Chinese Chemical Letters 34 (2023) 107964

Contents lists available at ScienceDirect

Chinese Chemical Letters

journal homepage: www.elsevier.com/locate/cclet

Cocrystal virtual screening based on the XGBoost machine learning model

Dezhi Yang[a,1], Li Wang[a,1], Penghui Yuan[a], Qi An[a], Bin Su[b], Mingchao Yu[a], Ting Chen[a], Kun Hu[a], Li Zhang[a,*], Yang Lu[a,*], Guanhua Du[c,*]

[a] Beijing City Key Laboratory of Polymorphic Drugs, Center of Pharmaceutical Polymorphs, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, China
[b] Shandong Soteria Pharmaceutical Co., Ltd., Laiwu 271100, China
[c] Beijing City Key Laboratory of Drug Target and Screening Research, National Center for Pharmaceutical Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, China

ARTICLE INFO

Article history:
Received 19 July 2022
Revised 18 August 2022
Accepted 28 October 2022
Available online 31 October 2022

Keywords:
Cocrystal
Machine learning
XGBoost
Molecular descriptor
Praziquantel
Nefiracetam

ABSTRACT

Co-crystal formation can improve the physicochemical properties of a compound, thus enhancing its druggability. Therefore, artificial intelligence-based co-crystal virtual screening in the early stage of drug development has attracted extensive attention from researchers. However, the complexity of developing and applying algorithms hinders it wide application. This study presents a data-driven co-crystal prediction method based on the XGBoost machine learning model of the scikit-learn package. The simplified molecular input line entry specification (SMILES) information of two compounds is simply inputted to determine whether a co-crystal can be formed. The data set includs the co-crystal records presented in the Cambridge Structural Database (CSD) and the records of no co-crystal formation from extant literature and experiments. RDKit molecular descriptors are adopted as the features of a compound in the data set. The developed model shows excellent performance in the proposed co-crystal training and validation sets with high accuracy, sensitivity, and F1 score. The prediction success rate of the model exceeds 90%. The model therefore provides a simple and feasible scheme for designing and screening co-crystal drugs efficiently and accurately.

Co-crystals are composed of an active pharmaceutical ingredient (API) and a suitable cocrystal former (CCF) [1], which are bound through non-covalent interactions, including hydrogen bonding, π-π stacking, van der Waals forces, and other weak interactions. Co-crystals can improve the physicochemical properties and biological activity of drugs without any chemical modification [2–4]. Therefore, co-crystals have elicited extensive interest in the field of medicine, and co-crystal screening has become an integral part of drug development. The use of suspension, slow solvent volatilization, and liquid addition milling methods for cocrystal experimental screening is often time consuming; therefore, fast, high-throughput, and convenient virtual screening techniques for co-crystals of APIs are necessary. A method based on the ΔpK_a value of API and CCF was proposed for the prediction of co-crystal formation [5,6]. The Hansen solubility parameter method was developed to predict the formation of drug co-crystals by predicting the degree of miscibility of API and CCF [7–10], and it only requires knowledge on the chemical structure of the molecule. In addition, the energy-based molecular electrostatic potential surface (MEPS) prediction method has elicited widespread interest [11–14]. MEPS analysis is based on density functional theory (DFT), and this method can be used for qualitative and quantitative analyses of weak interactions between different molecules [15,16]. However, because it does not consider the factors of conformation and steric hindrance, this method sometimes produces unsatisfactory prediction results. Therefore, MEPS combined with conformational analysis was developed to improve the prediction success rate for co-crystal formation [17].

Co-crystal screening based on theoretical computations is also time consuming and requires extensive professional theoretical knowledge. Therefore, the application of artificial intelligence (AI) in co-crystal screening has become a new research direction because it can shorten the experimental cycle, improve research efficiency, and reduce experimental costs [18]. Several researchers have used classification and regression algorithms in predicting

* Corresponding authors.
E-mail addresses: zhangl@imm.ac.cn (L. Zhang), luy@imm.ac.cn (Y. Lu), dugh@imm.ac.cn (G. Du).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cclet.2022.107964
1001-8417/

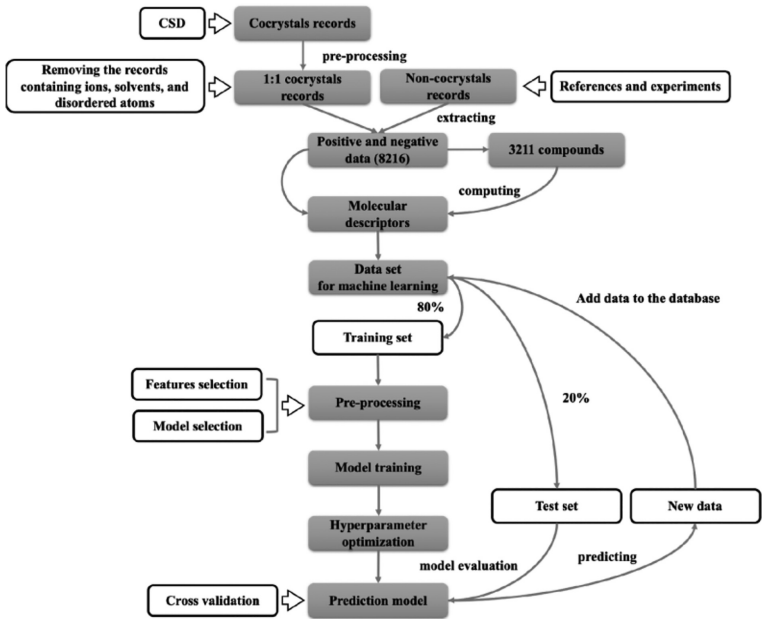

Fig. 1. Flowchart for building a XGBoost machine learning model for co-crystal screening.

co-crystal formation to accelerate co-crystal screening and obtain high-quality co-crystals [19–24].

In this work, we developed a data-driven co-crystal prediction method based on the eXtreme Gradient Boosting (XGBoost) model [25]. Fig. 1 showed the flowchart of the study. The co-crystal data in the Cambridge Structural Database (CSD) of the Cambridge Crystallographic Data Centre (CCDC) and the data recorded as no co-crystal formation in experimental screening were used as the data set for model training. The structures of API and CCF were represented by SMILES strings. RDKit molecular descriptors computed from the input SMILES strings were used as the features of the corresponding compound, which were computed by the ChemDes website [26]. The advantages of this co-crystal prediction method are simple implementation, intuitive results, a high success rate, and ease of use.

Python scripts were written using the Python interface provided by CCDC to obtain the co-crystal records from CSD, including the constituent molecules and their stoichiometry. The records containing ions, solvents, and disordered atoms were removed, and in accordance with the general stoichiometric ratio used in the co-crystal experimental screening, the records with API:CCF ratios of 1:1 were extracted as the positive data of co-crystals; 4061 co-crystal records were obtained. The records in which non-co-crystals were formed were obtained from references [27] and experimental data in our laboratory. These records (4155 non-co-crystal records) were summarized as the negative data set of co-crystals. These 8216 records were employed as a data set for the co-crystal prediction study.

A total of more than 3211 compounds were included in the data set. The compounds were stored using canonical SMILES strings, and their structural information was transformed to .sdf files for the computation of molecular descriptors on the ChemDes website. At present, many software or programs can be used to calculate molecular descriptors. The ChemDes website allows users to compute 3679 molecular descriptors from several open-source packages, including Chemopy Descriptors (1135) [28], CDK Descriptors (275) [29], RDKit Descriptors (196) [30], Pybel Descriptors (24) [31], BlueDesc Descriptors (174) [32], and PaDEL Descriptors (1875) [33]. This website also provides the computation of 59 types of molecular fingerprints. Preliminary tests have shown that simply increasing the number of molecular descriptors does not significantly improve the prediction success rate of the model, but it increases the computational workload. This study selected RDKit molecular descriptors as the features of compounds in data set.

Table 1
The scores and cross-validation scores of different classifiers in the preliminary evaluation.

Classifier	Score	Cross-validation score
XGB	0.9599	0.9630
RF	0.9580	0.9615
GB	0.9307	0.9407
KNN	0.8893	0.8659
DT	0.9179	0.9210
AB	0.8887	0.8908
LDA	0.8571	0.8663
QDA	0.7238	0.7325

The research used a classification method to carry out machine learning. The records of the co-crystals were labeled as 1, and 0 was used for non-co-crystals. Among the 196 RDKit descriptors, 80% of the calculated features with zeroes were removed, and the remaining 143 descriptors were used as compound feature vectors. Then, 80% of the records were used for training, and 20% was adopted for testing. This study performed a preliminary evaluation of nine classification models of scikit-learn (version 1.1.1) [34], namely, gradient boosting (GB), adaptive boosting (AB), extreme gradient boosting (XGB), random forest (RF), k-nearest neighbors (KNN), decision tree (DT), linear discriminant analysis (LDA), rectangular discriminant analysis (QDA), and multi-layer perception (MLP) classifiers. The nine classifiers with the default parameters were used for the preliminary evaluation of machine learning, and the results are shown in Table 1.

Except for QDA, the classifiers showed good classification performance for co-crystals and non-co-crystals in this data set, with scores and cross-validation scores of more than 80%. XGB, RF, GB and DT all had scores above 92%, but XGB performed the best among them. Therefore, XGB classifiers were selected as the machine learning model for further hyperparameter optimization. This decision was also based on the reported excellent performance of this model and the fact that it is the go-to algorithm of compe-

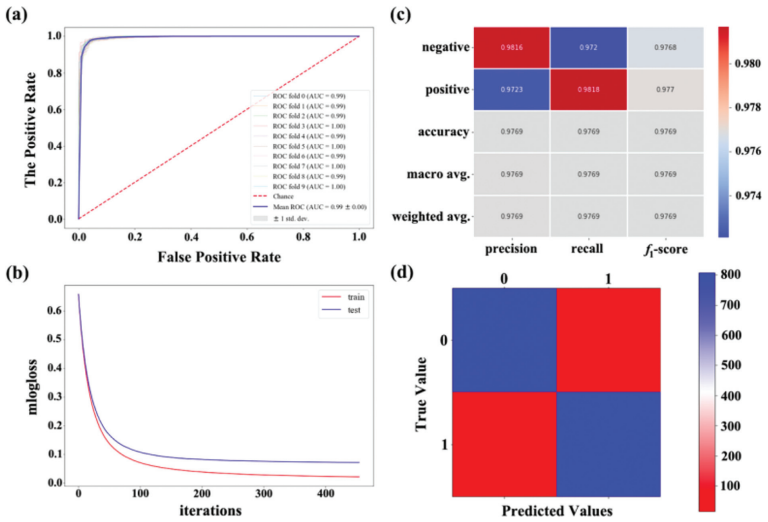

Fig. 2. The results of model evaluation and validation. (a) AUC-ROC curve; (b) logloss vs iteration curve; (c) classification report plot; (d) confusion matrix plot.

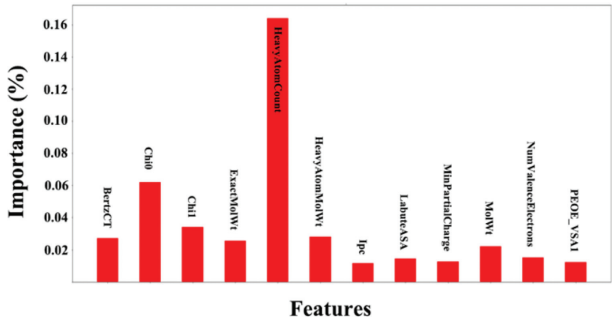

Fig. 3. Feature importance histogram of the top 12 descriptors.

tition winners on the Kaggle competitive data science platform. After adjusting the hyperparameters of the XBG model, including learning_rate, max_depth, n_estimators, min_child_weight, and gamma, the final model score reached 97.69%.

Various methods were used to evaluate the model and draw the receiver operating characteristic (ROC) curve. The area under the curve (AUC) reached 99%; notably, the closer AUC is to 1, the better the performance is in distinguishing the two classes (Fig. 2a). The logarithmic loss function (logloss) versus iteration times curve was plotted and is shown in Fig. 2b. The results showed that after 400 iterations, logloss was basically stable and maintained at a low level. The model accuracy, recall rate of prediction, and F1 score were also evaluated, and the scores were all above 97% (Fig. 2c). The prediction results of the test set were used to draw a confusion matrix to evaluate the model (Fig. 2d). The results revealed that among the 1644 (20%) predicted data items, 799 were true positives, 807 were true negatives, 15 were false negatives, and 23 were false positives. This finding indicates that the proportion of model error classification was small, and the model had excellent classification accuracy.

Furthermore, the feature importance of 143 molecular descriptors was analyzed. Fig. 3 shows a histogram of the importance of the top 12 features. The descriptor named HeavyAtomCount, which represents the number of heavy atoms of a molecule, was the most important. The descriptors Chi0 and Chi1 followed HeavyAtomCount; they belong to connectivity descriptors and are from Eqs. 1, 9 and 10 in Ref. [35]. The meanings of all descriptors are listed in Table 2 in the order of importance. The descriptor HeavyAtomCount appeared to be important mainly due to the higher weights of the heavy atoms (C, N and O) in CCF of co-crystal records in the dataset.

As shown in Fig. 4, a branch of the decision tree within the trained XGBoost model was plotted using Graphviz [37], which clearly showed the decision-making process of the model. Each decision-making process started with a feature and its value. When the predicted value was less than the value, the decision-making process developed downward along the left branch; when the predicted value was greater than or equal to the value, the process developed downward along the right branch. The process ended when the final judgment was given. Notably, the size of decision trees in the trained XGBoost can be tuned by the max_depth parameter.

Shapley additive translation (SHAP) is a model explanation package developed by Python [38], and it is a game theory approach to explaining each machine learning model's individual output. In this work, SHAP was used to visualize the prediction results of two cases belonging to two classes by using force and waterfall plots.

In the force plot, the average value of model output and training data values, namely, the base value, was calculated and found to be 0.4956 in this model. Starting from the base value, the features that pushed the prediction higher and lower were presented in red and blue, respectively. The longer the arrow was, the greater the impact of the features was on the output value. For example, as indicated in Fig. 5a, the feature BertzCT played the most important role in the red force that pushed the predicted value up, whereas the features ExactMolWt, Chi0, and HeavyAtomMolWt played the most important role in the blue force that pushed the predicted value down. The final force balance value was the predicted value of 0.00. The process was similar in another case, and the predicted value in Fig. 5b is 1.00.

The waterfall plot was also used to visualize the interpretation of individual predictions. The waterfall plot started with the expected value of the model output at the bottom, and each row showed the positive (red) or negative (blue) contribution of each feature, that is, how the features pushed the values from the model's expected output on the data set to the model's predicted output. The same cases used in the force plot were selected, and similar results were obtained (Figs. 5c and d).

Then, co-crystal prediction was conducted for the APIs of praziquantel (PZQ) and nefiracetam (NFC) with 8 and 3 CCFs, respectively, by using the XGB model. We are conducting co-crystal studies on the two drugs, and they can be used to validate the predic-

Table 2
The descriptions of the top 12 descriptors in order of feature importance.

No.	Descriptor name	Description	Dimention	Extended class
1	HeavyAtomCount	Number of heavy atoms of a molecule	1D	Constitutional descriptors
2	Chi0	From Eqs. 1, 9 and 10 of Ref. [27]	2D	Connectivity descriptors
3	Chi1	From Eqs. 1, 9 and 10 of Ref. [27]	2D	Connectivity descriptors
4	BertzCT	A topological index meant to quantify "complexity" of molecules [36]	2D	Topological descriptors
5	HeavyAtomMolWt	The average molecular weight of the molecule ignoring hydrogens	1D	Constitutional descriptors
6	ExactMolWt	The molecule's exact molecular weight	2D	Molecular property descriptors
7	MolWt	The average molecular weight of the molecule	2D	Molecular property descriptors
8	NumValenceElectrons	The number of valence electrons the molecule has	1D	Constitutional descriptors
9	LabuteASA	Labute's Approximate Surface Area (ASA from MOE)	2D	MOE-type descriptors
10	MinPartialCharge	Returns molecular charge descriptors	2D	Topological descriptors
11	PEOE_VSA1	MOE Charge VSA Descriptor 1 (-inf < x < -0.30)	2D	MOE-type descriptors
12	Ipc	The information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule	2D	Topological descriptors

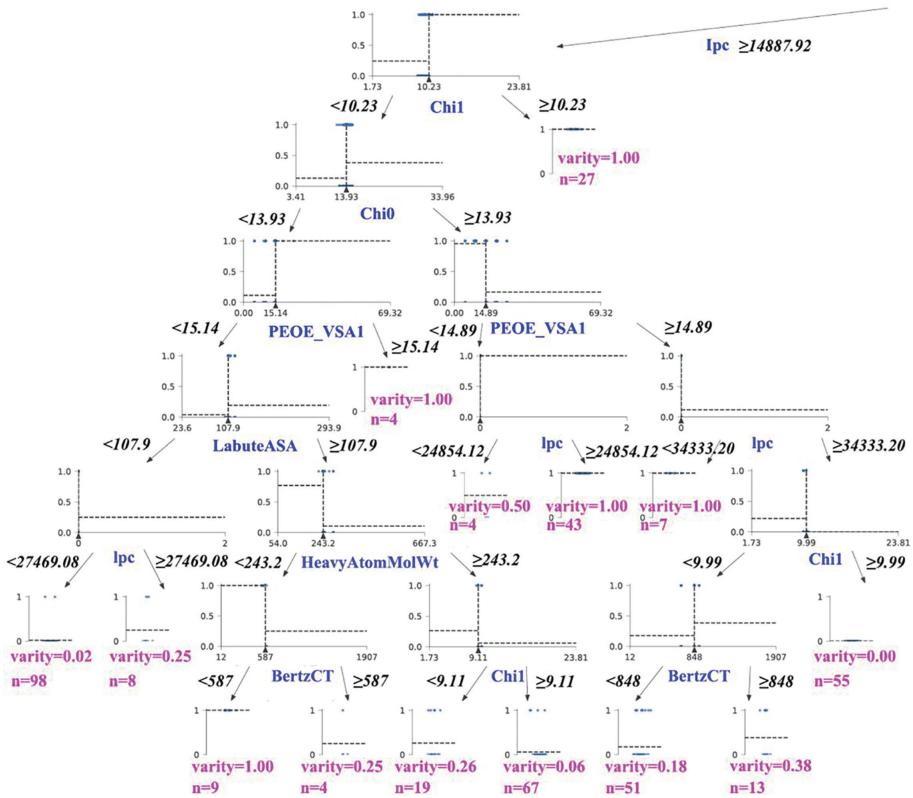

Fig. 4. A branch of the decision tree within the trained XGBoost model.

tion accuracy of the model. PZQ is the most effective and widely used drug of choice for treating schistosomiasis [39–41]. NFC is a nootropic compound typically administered as a cognitive enhancer [42,43]. Both drugs belong to biopharmaceutical classification system (BCS) II, that is, their solubilities are low. Therefore, the study of co-crystals is expected to improve their solubilities. The structural formulas of APIs and CCFs are shown in Fig. S1 (Supporting informatin). The model predicted that PZQ could form co-crystals with all 8 CCFs, which was consistent with the experimental results [17,44,45]. The prediction results indicated that NFC could form co-crystals with tartaric acid (TA) but not with maleic acid (MA) and fumaric acid (FA). However, the experimental results showed that NFC could form co-crystals with MA.

When analyzing the reasons for the failure of the co-crystal prediction of NFC–MA, we noticed that MA and FA were cis-trans isomers, and their SMILES files were the same. As a result, the developed model could not distinguish their 3D structures. In addition, in the co-crystal data set containing the SMILES file of butenedioic acid, the data of FA predominated. In the experimental screening, NFC and FA could not form co-crystals, which was the main reason for the prediction failure. This particular result suggests that the usage of 3D molecular descriptors to distinguish isomers could be one of the research directions in future model upgrade.

Fig. 6 shows the asymmetric unit and molecular packing in a unit cell of the co-crystals of NFC. Crystallographic information is listed in Table S1 (Supporting information). Additional characterization results of the co-crystals are also applied, including PXRD patterns (Figs. S1 and S2 in Supporting information), DSC curves (Fig. S3 in Supporting information) and IR spectrum (Fig. S4 in Supporting information).

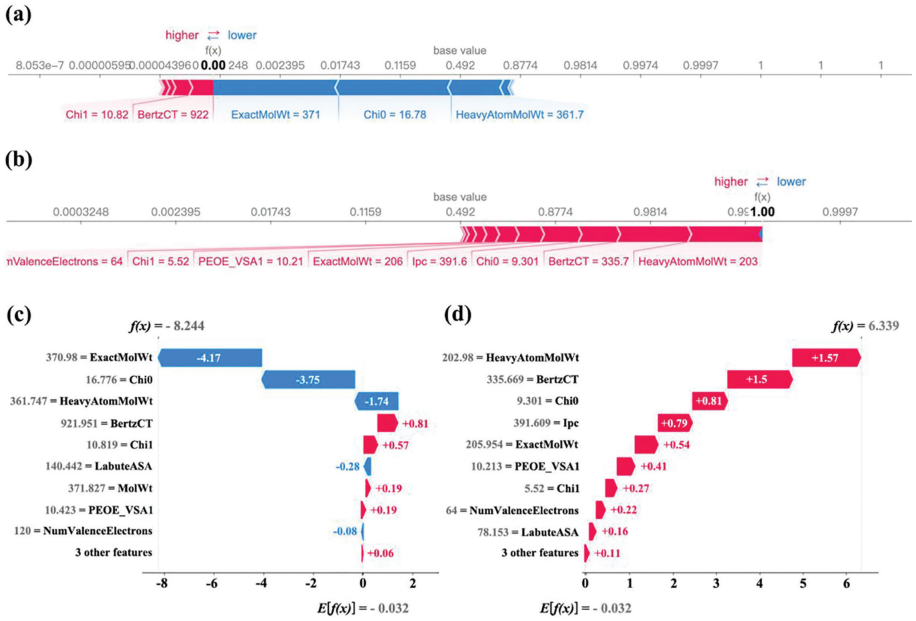

Fig. 5. (a, b) The force plot and (c, d) waterfall plot of the prediction results of two cases.

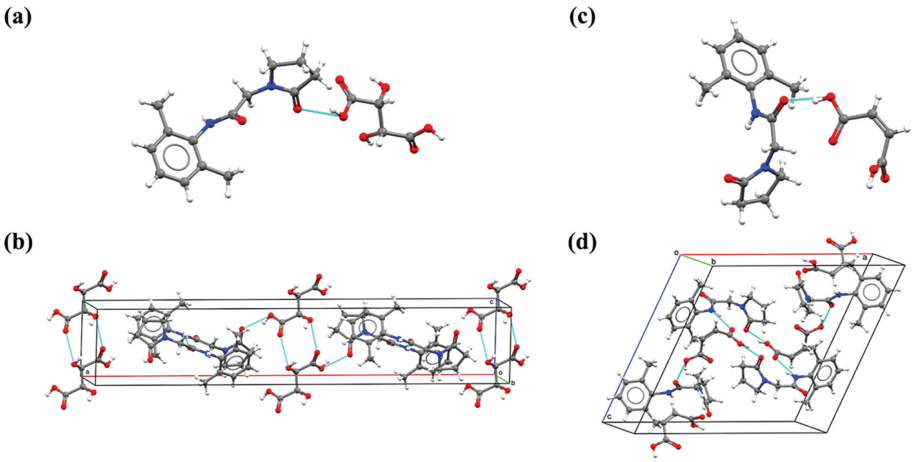

Fig. 6. The asymmetric unit (up) and molecular packing in a unit cell (down) of cocrystals of NFC: (a) NFC-TA; (b) NFC-TA; (c) NFC-MA; (d) NFC-MA.

The developed XGBoost machine learning model demonstrated excellent performance, and its prediction success rate exceeded 90%. All of the predicted co-crystals in this study were synthesized experimentally; among them, the co-crystals of PZQ have been published in previous research, but the co-crystals of NFC are reported here first. The model can be used as a powerful tool for co-crystal prediction and design in the field of drug research.

Declaration of competing interest

The authors report no declarations of interest.

Acknowledgment

The authors acknowledge the National Natural Science Foundation of China (No. 22278443), CAMS Innovation Fund for Medical Sciences (No. 2022-I2M-1-015), the Key R&D Program of Shan Dong Province (No. 2019JZZY020909), and the Xinjiang Uygur Autonomous Region Innovation Environment Construction Special Fund and Technology Innovation Base Construction Key Laboratory Open Project (No. 2022D04016) for the financial support.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cclet.2022.107964.

References

[1] S.H. Mithu, S.A. Ross, A.P. Hurt, et al., J. Drug Deliv. Sci. Technol. 63 (2021) 102508–102518.
[2] V.N. Saladi, B.R. Kammari, P.R. Mandad, et al., Cryst. Growth Des. 22 (2022) 1130–1142.
[3] D. Yang, H. Wang, Q. Liu, et al., Chin. Chem. Lett. 33 (2022) 3207–3211.
[4] F. Martin, M. Pop, I. Kacso, et al., Mol. Pharm. 17 (2020) 919–932.
[5] L.S. Reddy, B.R. Bhogala, A. Nangia, CrystEngComm 7 (2005) 206–209.
[6] A.J. Cruz-Cabeza, M. Lusi, H.P. Wheatcroft, et al., Faraday Discuss. 235 (2022) 446–466.
[7] C.M. Hansen, K. Skaarup, J. Paint Technol. 39 (1967) 511–520.
[8] A. Salem, S. Nagy, S. Pal, et al., Int. J. Pharm. 558 (2019) 319–327.
[9] P.S. Panzade, G.R. Shendarkar, Curr. Drug Deliv. 14 (2017) 1097–1105.
[10] M.A. Mohammad, A. Alhalaweh, S.P. Velaga, Int. J. Pharm. 407 (2011) 63–71.
[11] D. Musumeci, C.A. Hunter, R. Prohens, et al., Chem. Sci. 2 (2011) 883–890.
[12] Y.S. Mary, Y.S. Mary, Polycycl. Aromat. Compd. (2021) 1–12.
[13] R. Barbas, M. Font-Bardia, A. Paradkar, et al., Cryst. Growth Des. 18 (2018) 7618–7627.
[14] B.K. Mehta, S.S. Singh, S. Chaturvedi, et al., Cryst. Growth Des. 18 (2018) 1581–1592.

[15] H. Wu, Y. Sun, L. Sun, et al., Chin. Chem. Lett. 32 (2021) 3007–3010.
[16] M.U. Faroque, A. Mehmood, Noureen, et al., J. Mol. Struct. 1214 (2020) 128183.
[17] D. Yang, J. Cao, T. Heng, et al., Cryst. Growth Des. 21 (2021) 2292–2300.
[18] T. Heng, D. Yang, R. Wang, et al., ACS Omega 6 (2021) 15543–15550.
[19] M. Przybyek, T. Jeliński, J. Sabuszewska, et al., Cryst. Growth Des. 19 (2019) 3876–3887.
[20] B. Chabalenge, S. Korde, A.L. Kelly, et al., Cryst. Growth Des. 20 (2020) 4540–4549.
[21] J.J. Devogelaer, H. Meekes, P. Tinnemans, et al., Angew. Chem. Int. Ed. 59 (2020) 21711–21718.
[22] A. Vriza, A.B. Canaj, R. Vismara, et al., Chem. Sci. 12 (2021) 1702–1719.
[23] M. Przybyłek, P. Cysewski, Cryst. Growth Des. 18 (2018) 3524–3534.
[24] Y. Jiang, Z. Yang, J. Guo, et al., Nat. Commun. 12 (2021) 1–14.
[25] T. Chen, C. Guestrin, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[26] J. Dong, D.S. Cao, H.Y. Miao, et al., J. Cheminform. 7 (2015) 1–10.
[27] M.E. Mswahili, M.J. Lee, G.L. Martin, et al., Appl. Sci. 11 (2021) 1323.
[28] D.S. Cao, Q.S. Xu, Q.N. Hu, et al., Bioinformatics 29 (2013) 1092–1094.
[29] C. Steinbeck, Y. Han, S. Kuhn, et al., J. Chem. Inf. Comput. Sci. 43 (2003) 493–500.
[30] G. Landrum, URL http://www.rdkit.org/, https://github.com/rdkit/rdkit.149 (2016) 150.
[31] N.M. O'Boyle, C. Morley, G.R. Hutchison, Chem. Cent. J. 2 (2008) 1–7.
[32] University of Tübingen: BlueDesc. http://www.ra.cs.uni-tuebingen.de/software/bluedesc/.
[33] C.W. Yap, J. Comput. Chem. 32 (2011) 1466–1474.
[34] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., J. Mach. Learn. Res. 12 (2011) 2825–2830.
[35] K.B. Lipkowitz, D.B. Boyd, Reviews in Computational Chemistry, 2, Wiley-VCH, Hoboken, 1996, pp. 367–422. Volume.
[36] S.H. Bertz, J. Am. Chem. Soc. 103 (1981) 3599–3601.
[37] J. Ellson, E. Gansner, L. Koutsofios, et al., in: Proceedings of the International Symposium on Graph Drawing, Springer, 2001, pp. 483–484.
[38] I.E. Kumar, S. Venkatasubramanian, C. Scheidegger, et al., in: Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 5491–5500.
[39] P.T. LoVerde, Adv. Exp. Med. Biol. 1154 (2019) 45–70.
[40] D. Cioli, L. Pica-Mattoccia, Parasitol. Res. 90 (2003) S3–S9.
[41] P.H.H. Schneeberger, J.T. Coulibaly, G. Panic, et al., Parasites Vectors 11 (2018) 1–12.
[42] X. Buol, K. Robeyns, C.C Garrido, et al., Pharmaceutics 12 (2020) 653.
[43] X. Buol, C.C Garrido, K. Robeyns, et al., Cryst. Growth Des. 20 (2020) 7979–7988.
[44] S. Yang, Q. Liu, W. Ji, et al., Molecules 27 (2022) 2022.
[45] Q. Liu, D. Yang, T. Chen, et al., Cryst. Growth Des. 21 (2021) 6321–6331.