BMC Chemical Engineering

RESEARCH ARTICLE Open Access

Hard-threshold neural network-based prediction of organic synthetic outcomes

Haoyang Hu[1] and Zhihong Yuan[1,2*]

Abstract

Retrosynthetic analysis is a canonical technique for planning the synthesis route of organic molecules in drug discovery and development. In this technique, the screening of synthetic tree branches requires accurate forward reaction prediction, but existing software is far from completing this step independently. Previous studies attempted to apply a neural network to forward reaction prediction, but the accuracy was not satisfying. Through using the Edit Vector-based description and extended-connectivity fingerprints to transform the reaction into a vector, this study focuses on the update of the neural network to improve the template-based forward reaction prediction. Hard-threshold activation and the target propagation algorithm are implemented by introducing mixed convex-combinatorial optimization. Comparative tests were conducted to explore the optimal hyperparameter set. Using 15,000 experimental reaction data extracted from granted United States patents, the proposed hard-threshold neural network was systematically trained and tested. The results demonstrated that a higher prediction accuracy was obtained than that for the traditional neural network with backpropagation algorithm. Some successfully predicted reaction examples are also briefly illustrated.

Keywords: Medicine development, Retrosynthetic analysis, Outcome prediction, Hard-threshold neural network, Combinatorial optimization

Introduction

Drug discovery and development are two of the most important tasks of the pharmaceutical industry. To meet customers' increasing demands while guaranteeing good potency and minimal side effects, the structures of new drug molecules have become increasingly complicated. Meanwhile, drugs that have such complexity cannot be manufactured in an acceptable time period using existing R&D technology; that is, extensive pharma R&D activities are dramatically required, and a relatively short discovery-development-deployment cycle is desirable [1].

Over recent decades, as a rate-limiting factor, innovations in organic synthesis have significantly enabled the discovery and development of important life-changing medicines, thereby improving the health of patients worldwide [2, 3]. Nevertheless, innovations and excellence in organic synthesis are expected to be the most powerful driver for all phases of drug discovery and development. Recently, chemists enthusiastically applied advanced machine learning and artificial intelligence (AI) technologies toward the synthesis of drug molecules. For example, the AI-driven discovery of drug molecules [4–6], automated planning of synthetic routes [7–9], machine learning-driven optimization of reaction conditions [10–12] and autonomous assembly of synthetic processes [13–15].

Synthesis planning, which is regarded as the central element of organic synthesis, can be traced back to the 1960s [16]. Traditional computer-aided approaches for synthesis planning have different disadvantages, such as low efficiency, poor repeatability and high experimental cost. Additionally, many new compounds and reactions

* Correspondence: zhihongyuan@mail.tsinghua.edu.cn
[1]Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
[2]State Key Laboratory of Chemical Engineering, Tsinghua University, Beijing 100084, China

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

have been discovered, which highly requires novel synthesis planning approaches. For example, the numbers of reactions and compounds currently contained in the Reaxys database have exceeded 40 million and 100 million, respectively. During the last 3 years, a variety of machine learning and AI methods, such as random forest, automated reasoning, support vector machines, and more recently, deep learning, have demonstrated their capacity for organic molecule discovery, design and production [13, 17–19]. Clearly, the application of the aforementioned new technologies for end-to-end organic molecule discovery and development will be key to achieving fully automated synthesis planning [20].

Retrosynthetic analysis is a canonical technique used to plan the synthesis of small organic molecules for drug discovery [21]. Generally, retrosynthesis analysis consists of four steps: (1) determine the target compound; (2) disconnect certain bonds that are considered to be easy to form according to known chemical knowledge in the target compound as the reverse of the reaction, and search for possible precursors in this manner; (3) repeat step 2 for all the precursors to form a synthesis tree, and expand it until all precursors are available; and (4) evaluate all the branches of the synthetic tree individually, and then take the most possible branch as the optimal route. Because different groups of reaction sites in the same group of precursors may exist, that is, the difference between real reactions and expected reactions in the branches may occur, step 4 is therefore critical, but also the most difficult. Forward reaction prediction is necessary to guarantee the correct evaluation of each branch. The main aim of this paper is to present machine learning approaches for assisting forward reaction prediction.

The existing approaches for forward reaction prediction can be categorized as template-based and template-free methods. For example, Coley et al. [17] applied reaction templates to reactants to generate as many candidate reactions as possible, which were then used to train a neural network. Whereas, for template-free methods, Schwaller et al. [22] compared chemical reactions from reactants to products to translations from one language to another so that forward reaction prediction could be transformed into machine translation and solved by training a seq-to-seq recurrent neural network that directly took reactants' simplified molecular input line entry specification (SMILES) as input.

As the research target is to discover new reactions that are meaningful for organic synthesis according to existing reaction mechanisms, in this paper, the template-based approach is adopted to fully reuse the discovered reaction rules summarized from experimental results to date. Specifically, a novel hard-threshold-based deep neural network is adopted to improve the accuracy of forward reaction prediction. In detail, hard-threshold activation and the target propagation algorithm are implemented by introducing mixed convex-combinatorial optimization. Comparative tests are conducted to explore the optimal hyperparameter set. The reminder of the paper is structured as follows: forward reaction prediction is described first, and then a hard-threshold neural network is briefly described; next, results and discussions are presented; finally, the conclusion is presented.

Forward reaction prediction

The overall approach for template-based forward reaction prediction is summarized in Fig. 1.

Given a certain group of reactants to predict, reaction templates extracted from a known popular template set are applied to generate a group of candidate reactions. Then the candidate reactions are converted into vectors. The vectors of the group of candidates are then used for training the hard-threshold-based deep neural network. Generally, the data augmentation strategy, vectorized description of the reaction and candidate reaction selection are key components.

Data augmentation strategy

Considering that the existing chemical knowledge databases only contains real reactions that take place in practice, to train the neural network to identify real reactions, it is necessary to adopt a data augmentation strategy to expand the database. Specifically, real reactions are first transformed into SMILES, which are further converted to form the template set in the SMARTS format via a heuristic algorithm. Then, all feasible popular templates are applied to each group of reactants in real reactions to generate a large amount of fake reactions that actually cannot occur in practice. Finally, the augmented reaction dataset including real reactions labeled as positive examples along with fake reactions labeled as negative examples are provided to the hard-threshold-based deep neural network.

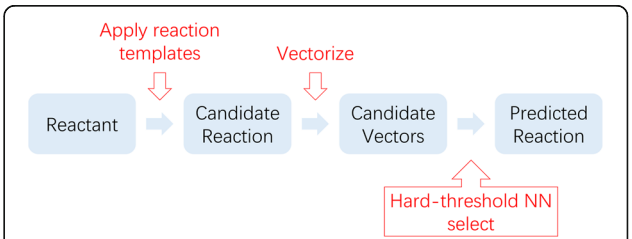


Fig. 1 Illustration of the template-based forward reaction prediction process

Vectorized description of the reaction

Because a neural network traditionally requires vector format input, the augmented reactions must be converted to vectors in an appropriate manner. Clearly, the strategy for choosing features to construct the vectors significantly affects the final prediction accuracy.

Following the "Edit Vector" format [17], an atom is described as a 32-dimensional feature vector and a bond is described as a four-dimensional feature vector. Each reaction is first decomposed into four types of basic Edits, that is, hydrogen loss, hydrogen gain, bond loss, and bond gain, and then described as a combination of feature vectors. Note that the loss and gain of non-hydrogen atoms are considered as the loss and gain of the bond between two atoms, respectively. As a result, the Edit Vector for either the bond loss or bond gain includes information from the two atoms and the bond. For example, hydrogen loss is described as a corresponding 32-dimensional atom feature vector, whereas bond loss is composed of two corresponding atom feature vectors and the corresponding bond feature vector, which thus can be described as a 68-dimensional feature vector.

To exactly explain how the Edit Vector describes the reaction, in the following is a simplified example based on the chemical reaction expressed in Fig. 2. The atom feature vector in this example has six dimensions [is_carbon, is_nitrogen, is_oxygen, is_chlorine, num_Hs, num_non-H_atoms] chosen from the aforementioned 32 dimensions, and the bond feature vector has four dimensions [is_single, is_aromatic, is_double, is_triple].

The above reaction can be decomposed into three Edits: atom 1 (nitrogen) loses a hydrogen, atoms 2 (carbon) and 3 (chlorine) lose a single bond, and atoms 1 and 3 gain a single bond. In the reactant, atom 1 has one hydrogen and two non-hydrogen neighbors, atom 2 has no hydrogen but is surrounded by three non-hydrogen neighbors, and atom 3 has one non-hydrogen neighbor without any hydrogen. Therefore, the feature vectors of atoms 1, 2 and 3, along with the two bonds, can be expressed as

$$a_1 = [0, 1, 0, 0, 1, 2]$$

$$a_2 = [1, 0, 0, 0, 0, 3]$$

$$a_3 = [0, 0, 0, 1, 0, 1]$$

$$b_{12} = [1, 0, 0, 0]$$

$$b_{23} = [1, 0, 0, 0].$$

The Edit Vectors for the hydrogen loss and gain (e_1 and e_2, respectively) are directly taken as the feature vectors of the corresponding atoms, and those of the bond loss and gain (e_3 and e_4, respectively) are taken as the connection of the feature vectors of the corresponding atoms and bonds:

$$e_1 = [a_1] = [[0, 1, 0, 0, 1, 2]]$$

$$e_2 = [\mathbf{0}] = [[0, 0, 0, 0, 0, 0]]$$

$$e_3 = [a_2/b_{23}/a_3] = [[1, 0, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1]]$$

$$e_4 = [a_1/b_{12}/a_2] = [[0, 1, 0, 0, 1, 2, 1, 0, 0, 0, 1, 0, 0, 0, 0, 3]],$$

where "/" represents the connection of feature vectors, and the outermost brackets indicate that the Edit Vector is the combination of all the vectors of the basic Edit in a reaction. For instance, if one reaction has four atoms that lose hydrogen in the reactant, its Edit Vector for the hydrogen loss (e_1) has four six-dimensional feature vectors.

Candidate reaction selection

The selection step uses a complex neural network that consists of several subnetworks, as shown in Fig. 3. For each candidate reaction, the aforementioned four Edit Vectors are calculated, and then provided as input to four corresponding subnetworks. The sum of outputs from the four subnetworks is then fed to the lower integrating subnetwork to produce scalar probability scores. The above steps are repeated for all candidate reactions, and then all the probability scores are normalized using the softmax method to estimate the probability of occurrence of each candidate reaction. Finally, all candidate reactions are sorted by the probability score, and the candidate that ranks first is regarded as the prediction result. The prediction is correct if its outcome has the same SMILES as the recorded reaction's, and vice versa.

To improve the prediction accuracy, a hybrid model that uses the Edit Vector and extended-connectivity

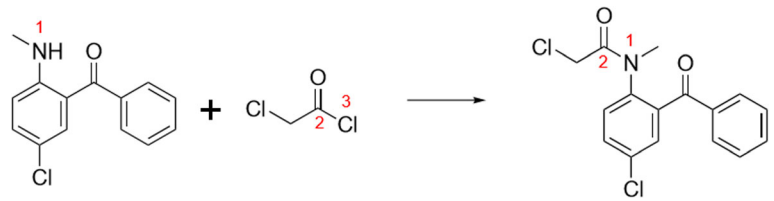


Fig. 2 Reaction between chloroacetyl chloride and 2-methylamino-5-chlorobenzophenone

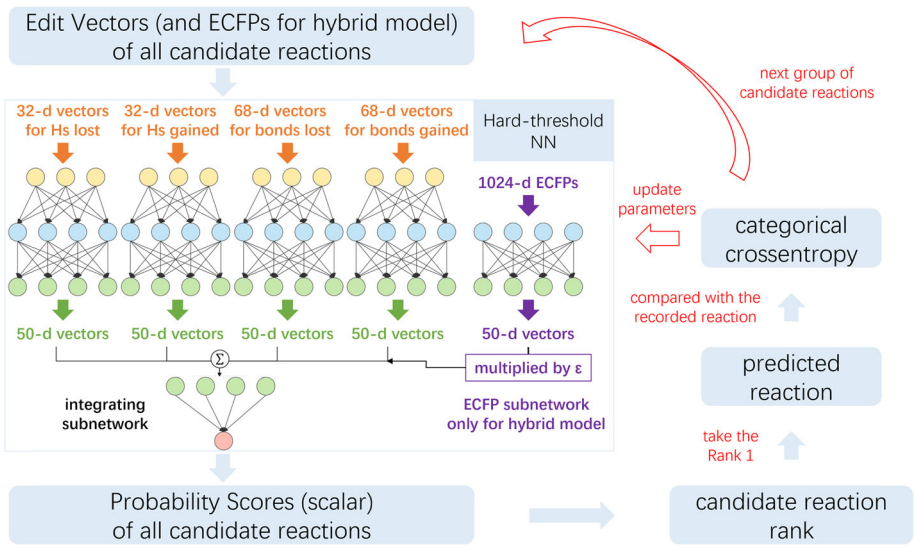


Fig. 3 Illustration of the selection process

fingerprint (ECFP) [23] is also considered in this paper. The only difference between the two models is that an extra subnetwork without a hidden layer, which evaluates the ECFP, is added to the hybrid model, as shown in the middle of Fig. 3. The output of the ECFP subnetwork is multiplied by ε when the subnetwork outputs are summed before input to the final integrating subnetwork, where ε is the mixing factor. By adjusting ε, the proportion of the ECFP subnetwork's output can be precisely controlled.

Hard-threshold neural network

The neural network, particularly the deep neural network learning, is currently the most popular machine learning algorithm and has powerful fitting capability [24]. However, with the ceaseless expansion of the size of the neural network, a series of problems, particularly gradient vanishing and gradient explosion, often occur. To eliminate this dilemma, in this paper, a hard-threshold neural network is applied to predict the outcomes of organic synthesis.

Constructing a hard-threshold neural network

"Hard-threshold neural network" refers to a neural network that has hard-threshold activation, which includes step activation and staircase activation shown in Fig. 4a and b, respectively. Staircase activation is the sum of

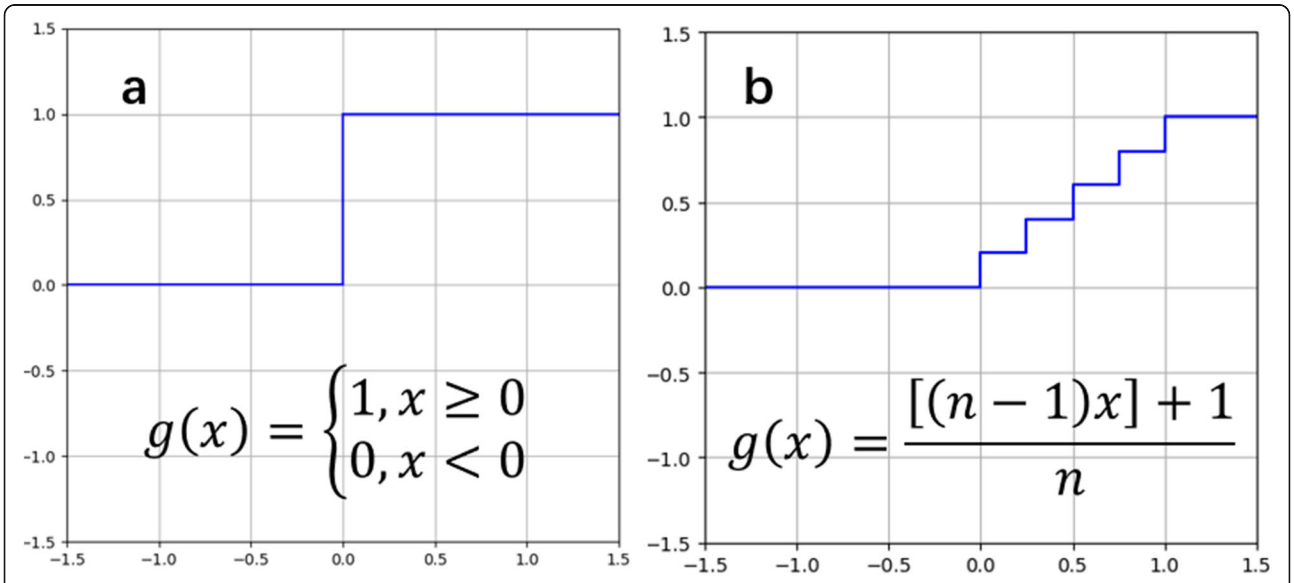


Fig. 4 Hard-threshold activations: **a** step activation and **b** staircase activation

some step activations. Hard-threshold activation was used in early contributions to perform binary classification before the neural network had been proposed. Hard-threshold activation has a constant derivative that is zero, which can effectively avoid gradient vanishing and gradient explosion. Additionally, the scale of the output is almost fixed and insensitive to the scale of the input, which helps to avoid certain abnormal propagation and simplify the computation. However, the zero derivative of hard-threshold activation also prevents it from being trained using traditional backpropagation.

Target propagation algorithm

A new backpropagation algorithm is required to train the hard-threshold neural network to bypass the zero derivative of hard-threshold activation.

Based on the "target propagation" concept [25], a new target propagation algorithm called FTPROP-MB was recently proposed [26]. Essentially, because a perceptron with step activation is trainable, the hard-threshold neural network could also be trainable if it could be decomposed into a perceptron. Specifically, a target vector t_d is introduced to represent what the d^{th} layer is supposed to output for all hard-threshold activation layers. After the normal forward propagation procedure for each layer, FTPROP-MB determines t_d first, and then introduces a layer loss L_d, which is used to compute the gradient, similar to training a perceptron, so that the weights can be updated.

Considering that the output of hard-threshold activation is a set of discrete values, the determination of t_d can be reduced to a combinatorial optimization problem. In detail, the question of how to optimize t_d regarding the overall loss and layer loss can be expressed in a standard form as follows:

$$\begin{aligned} &minL_d(\boldsymbol{z_d}, \boldsymbol{t_d}) \\ &\boldsymbol{z_d} = \boldsymbol{W_d t_{d-1}} \\ &s.t. \boldsymbol{t_{d-1}} \in \{0,1\}^n, \end{aligned} \tag{1}$$

where W_d and z_d represent the weight and pre-activation output of the d^{th} layer, respectively. The search space is large and discrete because all the components of t_d are restricted to 0 and 1, so it is difficult for common search algorithms to determine the optimal solution in a reasonable time horizon. Because layer loss is typically convex, FTPROP-MB determines the target vector t_d in the d^{th} layer using a heuristic method: compute the derivative of layer loss in the $(d+1)^{th}$ layer L_{d+1} with respect to the d^{th} layer's output h_d, and then set t_d according to the opposite sign of this derivative. This method can be mathematically formulated as

$$t_{dj} = r(h_{dj}) \triangleq sign\left[-\frac{\partial}{\partial h_{dj}} L_{d+1}(\boldsymbol{z_{d+1}}, \boldsymbol{t_{d+1}})\right]. \tag{2}$$

When the layer loss function is convex, the negative partial derivative of L_{d+1} on h_{dj} points to the global minimal of L_{d+1}. Consider $h_{dj} = -1$ as an example. If $r(h_{dj}) = -1$, which indicates that the partial derivative of L_{d+1} is positive, clearly L_{d+1} increases if $h_{dj} = +1$ when fixing the other components of $\boldsymbol{h_d}$; thus, $t_{dj} = r(h_{dj}) = -1$. By contrast, when $r(h_{dj}) = +1$, which means that the partial derivative of L_{d+1} is negative, it is not known exactly whether L_{d+1} increases or decreases if $h_{dj} = +1$. However, the difference between h_{dj} and $r(h_{dj})$ indicates that the current value of h_{dj} is a lack of confidence, so a natural choice is to lead z_{dj} to zero by adjusting t_{dj} to make it more possible for h_{dj} to flip, w.r.t $t_{dj} = r(h_{dj}) = +1$.

To summarize, the training process of an n-layer hard-threshold neural network has both an optimization problem on the weights and convex-combinatorial optimization problem on the target vectors; hence, a mixed convex-combinatorial optimization problem is formed. A block diagram of the target propagation algorithm is shown in Fig. 5.

Layer loss function

Till now we are still encountering the problem of choosing the layer loss function. According to related work [27], it is acceptable to adopt soft hinge loss and weighing according to the gradient, which is shown in Fig. 6.

Methods

Preparing the reaction database

The reaction datasets were originally extracted from the 1976–2013 USPTO dataset compiled by Lowe [28]. Based on popular template sets [29, 30], the original extracted reaction datasets were reduced to 15,000 groups of reactants corresponding to 15,000 real reactions. Then, approximately 5 million candidate reactions, including real and fake reactions, were generated using the aforementioned data augmentation strategy and stored in MongoDB format.

Structure of the edit vector

The atom features used in this paper are much more complex than the simplified example illustrated above, whereas the bond features are the same. The specific structure is shown in Table 1.

Structure of the ECFP for the hybrid model

A molecular fingerprint is also a common method for vectorizing molecules. A fingerprint is typically a 0–1 vector with an adjustable dimension, and is equivalent to the hash of a molecule. The ECFP proposed by Rogers et al. [23] in 2010 is a circular fingerprint based on the Morgan algorithm, and has become the de facto

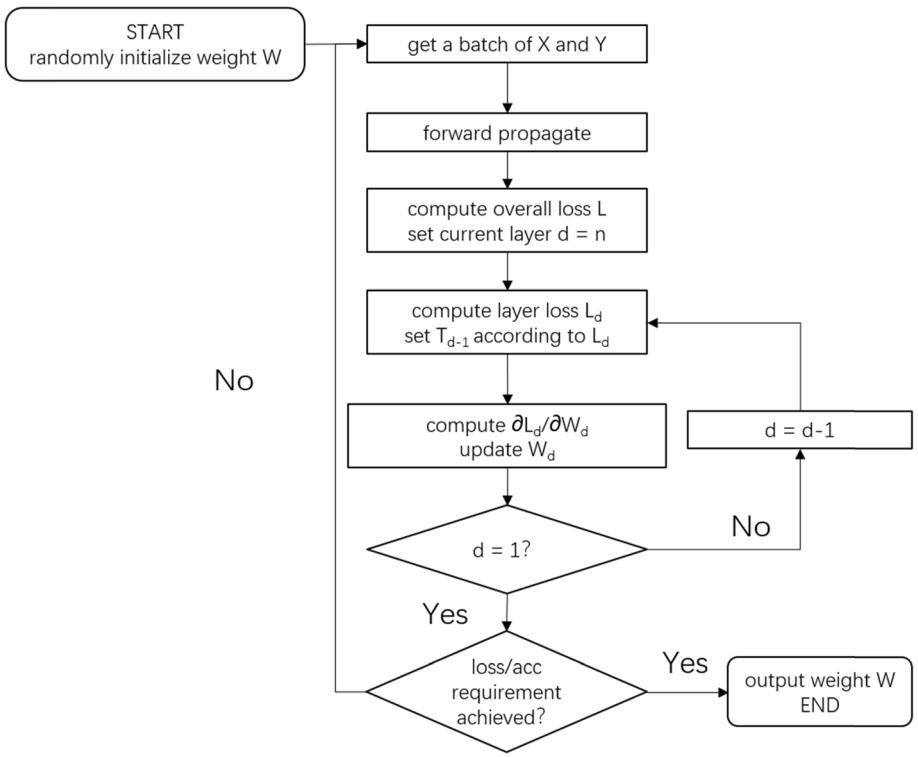


Fig. 5 Block diagram of the target propagation algorithm

standard in the industry. In this paper, the ECFP of the reactants and products is used with a radius of 2 and dimension of 1024 as a supplement to the Edit Vector to construct the hybrid model, where the radius and dimension are determined by convention.

Building the training platform

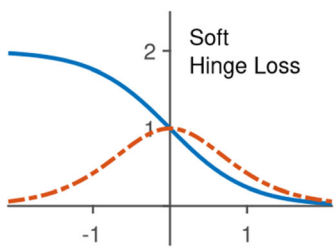
PyTorch is one of the most popular deep learning frameworks, and its high customizability makes it very convenient for implementing all types of "non-standard" backpropagation algorithms. Therefore, in this study, PyTorch and an NVIDIA GeForce GTX 1070 GPU were used to conduct all the experiments.

A dataset containing 15,000 groups of reactants (5 million candidate reactions) was divided by convention: the last 20% (3000 groups with 1 million candidate reactions) was the test set; 12.5% of the remaining 12,000 groups was randomly taken as the validation set (1500 groups with 0.5 million candidate reactions); the final 10,500 groups (3.5 million candidate reactions) were considered as the training set.

The initial hyperparameters for the hard-threshold neural network were as follows: the hidden node structure of the four subnetworks evaluating the Edit Vector was [200/100/50], whereas that of the integrating subnetwork was [50/1]; the activation was Tanh; and the optimizer was AdaDelta ($\rho = 0.95$). Each batch contained 20 groups of reactants, and each model was trained for 85 epochs.

Results and discussion

Through extensive observations of the prediction accuracy, we concluded that the model tended to be stable after 100 epochs, and the Adam optimizer made training



$$soft\_hinge(t_{dj}z_{dj}) = tanh(-t_{dj}z_{dj}) + 1$$

$$L_d = \sum_j [soft\_hinge(z_{dj}, t_{dj}) \cdot \left|\frac{\partial L_{d+1}}{\partial h_{dj}}\right|]$$

Fig. 6 Weighted soft hinge layer loss function

Table 1 Structure of the Edit Vector

Object	Index	Feature
Atom	0	Crippen logP contribution
	1	Crippen MR contribution
	2	TPSA contribution
	3	Labute ASA contribution
	4	Estate index
	5	Gasteiger partial charge
	6	Gasteiger H partial charge
	7–17	atomic number (1-hot)
	18–23	number of neighbors (1-hot)
	24–28	number of hydrogens (1-hot)
	29	formal charge
	30	is in ring
	31	is aromatic
Bond	0	is single bond
	1	is aromatic bond
	2	is double bond
	3	is triple bond

more stable than AdaDelta. Therefore, the following results and discussions are from the experiments performed using the Adam optimizer.

Edit vector-based model

The core of the hard-threshold neural network is activation, so its influence on the model was examined first. The results are shown in Table 2, where "(Soft/Hard)" in the first column indicates the type of activation.

Although naive step activation performed even worse than Tanh activation, the prediction accuracy gradually improved as the order of the hard-threshold activation increased. When the order reached 7, staircase activation achieved a higher prediction accuracy than traditional soft activation. However, the prediction accuracy did not continue to increase when the order was beyond 7. Thus, in the study, the following experiments were performed using 7-staircase activation.

Table 2 Influence of hard-threshold activation on the Edit Vector-based model

Activation Type	Training Accuracy	Validation Accuracy	Test Accuracy
Tanh(Soft)	80.0%	71.1%	70.0%
Step(Hard)	72.7%	71.5%	69.1%
3-Staircase(Hard)	76.4%	69.3%	69.8%
5-Staircase(Hard)	78.0%	68.9%	68.8%
7-Staircase(Hard)	**80.1%**	**70.0%**	**71.2%**
10-Staircase(Hard)	80.0%	69.8%	70.8%

Table 3 Influence of the subnetwork structure on the Edit Vector-based model

Subnetwork structure	Training Accuracy	Validation Accuracy	Test Accuracy
200/100/50	**80.1%**	**70.0%**	**71.2%**
250/100/50	**80.0%**	**72.0%**	**71.3%**
250/125/50	78.8%	71.7%	70.9%
300/100/50	78.7%	69.8%	69.8%
100/100/50/50/50	76.0%	70.4%	69.9%
100/100/50/50/50	74.7%	70.1%	69.5%

The influence of the subnetwork structure is shown in Table 3.

It can be seen that the prediction accuracy could not be significantly improved by deepening or widening the structure of the subnetworks. The subnetwork structure of 200/100/50 was identified to be sufficiently complicated in this task; that is, the accuracy was limited by overfitting rather than underfitting, and more hidden nodes would only disturb training. Therefore, the following experiments focus on how to avoid overfitting using the original subnetwork structure.

Dropout is a common and convenient strategy to avoid overfitting [31]. The original idea of dropout is very simple: in each forward propagation step, some outputs of the hidden nodes are forced to be zero. Then the hidden nodes are prevented from connecting incorrect partners, and overfitting can thus be avoided. Similar to regularization, the principle of dropout is also to reduce the number of non-zero parameters in the model. However, two additional advantages of dropout should be emphasized: first, the meaning of the dropout rate is relatively intuitive, so the proportion of parameters set to zero in the model can be directly adjusted using the dropout rate; and second, different from regularization that penalizes all non-zero parameters, the parameters are set to zero using dropout in a random manner during each forward propagation in training, which improves the robustness of the model. Note that the dropout rate must be set carefully: a dropout rate that is too low cannot avoid overfitting, whereas a dropout rate that is too high will lead to underfitting. The experimental results for the dropout rate are shown in Table 4.

Table 4 Influence of the dropout rate on the Edit Vector-based model

Dropout rate	Training Accuracy	Validation Accuracy	Test Accuracy
0	80.1%	70.0%	71.2%
0.01	77.3%	69.8%	70.1%
0.02	**79.9%**	**72.5%**	**72.7%**
0.1	75.4%	69.7%	70.8%

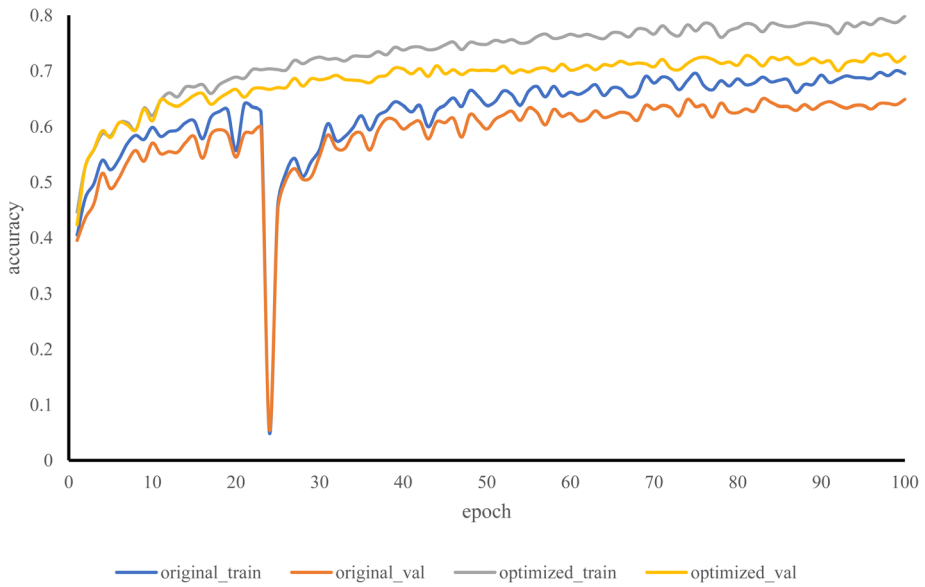


Fig. 7 Training and validation accuracy history of the original model and optimized model

As shown in Table 4, a high dropout rate (e.g., 0.1) damaged the model significantly, whereas a low dropout rate (e.g., 0.01) could not solve the overfitting problem appropriately, and the prediction accuracy did not improve in both cases. A dropout rate of 0.02 achieved a balance between the above two cases; that is, it moderated overfitting while not damaging the model too much. After extensive experiments, the test accuracy reached as high as 72.7%. Compared with the published contribution of 68.5% for the test accuracy [17], more than 120 reactions were correctly predicted additionally.

The training and validation processes of our model, and the model provided in the published literature [17], are shown in Fig. 7. Clearly, the hard-threshold neural network has the potential to approach a higher prediction accuracy while reducing the instability of the running processes.

Regarding training efficiency, the experiments above show that the hard-threshold neural network and corresponding optimization algorithm did not require many extra computing resources during training. It took approximately 13–14 h with 100 epochs, which is similar to traditional neural networks.

Hybrid model

As mentioned in the section "Candidate reaction selection," the mixing factor ε determines the proportion of

Table 5 Effect of the mixing factor ε and deactivation rate on the hybrid model

ε	Dropout rate	Training Accuracy	Validation Accuracy	Test Accuracy
0	0	80.1%	70.0%	71.2%
1.0000	0	99.9%	63.1%	61.6%
0.1000	0	99.7%	65.1%	66.9%
0.0200	0	98.8%	71.0%	70.8%
0.0010	**0**	**85.3%**	**71.9%**	**72.5%**
0.0008	**0**	**85.5%**	**75.3%**	**72.6%**
0.0005	0	83.6%	70.1%	72.3%
0.0010	0	85.3%	71.9%	72.5%
0.0010	0.01	85.7%	73.8%	73.0%
0.0010	**0.02**	**85.4%**	**73.7%**	**73.9%**
0.0010	0.05	83.4%	73.1%	73.1%
0.0010	0.1	82.6%	70.9%	72.7%
0.0010	0.2	77.9%	68.7%	70.3%

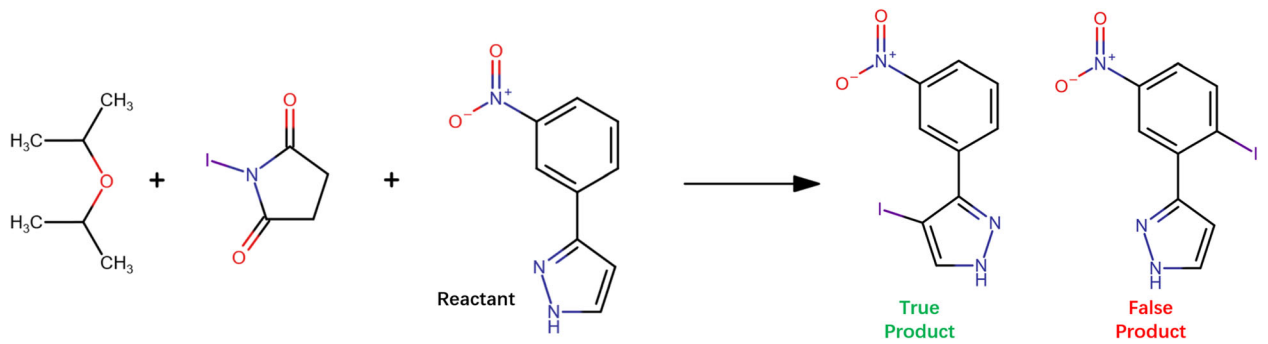


Fig. 8 Iodine reaction in 2-propan-2-yloxypropane between NIS and 3-(3-Nitrophenyl)-1H-pyrazole

the ECFP subnetwork's output in the sum fed to the integrating subnetwork, which directly determines how much the model relies on ECFP. According to the results in the section "Edit Vector-based model," a more complex subnetwork is meaningless for prediction, so only the influence of the mixing factor and dropout rate was examined for the hybrid model, and the results are shown in Table 5.

As shown in Table 5, a large mixing factor ε (e.g., 1 or 0.1), which made the model largely based on ECFP, caused serious overfitting. Additionally, the prediction accuracy decreased rapidly with more serious overfitting when the mixing factor ε was increased. Therefore, it can be rationally inferred that a purely ECFP-based model (ε is very large) would only achieve a lower prediction accuracy than that of the Edit Vector-based model ($\varepsilon = 0$); that is, the information from ECFP alone is not sufficient for good prediction, unlike that from the Edit Vector. However, when the mixing factor was gradually reduced to around 0.001, overfitting almost disappeared, and the prediction accuracy was even higher than that of the Edit Vector-based model ($\varepsilon = 0$), which means that the extra information introduced by the ECFP did result in positive effects on prediction.

Prediction examples

Next, two very important reactions are illustrated that the proposed model successfully predicted but the model in the literature failed to predict.

The reaction in Fig. 8 is taken from the synthesis route of certain substituted 3,4-diarylpyrazole compounds, which modulate the activity of protein kinases [32]. These compounds are very useful in therapy and in the treatment of diseases associated with dysregulated protein kinase activity, such as cancer. For this reaction, the substitution should occur on the pyrazole ring because of the strong electron withdrawing effect of the nitro group. The proposed model assigned a probability of 33.1% to the true product. By contrast, the model in the published literature assigned a probability of 1.7% to the true product and a probability of 31.6% to the wrong product.

The reaction in Fig. 9 is extracted from the synthesis route of novel P2X3 receptor antagonists that play a critical role in treating disease states associated with pain, in particular, peripheral pain, inflammatory pain and tissue injury pain [33]. For this reaction, because the hydrochloric acid-pyridine condition is weakly acidic, the imine hydroxyl group on the product should not dehydrate to form a cyano group. The proposed model

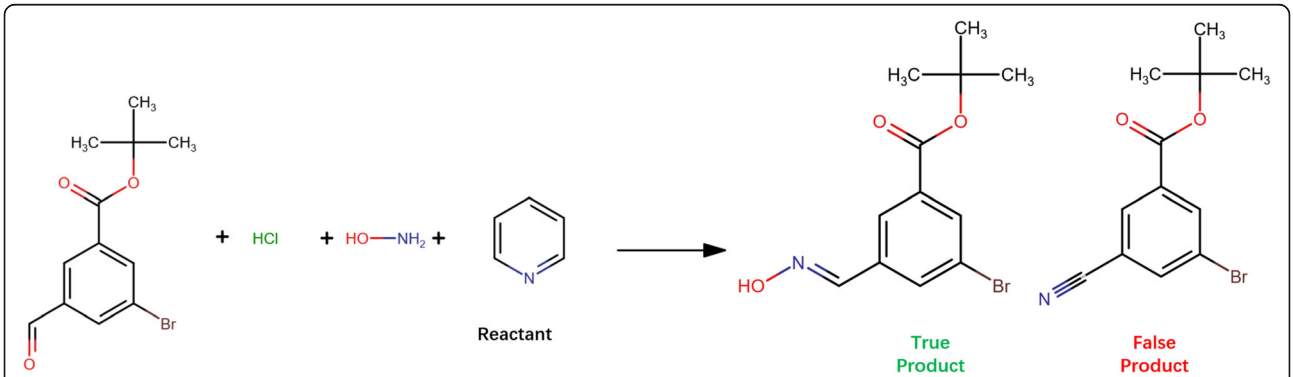


Fig. 9 Amination of hydrazines with aromatic aldehydes under the hydrochloric acid-pyridine condition

assigned a probability of 70.1% to the true product. By contrast, the model in the published literature assigned a probability of 47.1% to the true product and a probability of 48.5% to the wrong product.

Conclusions

In this paper, we implemented a vectorized description of a reaction using the Edit Vector and ECFP, and applied a hard-threshold neural network with the target propagation algorithm to the template-based forward reaction prediction. For the pure Edit Vector-based model, the prediction accuracy reached as high as 72.7%, which is higher than the published accuracy of 68.5%. We also found that the prediction accuracy benefited from the use of ECFP with the proper mixer factor. Although the implemented hard-threshold neural network, whose hyperparameters were adjusted using a heuristic approach, only improved the prediction accuracy by 4.2%, it provides a new alternative approach for computer-aided template-based forward reaction prediction of organic synthesis for drug discovery purposes. An automatic approach for adjusting the hyperparameters to improve the prediction accuracy is under investigation. Furthermore, novel approaches for describing the reaction for prediction purposes are also under consideration.

Abbreviations

SMILES: Simplified Molecular Input Line Entry Specification; SMARTS: SMiles ARbitrary Target Specification; ECFP: Extended-Connectivity FingerPrint

Acknowledgements

We would like to acknowledge the contributions of Si Zhaofeng from University of Chinese Academy of Sciences for introducing PyTorch and Cheng Yu from Department of Chemical Engineering, Tsinghua University for providing organic chemistry support.

Authors' contributions

Dr. ZY conceived and guided the project, and contributed to writing the manuscript. Mr. HH built the model, performed all experiments, and wrote the manuscript. All authors have read and approved the final manuscript.

Funding

The authors gratefully acknowledge financial support from the National Scientific Foundations of China (NSFC, Grant No. 21706143) and the State Key Laboratory of Chemical Engineering (Grant No. SKL-ChE-18 T01).

Availability of data and materials

The reaction datasets used in this article are available in the following repository:
https://figshare.com/articles/MongoDB_dump_compressed_/4833482
The implementation of FTPROP-MB used in this article is available in the following repository:
https://github.com/afriesen/ftprop
Other datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Received: 28 October 2019 Accepted: 18 March 2020
Published online: 08 April 2020

References

1. Scannell JW, Blanckley A, Boldon H, et al. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191–200.
2. Campos KR, Coleman PJ, Alvarez JC, et al. The importance of synthetic chemistry in the pharmaceutical industry. Science. 2019;363(6424):eaat0805.
3. Blakemore DC, Castro L, Churcher I, et al. Organic synthesis provides opportunities to transform drug discovery. Nat Chem. 2018;10(4):383–94.
4. Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97–113.
5. Button AL, Merk D, Hiss JA, et al. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. Nat Mach Intell. 2019;1(7):307–15.
6. Elton DC, Boukouvalas Z, Fuge MD, et al. Deep learning for molecular design—a review of the state of the art. Mol Syst Des Eng. 2019;4(4):828–49.
7. Segler MHS, Preuss M, Waller MP, et al. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018;555(7698):604–10.
8. Ahneman DT, Estrada JG, Lin S, et al. Predicting reaction performance in C-N cross-coupling using machine learning. Science. 2018;360(6385):186–90.
9. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
10. Butler KT, Davies DW, Cartwright HM, et al. Machine learning for molecular and materials science. Nature. 2018;559(7715):547–55.
11. Zhou Z, Li X, Zare RN. Optimizing chemical reactions with deep reinforcement learning. ACS Cent Sci. 2017;3(12):1337–44.
12. Gao H, Struble TJ, Coley CW, et al. Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci. 2018;4(11):1465–76.
13. Coley CW, Thomas DA, Lummiss JAM, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science. 2019;365(6453):eaax1566.
14. Steiner S, Wolf J, Glatzel S, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. Science. 2019;363(6423):eaav2211.
15. Trobe M, Burke MD. The molecular industrial revolution: automated synthesis of small molecules. Angew Chem Int Ed. 2018;57(16):4192–214.
16. Corey EJ, Wipke WT. Computer-assisted Design of Complex Organic Syntheses. Science. 1969;166:178–92.
17. Coley CW, Barzilay R, Jaakkola TS, et al. Prediction of organic reaction outcomes using machine learning. ACS Cent Sci. 2017;3(5):434–43.
18. Segler MHS, Waller MP. Modelling chemical reasoning to predict and invent reactions. Chem Eur J. 2017;23(25):6118–28.
19. Segler MHS, Waller MP. Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chem Eur J. 2017;23(25):5966–71.
20. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435.
21. Corey EJ, Long AK, Rubenstein SD. Computer-assisted analysis in organic synthesis. Science. 1985;228(4698):408–18.
22. Schwaller P, Gaudin T, Lanyi D, et al. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci. 2018;9:6091–8.
23. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50:742–54.
24. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
25. LeCun Y. Learning process in an asymmetric threshold network. In: Disordered systems and biological organization. Berlin, Heidelberg: Springer; 1986. p. 233–40.
26. Friesen AL, Domingos PM. Deep learning as a mixed convex-combinatorial optimization problem. In: International Conference on Learning Representations. Canada: Vancouver; 2018.
27. Wu Y, Liu Y. Robust truncated hinge loss support vector machines. J Am Stat Assoc. 2007;102(479):974–83.
28. Lowe DM. Extraction of chemical structures and reactions from the literature [dissertation]. Cambridge: University of Cambridge; 2012.
29. Law J, Zsoldos Z, Simon A, et al. Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. J Chem Inf Model. 2009;49(3):593–602.
30. Bøgevig A, Federsel HJ, Huerta F, et al. Route design in the 21st century: the IC SYNTH software tool as an idea generator for synthesis prediction. Org Process Res Dev. 2015;19(2):357–68.

31. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
32. Pulici M, Zuccotto F, Badari A, et al. 3,4-diarylpyrazoles as protein kinase inhibitors: WO2010010154A1. 2010-01-28.
33. Burgey CS, Nguyen DN, Paone DV, et al. P2x3 receptor antagonists for treatment of pain: WO2009058299A1. 2009-05-07.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions