ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

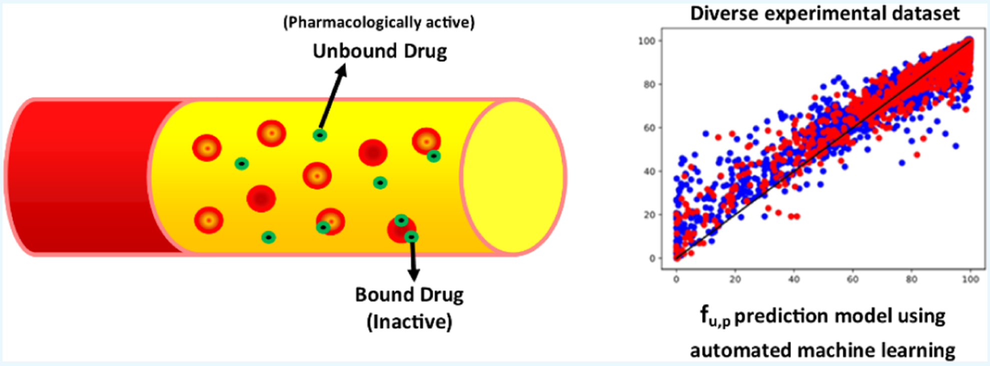
In Silico Prediction of Fraction Unbound in Human Plasma from Chemical Fingerprint Using Automated Machine Learning

Viswajit Mulpuru and Nidhi Mishra*

Cite This: *ACS Omega* 2021, 6, 6791−6797

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information


(Pharmacologically active)
Unbound Drug
Bound Drug
(Inactive)
Diverse experimental dataset
100
80
60
40
20
0
0
20
40
60
80
100
$f_{u,p}$ prediction model using
automated machine learning

ABSTRACT: Predicting the fraction unbound of a drug in plasma plays a significant role in understanding its pharmacokinetic properties during in vitro studies of drug design and discovery. Owing to the gaining reliability of machine learning in biological predictive models and development of automated machine learning techniques for the ease of nonexperts of machine learning to optimize and maximize the reliability of the model, in this experiment, we built an in silico prediction model of a fraction unbound drug in human plasma using a chemical fingerprint and a freely available AutoML framework. The predictive model was trained on one of the largest data sets ever of 5471 experimental values using four different AutoML frameworks to compare their performance on this problem and to choose the most significant one. With a coefficient of determination of 0.85 on the test data set, our best prediction model showed better performance than other previously published models, giving our model significant importance in pharmacokinetic modeling.

■ INTRODUCTION

Computational drug discovery has rapidly evolved into an alternative to the development of novel drugs,[1] and in the last few years, it has also been reported that the majority of new drugs have originated from academia.[2,3] Although a large number of drugs have been reported by academic researchers, the majority of them fail to interest the pharmaceutical industry due to the lack of proper pharmacokinetic and toxicity studies.[4] Given the above, accurate in silico prediction models of absorption, distribution, metabolism, excretion, and toxicity (ADMET) can be a valuable asset to pharmaceutical scientists. These models can help in screening the novel druglike compounds for consideration in clinical trials.

Plasma is an important component of the cardiovascular system that plays a significant role in the transportation of drugs throughout the body, especially during intravenous administration. Plasma is the straw-colored liquid portion of blood. About 55% of our blood is plasma, and the rest 45% comprises red blood cells (RBCs), white blood cells (WBCs), and platelets that are suspended in the plasma, occupying arterial and venous space including space within the tissues, Plasma helps in interconnecting different organs and tissues for transportation of nutrients, hormones, and proteins to the tissues and waste products back from the tissues for elimination (Figure 1).

As it is known that unbound drug in plasma is capable of showing pharmacological activity by interacting with the targets such as proteins, enzymes, receptors, and channels, for the construction of a pharmacokinetic model, the fraction unbound in plasma ($f_{u,p}$) of a drug is an important factor in determining the drug efficacy. It has also been reported that $f_{u,p}$ influences various other factors of drug efficacy and side effects[5] ranging from renal glomerular filtration to total clearance and hepatic metabolism.[6] Considering the importance of $f_{u,p}$ in determining the efficacy of the drug molecule, it is important to make a precise prediction of $f_{u,p}$ during drug development.

Received: December 1, 2020
Accepted: February 19, 2021
Published: March 5, 2021

https://dx.doi.org/10.1021/acsomega.0c05846

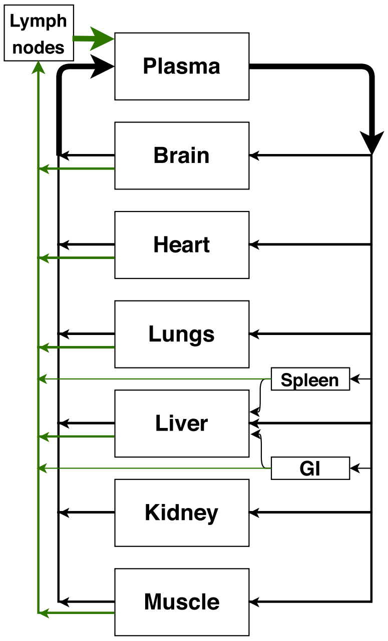


Figure 1. Structure of a simplified PBPK model for human body pharmacokinetics with tissues interconnected by plasma (black arrows) and lymph (green arrows) flows.

Owing to the importance of $f_{u,p}$ in drug discovery, various QSAR models of $f_{u,p}$ are available, but most of these models are focused only on individual plasma proteins,[7−9] confined to a narrow data set[10] or specific to categorical drugs.[11] Only a few models are generated using large data sets, but almost all of these models rely on commercial software to generate descriptors with almost none using freely available software.[12−15] In addition to these predictive models, there are some commercial software that predict $f_{u,p}$. To the best of our knowledge, no prediction model is available that uses free software and a fingerprint-based approach at this moment.

In the field of bioinformatics, the generation of predictive models using machine learning algorithms for building an accurate model to predict a variable of interest is gaining attention.[16,17] However, choosing an appropriate model requires sample characterization, fine-tuning of parameters, and comparing configurations.[18] These extensive steps to find a suitable model for the given data pose significant problems, especially to nonexperts of machine learning. Given this, recently, automated machine learning (AutoML) is being looked into, where AutoML takes advantage of data complexity and automatically identifies the most appropriate model along with the hyperparameters, thus simultaneously optimizing the performance and maximizing the reliability of the model.[19] Considering the advantages of AutoML, many different AutoML frameworks have been developed, with auto-sklearn,[20] Auto-WEKA,[19,21] AutoKeras,[22] H2O AutoML,[23] PyCaret,[24] and TPOT[25] being a few examples.

In this study, we built an $f_{u,p}$ prediction model trained on a comparatively very large data set of experimental values from 5471 compounds. We implemented an automated machine learning approach focusing on molecular-descriptor-free modeling for the ease of users. To create a large and diverse data set, we compiled the experimental values from the ChEMBL[26] database and processed the data for consistency using ad hoc scripts. Then, the PubChem fingerprint (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf) of each molecule was calculated to train the model avoiding any molecular descriptors to represent the chemical structures. Later, the prediction models were generated using the AutoML technique. With auto-sklearn being declared the overall winner of the ChaLearn AutoML Challenge twice, AutoKeras utilizing Tensorflow, which is backed by industrial experts, PyCaret being a low-code machine learning library that automates machine learning workflow, and TPOT being an automated machine learning tool that optimizes machine learning pipelines, we have utilized auto-sklearn v0.8.0, AutoKeras v1.0.4, PyCaret v2.2, and TPOT v0.11.2 to build the model and as well to evaluate their performance on this problem. The prediction models fared excellently when compared to previously available models.

■ MATERIALS AND METHODS

Data-Set Preparation. ChEMBL, a manually curated database of druglike molecules, was used in this study. For the construction of the prediction model, the reported values of PPB or $f_{u,p}$ obtained from ChEMBL were considered.

From the ChEMBL 27 database, about 16 million records of molecules containing activity data are filtered and downloaded. Then, the records containing human $f_{u,p}$ or PPB data were extracted from the obtained data using "PPB", "Unbound plasma", and "%PFU" as keywords in the "standard type" field, resulting in 9463 records. To filter records that are incomplete or that did not satisfy the inclusion criteria, the records with ranged values such as " > ", " < ", " > =", and "≥" were removed, resulting in 8810 records. If a molecule had more than one entry, the average of their values was taken into consideration, ultimately resulting in 5471 records.

Descriptor Calculation. Although a large number of descriptors can be calculated using various descriptor calculation software, considering the information available in a PubChem fingerprint, only PubChem fingerprints of the compounds were used to train the model. PaDELPy, a Python wrapper for PaDEL-Descriptor software, was used for calculating the PubChem fingerprints.

Data Analysis. Data analysis was performed in Python, and the results were visualized using Matplotlib.[27] The diversity of the data set was visualized by constructing a histogram of the number of compounds per 1% bracket of $f_{u,p}$ percentage values. Further, Simpson's diversity index (SDI)[28] was calculated to statistically evaluate the diversity of the data set. SDI measures the community diversity with a value ranging between 0 and 1, with values closer to 1 indicating a high diversity. The SDI was calculated by classifying the compounds into 20 different species based on their $f_{u,p}$ percentage value with classifying compounds in every 5% bracket as the same species.

Model Construction. Auto-sklearn, AutoKeras, PyCaret, and TPOT, the Python modules, and automated machine learning toolkits were used in building the prediction models. The PubChem fingerprints, each containing 881 binary values as calculated by PaDEL, were used in model construction. The data set was split into training (4103 records) and test (1368 records) sets using random selection at a ratio of 3:1 before constructing the regression model. Finally, the best models constructed by auto-sklearn, AutoKeras, and PyCaret were evaluated for their performance on the test set using a range of metrics, and further, the models were saved for predictions. To evaluate and find the most effective automated machine learning toolkit for a nonexpert of machine learning, all of the optional parameters were largely set to their default values.

Auto-Sklearn. Auto-sklearn is a robust AutoML library based on scikit-learn, a Python-based machine learning package. It generates the prediction model considering 14 feature preprocessors, 4 data preprocessors, and 15 classifiers, constructing a hypothesis space with 110 hyperparameters. To generate an efficient prediction model, auto-sklearn uses a meta-learning technique for optimization that is known to show a considerable boost in efficiency. It also includes an automated ensemble constructor to construct a prediction model considering all of the classifiers that were found by the optimizer, further increasing the efficiency.

To construct a prediction model using auto-sklearn, the training set was subjected to the regression module of the auto-sklearn toolkit. The model was constructed using a fivefold cross-validation resampling strategy, with other parameters set to default. Further, the model was evaluated using the test set and the graphs are plotted for visualization.

AutoKeras. AutoKeras is an AutoML library that is based on Keras, which is built on top of TensorFlow and is an industrial-grade framework that is used in scientific organizations of significant importance around the world. Just like other AutoML libraries, the goal of AutoKeras is to help domain experts who are not familiar with machine learning techniques to use the power of ML in their domain of expertise. Unlike many other libraries, AutoKeras focuses on model construction using deep learning techniques through efficient neural architecture search with network morphism.

To construct a prediction model using AutoKeras, the structured data regression function of AutoKeras was used to train the model. The training set was used to train the model specifying the test set as the validation data while setting all other parameters including maximum trials and maximum epochs per trial to the default value. The graphs of the constructed and hyperparameter tuned model were plotted for visualization.

PyCaret. PyCaret is a Python wrapper around several machine learning libraries and frameworks such as scikit-learn, XGBoost, Microsoft LightGBM, and spaCy, among others. Also featuring hyperparameter tuning and ensembling techniques to increase the efficiency of the identified model, it aims to reduce the time taken to construct a suitable machine learning model, thus helping researchers perform experiments quickly and efficiently.

To construct a prediction model using PyCaret, the whole data set was passed to the regression module of PyCaret 2.2, which splits the data set into training and test sets of 70% (3829) and 30% (1642) records, respectively, by default. Further, all of the models from the available machine learning libraries and frameworks were trained on the training set, and the top five models based on their R^2 score were selected for further steps. The selected models were subjected to hyperparameter tuning, specifying to optimize the R^2 score. Further, all of the tuned models were ensembled using the bagging method, which is known to improve the stability and accuracy of the regression models. After ensembling, the best of all of the models were calculated and selected using the AutoML function, further optimizing the R^2 score before finalizing the model for saving.

TPOT. TPOT, built on top of scikit-learn, is an AutoML library that is designed to select and optimize efficient machine learning models and their parameters using genetic programming. TPOT explores thousands of possible pipelines to identify and optimize the best analysis pipeline for the given data.

The training set was subjected to the regression module of TPOT, and the constructed pipeline was evaluated using the test set. The generations and population size were set to the default value of 100 each while setting all other parameters to default. The evaluation graphs of the pipeline are plotted for visualization.

RESULTS

Chemical Space Diversity Analysis. To visualize the diversity of the data set obtained from ChEMBL, a histogram plot of the number of compounds per 1% bracket of $f_{u,p}$ values was generated. The histogram plot with the y-axis plotted on a logarithmic scale for ease of visualization is shown in Figure 2. As seen, although the data set was largely diverse, it was skewed toward higher percentages of $f_{u,p}$ values. Therefore, to statistically confirm the diversity measure of the data set, the SDI of the data set was calculated. As stated, the SDI helps in measuring the diversity of the population. With an SDI of 0.728, the data set is considered to be highly diverse and can be used for training the prediction model.

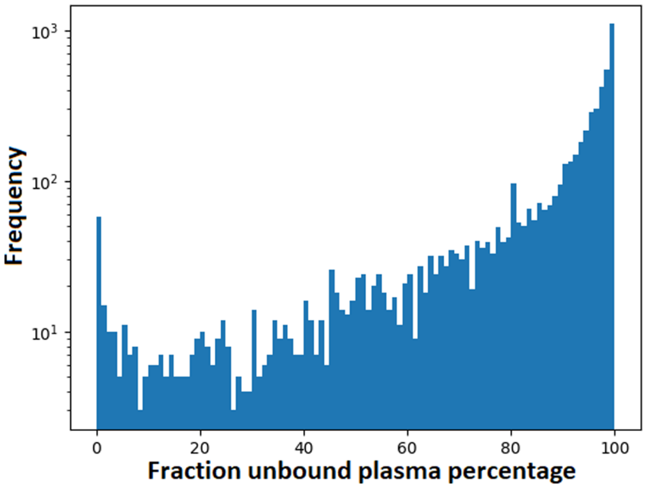


Figure 2. Histogram plot of the frequency of compounds from the data set concerning the $f_{u,p}$ value per 1% bracket on a logarithmic scale.

Performance of Regression Models. The regression models of $f_{u,p}$ were generated using auto-sklearn, AutoKeras, PyCaret, and TPOT. For a diverse data set, the coefficient of determination (R^2) along with RMSE is considered to be a reliable statistic for the evaluation of the prediction model. The statistical significance of the generated continuous models on the training set and the test set giving their coefficient of determination (R^2), the mean absolute error (MAE), and the root-mean-square error (RMSE) is given in Table 1.

Based on the statistical values, it appears that PyCaret generated a highly significant prediction model with an RMSE of 8.44 on the test set, while auto-sklearn and TPOT overfitted the prediction model on the training data. On the other hand,

Table 1. Statistical Values of the Generated Prediction Models on the Training and Test Sets

model	R^2: train	MAE: train	RMSE: train	R^2: test	MAE: test	RMSE: test
Auto-sklearn	0.94	3.52	5.54	0.55	10.23	15.80
AutoKeras	0.88	5.18	7.61	0.54	10.80	15.97
PyCaret	0.86	5.59	8.50	0.85	5.52	8.44
TPOT	0.92	4.21	6.59	0.51	9.40	15.02

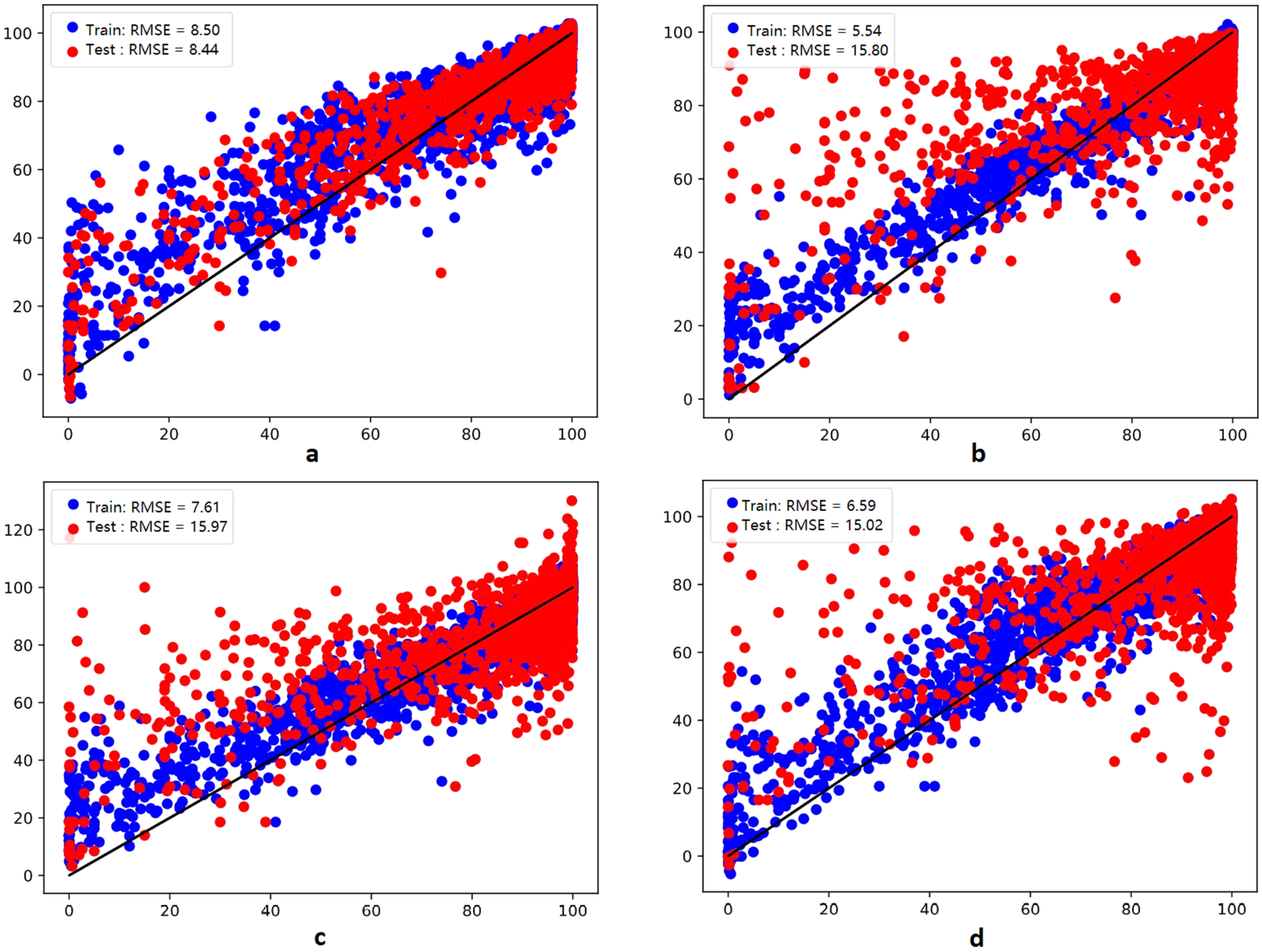


Figure 3. Plots of the experimental and predicted $f_{u,p}$ by the generated models on training and test sets: (a) PyCaret, (b) auto-sklearn, (c) AutoKeras, and (d) TPOT.

AutoKeras generated an insignificant and least reliable model with an RMSE of 15.97 on the test set. The plots between the experimental and predicted $f_{u,p}$ as predicted by the generated models are shown in Figure 3, where the X-axis and Y-axis represent the experimental and predicted values, respectively, and the blue and red colors represent the training and test sets, respectively. The black diagonal line represents the identity line.

The statistical values reveal that AutoKeras, auto-sklearn, and TPOT overfitted the model onto the training set that largely comprises compounds with high range values. The model generated by PyCaret is the best of all four and is also better than most of the previously published models. The saved file of the best model is provided in the Supporting Information file S1.zip along with the instructions for usage.

Considering the statistical results and the plots, the model generated by PyCaret can be used for prediction of fraction unbound in human plasma and PyCaret can also be used seamlessly for generating prediction models in bioinformatics.

Analysis and Comparison of the Best Model. To evaluate our best model further, the data diversity of the training and test sets as split by PyCaret is plotted as a frequency histogram of the number of compounds per 1% bracket on a log scale as shown in Figure 4, where the orange bars represent the training set, the green bars represent the test set, and the blue bars represent the complete data set. The frequency histogram plot of the training and test sets reveals that the split data sets are relatively unbiased toward high or low ranges when compared to the complete data set.

As it is known that PyCaret is a wrapper around several machine learning libraries and frameworks, the model constructed by PyCaret is further analyzed to understand the details of the best algorithm selected by the AutoML function. The best selected model is found to be the bagging regressor, with the base estimator as the LightGBM Regressor.

Further, our best model was compared with a previously published freely available prediction model using the DruMAP Web server.[10] Five hundred molecules were randomly selected from our test set, and the SMILES notion of each molecule was converted into a two-dimensional SDF format using Open Babel,[29] and their $f_{u,p}$ values were predicted using the DruMAP server. As the DruMAP server provides outputs in a range of 0−1, the predicted values are multiplied by 100 for consistency with our model and an experimental vs predicted value plot is plotted for visualization, as shown in Figure 5. The statistical RMSE value of 81.64 and the experimental vs predicted value plot on our test set suggest that the previously published model available on the DruMAP Web server is highly insignificant in predicting the $f_{u,p}$ values when compared to our model.

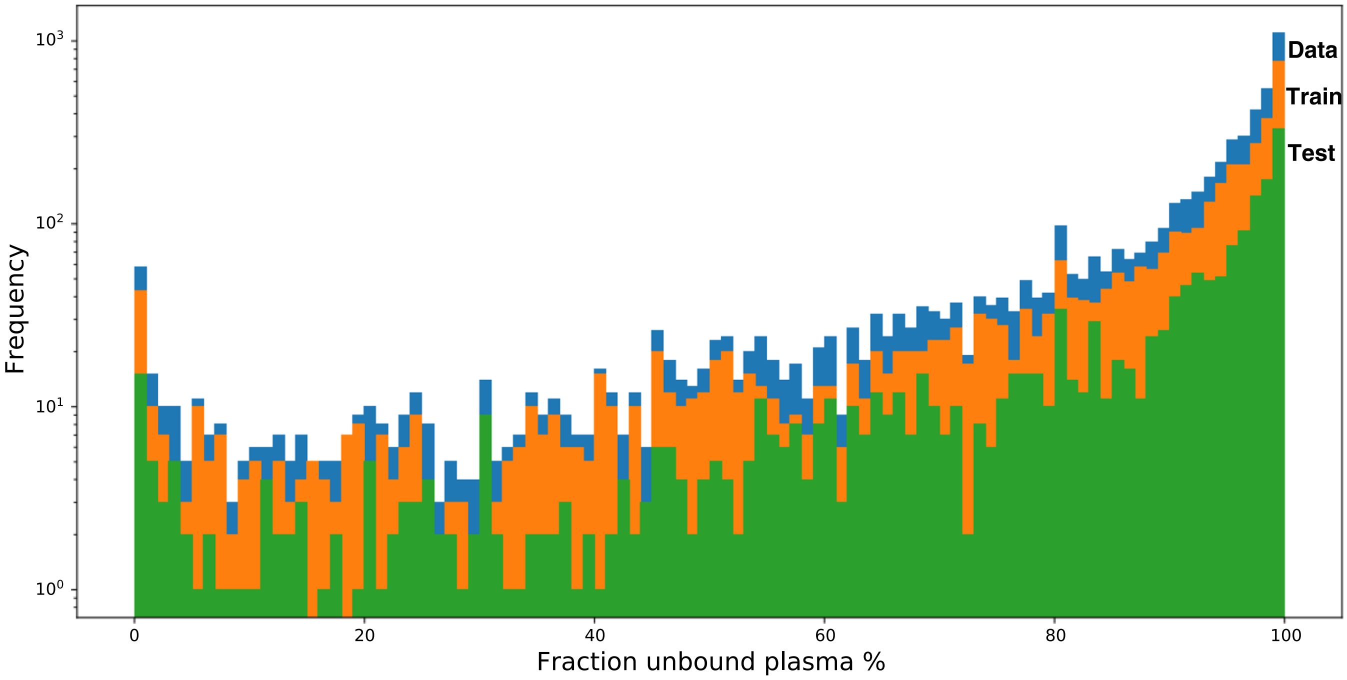


Figure 4. Histogram plot of the training and test sets of PyCaret on a logarithmic scale.

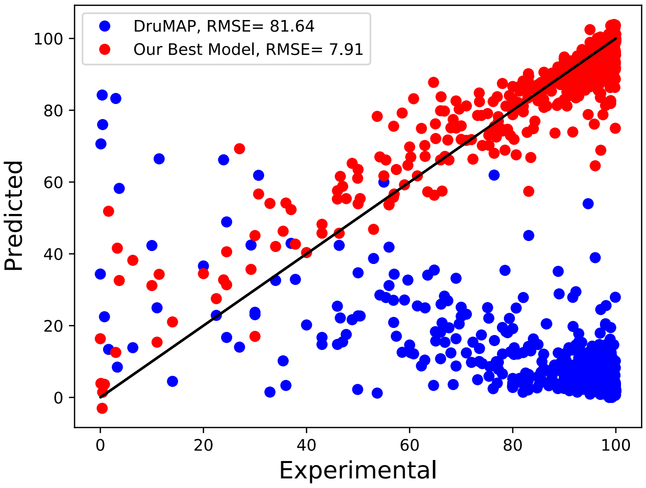


Figure 5. Plots of the experimental and predicted $f_{u,p}$ by the DruMAP server and our best model on 500 randomly selected compounds from the test set.

■ DISCUSSION

In recent years, the role of bioinformatics in drug discovery has gone a long way from target identification to diverse applications including commercial-level drug discovery.[30] To further the effectivity of academic research on in silico drug discovery, pharmacokinetic prediction models prove to be highly cost-effective in studying the pharmacokinetic properties of the drugs during the screening process. Although there are many prediction tools to study the pharmacokinetic properties of a compound, most of them are limited to lipophilicity and Lipinski's rule of five.[31]

It has been studied that protein plasma binding or fraction unbound in plasma can significantly affect the clearance, distribution, and effects of a drug, and because of this, accurate prediction models of $f_{u,p}$ are necessary to aid in in silico pharmacokinetic studies. Although $f_{u,p}$ values cannot directly ensure the success of a drug, it helps in deciding the dosage of the drug, which is a highly important aspect in toxicity studies. In this study, continuous regression models to predict the $f_{u,p}$ of the drug were constructed with a large and diverse data set of 5471 experimental records using an automated machine learning approach. To construct an optimized and reliability model, the automated machine learning technique was used to construct the model. To further help biologists who are nonexperts of machine learning in successfully constructing a prediction model, four well-known automated machine learning toolkits and libraries were evaluated for their performance on this problem. Further, considering the information represented in a PubChem fingerprint, only the PubChem fingerprints of the molecules were used to train the model.

As shown in Figure 2, the data set is weighted toward the highly unbound compounds. To confirm the statistical diversity of the data set, the SDI of the data set was calculated, which revealed a score of 0.73, confirming its diversity. The model constructed by auto-sklearn showed the lowest error rate among the four models on the training set, but its performance on test data along with TPOT was subpar when compared to other models. Statistically speaking, auto-sklearn was better than AutoKeras in all parameters, but the experimental vs prediction plot reveals that AutoKeras performed slightly better than auto-sklearn and TOPT in the lower ranges. On the other hand, although the RMSE of the model generated by PyCaret on the training set was lower than all of the other models, it performed excellently on the test set. This model also performed better than most of the previously available models and can be used for prediction of $f_{u,p}$ values. Conclusively, as only AutoKeras and PyCaret consider the test set to fine-tune the hyperparameters without directly training on them, these two AutoML frameworks can be considered to build a prediction model on biological data, with PyCaret proving better in this case. As it is known that $f_{u,p}$ plays a significant role in other pharmacokinetics such as clearance, elimination, and volume of distribution, we believe that this model will boost drug discovery in academia and this study will also help biological researchers in constructing an efficient prediction model by transiting toward the AutoML approach.

Although in recent times bioinformatics models are playing an important role in pharmaceutical and drug discovery studies, these models do have their own limitations, which have to be considered before employing their predictions in clinical studies. Major issues of the bioinformatics models are their inability to correctly model the complex biological parameters, biasness of the data obtained for the study, and overfitting of the models onto the data set. The inability to correctly predict the results of the new data when it does not fall in the training data distribution is also a major drawback of the bioinformatics prediction models.

■ CONCLUSIONS

In the present work, we curated ChEMBL, one of the largest manually curated chemical databases of bioactive druglike molecules, and generated a data set of compounds containing plasma protein binding or fraction unbound plasma data. This data set is used to develop a prediction model using automated machine learning methods, by only using the PubChem fingerprints of the compounds rather than chemical descriptors. With all of the models validated on the test set, the best model performed excellently in predicting the $f_{u,p}$ values on the test set. Our study evaluates four different AutoML toolkits on this problem to help biological researchers with no machine learning expertise transit toward machine learning and construct highly significant prediction models. In summary, we built a prediction model of $f_{u,p}$ that outperformed previously published models and can be a useful tool in pharmacokinetic modeling and in silico drug design and discovery.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c05846.

Pickle file of the best model for predictions, along with the instructions of usage and the PyCaret textual output of the best model (ZIP)

Complete data set along with training and test sets of PyCaret and the output of DruMAP used for comparative analysis (XLSX)

2D SDF format of the molecules used for DruMAP prediction (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Nidhi Mishra − *Department of Applied Sciences, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh 211015, India;* orcid.org/0000-0002-6028-7723; Email: nidhimishra@iiita.ac.in

Author

Viswajit Mulpuru − *Department of Applied Sciences, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh 211015, India;* orcid.org/0000-0001-6471-9143

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c05846

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research work was carried out in the laboratory of Chemistry & Nanotechnology, Department of Applied Sciences, Indian Institute of Information Technology Allahabad, Prayagraj, India. V.M. is thankful to the Ministry of Education, Government of India, for the fellowship to pursue doctoral degree.

■ REFERENCES

(1) Frearson, J.; Wyatt, P. Drug Discovery in Academia: The Third Way? *Expert Opin. Drug Discovery* **2010**, *5*, 909−919.

(2) Munos, B. Lessons from 60 Years of Pharmaceutical Innovation. *Nat. Rev. Drug Discovery* **2009**, *8*, 959−968.

(3) Stevens, A. J.; Jensen, J. J.; Wyller, K.; Kilgore, P. C.; Chatterjee, S.; Rohrbaugh, M. L. The Role of Public-Sector Research in the Discovery of Drugs and Vaccines. *N. Engl. J. Med.* **2011**, *364*, 535−541.

(4) Shamas-Din, A.; Schimmer, A. D. Drug Discovery in Academia. *Exp. Hematol.* **2015**, *43*, 713−717.

(5) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat. Rev. Drug Discovery* **2015**, *14*, 475−486.

(6) Bohnert, T.; Gan, L.-S. Plasma Protein Binding: From Discovery to Development. *J. Pharm. Sci.* **2013**, *102*, 2953−2994.

(7) Hall, L. M.; Hall, L. H.; Kier, L. B. QSAR Modeling of β-Lactam Binding to Human Serum Proteins. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 103−118.

(8) Lambrinidis, G.; Vallianatou, T.; Tsantili-Kakoulidou, A. In Vitro, in Silico and Integrated Strategies for the Estimation of Plasma Protein Binding. A Review. *Adv. Drug Delivery Rev.* **2015**, *86*, 27−45.

(9) Yamazaki, K.; Kanaoka, M. Computational Prediction of the Plasma Protein-binding Percent of Diverse Pharmaceutical Compounds. *J. Pharm. Sci.* **2004**, *93*, 1480−1494.

(10) Watanabe, R.; Esaki, T.; Kawashima, H.; Natsume-Kitatani, Y.; Nagao, C.; Ohashi, R.; Mizuguchi, K. Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Mol. Pharmaceutics* **2018**, *15*, 5302−5311.

(11) Zhivkova, Z.; Doytchinova, I. Quantitative Structure—Plasma Protein Binding Relationships of Acidic Drugs. *J. Pharm. Sci.* **2012**, *101*, 4627−4641.

(12) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure−Information Representation. *J. Med. Chem.* **2006**, *49*, 7169−7181.

(13) Zhu, X.-W.; Sedykh, A.; Zhu, H.; Liu, S.-S.; Tropsha, A. The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding. *Pharm. Res.* **2013**, *30*, 1790−1798.

(14) Ingle, B. L.; Veber, B. C.; Nichols, J. W.; Tornero-Velez, R. Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *J. Chem. Inf. Model.* **2016**, *56*, 2243−2252.

(15) Sun, L.; Yang, H.; Li, J.; Wang, T.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Compounds Binding to Human Plasma Proteins by QSAR Models. *ChemMedChem* **2018**, *13*, 572−581.

(16) Bhaskar, H.; Hoyle, D. C.; Singh, S. Machine Learning in Bioinformatics: A Brief Survey and Recommendations for Practitioners. *Comput. Biol. Med.* **2006**, *36*, 1104−1125.

(17) Valentini, G.; Tagliaferri, R.; Masulli, F. Computational Intelligence and Machine Learning in Bioinformatics. *Artif. Intell. Med* **2009**, *45*, 91−96.

(18) Blum, A. L.; Langley, P. Selection of Relevant Features and Examples in Machine Learning. *Artif. Intell.* **1997**, *97*, 245−271.

(19) Thornton, C.; Hutter, F.; Hoos, H. H.; Leyton-Brown, K. In *Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms*, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13, ACM Press: Chicago, Illinois, 2013; p 847.

(20) Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems*;Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; Garnett, R., Eds.; Curran Associates, Inc.: 2015; Vol. *28*, pp 2962−2970.

(21) Kotthoff, L.; Thornton, C.; Hoos, H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA. *J. Mach. Learn. Res.* **2017**, *18*, 1−5.

(22) Jin, H.; Song, Q.; Hu, X. In *Auto-Keras: An Efficient Neural Architecture Search System*, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM: Anchorage, AK, 2019; pp 1946−1956.

(23) LeDell, E.; Poirier, S. In *H2o Automl: Scalable Automatic Machine Learning*, Proceedings of the AutoML Workshop at ICML, 2020; Vol. 2020.

(24) Ali, M. PyCaret: An open source, low-code machine learning library in Python. https://www.pycaret.org (accessed Aug 10, 2020).

(25) Le, T. T.; Fu, W.; Moore, J. H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. *Bioinformatics* **2020**, *36*, 250−256.

(26) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(27) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90−95.

(28) Simpson, E. H. Measurement of Diversity. *Nature* **1949**, *163*, 688.

(29) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, No. 33.

(30) Vallance, P. Industry−Academic Relationship in a New Era of Drug Discovery. *J. Clin. Oncol.* **2016**, *34*, 3570−3575.

(31) Poda, G.; Tetko, I. *Abstracts of Papers*; 229th National Meeting of the American Chemical Society, San Diego, CA; American Chemical Society: Washington, DC, March 13−17, 2005; MEDI 514.