Received: 27 November 2020 | Accepted: 30 December 2020

DOI: 10.1002/biot.202000605

Biotechnology Journal

BIOTECH METHOD

A deep learning approach to evaluate the feasibility of enzymatic reactions generated by retrobiosynthesis

Yeji Kim[1,2,3,*] | **Jae Yong Ryu**[4,*] | **Hyun Uk Kim**[2,3,5,*] | **Woo Dae Jang**[1,2,3] | **Sang Yup Lee**[1,2,3]

[1] Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering, KAIST Institute for BioCentury, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

[2] Systems Metabolic Engineering and Systems Healthcare Cross-Generation Collaborative Laboratory, KAIST, Daejeon, Republic of Korea

[3] KAIST Institute for Artificial Intelligence, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea

[4] Data Convergence Drug Research Center, Korea Research Institute of Chemical Technology, Daejeon, Republic of Korea

[5] Systems Biology and Medicine Laboratory, Department of Chemical and Biomolecular Engineering, KAIST, Daejeon, Republic of Korea

Correspondence
Prof. Sang Yup Lee, Department of Chemical and Biomolecular Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea.
Email: leesy@kaist.ac.kr

[*] These authors contributed equally to this work.

Funding information
National Research Foundation of Korea, Grant/Award Numbers: NRF-2012M1A2A2026556, NRF-2012M1A2A2026557, 2018M3A9H3020459, NRF2019M3E5D4065860

Abstract

Retrobiosynthesis allows the designing of novel biosynthetic pathways for the production of chemicals and materials through metabolic engineering, but generates a large number of reactions beyond the experimental feasibility. Thus, an effective method that can reduce a large number of the initially predicted enzymatic reactions has been needed. Here, we present Deep learning-based Reaction Feasibility Checker (DeepRFC) to classify the feasibility of a given enzymatic reaction with high performance and speed. DeepRFC is designed to receive Simplified Molecular-Input Line-Entry System (SMILES) strings of a reactant pair, which is defined as a substrate and a product of a reaction, as an input, and evaluates whether the input reaction is feasible. A deep neural network is selected for DeepRFC as it leads to better classification performance than five other representative machine learning methods examined. For validation, the performance of DeepRFC is compared with another in-house reaction feasibility checker that uses the concept of reaction similarity. Finally, the use of DeepRFC is demonstrated for the retrobiosynthesis-based design of novel one-carbon assimilation pathways. DeepRFC will allow retrobiosynthesis to be more practical for metabolic engineering applications by efficiently screening a large number of retrobiosynthesis-derived enzymatic reactions. DeepRFC is freely available at https://bitbucket.org/kaistsystemsbiology/deeprfc.

KEYWORDS
deep learning, DeepRFC, enzymatic reaction, reaction feasibility, retrobiosynthesis

1 | INTRODUCTION

An increasing number of chemicals and materials have become microbially producible by metabolic engineering, which often involves

Abbreviations: DeepRFC, deep learning-based reaction feasibility checker; ECFP4, extended-connectivity fingerprints with a diameter of 4; MC, Monte Carlo; MCC, Matthews correlation coefficient; RDT, reaction decoder tool; SMILES, Simplified Molecular-Input Line-Entry System; VAE, variational autoencoder

the construction of novel biosynthetic pathways.[1] Among several useful computational tools available, retrobiosynthesis, in particular, can greatly facilitate this pathway construction process, which uses reaction rules to predict possible products that may be generated from a given substrate.[2,3] The reaction rules describe the conversion of a substrate to a product at an atomic level, and have been established based on a large number of already characterized enzymatic reactions that are often available at metabolic databases, such as KEGG,[4] MetaCyc,[5] and BRENDA.[6] Representative retrobiosynthesis tools developed so far include our in-house retrobiosynthesis tool,[7] METEOR,[8] PathPred,[9] GEM-Path,[10] BNICE,[11] novoStoic,[12] and RetroPath.[13] Some of these tools have been applied to metabolic engineering experiments, such as assimilation of formaldehyde,[14] production of 1,4-butanediol,[15] pinocembrin,[16] and 5-aminolevulinic acid.[17] One of the major problems in metabolic pathway design by retrobiosynthesis is that a large number of candidate reactions are suggested by these algorithms. To allow a more feasible application of retrobiosynthesis for metabolic engineering, more effective methods need to be developed that can substantially reduce a large number of retrobiosynthesis-derived enzymatic reactions to a set of reactions small enough for actual experimental validation.[2,3] Here, we suggest the evaluation of the reaction feasibility as a strategy to select candidate reactions for narrowing down the candidate reactions. The reaction feasibility herein is defined as the possibility of the biochemical conversion of a given substrate to each product predicted.

Several computational methods have been developed that help evaluating the feasibility of enzymatic reactions predicted by the retrobiosynthesis. A support vector machine model was previously developed that predicts the reactivity of small molecules involved in an enzymatic reaction;[18] it requires a set of atomic (e.g., electrostatic and topological properties) and molecular properties (e.g., shape, surface, energy, and charge distribution of a molecule) as an input, but preparation of the corresponding training dataset involved manual curation. More recently, deep learning is beginning to be applied to predict the reaction feasibility of non-enzymatic chemical reactions.[19–21] For example, the feasibility of copper(I)-catalyzed alkyne-azide cycloaddition reaction was predicted by a recurrent neural network with an accuracy of 80%.[19] Thus, it was reasoned that deep learning can be employed for efficiently predicting biochemical reaction feasibility as well.

In this study, we developed Deep learning-based Reaction Feasibility Checker (DeepRFC) that evaluates the feasibility of retrobiosynthesis-derived enzymatic reactions with high performance and speed. To develop the deep neural network (DNN), reactant pairs that correspond to the substrate-product pairs defined at KEGG[4] were used as a training dataset. Each of the enzymatic reactions predicted by retrobiosynthesis usually consists of a single substrate and a single product, and thus, the use of reactant pairs was considered ideal for evaluating the reaction feasibility. The performance of DeepRFC using a DNN was first evaluated in comparison with the versions using other representative machine learning methods in place of DNN. DeepRFC was compared with another in-house reaction feasibility checker that uses the concept of a reaction similarity instead of DNN.[22] These evaluation studies suggested that DeepRFC is expected to be a useful tool for effectively narrowing down a large number of retrobiosynthesis-derived enzymatic reaction candidates to several reactions that can be experimentally validated.

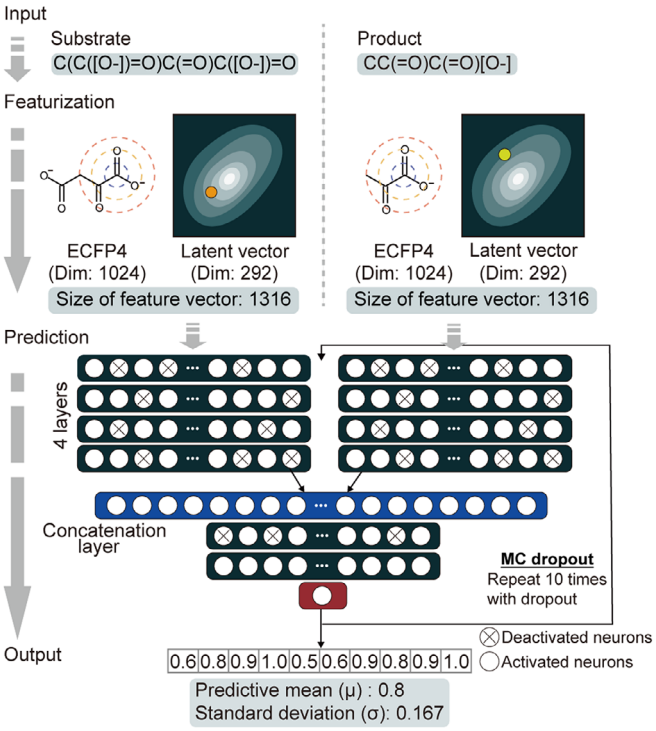

FIGURE 1 The overall scheme of DeepRFC. SMILES strings of a substrate and a product from a reactant pair serve as an input. Two feature vectors, each having 1316 dimensions, are generated from the substrate and the product by implementing ECFP4 and molecular VAE. The feasibility of the input reactant pair is classified by the DeepRFC model as an output. It should be noted that the output is determined based on the predictive mean, which, in turn, is obtained from the ten predictive values of the DeepRFC via Monte Carlo (MC) dropout. The reactant pair is considered feasible if the predictive mean is greater than the threshold of 0.32, and otherwise, it is considered infeasible

2 | RESULTS AND DISCUSSION

2.1 | Development of DeepRFC

The DeepRFC model (i.e., DNN) was constructed using "positive" (feasible) and "negative" (infeasible) datasets, each containing 4626 reactant pairs. DeepRFC receives two Simplified Molecular-Input Line-Entry System (SMILES) strings as inputs, each from a substrate and a product of a reactant pair of interest, and classifies whether the input reaction (provided as a reactant pair) is feasible (Figure 1). The DNN of DeepRFC was optimized by fine-tuning hyperparameters and featurization methods. First, the following three hyperparameters of DeepRFC were optimized with respect to the accuracy via Bayesian

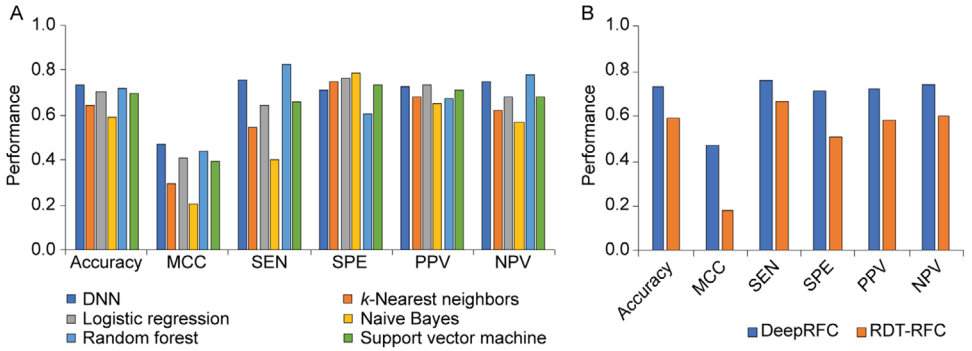

FIGURE 2 Performances of DeepRFC. (A) Performances of DeepRFC when using different machine learning methods, including deep neural network (DNN), *k*-nearest neighbors, logistic regression, Naive Bayes, random forest, and support vector machine. (B) Performances of DeepRFC and RDT-RFC. Abbreviated evaluation metrics are MCC, Matthews correlation coefficient; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; and NPV, negative predictive value

optimization:[23] learning rate, number of hidden layers for the initial two sets of the fully-connected layers, and number of nodes per hidden layer (Table S1). Five-fold cross-validation was conducted for hyperparameter optimization. By examining 24 corresponding architectures of the model, an optimal model appeared to have four hidden layers for the two sets of the fully-connected layers with 512 nodes per hidden layer, and a learning rate of 0.001 (Figure S1; Tables S1 and S2). Additionally, a dropout rate of 0.5 was used to avoid overfitting of the model.

Second, three different approaches for the featurization of the SMILES strings of a substrate and a product were applied to the DNN to find the best-performing featurization method, including extended-connectivity fingerprints with a diameter of 4 (ECFP4),[24] molecular variational autoencoder (VAE),[25] and a combination of ECFP4 and molecular VAE. As a result, the combination of ECFP4 with molecular VAE showed the highest performance in terms of accuracy, Matthews correlation coefficient (MCC), specificity, and positive predictive value (Table S3). Thus, the feature vectors of a substrate and a product were both generated using the combination of ECFP4 with molecular VAE in the optimized DNN of DeepRFC (Figure 1).

DeepRFC initially uses two distinct sets of four fully-connected layers, which independently process the feature vectors of a substrate and a product (Figure 1; Figure S1). The two sets of the fully-connected layers are concatenated into a single hidden layer, which is followed by two additional hidden layers and an output layer that generates a predictive value. The output neuron's predictive value ranges between 0 and 1, which is used to classify the feasibility of an input reactant pair; the threshold of 0.32 was used to determine the feasibility of the input reactant pair (Experimental Section). It should be noted that DeepRFC is implemented ten times for an input reactant pair via Monte Carlo (MC) dropout,[26] which allows measuring the uncertainty (via providing standard deviation) of the predictive values. Thus, the final binary classification of the reaction feasibility is performed based on the predictive mean of the ten predictive values (Figure 1); the output of DeepRFC additionally presents the predictive mean and standard deviations of the ten predictive values.

2.2 | Performance and key features of DeepRFC

2.2.1 | Comparison of DeepRFC with other representative machine learning methods

The performance of using the DNN in DeepRFC was compared with those DeepRFC versions that use other representative machine learning methods. The machine learning methods considered in place of DNN included *k*-nearest neighbors, logistic regression, Naive Bayes, random forest, and support vector machine (Figure 2A; Table S4). Hyperparameters for each of the machine learning models were optimized by grid search. The threshold for the binary classification of reaction feasibility was also optimized for these machine learning methods in the same manner as the DeepRFC with DNN (Table S4). As a result, DeepRFC with DNN showed the highest accuracy and MCC values, 0.73 and 0.47, respectively, which are the two most important evaluation metrics among those considered in this study (Figure 2A; Table S4). The reason why DeepRFC using the DNN did not show the best performance for the other four evaluation metrics (Figure 2A; Table S4) is because the accuracy was chosen as the key evaluation metric when optimizing the hyperparameters (Table S2) and the classification threshold (0.32) of the DeepRFC.

2.2.2 | Comparison of DeepRFC with RDT-RFC

The performance of DeepRFC was also compared with another in-house method, RDT-RFC, that examines the reaction feasibility by using the concept of reaction similarity (Figure 2B; Table S5). To the best of our knowledge, a computational method that evaluates the feasibility of an enzymatic reaction using up-to-date coverage of reactant pairs has not been developed. Thus, the reaction decoder tool (RDT)[22] was deployed together with additional features, developed in this study, to infer the feasibility of a reaction. RDT itself generates reaction centers in the form of SMILES by examining the structures of substrates and products for a given enzymatic reaction.[22] The

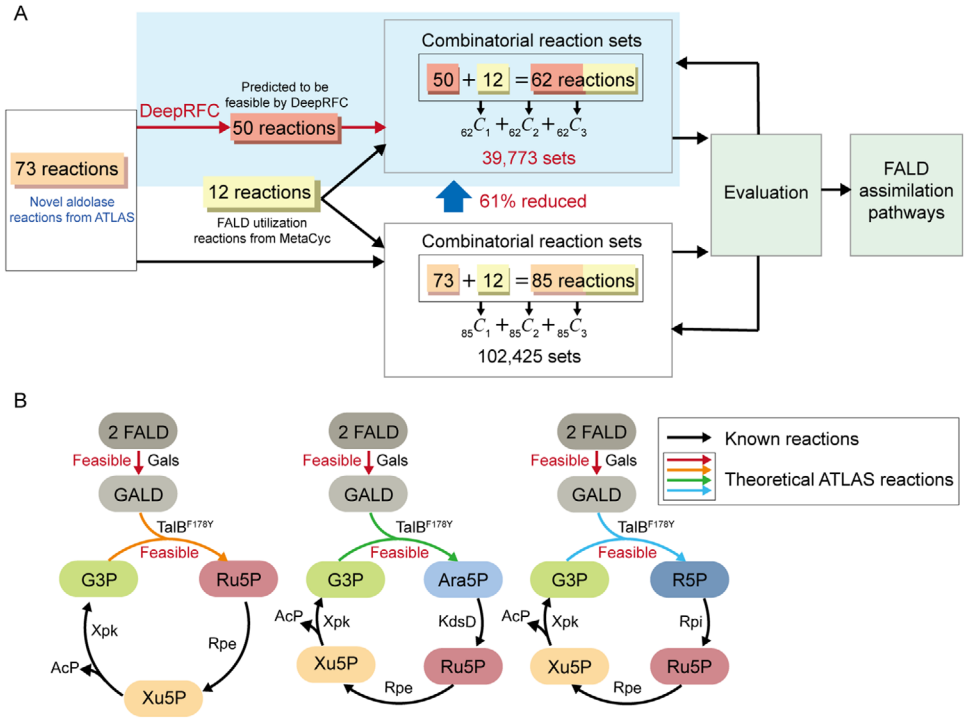

FIGURE 3 Application of DeepRFC for reducing the number of retrobiosynthesis-derived enzymatic reaction candidates. (A) A workflow for designing novel one-carbon (C_1) assimilation pathways that convert formaldehyde (FALD) to acetate as a target product without the requirements of ATP and NAD(P)H and the loss of CO_2.[14] In this study by Yang et al,[14] novel 73 retrobiosynthesis-derived aldolase reactions from ATLAS were considered in addition to the naturally occurring 12 FALD utilization reactions from MetaCyc to generate novel C_1 assimilation pathways. Aldolases are responsible for the first step of the C_1 assimilation pathways. With a total of 85 FALD utilization reactions from both MetaCyc and ATLAS, 102,425 reaction sets, each consisting of one, two, or three reactions, were generated for the evaluation. However, with the use of DeepRFC, the number of ATLAS reactions was reduced from 73 to 50, and consequently, the total number of the resulting reaction sets was substantially reduced to 39,773 (light blue box). (B) Three predicted FALD assimilation pathways that were experimentally validated by Yang et al.[14] The reactions catalyzed by Gals and TalB^{F178Y} were available in the 50 aldolase reactions from ATLAS, which were all predicted to be feasible by DeepRFC. Metabolite abbreviations are: AcP, acetyl-phosphate; Ara5P, arabinose 5-phosphate; G3P, glyceraldehyde 3-phosphate; GALD, glycolaldehyde; R5P, ribose 5-phosphate; Ru5P, ribulose 5-phosphate; and Xu5P, xylulose 5-phosphate. Enzyme abbreviations are: Gals, a mutant GALD synthase; KdsD, arabinose 5-phosphate isomerase; Rpe, ribulose-phosphate 3-epimerase; Rpi, ribose 5-phosphate isomerase; TalB^{F178Y}, engineered transaldolase; and Xpk, 5-phosphate phosphoketolase

resulting RDT-generated reaction centers of an input reaction and reactant pairs from the positive dataset were compared to determine the feasibility of the input reaction (Experimental Section). As a result, DeepRFC outperformed RDT-RFC for all the six evaluation metrics (Figure 2B; Table S5). In particular, DeepRFC showed 2.5-fold higher MCC (0.47) and 1.2-fold higher accuracy (0.73) than RDT-RFC (Figure 2B; Table S5). Besides the better classification performance, DeepRFC can be run using GPU, which allows substantially faster computation (Figure S2). Obviously, the fast computation of DeepRFC will be useful for handling a large number of candidate reactions generated from retrobiosynthesis.

2.2.3 | Application of DeepRFC for the retrobiosynthesis-based pathway design

As a demonstration for the pathway design process, DeepRFC was applied to 73 retrobiosynthesis-derived aldolase reactions from ATLAS,[11] which were previously used to predict novel one-carbon (C_1) assimilation pathways that would convert formaldehyde (FALD) to acetate as a target product (Figure 3A).[14] FALD is a key starting metabolite in the C_1 assimilation pathways that can be converted to various useful chemicals, and aldolases are responsible for the first step of the C_1 assimilation pathways. For a greater carbon yield, it is necessary to design novel FALD assimilation pathways that do not require ATP and NAD(P)H as well as those that do not generate CO_2. In the study by Yang et al,[14] novel 73 retrobiosynthesis-derived aldolase reactions from ATLAS were considered in addition to the naturally occurring 12 FALD utilization reactions from MetaCyc to generate novel C_1 assimilation pathways. With a total of 85 FALD utilization reactions from both MetaCyc and ATLAS, combinatorial reaction sets were generated, each having one, two, or three reactions, which gave rise to 102,425 sets that needed to be evaluated. However, with the use of DeepRFC, the 73 ATLAS reactions were reduced to 50 reactions, which consequently led to a substantially smaller number of sets, 39,773, for the evaluation (Figure 3A). Importantly, these 39,773 sets included the three predicted FALD assimilation pathways that were experimentally validated in the Yang et al.[14] (Figure 3B). This

example demonstrates that DeepRFC can be useful in narrowing down the retrobiosynthesis-derived enzymatic reaction candidates in addition to the established criteria, for example, thermodynamic analysis and the number of pathway steps.

3 | CONCLUSIONS

In this study, a deep learning-based DeepRFC was developed that classifies the feasibility of enzymatic reactions to narrow down a large number of retrosynthesis-derived candidate enzymatic reactions. The use of DNN, in comparison with the five other representative machine learning methods, was found to be the best for solving this binary classification problem by DeepRFC. DeepRFC outperformed another reaction feasibility checker that uses RDT. Finally, DeepRFC successfully reduced the number of retrobiosynthesis-derived enzymatic reaction candidates for novel FALD assimilation pathways, which were consistent with the previous experimental results.[14] With its high performance and speed, the DeepRFC is expected to facilitate the metabolic pathway design by screening and reducing a large number of enzymatic reaction candidates predicted by retrobiosynthesis algorithms. Further improvement of the DeepRFC performance is expected if more experimentally obtained positive (and also negative) datasets become available.

4 | EXPERIMENTAL SECTION

4.1 | Data preparation

Information on 4626 reactant pairs[4] in the positive dataset was obtained from the manually curated KEGG database (as of January 2020); the initial number of the unique reactant pairs retrieved was 5595. Reactant pairs were manually generated by retrieving substrates and products from already known enzymatic reactions, which share common substructures according to the previous study.[4] Structures of substrates and products were described using SMILES. Compounds with molecular weights of over 800 Da were removed from the dataset due to the computation time. Also, compounds described using more than 120 characters in the SMILES were removed for faster computations, which was previously reported to still achieve the reasonable model accuracy;[25] more than 90% of the entire compounds considered in this study were described using fewer than 100 characters in the SMILES. These 4626 reactant pairs constitute a "positive" dataset for training the DeepRFC model.

The negative dataset was constructed by generating artificial reactant pairs due to the lack of information on reactant pairs for infeasible reactions. To construct a negative dataset, artificial reactant pairs were newly prepared in this study by using reaction rules from RetroRules[27] and a Python package RDKit (ver. 2019.09.3.0) (https://www.rdkit.org/). An assumption for this approach is that a large number of reactions generated by using reaction rules would be false-positive predictions, hence infeasible. The reactions predicted using RetroRules and RDKit had 19.3 products per substrate on average, which contrasts with already known enzymatic reactions that usually have a very low number of products per substrate; for example, the 4626 reactant pairs in the positive dataset involve 2113 unique substrates, which gives 2.19 products per substrate on average. Among the 235,313 reaction rules available from the RetroRules, 145,544 reaction rules derived from the reactions with the fourth-level EC numbers were selected for generating a negative dataset. Using the RDKit, 36,098 hypothetical enzymatic reactions were first generated by using the 145,544 reaction rules and a unique set of 2113 substrates from the positive dataset. Among the newly generated reactions, 4626 reactions, the same number of reactions in the positive dataset, were randomly selected for the negative dataset if a corresponding product of each reaction was found in the PubChem database.[28]

For the DeepRFC development, 80% of the entire reactant pairs from both positive and negative datasets were used as a training dataset and 20% were used as a test dataset. The test dataset was used for evaluating the performance of DeepRFC versions that used DNN or the five representative machine learning methods.

4.2 | Featurization of the input data

The following three different featurization methods were examined for the optimal featurization of the SMILES strings of an input reactant pair: ECFP4,[24] molecular VAE,[25] and a combination of ECFP4 and molecular VAE (Table S1). Among the three methods, the combination of ECFP4 and molecular VAE was selected for the featurization of the input data. Molecular VAE is primarily aimed at generating a new molecule with desired properties via an encoder and a decoder, but only the encoder of molecular VAE was used in this study to convert SMILES into a feature vector that represents the unique characteristics of an input molecule. The resulting feature vectors from ECFP4 and molecular VAE have 1024 and 292 dimensions, respectively (Figure 1; Figure S1).

4.3 | Evaluation of DeepRFC

Ten predictive values were first obtained from the 10 MC dropout samples of the DeepRFC, and their mean was used to classify the feasibility of a given reactant pair. The threshold (predictive mean – standard deviation*0.5) of the mean predictive value was set to be 0.32 for the binary classification. This value was determined because it led to the best model accuracy among those from the range between 0 to 1 with an interval of 0.01. Performance of the DeepRFC was evaluated based on six metrics, including accuracy, MCC, sensitivity, specificity, positive predictive value, and negative predictive value (Table S4).

4.4 | Evaluation of RDT-RFC

To infer the feasibility of an input reaction using RDT-RFC, the RDT-generated reaction centers from an input reaction and each reactant

pair from the training subset of the positive dataset were compared. For these structural comparisons, the Tanimoto coefficient was used. From all the structural comparisons, if the best Tanimoto coefficient was greater than 0.3, the input reaction was considered to be feasible. This threshold of Tanimoto coefficient was determined by examining the accuracy of the RDT-RFC by changing the Tanimoto coefficient from 0 to 1 with an interval of 0.01; Tanimoto coefficient of 0.3 gave the best accuracy of the RDT-RFC. It should be noted that, with a Tanimoto coefficient threshold as well as the test subsets of the positive and negative datasets, it was possible to calculate the number of true positives, false-positives, true negatives, and false-negatives by evaluating the predicted feasibility of the input reaction, which consequently allowed the calculation of the prediction accuracy of the RDT-RFC.

4.5 | Development environment

DNN of DeepRFC (Figure 1) was constructed and trained to classify the reaction feasibility by using a Python package *Keras* (version 2.3.0) (https://keras.io/) with *TensorFlow* backend (version 1.13.1).[29] Specifications considered for developing DeepRFC are described in Figure S1. Source code for the featurization of input data, two SMILES strings of a reactant pair, are available at https://bitbucket.org/kaistsystemsbiology/deeprfc. A Python package *scikit-learn* (0.22.1)[30] was used to evaluate the performances of DeepRFC and RDT-RFC.

ACKNOWLEDGMENTS

Y.K., J.Y.R., and H.U.K. contributed equally to this work. This work was supported by the Technology Development Program to Solve Climate Changes on Systems Metabolic Engineering for Biorefineries (NRF-2012M1A2A2026556 and NRF-2012M1A2A2026557) from the Ministry of Science and ICT (MSIT), through the National Research Foundation (NRF) of Korea. This work was also supported by the NRF grants funded by the Korean government (2018M3A9H3020459 and NRF2019M3E5D4065860).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Yeji Kim: Conceptualization; Investigation; Software; Writing-original draft; Writing-review & editing. Jae Yong Ryu: Conceptualization; Investigation; Software; Writing-original draft; Writing-review & editing. Hyun Uk Kim: Conceptualization; Investigation; Supervision; Writing-original draft; Writing-review & editing. Woo Dae Jang: Investigation; Validation; Writing-original draft; Writing-review & editing. Sang Yup Lee: Conceptualization; Funding acquisition; Project administration; Supervision; Writing-original draft; Writing-review & editing.

DATA AVAILABILITY STATEMENT

DeepRFC is freely available at https://bitbucket.org/kaistsystemsbiology/deeprfc. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Yeji Kim https://orcid.org/0000-0001-9801-917X
Jae Yong Ryu https://orcid.org/0000-0003-0603-1599
Hyun Uk Kim https://orcid.org/0000-0001-7224-642X
Woo Dae Jang https://orcid.org/0000-0003-1649-0174
Sang Yup Lee https://orcid.org/0000-0003-0599-3091

REFERENCES

1. Ko, Y. S., Kim, J. W., Lee, J. A., Han, T., Kim, G. B., Park, J. E., & Lee, S. Y. (2020). Tools and strategies of systems metabolic engineering for the development of microbial cell factories for chemical production. *Chem. Soc. Rev.*, *49*, 4615.
2. Hadadi, N., & Hatzimanikatis, V. (2015). Design of computational retrobiosynthesis tools for the design of *de novo* synthetic pathways. *Curr. Opin. Chem. Biol.*, *28*, 99.
3. Wang, L., Dash, S., Ng, C. Y., & Maranas, C. D. (2017). A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.*, *2*, 243.
4. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, *45*, D353.
5. Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., … Karp, P. D. (2020). The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.*, *48*, D445.
6. Jeske, L., Placzek, S., Schomburg, I., Chang, A., & Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.*, *47*, D542.
7. Cho, A., Yun, H., Park, J. H., Lee, S. Y., & Park, S. (2010). Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol.*, *4*, 35.
8. Marchant, C. A., Briggs, K. A., & Long, A. (2008). In silico tools for sharing data and knowledge on toxicity and Metabolism: Derek for windows, meteor, and vitic. *Toxicol. Mech. Methods*, *18*, 177.
9. Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., & Kanehisa, M. (2010). PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, *38*, W138.
10. Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O., & Feist, A. M. (2014). Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path. *Metab. Eng.*, *25*, 140.
11. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., & Hatzimanikatis, V. (2016). ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. *ACS Synth. Biol.*, *5*, 1155.
12. Kumar, A., Wang, L., Ng, C. Y., & Maranas, C. D. (2018). Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.*, *9*, 184.
13. Delepine, B., Duigou, T., Carbonell, P., & Faulon, J. L. (2018). RetroPath2.0: a retrosynthesis workflow for metabolic engineers. *Metab. Eng.*, *45*, 158.
14. Yang, X., Yuan, Q., Luo, H., Li, F., Mao, Y., Zhao, X., & Ma, Y. (2019). Systematic design and in vitro validation of novel one-carbon assimilation pathways. *Metab. Eng.*, *56*, 142.
15. Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., & Van Dien, S. (2011). Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat. Chem. Biol.*, *7*, 445.
16. Feher, T., Planson, A. G., Carbonell, P., Fernandez-Castane, A., Grigoras, I., Dariy, E., … Faulon, J. L. (2014). Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering. *Biotechnol. J.*, *9*, 1446.
17. Ren, J., Zhou, L., Wang, C., Lin, C., Li, Z., & Zeng, A. P. (2018). An unnatural pathway for efficient 5-aminolevulinic acid biosynthesis

with glycine from glyoxylate based on retrobiosynthetic design. *ACS Synth. Biol.*, *7*, 2750.
18. Mu, F., Unkefer, C. J., Unkefer, P. J., & Hlavacek, W. S. (2011). Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics*, *27*, 1537.
19. Su, S., Yang, Y., Gan, H., Zheng, S., Gu, F., Zhao, C., & Xu, J. (2020). Predicting the feasibility of copper(I)-catalyzed alkyne–azide cycloaddition reactions using a recurrent neural network with a self-attention mechanism. *J. Chem. Inf. Model.*, *60*, 1165.
20. Fooshee, D., Mood, A., Gutman, E., Tavakoli, M., Urban, G., Liu, F., … Baldi, P. (2018). Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.*, *3*, 442.
21. Baylon, J. L., Cilfone, N. A., Gulcher, J. R., & Chittenden, T. W. (2019). Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *J. Chem. Inf. Model.*, *59*, 673.
22. Rahman, S. A., Torrance, G., Baldacci, L., Martínez Cuesta, S., Fenninger, F., Gopal, N., … Steinbeck, C. (2016). Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics*, *32*, 2065.
23. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.*, *2*, 2951.
24. Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, *50*, 742.
25. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., … Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.*, *4*, 268.
26. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Int. Conf. Mach. Learn.*, *48*, 1050.
27. Duigou, T., du Lac, M., Carbonell, P., & Faulon, J. L. (2019). RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Res.*, *47*, D1229.
28. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., & Shoemaker, B. A. (2016). PubChem substance and compound databases. *Nucleic Acids Res.*, *44*, D1202.
29. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467v2.
30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Dubourg, V. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, *12*, 2825.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kim Y, Ryu JY, Kim HU, Jang WD, Lee SY. A deep learning approach to evaluate the feasibility of enzymatic reactions generated by retrobiosynthesis. *Biotechnol J*. 2021;*16*:e2000605. https://doi.org/10.1002/biot.202000605