QSAR modeling of PTP1B inhibitor by using Genetic algorithm-Neural network methods

Isman Kurniawan, Dede Tarwidi, Jondri
School of Computing, Telkom University, Terusan Buah Batu, Bandung, 40257 Indonesia
E-mail: ismankrn@telkomuniversity.ac.id

Abstract.
Type-2 diabetes mellitus is an epidemic disease that is characterized by the chronic increase of glucose level. The insulin hormone is known to correspond to this disease, while PTP1B involved in the regulation of this hormone. Hence, PTP1B has become the primary target of drug development to treat this disease. In this study, we aim to develop QSAR model to predict PTP1B inhibitor by using a neural network method. Genetic algorithm (GA) method was used to select the set of the molecular descriptor. We improved the performance of the models by performing a hyperparameter tuning procedure. From the results of validation analysis, we found that the model 2 containing 5 descriptors as the best model. This confirms by the value of MCC (0.68) and AUC (0.89) of this model that is higher than the others. Also, the additional y-scrambled analysis confirms that this model does not correspond to coincidental correlation, indicated by a very low value of MCC of the scrambled model.

1. Introduction
Type-2 diabetes mellitus is a kind of disease that is recognized by WHO as epidemic due to the diffusion in the worldwide.[1] This disease characterized by the chronic increasing of the level of glucose. The increasing of type-2 diabetes mellitus incident and obesity lead to the intense search for new therapeutic treatment procedure.[2] Type-2 diabetes[1] and obesity[3] characteristics is related to the resistance to the hormone insulin in the liver, muscle and central nervous system.[4, 5] So, one of the criteria of the drug that will be effective to treat this disease is the ability to alleviate this resistance.

Regarding the regulation the hormone insulin, it has demonstrated that insulin receptor can be dephosphorylated by protein tyrosine phosphatase 1B (PTP1B). Hence, this protein can act as negative regulator of insulin signaling.[6, 7] A study shows that the removing PTP1B in mice leads to the increase of insulin receptor phosphorylation and enhanced the sensitivity to insulin. Further, the lack of PTP1B in mice also leads the protection from diet-induced obesity.[8, 9] Also, these mice were observed in normal and healthy condition, which point out that certain inhibitors can be used to remove the effect of PTP1B. Hence, PTP1B has become a main therapeutic target to treat type-2 diabetes and obesity, and many small molecules have been developed by targeting this enzyme.[10, 11, 12, 13, 14]

To optimize the inhibitor structure, quantitative structure-activity relationship (QSAR) have a major role by finding the relationship between the features of a compound and their biological activities, such as IC50 and LD50. QSAR performs a multivariate analysis to constructs a model that has an ability to predictive bioactivity or the class of unknown compound. The concept of
QSAR is known since the seminal work by Hansch and coworkers that looking for the correlation between the biological activity of plant growth regulators and Hammett constants and partition coefficients.[15] Recently, QSAR has been utilized to model not only biological activities,[16] but also chemical properties[17] and medical condition.[18]

In this study, we aim to build QSAR model to predict a compound as the PTP1B inhibitor. A series of the PTP1B inhibitor with activity value (IC50) were utilized to construct the model. Those inhibitors were classified into active and inactive by using certain cut-off value. We calculated a number of 2-dimension (2D) molecular descriptor of PTP1B inhibitor to provide a numerical feature of the inhibitors. The used descriptors is selected by using genetic algorithm (GA) method to produce the set of the descriptor with low cross-entropy loss value. Here, we used the GA method because this method is commonly used in feature selection procedure of QSAR analysis.[19, 20, 21] The GA method provides a set of feature that produces a result that is more accurate compared to other methods, such as forward selection or backward elimination.[22] Finally, the QSAR model was constructed to correlate the descriptor to the class of compound by using a neural network (NN) with optimized hyperparameter.

2. Material and methods

2.1. Data sets preparation

The data set of PTP1B inhibitor compounds were retrieved from ChEMBL database[23] (up to year 2017). We refined the data set by selecting human PTP1B inhibitor assays only. The data set were classified into two sub-datasets by using IC50 cut-off value as classification criteria. The first sub-dataset contain active inhibitor with IC50 value < 5 µM, while the other one contain inactive inhibitor with IC50 value > 500 µM. Here, an inhibitor will be marked as '1' if it classified as an active inhibitor, otherwise with '0'. The range of IC50 between 5 µM - 500 µM were defined as ‘gray zone’ and the inhibitors with IC50 value lies in this range were discarded. By filtering with IC50 cut-off value, we obtained a data set consisting of 1298 active inhibitors and 652 inactive inhibitors. Finally, the data set were randomly divided into training and test data with the ratio of data set number is 3:1.

2.2. Molecular descriptor calculation and selection

The total of 1613 of 2-dimensional (2D) descriptors were calculated by using Mordred program.[24] Those molecular descriptors represent structural, topological, and electrostatic properties of each inhibitor. Since a large number of descriptor corresponding to the complexity of the model, we have to select a series of the descriptor that can predict the class of inhibitor with satisfied accuracy. Thus, we performed a descriptor selection procedure in two consecutive steps. In the first step, a statistical analysis was conducted to select a descriptor with valuable information. So, the descriptor with zero variance and containing zero values more than 25 % were deleted. Also, Pearson correlation analysis was performed to reduce noise and also to remove the descriptors with similar information. The descriptor that has a weak correlation (Pearson correlation coefficient < 0.1) with pIC50 or has a strong correlation (Pearson correlation coefficient > 0.4) with other descriptor was discarded. Between two descriptors with strong correlation, the one with a stronger correlation with pIC50 will be selected. The series of descriptor obtained from this step will have valuable information.

Before continue to the second step, the set of descriptor were scaled by performing standardization procedure as formulated in the following equation

\[ x_{std}^{i} = \frac{x_{i} - \bar{x}_{i}}{\sigma_{i}} \]  

where \( x_{std}^{i} \) and \( x_{i} \) represent standardized and non-standardized value of \( i^{th} \) descriptor, while \( \bar{x}_{i} \) and \( \sigma_{i} \) represent mean and standard deviation of \( i^{th} \) descriptor. The standardization procedure is
important to make sure that all descriptor lied in similar range of mean and standard deviation. This condition is needed because neural network (NN) method will be used in the next descriptor selection procedure.

In the second step, the best combination of descriptor was selected by using Genetic Algorithm (GA) method. The GA descriptor selection was started by defining a series of a chromosome as a sequence of a bit. The total number of bit is equal to the number of a descriptor that is possible to be selected. The bit value of a descriptor will be '1' if the descriptor is selected, otherwise '0'. We iterate the GA procedure with 100 iterations and keep 20 best chromosomes in each iteration. To yield a new chromosome, we performed crossover procedure at each bit point of the chromosome. The yielded chromosome will be accepted if it contains the desired descriptor number and different with existing chromosome in the population. Then, the mutation process was performed with the probability is 0.2. To determine the best chromosomes amongst the population, we minimized the fitness function that is defined as cross-entropy loss as shown in Equation 2

$$\text{loss} = -(y \log(p) + (1 - y) \log(1 - p))$$

with $y$ and $p$ represent true target value and prediction value, respectively. The prediction value was generated by using NN method with default hyperparameter values. The GA method was used to produce 5 model of descriptor combination containing the number of descriptor of 4, 5, 6, 7, and 8, respectively.

2.3. Prediction model
We build a model to predict the target value by using NN method. This method is classified as supervised machine learning method that mimics the system of human brain. The neuronal nodes in NN system can represent the human neuron, while the interconnection between the dendrites and axons are computationally represented by synaptic weights. During the learning process, the weights will be adjusted by using a certain algorithm, e.g. forward and backward propagation, to predict the target with satisfied accuracy.

The structure of NN consists of three main layers, i.e. input, hidden and output layer. Here, the value of descriptor, defined in the input layer, will be transferred to the hidden layer and summarized as predicted value in the output layer. In the classification problem, the output layer will produce the predicted probability that will be converted to a binary value by using a function, such as sigmoid function. An activation function is also implemented in each node of a hidden layer to limit the yielded value between 0 and 1.

To improve the performance of NN, we tuned the hyperparameter of NN consists of network structure parameter, such as hidden layer, hidden unit and dropout rate, and training algorithm hyperparameter, such as learning rate and the number of epoch. The hidden layer and hidden unit was optimized by selecting the value in the range of 1 to 5 and 10 to 100, respectively. The optimal value of the hyperparameters was obtained by systematic search, while cross-entropy loss is used to measure predictive performance. As for activation function, we used ReLU function and sigmoid function for the activation function in the node of hidden layer and output layer, respectively.

2.4. Validation method
We assessed the performance of our models by evaluating the prediction accuracy of the binary classification with several statistical parameters, e.g. true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), sensitivity (SE), specificity (SP), the overall prediction accuracy ($Q$), and Matthews correlation coefficient (MCC). Here, the sensitivity measures the ability of a model to recognize the data when the positive condition is present, while the specificity measures the ability of a model to correctly exclude the data when the condition
is absent. MCC is the measure of quality of binary classification. The MCC is a coefficient that represents the correlation between the observed and predicted binary classifications. So, MCC can be used to measure overall quality of binary classification. The values of those parameters are evaluated by using Equations 3 - 6.

\[
\text{SE} = \frac{TP}{TP + FN} \tag{3}
\]

\[
\text{SP} = \frac{TN}{TN + FP} \tag{4}
\]

\[
Q = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}
\]

\[
\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{6}
\]

The models were also evaluated by visualizing receiver operating characteristic (ROC) curve. This curve illustrates the rate of success and error prediction observed in a classification model. ROC curve is plotted by taking true positive rate and false positive rate on the y-axis and x-axis, respectively. The precision of prediction can be easily recognized by analyzing the characteristic of the curve.[25] From this curve, we can also calculate another parameter, namely area under the ROC curve (AUC). The AUC measure the ability of a model to distinguish between two classification groups and thus represent the accuracy of prediction.

After we obtained the model with the best performance, a y-scrambling experiment was performed on the best model to prove that the performance of the model did not correspond to a coincidental correlation. The 10 random scrambled models were generated by shuffling the target value while keeping the descriptor. The performance of the scrambled models was evaluated by calculating MCC value and comparing the value with the non-scrambled model.

### 3. Results and discussions

#### 3.1. Dataset diversity

The chemical diversity of the training and test set of inhibitor has a vital impact on the accuracy of predictive model. The scope of model applicability will be limited if the data set involve in the training just covers a small area of chemical space. Here, we defined the chemical space by using AlogP and MW of the training and test set in a separate way. As shown in Figure 1, AlogP in the training and test set was found in the range of -7.5 to -12.5, while MW found in the range of 100 to 1000. The range of AlogP and MW in this study is also comparable with the range of similar data set in the Ref. [26]. This indicate that our data set have a large chemical space due to the broad range of AlogP and MW values. We also confirm that the active and inactive data set distribute in a proper way in training and test set.

#### 3.2. Descriptor selection

We selected the descriptor by using statistical analysis and genetic algorithm (GA) method. The number of descriptor decrease from 1613 to 650 after we removed the descriptors with zero variance and the descriptor containing zero values more than 25 %. Then, by removing the descriptor with a weak correlation to target (pIC50) and a strong correlation to other descriptors, we obtained 17 descriptors. We assumed that the selected descriptor contain enough information to be used in the model. However, the number of the selected descriptor is too large to be used for building a model, which can lead to an overfitting problem.

Hence, a wrapper type of selection method is required to select the combination of descriptors that produce satisfying results. Usually, stepwise regression method is used for this purpose.[26, 27] Unfortunately, this method is computationally expensive because this method
Figure 1. The distribution of data set diversity of the (a) training set (n = 1300 compounds) and (b) test set (n = 650 compounds).

searches the combination of all possibility in a systematic way. Therefore, we utilized the GA method, instead of stepwise regression, to select the best combination of descriptors. The GA method is less expensive because this method uses a stochastic approximation to search the combination.

We obtained 5 model from GA selection that is summarized in Table 1. Several descriptors, such as ATS2Z and GATS1se, were found to be selected in several models. This indicates that those descriptors have valuable information to be used for building the model. By comparing the loss values, we found that the addition of descriptor increases the performance of the model. However, the performance of each model can be improved by performing hyperparameter tuning. As a consequence, the model with the smaller number of descriptor can performed better then the larger one.

| Model No. | Number of Desc. | Selected Desc. | Loss  |
|-----------|----------------|----------------|-------|
| 1         | 4              | ATS2Z, MATS4i, GATS1se, IC1 | 7.57  |
| 2         | 5              | ATS2Z, AATSC1m, GATS1se, Xch-6dv, EState_VSA8 | 6.59  |
| 3         | 6              | ATS2Z, AATSC1m, MATS4i, Xch-6dv, SlogP_VSA5, EState_VSA3 | 6.32  |
| 4         | 7              | ATS2Z, ATSC2dv, MATS4i, MATS5i, GATS1se, Xch-6dv, EState_VSA8 | 5.53  |
| 5         | 8              | ATS2Z, ATSC3se, AATSC1m, MATS5i, Xch-6dv, PEOE_VSA9, EState_VSA5, piPC9 | 4.91  |

Fifteen descriptors selected in all GA models were analyzed in relation to data distribution and correlation coefficient amongst of descriptor and target. The distribution of normalized data set values of the selected descriptor was presented in Figure 2. We found that the data set
Figure 2. The distribution of normalized value of selected descriptor.

Figure 3. The correlation matrix amongst the selected descriptor and target.

was distributed in a larger range of values, while another data set was distributed in a smaller range of values. Meanwhile, we also present the correlation matrix of descriptors and target
that represent the correlation strength of them, as shown in Figure 3.

In the case of correlation coefficient analysis, we found that the correlation amongst the
descriptor in all model is less than 0.4. This confirmed that the selection by statistical analysis
had been performed successfully. We also found that only one descriptor that has a strong
correlation to the target, while the correlation of the others is relatively small.

### 3.3. Neural network model

We built a neural network model for 5 descriptor combinations obtained from genetic algorithm
(GA) selection. Hyperparameter tuning was performed to obtain optimal hyperparameter values
and improve the performance of each model. The result of the hyperparameter tuning process
was presented in Table 2. We found that the optimized number of layer number for all model
is 5. Since this number is the maximum limit of layer number, it indicates that the addition of
layer number increases the performance of NN model. Meanwhile, the value of the optimized
hidden unit was found to be different for each model.

We also found that optimized dropout rate is zero for all model, except model 2. This point
out that, in this study, the implementation of dropout make the model become simpler and
decrease the performance of the model. However, the original network of model 2 seems to
be overfitting, and thus the dropout procedure can reduce the complexity of this model. The
number of the epoch was determined by plotting the loss value vs. epoch number, as shown
in Figure 4. Interestingly, the model 5 minimize the loss value in a less epoch number than
other models. This corresponds to the number of descriptors where model 5 is comprised by the
largest number of the descriptor. We choose the optimized epoch number as 1000 because the
loss value of all model was quite stable from this point.

| Model No. | Layer Number | Hidden Unit | Dropout Rate | Learning Rate |
|-----------|--------------|-------------|--------------|---------------|
| 1         | 5            | 100         | 0.0          | 0.0017        |
| 2         | 5            | 80          | 0.1          | 0.0010        |
| 3         | 5            | 90          | 0.0          | 0.0077        |
| 4         | 5            | 100         | 0.0          | 0.0077        |
| 5         | 5            | 90          | 0.0          | 0.0010        |

### 3.4. Validation results

The validation results of model performance are presented in Table 3. In the case of training
data, we found that all of the models can accurately predict the target value, which is indicated
by the high value of SE, SP, Q, and MMC. However, model 2 and 5 seems to produce the higher
value of the results compare to the others. By using model 2, the resulted values of SE, SP, Q,
and MCC are 0.97, 0.99, 0.98 and 0.95, respectively, while those values by using model 5 are
0.98, 0.98, 0.98 and 0.96, respectively. This indicated that both models predict the target in
training data with the better results.

The test set validation is found to produce the values of the validation parameter that are
lower than training set validation. According to the results, we found that the test set prediction
by using model 2 is more accurate than other models. This is indicated by the value of the
validation parameter that is higher than other models. The values of SE, SP, Q, and MCC
obtained from the prediction by using model 2 are 0.86, 0.84, 0.86 and 0.68, respectively. Also,
Figure 4. The plot of cross-entropy loss as a function of epoch number.

Table 3. The summary of validation results of the prediction on training and test set.

| Model | TP  | FP  | TN  | FN  | SE  | SP  | Q   | MCC |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|
|       |     |     |     |     |     |     |     |     |
| Training set |     |     |     |     |     |     |     |     |
| 1     | 861 | 17  | 390 | 32  | 0.96| 0.96| 0.96| 0.91|
| 2     | 872 | 6   | 399 | 23  | 0.97| 0.99| 0.98| 0.95|
| 3     | 869 | 9   | 392 | 30  | 0.97| 0.98| 0.97| 0.93|
| 4     | 813 | 65  | 389 | 33  | 0.96| 0.86| 0.92| 0.83|
| 5     | 868 | 10  | 407 | 15  | 0.98| 0.98| 0.98| 0.96|
| Test set |     |     |     |     |     |     |     |     |
| 1     | 354 | 66  | 157 | 73  | 0.83| 0.70| 0.79| 0.53|
| 2     | 388 | 32  | 169 | 61  | 0.86| 0.84| 0.86| 0.68|
| 3     | 383 | 37  | 165 | 65  | 0.85| 0.82| 0.84| 0.65|
| 4     | 357 | 63  | 177 | 53  | 0.87| 0.74| 0.82| 0.61|
| 5     | 375 | 45  | 172 | 58  | 0.87| 0.79| 0.84| 0.65|

To confirm the performance of the models, we also present the receiver operating characteristic (ROC) curve and the value of the area under the curve (AUC), as shown in Figure 5. We found that the characteristic of the ROC curve is quite similar for all model. However, the value of AUC of model 2 is found as around 0.89 that is higher than the others. This confirms that model 2 is the best model in predicting the activity class of inhibitors.

Then, a y-scrambling analysis was performed on model 2 to confirm that the results do not correspond to a coincidental correlation. The MCC values of model 2 that is used to predict 10 scrambled data set presented in Figure 6. For the comparison reason, we also present the MCC value of model 2 for unscrambled data. We found that the MCC value for 10 scrambled data is lower than 0.1, while the MCC value for unscrambled data is around 0.67. This confirms that the accuracy of model 2 does not related to a coincidental correlation.
Figure 5. Receiver operating characteristic (ROC) plot of 5 neural network models.

Figure 6. Y-scrambling result of model 2.

4. Conclusion
We have selected the set of molecular descriptor appeared in QSAR model by using genetic algorithm (GA) methods. The QSAR models were developed by using a neural network method to predict the class of inhibitor. The procedure of hyperparameter tuning was performed to improve the performance of the models. According to validation results, we confirmed that the model 2, that contains 5 descriptors, produce the most accurate results. This is indicated by the higher value of MCC (0.68) and AUC (0.89) compare to the others. Also, we confirm that the model does not correspond to coincidental correlation, which indicates by an extremely low value of MCC of the scrambled model.

References
[1] DeFronzo R A, Bonadonna R C and Ferrannini E 1992 Diabetes Care 15 318–368 ISSN 0149-5992
[2] Zimmet P, Alberti K G and Shaw J 2001 Nature 414 782–787 ISSN 0028-0836
[3] Johnson T O, Ermoliev J and Jirousek M R 2002 Nature Reviews. Drug Discovery 1 696–709 ISSN 1474-1776
[4] Saltiel A R and Kahn C R 2001 Nature 414 799–806 ISSN 0028-0836
[5] Obici S, Feng Z, Karkanias G, Baskin D G and Rossetti L 2002 Nature Neuroscience 5 566–572 ISSN 1097-6256
[6] Byon J C, Kusari A B and Kusari J 1998 Molecular and Cellular Biochemistry 182 101–108 ISSN 0300-8177
[7] Goldstein B J, Bittner-Kowalczyk A, White M F and Harbeck M 2000 The Journal of Biological Chemistry 275 4283–4289 ISSN 0021-9258
[8] Elchebly M, Payette P, Michalisyn E, Cromlish W, Collins S, Loy A L, Normandin D, Cheng A, Himms-Hagen J, Chan C C, Ramachandran C, Gresser M J, Tremblay M L and Kennedy B P 1999 Science (New York, N.Y.) 283 1544–1548 ISSN 0036-8075
[9] Klaman L D, Boss O, Peroni O D, Kim J K, Martino J L, Zabolotny J M, Moghal N, Lubkin M, Kim Y B, Sharpe A H, Stricker-Krongrad A, Shulman G I, Neel B G and Kahn B B 2000 Molecular and Cellular Biology 20 5479–5489 ISSN 0270-7306
[10] Yue E W, Wayland B, Douty B, Crawley M L, McLaughlin E, Takvorian A, Wasserman Z, Bower M J, Wei M, Li Y, Ala P J, Gonneville L, Wynn R, Burn T C, Liu P C C and Combs A P 2006 Bioorganic & Medicinal Chemistry 14 5833–5849 ISSN 0968-0896
[11] Zhao H, Liu G, Xin Z, Serby M D, Pei Z, Szczepankiewicz B G, Hajduk P J, Abad-Zapatero C, Hutchins C W, Lubben T H, Ballaron S J, Haasch D L, Kaszubska W, Rondinone C M, Trevillyan J M and Jirousek M R 2004 Bioorganic & Medicinal Chemistry Letters 14 5543–5546 ISSN 0960-894X
[12] Cho S Y, Baek J Y, Han S S, Kang S K, Ha J D, Ahn J H, Lee J D, Kim K R, Cheon H G, Rhee S D, Yang S D, Yon G H, Pak C S and Choi J K 2006 Bioorganic & Medicinal Chemistry Letters 16 499–502 ISSN 0960-894X
[13] Sparks R B, Polam P, Zhu W, Crawley M L, Takvorian A, McLaughlin E, Wei M, Ala P J, Gonneville L,
Taylor N, Li Y, Wynn R, Burn T C, Liu P C C and Combs A P 2007 *Bioorganic & Medicinal Chemistry Letters* **17** 736–740 ISSN 0960-894X

[14] Dixit M, Tripathi B K, Tamrakar A K, Srivastava A K, Kumar B and Goel A 2007 *Bioorganic & Medicinal Chemistry* **15** 727–734 ISSN 0968-0896 URL http://www.sciencedirect.com/science/article/pii/S0968089606008856

[15] Hansch C, Muir R M, Fujita T, Maloney P P, Geiger F and Streich M 1963 *Journal of the American Chemical Society* **85** 2817–2824 ISSN 0002-7863 URL https://doi.org/10.1021/ja00901a033

[16] Thippakorn C, Sukhirachavalit T, Nantasenamat C, Tantimongcolwat T, Isarakura-Na-Ayudhya C, Naenna T, Prachayasittikul V, Thippakorn C, Sukhirachavalit T, Nantasenamat C, Tantimongcolwat T, Isarakura-Na-Ayudhya C, Naenna T and Prachayasittikul V 2009 *Molecules* **14** 1869–1888 URL http://www.mdpi.com/1420-3049/14/5/1869

[17] Nantasenamat C, Isarakura-Na-Ayudhya C, Naenna T and Prachayasittikul V 2007 *Biosensors and Bioelectronics* **22** 3309–3317 ISSN 0956-5663 URL http://www.sciencedirect.com/science/article/pii/S0956566307000358

[18] Worachartcheewan A, Nantasenamat C, Isarakura-Na-Ayudhya C, Piletscha P and Prachayasittikul V 2010 *Diabetes Research and Clinical Practice* **90** e15–e18 ISSN 0168-8227 URL http://www.sciencedirect.com/science/article/pii/S016822710003062

[19] Hasegawa K, Miyashita Y and Funatsu K 1997 *Journal of Chemical Information and Computer Sciences* **37** 306–310 ISSN 0095-2338 URL https://doi.org/10.1021/ci960047x

[20] Sagrado S and Cronin M T D 2008 *Analytica Chimica Acta* **609** 169–174 ISSN 0003-2670 URL http://www.sciencedirect.com/science/article/pii/S0003267008000846

[21] Gao H, Lajiness M S and Drie J V 2002 *Journal of Molecular Graphics and Modelling* **20** 259–268 ISSN 1093-3263 URL http://www.sciencedirect.com/science/article/pii/S109332630100122X

[22] Xu L and Zhang W J 2001 *Analytica Chimica Acta* **446** 475–481 ISSN 0003-2670 URL http://www.sciencedirect.com/science/article/pii/S0003267001012715

[23] Gaulton A, Hersey A, Nowotka M, Bento A P, Chambers J, Mendez D, Mutowpo P, Atkinson F, Bellis L J, Chirin-Uhalte E, Davies M, Dedman N, Karlsson A, Magarios M P, Overington J P, Papadatos G, Smitt I and Leach A R 2017 *Nucleic Acids Research* **45** D945–D954 ISSN 0305-1048 URL https://academic.oup.com/nar/article/45/D1/D945/2605707

[24] Moriwaki H, Tian Y S, Kawashita N and Takagi T 2018 *Journal of Cheminformatics* **10** 4 ISSN 1758-2946 URL https://doi.org/10.1186/s13321-018-0258-y

[25] Fawcett T 2006 *Pattern Recognition Letters* **27** 861–874 ISSN 0167-8655 URL http://www.sciencedirect.com/science/article/pii/S016786550500303X

[26] Chandra S, Pandey J, Tamrakar A K and Siddiqi M I 2017 *Journal of Molecular Graphics and Modelling* **71** 242–256 ISSN 1093-3263 URL http://www.sciencedirect.com/science/article/pii/S1093326316303308

[27] Cai J, Li C, Liu Z, Du J, Ye J, Gu Q and Xu J 2017 *Journal of Computer-Aided Molecular Design* **31** 393–402 ISSN 0920-654X, 1573-4951 URL https://link.springer.com/article/10.1007/s10822-017-0009-6