1   **Combining Multi-Dimensional Molecular Fingerprints to Predict hERG**

2   **Cardiotoxicity of Compounds**

3   Weizhe Ding [1][#], Li Zhang[1,2,3][*][#], Yang Nan[1], Juanshu Wu[1], Xiangxin Xin[1], Chenyang

4   Han[1], Siyuan Li[1], Hongsheng Liu[2,3,4][*]

5   [1]School of Life Sciences, Liaoning University, Shenyang 110036, China

6   [2]Technology Innovation Center for Computer Simulating and Information Processing

7   of Bio-macromolecules of Liaoning Province, Shenyang 110036, China

8   [3]Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules

9   of Liaoning, Liaoning University, Shenyang, 110036, China

10   [4]School of Pharmaceutical Sciences, Liaoning University, Shenyang 110036, China

11   [*]Corresponding author

12   Hongsheng Liu

13   School of Pharmaceutical Sciences, Liaoning University

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

15   liuhongsheng@lnu.edu.cn

16   Li Zhang

17   School of Life Sciences, Liaoning University

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

19   lizhang@lnu.edu.cn

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

21 **Abstract**

22   At present, drug toxicity has become a critical problem with heavy medical and

23 economic burdens. acLQTS (acquired Long QT Syndrome) is acquired cardiac ion

24 channel disease caused by drugs blocking the hERG channel. Therefore, it is

25 necessary to avoid cardiotoxicity in the drug design and computer models have been

26 widely used to fix this plight. In this study, we present a molecular fingerprint based

27 on the molecular dynamic simulation and uses it combined with other molecular

28 fingerprints (multi-dimensional molecular fingerprints) to predict hERG

29 cardiotoxicity of compounds. 203 compounds with hERG inhibitory activity (pIC50)

30 were retrieved from a previous study and predicting models were established using

31 four machine learning algorithms based on the single and multi-dimensional

32 molecular fingerprints. Results showed that MDFP has the potential to be an

33 alternative to traditional molecular fingerprints and the combination of MDFP and

34 traditional molecular fingerprints can achieve higher prediction accuracy. Meanwhile,

35 the accuracy of the best model, which was generated by consensus of four algorithms

36 with multi-dimensional molecular fingerprints, was 0.694 (RMSE) in the test dataset.

37 Besides, the number of hydrogen bonds from MDFP has been determined as a critical

38 factor in the predicting models, followed by rgyr and sasa. Our findings provide a new

39 sight of MDFP and multi-dimensional molecular fingerprints in building models of

40 hERG cardiotoxicity prediction.

41 **Keywords**: Molecular dynamic simulation; Molecular fingerprint; Machine learning;

42 hERG;

43 **1. Introduction**

44   Drug-induced toxicity has become a critical reason for the failure of drug

45 discovery and development in recent years (Wallace, 2015). A previous study showed

46 that there were more than half of drugs failed (54%) in clinical development among

47 640 novel therapeutics, while 17% of them failed because of drug-induced toxicity

48 (Hwang et al., 2016). Besides, it has also been reported that the mean costs required to

49 bring a new drug to market increased from $374.1 million to $1335.9 million after

50 counting for costs of failed trials (Wouters et al., 2016). Thus, it has become an urgent

51 task to find ways to identify the toxicity of compounds on a large scale in drug

52 development.

53   Acquired Long QT syndrome (acLQTS), one of the most important diseases

54 caused by drug-induced toxicity, is a potentially life-threatening cardiac arrhythmia

55 disease that increases the risk for syncope, sudden cardiac death (SCD), and seizures

56 (Tester & Ackerman, 2014). The hERG protein is a tetrameric potassium ion channel

57 and mainly relates to cardiotoxicity and acLQTS (Liu et al., 2020). It has been

58 reported that the potassium ion channel (hERG channel) may be blocked caused by

59 antiarrhythmic drug binding, which leads to prolonged repolarization time and

60 acLQTS (Witchel, 2007). At present, multiple drug candidates have failed due to the

61 cardiotoxicity of hERG, such as cisapride, terfenadine, sertindole, pimozide, and

62 astemizole, which have become a significant limiting factor in drug discovery and

63 development (Bergström & Lindmark, 2019; Villoutreix & Taboureau, 2019).

64   Computer-aided drug design (CADD) has been thought of as an alternate choice

65 to reduce the amount of time and money in the development of drug design,

66 especially in predicting drug toxicity (Maia et al., 2020). Molecular fingerprints are a

67 way of CADD and are used to encoding the structure of molecules (O'Boyle et al.,

68 2011). It has been deployed as descriptors for predicting biological activities and

69 compound properties (Muegge & Mukherjee, 2014). Frequently used molecular

70 fingerprints are structure-based and property-based (Kelley, 2018; Rogers & Hahn,

71 2010; Riniker & Landrum, 2013; Riniker, 2017). A previous study of hERG

72 cardiotoxicity prediction showed that the accuracy of the best model developed by

73 molecular descriptors reached 0.54 (R^2), while RMSE was 0.63 (Johnson et al., 2007).

74 Another study of the hERG channel also showed that the accuracy of the regression

75 model by descriptors was 0.60 (Q^2) and 0.55 (RMSE) for pIC50 (Radchenko et al.,

76 2017). These results showed the practicalities and effectiveness based on commonly

77 used molecular fingerprints. However, there are still no fingerprints that considered

78 the time factor applied on the cardiotoxicity prediction of hERG.

79   Molecular dynamics fingerprints (MDFP) are the fingerprints based on

80 calculating the trajectory of molecular dynamic simulation and have rapidly become a

81 hotspot. After adding the dimension of time, MDFP can be seen as a choice of the

82 traditional molecular fingerprint. The study of p-glycoprotein substrates prediction

83 showed that gradient tree boosting (GTB) methods in combination with MDFP was

84 the only model which achieved a good accuracy on the in-house dataset (Esposito et

85 al., 2020). Meanwhile, the research of free-energy prediction showed good

86 performance with a heterogeneous fusion model by MDFP (Riniker, 2017). Besides,

87   studies of self-solvation free energies and application of MDFP in SAMPL6

88   octanol–water log P blind challenge also revealed a high prediction rate (Gebhardt et

89   al., 2020; Wang & Riniker, 2019). As a consequence, MDFP can be an alternative

90   choice of traditional molecular fingerprints and has great application potential on the

91   cardiotoxicity prediction of hERG.

92   Multi-dimensional molecular fingerprints are indicated as multiple molecular

93   fingerprints combining together in order to predict more accurately. Previous studies

94   showed that multi-dimensional molecular fingerprints were better than the single

95   molecular fingerprint in drug development (Kyaw et al., 2020). Thus, in this study, we

96   studied MDFP and multi-dimensional molecular fingerprints (MDFP with other

97   molecular fingerprints) in predicting hERG cardiotoxicity of compounds. The

98   extensive open dataset of hERG compounds with IC50 values has been collected from

99   previous studies. Then, molecular dynamic simulation was conducted to generate

100   MDFP and traditional molecular fingerprints have also been generated by Baseline2D,

101   ECFP4, and PropertyFP. Finally, the regression models were built by machine

102   learning with four algorithms. Our study provides new sights on the combination of

103   multi-dimensional molecular fingerprints and the research of predicting the hERG

104   cardiotoxicity of compounds.

105   **2.   Methods**

106   **2.1. Toxicity Datasets**

107   A high-quality hERG inhibitor dataset has been collected from the previous

108   study (Munawar et al., 2019). The IC50 value is the biochemical half-maximal

109   inhibitory concentration and has been used to represent the inhibiting abilities of

110   compounds on hERG in this dataset (Kalliokoski et al., 2013). The data of toxicity

111   have been eliminated if the name and IC50 values were repeated. The repeated

112   molecules have also been averaged if the difference IC50 values were less than one

113   order of magnitude (Feng et al., 2021). Finally, 203 compounds have been collected

114   with specific IC50 values of the hERG. The distribution of training and testing sets

115   followed by 80% and 20%, respectively. The training sets were used for 5-fold

116   cross-validation and the testing sets were used to check the prediction performance of

117   the established model for new compounds. Besides, pIC50 is the negative log unit of

118   the IC50 values and has been used to represent inhibiting abilities better than IC50

119   (Cortés-Ciriano et al., 2020). Therefore, IC50 of compounds was converted to pIC50.

120   **2.2. MD Simulations**

121   Molecular dynamics (MD) simulation was performed by GROMACS (2020.4).

122   For compounds in the dataset, mol2 files were obtained from Zinc15

123   (http://zinc15.docking.org/) by using SMILES files. The topology of compounds was

124   generated with AMBER14SB force field by ACPYPE (https://www.bio2byte.be/)

125   (Sousa da Silva et al., 2012). Afterward, the compounds were placed in a

126   dodecahedron box with a size of 1.0 nm centrally and solvated with the TIP3P water

127   model. Then, the descent energy minimization with 100ps was applied to the system.

128   An additional equilibration of 1ns under NVT and NPT conditions was carried out,

129   while the constant temperature was 300 K and the constant pressure was 1 bar,

130   respectively (Sun et al., 2020). Finally, the system was performed with running 5 ns

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

132   **2.3. 2D Molecular Fingerprints**

133   Three types of molecular fingerprints have been used in this study. Baseline2D

134   was obtained using RDKit and its elements mainly consisted of 10 counts: number of

135   heavy atoms, number of rotatable bonds, number of N, O, F, P, S, Cl, Br, and I atoms

136   (Riniker, 2017; Wang & Riniker, 2019). The PropertyFP fingerprint was also obtained

137   using the Descriptastorus package from RDKit (Kelley, 2018). It contained nearly 200

138   atoms features and properties. Besides, ECFP4 was generated using the RDKit

139   implementation of the Morgan algorithm with a vector length of 2048 and a radius of

140   2 (Rogers & Hahn, 2010).

141   **2.4. MD Fingerprints**

142   The MD trajectories were analyzed by the GROMACS toolkit (Ogunwa, 2019).

143   Following features has been generated: radius of gyration (rgyr), solvent-accessible

144   surface area (sasa), root mean squared error (rmsd), total energy (tenergy), hydrogen

145   bonds (hbond), kinetic energy (kinetic), Lennard-Jones short-range energies (LJ-SR)

146   and Lennard-Jones 1-4 energies (LJ-14). The average (avr), median (mid), and

147   standard deviation (std) of features were calculated using the R version 3.6.1 (Team,

148   2013). Fig. 1 showed the MDFP with all properties.

149   **2.5. Feature Selection**

150   Feature selection is critically important for predictive models, especially in

151   machine learning (Johnson et al., 2018). It provides an effective way to reduce the

152   dimensionality of data sets, identify informative features, and remove irrelevant

153 features, improving the learning accuracy of machine learning models (Holder et al.,

154 2017). In this study, zero variation and near-zero variation features were deleted

155 firstly using the nearZeroVar function in the R package caret (version 6.0–84) (Kuhn,

156 2008). Then, recursive feature elimination (RFE) was performed to select the optimal

157 feature subset using the rfe function in the caret package in a 10 times 5-fold

158 cross-validation setting (Darst et al., 2018). In the RFE process, all features are first

159 ranked according to the feature importance values obtained by the random forest (RF)

160 algorithm, and then RF models are trained iteratively on the features that are gradually

161 reduced according to the ranking to evaluate the performance of the feature subsets

162 (Tang et al., 2020).

163 **2.6. Model Construction**

164   In this study, RF, SVM, gradient boosting machine (GBM), and partial least

165 square regression (PLS) was used for machine learning model construction. All

166 models were executed beyond R (version 3.6.1) with using the randomForest (version

167 4.6–12) (Liaw & Wiener, 2002), the kernlab (version 0.9-25) (Karatzoglou et al.,

168 2004), the gbm (version 2.1.5) (Brandon et al., 2019), and the pls (version 2.7-1)

169 packages (Bjørn-Helge et al., 2019), respectively.

170 *2.6.1 Random forest*

171   RF is the machine learning ensemble classifier and has been applied in many

172 fields (Breiman, 2001). By constructing multiple decision trees, the RF classifier has

173 been considered as better performance than the single decision tree (Gandhi et al.,

174 2018). In the current study, the randomforest function has been used to build RF

175 classifiers. The number of classification trees and variables randomly selected for

176 each node spilt have been set as ntree = 500, while mtry was optimized from one-third

177 of the number of features minus 10 to plus 15. The relative importance of molecular

178 fingerprints has also been calculated by the important function of the package.

179 *2.6.2 Support vector machine*

180   SVM is a generalized linear classifier based on the principle of structural risk

181 reduction for pattern recognition (Huang et al., 2018). It is well known as a supervised

182 learning algorithm that analyzes data and recognizes patterns (Nedaie et al., 2018). In

183 this study, the radial basis function (RBF) kernel was used for building the SVM

184 classifier. Meanwhile, the random search method (Bergstra & Bengio, 2012) was also

185 applied to optimize specific SVM parameters with the regularization parameter C and

186 σ parameter by using the caret package, while C was from e^{-2} to e^{6}, σ was from e^{-7}

187 to e with the step of $e^{0.5}$.

188 *2.6.3 Gradient boosting machine*

189   GBM is also a tree-based machine learning model. It has been considered as a

190 step-wise, additive type model which sequentially fits new-tree-based models (Golden

191 et al., 2019). Meanwhile, it also has many advantages, especially worked well in

192 practice (Cho et al., 2019). In this study, the total number of trees (n.trees) and the

193 maximum depth of each tree (interaction. depth) have been optimized by using the

194 caret package and have been set from 1 to 3000 and 1 to 10, respectively. Besides,

195 shrinkage and n.minobsinnode were set as 0.005 and 10.

196 *2.6.4 Partial least square regression*

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

198   maximally and determines the relationship between the input and output data (Foodeh

199   et al., 2020). It is a stretch of the multiple linear regression models and is widely used

200   in many domains (Wu et al., 2020). Unlike multiple linear regression (MLR), it can

201   handle the data with noisy, strongly collinear, and X-variables (Dong et al., 2018). In

202   this study, n_components for PLS were optimized from 1 to the greatest features or

203   sample sizes.

204   **2.7. Model Evaluation**

205   In order to test the predictive performance of the models, 5-fold cross-validation

206   with 10 repeats has been used to evaluate the models. After randomly divided the

207   original dataset into five equal subsets, four of them were used for training and the

208   other was used for testing. Then the 5-fold cross-validation was repeated ten times to

209   reduce the randomness. This cross-validation progress was performed 10 times with

210   different random seeds of 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024. Then, average

211   values were calculated to evaluate the prediction performance of the models.

212   Root-mean-squared error (RMSE), mean unsigned error (MUE), and R^2 has been

213   used to evaluate the predictive performance of the models. These indicators were

214   calculated by the following formulas:

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

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

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

215   Where P, $\bar{E}$, E, n represent predictive value, the average of experimental value,

216   experimental value, and compound numbers, respectively.

217   **3.   Results and discussion**

218   **3.1. Feature selection**

219   In this study, 203 compounds were collected from the previous study and divided

220   into training and testing datasets with 80% to 20%, respectively. In order to build

221   models to predict hERG cardiotoxicity, MDFP, Baseline2D, ECFP4, and PropertyFP

222   have been calculated for the compounds in the dataset. Table 1 illustrated the number

223   of features calculated from each type of molecular fingerprint and the detailed

224   description of these features is shown in the supplementary files (Table S1 and Table

225   S2). After the feature selection by RF-RFE, 11 and 6 features have been selected from

226   MDFP and Baseline2D, respectively. Meanwhile, there were also 99 features selected

227   from ECFP4 and 71 from PropertyFP. Percentage increase in MSE (%lncMSE)

228   obtained by RF was used to evaluate the importance of features. Fig. 2 showed the top

229   ten features (Baseline2D for six) which important to the prediction models. The

230   results of MDFP showed that the number of hydrogen bonds between compounds and

231   water has a significant effect on predicting hERG cardiotoxicity, followed by kinetic

232   energy and surface area. Besides, the results of 2D molecular fingerprints indicated

233   that the number of heavy atoms, number of O atoms (oxygens), and number of F

234   atoms (fluorines) were the most important features in Baseline2D, while MolLog P in

235   PropertyFP and 3218693969 in ECFP4. Above all, after calculating features in all

236   molecular fingerprints, the following features have been selected as the most critical

237   with heavyatoms, oxygens, fluorines, the median of hydrogen bonds, and 3218693969.

238   These features may be played important roles in predicting the hERG cardiotoxicity

239   and should be paid extra attention in the development of drug candidates.

240   **3.2. Prediction performance of the models**

241   After performing feature selection, the GBM, PLS, RF, and SVM algorithms

242   were used for generating ML models based on the resulting fingerprints. The

243   performance of these machine learning models was evaluated by 10 times 5-fold

244   cross-validation and their performances were presented in Table 2. The results showed

245   that the RMSE of each machine learning model based on PropertyFP is the lowest,

246   with a range of 0.860-0.960, followed by MDFP, with a range of 0.967-1.039, while

247   ECFP4 and Baseline2D are poor quality. R^2 and MUE also showed the same pattern.

248   Table 3 illustrated the performance of these models which were used to predict the

249   pIC50 of the molecules in the testing set. In general, the models show better RMSE

250   values in the test set than in the 5-fold cross-validation, indicating that the model has

251   not been overfitted. Meanwhile, compared with the models based on different

252   molecular fingerprints, the performance in the testing set was similar, while

253   Baseline2D was slightly better (RMSE=0.721 to 0.795) and MDFP also obtained a

254   good performance (RMSE=0.755 to 0.819). These results indicated that MDFP can

255   effectively predict the activity of hERG inhibitors, and the predictive performance of

256   the MDFP was similar to the traditional molecular fingerprints.

257   The predictive performance of the MDFP model combined with other molecular

258   fingerprints was also investigated in this study. Table 4 and Table 5 showed the

259   performance of models in the 5-fold cross-validation sets and testing sets while

260   MDFP combined with other molecular fingerprints, respectively. The results showed

261   that the combination of MDFP and other molecular fingerprints can obtain a model

262   with better prediction performance. For example, the model established by the single

263   molecular fingerprint (MDFP or PropertyFP) in the 5-fold cross-validation had the

264   best performance as PropertyFP-SVM (RMSE=0.860). However, the model

265   established by multi-dimensional molecular fingerprints (MDFP and PropertyFP) was

266   MDFP+PropertyFP-SVM (RMSE=0.837), which showed a better performance than

267   using the single molecular fingerprints. Besides, models combining MDFP with other

268   molecular fingerprints also showed better predictive performance in the testing set

269   (Table 5), while the best model was the SVM model trained on MDFP++ (MDFP with

270   all other fingerprints) (RMSE=0.696±0.015). These results illustrated that the

271   performance of multi-dimensional molecular fingerprints was better than the single

272   molecular fingerprints and MDFP may provide additional effective predictors for the

273   prediction of hERG inhibitor activity.

274   In order to improve the prediction performance of the model, we further

275   averaged the prediction results of the four machine learning models to obtain a

276   consensus value. The prediction performance was shown in Table 3 and Table 5. Fig.

277   3 and Fig. 4 showed the predicted values vs experimental values for MDFP and

278   MDFP++, respectively. The values of other molecular fingerprints have been

279   demonstrated in the supplementary files (Fig. S1 to S6). It was found that the

280   performance of consensus models was significantly better than the other models

281   (except PropertyFP). Among the models established by a single molecular fingerprint,

282   the consensus model based on Baseline2D had the highest accuracy (RMSE=0.713),

283   while the consensus model based on MDFP also obtained a better RMSE of

284   0.745. Meanwhile, in the model based on the multi-dimensional molecular

285   fingerprints, MDFP+ECFP4 and MDFP++ obtained high accuracy with RMSE of

286   0.694 and 0.695, respectively. These results indicated that the integrated model can

287   obtain a better method for predicting the activity of hERG inhibitors.

288   In summary, these results illustrated that the MDFP was effective compared with

289   traditional molecular fingerprints and can truly be an alternative to the other

290   molecular fingerprints. Meanwhile, the prediction accuracies of all ML models on

291   multi-dimensional molecular fingerprints were better than the single molecular

292   fingerprints in predicting the hERG cardiotoxicity. Besides, the integrated models

293   showed the best prediction than the single models among most of the molecular

294   fingerprints. Thus, the models obtained by multiple machine learning methods could

295   be more accurate in predicting the hERG cardiotoxicity of compounds.

296   **3.3. MDFP features associated with cardiotoxicity**

297   To further reveal the contributions of fingerprint features associated with

298   cardiotoxicity, the correlation coefficient has been used to determine the feature

299   between MDFP and pIC50. Correlation is a measure of a monotonic association

300   between 2 variables and Pearson's correlation coefficient has become one of the most

301   frequently used statistics (Armstrong, 2019). In this study, Pearson, Kendall, and

302   Spearman correlations were used to evaluate the important features of MDFP with

303   pIC50. Table 6 showed the correlation coefficient between the feature of MDFP and

304   pIC50. The median of rgyr has been determined as the most relevant feature with

305   pIC50 (Kendall = 0.35, Pearson = 0.51, and Spearman = 0.49), followed by the

306   median of sasa and kinetic with the high correlation coefficient. These results showed

307   the features which extracted from MDFP had strong correlations with pIC50 and can

308   be used to predict cardiotoxicity in the future study.

309   **3.4. Compared with other models**

310   Recently, a couple of computational models have been developed for toxicity

311   prediction. Among them, cardiotoxicity prediction has become a hotspot with multiple

312   studies. Table 7 showed the comparisons between our model and other models for

313   cardiotoxicity prediction. Compared with other models, the consensus model with

314   MDFP and ECFP4 showed the lowest RMSE and MUE, with higher R^2. Meanwhile,

315   the molecular fingerprints of previous studies were used by only one dimension,

316   which may prove that multi-dimensional fingerprints performed well in predicting the

317   cardiotoxicity of hERG. Besides, although it was lower than QSAR-SVM, the

318   consensus with MDFP still better than the other models as 0.745±0.005 (RMSE),

319   which illustrated the advantages of MDFP. These findings showed that MDFP and

320   multi-dimensional molecular fingerprints with machine learning methods can be an

321   outstanding model in predicting cardiotoxicity.

322   **4.   Conclusion**

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

324   building machine learning models to predict the hERG cardiotoxicity of compounds.

325   203 compounds were firstly identified to establish the 5-fold cross-validation and

326   testing datasets. Then molecular dynamic simulation has been used to generate

327   molecular dynamic molecular fingerprints. Baseline2D, ECFP4, and PropertyFP were

328   used to generate traditional molecular fingerprints. After that, critical features have

329   been selected by RF-RFE and 4 machine learning algorithms, namely RF, SVM,

330   GBM, and PLS were used for building predicting models based on the single

331   fingerprints and multi-dimensional molecular fingerprints. Besides, the correlation

332   between MDFP and pIC50 has also been surveyed. Results showed that MDFP has

333   the potential to be an alternative choice of molecular fingerprints and

334   multi-dimensional molecular fingerprints are better than single fingerprints in

335   predicting cardiotoxicity. It also illustrated that the consensus model with MDFP and

336   ECFP4 has the optimum prediction effect and hydrogen bonds are critically important

337   in the models with MDFP. Our finding provides a new sight into the application of

338   MDFP and multi-dimensional molecular fingerprints in predicting the hERG

339   cardiotoxicity of compounds. Cell and animal experiments will be carried out to

340   validate further.

341   **Conflict of interests**

342   The authors declare that they have no conflict of interests.

343   **Acknowledgements**

346 Scientific Research Project from Department of Education of Liaoning Province (No.

347 LQN201906), Shenyang Science and Technology Plan Project (No. 17-65-7-00,

348 19-302-3-04).

349 **Data Availability Statement**

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

351 article.

352 **Author contributions**

353 WZD, LZ, and HSL conceived the project, developed the prediction method, designed,

354 and implemented the experiments, analyzed the result, and wrote the paper. YN, JSW,

355 and XXX implemented the experiments, analyzed the result, and wrote the paper.

356 SYH and SYL analyzed the result. All authors read and approved the final manuscript.

357 **References**

358 Armstrong RA., 2019. Should Pearson's correlation coefficient be avoided? Ophthalmic Physiol

359 Opt. 39, 316-327. https://doi.org/10.1111/opo.12636

360 Bergstra J., Bengio Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn.

361 Res. 13, 281–305.

362 Bergström F., Lindmark B., 2019. Accelerated drug discovery by rapid candidate drug

363 identification. Drug Discov Today. 24, 1237-1241.

364 https://doi.org/10.1016/j.drudis.2019.03.026.

365 Bjørn-Helge M., Ron W., and Kristian L., 2019. Partial Least Squares (PLS) and Principal

366 Component Regression. R package v2.7.1 (version 2.7.1).

367 https://CRAN.R-project.org/package=pls

368   Brandon G., Bradley B., Jay C., and GBM Developers., 2019. Generalized Boosted Regression

369   Models (GBM). R package v2.1.5 (version 2.1.5). https://CRAN.R-project.org/package=gbm

370   Breiman   L.,   2001.   Random   Forests.   Mach.   Learn.   45,   5–32.

371   https://doi.org/10.1023/A:1010933404324.

372   Cho G., Yim J., Choi Y., Ko J., Lee SH., 2019. Review of Machine Learning Algorithms for

373   Diagnosing   Mental   Illness.   Psychiatry   Investig.   16,   262-269.

374   https://doi.org/10.30773/pi.2018.12.21.2.

375   Cortés-Ciriano I., Škuta C., Bender A., Svozil D., 2020. QSAR-derived affinity fingerprints (part

376   2):   modeling   performance   for   potency   prediction.   12,   41.

377   https://doi.org/10.1186/s13321-020-00444-5.

378   Darst BF., Malecki KC., Engelman CD., 2018. Using recursive feature elimination in random

379   forest to account for correlated variables in high dimensional data. BMC Genet. 19, 65.

380   https://doi.org/10.1186/s12863-018-0633-8.

381   Dong J., Wang NN., Yao ZJ., Zhang L., Cheng Y., Ouyang D., Lu AP., Cao DS., 2018.

382   ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively

383   collected   ADMET   database.   J   Cheminform.   10,   29.

384   https://doi.org/10.1186/s13321-018-0283-x.

385   Esposito C., Wang S., Lange UEW., Oellien F., Riniker S., 2020. Combining Machine Learning

386   and Molecular Dynamics to Predict P-Glycoprotein Substrates. J Chem Inf Model. 60,

387   4730-4749. https://doi.org/10.1021/acs.jcim.0c00525.

388   Feng H., Zhang L., Li S., Liu L., Yang T., Yang P., Zhao J., Arkin IT., Liu H., 2021. Predicting

389   the reproductive toxicity of chemicals using ensemble learning methods and molecular

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

391   Foodeh R., Ebadollahi S., Daliri MR., 2020. Regularized Partial Least Square Regression for

392   Continuous Decoding in Brain-Computer Interfaces. Neuroinformatics. 18, 465-477.

393   https://doi.org/10.1007/s12021-020-09455-x.

394   Gandhi K., Schmidt B., Ng A.H., 2018. Towards data mining based decision support in

395   manufacturing maintenance. Procedia CIRP. 72, 261–265.

396   http://doi.org/10.1016/j.procir.2018.03.076.

397   Gebhardt J., Kiesel M., Riniker S., Hansen N., 2020. Combining Molecular Dynamics and

398   Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients.

399   J Chem Inf Model. 60, 5319-5330. http://doi.org/10.1021/acs.jcim.0c00479.

400   Golden CE., Rothrock MJ Jr., Mishra A., 2019. Comparison between random forest and gradient

401   boosting machine methods for predicting Listeria spp. prevalence in the environment of

402   pastured poultry farms. Food Res Int. 122, 47-55.

403   http://doi.org/10.1016/j.foodres.2019.03.062.

404   Holder LB., Haque MM., Skinner MK., 2017. Machine learning for epigenetics and future

405   medical applications. Epigenetics. 12, 505-514.

406   http://doi.org/10.1080/15592294.2017.1329068.

407   Huang S., Cai N., Pacheco PP., Narrandes S., Wang Y., Xu W., 2018. Applications of Support

408   Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 15,

409   41-51. http://doi.org/10.21873/cgp.20063.

410   Hwang TJ., Carpenter D., Lauffenburger JC., Wang B., Franklin JM., Kesselheim AS., 2016.

411   Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial

412   Results. JAMA Intern Med. 176, 1826-1833.

413   http://doi.org/10.1001/jamainternmed.2016.6008.

414   Johnson SR., Yue H., Conder ML., Shi H., Doweyko AM., Lloyd J., Levesque P., 2007.

415   Estimation of hERG inhibition of drug candidates using multivariate property and

416   pharmacophore SAR. Bioorg Med Chem. 15, 6182-6192.

417   http://doi.org/10.1016/j.bmc.2007.06.028.

418   Johnson KW., Torres Soto J., Glicksberg BS., Shameer K., Miotto R., Ali M., Ashley E., Dudley

419   JT., 2018. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 71, 2668-2679.

420   http://doi.org/10.1016/j.jacc.2018.03.521.

421   Kalliokoski T., Kramer C., Vulpetti A., Gedeck P., 2013. Comparability of mixed IC_{50} data - a

422   statistical analysis. PLoS One. 8, e61007. http://doi.org/10.1371/journal.pone.0061007.

423   Karatzoglou A., Smola A., Hornik K., Zeileis A., 2004. Kernel-an S4 package for kernel methods

424   in R. J. Stat. Softw. 11, 1–20.

425   Kelley B. Descriptor Computation(Chemistry) and (Optional) Storage for Machine Learning.

426   DescriptaStorus, version 2.2.0. https://github.com/bp-kelley/descriptastorus.

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

428   Kyaw Zin PP., Borrel A., Fourches D., 2020. Benchmarking 2D/3D/MD-QSAR Models for

429   Imatinib Derivatives: How Far Can We Predict? J Chem Inf Model. 60, 3342-3360.

430   http://doi.org/10.1021/acs.jcim.0c00200.

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

432   Liu M., Zhang L., Li S., Yang T., Liu L., Zhao J., Liu H., 2020. Prediction of hERG potassium

433   channel blockage using ensemble learning methods and molecular fingerprints. Toxicol Lett.

434   332, 88-96. http://doi.org/10.1016/j.toxlet.2020.07.003.

435   Maia EHB., Assis LC., de Oliveira TA., da Silva AM., Taranto AG., 2020. Structure-Based

436   Virtual Screening: From Classical to Artificial Intelligence. Front Chem. 8, 343.

437   http://doi.org/10.3389/fchem.2020.00343.

438   Muegge I., Mukherjee P., 2016. An overview of molecular fingerprint similarity search in virtual

439   screening.   Expert   Opin   Drug   Discov.   11,   137-148.

440   http://doi.org/10.1517/17460441.2016.1117070.

441   Munawar S., Vandenberg JI., Jabeen I., 2019. Molecular Docking Guided Grid-Independent

442   Descriptor Analysis to Probe the Impact of Water Molecules on Conformational Changes of

443   hERG Inhibitors in Drug Trapping Phenomenon. Int J Mol Sci. 20, 3385.

444   http://doi.org/10.3390/ijms20143385.

445   Nedaie A., Najafi AA., 2018. Support vector machine with Dirichlet feature mapping. Neural

446   Netw. 98, 87-101. http://doi.org/10.1016/j.neunet.2017.11.006.

447   O'Boyle NM., Banck M., James CA., Morley C., Vandermeersch T., Hutchison GR., 2011. Open

448   Babel: An   open   chemical   toolbox.   J   Cheminform.   3,   33.

449   http://doi.org/10.1186/1758-2946-3-33.

450   Ogunwa TH., Laudadio E., Galeazzi R., Miyanishi T., 2019. Insights into the Molecular

451   Mechanisms of Eg5 Inhibition by (+)-Morelloflavone. Pharmaceuticals (Basel). 12, 58.

452   http://doi.org/10.3390/ph12020058.

453   Radchenko EV., Rulev YA., Safanyaev AY., Palyulin VA., Zefirov NS., 2017. Computer-aided

454   estimation of the hERG-mediated cardiotoxicity risk of potential drug components. Dokl

455   Biochem Biophys. 473, 128-131. http://doi.org/10.1134/S1607672917020107.

456   Riniker S., Landrum GA., 2013. Open-source platform to benchmark fingerprints for ligand-based

457   virtual screening. J Cheminform. 5, 26. http://doi.org/10.1186/1758-2946-5-26.

458   Riniker S., 2017. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data

459   To   Predict   Free-Energy   Differences.   J   Chem   Inf   Model.   57,   726-741.

460   http://doi.org/10.1021/acs.jcim.6b00778.

461   Rogers D., Hahn M., 2010. Extended-connectivity fingerprints. J Chem Inf Model. 50, 742-754.

462   http://doi.org/10.1021/ci100050t.

463   Sousa da Silva AW., Vranken WF., 2012. ACPYPE - AnteChamber PYthon Parser interfacE.

464   BMC Res Notes. 5, 367. http://doi.org/10.1186/1756-0500-5-367.

465   Subramanian G., Ramsundar B., Pande V., Denny RA., 2016. Computational Modeling of

466   β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. J Chem Inf Model. 56,

467   1936-1949. http://doi.org/10.1186/10.1021/acs.jcim.6b00290.

468   Sun CP., Yan JK., Yi J., Zhang XY., Yu ZL., Huo XK., Liang JH., Ning J., Feng L., Wang C.,

469   Zhang BJ., Tian XG., Zhang L., Ma X., 2019. The study of inhibitory effect of natural

470   flavonoids toward β-glucuronidase and interaction of flavonoids with β-glucuronidase. Int J

471   Biol Macromol. 143, 349-358. http://doi.org/10.1016/j.ijbiomac.2019.12.057.

472   Tang J., Wang Y., Luo Y., Fu J., Zhang Y., Li Y., Xiao Z., Lou Y., Qiu Y., Zhu F., 2020.

473   Computational advances of tumor marker selection and sample classification in cancer

474   proteomics.   Comput   Struct   Biotechnol   J.   18,   2012-2025.

475   http://doi.org/10.1016/j.csbj.2020.07.009.

476   R Core Team., 2013. R: A language and environment for statistical computing. R Foundation for

477   Statistical Computing. http://www.R-project.org

478   Tester DJ., Ackerman MJ., 2014. Genetics of long QT syndrome. Methodist Debakey Cardiovasc

479   J. 10, 29-33. http://doi.org/10.14797/mdcj-10-1-29.

480   Villoutreix BO., Taboureau O., 2015. Computational investigations of hERG channel blockers:

481   New insights and current predictive models. Adv Drug Deliv Rev. 86, 72-82.

482   http://doi.org/10.1016/j.addr.2015.03.003.

483   Wallace KB., 2015. Multiple Targets for Drug-Induced Mitochondrial Toxicity. Curr Med Chem.

484   22, 2488-2492. http://doi.org/10.2174/0929867322666150514095424.

485   Wang S., Riniker S., 2020. Use of molecular dynamics fingerprints (MDFPs) in SAMPL6

486   octanol-water log P blind challenge. J Comput Aided Mol Des. 34, 393-403.

487   http://doi.org/10.1007/s10822-019-00252-6.

488   Witchel HJ., 2007. The hERG potassium channel as a therapeutic target. Expert Opin Ther Targets.

489   11, 321-336. http://doi.org/10.1517/14728222.11.3.321.

490   Wouters OJ., McKee M., Luyten J., 2020. Estimated Research and Development Investment

491   Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 323, 844-853.

492   http://doi.org/10.1001/jama.2020.1166.

493   Wu ML., Wang YT., Cheng H., Sun FL., Fei J., Sun CC., Yin JP., Zhao H., Wang YS., 2020.

494   Phytoplankton community, structure and succession delineated by partial least square

495   regression in Daya Bay, South China Sea. Ecotoxicology. 29, 751-761.

496   http://doi.org/10.1007/s10646-020-02188-2.
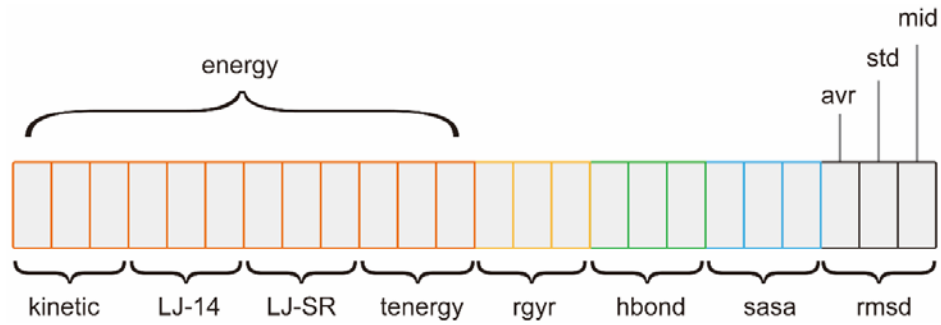
Fig.1. Schematic representation of the MDFP variant with all properties: kinetic, LJ-14, LJ-SR, tenergy, rgyr, hbond, sasa, rmsd. Each property is represented by the avr (average), std (standard deviation), and mid (median).
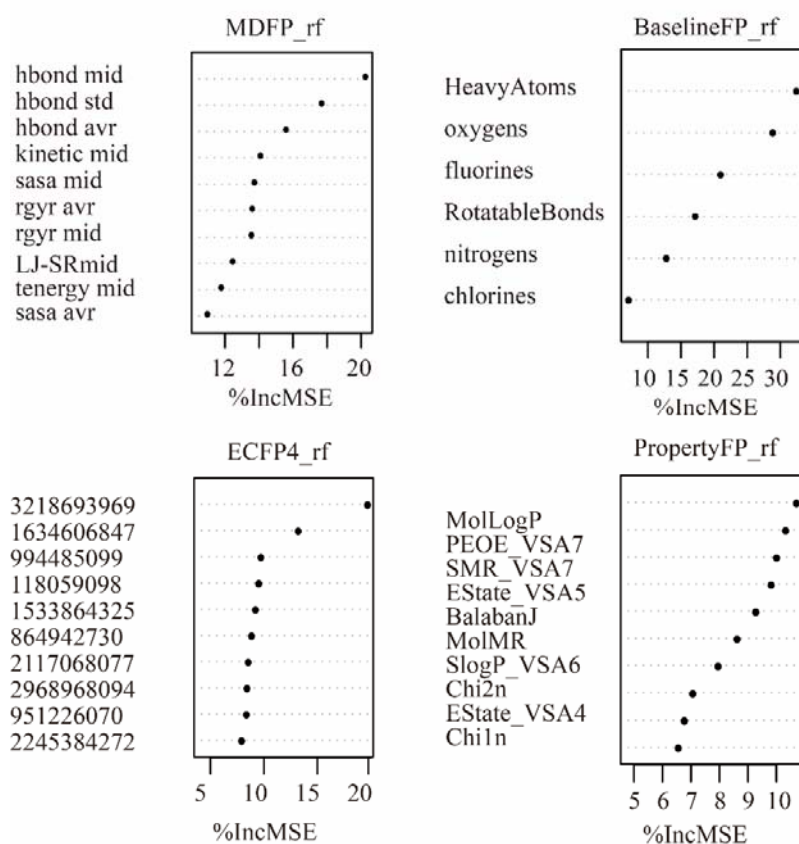
Fig. 2. The most important features selected by RF-RFE from MDFP, Baseline2D, ECFP4, and PropertyFP fingerprints.
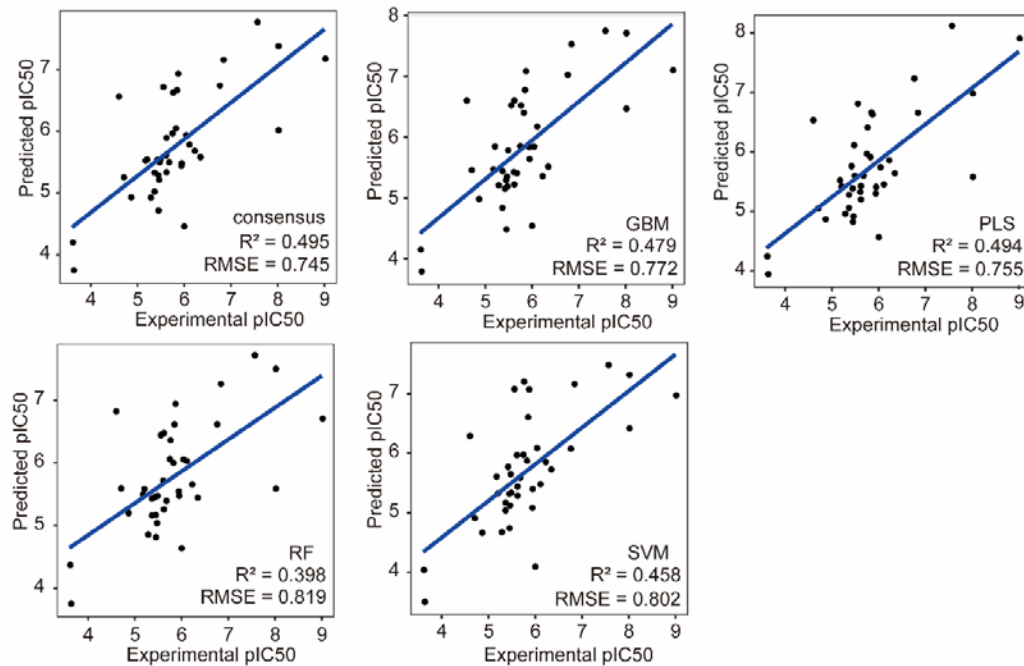
Fig. 3. pIC50: The experimental values of the 10th operation for the data set. Predictions were generated using consensus, GBM, PLS, RF, SVM trained on MDFP. The linear regression lines are shown in blue.
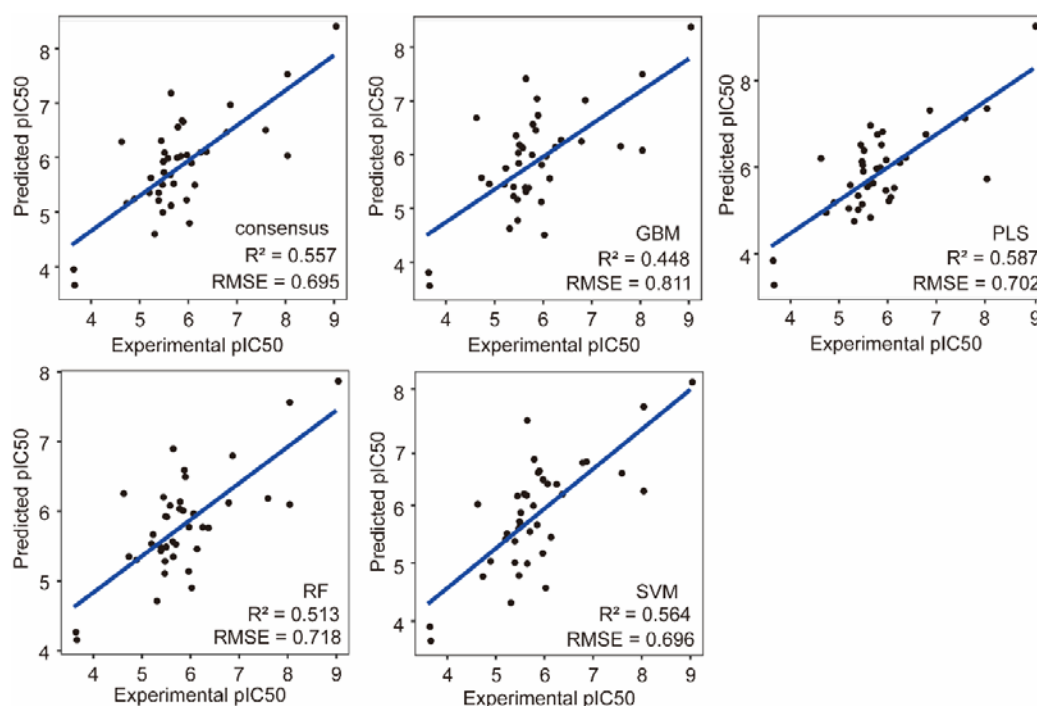
Fig. 4. pIC50: The experimental values of the 10th operation for the data set. Predictions were generated using consensus, GBM, PLS, RF, SVM trained on MDFP++. The linear regression lines are shown in blue. MDFP++ including MDFP, Baseline2D, ECFP4, and PropertyFP.

Table 1 The number of features for the different molecular fingerprints.

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

Table 2 Cross-validation performance for models trained using different ML algorithms on the molecular fingerprints (MDFP, Baseline2D, ECFP4, PropertyFP). Performance metrics are represented as average and standard deviation of 10 times 5-fold cross-validation runs of different random seeds.

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

Table 3 Cross-validation performance for models tested using different ML algorithms on the molecular fingerprints (MDFP, Baseline2D, ECFP4, PropertyFP). Performance metrics are represented as average and standard deviation of 10 times 5-fold cross-validation runs of different random seeds.

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

Table 4 Cross-validation performance for models trained using different ML algorithms on the molecular fingerprints (MDFP + Baseline2D, MDFP + ECFP4, MDFP + PropertyFP, MDFP++). Performance metrics are represented as average and standard deviation of 10 times 5-fold cross-validation runs of different random seeds.

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

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

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

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	