure fingerprint** records the existence of substructure after separating the drug molecules into fragments. Moreover, the structural properties of the drug molecules were encoded in binary bits, which can directly know whether specific substructure fragments in the drug molecules exist or not. Given a specific drug molecular, the vector is set to be 1 if the substructure is present or the vector is set to be 0. **Position-Specific Scoring Matrix**, is a type of scoring matrix used in protein BLAST searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment. PSSM scores are generally shown as positive or negative integers. Positive scores indicate that the given amino acid substitution occurs more frequently in the alignment than expected by chance, while negative scores indicate that the substitution occurs less frequently than expected. Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) which can search and compare the homologous sequence of each target protein sequence is adopted to create PSSM of each target protein sequence. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins. The **Local Optimal Oriented Pattern** (LOOP) is a texture descriptor which encodes repeated local patterns in images as binary codes, and it is a popular type of feature used for classification in computer vision. Because of the disadvantage of local binary pattern (LBP) and local derivative pattern (LDP) is the arbitrary sequence of binarization weights that adds dependency to orientation. Thus, LOOP presents a nonlinear amalgamation of LBP and LDP that overcomes these drawbacks while preserving these strengths. It integrates the strength of two texture descriptors LDP and LBP for assigning weights and binding the intensity differences. In LOOP algorithm, The LOOP feature is obtained by calculating for each image pixel using a 3x3 neighbourhood around each pixel. Encodes rotation invariance into the main formulation itself. This makes any post processing stage for rotation invariance redundant and improves on both accuracy and time complexity. The **Rotation Forest algorithm** focuses on improving the difference and accuracy of the base classifier. RF algorithm focuses on improving the difference and accuracy of the base classier. In this work, we adopt RF as a classification model for predicting DTIs. Specifically, the RF randomly divides the entire sample set into K subsets, and a principal component analysis (PCA) method is adopted to transform the subsets which make the difference between each subset. Finally, the prediction score is obtained after training different base classifiers.

**Architecture diagram:**

Colour code:

* Existing Algorithm without modification
* Existing Algorithm with modification
* Our Contribution

**Protein Sequence**



**Gapped BLAST search**

**Conversion to low-complexity sequence**

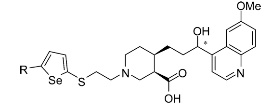
r

**2D Matrix (PSSM)**

**Alignment**

PSI BLAST

PSSM Matrix conversion

****

**Non-Linear amalgamation of LBP & LDP**

**Encoding**

**Drug Molecular structure**

**Fingerprint vector Conversion**

**Feature description**

Drug Target Vectorization

LOOP representation

**Rotation Forest Classification**

**Ligand based drug feature**

Fig 2: Complete Architecture Diagram

**Details of modules:**

1. **PSSM Matrix Conversion**

A PSSM, or Position-Specific Scoring Matrix, is a type of scoring matrix used in protein BLAST searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment.

As a modification we first align and do PSI-Blast which is more efficient.

**Input**: protein sequence

e.g:MAESASPPSSSAAAPAAEPGVTTEQPGPRSPPSSPPGLEEPLDGADPHVPHPDLAPIAFFCLRQTTSPRNWCIKMVCNPWFECVSMLVILLNCVTLGMYQPCDDMDCLSDRCKILQVFDD

Each letter represents an amino acid. Protein is generally a sequence of amino acid.

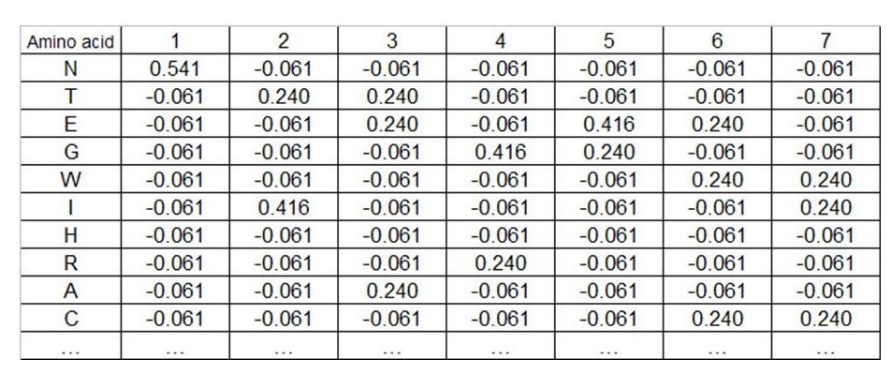
**Output:** 2D Matrix

Fig 3: example of PSSM matrix

A matrix in which the row represents the length of the protein sequence and column is 20 (Number of available amino acids). The value contains the frequency of that amino acid at that position in that sequence.

**Algorithm:**

1. Input protein sequence
2. Run PSI-Blast with 3 iteration
3. Normalize each value in range 0-1
4. Sum up all rows to same amino acid in primary sequence
5. Resultant matrix is 20x20

**2D Matrix**

**PSI BLAST search**

**Alignment**

**Protein Sequence**

Fig 4: Architecture diagram of PSSM conversion

1. **Fingerprint Vectorization**

Molecular fingerprints are a way to represent molecules as mathematical objects. By doing this, we can perform statistical analyses and/or machine learning techniques on the set of molecules to gain new insights that we could not gain as humans. One of the most common molecular fingerprinting methods is **Extended Connectivity Finger Printing (ECFP).**

As a modification we use Morgan fingerprint (2048 bits) rather than PubChem (881 bits) based fingerprint which is proposed in our paper.

**Input:** SMILES (Chemical formula)

**e.g:** N[C@@H](CC(O)=O)C(O)=O

**Output: Morgan Fingerprint (2048 bit vector)**

**01000000000000000000000000000000000000000100000000000000000000000000000000000001100000000000000000000000000010000000010000000000000000000000100100000000000000000000000000000100000000000000000001000100000000000000000000000000000000000000000000100000000000000000000000000100000000000001000000000010000000000000001000000000000000000000000000000000000000000010000000000000000000001000000000000100000000000000000000000000000000000000000000000000000000100000000000000000000000000000000000000000100000000000000000000000000000000000000000010000000000000000000000000000000000000000010000000000000000000000000000000000000000000000000110000000000000000000000000100000000000000001000000000000100000000100000000000000000010000000000000000000000000001001000001000000000000010000000000000000000000000010000000000000000000010000000000000000000000010000010000000000010000000100000000000000000100000000000000000010000000000000000000000000000000110000000100000000000000000101000000000000000000000000000100000000000000000000000000000000010100000000100000000000000000000000000001000000001000000000000000000000100000000000000010010010000000100000000000000000000001000000000010000000000000000001000000000000000000000000000100000000000000000000000000000000000000000000000000000000000000000000000000000000000000010010000000000000000000000100000000000100000000000000000000000000000000000000000000000000000010000000000001000000000000000000000000000010000000000100000010000000000100000000000000000000000000001010000000000000000000000000000000000000000000000010001000000000100000000000000000001000000000010000000000000000000000000010000000000000101000001000000000000000000000000000000000000000000000000000000000000000100000000000000000000000000100010000000000000000010000000000001000110000000000000000000000100000010000000000000000000000000000000000000100000000000000000000100101000000000000000000000001001000000000000000000001100000000000010000010000000000000000000000000000000000000000010000000000000000000000000000000000000000010000100000000010001000000000000000000000000000**

**Each set bit represent a molecular compound.**

**Algorithm:**

1. Assign each atom with identifier.
2. Update each atom’s identifiers based on its neighbours.
3. Remove Duplicates.
4. Fold list of identifiers into 2048-bit vector (a Morgan Fingerprint)

**SMILES**

**Hashed 2048bit vector**

**Molecule**

Fig 5: Architecture diagram of fingerprint vectorization

1. **LOOP formation:**

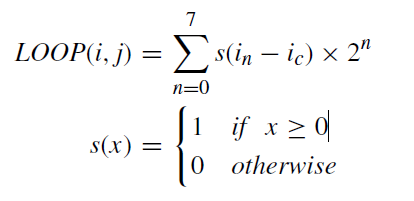
This letter introduces the LOOP binary descriptor (local optimal oriented pattern) that encodes rotation invariance into the main formulation itself. This makes any post processing stage for rotation invariance redundant and improves on both accuracy and time complexity.

**Input:** 2D PSSM matrix

**Output:** 256 feature vector

**Algorithm:** Kirsch laplacian algorithm

In LOOP algorithm, The LOOP feature is obtained by calculating for each image pixel using a 3x3 neighbourhood around each pixel. Firstly, it computed the eight responses of the Kirsch masks, (ln; n D 0,1,..,7) corresponding to pixels with intensities (in; D 0,1,..,7) to obtain the intensity variation in the eight directions. Secondly, based on the rank of the magnitude value ln, each pixel obtained different weight n. Finally, computing the LOOP code for the centre pixel (i, j) as follows:



**Implementation:**

As for the 30% implementation we have done the data pre-processing and fingerprint vector conversion.

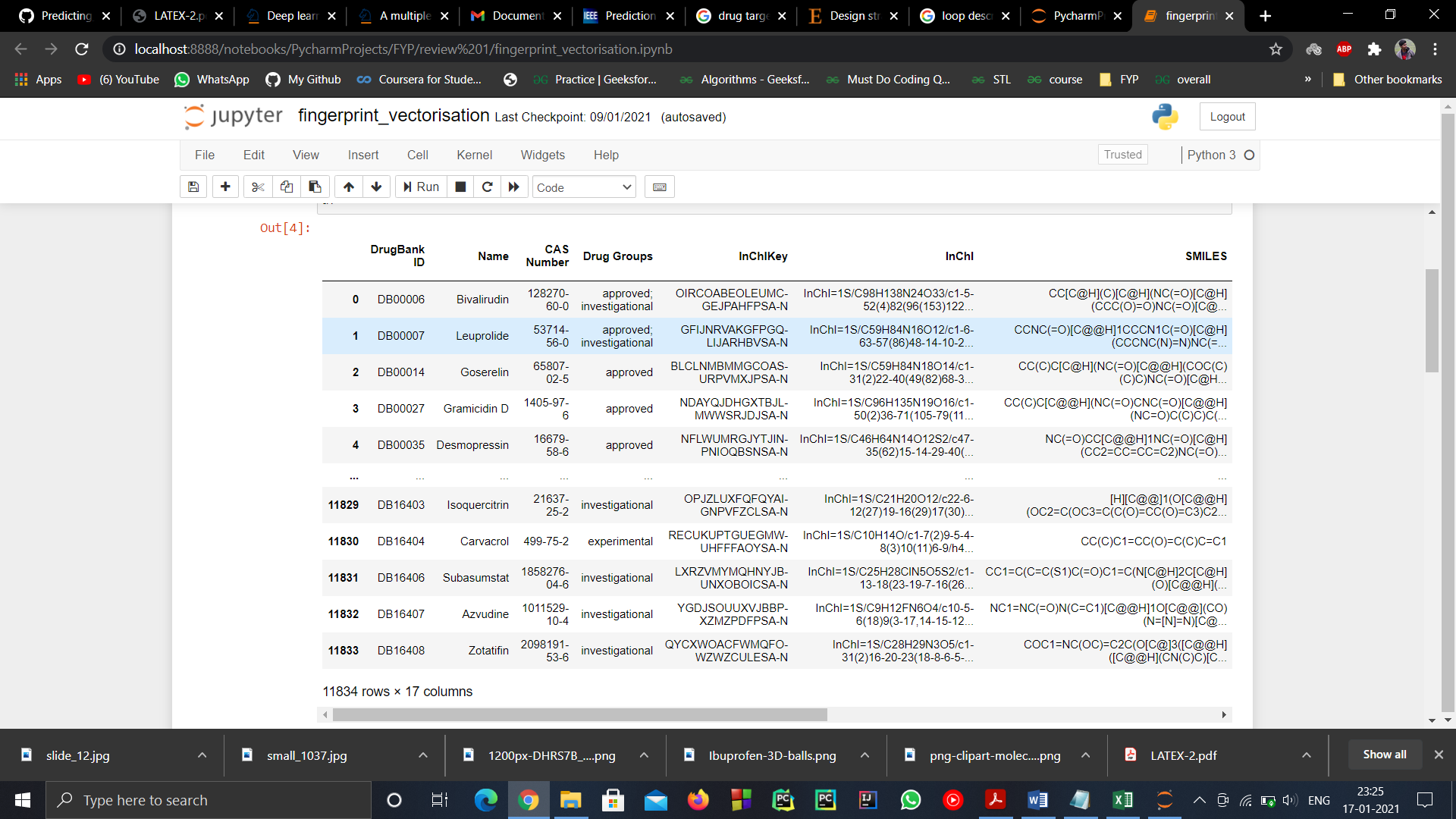


Fig 6: Drug data without Morgan fingerprint

After deriving the Morgan fingerprint from the SMILES, it is added to the existing drug data for future purpose.

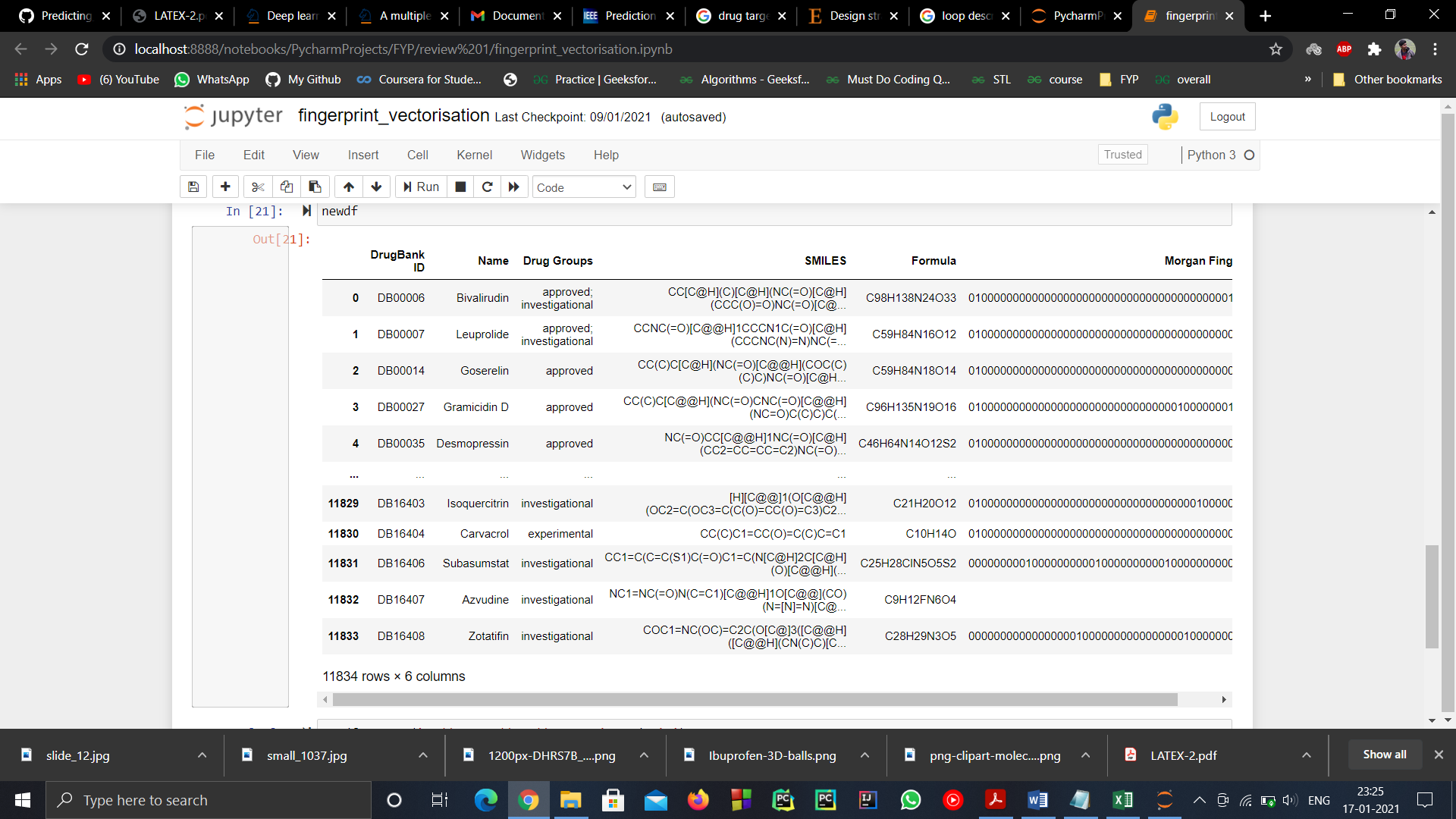


Fig 7: Modified drug data with fingerprint

**Metrics for Evaluation:**

In order to evaluate the performance of the propose method, we use the evaluation measures such as the overall prediction accuracy (Accu.), sensitivity (Sens.), precision (Prec.), and Matthews correlation coefficient (MCC). When predicting the DTIs datasets of enzyme, ion channels, GPCRs and nuclear receptor, five-fold cross-validation would be adopted in this work in order to avoid the over-fitting of the prediction model.

Accuracy = (TN+TP) / (TN+TP+FN+FP)

Precision = TP / (TP+FP)

Sensitivity = TP / (TP+FN)

Specificity = TN / (TN+FP)

**Test cases:**

1. Bivalirudin (DB00006)
   1. Expected output: 010000000000000000000000000000000000000001000000000000000000000000000000000000011000000000000000000…00000000000000000000000
   2. Derived output:

010000000000000000000000000000000000000001000000000000000000000000000000000000011000000000000000000…00000000000000000000000

* 1. Result: Matched

1. Goserelin (DB00014)
   1. Expected output: 010000000000000000000000000000000000000000000000000000000000000000000000000000011100000000000010000000000…0000000000000000000
   2. Derived output:
   3. 010000000000000000000000000000000000000000000000000000000000000000000000000000011100000000000010000000000…0000000000000000000
   4. Result: Matched

**Datasets:**

In this study, we execute the experiment for predicting DTIs on four golden standard datasets, namely enzyme, ion channels, GPCRs, and nuclear receptor, respectively. These datasets are collected from DrugBank, KEGG BRITE, SuperTarget & Matador, and BRENDA which were considered as high-reliability databases.

Drugbank- <https://go.drugbank.com/releases/latest>

**References:**

1. Xinkie Zhan , Zhu-Hong You (Member, IEEE), Jinfan Cai, Liping Li, Chanqing Yu, Jie Pan, and Jiangkun Kong “Prediction of Drug-Target Interactions by Ensemble Learning Method from Protein Sequence and Drug Fingerprint”.

2. Nelson R. C. Monteiro, Bernardete Ribeiro, Joel P. Arrais “Drug-Target Interaction Prediction: End-to-End Deep Learning Approach”.

3. Y. Yamanishi, E. Pauwels, and M. Kotera, ``Drug side-effect prediction based on the integration of chemical and biological spaces, Nov. 2012.

4. M. Kuhn, M. Al Banchaabouchi, M. Campillos, L. J. Jensen, C. Gross, A. Gavin, and P. Bork, ``Systematic identification of proteins that elicit drug side effects,'' Jan. 2013.

5. A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, ``Drug-target interaction prediction via class imbalance-aware ensemble learning,''Dec. 2016.

6. S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, ``How to improve R&D productivity: The pharmaceutical industry's grand challenge,''Feb. 2010.

7. F. Zhu, B. Han, P.Kumar, X. Liu, X. Ma, X.Wei, L.Huang,Y. Guo, L. Han,C. Zheng, and Y. Chen, ``Update of TTD: Therapeutic target database,''Jan. 2010.

8. Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z.Wang, R. Zhang, J. Zhu,Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, and F. Zhu, ``Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics,'' Jan. 2020.

9. S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C.Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner, ``Super-Target and matador: Resources for exploring drug-target relationships,''Jan. 2007.

10. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki Kinoshita, M. Itoh,S. Kawashima, T. Katayama, and M. Hirakawa, ``From genomics to chemical genomics: New developments in KEGG,''Jan. 2006.

11. D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, ``DrugBank: A knowledgebase for drugs, drug actions and drug targets,''Jan. 2008.

12. D. S.Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J.Woolsey, ``DrugBank: A comprehensive resource for in silico drug discovery and exploration,''Jan. 2006.

13. D.-L. Ma, D. S.-H. Chan, and C.-H. Leung, ``Drug repositioning by structure-based virtual screening,''2013.

14. B. Waszkowycz, D. E. Clark, and E. Gancia, ``Outstanding challenges in proteinfiligand docking and structure-based virtual screening,''Mar. 2011.

15. Y.-F. Dai and X.-M. Zhao, ``A survey on the computational approaches to identify drug targets in the postgenomic era,''Jan. 2015.

16. P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su, and M. A. Murcko, ``Chemogenomic approaches to drug discovery,''Aug. 2001.