JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau
Article

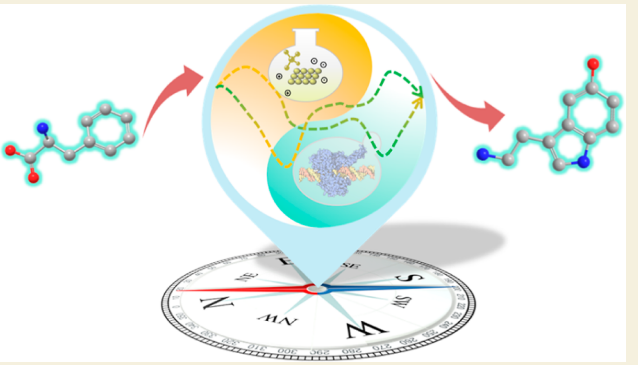
Developing BioNavi for Hybrid Retrosynthesis Planning

Tao Zeng, Zhehao Jin, Shuangjia Zheng, Tao Yu, and Ruibo Wu*

Cite This: JACS Au 2024, 4, 2492−2502
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Illuminating synthetic pathways is essential for producing valuable chemicals, such as bioactive molecules. Chemical and biological syntheses are crucial, and their integration often leads to more efficient and sustainable pathways. Despite the rapid development of retrosynthesis models, few of them consider both chemical and biological syntheses, hindering the pathway design for high-value chemicals. Here, we propose BioNavi by innovating multitask learning and reaction templates into the deep learning-driven model to design hybrid synthesis pathways in a more interpretable manner. BioNavi outperforms existing approaches on different data sets, achieving a 75% hit rate in replicating reported biosynthetic pathways and displaying superior ability in designing hybrid synthesis pathways. Additional case studies further illustrate the potential application of BioNavi in a de novo pathway design. The enhanced web server (http://biopathnavi.qmclab.com/bionavi/) simplifies input operations and implements step-by-step exploration according to user experience. We show that BioNavi is a handy navigator for designing synthetic pathways for various chemicals.



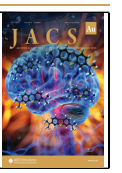
KEYWORDS: *retrosynthesis, hybrid synthesis, deep learning, chemo-enzymatic synthesis, reaction pathway*

INTRODUCTION

Producing high-value-added chemicals from biobased building blocks such as CO_2, fermentable sugars, and primary metabolites from microorganisms has caught much attention due to increasing concerns about resource shortages and climate change.[1] Inspired by nature, biological synthesis, encompassing biosynthesis and biocatalysis, utilizes enzymes to catalyze reactions for the production of complex natural products or their analogs,[2,3] such as catharanthine[4] and jasmonates[5] (Figure 1A). Although enzymatic reactions can be efficient and environmentally friendly,[6] they cannot be responsible for the industrial production of all chemicals due to the limited enzymes or technical issues.[1] Another complementary approach is chemical synthesis, where enzymes can be replaced with inorganic catalysts or under extreme conditions such as high temperature and pressure. Chemical synthesis also expands the chemical space of molecules in more flexible ways, while the control of regio- and stereoselectivity for complex structures is still challenging.[7] Merging biological and chemical synthesis, such as semisynthesis, can provide facile access to complex structures, especially natural products.[8,9] Early examples can be traced back to the synthesis of D-mannitol with combined enzyme and metal catalysts,[10] as well as later examples such as the hybrid organic-biocatalytic synthesis of highly oxidized diterpenes[11] and the hybrid synthesis of non-natural antiviral agents.[12] Thus, it is promising to consider both biological and chemical synthesis when designing synthetic routes for high-value-added chemicals.[13]

Currently, retrosynthesis planning tools that predict pathways for chemicals are developing rapidly, especially those based on machine/deep learning methods.[14−16] In general, as shown in Figure 1B, retrosynthesis planning consists of a single-step prediction model and a multistep search engine.[17] For a given target molecule, potential precursors are generated by a single-step model and then fed back into the model to produce precursors for the next step. This procedure is repeated iteratively until a termination condition is triggered (e.g., when the precursors are readily available or when the specified number of iterations is reached). Since multiple precursors are generated in one step, a scoring method is usually used to rank the precursor candidates to obtain reliable paths in the shortest time. To efficiently explore the search space in multistep pathway prediction, many searching techniques, such as Monte Carlo Tree Search[18] and Retro*,[19] have been used to find the optimal precursors with minimal time costs for each iteration step. Most single-step models are developed based on various reaction databases, for example, USPTO[20] and BRENDA,[21] which can be divided into template-based[18,22−24] and template-free[25−29] approaches.

Received: March 11, 2024
Revised: June 18, 2024
Accepted: June 20, 2024
Published: July 3, 2024



https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

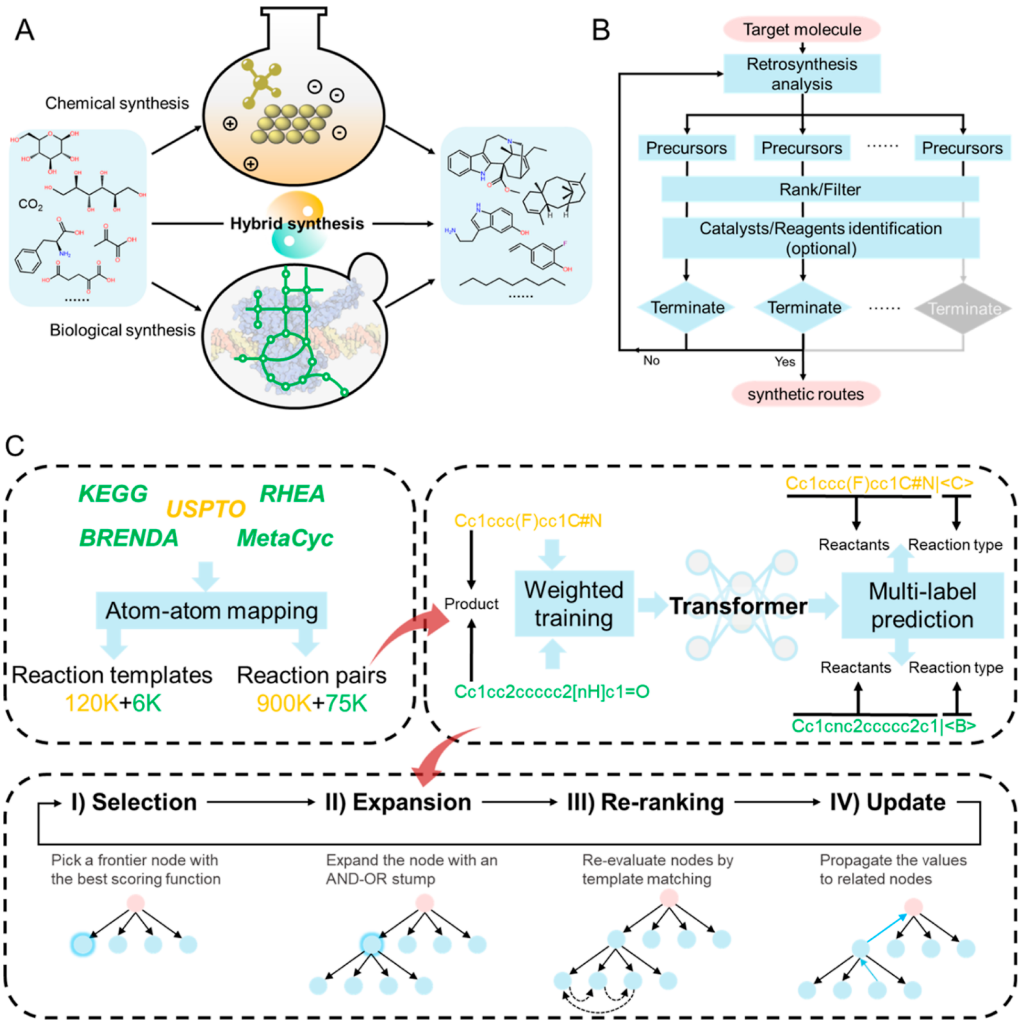


Figure 1. (A) Production of chemicals from simple building blocks by chemical synthesis and biosynthesis. (B) General workflow for the retrosynthesis planning. (C) Overview of data collection and model construction of BioNavi; yellow and green colors indicate the processes of chemical synthesis and biological synthesis, respectively.

Template-based approaches use reaction templates extracted from databases to make predictions. These predictions are generated by applying optimal templates, which are determined through methods such as structure similarity[23] or deep learning-based scoring[18] to the target compounds. While template-free approaches predict precursors by utilizing a deep learning model trained directly from the reaction pairs.[30] Recently, inspired by the concept of "synthon", semitemplate-based were proposed,[31,32] where the target molecule is first broken into "synthon" and then the "synthon" is completed in the reactant.

However, most approaches leverage only either chemical or biological reactions for pathway planning, which limits their ability to design hybrid synthetic pathways. Recently, Levin et al.[33] merged enzymatic and nonenzymatic reaction templates with computational synthetic planning (integrated into the ASKCOS platform), identifying more efficient and shorter routes for the production of dronabinol and arformoterol. Sankaranarayanan and Jensen[34] planned chemoenzymatic pathways by identifying the enzymatic steps from the chemical synthetic pathways suggested by ASKCOS. Nevertheless, predefined reaction templates cannot capture the reaction patterns beyond the database. More importantly, the reaction templates rely on either manual extraction by human experts (which is time-consuming) or automated generation by tools like RDChiral[35] (which faces challenges in balancing specificity and generality). Deep learning-based language models trained from reaction pairs can output reactants from an input product in an end-to-end manner without templates. In our previous work,[26] the constructed deep learning-based bioretrosynthesis tool, BioNavi-NP, outperformed the template-based methods in the biosynthetic pathway prediction of natural products, despite the existence of missing fragments and unreasonable reactions.

Herein, we promote BioNavi-NP to address the above issues and extend its application to the biobased hybrid synthesis prediction for high-value-added chemicals (Figure 1C). All chemical and biological reactions used in the model training are collected from public databases. The atom−atom mapping[36] strategy is introduced to extract the principal components from the original reactions, by which the cofactors can be removed and all necessary reactants are kept for a specific product (Methods). Then, a multitask learning strategy (weighted training) is conducted to balance the accuracy of chemical and biological reaction prediction. To automatically indicate the reaction type during pathway search, an additional label (representing a chemical or biological reaction) is output along with the reactants (multilabel prediction). Furthermore,

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

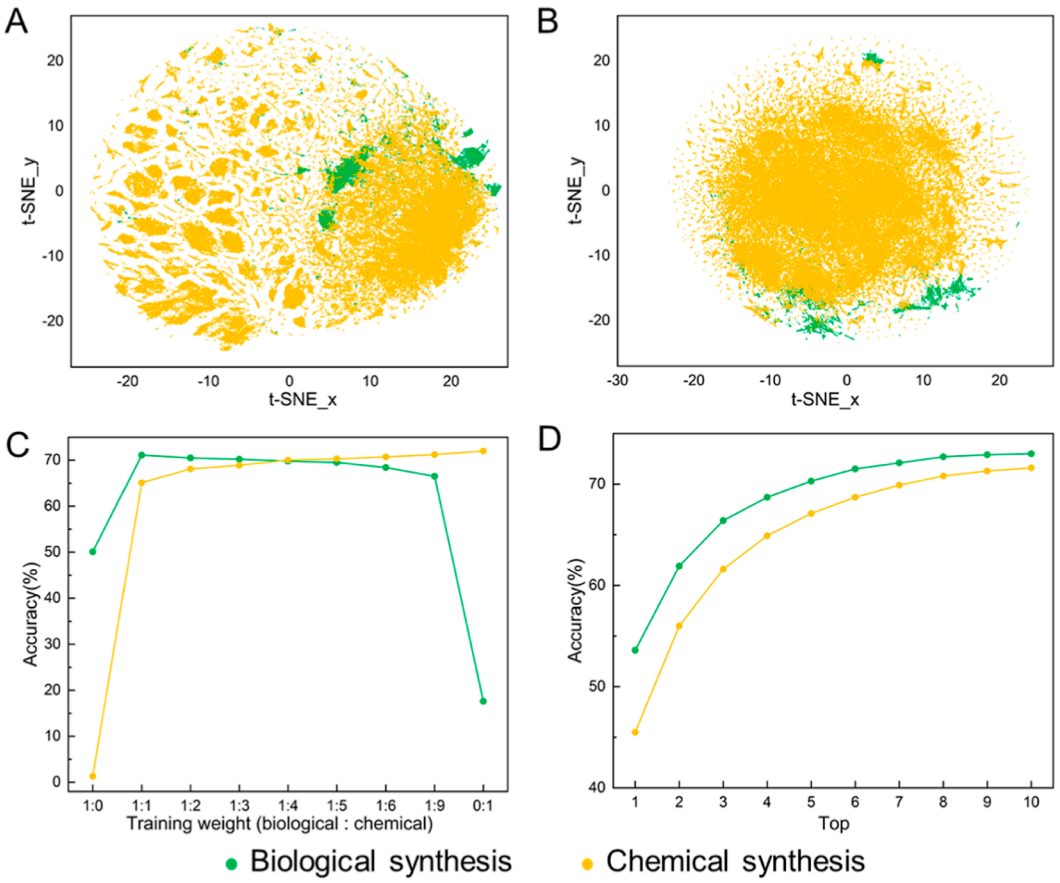


Figure 2. t-SNE distribution of reaction fingerprints (A) and molecule fingerprints (B) from different data sets. (C) Top-10 accuracy on test set with different training weights. (D) Top accuracy of the ensemble model on the test set.

reaction templates are employed to estimate the reaction feasibility, making the deep learning model interpretable and reliable for pathway ranking and enzyme selection. The web server has also been optimized and named BioNavi, which is deployed at http://biopathnavi.qmclab.com/bionavi/. Case studies demonstrate that BioNavi not only improves significantly in the biosynthetic pathway prediction for natural products but also exhibits the potential for chemical synthesis prediction, making it a promising tool for hybrid synthesis pathway design.

RESULTS

Single-Step Prediction

All reactions extracted from the public databases were filtered by removing unbalanced reactions and duplicates. A total of 75,012 enzymatic and 889,557 nonenzymatic reactants−product pairs were generated from the reactions after atom−atom mapping, which were divided into biological and chemical synthesis data sets, respectively (see Methods). Meanwhile, 6027 and 119,632 enzymatic and nonenzymatic reaction templates were extracted. The transformer model was used for single-step prediction, where the product was input and the reactants were output. It is reported that natural products and synthetic molecules exhibit different properties and are located in different chemical spaces,[37] so we speculate that the enzymatic and nonenzymatic reactions also capture different chemical patterns. This can be supported by the distribution of the chemical space of reactions and their components, where most of the data points from the biological data set are clustered in distinct regions (Figure 2A,B). The model can be biased if trained directly from data sets with an imbalanced size. Therefore, we weighted different corpus (i.e., the biological and chemical synthesis data sets) when training the Transformer models.

Figure 2C shows that the model does not perform well on the chemical synthesis set (1.3%) if trained only with the biological set, and vice versa (17.6%). As the proportion of chemical reaction pairs increases, the accuracy of the chemical synthesis set improves, while the accuracy of biological synthesis first significantly increases and then decreases. This can be due to the size of the biological data set that limits the model performance, which indicates the importance of data augmentation for bioretrosynthesis prediction.[25,26] Considering the model performance on both data sets, weight 1 on the biological synthesis data set and weight 4 on the chemical synthesis data set were selected to train the single-step model (69.8 and 70.0% for biological and chemical synthesis data sets, respectively). Four models with different training hyperparameters were selected as the model ensemble as our previous work did, which improved the performance with top-10 accuracy achieving 73.0 and 71.6% for biological and chemical synthesis data sets, respectively (Figure 2D).

Multistep Pathway Search

The ability of multistep pathway prediction was evaluated on two data sets containing natural products and drugs (mainly synthetic compounds, see Methods). The target compounds in the natural products data set are the same as those used in our previous work;[26] however, the pathways contain more

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

branches and components since all necessary components are kept for products based on atom−atom mapping in this work (Figure S1). Compared to other bioretrosynthesis approaches, the results show that BioNavi can generate pathways connecting the target structures and building blocks for 97% (success rate) of the natural products, achieving the highest hit rate for the reported pathways (75%, Table 1). For drugs, BioNavi also performed the best, with a success rate of 76.4%, demonstrating its robustness. Although BioNavi requires a longer time to make predictions, it is acceptable for each molecule to take around 10 min on average. The predicted solutions can be used to estimate pathway diversity, which is also important to pathway reconstruction and design. Compared with other models, BioNavi tends to produce more alternatives for both natural and synthetic molecules. To further investigate the diversity of the pathway collections, the Simpson concentration index,[38] commonly used in ecology to measure species diversity, was introduced (see Methods). Figure 3 A and B show the diversity distribution of pathways from natural products and drug data sets. For natural products, the prediction of RetroPathRL and BioNavi exhibits higher diversity than the other methods. For drugs, the diversity of the four approaches is similar, with BioNavi being slightly better than the others. In the results of the natural products and drugs data sets, 3093 (out of 10,984, 28.2%) and 684 (out of 3870, 17.7%) reactions, respectively, are not covered by the templates. Coverage failure does not necessarily mean poor feasibility but indicates that these reactions need to be evaluated seriously. More importantly, this highlights the advantage of deep learning models for the exploration of the reaction space beyond the reaction templates.

Table 1. Performance of BioNavi and Other Approaches on Different Datasets[a]

	natural products data set (368)			
	success rate (%)	hit rate of pathways (%)	avg. solution[b]	average time (minutes)
RetroPathRL	59.8	3.8	6.1	4.4[c]
RetroBioCat	33.2	1.1	9.4	2.5
RXN4Chem[d]	42.7	2.7	3.2	**0.8**
BioNavi-NP	89.4	48.1	9.3	2.9
ASKCOS	36.4	18.8	8.7	3.3
BioNavi	**97.0**	**75.0**	**9.5**	4.4
	top retail sales drugs data set (110)			
	success rate (%)	avg. solution	average time (minutes)	
LocalRetro	49.1	9.3	**1.6**	
R-SMILES	75.4	9.6	6.5	
ASKCOS	57.1	9.0	3.4	
BioNavi	**76.4**	**9.7**	10.9	

[a]The best-performing method for each metric is shown in bold. [b]A maximum of 10 pathways were considered for all approaches. [c]The runtime of RetroPathRL was controlled by a user-defined parameter, which was set to 5 min in this work. [d]Abbreviated representation of RXN for the Chemistry method.

Since BioNavi and ASKCOS are designed for hybrid synthesis pathway prediction, further analysis and comparison are performed based on their outputs. For BioNavi, there are reactions belonging to both chemical and biological categories since the top 10 results are output in single-step predictions, where the same reactants can be paired with different reaction types. In the natural products data set, it is not surprising that a minority of the reactions (6.4%, Figure 3C) and pathways (2.4%, Figure 3D) are chemical since currently chemical synthesis is not widely used in natural product synthesis. In the results of ASKCOS, the proportions of chemical reactions (50.8%) and pathways (29.8%) are slightly higher than those of biological reactions (49.2 and 27.3%, respectively). It should be noted that 64.7% of the pathways predicted by BioNavi are hybrid, which is higher than that of ASKCOS (42.9%). In the drug data set, chemical reactions and pathways are the majority for both BioNavi and ASKCOS predictions. Again, BioNavi outputs more hybrid pathways than ASKCOS. Sankaranarayanan et al.[34] proposed a complementary algorithm (refer to as ASKCOS-CE in this work) for performing multistep chemoenzymatic retrosynthesis based on the chemical synthesis results of ASKCOS, which can identify more biocatalytic steps. Herein, the algorithm was directly applied to the reaction networks generated by ASKCOS, which discovered more potential hybrid synthesis pathways, especially for drugs. The results demonstrate the adaptability of BioNavi in selecting reaction types and its ability to design hybrid synthesis pathways for both natural products and synthetic compounds.

Case Studies

We first investigated the synthetic pathway of the top-1 small-molecule drug nirmatrelvir, which is an anti-COVID-19 agent. As previously described by Pfizer[41] (Figure 4A), nirmatrelvir was synthesized by two key intermediates, A5 and A6. A6 was obtained by amine ester exchange from A7, followed by the removal of Boc. A5 was obtained through condensation (A2), hydrolysis, and Boc removal of A1 (N-Boc-protected A3) and then condensation with A4. Although BioNavi did not trace back to the predefined building blocks, one of the predicted pathways (ranked third) is consistent with the reported pathway, in which the reported reagents and key intermediates are reproduced (Figure 4A, the complete result can be found in Figure S2). What is different is that the synthesis order was A5, where building block A3 was condensed with A4 first and then A2 in the reported pathway. This can be the alternate pathway since it is shorter than the reported one. Another difference is the N-protection strategy (A7 and A8) in which Boc and Cbz are used in reported and predicted pathways, both of which are commonly used in chemical synthesis.

Except for the non-natural molecule, jomthonic acid A (JAA), a natural product first isolated from the culture broth of a soil-derived actinomycete of the genus Streptomyces[42] was then tested. JAA is an interesting modified amino acid that contains rare structural features and shows antidiabetic and antiatherogenic activities. Although JAA is a natural product, chemical reactions widely exist in the top 10 synthetic pathways (Figure S3). In the top-ranked (1st and second) pathways, JAA is synthesized from three biobased building blocks by a hybrid pathway. This is almost the same as the chemoenzymatic pathway reported by a recent work,[43] where three building blocks (B1, B3, and B5) were used, and an aromatic amino acid aminotransferase was developed for the preparation of β-branched aromatic α-amino acids (B2) with high diastereo- and enantioselectivity (Figure 4B).

The above cases demonstrate the pathway navigation ability of BioNavi in autonomously predicting precursors and the reaction types for the target molecule. Alternatively, users can also freely choose to output pathways that only include biological or chemical synthesis. Taking syringic acid (SA) as

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

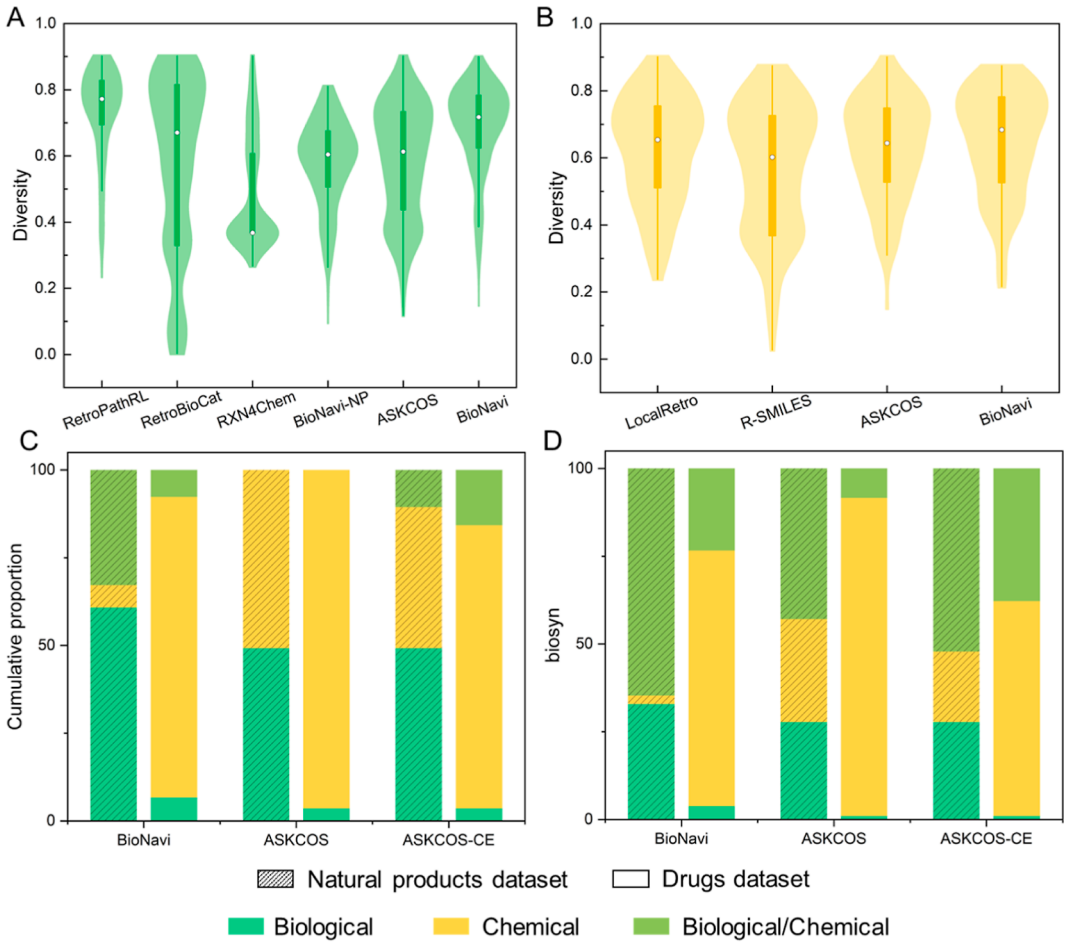


Figure 3. Pathway diversity distribution of natural products data set (A) and top drugs data set (B). 1 represents the highest diversity, and 0 represents the lowest diversity. The proportion of reactions (C) and pathways (D) in outputs using different approaches on different data sets.

an example, we investigated if BioNavi can be used for biosynthetic pathway design by limiting the precursor prediction to biological reactions. SA is a phenolic compound of natural origin and exhibits various biological activities such as antioxidant, anti-inflammatory, anticancer, and antidiabetic.[44] As shown in Figure 4C, SA can be derived from the shikimic acid pathway through a series of metabolites like phenylalanine (C2), cinnamic acid (C3), and sinapinic acid (C4). To explore the biosynthesis of SA, BioNavi was used to predict the synthetic pathways, with reactions being limited to biological ones and the building block being limited to 3-dehydroshikimate (C1), which is the key intermediate upstream of the shikimic acid pathway. A total of five pathways were obtained (Figure S4); the top 3 pathways start from the reduction of 3-dehydroshikimate and then a few steps of oxidation and methylation (Figure 4C). Considering that the biosynthesis from 3-dehydroshikimate to phenylalanine requires another 7 steps,[45] the predicted pathways extremely simplify the synthesis of SA (3 vs 15). It should be noted that the reported SA synthetic pathway from phenylalanine can also be reproduced by manually selecting the "right" precursor in every single step (Figure S5) and most of the precursors are ranked near the top except for the penultimate step (ranked eighth). This indicates that, in addition to relying on the top routes provided by the model, the selection of precursors based on expert experience is also worth noting. That is why we provide a convenient one-step prediction module for users to utilize (as described in the Web server section).

The predicted SA biosynthesis pathways were constructed using *Saccharomyces cerevisiae* to verify the feasibility of the predictions (Figure S6). The pathway 1 (C1−C7−C9) is the shortest, where C1 can convert to C7 directly. Although BioNavi did not identify the enzyme of this step, and it is not verified in this work, it is reported that a bifunctional enzyme can catalyze it.[46] For pathway 2 (C1−C6−C7−C9), 3-dehydroshikimate dehydratase (3DSD) was first integrated and overexpressed in *S. cerevisiae* to construct the protocatechuic acid (C6) biosynthetic pathway (Le01 strain, Figure 4D). Then, the *p*-hydroxybenzoate hydroxylase (PobA) and caffeate O-methyltransferase (COMT) were sequentially integrated into the Le01 strain (L02 strain and L03 strain, respectively), leading to the biosynthesis of gallic acid (C7) (Figure 4E). Unfortunately, SA was not detected in the L03 strain (Figure S7), indicating that the COMT did not work. Previous studies realized the conversion from C6 to C8 with OMT from *Homo sapiens*[47] or mutated OMTs from *Medicago sativa*,[48] reminding us that additional efforts to find the adaptive enzymes are needed for the heterologous biosynthesis of SA. Nevertheless, BioNavi is a great inspiration tool for pathway construction with excellent pathway exploration ability. In summary, the case studies demonstrate that BioNavi can be used for synthetic pathway design and reconstruction

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

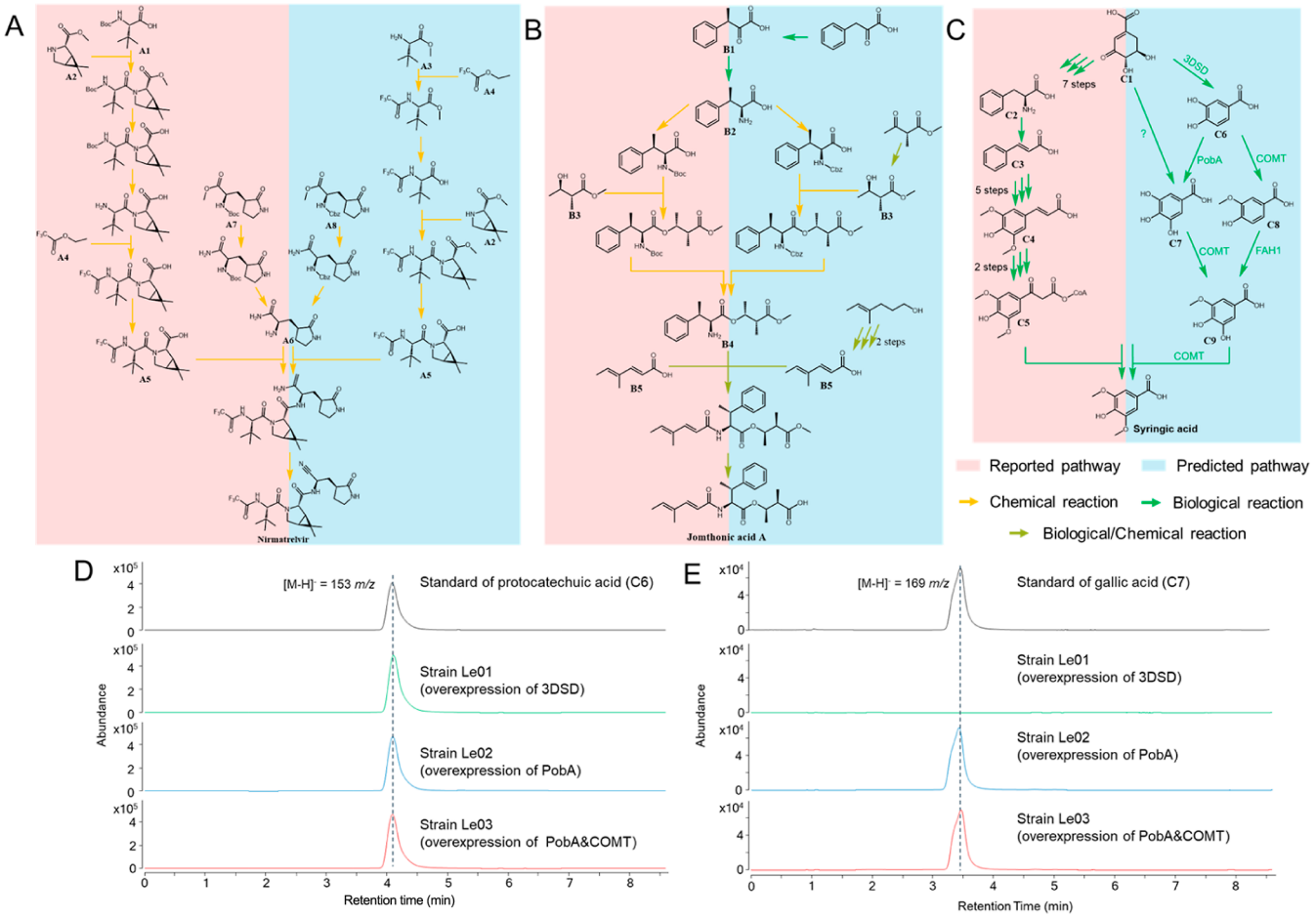


Figure 4. Reported and predicted synthetic pathways of nirmatrelvir (A), jomthonic acid A (B), and SA (C). Selected ion chromatograms for protocatechuic acid (D) and gallic acid (E) from LC−MS analysis. The parent strain is the unpublished protocatechuic acid (C6) biosynthesis strain. 3DSD, 3-dehydroshikimate dehydratase; PobA, *p*-hydroxybenzoate hydroxylase; COMT, caffeate O-methyltransferase; and FAH1, ferulic acid 5-hydroxylase 1.

for both natural and non-natural molecules in a chemo-enzymatic manner.

Web Server

The BioNavi web server (http://biopathnavi.qmclab.com/bionavi/) provides a better user experience compared to BioNavi-NP (Figure 5). Some user-defined parameters (e.g., expansion time and max depth) are no longer required as they would be automatically set as the best value according to the target molecule. Every reaction in the pathway network will be checked to see if it can be reproduced by the reaction templates (see Methods), and then the template and most similar reference reaction will be displayed on the web page, which can be used to search for potential enzymes. Meanwhile, the interfaces of two tools (Selenzyme[49] and E-zyme 2[50]) used in BioNavi-NP are preserved for enzyme selection. For chemical synthesis reactions, a deep learning-based condition prediction tool (Parrot[51]) was also integrated into the web server, by which the catalysts, solvents, and reagents can be predicted for each reaction. To allow users to score the precursors and determine the search direction based on their own experience, a step-by-step mode was provided to complement the original pathway navigation module (Figure 5C). This gives users more options to improve the poor performance of the scoring method in some cases and to adapt the search direction to various building blocks.

DISCUSSION

We showcase here the development of the hybrid retrosynthesis planning approach (BioNavi) by leveraging the advantages of deep learning and reaction templates. In particular, the atom−atom mapping strategy is used to improve the data quality, thus enhancing the precursor prediction accuracy. Multitask learning is introduced to balance the prediction of chemical and biological synthesis. Furthermore, integrating reaction templates into the scoring evaluation makes the prediction more interpretable and guides the enzyme selection. Extensive tests demonstrate that BioNavi not only consistently and comprehensively outperforms current approaches in predicting the biosynthetic pathways of natural products but also can be utilized to seek the chemical, biological, or even hybrid pathways with higher diversity for any molecule. Besides, the revamped web server makes BioNavi easier and more adaptive for navigating the potent synthetic pathways for target compounds, promoting efficient production of high-value-added chemicals.

Nevertheless, there remains a considerable distance to traverse from the pathway design to efficient production. One of the challenges lies in the available reaction data. First, most of the data-driven pathway planning tools (including BioNavi) predominantly rely on experimentally validated reactions (positive data) during model development. However,

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

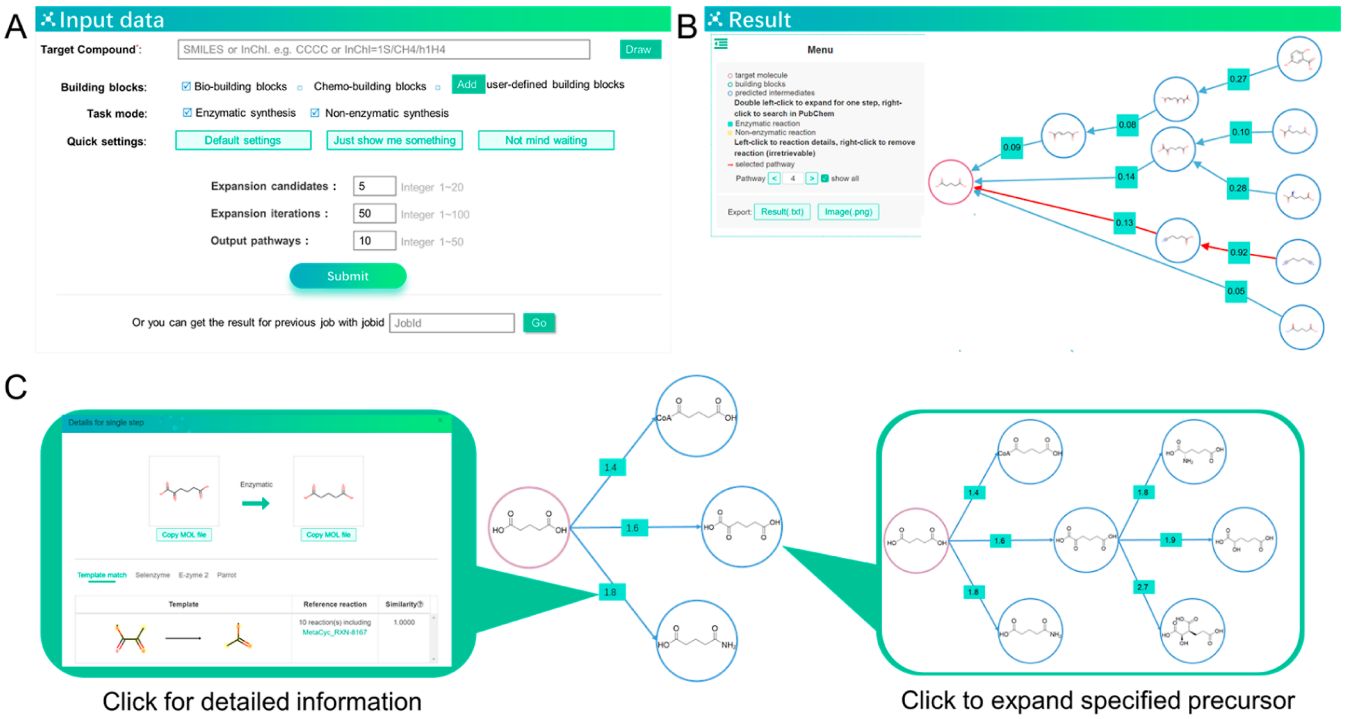


Figure 5. BioNavi input (A) and output (B) interfaces. (C) Details of the result panel of pathway planning. Any precursor or intermediate can be expanded according to the score, enzyme information, or personal experience.

the reactions with low or no yield (negative data) are also important for models to capture chemical patterns.[52] It is challenging to collect negative data while considering that the reaction yield can be an option and that it will be helpful to the pathway evaluation. Although many models have been proposed to predict reaction yields,[53] they are independent of pathway planning tools, and future works can be focused on the fusion of reaction yields in retrosynthesis prediction.[54] Second, stereochemistry is often missing (partial or complete) in existing reactions. Although the SMILES representation can encode the stereochemistry of structures, the prerequisite is that the data contain the correct stereochemistry. Especially for biological reactions, where stereoselectivity is an important feature, stereochemistry needs to be taken seriously in future reaction database curation. Another challenge lies in the condition selection (such as temperature, solvent, and catalysts) for the reactions along the pathways. Substantial efforts have been made to predict the reaction conditions or chemical contexts with physical-based (such as quantum chemistry) and data-driven (machine learning) models.[55] Retrosynthesis planning models can benefit from the recommendations of such tools. Enzyme engineering approaches are also necessary to improve the catalytic activity of specific biological reactions. Recent computational approaches such as fitness prediction[56] and protein generation[57] will also accelerate the process of enzyme selection.

METHODS

Data Set

The biological reactions were collected from MetaCyc,[58] KEGG,[59] Rhea,[60] and BRENDA,[21] and chemical reactions were retrieved from the USPTO[20] data set. All reactions with an unbalanced number of carbon atoms were removed, followed by atom−atom mapping with RXNMapper.[61] Then, reaction templates were extracted with the RDChiral package,[35] and the reactions sharing the same template (reference reactions) will be collected. In most retrosynthesis prediction scenarios, one can only provide the specific target molecule, so the retrosynthesis model takes only one molecule as input. Herein, the reactions with multiple products were first split into multiple reactions that kept all substrates and only one of the products (i.e., the reaction A + B ≫ C + D will be split into A + B ≫ C and A + B ≫ D). To simplify the influence of cofactors and coenzymes, only the reactants containing carbon atoms with the same indexes as the product were preserved (Figure S8). This allows us to keep all necessary reactants consisting of the product while minimizing data complexity. Finally, all reactants−product pairs were standardized by calculating canonical SMILES with RDKit and then deduplicated. On the side of reactants, a label indicating the reaction type was added to the end of the SMILES with "|" as separation; for example, a biological reaction was represented as "Cc1cnc2ccccc2c1|⟨B⟩ ≫ Cc1cc2ccccc2-[nH]c1=O", and a chemical reaction can be "Cc1ccc(F)cc1C#N|⟨C⟩ ≫ Cc1ccc(F)cc1C#N".

Computational Model

All data pairs were randomly split into training, validation, and test data sets (8:1:1), and the products in the test set do not appear in the training set. The Transformer model was trained with SMILES of the product as input and reactants (along with a reaction-type label) as the output. The corpora labeled with chemical and biological reactions were loaded and trained with different weights to achieve multitask learning. Different weights have been tested, and 4:1 was selected for pathway evaluation. The weighted training was realized by the OpenNMT framework on Nvidia RTX 3090 with hyperparameters listed in Table S1, and four models trained with different random seeds constitute the ensemble model.

A score (P) output along with each precursor candidate can be used to estimate the probability that the model will output the specific candidate for the target molecule. Considering the intrinsic drawback of the end-to-end approach to SMILES that minor changes in strings can cause significant structural changes, the model sometimes makes unreasonable predictions. We rerank the predicted candidates by reproducing the reactions with templates, the score will be updated by multiplying by a coefficient (range from 0 to 1) that is related to the number of reference reactions (N) and the maximum similarity (s,

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

calculated according to molecular fingerprint ECFP[62]) of the target molecule, and products in reference reactions. Generally speaking, the bigger N means a stronger generality of the reaction rule, indicating the predicted reaction is more likely to occur. Meanwhile, the higher the s, the more likely the predicted reaction is to occur under the same conditions. Thus, the coefficient should increase with the increases in N and s. We designed the coefficient as follows

$$\text{score} = \begin{cases} 0.5P, & N = 0 \\ \dfrac{s+N}{1+N}P, & N > 0 \end{cases}$$

If the predicted reaction cannot be reproduced by templates (i.e., $N = 0$), it does not necessarily mean that the reaction is unreasonable; it may be a new reaction outside of the template library. Therefore, we only multiplied the original probability by a smaller coefficient (0.5, since the coefficient is greater than 0.5 when N is greater than 0). For reactions that can be reproduced, the reference reaction with maximum similarity will be output and shown on the web page, providing a reference for enzyme selection. After reranking, the Retro* algorithm will estimate the synthesis cost from building blocks to specific precursors with a pretrained value function as described in the original work,[19] and iteratively make single-step predictions until the building blocks are reached. By default, 387 metabolites from *Escherichia coli* (iML1515 model)[63] and molecules with less than 4 carbon atoms were defined as biobased building blocks. On the BioNavi web server, there is an option to select commercially available or user-defined structures as building blocks.

Evaluation Test

The biosynthetic pathways of 368 natural products used in our previous work[26] were collected from the reprocessed biological data set in this work, and the pathways have been verified by the public databases. Three template-based (RetroPathRL,[22] RetroBioCat[23] and ASKCOS[33]) and three template-free approaches (RXN for Chemistry,[25] BioNavi-NP[26] and BioNavi) were investigated on this data set to evaluate the biosynthetic pathway exploration power for natural products. Besides, 110 molecules from the top-100 small-molecule drugs by sales in 2022[39] were also tested (some brand drugs contain multiple major constituents; for example, Paxlovid contains nirmatrelvir and ritonavir[40]). Since most of the drugs are non-natural products and tend to be synthesized by chemical steps, template-based (LocalRetro[24]) and template-free (R-SMILES[28]) chemical retrosynthesis methods, along with ASKCOS and BioNavi, were used to make predictions. For the natural products data set, the biobuilding blocks were set as molecules with less than 4 carbon atoms and another 387 metabolites, as mentioned above, while for the top drugs data set, another 106,750 buyable molecules used in ASKCOS[33] were added to the building block list. Retro*[19] was used for pathway planning, with the expansion number being 10, iteration being 100, and maximum pathway being 10 unless otherwise specified. For a specific molecule, if one of the output pathways terminates with predefined building blocks, it is labeled as "successful", which is related to the success rate. If one of the output pathways contains exactly all components along the reported pathway, it is labeled as "hit", which is related to the pathway hit rate. Simpson concentration[38] indicates the probability that two individuals chosen at random and independently from the population will be found to belong to the same group. Herein, for a target molecule, if the model outputs k pathways which include a collection of reactions, the pathway diversity (D) is defined as the reciprocal of the probability that two reactions chosen at random and independently from the collection will be found to belong to the same pathway

$$D = \frac{1}{\sum_{i=1}^{k}\left(\frac{n_i}{N}\right)^2} \quad (1)$$

N is the total number of reactions, and n_i (i = 1, 2, 3, ..., k) is the number of reactions for a specific pathway.

All approaches, except for RXN4Chem, were installed locally by the source codes provided in the original publications for evaluation. The API of RXN4Chem was accessed using a Python wrapper (https://github.com/rxn4chemistry/rxn4chemistry). The structures used to determine termination for all tools (also except for RXN4Chem, which cannot be changed) were set as the building blocks described above. For RetroPathRL, the configuration of "Golden Default" was used with "time budgets" changed to 300. For BioNavi-NP, ASKCOS, LocalRetro, and R-SMILES, the single-step models were paired with Retro* for multistep pathway prediction, with the same parameters set as BioNavi. For RetroBioCat, the top 10 pathways were kept. All of the remaining parameters have retained the default settings. The evaluation of locally installed approaches was performed on 1 Nvidia RTX 3090 GPU and 2 AMD EPYC 7282 (16 cores) CPUs.

EXPERIMENTAL MATERIALS

Strain Growth Condition

Yeast strains were cultivated in YPD broth (Sangon Biotech., China) with 20 g L-1 glucose as the carbon source. The *URA*3 marker-based selection consist of a synthetic complete medium without uracil of SC/-Ura Broth containing YNB, ammonium sulfate, and amino acids (Coolaber Science & Tech., China) with 20 g L-1 glucose as the carbon source. The solid agar plates were used 2% of Bacto Agar (BD Biosciences, USA).

Reconstitution of Gallic Acid Biosynthesis in Yeast

The strains undergo high-level expression of heterologous genes by the CRISPR/*Cas*9 system.[64] The genetic integration of the expression cassette was performed by fusion PCR[65] for the upstream homologous region of (XI-1 UP)—Promoter (*TDH* 3*p*)—yeast codon usage optimized and synthesized 3*DSD*—Terminator (*CYC*1*t*)—Downstream of homologous region (XI-1 DW) to construct the Le01 strain. The Le02 strain was constructed by integration of the Upstream homologous region (XII-1 UP)—Terminator (*TPS*1*t*)—yeast codon usage optimized and synthesized *PobA*—Promoter (*CCW*12*p*)—Downstream homologous region (XII-1DW). The Le03 strain was constructed on the XII-1 integration point, where both genes of *PobA* and *COMT* were simultaneously overexpressed. The strain construction stratagem and schematic are illustrated in Figure S6. The constructed cassette was then transformed into the previously engineered parent strain of QL35[66] by using the LiAc/SS carrier DNA/PEG method.[67] All transformants were grown at 30 °C on a SC/-Ura agar plate for select positive transformants. The selected single colony was inoculated into a test tube containing 1 mL of the SC/-Ura liquid medium supplemented with 2% glucose and cultured for 2 days. The expression cassette assembled using primer pairs was synthesized by Sangon Biotech. (China) in Tables S2 and S3.

Yeast Fermentation

The selected single colony was inoculated in YPD medium as a seed culture transferred into 100 mL of Erlenmeyer flasks, each flask containing 20 mL of YPD medium. The inoculated cell density (optical density of 600 nm wavelength, OD_{600}) for consistency was normalized to 0.2. The cell density of each sample was measured using an Ultrospec 10 cell density meter (Biochrom, USA) with a 10 mm long cuvette (Fisher, USA). The flasks cultured in a shaking incubator set at 30 °C and 200 rpm for 4 days. All strains were inoculated into three biological replicates.

Metabolites Analysis by LC−MS

The LC−MS system was composed of Agilent 6470 with 6495 triple quadrupole mass spectrometers (Agilent Technologies). Reverse phase separation of metabolites was performed on a Phenomenex Kinetex C18 column, particle size of 2.6 μm, 100 × 2.1 mm. The metabolites were subjected through electrospray ionization to mass spectroscopy on selected ion monitoring (SIM) by the negative mode. The mobile phase consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). The solvent A linear gradient of 5−95% of solvent B in solvent A over 12.0 min at a flow

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

rate of 0.2 mL/min was used. The total analytical time was 15.0 min per run, and 2 μL was injected into the LC system. The metabolites used the negative ion mode $[M - H]^-$ for SIM of protocatechuic acid 153 m/z, gallic acid 169 m/z, and SA 197 m/z. The Agilent MassHunter Quantitative (version 10.1) analysis software was used for the ion current spectrum and mass fragment data processing.

■ ASSOCIATED CONTENT

Data Availability Statement

BioNavi web server is freely available at http://biopathnavi.qmclab.com/bionavi/, the code, pretrained models, and all data sets mentioned in the manuscript are available at https://github.com/zengtsysu/BioNavi.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.4c00228.

Hyperparameters of the computational model; primers and sequences used to verify the predicted pathway; comparison of natural products data sets in BioNavi and BioNavi-NP; raw outputs of the BioNavi web server for three cases; step-by-step pathway of SA; construction of engineered strains; detection of SA; and illustration of the reaction split according to atom−atom mapping (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ruibo Wu − *School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou 510006, P. R. China;* orcid.org/0000-0002-1984-046X; Email: wurb3@mail.sysu.edu.cn

Authors

Tao Zeng − *School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou 510006, P. R. China*

Zhehao Jin − *Center for Synthetic Biochemistry, CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), Shenzhen 518055, P. R. China*

Shuangjia Zheng − *Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*

Tao Yu − *Center for Synthetic Biochemistry, CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), Shenzhen 518055, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c00228

Author Contributions

R.W. designed and supervised the whole research. T.Z. collected, processed, and analyzed the data. Z.J. performed the in vivo experiments. T.Y. discussed the experimental results. T.Z., S.Z., and R.W. discussed and contributed the computational model as well as the web server. T.Z. and Z.J. wrote the primary manuscript, and all authors contributed to revising it.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Key-Area Research and Development Program of Guangdong Province, China (2022B1111080005), and the National Key Research and Development Program of China (2023YFC3404900). We also thank for the support from the Top-Notch Young Talents Program of China and the Guangzhou Supercomputer Center for providing computational source.

■ REFERENCES

(1) Lee, S. Y.; Kim, H. U.; Chae, T. U.; Cho, J. S.; Kim, J. W.; Shin, J. H.; Kim, D. I.; Ko, Y. S.; Jang, W. D.; Jang, Y. S. A comprehensive metabolic map for production of bio-based chemicals. *Nat. Catal.* **2019**, *2*, 18−33.

(2) Zhou, H.; Xie, X.; Tang, Y. Engineering natural products using combinatorial biosynthesis and biocatalysis. *Curr. Opin. Biotechnol.* **2008**, *19*, 590−596.

(3) Miller, D. C.; Athavale, S. V.; Arnold, F. H. Combining chemistry and protein engineering for new-to-nature biocatalysis. *Nat. Synth.* **2022**, *1*, 18−23.

(4) Gao, J.; Zuo, Y.; Xiao, F.; Wang, Y.; Li, D.; Xu, J.; Ye, C.; Feng, L.; Jiang, L.; Liu, T.; et al. Biosynthesis of catharanthine in engineered Pichia pastoris. *Nat. Synth.* **2023**, *2*, 231−242.

(5) Tang, H.; Lin, S.; Deng, J.; Keasling, J. D.; Luo, X. Engineering yeast for the de novo synthesis of jasmonates. *Nat. Synth.* **2023**, *3*, 224−235.

(6) Liu, J.-H.; Yu, B.-Y. Biotransformation of Bioactive Natural Products for Pharmaceutical Lead Compounds. *Curr. Org. Chem.* **2010**, *14*, 1400−1406.

(7) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; et al. Biocatalysis. *Nat. Rev. Methods Primers* **2021**, *1*, 46.

(8) Kirschning, A.; Hahn, F. Merging Chemical Synthesis and Biosynthesis: A New Chapter in the Total Synthesis of Natural Products and Natural Product Libraries. *Angew. Chem. Int. Ed* **2012**, *51*, 4012−4022.

(9) Rudroff, F.; Mihovilovic, M. D.; Gröger, H.; Snajdrova, R.; Iding, H.; Bornscheuer, U. T. Opportunities and challenges for combining chemo- and biocatalysis. *Nat. Catal.* **2018**, *1*, 12−22.

(10) Makkee, M.; Kieboom, A. P.; Van Bekkum, H.; Roels, J. A. *Combined Action of Enzyme and Metal Catalyst, Applied to the Preparation of D-Mannitol*; Delft University Press, 1984.

(11) Zhang, X.; King-Smith, E.; Dong, L. B.; Yang, L. C.; Rudolf, J. D.; Shen, B.; Renata, H. Divergent synthesis of complex diterpenes through a hybrid oxidative approach. *Science* **2020**, *369*, 799−806.

(12) Slagman, S.; Fessner, W.-D. Biocatalytic routes to anti-viral agents and their synthetic intermediates. *Chem. Soc. Rev.* **2021**, *50*, 1968−2009.

(13) Chainani, Y.; Bonnanzio, G.; Tyo, K. E. J.; Broadbelt, L. J. Coupling chemistry and biology for the synthesis of advanced bioproducts. *Curr. Opin. Biotechnol.* **2023**, *84*, 102992.

(14) Hadadi, N.; Hatzimanikatis, V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **2015**, *28*, 99−104.

(15) Dong, J.; Zhao, M.; Liu, Y.; Su, Y.; Zeng, X. Deep learning in retrosynthesis planning: datasets, models and tools. *Briefings Bioinf.* **2022**, *23*, bbab391.

(16) Boob, A. G.; Chen, J.; Zhao, H. Enabling pathway design by multiplex experimentation and machine learning. *Metab. Eng.* **2024**, *81*, 70−87.

(17) Yu, T.; Boob, A. G.; Volk, M. J.; Liu, X.; Cui, H.; Zhao, H. Machine learning-enabled retrobiosynthesis of molecules. *Nat. Catal.* **2023**, *6*, 137−151.

(18) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.

(19) Chen, B.; Li, C.; Dai, H.; Song, L. Retro*: learning retrosynthetic planning with neural guided A* search. In *International Conference on Machine Learning*; PMLR, 2020.
(20) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Doctoral thesis, University oc Cambridge, 2012.
(21) Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* **2021**, *49*, D498−D508.
(22) Koch, M.; Duigou, T.; Faulon, J. L. Reinforcement Learning for Bioretrosynthesis. *ACS Synth. Biol.* **2020**, *9*, 157−168.
(23) Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **2021**, *4*, 98−104.
(24) Chen, S.; Jung, Y. Deep Retrosynthetic Reaction Prediction using Local Reactivity and Global Attention. *JACS Au* **2021**, *1*, 1612−1620.
(25) Probst, D.; Manica, M.; Teukam, Y. G. N.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed synthesis planning using data-driven learning. *Nat. Commun.* **2022**, *13*, 964.
(26) Zheng, S.; Zeng, T.; Li, C.; Chen, B.; Coley, C. W.; Yang, Y.; Wu, R. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat. Commun.* **2022**, *13*, 3342.
(27) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103−1113.
(28) Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Wu, M.; Hou, T.; Song, M. Root-aligned SMILES: a tight representation for chemical reaction prediction. *Chem. Sci.* **2022**, *13*, 9023−9034.
(29) Ucak, U. V.; Ashyrmamatov, I.; Ko, J.; Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **2022**, *13*, 1186.
(30) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.
(31) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A graph to graphs framework for retrosynthesis prediction. In *Proceedings of the 37th International Conference on Machine Learning*; JMLR.org, 2020.
(32) Wang, Y.; Pang, C.; Wang, Y.; Jin, J.; Zhang, J.; Zeng, X.; Su, R.; Zou, Q.; Wei, L. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks. *Nat. Commun.* **2023**, *14*, 6155.
(33) Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W. Merging enzymatic and synthetic chemistry with computational synthesis planning. *Nat. Commun.* **2022**, *13*, 7747.
(34) Sankaranarayanan, K.; Jensen, K. F. Computer-assisted multistep chemoenzymatic retrosynthesis using a chemical synthesis planner. *Chem. Sci.* **2023**, *14*, 6467−6475.
(35) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59*, 2529−2537.
(36) Rahman, S. A.; Torrance, G.; Baldacci, L.; Cuesta, S. M.; Fenninger, F.; Gopal, N.; Choudhary, S.; May, J. W.; Holliday, G. L.; Steinbeck, C.; et al. Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics* **2016**, *32*, 2065−2066.
(37) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(38) Simpson, E. H. Measurement of Diversity. *Nature* **1949**, *163*, 688.
(39) McGrath, N. A.; Brichacek, M.; Njardarson, J. T. A Graphical Journey of Innovative Organic Architectures That Have Improved Our Lives. *J. Chem. Educ.* **2010**, *87*, 1348−1349.
(40) Lamb, Y. N. Nirmatrelvir Plus Ritonavir: First Approval. *Drugs* **2022**, *82*, 585−591.
(41) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; et al. An oral SARS-CoV-2 Mpro inhibitor clinical candidate for the treatment of COVID-19. *Science* **2021**, *374*, 1586−1593.
(42) Igarashi, Y.; Yu, L.; Ikeda, M.; Oikawa, T.; Kitani, S.; Nihira, T.; Bayanmunkh, B.; Panbangred, W. Jomthonic Acid A, a Modified Amino Acid from a Soil-Derived Streptomyces. *J. Nat. Prod.* **2012**, *75*, 986−990.
(43) Li, F.; Yang, L.-C.; Zhang, J.; Chen, J. S.; Renata, H. Stereoselective Synthesis of β-Branched Aromatic α-Amino Acids by Biocatalytic Dynamic Kinetic Resolution**. *Angew. Chem. Int. Ed* **2021**, *60*, 17680−17685.
(44) Srinivasulu, C.; Ramgopal, M.; Ramanjaneyulu, G.; Anuradha, C. M.; Kumar, C. S. Syringic acid (SA) − A Review of Its Occurrence, Biosynthesis, Pharmacological and Industrial Importance. *Biomed. Pharmacother.* **2018**, *108*, 547−557.
(45) Tohge, T.; Watanabe, M.; Hoefgen, R.; Fernie, A. R. Shikimate and phenylalanine biosynthesis in the green lineage. *Front. Plant Sci.* **2013**, *4*, 62.
(46) Muir, R. M.; Ibáñez, A. M.; Uratsu, S. L.; Ingham, E. S.; Leslie, C. A.; McGranahan, G. H.; Batra, N.; Goyal, S.; Joseph, J.; Jemmis, E. D.; et al. Mechanism of gallic acid biosynthesis in bacteria (Escherichia coli) and walnut (Juglans regia). *Plant Mol. Biol.* **2011**, *75*, 555−565.
(47) Kunjapur, A. M.; Prather, K. L. J. Development of a Vanillate Biosensor for the Vanillin Biosynthesis Pathway in E. coli. *ACS Synth. Biol.* **2019**, *8*, 1958−1967.
(48) Kota, P.; Guo, D.; Zubieta, C.; Noel, J.; Dixon, R. A. O-Methylation of benzaldehyde derivatives by ″lignin specific″ caffeic acid 3-O-methyltransferase. *Phytochemistry* **2004**, *65*, 837−846.
(49) Carbonell, P.; Wong, J.; Swainston, N.; Takano, E.; Turner, N. J.; Scrutton, N. S.; Kell, D. B.; Breitling, R.; Faulon, J. L. Selenzyme: enzyme selection tool for pathway design. *Bioinformatics* **2018**, *34*, 2153−2154.
(50) Moriya, Y.; Yamada, T.; Okuda, S.; Nakagawa, Z.; Kotera, M.; Tokimatsu, T.; Kanehisa, M.; Goto, S. Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate−Product Pairs. *J. Chem. Inf. Model.* **2016**, *56*, 510−516.
(51) Wang, X.; Hsieh, C. Y.; Yin, X.; Wang, J.; Li, Y.; Deng, Y.; Jiang, D.; Wu, Z.; Du, H.; Chen, H.; et al. Generic Interpretable Reaction Condition Predictions with Open Reaction Condition Datasets and Unsupervised Learning of Reaction Center. *Res.* **2023**, *6*, 0231.
(52) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P. O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *Org. Lett.* **2023**, *25*, 2945−2947.
(53) Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. V. When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges. *J. Chem. Inf. Model.* **2024**, *64*, 42−56.
(54) Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **2023**, *9*, 2196−2204.
(55) Andronov, M.; Voinarovska, V.; Andronova, N.; Wand, M.; Clevert, D.-A.; Schmidhuber, J. Reagent prediction with a molecular transformer improves reaction data quality. *Chem. Sci.* **2023**, *14*, 3235−3246.
(56) Luo, Y.; Jiang, G.; Yu, T.; Liu, Y.; Vo, L.; Ding, H.; Su, Y.; Qian, W. W.; Zhao, H.; Peng, J. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun.* **2021**, *12*, 5743.
(57) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089−1100.
(58) Caspi, R.; Billington, R.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Midford, P. E.; Ong, W. K.; Paley, S.; Subhraveti, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **2020**, *48*, D445−D453.

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502

(59) Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **2014**, *42*, D199−D205.
(60) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; et al. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **2022**, *50*, D693−D700.
(61) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **2021**, *7*, No. eabe4166.
(62) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(63) Monk, J. M.; Lloyd, C. J.; Brunk, E.; Mih, N.; Sastry, A.; King, Z.; Takeuchi, R.; Nomura, W.; Zhang, Z.; Mori, H.; et al. iML1515, a knowledgebase that computes Escherichia coli traits. *Nat. Biotechnol.* **2017**, *35*, 904−908.
(64) Mans, R.; van Rossum, H. M.; Wijsman, M.; Backx, A.; Kuijpers, N. G.; van den Broek, M.; Daran-Lapujade, P.; Pronk, J. T.; van Maris, A. J.; Daran, J. M. G. CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in Saccharomyces cerevisiae. *FEMS Yeast Res.* **2015**, *15*, fov004.
(65) Shao, Z.; Zhao, H.; Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* **2009**, *37*, No. e16.
(66) Liu, Q.; Yu, T.; Li, X.; Chen, Y.; Campbell, K.; Nielsen, J.; Chen, Y. Rewiring carbon metabolism in yeast for high level production of aromatic chemicals. *Nat. Commun.* **2019**, *10*, 4976.
(67) Gietz, R. D.; Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2007**, *2*, 31−34.

https://doi.org/10.1021/jacsau.4c00228
JACS Au 2024, 4, 2492−2502