IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 33, No. 3, June, 2019, pp. 227-233
Journal homepage: http://iieta.org/journals/ria

A Machine Learning Prediction Model for the Affinity Between Glucose and Binder

Rajesh Kondabala[1*], Vijay Kumar[2], Amjad Ali[3]

[1] School of Chemistry and Biochemistry, Thapar Institute of Engineering and Technology, Punjab 147004, India
[2] Department of Computer Science and Engineering, National Institute of Technology, Hamirpur 177005, India
[3] School of Chemistry and Biochemistry, Thapar Institute of Engineering and Technology, Punjab 147004, India

Corresponding Author Email: rajesh.kondabala@thapar.edu

https://doi.org/10.18280/ria.330309

Received: 6 April 2019
Accepted: 11 June 2019

Keywords:
machine learning, regression, prediction models, glucose binder, binding affinity

ABSTRACT

The glucose is an important source of fuel for the body. The binding affinity is an essential indicator of the interaction of a glucose molecule with its binder. This paper proposes a novel machine learning model for predicting the binding affinity of a small glucose molecule with the binder. Seven regression algorithms were compared on a dataset is generated based on Molecular Mechanics-Generalized Born and Surface Area (MM-GBSA). Through the comparison, Random Forest and Decision Tree were selected for our model, in light of their robustness and accuracy. The established model predicts binding affinity from the interaction properties of compounds and glucose, which are obtained through GLIDE program from Schrödinger software suite 2018-4. Finally, the prediction accuracy of our model was confirmed through k-fold cross-validation. Our research provides an efficient and low-cost method for screening of molecules during the development of glucose binders.

1. INTRODUCTION

Supra molecular chemistry continues to grow with an accelerated pace in biomolecular recognition [1]. It was made possible in developing a synthetic receptor for glucose recognition by mimicking the properties of naturally available glucose binding proteins [2]. Glucose is an essential carbohydrate and standard part of the daily meal. The brain of an organism uses half of the glucose amount present in our body. It involves in many diseases such as diabetes [3], hyponatremia [4], seizures [4], and cancer [5, 6]. It plays a significant role in the biological process such as the development of multicellular organisms, cell infection pathogens, distribution, and reactivity of proteins within host cells [7, 8]. Homeostasis of glucose concentration in tissues is essential for the normal functioning of our body. The imbalance of glucose concentration leads to severe health conditions. Monitoring glucose levels/concentration helps patients to keep control over their glucose levels through diet. Hence, there is a need to develop new techniques that will help in improving human health quality [9, 10]. Novel glucose binding molecules are required for developing these types of techniques. Binding affinity plays a crucial role that decides the fate of compound interaction towards glucose molecules. It is a challenging issue to recognize the glucose molecule in a solvent. Because of its abundant hydroxyl groups on glucose molecules and their interaction with the surrounding solvent medium makes a difficult task for designing a novel synthetic receptor [11, 12].

In this paper, a novel computational model is developed for predicting the binding affinity of a small glucose binder molecule. The proposed prediction model reduces not only computational time but also reduces the preliminary cost for the screening of molecules during the development of a novel glucose binder. The proposed dataset is generated by utilizing the basic concept of Molecular Mechanics-Generalized Born and Surface Area (MM-GBSA) [13]. The proposed computational model predicts binding affinity from the interaction properties of compounds and glucose. Those molecular interaction properties between glucose and small molecules are obtained through GLIDE [14] program from Schrödinger software suite 2018-4. K-fold cross-validation [15] is performed to measure the robustness of the best predictive model. The two well-known prediction techniques namely Random Forest [16], and Decision Tree [17] are utilized for the construction of the proposed framework which is more accurate and robust.

The remaining structure of this paper is organized as follows. Section 2 describes the preliminary concepts of the glucose molecule. The proposed computational prediction model is presented in Section 3. Experimental results and discussions are presented in Section 4. The concluding remarks are drawn in Section 5.

2. BACKGROUND

This section describes the natural glucose binding proteins followed by machine learning techniques.

2.1 Natural glucose binding molecule

Lectins are natural glucose binding proteins. They are widely present in plants, animals, and bacteria [18-20]. However, they have excellent selectivity and a very low binding affinity towards sugars [21, 22]. Hydrogen bonds play a vital role in specificity and affinity during receptor sugar interactions. These interactions are generally stable and exhibit optimal geometries. These are categorized into three main types, namely cooperative hydrogen bonds, bidentate

hydrogen bonds, and hydrogen networks hydrogen bonds. The residues are present in sugars binding sites that have polar side chains with at least two functional groups having three hydrogen bond types. Sugars show other interactions like van der Waals forces with carboxylate side chains and aromatic residues in π−π stacking with sugar rings [23, 24]. MM-GBSA [13] is used to estimate relative binding affinity from the PRIME [25] program. Binding energies are computed by MM-GBSA [13] method. The small binding molecules are expected to provide a good binding affinity.

2.2 Motivation

In the past few years, supramolecular chemistry boomed up in the field of diagnosis and therapy [1]. Davis et al. [2, 26-29] proposed different biomimetic receptors for carbohydrates known as synthetic lectins. Jiang et al. [30-33] proposed simple boronic acids and their aggregates for saccharide sensing. However, the receptors have their disadvantages. Such as the performance of the existing glucose binding molecules is not optimal. To overcome these problems, a novel computational prediction model is developed that predicts the binding affinity of small molecules towards glucose. It helps researchers to design and develop a selective glucose binding molecule. According to the best of the author's knowledge, there is no prediction model available in the literature for glucose binding small molecules.

3. PROPOSED APPROACH

3.1 Proposed prediction model

The proposed computational prediction model consists of three main phases, namely data generation, model development, and prediction. In the data generation and preparation phase, the annoying and biased features are removed. The importance of features is computed as these are dramatical effects on the performance of the proposed model. In the second phase, the prediction model is developed. The predictive model is based on molecular interaction and binding energy properties. The last phase is the prediction of the binding affinity of glucose binder by the proposed prediction model. Figure 1 shows the proposed prediction model. The detail description of these phases is mentioned in the preceding subsections.

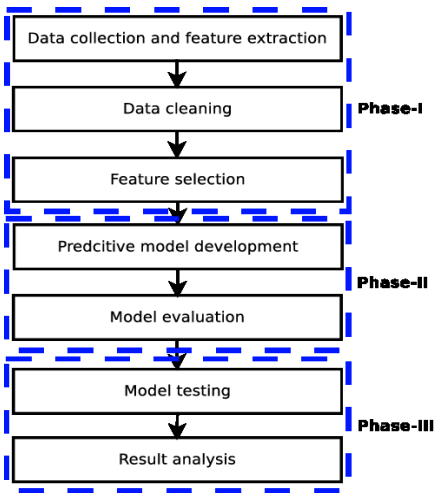


Figure 1. Proposed computational prediction model

3.2 Data generation and data pre-processing

The computational model is developed for the prediction of binding affinity from interaction data. The data cleaning process removes duplicate tuples and missing values. The feature selection process is carried out through the Pearsons Correlation method after the cleaning dataset. And Selected features were used for constructing a prediction model. Seventy percentage data from the entire dataset is used for training the model. The model is tested with the remaining thirty percentage testing data. Figure 2 shows the methodology that is used to generate molecular interaction and binding affinity.

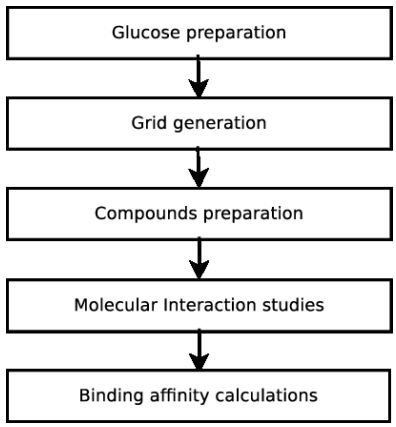


Figure 2. Data collection procedure through the Molecular and Binding affinity calculation approach

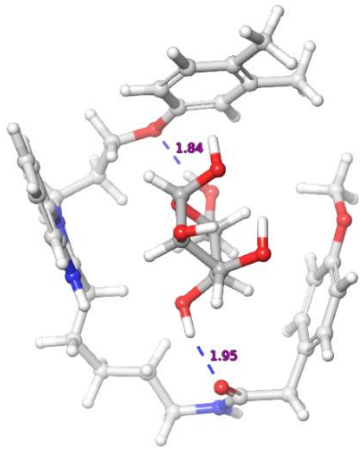


Figure 3. Hydrogen bonding (blue lines) between electronegative atoms of glucose and compound with ZINC ID ZINC40222112

3.2.1 Glucose structure preparation and grid generation

Glucose is a non-amino acid biomolecule. 3D structure of glucose in glucopyranose form was retrieved from PubChem [27] compound database. Macro Model [34] program is used for conformational search and optimizing glucose molecule structure using OPLS3e [35] forcefield. A grid box is generated on prepared glucose molecule with dimensions 40, 40, and 40 as x, y, and z coordinates, respectively.

3.2.2 Small molecule preparation

Forty thousand natural compounds are extracted from the ZINC database [36] and prepared using the LigPrep program [37]. It filters compounds that are based on drug-likeness rules

by using QikProp and Epik module [38]. It is used for generating possible protonation states between pH 7 +/- 2.

3.2.3 Molecular Interaction and Binding affinity calculations

The interaction studies and scoring of small compounds with glucose molecules are performed by the Standard Precision (SP) method of GLIDE [14] program. From 40,000 molecules, only 13,000 are screened out. After that, the screened-out compounds are submitted to MM-GBSA [13] method that presents in the PRIME [25] program. OPLS3e [35] is used to compute the binding affinity of screened out compounds. Figure 3 shows the interactions between glucose and compound with ZINC ID ZINC40222112.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Dataset and features used

Forty thousand small molecules of natural compounds are taken from the ZINC [36] compound database. After that, thirteen thousand small particles are screened out from forty thousand using the SP method. The molecular interactions towards glucose are generated using GLIDE [14]. Based on small molecules interact with glucose, the binding affinity energies are produced using Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) method [13]. Sixty-nine features are selected using PRIME [25]. MM-GBSA [13] calculates the binding free energies for molecules by combining molecular mechanics calculations and continuum (implicit) solvation models. The implicit solvent models are often used to estimate free energies of solute-solvent interactions. Table 1 shows the sample dataset of small glucose binding molecules.

Table 1. Sample dataset

Affinity (KCal/mol)	Coulomb	….	Hbond
-15.90	-13.57	….	-1.38
-15.27	-10.47	….	-0.51
-17.11	-7.29	….	-0.53
-16.82	-14.10	….	-0.68
-13.44	-10.55	….	-0.85

4.2 Performance measures

To demonstrate the performance of the proposed model, it is evaluated on four well-known performance measures such as root mean square error (RMSE) [40], correlation [41], coefficient of determination (R^2) [42], and accuracy [43]. For the measurement of the robustness of the proposed predictive model, the K-fold cross-validation [15] method is used. To perform this validation, the original dataset is randomly portioned into k equal size subsets. A single subset is kept as the validation data for testing purposes, and remaining subsets are used for training. This validation process is repeated for k times(folds). By this random sub-sampling, all observations are used for both training and validation.

4.3 Algorithms used for model development and their parameter setting

The eleven machine learning algorithms are used for the prediction of binding affinity of small compounds towards glucose molecules. The methods are available in Python [44] open-source licensed under GNU GPL. Scikit-learn [45] package from python is used for building the prediction models. These are Random Forest (RF) [16], Decision Tree (DT) [17], Support Vector Machine (SVM) [17, 46], Linear Method(LM) [17], Neural Network(NN) [17], Elastic Net(EN) [17, 47], K-Nearest Neighbors (KNN) [17, 48], Lasso [49], Random Sample Consensus(RANSAC) [17], Ridge [17], Stochastic Gradient Descent(SGD) [17]. Table 2 shows the parameters used for model development.

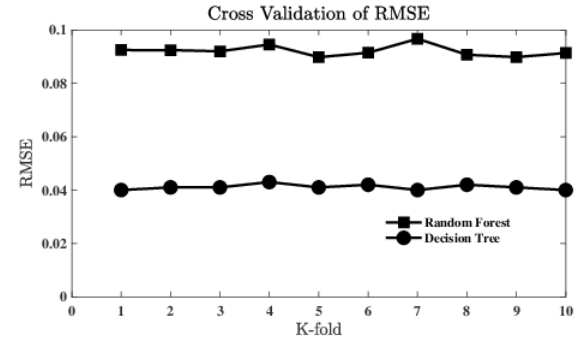
Table 2. Parameter setting of the involved computational techniques

Model	Method	Package	Parameters
DT	tree	Sklearn.tree	MaxDepth=30
RF	rf	randomForest	MaxDepth=30 Randstate=7n
SVM	svr	e1071	nu=10
LM	lm	glm	None
NN	mlp	MLP	hlayers=10 MaxNWts=10000
Enet	ElasticNet	Sklearn.linear	Max iter=1000 Random state=7
KNN	KNRegressor	Sklearn.neighbors	n-neighbors=10 n-jobs=1000
Lasso	Lasso	Sklearn.linear	Max iter=1000
RANSAC	RANSAC Regressor	Sklearn.linear	Max trails=100
Ridge	Ridge	Sklearn.linear	Normalize-True Max iter=1000
SGD	SGD Regressor	Sklearn.linear	None

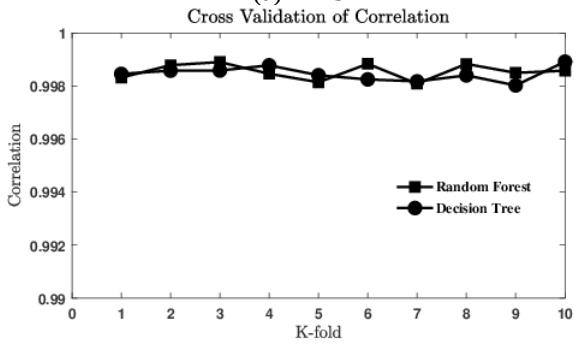
4.4 Performance analysis

The eleven machine learning methods are used for the prediction of binding affinity. From the original dataset, seventy percentage data are used for training and the remaining thirty percentage is used for testing for all the methods mentioned above. Table 4. shows the performance comparison of these methods in the prediction of binding affinity based on RMSE [40], correlation [41], R^2 [42], and accuracy [43]. The results reveal that the Random Forest [16] method performs better than the other machine learning methods. Here, 10-fold cross-validation [15] is used to measure the robustness of the predictive model. RMSE, correlation, R^2, and accuracy for ten folds in the prediction of binding affinity energy values. Cross-validation provides that Random Forest and Decision Tree models are better results than the other prediction models. Figure 4 shows 10-fold cross-validation on testing dataset in the prediction of Binding affinity using Random Forest and Decision Tree models.

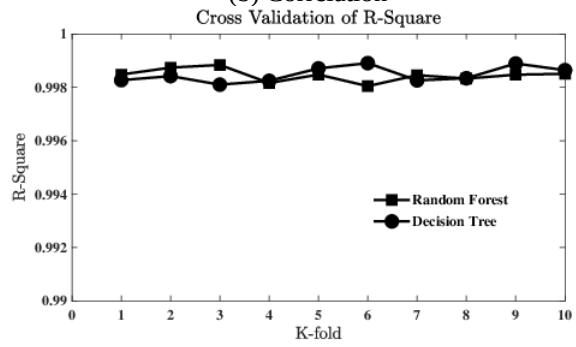
The value of RMSE obtained from Random Forest, and Decision Tree is 0.04 and 0.09, respectively. The correlation predicted by the Random Forest model is 0.99, and the Decision Tree shows 0.99, respectively. The calculated R^2 values for the Random Forest method and Decision Tree are 0.99 and 0.99, respectively. The computed accuracy of the Random Forest method with ±0.5 acceptance error is 99.03 % and the Decision Tree method is 97.14 % with ±0.5 acceptance error respectively. Figure 5 shows a scatter plot of regression graphs for Random Forest and Decision Tree models.



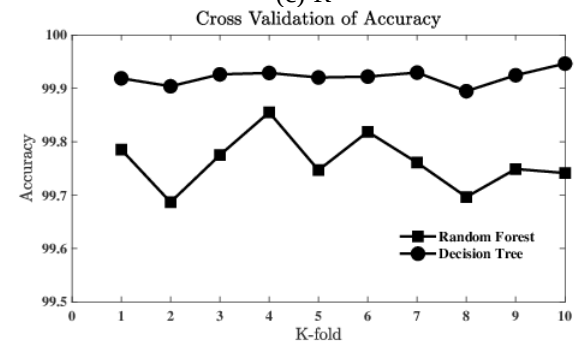
(a) RMSE



(b) Correlation

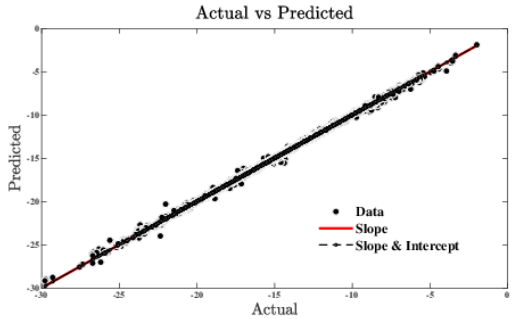


(c) R^2

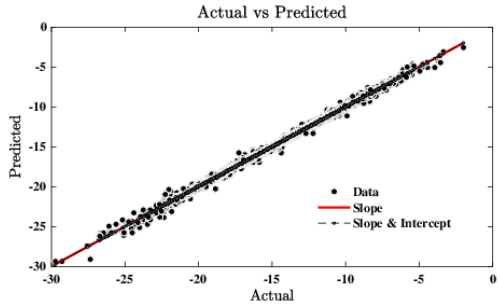


(d) Accuracy

Figure 4. 10-fold cross-validation on the testing dataset in the prediction of Binding affinity using random forest and decision tree



(a) Random forest prediction model



(b) Decision Tree prediction model

Figure 5. Scatter plot between actual vs. predictive values of binding affinity on the testing dataset

Table 3. Comparison of computational methods in the prediction of MM-GBSA (binding affinity) with an acceptance error ±0.5

Model	RMSE	Corr	R^2	Acc in %
DT	0.09	0.99	0.99	97.14
RF	0.04	0.99	0.99	99.03
SVM	0.35	0.98	0.96	83.96
LM	0.00	0.99	0.99	100
NN	0.44	0.92	0.84	71.03
ENet	1.01	0.97	0.95	31.87
K-NN	1.08	0.93	0.87	31.60
Lasso	1.22	0.97	0.95	33.24
Ransac	331.29	0.02	0.00	20.92
Ridge	0.68	0.98	0.96	47.33
SGD	1153.38	-0.00	2.78e	0.74

Note: 1. Corr = Correlation between actual and predicted values.
2. Acc = Accuracy in % with ±0.5 acceptance error.

4.5 Validation

The proposed prediction framework is validated over Inter Bio (IB) Screen [50] dataset. The values of correlation obtained and R^2 are 0.99 and 0.98, respectively. The RMSE and accuracy derived from the proposed framework are 0.08 and 95 %, respectively. Table 4 shows the actual and predicted binding energies of compounds from IB Screen databases by the Random Forest prediction model.

Table 4. Actual and predicted binding energies of validation compounds

Actual value of binding affinity	The predicted value of binding affinity
-18.19	-18.18
-13.61	-13.66
-17.00	-17.02
-16.46	-16.57
-16.32	-16.33
-5.43	-4.77
-21.08	-21.02
-9.56	-9.33
-11.51	-11.65

However, there is a significant increase in the field of supra-chemistry in the recognition of simple glucose molecules [1]. There are failures during experimental evaluation by wastage of funds and chemicals. Glucose recognition in the aqueous condition is a challenging task because of abundant hydroxyl groups around glucose, that can be easily embedded in the solvent [11, 12]. We screened out thirteen thousand small

molecules from forty thousand from the ZINC database by using the SP method of GLIDE module. The screened molecules are subjected to energy calculations using the MM-GBSA method present in the PRIME tool of Schrödinger suite.

The prediction models showed promising results in diagnosing and drug discovery. Therefore, there is a high need for prediction approaches for the screening and selection of a suitable glucose binder. These prediction approaches are more advanced than the traditional computational programs that consume a lot of time and computational power.

5. CONCLUSIONS

In this paper, a novel computational prediction model is used to predict the binding affinity of a glucose molecule with its binder. Eleven different learning methods are utilized in developing this predictive model. It is found that Random Forest and Decision Tree provide better predictions than the other methods. The robustness of the best prediction model is cross-validated through the K-fold method. However, the Linear Regression model provides the lowest RMSE, high R^2, and high correlation even with ±0.5 acceptance error. Linear Regression model is neglected due to over-fitting, which is not ideal for the prediction model. The proposed prediction model reduces the computational cost of running Molecular Mechanics (MM) and Molecular Dynamics (MD). The computational experiment results provided a prediction framework with RMSE from Random Forest, and Decision Tree is 0.04 and 0.09, respectively. The correlation Random Forest model is 0.99, and the Decision Tree shows 0.99, respectively. The R^2 values for the Random Forest method and Decision Tree are 0.99 and 0.99, respectively. And the accuracy of the Random Forest method with ±0.5 acceptance error is 99.03 % and the Decision Tree method is 97.14 % with ±0.5 acceptance error respectively.

ACKNOWLEDGMENT

This research is funded by the Early Career Research Award Program through the Science and Engineering Research Board (SERB) grant number ECR/2016/001231.

We thank Schrödinger company for providing PRIME module and OPLSe forcefield temporary license

REFERENCES

[1] Sun, X., James, T.D. (2015). Glucose sensing in supramolecular chemistry. Chemical Reviews, 115(15): 8001-8037. http://dx.doi.org/10.1021/cr500562m

[2] Tromans, R.A., Carter, T.S., Chabanne, L., Crump, M.P., Li, H., Matlock, J.V., Orchard, M.G., Davis, A.P. (2019). A biomimetic receptor for glucose. Nature Chemistry, 11(1): 52-56. http://dx.doi.org/10.1038/s41557-018-0155-z

[3] Association, A.D. (2015). 2. classification and diagnosis of diabetes. Diabetes Care, 38(1): 8-16. http://dx.doi.org/10.2337/dc17-S005

[4] De Vivo, D.C., Trifiletti, R.R., Jacobson, R.I., Ronen, G.M., Behmand, R.A., Harik, S.I. (1991). Defective glucose transport across the blood-brain barrier as a cause of persistent hypoglycorrhachia, seizures, and developmental delay. New England Journal of Medicine, 325(10): 703-709. http://dx.doi.org/10.1056/NEJM199109053251006

[5] Cheng, C., Ru, P., Geng, F., Liu, J., Yoo, J.Y., Wu, X., Cheng, X., Euthine, V., Hu, P., Guo, J.Y., Lefai, E. (2015). Glucose-mediated n-glycosylation of scap is essential for srebp-1 activation and tumor growth. Cancer Cell, 28(5): 569-581. http://dx.doi.org/10.1016/j.ccell.2015.09.021

[6] Hamanaka, R.B., Chandel, N.S. (2012). Targeting glucose metabolism for cancer therapy. Journal of Experimental Medicine, 209(2): 211-215. http://dx.doi.org/10.1083/JCB1964OIA3

[7] Nosadini, R., Tonolo, G. (2004). Relationship between blood glucose control, pathogenesis and progression of diabetic nephropathy. Journal of the American Society of Nephrology, 15(1): 1-5. http://dx.doi.org/10.1097/01.ASN.0000093372.84929.BA

[8] Gallacher, S., Thomson, G., Fraser, W., Fisher, B., Gemmell, C., MacCuish, A. (1995). Neutrophil bactericidal function in diabetes mellitus: Evidence for association with blood glucose control. Diabetic Medicine, 12(10): 916-920. http://dx.doi.org/10.1111/j.1464-5491.1995.tb00396.x

[9] Crane, B.C., Barwell, N.P., Gopal, P., Gopichand, M., Higgs, T., James, T.D., Jones, C.M., Mackenzie, A., Mulavisala, K.P., Paterson, W. (2015). The development of a continuous intravascular glucose monitoring sensor. Journal of Diabetes Science and Technology, 9(4): 751-761. http://dx.doi.org/10.1177/1932296815587937

[10] Mauras, N., Fox, L., Englert, K., Beck, R.W. (2013). Continuous glucose monitoring in type 1 diabetes. Endocrine, 43(1): 41-50. http://dx.doi.org/10.1007/s12020-012-9765-1

[11] Ke, C., Destecroix, H., Crump, M.P., Davis, A.P. (2012). A simple and accessible synthetic lectin for glucose recognition and sensing. Nature Chemistry, 4(9): 718-723. http://dx.doi.org/10.1038/nchem.1409

[12] Sears, P., Wong, C.H. (1999). Carbohydrate mimetics: A new strategy for tackling the problem of carbohydrate-mediated biological recognition. Angewandte Chemie International Edition, 38(16): 2300-2324. http://dx.doi.org/10.1002/(SICI)1521-3773(19990816)38:16%3C2300::AID-ANIE2300%3E3.0.CO;2-6

[13] Hou, T., Wang, J., Li, Y., Wang, W. (2010). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. Journal of Chemical Information and Modeling, 51(1): 69-82. http://dx.doi.org/10.1021/ci100275a

[14] Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., Shaw, D.E. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. Journal of Medicinal Chemistry, 47(7): 1739-1749. http://dx.doi.org/10.1021/jm0306430

[15] Arlot, S., Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4: 40-79. http://dx.doi.org/10.1214/09-SS054

[16] Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5-32.

http://dx.doi.org/10.1023/A:1010933404324
[17] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016). Data mining: Practical machine learning tools and techniques. ACM SIGMOD Record, 31(1): 76-77. http://dx.doi.org/10.1145/507338.507355
[18] Lis, H., Sharon, N. (1998). Lectins: Carbohydrate-specific proteins that mediate cellular recognition. Chemical Reviews, 98(2): 637-674. http://dx.doi.org/10.1021/cr940413g
[19] Goldstein, I.J., Hayes, C.E. (1978). The lectins: Carbohydrate-binding proteins of plants and animals. Advances in Carbohydrate Chemistry and Biochemistry, 35: 127-340. http://dx.doi.org/10.1016/S0065-2318(08)60220-6
[20] Sharon, N., Lis, H. (1972). Lectins: Cell-agglutinating and sugar-specific proteins. Science, 177(4053): 949-959. http://dx.doi.org/10.1126/science.177.4053.949
[21] Weis, W.I., Drickamer, K. (1996). Structural basis of lectin-carbohydrate recognition. Annual Review of Biochemistry, 65(1): 441-473. http://dx.doi.org/10.1146/annurev.bi.65.070196.002301
[22] Toone, E.J. (1994). Structure and energetics of protein-carbohydrate complexes. Current Opinion in Structural Biology, 4(5): 719-728. http://dx.doi.org/10.1016/S0959-440X(94)90170-8
[23] Jin, S., Cheng, Y., Reid, S., Li, M., Wang, B. (2010). Carbohydrate recognition by boronolectins, small molecules, and lectins. Medicinal Research Reviews, 30(2): 171-257. http://dx.doi.org/10.1002/med.20155
[24] Wang, B., Boons, G.J. (2011). Carbohydrate Recognition: Biological Problems, Methods, and Applications. John Wiley & Sons. http://dx.doi.org/10.1002/9781118017586
[25] Prime, S. (2018). Schrödinger Release 2018-4: Prime. Schrödinger, New York, NY.
[26] Mooibroek, T.J., Crump, M.P., Davis, A.P. (2016). Synthesis and evaluation of a desymmetrised synthetic lectin: An approach to carbohydrate receptors with improved versatility. Organic & Biomolecular Chemistry, 14(6): 1930-1933. http://dx.doi.org/10.1039/C6OB00023A
[27] Mooibroek, T.J., Casas-Solvas, J.M., Harniman, R.L., Renney, C.M., Carter, T.S., Crump, M.P., Davis, A.P. (2016). A threading receptor for polysaccharides. Nature Chemistry, 8(1): 69-74. http://dx.doi.org/10.1038/nchem.2395
[28] Ferrand, Y., Crump, M.P., Davis, A.P. (2007). A synthetic lectin analog for biomimetic disaccharide recognition. Science, 318(5850): 619-622. http://dx.doi.org/10.1126/science.1148735
[29] Ríos, P., Mooibroek, T.J., Carter, T.S., Williams, C., Wilson, M.R., Crump, M.P., Davis, A.P. (2017). Enantioselective carbohydrate recognition by synthetic lectins in water. Chemical Science, 8(5): 4056-4061. http://dx.doi.org/10.1039/C6SC05399H
[30] Wu, X., Li, Z., Chen, X.X., Fossey, J.S., James, T.D., Jiang, Y.B. (2013). Selective sensing of saccharides using simple boronic acids and their aggregates. Chemical Society Reviews, 42(20): 8032-8048. http://dx.doi.org/10.1039/c3cs60148j
[31] Xu, S.Y., Wang, H.C., Flower, S.E., Fossey, J.S., Jiang, Y.B., James, T.D. (2014). Suzuki homo-coupling reaction based fluorescent sensors for monosaccharides. RSC Advances, 4(66): 35238-35241. http://dx.doi.org/10.1039/C4RA07331B
[32] Huang, Y.J., Ouyang, W.J., Wu, X., Li, Z., Fossey, J.S., James, T.D., Jiang, Y.B. (2013). Glucose sensing via aggregation and the use of "Knock-Out" binding to improve selectivity. Journal of the American Chemical Society, 135(5): 1700-1703. http://dx.doi.org/10.1021/ja311442x
[33] Guo, L.E., Hong, Y., Zhang, S.Y., Zhang, M., Yan, X.S., Cao, J.L., Li, Z., James, T.D., Jiang, Y.B. (2018). Proline-based boronic acid receptors for chiral recognition of glucose. The Journal of Organic Chemistry, 83(24): 15128-15135. http://dx.doi.org/10.1021/acs.joc.8b02425
[34] MacroModel, S. (2018). Schrödinger Release 2018-4: MacroModel. Schrödinger, New York, NY.
[35] Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J.Y., Wang, L., Lupyan, D., Dahlgren, M.K., Knight, J.L., Kaus, J.W., Cerutti, D.S., Krilov, G., Jorgensen, W.L., Abel, R., Friesner, R.A. (2016). Opls3: A force field providing broad coverage of drug-like small molecules and proteins. Journal of Chemical Theory and Computation, 12(1): 281-296. http://dx.doi.org/10.1021/acs.jctc.5b00864
[36] Sterling, T., Irwin, J.J. (2015). Zinc 15–ligand discovery for everyone. Journal of Chemical Information and Modeling, 55(11): 2324-2337. http://dx.doi.org/10.1021/acs.jcim.5b00559
[37] LigPrep, S. (2018). 2: Ligprep, schrödinger, llc, New York, NY, 2018. New York, NY.
[38] Epik, S. (2018). Schrödinger Release 2018-4: Epik. Schrödinger, New York, NY.
[39] Joshi, G., Davis, A.P. (2012). New h-bonding patterns in biphenyl-based synthetic lectins; pyrrolediamine bridges enhance glucose-selectivity. Organic & Biomolecular Chemistry, 10(30): 5760–5763. http://dx.doi.org/10.1039/c2ob25900a
[40] Wang, W., Lu, Y. (2018). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In Materials Science and Engineering Conference Series, 324: 012049. http://dx.doi.org/10.1088/1757-899X/324/1/012049
[41] Schober, P., Boer, C., Schwarte, L.A. (2018). Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia, 126(5): 1763-1768. http://dx.doi.org/10.1213/ANE.0000000000002864
[42] Barrett, J.P. (1974). The coefficient of determination-some limitations. The American Statistician, 28(1): 19-20. http://dx.doi.org/10.2307/2683523
[43] Cai, T.T., Guo, Z. (2018). Accuracy assessment for high-dimensional linear regression. The Annals of Statistics, 46(4): 1807-1836. http://dx.doi.org/10.1214/17-AOS1604
[44] Zelle, J. (2016). Python Programming: An Introduction to Computer Science. Franklin, Beedle & Associates Inc.
[45] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12(Oct): 2825-2830. https://arxiv.org/abs/1201.0490
[46] Cherkassky, V., Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1): 113-126.

http://dx.doi.org/10.1016/S0893-6080(03)00169-2
[47] Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. Journal of the American Statistical Association, 106(496): 1383-1393. http://dx.doi.org/10.1198/jasa.2011.tm09241
[48] Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3): 175-185. http://dx.doi.org/10.2307/2685209
[49] Hans, C. (2009). Bayesian lasso regression. Biometrika, 96(4): 835-845. http://dx.doi.org/10.1093/biomet/asp047
[50] InterBioScreen (2018). Natural compound database. https://www.ibscreen.com/natural-compounds, accessed on 16 October, 2018.