An Effective Method of Identifying Lead Compounds based on SF-CNN for Hand-Foot-Mouth Disease

Anh Le-Phan1,2 , Phuc-Chau Do3,2 , Loan T.T. Nguyen1,2 , and Nga Ly-Tu1,2  
  
1School of Computer Science and Engineering, International University, Ho Chi Minh City, Viet Nam  
 2Vietnam National University, Ho Chi Minh City, Viet Nam

3School of Biotechnology, International University, Ho Chi Minh City, Viet Nam

[anhphanle271@gmail.com](mailto:anhphanle271@gmail.com), [dnpchau@hcmiu.edu.vn](mailto:dnpchau@hcmiu.edu.vn), nttloan@hcmiu.edu.vn, ltnga@hcmiu.edu.vn

ABSTRACT

Medical drug development has always been one of the major targets in the medical department, due to the complexity of the structure and arrangement of atoms in an amino acid-chain molecule. In order to create the drugs, a procedure needs to be followed carefully with great experience in the field and costs a lot, therefore, computer-aided drug design approach provides an effective method of identifying lead compounds. However, the computational calculation will generate a vast amount of data causing problems for manual analysis, we propose and build upon the architecture of Scoring Function Convolutional Neural Network (SFCNN) which is based on 3D Convolutional Neural Network (3DCNN) to predict protein-ligand Binding Affinity (BA) and classify *desired* compounds, namely *Hit* compounds. The deep learning models are trained with Hand-Foot-Mouth Disease (HFMD) to evaluate the differences between learning the same protein family and learning from 8 kinds of protein families. We perform extent validation to test the improvement of our models with modifications from the original SFCNN. Our works verify the results from analyzing the compatibilities of learning and adapting to the dataset from predicting BA with identifying Hit. Moreover, the Pearson correlation coefficients between the predicted BA by our models achieved over 0.75 and the Root Mean Square Error (RMSE) evaluations are below 0.4 on the validation set, respectively. Our models predict *Top 10% Leading* compounds, namely *Hit-to-Lead* (H2L) compounds, and BA of them and produce them to csv files for examination.

KEYWORDS

Hit, Binding affinity, 3D Convolutional Neural Network, Scoring Function Convolutional Neural Network, Hand-Foot-Mouth Disease.

1 Introduction

Hand, foot, and mouth disease is a mild and contagious viral infection that commonly affects young children under the age of 5. To address this health concern without putting children at risk. Artificial Intelligence (AI) is employed for disease simulation and management. Decades of research and experimentation with molecules have yielded a substantial and reliable dataset, which serves as the foundation for training machine learning models.

The complexity of these molecules presents a challenge for many neural networks, except for the 3D Convolution Neural Network (3DCNN) [1]. In our approach, we disassemble the compound features into a graph structure, allowing us to train the model to predict the binding affinity of protein-ligand complexes obtained through docking.

Our dataset is enriched with comprehensive literature reviews and studies on Hand-Foot-Mouth Disease (HFMD) [2]. Combining this dataset with the Python Molecular View (PMV) [3] tool enables us to predict Binding Affinity (BA) and identify potential hit compounds by confirming the docking sites in a 3D environment.

The contribution of this paper are as follows:

Firstly, we study drug discovery and potency, exploring how to extract the molecule features for identifying hit-to-lead compounds. We collect protein-ligand complexes in PDB [12] format and employ the RDKit tool [4] to distinguish all the significant features of these compounds. Subsequently, we transfer this data to neural network models for training. While several neural network approaches can handle such data, we propose modified Scoring Function Convolutional Neural Network (SFCNN) and SFCNN-Hit models (refer to Section 3.4) with Hit (desired) classification to simultaneously predict BA and Hit (desired) status. Secondly, we introduce a training method utilizing a random selection dataset and permutation random model dataset (outlined in Section 3.5). This approach helps us uncover patterns to accurately predict Hit-to-Lead (H2L) or *Top 10% leading* compounds. Through this method, we aim to discover preferred binding sites and amino acids with reduced cost and time. Next, we offer a bioinformatics model training methodology that involves processing and initializing datasets, ultimately converting them into CSV files for in-depth analysis. Finally, we provide a visualization method in 3D to facilitate the study and verification of our predictions.

2 Literature review

2.1 Related works

The well-known docking tools, AutoDock [5] and AutoDock Vina [6], both are open-source software created by the Olson Lab at the Scripps Research Institute and released in January 2002 and May 2010, respectively, by the Scripps Research Institute. The two in collaboration with PMV [3] to create an easy way see molecules and compounds in 3D. The AutoDock Vina [6] was mainly used to generate protein-ligand complexes for training. It is a global docking based on the geometry of the protein and ligand for fit-to-match results. For each time of running, there are maximum 20 complexes generating. The application of finding the active site in a biological particle requires many replications resulting in many unrelated calculations, which causes difficulty in final analysis, such as manual evaluation and time consuming.

Currently, we consider and propose the SFCNN models with modification from adding complex layers and *Hit* classification to train with the HFMD dataset [2] based on two kinds of random sets (see Section 3.5) to potentially identify H2L or *top 10% leading* compounds and higher BA.

2.1 Background

*Protein, Ligand and Docking:*

Proteins are the building blocks, the gears, and the systems of all organisms on this planet as they work inside the cells and deliver nutrients throughout the body. The foundation of proteins is built upon by hundreds to thousands of smaller units called amino acids, which are made from the most abundant atoms in the world, and they are connected into long chains. These chains can be formed from 20 different types of amino acids and their sequence creates each protein’s unique 3-dimensional structure and functionalities. Protein is one of the major types of large biomolecules as it is responsible for catalyzing the biochemical reactions that sustain functionalities of the host. Thanks to evolution, furthermore, proteins have strengthened their ability to bind with other proteins or ligands to produce desirable results as they bind to receptors which are usually referred to as ligands. In protein-ligand binding complexes, the ligand is often than not molecules which bind with its target by docking in protein sites and produce reactions with amino acids.

Drug discovery must be followed step by step and its early phase requires finding candidate protein-ligand complexes by applying H2L method which demands the high drug potency shown in Fig.1 which as the complex with rank 3 is highlighted, meaning it was chosen as the preferred compounds over the 2 ranks above it because of the star-shaped ligand connects to the protein almost perfectly. Potency is a result of the combined score from ligand efficacy and binding affinity. Ligand’s ability to manufacture biological response from linking up with amino acids of the target and the quantitative magnitude of the previous response are cited as Ligand Efficacy (LE). This response may be as an agonist, antagonist, or inverse agonist, depending on the physiological response produced. The more fitting the complex is, the better the complex is considered to be qualified as *Hit* or *desired*.

A picture containing screenshot, text, diagram

Description automatically generated

Figure 1: Drug potency [9]

The other requirement of potency, BA, measures the tendency or strength of the effect. In general, the higher the affinity the greater attractive forces the better interaction between the ligand and the receptor. This rate can be predicted by utilizing the Docking method which orders the ligand to search for docking sites and rotate itself to create the best results shown in Fig.2 as in the shape of the ligand has to be changed and rotated to match the binding site of the target and form the strongest connection.

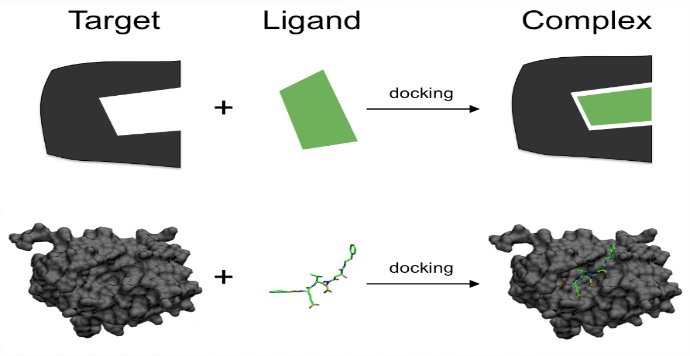


Figure 2: Docking method [10]

*Environment and Tools:*

3Dmol.js [11] is an object-oriented, webGL based JavaScript library or online molecular visualization without the need to install browser plugins or Java. It provides many features to analyze and visualize the protein-ligand complexes from any format the molecules are in such as pdb, mol2, xyz, sdf and cube. Molecules have multiple structure models to study on as shown in Fig 3, we follow the standard procedure benchmark of using secondary structure which is all the structures below. The protein uses the string model while the ligand uses the ribbon model for clarity.

A picture containing art, graphics, fractal art, graphic design

Description automatically generated

Figure 3: Molecule styles [12]

RDKit [4] is an open-source toolkit which is supported in C++ and Python languages, fortunately for cheminformatics and machine learning software. It scans data in the previously mentioned formats then proceeds to showcase the structure and all features of the molecule. Not only that, but we can also add draw connections of the atoms in 2D or 3D.

A diagram of a molecule

Description automatically generated

Figure 4: Input data for SFCNN [8]

We can either use Pytorch or Keras to build machine learning modoels, we choose Keras as it’s easy to use. Keras is also an open-source Python library that helps understanding and developing neural networks which has been known since 2005 by Francois Chollet with the help of other libraries like Theano, TensorFlow and CNTK. Keras reduces the workload of developers to integrate their ideas freely while maintaining advanced workflows. With its industry-strength performance and scalability, well-known companies and projects have been utilizing it such as Netflix, Uber, NASA, etc.

*SF Convoluational Neural Network (SFCNN):*

Figure 4 shows the SFCNN input data [8] gets built by extracting features from RDKit [4] and applying empty space discovering algorithm to determine the availability of ligand to the protein when docking and learn to predict BA better.

*Metrics:*

For classification, the confusion matrix [15] is brought into play to express the performances smoothly. We compare the BA predictions mostly throughout the whole process, so we apply the Matthews Correlation Coefficient (MCC) [14] as it is a more reliable and dependable metric for taking in the all the four categories of the confusion matrix proportionally to the number of positive and negative elements within the dataset.

(1)

where TP, FP, TN, and FN are True Positive, False Positive, Trure Negative and False Negative, respectively.

When the denominator is arbitrarily set to one and any one of the four sums in the denominator is zero, the MCC is zero, which can be demonstrated to be the proper limiting value. The MCC, which considers the balancing ratios of the four confusion matrix categories, is more insightful in assessing binary classification difficulties than F1 score and accuracy.

Mean Absolute Error (MAE) is calculated as the sum of absolute errors divided by the sample size: the true value. Alternative formulations may include relative frequencies as weight factors. The MAE uses the same scale as the data being measured as follows.

(2)

where is the total size of the data.

RMSE represents the root mean square error, N is the total number of observations or data points; and represent the individual predicted values and the actual values, respectively. It is widely used in various fields, such as machine learning, statistics, and predictive modeling, to evaluate the accuracy and performance of regression models.

(3)

Pearson correlation coefficient (r) is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

(4)

where and are the mean of the vector x and vector y, respectively.

3 Proposal

Our proposal consists of training proposal models, 3D visualization front-end and back-end. The front-end is a web application from using Flask framework with python environment and implementing 3Dmol.js from py3dmol and other tools to visual 3D molecule with H2L predictions. The evaluation is completed by utilizing pre-trained proposal models running through preprocessed dataset extracted from RDKit tool [4].

3.1 Preprocessing

Our dataset is originally taken from RCSB Protein Data Bank (1) and PDBbind (2) which contain thousands of protein-ligand complexes, up to 16000. The raw data goes through its first preprocessing through [2]. By manually analyzing every single molecule from the 16000 data using AutoDock Vina and PMV [3] to provide csv files containing the name of the models, the BA score and hit/no hit status of each molecule shown in Figure 5.

A picture containing text, number, screenshot, font

Description automatically generated

Figure 5: Original Labels for HFMD

A screenshot of a computer

Description automatically generated with low confidence

Figure 6: NN Model Prediction for HFMD

Figure 6 shows these two attributes, namely BA and Hit/No\_hit, are used for training our proposal neural network model to produce the predictions on the very same attributes. During our validation and testing phase, the models take in the datasets to export the csv files like below, while the remaining attributes are used for matching the data.

3.2 Top 10% Leading Compound (H2L)

Hit-to-lead (H2L) is a stage in early drug discovery where small molecule hits from a high throughput screen (HTS) are evaluated and undergo limited optimization to identify promising lead compounds. In order to obtain the best ligand model, the ligand is required to arrange and dock itself to the protein with the best structure and amino acids to be considered as “Hit”. The binding site is also vital to determine the “Hit” status as the better the shape the easier the ligand can dock in the site.

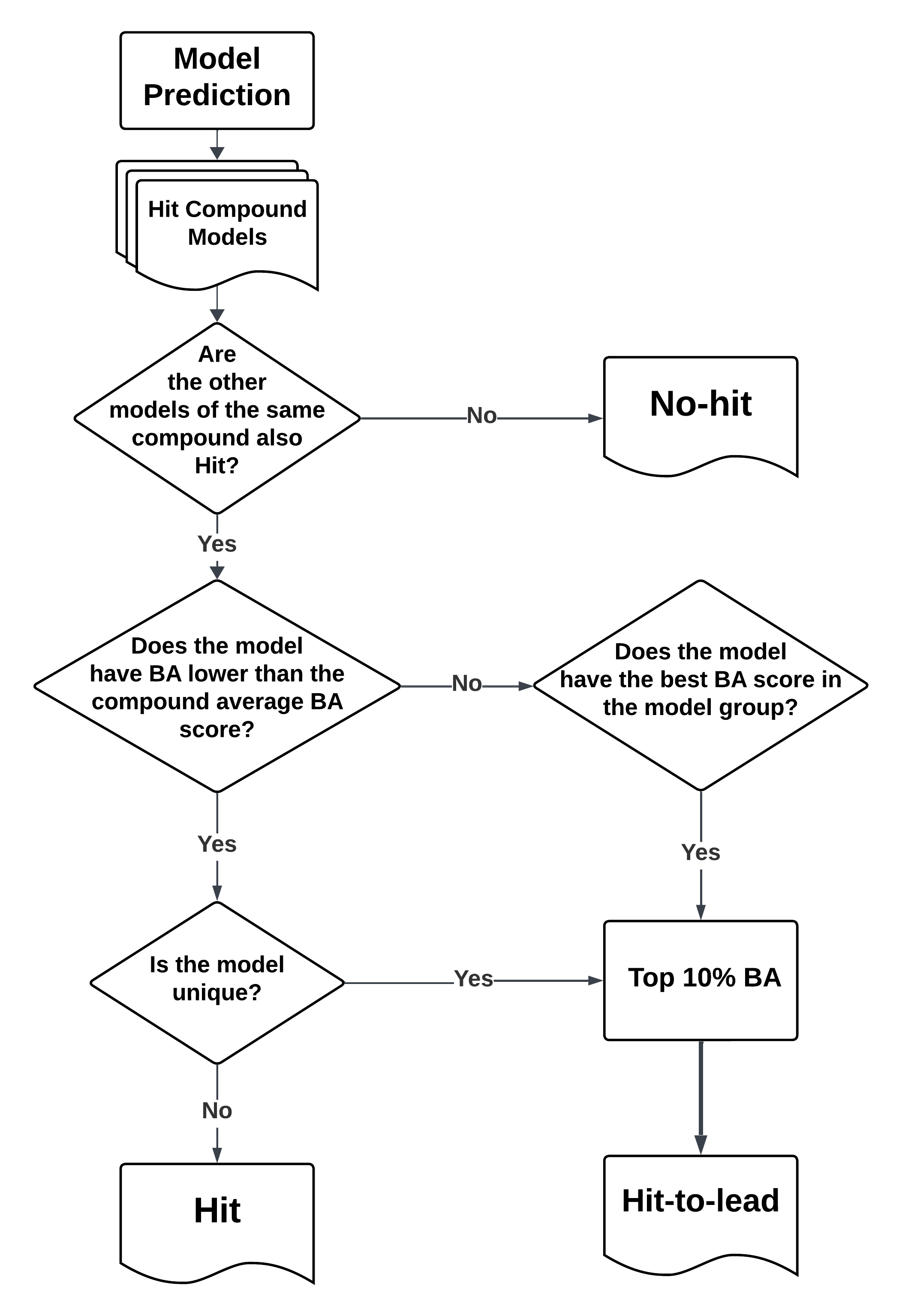


Figure 7: H2L criteria

Despite the model having great predicting capability, there are some results still considered No-hit if the following models of the compounds aren’t classified as Hit as well. Hit-to-lead requires the best models with the best BA scores to have the greatest chance to pass clinical trials for safety on humans. From this sorted H2L, the compounds will be tested and return the comparisons with the lab results as training data for fine-tuning.

3.3 Build SFCNN model

The SFCNN model is constructed using the 3DCNN model, which was initially developed to identify docking sites by detecting hit molecules. The 3DCNN model has convolutional layers that are designed to analyze the grid as if it were learning the positions of amino acids and atoms within the grid. The model accepts a single 4D grid data input and produces a prediction of the binding affinity score and hit status shown in Fig.7.

Diagram

Description automatically generated

Figure 8: 4D Molecule data [15]

The SFCNN model has an approach that focuses on the effectiveness of ligand docking in protein instead of finding docking sites for ligands is shown in Fig.9. With its much smaller input shape, the model can demand the convolutional layers to thoroughly compute the transformation of the matrices in hoping for understanding the relationship between ligand and protein that can get the *top 10% leading*. The model receives 4D grid data and pushes out BA at first but also Hit classification.

A diagram of a cube

Description automatically generated

Figure 9: SFCNN One-Hot Encoding [8]

The moment the models have finished training, they are selected with py3dmol to demonstrate their effectiveness and efficiency in classifying H2L compounds or *top 10% leading* compounds. In addition to this, there are available algorithms to evaluate the performance of the models on testing dataset and export a csv file with details of each prediction on each molecule.

To showcase the H2L (*top 10% leading*) predictions in detail and the evaluation of the model on the sample size, we have to implement an algorithm in identifying unique complexes within the dataset because the ligand binding with the same protein can appear multiple times due to its efficacy relying on the position and the rotation the ligand docking to.

3.4 Proposed architecture

The SFCNN model from paper [16] is similar to the 3DCNN model in structure and components. However, the SFCNN model focuses on learning the binding sites of ligands instead of the whole molecule by constructing input data from the positioning of the ligands. Although SFCNN and 3DCNN use the same building blocks and steps in training, their input data shape and receiving data are the main differences in achieving distinguishable results and evaluation.

Figure 10 shows the SFCNN model constructed upon multiple convolution layers that start with an input layer of (20, 20, 20, 28). The model then repeats the pattern of Conv, Batch Norm, Activation, and Pooling until it diverges into Fully connected layers to predict BA and Global pooling to classify “Hit” or no “Hit”.

A picture containing text, diagram, sketch, black and white

Description automatically generated

Figure 10: SFCNN Architecture

A diagram of a computer program

Description automatically generated with medium confidence

Figure 11: Hit Classification

Our proposed SFCNN model obtains a classification branch in its architecture to learn Hit status and reduce the stress and fine tuning of the BA layers shown in Fig.11.

A diagram of a program

Description automatically generated

Figure 12: Modified SFCNN

Figure 12 shows another of our proposal model, namely modified SFCNN, contains more layers aiding the calculation of the models and enhance the output of CNN layers Our method is built with the same building blocks 3D Convolutional layer, Activation layer, Batch Normalization layer, Dropout layer, and Pooling layers which can be accessed by Keras and Tensorflow libraries. This model also contains the 2 branches for Hit classification and BA.

3.5 Random Datasets

We perform 2 random datasets for the models to train and test their capabilities. The first one is Random protein selection dataset as we split the 21 ligands binding with the 8 proteins into 15 ligands for training, 3 for validation and 3 for testing. The models are forced to learn every positioning 15 ligands can arrange themselves to bind and get the desired outcome while the models correct their mistakes from the 3-validation dataset. Finally, the models will be introduced with totally brand-new ligands to test, it is picked randomly as despite each ligand’s goal is to eliminate HFMD, each one still generates its own BA result and success of docking into the proteins perfectly.

The second dataset is randomized by Mersenne Twister pseudo-random number generator for all 21 ligands over 8 proteins. Fortunately, the generator provides a fixed method to produce the same randomized data for and test for evenly distributed ligands for training, validation, and testing datasets. To compare with the Random protein selection dataset on equal footing, we match the data length for the datasets.

Our aim for training the models on these datasets is to determine the dependency of the models towards the full understanding of the ligands themselves or the rearrangement of the ligands themselves when docking to the proteins.

4 RESULTS

These two libraries are written in Python language; thus, our models are trained in Python environment with Jupyter kernel to run each function individually. As the names of 3DCNN suggested, the models work with 3D Convolution which translates to the very 3D Convolutional layer only accepts 4D data grid or 5D data tensor, meaning the data is massive for one molecule meaning Local IDEs would take too long to process the training phase of machine learning. Therefore, only required to perform training from any platform that has Python environment to import Keras and RDKit to take over the entire process from start to end of machine learning. Google Colab with its access to Colab Pro allows us to work with more powerful compute units, we mainly use T4 GPU with High-RAM runtime while withholding 25.5 GB system RAM, and 166.8 GB disk to freely handle machine learning and data analysis. Before starting the machine learning process, the data enters preprocessing phase to match each model’s input shape, SFCNN with (20, 20, 20,28) and contains all the features of individual atom within the molecule and place them to last dimension of the grid just from doing one-hot encoding.

4.1 Dataset

The original dataset provided [2] contains 8 proteins of the same family having already been bound with ligands. More precisely, each of the proteins only bind with one ligand then get converted into a pdb file for RDKit tool to access.

A picture containing text, screenshot, font, number

Description automatically generated A picture containing text, screenshot, font, number

Description automatically generated

Figure 13: Original data of HFMD

The proteins start their structure with ATOM class while the ligands begin with HETATM shown in Fig.13. Not only that, each ligand traverses on the protein and dock itself one hundred times in order to find the best binding sites, hence there are 100 protein-ligand complexes with the same name. Alongside the pdb files, the author [2] provides a csv file containing crucial info of each molecule, BA and Hit/No\_hit (namely desired/undesired, respectively), as labels for training our proposed models shown in Fig.14.

A picture containing text, screenshot, number, font

Description automatically generated

Figure 14: BA and Hit (desired) /No hit (undesired) status of HFMD.

1. *The* performances *of SFCNN on 2400 samples from different Random Dataset*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **r (%)** | **MAE (%)** | **RMSE (%)** | **MCC (%)** |
| *Random protein selection dataset-2400 samples* | | | | |
| SFCNN [8] | 61.384 | 35.416 | 46.129 | NaN |
| SFCNN-Hit (proposal) | 69.429 | 31.968 | 40.528 | 96.79 |
| Modify SFCNN (proposal) | **82.354** | **30.4** | **40.5** | **98.01** |
| *Random Permutation models dataset-2400 samples* | | | | |
| SFCNN [8] | 69.424 | 31.968 | 40.538 | NaN |
| SFCNN-Hit (proposal) | 76.11 | 30.331 | 39.754 | **98.39** |
| Modify SFCNN (proposal) | **83.577** | **29.59** | **38.079** | 95.941 |

The models used the same hyperparameters, except for the structure, like Adam optimizer with learning rate of 0.0001 and batch size of 32 while running 200 epochs with early stopping under the same Python environment and T4 GPU RAM. We evaluate our proposal models with SFCNN on both Random protein selection and Random Permutation models to see which training approach fits with the data and the preprocess datasets the most and compare the performance of the models with each other if adding classification layers and make the layers denser help at all.

After simulating 2400 samples on both datasets, we obtain the performances mostly on Regression for the 3 models since original SFCNN models only train and return BA. We have an additional metric to track the Hit (desired) classification for the proposal models.

Table I displays the adaptability of the proposed models to learn the pattern of docking better with by adding the Hit classification layers, it improves the performance a decent amount and by modifying the layers to account for both BA and Hit status, the model outperforms expectations.

4.2 3D Visualization Evaluation

After receiving the predicted results and its evaluation, we acquire py3Dmol to visualize them. This enables professionals to analyze and examine the results and observe the progression of the models. Lastly, based on Unique Complex Filter to sort out the best 10% of the H2L (*top 10% leading*) predictions to distinguish the H2L compounds so we can observe its docking site with 3Dmol.js.

A close-up of a cell

Description automatically generated

Figure 15: H2L (top 10% leading) compounds – 5c1u-CR\_models

A structure of a molecule

Description automatically generated with medium confidence

Figure 16: No-Hit (undesired) compounds –5c1u-FOPMC-models

Figure 15 is a demonstration of protein and ligands forming “Hit”/”H2L” compounds from docking to correct binding site. The orange ribbon molecules represent the ligands binding to the blue area which is determined in paper [2] as the active site for developing “Hit” (desired) and “H2L” (top 10% leading) compounds, namely *desired* and *top 10% leading* compounds, respectively. Not only that, but we also take the BA predictions to sort out the proper “H2L” compounds since the proteins can suffer mutation or the rotation model is incompatible with the binding site. Those exceptions will be rejected and later recognized as “No Hit” (undesired) compounds which are the green ribbon molecules shown in Figure 16.

Anything outside of the blue area is appraised as undesired compounds immediately with no exceptions. This data will be reusable once getting approval from the professional testing the predictions on new combinations of ligands to this protein family.

5 Conclusion

In this paper, we have investigated and understood the drug discovery and drug potency in the making, what are the challenges and the tools we have implemented to solve those challenges. With the help of 3D visualization from 3dmol.js, we can provide means to convey the topic better for everyone to understand. We also propose modified SFCNN and SFCNN-Hit to handle and assist in searching for *top 10% leading* compounds faster. In the future, we will fully implement web applications and enhance the SFCNN to predict BA better. Not only that, but we will also implement Graphic Neuron Network (GNN) into the modified models and other methods to perform Hit-to-Lead (*top 10% leading*) predictions on other proteins in the same family and potentially work on different family of proteins as well.

References

[1] 3DCNN Rao, C., & Liu, Y. (2020). Three-dimensional convolutional neural network (3D-CNN) for heterogeneous material homogenization. Computational Materials Science, 184, 109850.

[2] Le, T. T. V., & Do, P. C. (2022). Molecular docking study of various Enterovirus—A71 3C protease proteins and their potential inhibitors. Frontiers in Microbiology, 13.

[3] PMV Documentation, Retrieved from: <https://www.autopack.org/install/python-molecular-viewer-pmv-installation>, last accessed 2023/3/14.

[4] The RDKit Documentation, Retrieved from: <https://www.rdkit.org/docs/index.html>, last accessed 2023/3/14.

[5] The Scripps Research Institute. (2021). AutoDock. Retrieved from <http://autodock.scripps.edu/>

[6] Trott, O., & Olson, A. J.: AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, 31(2), 455-461 (2010).

[8] Wang, Yu, et al.: Sfcnn: A Novel Scoring Function Based on 3D Convolutional Neural Network for Accurate and Stable Protein–ligand Affinity Prediction. BMC Bioinformatics, vol. 23, no. 1, Springer Science and Business Media LLC, (June 2022).

[9] WiKiDoc Users. “Ligand (Biochemistry) - Wikidoc.” Ligand (Biochemistry) - Wikidoc, 9 Aug. 2012.

[10] Meng, E. C., Shoichet, B. K., & Kuntz, I. D. (1992). Automated docking with grid‐based energy evaluation. Journal of computational chemistry, 13(4), 505-524.

[11] 3dmol js Nicholas Rego , David Koes, 3Dmol.js: molecular visualization with WebGL, Bioinformatics, Volume 31, Issue 8, April 2015, Pages 1322–1324, https://doi.org/10.1093/bioinformatics/btu829.

[12] Protein Data Bank, <https://www.rcsb.org/>, last accessed 30/5/2023

[13] Paciotti, Roberto & Agamennone, Mariangela & Coletti, Cecilia & Storchi, Loriano. (2020). Characterization of PD-L1 binding sites by a combined FMO/GRID-DRY approach. Journal of Computer-Aided Molecular Design. 34. 10.1007/s10822-020-00306-0.

[14] Huang, S.-Y., and Zou, X. (2010). Advances and challenges in protein-ligand docking. Int. J. Mol. Sci. 11, 3016–3034. doi: 0.3390/ijms11083016.

[15] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. Nucleic Acids Res. 28, 235–242. doi: 10.1093/nar/28.1.235.

[16] Savio, D., Pastewka, L., & Gumbsch, P. (2016). Boundary lubrication of heterogeneous surfaces and the onset of cavitation in frictional contacts.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00