Organic & Biomolecular Chemistry

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Org. Biomol. Chem.*, 2021, **19**, 6267

Accelerating the optimization of enzyme-catalyzed synthesis conditions *via* machine learning and reactivity descriptors†

Zhongyu Wan, *[a,b] Quan-De Wang, *[a] Dongchang Liu[c,e] and Jinhu Liang[d]

Enzyme-catalyzed synthesis reactions are of crucial importance for a wide range of applications. An accurate and rapid selection of optimal synthesis conditions is crucial and challenging for both human knowledge and computer predictions. In this work, a new scenario, which combines a data-driven machine learning (ML) model with reactivity descriptors, is developed to predict the optimal enzyme-catalyzed synthesis conditions and the reaction yield. Fourteen reactivity descriptors in total are constructed to describe 125 reactions (classified into five categories) included in different reaction mechanisms. Nineteen ML models are developed to train the dataset and the Quadratic support vector machine (SVM) model is found to exhibit the best performance. The Quadratic SVM model is then used to predict the optimal reaction conditions, which are subsequently used to obtain the highest yield among 109 200 reaction conditions with different molar ratios of substrates, solvents, water contents, enzyme concentrations and temperatures for each reaction. The proposed protocol should be generally applicable to a diverse range of chemical reactions and provides a black-box evaluation for optimizing the reaction conditions of organic synthesis reactions.

Received 2nd June 2021,
Accepted 16th June 2021

DOI: 10.1039/d1ob01066b

rsc.li/obc

1. Introduction

In recent years, bio-enzyme-catalyzed synthesis reactions have attracted significant attention because they can greatly reduce the production of unnecessary products.[1,2] In addition, enzyme-catalyzed reactions can occur at 0–100 °C under normal pressure, and they are much more energy-efficient than traditional synthesis processes, generally with cumbersome post-reaction treatments and environmental pollution problems.[3,4] However, the ability to accurately and rapidly select an optimal reaction condition remains challenging for all organic synthesis reactions, especially for enzyme-catalyzed synthesis reactions due to large affecting factors such as the molar ratio of substrates (mr), solvent oil–water partition coefficient (log P), water content (W), enzyme concentration (c), temperature (T) and time (H). Despite many successes in experimental research on optimal reaction conditions for typical reactions, the fast evaluation of appropriate reaction conditions and the prediction of the corresponding reaction yield remain challenging due to the complex relationship between different factors.

With the rapid development of the chemical space, traditional synthesis methods based on human knowledge cannot meet the needs of accelerated reaction discovery.[5,6] This encourages chemists to assess chemical reactivity by using computer predictions. Quantum chemical methods, especially density functional theory (DFT), provide powerful tools to predict reactivity trends of organic reactions, and have been widely employed to study reaction mechanisms.[7] However, such predictions can only predict reactivity information under ideal conditions. For typical organic synthesis reactions, reaction yields and products are affected by reaction conditions, such as temperatures, pressures, substrates, concentrations, *etc.*, which are hard to describe by traditional DFT calculations.

Machine learning (ML) models aiming to learn the correlation between a sequence of descriptors and chemical reactivity now receive significant attention due to the rapid expansions of the chemical space.[8,9] The earlier ML model used to predict chemical reactivity was usually recognized as the quan-

[a]*Jiangsu Key Laboratory of Coal-based Greenhouse Gas Control and Utilization, Low Carbon Energy Institute and School of Chemical Engineering, China University of Mining and Technology, Xuzhou, 221008, People's Republic of China. E-mail: quandewang@cumt.edu.cn, zhongyuwanxzit@163.com*
[b]*School of Science, City University of Hong Kong, Hong Kong SAR 999077, People's Republic of China*
[c]*School of Science, Xi'an Polytechnic University, Xi'an 710048, People's Republic of China*
[d]*School of Environment and Safety Engineering, North University of China, Taiyuan 030051, People's Republic of China*
[e]*Department of Physics, Sungkyunkwan University, Suwon 16419, Korea*

†Electronic supplementary information (ESI) available. See DOI: 10.1039/d1ob01066b

titative structure–activity relationship (QSAR) method. For example, Norrby and co-workers used QSAR and steric descriptors to predict the regio- and stereo-selectivity in palladium-catalyzed allylation reactions.[10] Thereafter, researchers have developed quantitative structure–selectivity relationships to predict the various properties of chemical reactions.[11–13]

Recent advances in high-level quantum chemical calculations together with high-throughput experimentations and data-mining techniques increase the data quality and also expand the dataset, which significantly promote the use of ML models in chemical reaction predictions. Doyle *et al.* used the random forest (RF) algorithm to predict the yield of the Buchwald–Hartwig coupling reaction at a specific temperature and in a specific solvent using a high-throughput dataset.[14] Denmark *et al.* carried out the accurate prediction of the selectivity of chiral phosphoric acid catalysts for specific reactions using artificial neural networks (ANN).[15] Chen *et al.* developed an efficient ML model to predict the reaction yields of typical electro-organic synthesis reactions by the introduction of three electro-descriptors.[16]

However, as one of the promising synthesis reactions, enzyme-catalyzed reactions have received little attention, especially for the prediction of optimal reaction conditions and yields. The major difficulties in the development of an ML model for such predictions are: (1) complex factors affecting the reaction process and (2) the appropriate descriptor definition for these factors. To this end, relatively simple descriptors are introduced in the present work to describe different factors of reaction conditions. These descriptors are then used as the input to develop an efficient ML model, aiming to predict the optimal reaction conditions of enzyme-catalyzed synthesis reactions ultimately.

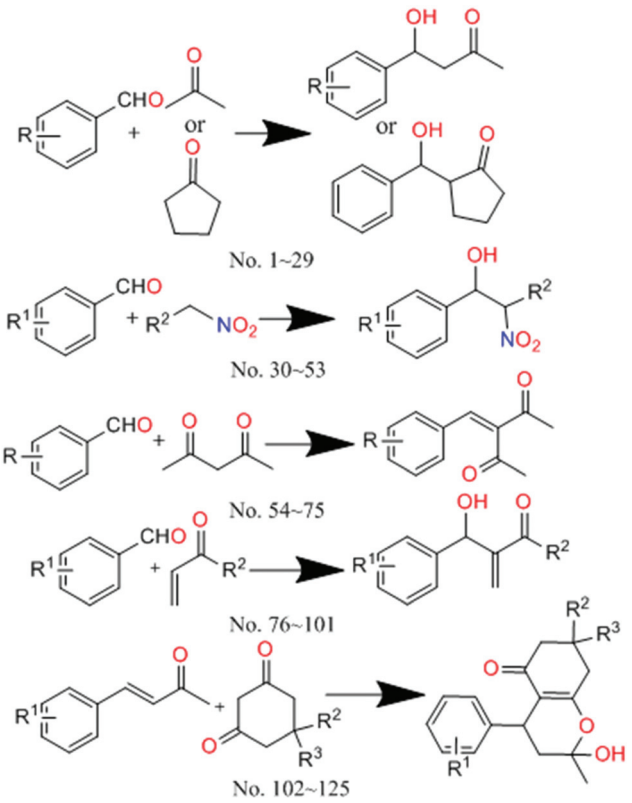

Fig. 1 Five types of chemical reactions in the dataset.

2. Methodologies

The first step of our work is developing a general framework. 125 reactions classified into five categories are selected, including the aldol reaction, nitro-aldol reaction, Knoevenagel condensation reaction, Baylis–Hillman reaction and Michael addition reaction, only involving an *Escherichia coli* enzyme (BioH) with their related reaction mechanisms (as shown in Fig. 1).[17–21] Four functional groups (aldehyde groups, nitro groups, double bonds and halo groups) and five- and six-membered rings are contained in the structures of the substrates. It can be seen that the structures of the substrates in the dataset are diverse. To incorporate the information of reaction mechanisms, descriptors from the molecular frontier orbital theory are introduced. Considering the large number of species and the molecular size of these species, the semi-empirical PM7 method[22] is used to optimize the geometries and compute the descriptors, *i.e.*, the highest occupied molecular orbital energy (E^i_{HOMO}), the lowest unoccupied molecular orbital energy (E^i_{LUMO}), the cavity surface (S^i_{COSMO}) and the cavity volume (V^i_{COSMO}) based on the conduct-like screening model for each substrate (i = 1, 2). Specifically, most of the initial structures of the reactants/products are adopted from the NIST Chemistry WebBook,[23] while the others are derived using the Avogadro software.[24] Then, we perform geometry optimization using the semi-empirical PM7 method. Currently, it is hard to consider the conformer effect on the computed properties, *i.e.*, HOMO and LUMO energies; thus, we use the lowest energy structures to compute quantum chemical descriptors. The optimized structures with conformers are rechecked to ensure that they are generally in good consistency with the general knowledge of organic chemistry, *i.e.*, the optimized geometries are in trans-structures. All quantum chemical calculations are performed using the MOPAC software.[25] In addition to the substrates with different structures, mr, log P, W, c, T and H also affect the reaction. Therefore, a total number of 14 descriptors related to reaction yields are adopted.

The entire dataset consists of a training set (100 chemical reactions) and a test set (25 chemical reactions) in a ratio of 80% and 20% at random. The training set is used to fit the known data to obtain the prediction model, and the test set is used to verify and evaluate the prediction effect. In the development of ML models, redundant information may be carried by some descriptors, which further increase the complexity of ML models. Thus, the number of descriptors selected in the models should be as small as possible to promote ML model development and reduce computational cost. For this purpose, correlation analysis is firstly employed to select important

descriptors used for the development of ML models. Then, we use the selected important descriptors as the input for 19 ML models to explore the most accurate ML model for the optimization of reaction conditions. Finally, the selected ML model is employed for the prediction of optimal conditions *via* high-throughput virtual screenings.

3. Results and discussion

3.1. Correlation analysis

In order to ensure the rationality of the prediction model, there should be no serious overlap between the descriptors in the model. Thus, the correlation degree needs to be evaluated. The Pearson correlation coefficient (r), which can intuitively reflect the degree of collinearity between two variables, is used as the indicator. The correlation coefficient matrix of 14 descriptors is computed as shown in Fig. 2.

Generally, if the correlation coefficient between two descriptors satisfies $|r| > 0.8$, it indicates that they have a strong linear overlap relationship. Hence, it is necessary to delete one of the descriptors. By calculating the averaged value of $|r|$ between one descriptor and the other descriptors $|\bar{r}|$, the descriptors with larger ones are eliminated. In the end, we eliminated three descriptors V^1_{COSMO}, E^1_{HOMO}, and V^2_{COSMO}. The remaining 11 descriptors are used to further establish the prediction model of yields.

3.2. Machine learning models

Most of the relationships between descriptors and properties usually tend to be nonlinear. ML algorithms are very suitable to fit nonlinear mathematical relationships. However, different types of ML models have different core algorithms, which are suitable for different datasets. Thus, the selection of appropriate ML algorithms is crucial for the accurate prediction of reaction yields. It is worth noting that it is hard to consider all existing ML methods due to the rapid development of various algorithms. Herein, we use 19 typical ML models based on their availability in the widely used MATLAB software.[26] These models include three regression trees (RT), six support vector machines (SVM), four Gaussian process regressions (GPR), two ensemble trees (ET), and four artificial neural networks (ANN) for model comparison as listed in Table 1. These models cover the widely used ML algorithms in chemistry nowadays. For ANN based on an error back propagation algorithm (BP-ANN), the Andrea rule and Xu Lu rule are used to determine the number of hidden layers.[27,28] Trainlm is selected as the training function, and the activation function in the neuron is Sigmoid. The remaining hyper-parameters of other ML models adopt the system default values; all ML model development and validation are carried out using the MATLAB software.[26]

The coefficient of determination (R^2) and the root mean square error (RMSE) of the test set are used to describe the accuracy of the prediction results. The larger the R^2 value, the smaller the RMSE value, indicating a better external predictive

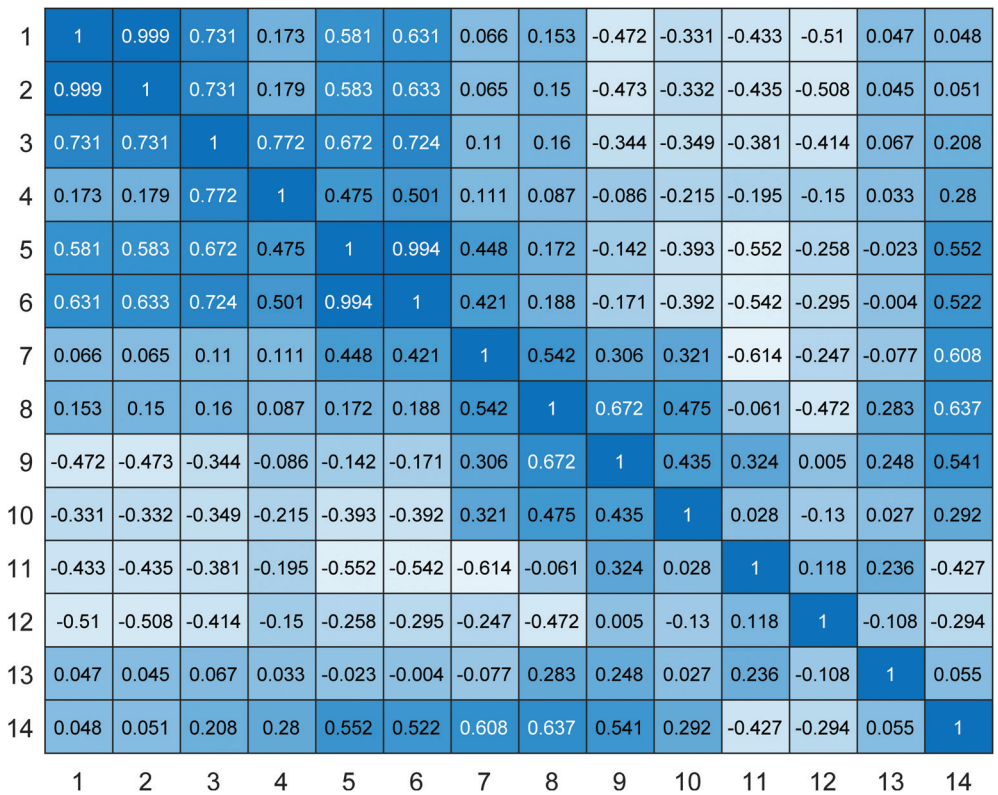

Fig. 2 The computed correlation coefficient matrix of the 14 descriptors.

Table 1 The R^2 and RMSE values from different ML models for the test set

Model	R^2	RMSE (%)
Coarse Gaussian SVM	0.27	26.7
Elman ANN	0.29	26.1
RBF ANN	0.33	25.6
Simple Tree	0.36	25.1
Cubic SVM	0.4	24.4
Linear SVM	0.41	24.1
Medium Tree	0.46	23.1
Bagged Trees	0.48	22.6
GRNN	0.51	21.4
Complex Tree	0.54	21.2
BP ANN	0.57	20.3
Boosted Trees	0.6	19.8
Fine Gaussian SVM	0.61	19.6
Matern 5/2 GPR	0.66	18.3
Medium Gaussian SVM	0.66	18.1
Rational Quadratic GPR	0.7	17.1
Exponential GPR	0.71	16.8
Squared Exponential GPR	0.74	15.9
Quadratic SVM	0.86	10.1

ability of the model. Table 1 also shows the test results of the two values from different ML models. Generally, a model with a large R^2 value is accompanied by a small RMSE value. It is found that the Quadratic SVM model has the highest R^2 value and the lowest RMSE value (R^2_{Test} = 0.86; RMSE = 10.1%) among the 19 ML models, indicating that this model has an excellent accuracy for predicting the yield of enzyme-catalyzed synthesis in the dataset. In addition, it can be seen that only the Quadratic SVM model exhibits an R^2 value exceeding 0.8. Thus, this model is employed for further studies.

The Quadratic SVM model is then used to evaluate its performance in the prediction of reaction yields for both the training and test datasets as shown in Fig. 3. For both the training and test datasets, most of the scattered points are evenly distributed around the straight line $y = x$ (*i.e.*, predicted value = experimental value), indicating that the developed ML model shows good prediction accuracy. The value of R^2_{Train} is 0.88, indicating that the model fits the relationship between the descriptor and the yield well. The difference of R^2 between the training set and the test set is only 0.02, which implies that the model has overcome the problem of overfitting during the training process. As the Quadratic SVM model is an ML model with a quadratic function as the basis function, the relationship between the descriptors and the yield tends to be quadratic.

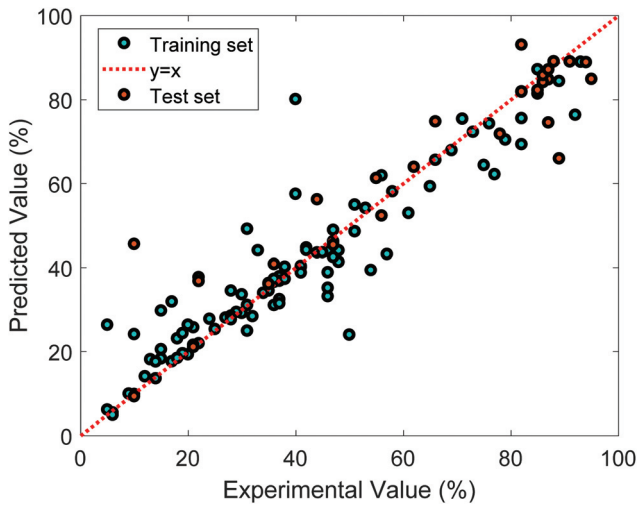

Fig. 3 Predicted reaction yields using the Quadratic SVM model for both the training and test datasets.

In order to show the necessity of deleting high-relevance descriptors, Quadratic SVM is used to establish a prediction model for the 14 original descriptors and the remaining 11 descriptors after removing the high-relevance descriptors. The R^2_{Test} value of the former is 0.87 and that of the latter is 0.86. The difference between them is only 0.01, but there is a significant difference in the prediction speed. The prediction speed of the model without descriptors removed is ~2500 per second. After excluding high-correlation descriptors, the prediction speed is increased to 3100 per second. This shows that the introduction of high-relevance descriptors will not significantly improve the prediction accuracy, but it will have a serious impact on the speed of modeling.

In order to further prove that the Quadratic SVM model has a universal predictive ability for different types of reactions, we leave one type out as a test set. The predicted results for different reactions are shown in Table 2.

It can be seen in Table 2 that for different types of chemical reactions, the R^2 value of the test set ranges from 0.77 to 0.87, which is very close to 1, indicating that Quadratic SVM can effectively predict different types of reactions, and this model has excellent robustness.

Considering that the division of the training set and test set is random, the model will be affected by random factors. As one of the most commonly used verification methods, 10-fold cross-validation can effectively avoid the influence of randomness. It divides the entire dataset into 10 parts, 9 of which are used for training, and the other part is used for performance evaluation. The advantage of this method is that each sample can be used as test data for training and verification only once, which has greater credibility than random sampling. The cross-validation coefficients (Q^2) of the 19 ML models are shown in Table 3.

The model with a higher Q^2 value has better stability; it can be seen in Table 3 that the Quadratic SVM model has the highest stability, indicating that the choice of it is reasonable.

Table 2 Predictive performance of the Quadratic SVM model for different types of chemical reactions

No.	Training set	Test set	R^2_{Test}
1	30–125	1–29	0.84
2	1–29, 54–125	30–53	0.79
3	1–53, 76–125	54–75	0.81
4	1–75, 102–125	76–101	0.87
5	1–101	102–125	0.77

Table 3 The Q^2 values from different ML models

Model	Q^2
Cubic SVM	0.25
Elman ANN	0.29
Linear SVM	0.29
Coarse Gaussian SVM	0.31
Medium Tree	0.35
Simple Tree	0.36
RBF ANN	0.38
Complex Tree	0.39
Bagged Trees	0.43
Boosted Trees	0.49
GRNN	0.51
Fine Gaussian SVM	0.58
Medium Gaussian SVM	0.59
Rational Quadratic GPR	0.62
BP ANN	0.62
Matern 5/2 GPR	0.63
Squared Exponential GPR	0.63
Exponential GPR	0.64
Quadratic SVM	0.75

3.3. Optimizing reaction conditions *via* high-throughput virtual screenings

As shown previously, the ultimate goal of developing appropriate ML models is to accurately predict the optimal reaction conditions to obtain the highest reaction yield for specific products. Herein, three enzyme-catalyzed reactions outside the dataset shown in Fig. 4 are used to further validate the Quadratic SVM model. They have different but similar structures to reactions 1, 102 and 54 in Table S1,† respectively. Reactions 1, 2, and 3 in Table 4 have the same reaction conditions as reactions 1, 102, and 54 in Table S1,† but the positions of the nitro substituents are changed from *para* to *meta*, *ortho*, and *ortho* positions, respectively. They are also used for reaction condition optimization. Table 4 lists the corresponding initial reaction conditions for the three reactions. The Quadratic SVM model is used to predict the reaction yields against the experimental results as shown in Fig. 4.

It can be seen that the developed Quadratic SVM ML model can accurately predict the reaction yields compared with the experimental results. Therefore, it is proved that the developed Quadratic SVM ML model not only exhibits good external prediction ability for the dataset in this work, but also shows good prediction accuracy for samples outside the dataset. To further evaluate the applicability of the developed ML model, we perform high-throughput virtual screenings for the three reactions by changing the different reaction factors that are affecting the reaction yield to obtain the optimal reaction conditions. It is worth noting that the optimal conditions obtained in the literature are relative rather than absolute. This work traverses all the conditions through the ML model, and can obtain the potential optimal conditions. Especially for reactions 2 and 3 in Table 4, we have found new optimal conditions, which can provide a reference and guidance for experimenters to carry out efficient synthesis. Considering that reaction 2 in Table 4 and reaction 102 in Table S1† belong to the same reaction type and have similar structures, reactions 102–108 in Table S1† are the process of single-factor optimization of the reaction concentration. It can be seen that as the

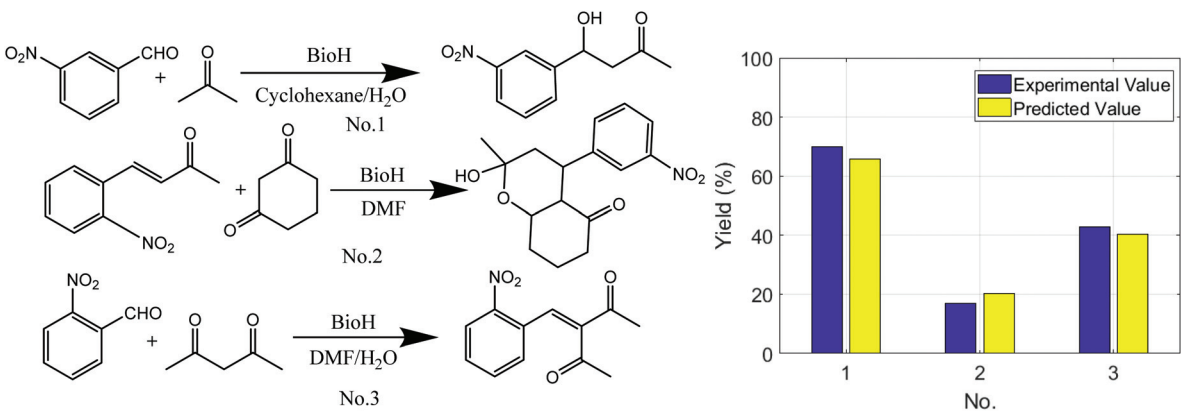

Fig. 4 The three enzyme-catalyzed synthesis reactions used for ML model validation and reaction condition optimization.

Table 4 Names and values of different factors

No.	Reaction	mr	Solvent	W	c	T	H	Predicted yield (%)
1	Initial conditions	15	Cyclohexane	20	1.5	37	200	66
	Optimal conditions	15	Cyclohexane	20	2	45	200	73
2	Initial conditions	1	Dimethyl formamide	0	1	37	120	20
	Optimal conditions	15	Dimethyl sulfoxide	0	5	50	120	70
3	Initial conditions	15	Dimethyl formamide	20	3	37	168	40
	Optimal conditions	10	Dimethyl sulfoxide	20	6	40	168	71

Table 5 Names and values of different factors

Condition	Value	Numbers
Substrate molar ratio	mr = 1, 5, 10, 15, 20, 25, 30	7
Solvent	Log P = −1.35 (dimethyl sulfoxide), −1.3 (dimethyl formamide), −0.5 (1,4-dioxane), −0.44 (1,2-dimethoxyethane), −0.33 (acetonitrile), 0.46 (tetrahydrofuran), 0.94 (*tert*-butyl methyl ether), 2.56 (toluene), 3.35 (cyclohexane), 3.9 (hexane)	10
Percent water content (%)	W = 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 80	13
Enzyme concentration (mg mL^{-1})	c = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4, 5, 6	12
Temperature (°C)	T = 10, 15, 20, 25, 30, 37, 40, 45, 50, 60	10

concentration increases, the yield of the reaction also increases, indicating that in the predicted optimal conditions, the result of c = 5 is credible to a certain extent. For reaction 3 in Table 4, its initial conditions are the same as those of reaction 54 in Table S1,† and their reactants are relatively similar in structure. Reactions 54–59 in Table S1† are the process of single-factor optimization of the solvent. It can be seen that as the log P value increases, the yield of the reaction continues to decrease. Therefore, it is reasonable to choose dimethyl sulfoxide with the smallest log P value in the optimal conditions of prediction. Reactions 66–70 and 54 in Table S1† are the single-factor optimization of the concentration, and it can be seen that the increase of the concentration leads to an increase of the yield. Therefore, in the predicted optimal conditions, c = 6 is also acceptable. However, there are also mutual influences between different reaction factors. However, the results of theoretical predictions still need to be verified by experiments.

Table 5 lists the factors and their corresponding ranges used for high-throughput virtual screenings. The changes in the substrate structure and the reaction time of the chemical reactions are not considered in the screening process, since they usually involve reaction mechanism studies and are not changed too much in real synthesis experiments. Five reaction factors, including seven kinds of substrate molar ratios, ten kinds of solvents, thirteen kinds of water contents, twelve kinds of enzyme concentrations and ten kinds of temperatures, are shown in Table 5. A total number of 109 200 reaction conditions for each reaction are calculated by combining these factors. The developed Quadratic SVM model is used to predict the reaction yields of these different conditions, and the predicted optimal conditions with the highest yields for the three reactions are listed in Table 4. The number of processor cores of the computer is 2, the memory is 128 MB, and the CPU model is Intel i5-2415M, which can reach a speed of predicting 60 000 responses within one minute.

4. Conclusion

This work proposes a novel scenario that combines the data-driven machine learning method and descriptors to optimize the reaction conditions of enzyme-catalyzed synthesis. The descriptors related to the synthesis yield are obtained *via* quantum chemistry calculation and the collection of reaction conditions. Correlation analysis is used to delete overlapping descriptors. The remaining eleven descriptors are used to build a machine learning model for the prediction of reaction yields. The Quadratic SVM model exhibits the best agreement with the experimental results. It is further used to predict the yield of three reactions with 109 200 conditions, and then the optimal conditions corresponding to the highest yield are calculated. This work can quickly find the optimal reaction conditions and provide guidance for the design of organic synthesis reactions.

Code availability

The code is available on the GitHub website (https://github.com/phydcliu/OBC).

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities of China (No. 2020ZDPYMS05).

References

1 A. O. Brachmann, S. I. Probst, J. Rüthi, D. Dudko, H. B. Bode and J. Piel, A Desaturase-Like Enzyme Catalyzes Oxazole Formation in Pseudomonas Indolyloxazole Alkaloids, *Angew. Chem., Int. Ed.*, 2021, **60**(16), 8781–8785.
2 R. Cruz-Valencia, A. A. Arvizu-Flores, J. A. Rosas-Rodríguez and E. M. Valenzuela-Soto, Effect of the drug cyclophosphamide on the activity of porcine kidney betaine aldehyde dehydrogenase, *Mol. Cell. Biochem.*, 2021, **476**, 1467–1475.
3 R. Jójárt, S. A. S. Tahaei, P. Trungel-Nagy, Z. Kele, R. Minorics, G. Paragi, I. Zupkó and E. Mernyák, Synthesis and evaluation of anticancer activities of 2- or 4-substituted 3-(N-benzyltriazolylmethyl)-13α-oestrone derivatives, *J. Enzyme Inhib. Med. Chem.*, 2021, **36**(1), 58–67.
4 Y. Li, L. Yi, S. Cheng, Y. Wang and X. Xu, Inhibition of canine distemper virus replication by blocking pyrimidine nucleotide synthesis with A77 1726, the active metabolite

of the anti-inflammatory drug leflunomide, *J. Gen. Virol.*, 2021, **102**(3), 001534.

5 C. W. Coley, N. S. Eyke and K. F. Jensen, Autonomous Discovery in the Chemical Sciences Part II: Outlook, *Angew. Chem.*, 2020, **59**(52), 23414–23436.

6 C. W. Coley, N. S. Eyke and K. F. Jensen, Autonomous Discovery in the Chemical Sciences Part I: Progress, *Angew. Chem.*, 2020, **59**(51), 22858–22893.

7 P. H.-Y. Cheong, C. Y. Legault, J. M. Um, N. Çelebi-Ölçüm and K. N. Houk, Quantum Mechanical Investigations of Organocatalysis: Mechanisms, Reactivities, and Selectivities, *Chem. Rev.*, 2011, **111**(8), 5042–5137.

8 N. S. Eyke, W. H. Green and K. F. Jensen, Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening, *React. Chem. Eng.*, 2020, **5**(10), 1963–1972.

9 Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors, *Chem. Sci.*, 2021, **12**, 2198–2208.

10 E. Hansen, A. R. Rosales, B. Tutkowski, P.-O. Norrby and O. Wiest, Prediction of Stereochemistry using Q2MM, *Acc. Chem. Res.*, 2016, **49**(5), 996–1005.

11 J. Ohyama, S. Nishimura and K. Takahashi, Data Driven Determination of Reaction Conditions in Oxidative Coupling of Methane via Machine Learning, *ChemCatChem*, 2019, **11**(17), 4307–4313.

12 M. H. S. Shegler and M. P. Waller, Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction, *Chem. – Eur. J.*, 2017, **23**(5), 5966–5971.

13 K. Takahashi and I. Miyazato, Rapid estimation of activation energy in heterogeneous catalytic reactions via machine learning, *J. Comput. Chem.*, 2018, **39**(28), 2405–2408.

14 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, Predicting reaction performance in C–N cross-coupling using machine learning, *Science*, 2018, **360**(6385), 186–190.

15 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning, *Nature*, 2019, **363**(6424), eaau5631.

16 Y. Chen, B. Tian, Z. Cheng, X. Li, M. Huang, Y. Sun, S. Liu, X. Cheng, S. Li and M. Ding, Electro-Descriptors for the Performance Prediction of Electro-Organic Synthesis, *Angew. Chem., Int. Ed.*, 2021, **60**(8), 4199–4207.

17 L. Jiang and H. W. Yu, An example of enzymatic promiscuity: the Baylis–Hillman reaction catalyzed by a biotin esterase (BioH) from Escherichia coli, *J. Biotechnol. Lett.*, 2014, **36**(1), 99.

18 J. Ling and Y. U. Hongwei, Enzymatic promiscuity: Escherichia coli BioH esterase-catalysed Aldol reaction and Knoevenagel reaction, *Chem. Res. Chin. Univ.*, 2014, **30**(002), 289–292.

19 L. Jiang, B. Wang, R.-R. Li, S. Shen and H.-W. Yu, "Amano" lipase DF-catalyzed efficient synthesis of 2,2′-arylmethylene dicyclohexane-1,3-dione derivatives in anhydrous media, *Chin. Chem. Lett.*, 2014, **25**(8), 1190–1192.

20 L. Jiang, B. Wang, R. R. Li, S. Shen, H. W. Yu and L. D. Ye, Catalytic promiscuity of Escherichia coli BioH esterase: Application in the synthesis of 3,4-dihydropyran derivatives, *Process Biochem.*, 2014, **49**(7), 1135–1138.

21 L. Jiang, *Studies on the Catalytic Promiscuity of Esterase/Lipase in Carbon-Carbon Bond Formation and Heterocycle Compounds Preparation*, Doctor Thesis, Zhejiang University, Hangzhou, Zhejiang, China, 2014.

22 J. J. P. Stewart, Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, *J. Mol. Model.*, 2013, **19**, 1–32.

23 NIST Chemistry WebBook, *NIST Standard Reference Database Number 69*, ed. P. J. Linstrom and W. G. Mallard, National Institute of Standards and Technology, retrieved June 6, 2021.

24 M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison, Avogadro: an advanced semantic chemical editor, visualization, and analysis platform, *J. Cheminf.*, 2012, **4**(1), 17.

25 J. J. P. Stewart, *MOPAC2016*, Stewart Computational Chemistry, Colorado Springs, CO, USA, http://OpenMOPAC.net2016.

26 MATLAB, *version R2019b*, The MathWorks Inc., 2019.

27 T. A. Andrea and H. Kalayeh, Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors, *J. Med. Chem.*, 1991, **34**(9), 2824–2836.

28 D. Liu, W. Zhang and L. Xu, Quantitative Structure-Activity/Property Relationships for Chiral Hydroxy Acids and Amino Acids, *Acta Chim. Sin.*, 2009, **67**(2), 145–150.