bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Combining Multi-Dimensional Molecular Fingerprints to Predict hERG Cardiotoxicity of Compounds

Weizhe Ding [1#], Li Zhang[1,2,3*#], Yang Nan[1], Juanshu Wu[1], Xiangxin Xin[1], Chenyang Han[1], Siyuan Li[1], Hongsheng Liu[2,3,4*]

[1]School of Life Sciences, Liaoning University, Shenyang 110036, China

[2]Technology Innovation Center for Computer Simulating and Information Processing of Bio-macromolecules of Liaoning Province, Shenyang 110036, China

[3]Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning, Liaoning University, Shenyang, 110036, China

[4]School of Pharmaceutical Sciences, Liaoning University, Shenyang 110036, China

[*]Corresponding author

Hongsheng Liu

School of Pharmaceutical Sciences, Liaoning University

Address: No. 66, Chongshan Zhonglu, Shenyang, Liaoning, 110036, China

liuhongsheng@lnu.edu.cn

Li Zhang

School of Life Sciences, Liaoning University

Address: No. 66, Chongshan Zhonglu, Shenyang, Liaoning, 110036, China

lizhang@lnu.edu.cn

[#]These co-first authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

At present, drug toxicity has become a critical problem with heavy medical and economic burdens. acLQTS (acquired Long QT Syndrome) is acquired cardiac ion channel disease caused by drugs blocking the hERG channel. Therefore, it is necessary to avoid cardiotoxicity in the drug design and computer models have been widely used to fix this plight. In this study, we present a molecular fingerprint based on the molecular dynamic simulation and uses it combined with other molecular fingerprints (multi-dimensional molecular fingerprints) to predict hERG cardiotoxicity of compounds. 203 compounds with hERG inhibitory activity (pIC50) were retrieved from a previous study and predicting models were established using four machine learning algorithms based on the single and multi-dimensional molecular fingerprints. Results showed that MDFP has the potential to be an alternative to traditional molecular fingerprints and the combination of MDFP and traditional molecular fingerprints can achieve higher prediction accuracy. Meanwhile, the accuracy of the best model, which was generated by consensus of four algorithms with multi-dimensional molecular fingerprints, was 0.694 (RMSE) in the test dataset. Besides, the number of hydrogen bonds from MDFP has been determined as a critical factor in the predicting models, followed by rgyr and sasa. Our findings provide a new sight of MDFP and multi-dimensional molecular fingerprints in building models of hERG cardiotoxicity prediction.

Keywords: Molecular dynamic simulation; Molecular fingerprint; Machine learning; hERG;

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Introduction

Drug-induced toxicity has become a critical reason for the failure of drug discovery and development in recent years (Wallace, 2015). A previous study showed that there were more than half of drugs failed (54%) in clinical development among 640 novel therapeutics, while 17% of them failed because of drug-induced toxicity (Hwang et al., 2016). Besides, it has also been reported that the mean costs required to bring a new drug to market increased from $374.1 million to $1335.9 million after counting for costs of failed trials (Wouters et al., 2016). Thus, it has become an urgent task to find ways to solve drug-induced toxicity.

Acquired Long QT syndrome (acLQTS), one of the most important diseases caused by drug-induced toxicity, is a potentially life-threatening cardiac arrhythmia disease that increases the risk for syncope, sudden cardiac death (SCD), and seizures (Tester & Ackerman, 2014). The hERG protein is a tetrameric potassium ion channel and mainly relates to cardiotoxicity and acLQTS (Liu et al., 2020). It has been reported that the potassium ion channel (hERG channel) may be blocked caused by antiarrhythmic drug binding, which leads to prolonged repolarization time and acLQTS (Witchel, 2007). At present, multiple drug candidates have failed due to the cardiotoxicity of hERG, such as cisapride, terfenadine, sertindole, pimozide, and astemizole, which have become a significant limiting factor in drug discovery and development (Bergström & Lindmark, 2019; Villoutreix & Taboureau, 2019).

Computer-aided drug design (CADD) has been thought of as an alternate choice to reduce the amount of time and money in the development of drug design,

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

especially in predicting drug toxicity (Maia et al., 2020). Molecular fingerprints are a way of CADD and are used to encoding the structure of molecules (O'Boyle et al., 2011). It has been deployed as descriptors for predicting biological activities and compound properties (Muegge & Mukherjee, 2014). Frequently used molecular fingerprints are structure-based and property-based (Kelley, 2018; Rogers & Hahn, 2010; Riniker & Landrum, 2013; Riniker, 2017). A previous study of hERG cardiotoxicity prediction showed that the accuracy of the best model developed by molecular descriptors reached 0.54 (R^2), while RMSE was 0.63 (Johnson et al., 2007). Another study of the hERG channel also showed that the accuracy of the regression model by descriptors was 0.60 (Q^2) and 0.55 (RMSE) for pIC50 (Radchenko et al., 2017). These results showed the practicalities and effectiveness based on commonly used molecular fingerprints. However, there are still no fingerprints that considered the time factor applied on the cardiotoxicity prediction of hERG.

Molecular dynamics fingerprints (MDFP) are the fingerprints based on calculating the trajectory of molecular dynamic simulation and have rapidly become a hotspot. After adding the dimension of time, MDFP can be seen as a choice of the traditional molecular fingerprint. The study of p-glycoprotein substrates prediction showed that gradient tree boosting (GTB) methods in combination with MDFP was the only model which achieved a good accuracy on the in-house dataset (Esposito et al., 2020). Meanwhile, the research of free-energy prediction showed good performance with a heterogeneous fusion model by MDFP (Riniker, 2017). Besides, studies of self-solvation free energies and application of MDFP in SAMPL6

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

octanol–water log P blind challenge also revealed a high prediction rate (Gebhardt et al., 2020; Wang & Riniker, 2019). As a consequence, MDFP can be an alternative choice of traditional molecular fingerprints and has great application potential on the cardiotoxicity prediction of hERG.

Multi-dimensional molecular fingerprints are indicated as multiple molecular fingerprints combining together in order to predict more accurately. Previous studies showed that multi-dimensional molecular fingerprints were better than the single molecular fingerprint in drug development (Kyaw et al., 2020). Thus, in this study, we studied MDFP and multi-dimensional molecular fingerprints (MDFP with other molecular fingerprints) in predicting hERG cardiotoxicity of compounds. The extensive open dataset of hERG compounds with IC50 values has been collected from previous studies. Then, molecular dynamic simulation was conducted to generate MDFP and traditional molecular fingerprints have also been generated by Baseline2D, ECFP4, and PropertyFP. Finally, the regression models were built by machine learning with four algorithms. Our study provides new sights on the combination of multi-dimensional molecular fingerprints and the research of predicting the hERG cardiotoxicity of compounds.

2. Methods

2.1. Toxicity Datasets

A hight quality hERG inhibitor dataset has been collected from the previous study (Munawar et al., 2019). The IC50 value is the biochemical half-maximal inhibitory concentration and has been used to represent the inhibiting abilities of

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compounds on hERG in this dataset (Kalliokoski et al., 2013). The data of toxicity have been eliminated if the name and IC50 values were repeated. The repeated molecules have also been averaged if the difference IC50 values were less than one order of magnitude (Feng et al., 2021). Finally, 203 compounds have been collected with specific IC50 values of the hERG. The distribution of training and testing sets followed by 80% and 20%, respectively. The training sets were used for 5-fold cross-validation and the testing sets were used to check the prediction performance of the established model for compounds. Besides, pIC50 is the negative log unit of the IC50 values and has been used to represent inhibiting abilities better than IC50 (Cortés-Ciriano et al., 2020). Therefore, IC50 of compounds was converted to pIC50 and all data has been double-checked.

2.2. MD Simulations

Molecular dynamics (MD) simulation was performed by GROMACS (2020.4). For each compound in the dataset, mol2 files were obtained from Zinc15 (http://zinc15.docking.org/) by using SMILES files. The topology of compounds was generated with AMBER14SB force field by ACPYPE (https://www.bio2byte.be/) (Sousa da Silva et al., 2012). Afterward, the compounds were placed in a dodecahedron box with a size of 1.0 nm centrally and solvated with the TIP3P water model. Then, the descent energy minimization with 100ps was applied to the system. An additional equilibration of 1ns under NVT and NPT conditions was carried out, while the constant temperature was 300 K and the constant pressure was 1 bar, respectively (Sun et al., 2020). Finally, the system was performed with running 5 ns

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MD simulation and coordinates were written every 10ps, energies every 1ps.

2.3. 2D Molecular Fingerprints

Three types of molecular fingerprints have been used in this study. Baseline2D was obtained using RDKit and its elements mainly consisted of 10 counts: number of heavy atoms, number of rotatable bonds, number of N, O, F, P, S, Cl, Br, and I atoms (Riniker, 2017; Wang & Riniker, 2019). The PropertyFP fingerprint was also obtained using the Descriptastorus package from RDKit (Kelley, 2018). It contained nearly 200 atoms features and properties. Besides, ECFP4 was generated using the RDKit implementation of the Morgan algorithm with a vector length of 2048 and a radius of 2 (Rogers & Hahn, 2010).

2.4. MD Fingerprints

The MD trajectories were analyzed by the GROMACS toolkit (Ogunwa, 2019). Following features has been selected: radius of gyration (rgyr), solvent-accessible surface area (sasa), root mean squared error (rmsd), total energy (tenergy), hydrogen bonds (hbond), kinetic energy (kinetic), Lennard-Jones short-range energies (LJ-SR) and Lennard-Jones 1-4 energies (LJ-14). The average (avr), median (mid), and standard deviation (std) of features were calculated using the R version 3.6.1 (Team, 2013). Fig. 1 showed the MDFP with all properties.

2.5. Feature Selection

Feature selection is critically important for predictive models, especially in machine learning (Johnson et al., 2018). It provides an effective way to reduce the dimensionality of data sets, identify informative features, and remove irrelevant

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

features, in order to improve the learning accuracy of machine learning models (Holder et al., 2017). In this study, zero variation and near zero variation features were deleted firstly using the nearZeroVar function in the R package caret (version 6.0–84) (Kuhn, 2008). Recursive feature elimination (RFE) was initially to perform feature selection by iteratively training in support vector machine (SVM) and has been widely used currently, especially in the random forest (RF) (Darst et al., 2018). Random Forest with Recursive Feature Elimination (RF-RFE) was a recursive backward feature elimination feature and has been used applied to predict the hERG cardiotoxicity in this study (Tang et al., 2020). Various threshold values were tested to identify the optimal RFE on the performance of a random forest model in 10 times 5-fold cross-validation.

2.6. Model Construction

In this study, RF, SVM, gradient boosting machine (GBM), and partial least square regression (PLS) was used for machine learning model construction. All models were executed beyond R (version 3.6.1) with using the randomForest (version 4.6–12) (Liaw & Wiener, 2002), the kernlab (version 0.9-25) (Karatzoglou et al., 2004), the gbm (version 2.1.5) (Brandon et al., 2019), and the pls (version 2.7-1) packages (Bjørn-Helge et al., 2019), respectively.

2.6.1 Random forest

RF is the machine learning ensemble classifier and has been applied in many fields (Breiman, 2001). By constructing multiple decision trees, the RF classifier has been considered as better performance than the single decision tree (Gandhi et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2018). In the current study, the randomforest function has been used to build RF classifiers. The number of classification trees and variables randomly selected for each node spilt have been set as ntree = 500 and mtry = the square root of the number of features, respectively. The relative importance of molecular fingerprints has also been calculated by the important function of the package.

2.6.2 Support vector machine

SVM is a generalized linear classifier based on the principle of structural risk reduction for pattern recognition (Huang et al., 2018). It is well known as a supervised learning algorithm that analyzes data and recognizes patterns (Nedaie et al., 2018). In this study, the radial basis function (RBF) kernel was used for building the SVM classifier. Meanwhile, the random search method (Bergstra & Bengio, 2012) was also applied to optimize specific SVM parameters with the regularization parameter C and σ parameter by using the caret package.

2.6.3 Gradient boosting machine

GBM is also a tree-based machine learning model. It has been considered as a step-wise, additive type model which sequentially fits new-tree-based models (Golden et al., 2019). Meanwhile, it also has many advantages, especially worked well in practice (Cho et al., 2019). In this study, the total number of trees (n.trees) and the maximum depth of each tree (interaction. depth) have been optimized by using the caret package.

2.6.4 Partial least square regression

PLS calculates a group of latent variables in connection with the output

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

maximally and determines the relationship between the input and output data (Foodeh et al., 2020). It is a stretch of the multiple linear regression models and widely used in many domains (Wu et al., 2020). Unlike multiple linear regression (MLR), it can handle the data with noisy, strongly collinear, and X-variables (Dong et al., 2018). In this study, n_components for PLS were optimized from 1 to 100.

2.7. Model Evaluation

In order to test the predictive performance of the models, 5-fold cross-validation with 10 repeats has been used to evaluate the models. After randomly divided the original dataset into five equal subsets, four of them were used for training and the other was used for testing. Then the 5-fold cross-validation was repeated ten times in order to reduce the randomness. This cross-validation progress was performed 10 times with different random seeds of 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024. Then, average values were calculated to evaluate the prediction performance of the models.

Root-mean-squared error (RMSE), mean unsigned error (MUE), and R^2 have been used to evaluate the predictive performance of the models. These indicators were calculated by the following formulas:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P-E)^2}$$

$$MUE = \frac{1}{n}\sum_{i=1}^{n}|P-E|$$

$$R^2 = 1 - \frac{\sum_i (P-E)^2}{\sum_i (\bar{E}-E)^2}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Where P, $\overline{E}$, E, n represent predictive value, the average of experimental value, experimental value, and compound numbers, respectively.

3. Results and discussion

3.1. Feature selection

In this study, 203 compounds were collected from the previous study and divided into training and testing datasets with 80% to 20%, respectively. In order to build models to predict hERG cardiotoxicity, MDFP, Baseline2D, ECFP4, and PropertyFP have been calculated for the compounds in the dataset. Table 1 illustrated the number of features calculated from each type of molecular fingerprint and the detailed description of these features is shown in the supplementary files (Table S1 and Table S2). After the feature selection by RF-RFE, 11 and 6 features have been selected from MDFP and Baseline2D, respectively. Meanwhile, there were also 99 features selected from ECFP4 and 71 from PropertyFP. Fig. 2 showed the top ten features (Baseline2D for six) which close to the prediction models by RF-RFE using %IncMSE. The results of MDFP showed that the number of hydrogen bonds between compounds and water has a significant effect on predicting hERG cardiotoxicity, followed by kinetic energy and surface area. Besides, the results of 2D molecular fingerprints indicated that the number of heavy atoms, number of O atom (oxygens), and number of F atom (fluorines) were the most important features in Baseline2D, while MolLog P in PropertyFP and X3218693969 in ECFP4. Above all, after calculating features in all molecular fingerprints, the following features have been selected as the most critical with heavyatoms, oxygens, fluorines, the median of hydrogen bonds, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

X3218693969. These features may be played important roles in predicting the hERG cardiotoxicity and should be paid extra attention in the development of drug candidates.

3.2. Prediction performance of the models

After performing feature selection, the GBM, PLS, RF, and SVM algorithms were used for generating ML models based on the resulting fingerprints. The consensus algorithm was the model which gathered the GBM, PLS, RF, and SVM models. The performance of these machine learning models was evaluated by 10 times 5-fold cross-validation and their performances were presented as follows. Table 2 and Table 3 showed the performance of models with MDFP, Baseline2D, ECFP4, and PropertyFP in the 5-fold cross-validation sets and testing sets, respectively. All models were not over-filtering and have been double-checked. Results showed that MDFP got a similar performance with other molecular fingerprints. The RMSE of all models and molecular fingerprints was between 0.860 and 1.189, 0.709 and 0.865 in the 5-fold cross-validation and testing sets, respectively. The best model in the MDFP was consensus with 0.745±0.005 (RMSE). Meanwhile, Table 4 and Table 5 showed the performance of models in the 5-fold cross-validation sets and testing sets while MDFP combined with other molecular fingerprints, respectively. The best model in the 5-fold cross-validation sets and testing sets was the SVM model of MDFP with PropertyFP (RMSE=0.837±0.006) and the consensus model of MDFP with ECFP4 (RMSE=0.694±0.002), respectively. It illustrated that the performance of multi-dimensional molecular fingerprints was better than the single molecular

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fingerprints. Fig. 3 and Fig. 4 showed the predicted values vs experimental values for MDFP and MDFP++ (MDFP, Baseline2D, ECFP4, and PropertyFP), respectively. The values of other molecular fingerprints have been demonstrated in the supplementary files (Fig. S1 to S6). As shown in Fig. 3, the consensus model was the best model which was also confirmed by the tables in MDFP. Fig. 4 also illustrated that the consensus was the best model with RMSE, followed by the SVM and PLS model.

In summary, these results illustrated that the MDFP was effective compared with traditional molecular fingerprints and can truly be an alternative choice of the other molecular fingerprints. Meanwhile, the prediction accuracies of all ML models on multi-dimensional molecular fingerprints were better than the single molecular fingerprints in predicting the hERG cardiotoxicity. Besides, the results of each model were similar, while the consensus model showed the best prediction in the testing dataset, followed by SVM and PLS model. Thus, the models obtained by multiple machine learning methods could be more accurate in predicting the hERG cardiotoxicity of compounds.

3.3. MDFP features associated with cardiotoxicity

To further reveal the contributions of fingerprint features associated with cardiotoxicity, the correlation coefficient has been used to determine the feature between MDFP and pIC50 by applying the RF algorithm. Correlation is a measure of a monotonic association between 2 variables and Pearson's correlation coefficient has become one of the most frequently used statistics (Armstrong, 2019). In this study,

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pearson, Kendall, and Spearman correlation were used to evaluate the important features of MDFP with pIC50. Table 6 showed the correlation coefficient between the feature of MDFP and pIC50. The median of rgyr has been determined as the most relevant feature with pIC50 (Kendall = 0.35, Pearson = 0.51, and Spearman = 0.49), followed by the median of sasa and kinetic. These results showed the features which extracted from MDFP had strong correlations with pIC50 and can be used to predict cardiotoxicity in the future study.

3.4. Compared with other models

Recently, a couple of computational models have been developed for toxicity prediction. Among them, cardiotoxicity prediction has become a hotspot with multiple studies. Table 7 showed the comparisons between our model and other models for cardiotoxicity prediction. Compared with other models, the consensus model with MDFP and ECFP4 showed the lowest RMSE and MUE, with higher R^2. Meanwhile, the molecular fingerprints of previous studies were used by only one dimension, which may prove that multi-dimensional fingerprints performed well in predicting the cardiotoxicity of hERG. Besides, although it was lower than QSAR-SVM, the consensus with MDFP still better than the other models as 0.745±0.005 (RMSE), which illustrated the advantages of MDFP. These findings showed that MDFP and multi-dimensional molecular fingerprints with machine learning methods can be an outstanding model in predicting cardiotoxicity.

4. Conclusion

In this study, MDFP and multi-dimensional molecular fingerprints were used for

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

building machine learning models to predict the hERG cardiotoxicity of compounds. 203 compounds were firstly identified to establish the 5-fold cross-validation and testing datasets. Then molecular dynamic simulation has been used to generate molecular dynamic molecular fingerprints. Baseline2D, ECFP4, and PropertyFP were used to generate traditional molecular fingerprints. After that, critical features have been selected by RF-RFE and 4 machine learning algorithms, namely RF, SVM, GBM, and PLS were used for building predicting models based on the single fingerprints and multi-dimensional molecular fingerprints. Besides, the correlation between MDFP and pIC50 has also been surveyed. Results showed that MDFP has the potential to be an alternative choice of molecular fingerprints and multi-dimensional molecular fingerprints are better than single fingerprints in predicting cardiotoxicity. It also illustrated that the consensus model with MDFP and ECFP4 has the optimum prediction effect and hydrogen bonds are critically important in the models with MDFP. Our finding provides a new sight into the application of MDFP and multi-dimensional molecular fingerprints in predicting the hERG cardiotoxicity of compounds. Cell and animal experiments will be carried out to validate further.

Conflict of interests

The authors declare that they have no conflict of interests.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 82003655), the Key R&D Program of Liaoning Province (No. 2019JH2/10300041),

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Scientific Research Project from Department of Education of Liaoning Province (No. LQN201906), Shenyang Science and Technology Plan Project (No. 17-65-7-00, 19-302-3-04).

Data Availability Statement

All data and models generated or used during the study appear in the submitted article.

Author contributions

WZD, LZ, and HSL conceived the project, developed the prediction method, designed, and implemented the experiments, analyzed the result, and wrote the paper. YN, JSW, and XXX implemented the experiments, analyzed the result, and wrote the paper. SYH and SYL analyzed the result. All authors read and approved the final manuscript.

References

Armstrong RA., 2019. Should Pearson's correlation coefficient be avoided? Ophthalmic Physiol Opt. 39, 316-327. https://doi.org/10.1111/opo.12636

Bergstra J., Bengio Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Bergström F., Lindmark B., 2019. Accelerated drug discovery by rapid candidate drug identification. Drug Discov Today. 24, 1237-1241. https://doi.org/10.1016/j.drudis.2019.03.026.

Bjørn-Helge M., Ron W., and Kristian L., 2019. Partial Least Squares (PLS) and Principal Component Regression. R package v2.7.1 (version 2.7.1). https://CRAN.R-project.org/package=pls

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Brandon G., Bradley B., Jay C., and GBM Developers., 2019. Generalized Boosted Regression Models (GBM). R package v2.1.5 (version 2.1.5). https://CRAN.R-project.org/package=gbm

Breiman L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cho G., Yim J., Choi Y., Ko J., Lee SH., 2019. Review of Machine Learning Algorithms for Diagnosing Mental Illness. Psychiatry Investig. 16, 262-269. https://doi.org/10.30773/pi.2018.12.21.2.

Cortés-Ciriano I., Škuta C., Bender A., Svozil D., 2020. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. 12, 41. https://doi.org/10.1186/s13321-020-00444-5.

Darst BF., Malecki KC., Engelman CD., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19, 65. https://doi.org/10.1186/s12863-018-0633-8.

Dong J., Wang NN., Yao ZJ., Zhang L., Cheng Y., Ouyang D., Lu AP., Cao DS., 2018. ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. J Cheminform. 10, 29. https://doi.org/10.1186/s13321-018-0283-x.

Esposito C., Wang S., Lange UEW., Oellien F., Riniker S., 2020. Combining Machine Learning and Molecular Dynamics to Predict P-Glycoprotein Substrates. J Chem Inf Model. 60, 4730-4749. https://doi.org/10.1021/acs.jcim.0c00525.

Feng H., Zhang L., Li S., Liu L., Yang T., Yang P., Zhao J., Arkin IT., Liu H., 2021. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fingerprints. Toxicol Lett. 340, 4-14. https://doi.org/10.1016/j.toxlet.2021.01.002.

Foodeh R., Ebadollahi S., Daliri MR., 2020. Regularized Partial Least Square Regression for Continuous Decoding in Brain-Computer Interfaces. Neuroinformatics. 18, 465-477. https://doi.org/10.1007/s12021-020-09455-x.

Gandhi K., Schmidt B., Ng A.H., 2018. Towards data mining based decision support in manufacturing maintenance. Procedia CIRP. 72, 261–265. http://doi.org/10.1016/j.procir.2018.03.076.

Gebhardt J., Kiesel M., Riniker S., Hansen N., 2020. Combining Molecular Dynamics and Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients. J Chem Inf Model. 60, 5319-5330. http://doi.org/10.1021/acs.jcim.0c00479.

Golden CE., Rothrock MJ Jr., Mishra A., 2019. Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. Food Res Int. 122, 47-55. http://doi.org/10.1016/j.foodres.2019.03.062.

Holder LB., Haque MM., Skinner MK., 2017. Machine learning for epigenetics and future medical applications. Epigenetics. 12, 505-514. http://doi.org/10.1080/15592294.2017.1329068.

Huang S., Cai N., Pacheco PP., Narrandes S., Wang Y., Xu W., 2018. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 15, 41-51. http://doi.org/10.21873/cgp.20063.

Hwang TJ., Carpenter D., Lauffenburger JC., Wang B., Franklin JM., Kesselheim AS., 2016. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results. JAMA Intern Med. 176, 1826-1833. http://doi.org/10.1001/jamainternmed.2016.6008.

Johnson SR., Yue H., Conder ML., Shi H., Doweyko AM., Lloyd J., Levesque P., 2007. Estimation of hERG inhibition of drug candidates using multivariate property and pharmacophore SAR. Bioorg Med Chem. 15, 6182-6192. http://doi.org/10.1016/j.bmc.2007.06.028.

Johnson KW., Torres Soto J., Glicksberg BS., Shameer K., Miotto R., Ali M., Ashley E., Dudley JT., 2018. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 71, 2668-2679. http://doi.org/10.1016/j.jacc.2018.03.521.

Kalliokoski T., Kramer C., Vulpetti A., Gedeck P., 2013. Comparability of mixed IC□□ data - a statistical analysis. PLoS One. 8, e61007. http://doi.org/10.1371/journal.pone.0061007.

Karatzoglou A., Smola A., Hornik K., Zeileis A., 2004. Kernel-an S4 package for kernel methods in R. J. Stat. Softw. 11, 1–20.

Kelley B. Descriptor Computation(Chemistry) and (Optional) Storage for Machine Learning. DescriptaStorus, version 2.2.0. https://github.com/bp-kelley/descriptastorus.

Kuhn M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 26, 1–26.

Kyaw Zin PP., Borrel A., Fourches D., 2020. Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict? J Chem Inf Model. 60, 3342-3360. http://doi.org/10.1021/acs.jcim.0c00200.

Liaw A., Wiener M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Liu M., Zhang L., Li S., Yang T., Liu L., Zhao J., Liu H., 2020. Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. Toxicol Lett.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

332, 88-96. http://doi.org/10.1016/j.toxlet.2020.07.003.

Maia EHB., Assis LC., de Oliveira TA., da Silva AM., Taranto AG., 2020. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. Front Chem. 8, 343. http://doi.org/10.3389/fchem.2020.00343.

Muegge I., Mukherjee P., 2016. An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov. 11, 137-148. http://doi.org/10.1517/17460441.2016.1117070.

Munawar S., Vandenberg JI., Jabeen I., 2019. Molecular Docking Guided Grid-Independent Descriptor Analysis to Probe the Impact of Water Molecules on Conformational Changes of hERG Inhibitors in Drug Trapping Phenomenon. Int J Mol Sci. 20, 3385. http://doi.org/10.3390/ijms20143385.

Nedaie A., Najafi AA., 2018. Support vector machine with Dirichlet feature mapping. Neural Netw. 98, 87-101. http://doi.org/10.1016/j.neunet.2017.11.006.

O'Boyle NM., Banck M., James CA., Morley C., Vandermeersch T., Hutchison GR., 2011. Open Babel: An open chemical toolbox. J Cheminform. 3, 33. http://doi.org/10.1186/1758-2946-3-33.

Ogunwa TH., Laudadio E., Galeazzi R., Miyanishi T., 2019. Insights into the Molecular Mechanisms of Eg5 Inhibition by (+)-Morelloflavone. Pharmaceuticals (Basel). 12, 58. http://doi.org/10.3390/ph12020058.

Radchenko EV., Rulev YA., Safanyaev AY., Palyulin VA., Zefirov NS., 2017. Computer-aided estimation of the hERG-mediated cardiotoxicity risk of potential drug components. Dokl Biochem Biophys. 473, 128-131. http://doi.org/10.1134/S1607672917020107.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Riniker S., Landrum GA., 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform. 5, 26. http://doi.org/10.1186/1758-2946-5-26.

Riniker S., 2017. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. J Chem Inf Model. 57, 726-741. http://doi.org/10.1021/acs.jcim.6b00778.

Rogers D., Hahn M., 2010. Extended-connectivity fingerprints. J Chem Inf Model. 50, 742-754. http://doi.org/10.1021/ci100050t.

Sousa da Silva AW., Vranken WF., 2012. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes. 5, 367. http://doi.org/10.1186/1756-0500-5-367.

Subramanian G., Ramsundar B., Pande V., Denny RA., 2016. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. J Chem Inf Model. 56, 1936-1949. http://doi.org/10.1186/10.1021/acs.jcim.6b00290.

Sun CP., Yan JK., Yi J., Zhang XY., Yu ZL., Huo XK., Liang JH., Ning J., Feng L., Wang C., Zhang BJ., Tian XG., Zhang L., Ma X., 2019. The study of inhibitory effect of natural flavonoids toward β-glucuronidase and interaction of flavonoids with β-glucuronidase. Int J Biol Macromol. 143, 349-358. http://doi.org/10.1016/j.ijbiomac.2019.12.057.

Tang J., Wang Y., Luo Y., Fu J., Zhang Y., Li Y., Xiao Z., Lou Y., Qiu Y., Zhu F., 2020. Computational advances of tumor marker selection and sample classification in cancer proteomics. Comput Struct Biotechnol J. 18, 2012-2025. http://doi.org/10.1016/j.csbj.2020.07.009.

R Core Team., 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tester DJ., Ackerman MJ., 2014. Genetics of long QT syndrome. Methodist Debakey Cardiovasc J. 10, 29-33. http://doi.org/10.14797/mdcj-10-1-29.

Villoutreix BO., Taboureau O., 2015. Computational investigations of hERG channel blockers: New insights and current predictive models. Adv Drug Deliv Rev. 86, 72-82. http://doi.org/10.1016/j.addr.2015.03.003.

Wallace KB., 2015. Multiple Targets for Drug-Induced Mitochondrial Toxicity. Curr Med Chem. 22, 2488-2492. http://doi.org/10.2174/0929867322666150514095424.

Wang S., Riniker S., 2020. Use of molecular dynamics fingerprints (MDFPs) in SAMPL6 octanol-water log P blind challenge. J Comput Aided Mol Des. 34, 393-403. http://doi.org/10.1007/s10822-019-00252-6.

Witchel HJ., 2007. The hERG potassium channel as a therapeutic target. Expert Opin Ther Targets. 11, 321-336. http://doi.org/10.1517/14728222.11.3.321.

Wouters OJ., McKee M., Luyten J., 2020. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 323, 844-853. http://doi.org/10.1001/jama.2020.1166.

Wu ML., Wang YT., Cheng H., Sun FL., Fei J., Sun CC., Yin JP., Zhao H., Wang YS., 2020. Phytoplankton community, structure and succession delineated by partial least square regression in Daya Bay, South China Sea. Ecotoxicology. 29, 751-761. http://doi.org/10.1007/s10646-020-02188-2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

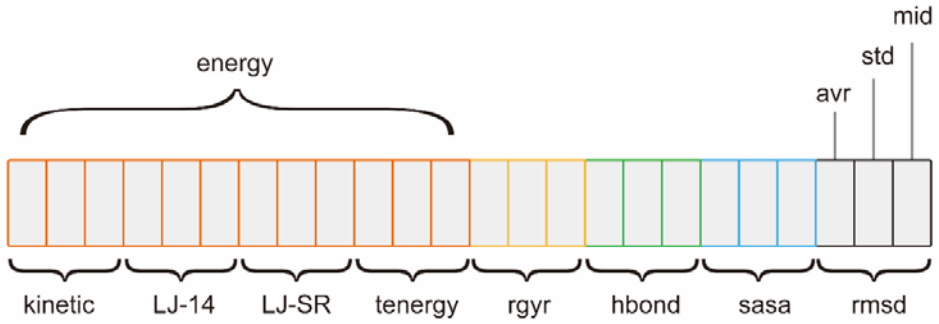

Fig.1. Schematic representation of the MDFP variant with all properties: kinetic, LJ-14, LJ-SR, tenergy, rgyr, hbond, sasa, rmsd. Each property is represented by the avr (average), std (standard deviation), and mid (median).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

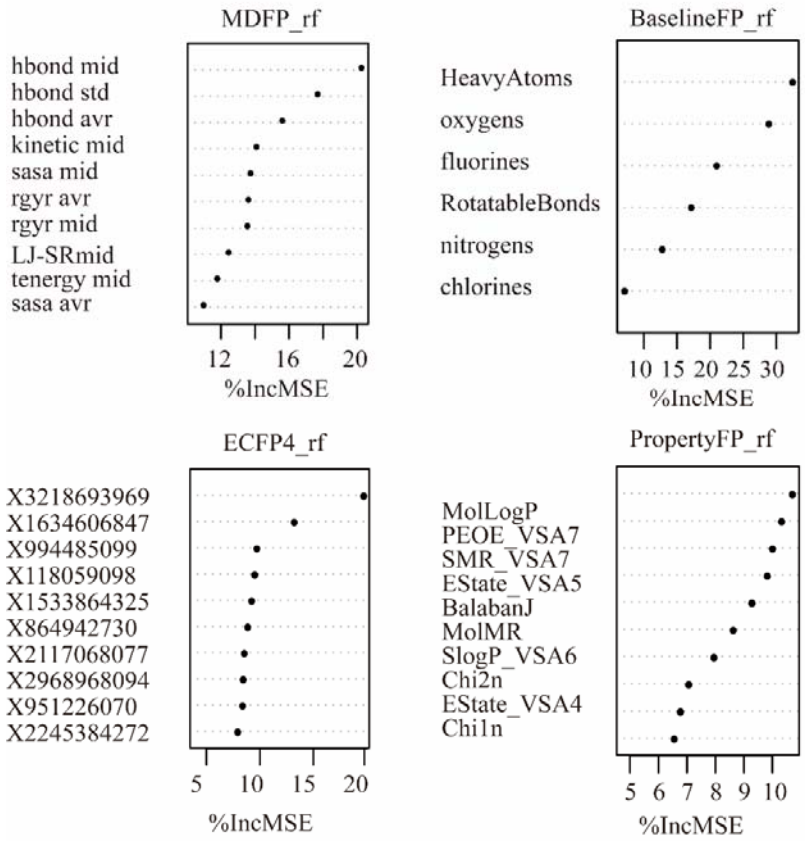

Fig. 2. The most important features selected by RF-RFE from MDFP, Baseline2D, ECFP4, and PropertyFP fingerprints.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

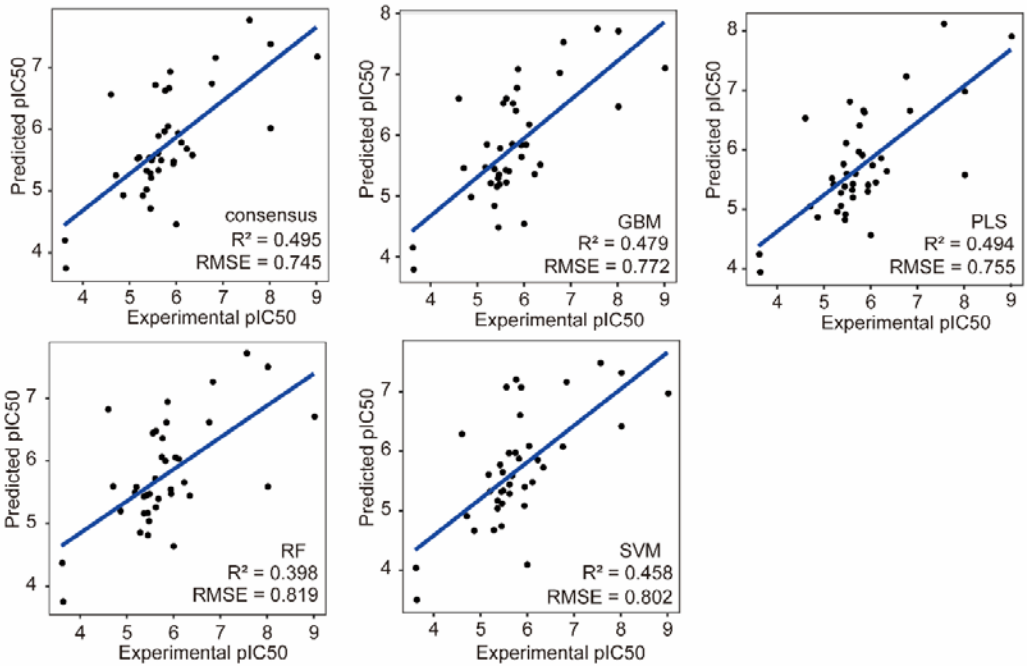

Fig. 3. pIC50: The experimental values of the 10th operation for the data set. Predictions were generated using consensus, GBM, PLS, RF, SVM trained on MDFP. The linear regression lines are shown in blue.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

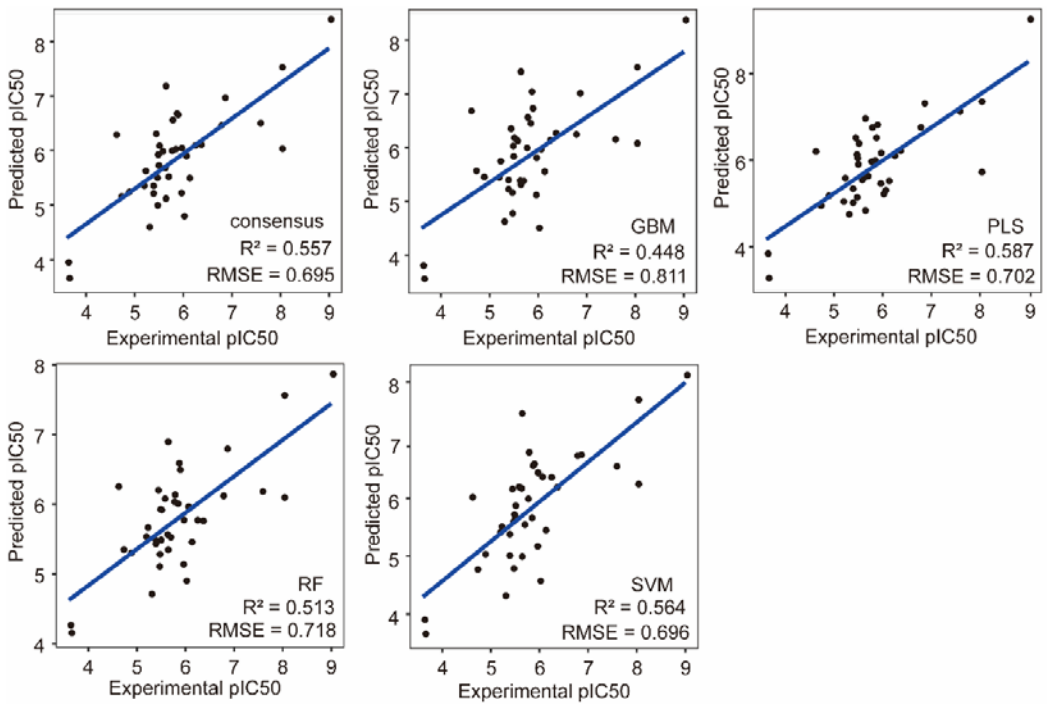

Fig. 4. pIC50: The experimental values of the 10th operation for the data set. Predictions were generated using consensus, GBM, PLS, RF, SVM trained on MDFP++. The linear regression lines are shown in blue. MDFP++ including MDFP, Baseline2D, ECFP4, and PropertyFP.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1 The number of feature selections for the different molecular fingerprints based on RF-RFE.

fingerprints	number of features	feature selection
MDFP	24	11
Baseline2D	10	6
ECFP4	2298	99
PropertyFP	200	71

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2 Predictions were generated using different ML models trained on the molecular fingerprints (MDFP, Baseline2D, ECFP4, PropertyFP) in 5-fold cross-validation.

fingerprint	ML models	RMSE	R^2	MUE
MDFP	GBM	0.985±0.005	0.523±0.004	0.774±0.005
	PLS	1.039±0.005	0.482±0.006	0.797±0.003
	RF	0.977±0.005	0.534±0.006	0.768±0.004
	SVM	0.967±0.007	0.541±0.006	0.745±0.007
Baseline2D	GBM	1.112±0.009	0.394±0.009	0.884±0.006
	PLS	1.189±0.004	0.321±0.007	0.956±0.003
	RF	1.036±0.011	0.465±0.011	0.813±0.008
	SVM	1.014±0.006	0.492±0.006	0.791±0.006
ECFP4	GBM	1.072±0.006	0.433±0.007	0.837±0.007
	PLS	1.084±0.004	0.433±0.004	0.850±0.006
	RF	1.043±0.004	0.464±0.004	0.827±0.004
	SVM	1.009±0.004	0.497±0.004	0.800±0.004
PropertyFP	GBM	0.941±0.008	0.562±0.006	0.747±0.007
	PLS	0.959±0.008	0.551±0.006	0.776±0.008
	RF	0.960±0.004	0.559±0.004	0.763±0.005
	SVM	0.860±0.006	0.634±0.006	0.676±0.009

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3 Predictions were generated using different ML models trained on the molecular fingerprints (MDFP, Baseline2D, ECFP4, PropertyFP) in test.

fingerprint	ML models	RMSE	R^2	MUE
MDFP	GBM	0.772±0.008	0.479±0.008	0.582±0.009
	PLS	0.755±0	0.494±0	0.564±0
	RF	0.819±0.011	0.398±0.012	0.570±0.006
	SVM	0.802±0.010	0.458±0.007	0.586±0.005
	consensus	0.745±0.005	0.495±0.005	0.524±0.003
Baseline2D	GBM	0.794±0.005	0.472±0.004	0.568±0.004
	PLS	0.772±0.000	0.441±0.000	0.548±0.000
	RF	0.795±0.015	0.423±0.015	0.545±0.011
	SVM	0.721±0.005	0.525±0.005	0.520±0.011
	consensus	0.713±0.003	0.520±0.004	0.507±0.002
ECFP4	GBM	0.858±0.008	0.348±0.010	0.664±0.009
	PLS	0.752±0.001	0.495±0.010	0.578±0.006
	RF	0.865±0.009	0.315±0.011	0.635±0.011
	SVM	0.737±0	0.491±0	0.553±0
	consensus	0.761±0.001	0.457±0.003	0.571±0.003
PropertyFP	GBM	0.813±0.005	0.432±0.006	0.632±0.008
	PLS	0.764±0.002	0.492±0.003	0.596±0.001
	RF	0.709±0.006	0.529±0.009	0.540±0.006
	SVM	0.761±0.035	0.488±0.040	0.605±0.033
	consensus	0.730±0.008	0.508±0.010	0.560±0.008

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4 Predictions were generated using different ML models trained on MDFP combined with multi-dimensional molecular fingerprints (MDFP + Baseline2D, MDFP + ECFP4, MDFP + PropertyFP, MDFP++) in 5-fold cross-validation. MDFP++ including MDFP, Baseline2D, ECFP4, and PropertyFP.

fingerprint	ML models	RMSE	R^2	MUE
MDFP + Baseline2D	GBM	0.991±0.005	0.516±0.005	0.767±0.005
	PLS	1.068±0.007	0.458±0.004	0.820±0.004
	RF	0.950±0.006	0.560±0.006	0.738±0.005
	SVM	0.938±0.008	0.568±0.007	0.717±0.008
MDFP + ECFP4	GBM	0.975±0.005	0.529±0.006	0.745±0.006
	PLS	1.021±0.010	0.509±0.005	0.797±0.009
	RF	0.945±0.005	0.566±0.004	0.740±0.005
	SVM	0.935±0.005	0.569±0.004	0.740±0.005
MDFP + PropertyFP	GBM	0.915±0.008	0.585±0.006	0.722±0.009
	PLS	0.948±0.011	0.568±0.009	0.754±0.011
	RF	0.944±0.005	0.578±0.004	0.742±0.004
	SVM	0.837±0.006	0.654±0.006	0.659±0.007
MDFP++	GBM	0.920±0.008	0.580±0.006	0.723±0.008
	PLS	0.958±0.007	0.556±0.005	0.754±0.007
	RF	0.940±0.005	0.578±0.004	0.742±0.005
	SVM	0.873±0.007	0.623±0.005	0.686±0.007

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 5 Predictions were generated using different ML models trained on MDFP combined with multi-dimensional molecular fingerprints (MDFP + Baseline2D, MDFP + ECFP4, MDFP + PropertyFP, MDFP++) in test. MDFP++ including MDFP, Baseline2D, ECFP4, and PropertyFP.

fingerprint	ML models	RMSE	R^2	MUE
MDFP + Baseline2D	GBM	0.728±0.008	0.525±0.008	0.544±0.008
	PLS	0.751±0.007	0.502±0.011	0.559±0.006
	RF	0.789±0.009	0.427±0.011	0.560±0.008
	SVM	0.781±0.003	0.494±0.002	0.551±0.001
	consensus	0.721±0.003	0.524±0.003	0.518±0.003
MDFP + ECFP4	GBM	0.758±0.007	0.491±0.007	0.569±0.004
	PLS	0.702±0	0.555±0	0.535±0
	RF	0.750±0.012	0.472±0.016	0.553±0.007
	SVM	0.698±0.003	0.550±0.004	0.522±0.008
	consensus	0.694±0.002	0.548±0.003	0.515±0.004
MDFP + PropertyFP	GBM	0.799±0.009	0.456±0.010	0.615±0.008
	PLS	0.794±0.000	0.481±0.004	0.610±0.003
	RF	0.709±0.008	0.527±0.011	0.549±0.009
	SVM	0.723±0.011	0.518±0.012	0.578±0.012
	consensus	0.719±0.003	0.523±0.003	0.554±0.003
MDFP++	GBM	0.811±0.008	0.448±0.009	0.619±0.009
	PLS	0.702±0	0.587±0	0.526±0
	RF	0.718±0.010	0.513±0.014	0.554±0.006
	SVM	0.696±0.015	0.564±0.011	0.516±0.017
	consensus	0.695±0.004	0.557±0.004	0.518±0.003

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 6 Correlation coefficient between the features of MDFP and pIC50.

feature	kendall	pearson	spearman
rgyr mid	0.35	0.51	0.49
sasa mid	0.30	0.41	0.43
kinetic mid	0.28	0.32	0.42
LJ-SR mid	0.28	0.25	0.41
rgyr avr	0.23	0.20	0.33
sasa avr	0.20	0.16	0.29
sasa std	0.17	0.01	0.25
hbond avr	-0.08	-0.14	-0.12
hbond std	-0.09	-0.09	-0.13
hbond mid	-0.12	-0.19	-0.17
tenergy mid	-0.28	-0.42	-0.41

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.06.447291; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table. 7 Performance indicators of several cardiotoxicity prediction models reported in the literature.

models	RMSE	R^2	MUE	Reference
QSAR-SVM	0.79 ± 0.05	0.58 ± 0.05	-	(Simeon & Jongkon, 2019)
QSAR-DNN	0.90 ± 0.06	0.49 ± 0.04	-	
MLR-Canvas	1.186	0.191	0.941	(Subramanian et al., 2016)
DNN-DeepChem	1.03	0.351	0.763	
PLS-FFD	1.07	0.48	-	(Munawar et al., 2019)
consensus-MDFP	0.745±0.045	0.495±0.005	0.524±0.003	
consensus-MDFP+ECFP4	0.694±0.002	0.548±0.003	0.515±0.004	