Hybrid deep learning model for evaluations of protein-ligand binding kinetic property

Yujing Zhao, Qilei Liu\*, Yu Zhuang, Yachao Dong, Linlin Liu, Jian Du, Qingwei Meng, Lei Zhang\*

State Key Laboratory of Fine Chemicals, Frontiers Science Center for Smart Materials Oriented Chemical Engineering, Institute of Chemical Process Systems Engineering, School of Chemical Engineering, Dalian University of Technology, Dalian 116024, China

Qilei Liu (liuqilei@dlut.edu.cn), Lei Zhang (keleiz@dlut.edu.cn)

Abstract

A growing consensus is emerging that optimizing the binding kinetic parameters is crucial to ensure better drug efficacy. Therefore, *in silico* methods are necessary to predict the kinetic parameters of drug-protein binding. In this study, we demonstrate the application of an attention mechanism to derive a deep learning-based structure-kinetics relationship model for the dissociation rate constants (*k*off) of inhibitors targeting drug-protein interactions. This model has the capability to provide accurate predictions for evaluating the *k*off of protein-ligand complexes. To the best of our knowledge, the deep learning model developed in this work achieves the state-of-the-art prediction accuracy and outperforms other commonly used machine learning-based generic dissociation kinetic models (e.g., random forest models) in terms of precision and efficiency.

**Keywords**: drug binding kinetics, deep learning, dissociation kinetic constant, Mol2vec, attention mechanism.

* 1. Introduction

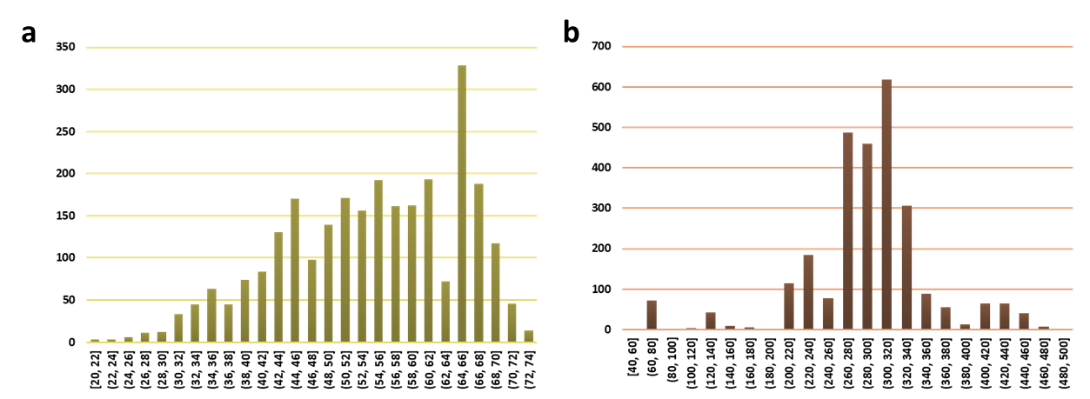
Drugs play a significant role in disease prevention and safeguarding public health. The process of discovering a new drug is an arduous and costly endeavour, often surpassing a duration of 10 years and requiring a budget exceeding one billion dollars. The binding/unbinding kinetic properties of protein-ligand complexes, such as the dissociation rate constant (*k*off), are increasingly recognized as crucial factors influencing drug efficacy and safety during the drug optimization process. Although various experimental methods, molecular dynamics simulations, and surrogate modeling techniques have been developed to measure or estimate *k*off, the current approaches, both experimental and *in silico*, suffer from limitations in providing high-throughput and accurate predictions. Therefore, this study aims to address this gap by introducing a novel data-driven dissociation kinetic model. This model employs a hybrid deep learning architecture that combines a convolutional neural network with the attention mechanism for the prediction of *k*off values in protein-ligand complexes across multiple targets.

* 1. Development of the Dissociation Kinetic Model
     1. Dissociation Kinetic Data Preparation

In this study, a comprehensive dissociation kinetic database is established by collecting dissociation kinetic data from four datasets, namely KIND (Schuetz et al., 2020), BindingDB (Gilson et al., 2016), PDBbind-koff-2020 (Liu et al., 2022), and the work by Amangeldiuly et al. (2020). For the database construction, the dissociation kinetic data contributed by the aforementioned datasets are pre-processed based on the following criteria:

* Entries involving ligand SMILES that cannot be read by the RDKit tool (Landrum, 2006) or the ligand Morgen fragments cannot be found in the Mol2vec method (Jaeger et al., 2018) are deleted.
* Entries associated with protein-ligand complexes having a Morgen fragment number between 20 and 74 and an amino acid number between 40 and 500 are retained.
* Entries related to protein-ligand complexes with *pk*off values ranging from 0 to 5 and without uncertain inequality signs (e.g., “>”, “<”) are retained.
* Entries involving protein-ligand complexes with more than one protein chain are excluded.
* Ligands conforming to Lipinski’s “Rule of Five” are retained.
* Entries involving charged or ionic ligands are excluded.
* Duplicate entries are retained only once.

After applying the above criteria, a total of 2,716 unique protein-ligand *pk*off values are obtained for our database, as well as the corresponding SMILES strings of ligands and FASTA strings of proteins. This dataset is divided into a training dataset (2,175 samples), a validation dataset (270 samples), and a test dataset (271 samples), following an 8:1:1 ratio, respectively. The *pk*off values in our database cover a wide range from 0 to 5, representing the majority of cases in terms of *pk*off values. A higher *pk*off value indicates a longer duration of the inhibitor candidate binding to the protein. In **Figures 1(a-b)**, the distributions of the numbers of Morgan fragments for ligands and the amino acids for proteins in our database are presented, respectively. These distributions approximate the normal distribution, indicating that our dissociation kinetic database encompasses diverse samples of ligands and proteins.



**Figure 1**. **a** The distributions of the numbers of the Mogan fragments for ligands in the established database. **b** The distributions of the numbers of the amino acids for proteins in the established database.

* + 1. Input Representation

In our dissociation kinetic model, the protein-ligand complexes are represented using text-based strings because accessing their three-dimensional structures is time-consuming. The ligand is represented using the Mol2vec descriptor, which is a unique set of Morgan fragments with embedded feature vectors for each molecule. To obtain the Mol2vec descriptor, the SMILES string of the ligand is first transformed into fragments using the Morgan algorithm with a radius of 1. Then, a search approach is employed to embed the fragments with feature vectors, where the default dimension is 300. It is important to note that the Morgan fragments used in Mol2vec provide greater interpretability compared to plain text-based SMILES strings. Each ligand sample is represented by a two-dimensional matrix with dimensions of 74×300. The complete ligand feature is obtained by summing the vectors of individual Morgan fragments. For more details about the Mol2vec descriptor, please refer to Jaeger et al. (2018).

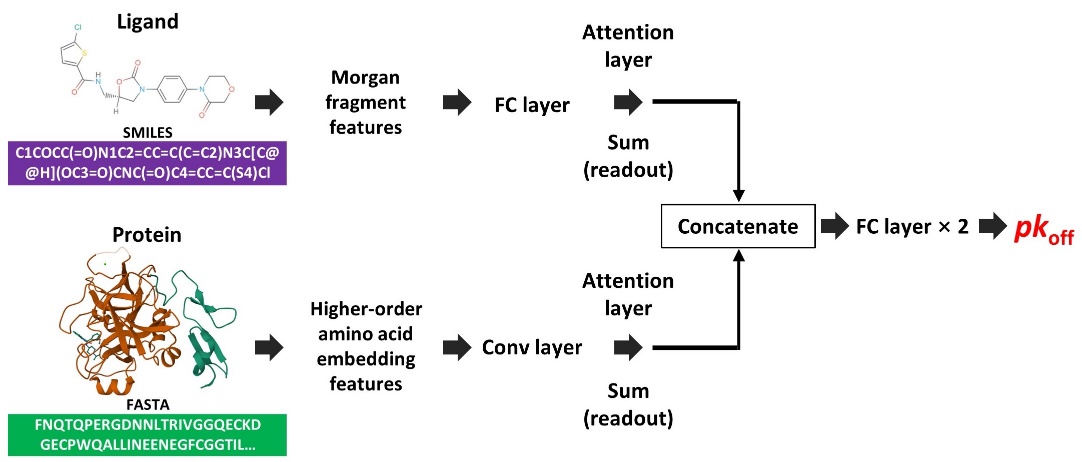
The protein is represented using higher-order amino acid strings with embedded feature vectors. Each higher-order amino acid consists of three adjacent amino acids in the protein sequence, which is provided in the FASTA format. This representation captures the interactions among amino acids. A total of 6,715 higher-order amino acid strings are generated from the proteins in our dissociation kinetic database. Each higher-order amino acid string is then embedded with a feature vector using the embedding method in the PyTorch library (Paszke et al., 2019). The feature size for each embedded vector is set to 30. Based on the higher-order amino acid strings, each protein sample is represented by a two-dimensional matrix with dimensions of 500×30. The complete protein feature is obtained by summing the vectors of individual higher-order amino acid strings.

* + 1. Implementation of the Dissociation Kinetic Model

The architecture of the end-to-end text-based deep learning model for predicting *pk*off is illustrated in **Figure 2**. Take a protein-ligand complex as an example. The deep learning model begins with the SMILES string of the ligand and the amino acid sequence (FASTA string) of the protein. These inputs are then transformed into the ligand matrix (Mol2vec descriptor: Morgan fragments with identifier features) and the protein matrix (higher-order amino acids with embedding features), respectively. For the ligand matrix, it undergoes feature extraction through a fully connected (FC) layer and an attention layer. On the other hand, the protein matrix is processed by a one-dimensional convolution (Conv) layer and a MaxPool layer to reduce the row size before being passed through an attention layer. This step helps reduce unnecessary computational costs in the attention layer. At last, the processed features of the ligand matrix are aggregated using a readout operation that adds up the matrix rows. The same process is applied to the protein matrix. The resulting features from both matrices are concatenated and passed through subsequent FC layers to predict the *pk*off values for the protein-ligand complex.

* 1. Leverage a Hybrid Deep Learning Model for Dissociation Rate Evaluations

A novel data-driven dissociation kinetic model is developed in this work to predict *pk*off using a hybrid deep learning architecture comprising a convolutional neural network and the attention layer. The hyperparameters (epoch, batch size, dropout rate, etc.) are determined through a trial-and-error approach based on our empirical knowledge. The loss function employed in our dissociation kinetic model is the mean squared error (*MSE*), commonly used in regression tasks. To minimize the loss function, we optimize the model parameters using the adaptive moment optimizer (Kingma and Ba, 2014), with the learning rates ranging from 0.0005 to 0.00005. A decay factor of 0.5 is applied to update the learning rates when the loss function of the validation set does not decrease with epochs. The evaluation criteria for assessing the prediction results of our dissociation kinetic model in estimating *pk*off are the Pearson coefficient (*r*) and the root mean square error (*RMSE*). Higher values of *r* and lower values of *RMSE* indicate better predictive performance of the dissociation kinetic model. The deep learning-based dissociation kinetic model is trained 20 times to prevent fortuitously exceptional outcomes. The dissociation kinetic model is implemented using the PyTorch library (Paszke et al., 2019) and the Python language (Oliphant, 2007). The training and prediction results for *pk*off of the deep learning model are shown in **Figure. 3**.



**Figure 2**. The architecture of the deep learning model.

图形用户界面, 图表, 散点图

描述已自动生成

**Figure 3**. Training and prediction results for *pk*off of the deep learning model. **a** The loss function (*MSE*) varying with the epochs during the training process of the deep learning model. **b** The metrics function (*r*) varying with the epochs during the training process of the deep learning model. **c** The prediction results of the deep learning model for *pk*off.

The *MSE* and *r* values of our deep learning model for the training, validation, test sets are as follows: 0.238±0.018, 0.299±0.020, 0.292±0.014 and 0.869±0.010, 0.838±0.012, 0.836±0.009, respectively. **Figures 3(a-b)** depict the variations of the loss function (*MSE*) and the metrics function (*r*) varying with the epochs during the training process of the deep learning model, respectively. These results demonstrate that our developed deep learning model does not suffer from overfitting and can effectively extrapolate to new samples. Furthermore, the incorporation of the attention mechanism accelerates the learning process of our model. The deep learning model achieves convergence at around 40th epoch when attention layers are utilized, whereas it requires more than 80 epochs to achieve similar results without attention layers. **Figure 3(c)** illustrates the prediction results of our optimal deep learning model on the test set, demonstrating an acceptable prediction ability with an *r* value of 0.840. This indicates that our deep learning model can provide qualitative evaluations of the dissociation kinetics of protein-ligand complexes. Importantly, our model is lightweight, with only 15,603 training parameters, compared to other affinity-based deep learning models such as KDEEP (Jiménez et al., 2018), which typically have around 1,000,000 training parameters. This highlights the efficiency of our model in terms of training and extrapolations.

To highlight the superior prediction accuracy of our deep learning model in estimating *pk*off values, we compare it with other machine learning-based generic dissociation kinetic models described in the literature (**Table 1**). Amangeldiuly et al. (2020) compiled a database containing 501 protein-ligand complexes with experimental *k*off from the public literature. They used this dataset to develop an optimal random forest (RF) algorithm, achieving an *r* value of 0.78 and *RMSE* of 0.82 on the validation set, and an *r* value of 0.75 and *RMSE* of 1.10 on the test set. Liu et al. (2022) created a database consisting of 680 experimental *k*off values and their corresponding protein-ligand complex structures. They developed a general RF model based on this database to predict *pk*off values. The final model was selected using the evaluation criteria of *r* and *RMSE* values: for the training set, *r* = 0.968 and *RMSE* = 0.474; for the validation set, *r* = 0.706 and *RMSE* = 0.986; and for the test set, *r* = 0.501 and *RMSE* = 0.891. The model developed by Liu et al. (2022) shows signs of overfitting, as indicated by its performance on the training, validation, and test datasets. On the other hand, it is unclear whether the model proposed by Amangeldiuly et al. (2020) suffers from overfitting, as they have not provided the performance on the training dataset. In comparison, our deep learning model demonstrates superior performance on the training set (*r* = 0.869±0.010 and *RMSE* = 0.488±0.018), validation set (*r* = 0.838±0.012 and *RMSE* = 0.546±0.018), and test set (*r* = 0.836±0.009 and *RMSE* = 0.540±0.013). This improvement can be attributed to the advantages of our deep learning architecture and the abundance of our database. Moreover, our model enjoys faster computational speed compared to the aforementioned models, thanks to the utilization of text descriptors. In contrast, the calculations of descriptors in the previous models rely on three-dimensional structures of protein-ligand complexes. These analyses strongly support the conclusion that our deep learning model outperforms other known machine learning-based generic dissociation kinetic models in predicting *pk*off values.

**Table 1**. Performance comparisons between the deep learning model developed in this work and the existing models in literature.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Dataset** | ***r*** | ***RMSE*** |
| This work | Training | 0.869±0.010 | 0.488±0.018 |
|  | Validation | 0.838±0.012 | 0.546±0.018 |
|  | **Test** | **0.836±0.009** | **0.540±0.013** |
| RF model (Amangeldiuly et al., 2020) | Training | Unknown | Unknown |
|  | Validation | 0.78 | 0.82 |
|  | Test | 0.75 | 1.10 |
| RF model (Liu et al., 2022) | Training | 0.968 | 0.474 |
|  | Validation | 0.706 | 0.986 |
|  | Test | 0.501 | 0.891 |

* 1. Conclusions

A comprehensive dissociation kinetic database is constructed by gathering 2,716 unique *k*off values for various ligands and targets from the literature. Utilizing this database, an end-to-end deep learning-based model is developed for rapid *k*off predictions. Our deep learning model achieves superior performance compared with the existing machine learning-based generic dissociation kinetic models, as evidenced by the *MSE* and *r* values obtained on the training, validation, and test sets. Specifically, the *MSE* values are 0.238±0.018, 0.299±0.020, and 0.292±0.014, while the *r* values are 0.869±0.010, 0.838±0.012, and 0.836±0.009, respectively. These results demonstrate that our model surpasses other known machine learning-based generic dissociation kinetic models and achieves the state-of-the-art prediction accuracy in predicting *pk*off.

Acknowledgement

The authors are grateful for financial supports of “the Fundamental Research Funds for the Central Universities” [DUT22YG218], China Postdoctoral Science Foundation [2022M710578], and National Natural Science Foundation of China [22208042].

References

T.E. Oliphant, 2007. Python for scientific computing. Computing in science & engineering, 9(3), pp.10-20.

S. Jaeger, S. Fulle, and S. Turk, 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. Journal of chemical information and modeling, 58(1), pp.27-35.

N. Amangeldiuly, D. Karlov, and M.V. Fedorov, 2020. Baseline model for predicting protein–ligand unbinding kinetics through machine learning. Journal of Chemical Information and Modeling, 60(12), pp.5946-5956.

M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, 2016. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 44(D1), pp.D1045-D1053.

J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis, 2018. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. Journal of chemical information and modeling, 58(2), pp.287-296.

H. Liu, M. Su, H.X. Lin, R. Wang, and Y. Li, 2022. Public data set of protein–ligand dissociation kinetic constants for quantitative structure–kinetics relationship studies. ACS omega, 7(22), pp.18985-18996.

G. Landrum, 2006. RDKit: Open-source cheminformatics. 2006. Google Scholar.

D.P. Kingma, and J. Ba, 2014. Adam: A method for stochastic optimization. arxiv preprint arxiv:1412.6980.

D.A. Schuetz, L. Richter, R. Martini, and G.F. Ecker, 2020. A structure–kinetic relationship study using matched molecular pair analysis. RSC Medicinal Chemistry, 11(11), pp.1285-1294.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N., Gimelshein, L. Antiga, and A. Desmaison, 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.