ACS Synthetic Biology

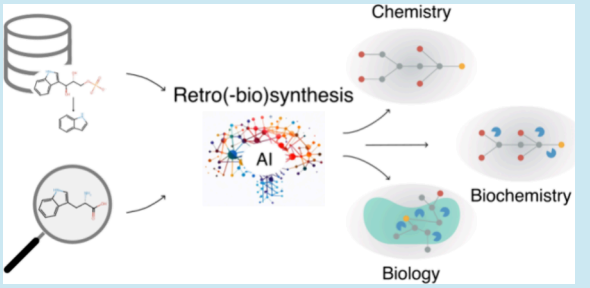
pubs.acs.org/synthbio
Review

Artificial Intelligence Methods and Models for Retro-Biosynthesis: A Scoping Review

Guillaume Gricourt, Philippe Meyer, Thomas Duigou, and Jean-Loup Faulon*

Cite This: https://doi.org/10.1021/acssynbio.4c00091

ABSTRACT: Retrosynthesis aims to efficiently plan the synthesis of desirable chemicals by strategically breaking down molecules into readily available building block compounds. Having a long history in chemistry, retro-biosynthesis has also been used in the fields of biocatalysis and synthetic biology. Artificial intelligence (AI) is driving us toward new frontiers in synthesis planning and the exploration of chemical spaces, arriving at an opportune moment for promoting bioproduction that would better align with green chemistry, enhancing environmental practices. In this review, we summarize the recent advancements in the application of AI methods and models for retrosynthetic and retro-biosynthetic pathway design. These techniques can be based either on reaction templates or generative models and require scoring functions and planning strategies to navigate through the retrosynthetic graph of possibilities. We finally discuss limitations and promising research directions in this field.



KEYWORDS: *retrosynthesis, retro-biosynthesis, artificial intelligence*

INTRODUCTION

Retrosynthesis[1] is essential for the development of new compounds in the fields of pharmaceuticals and organic chemistry, providing chemists with the ability to access complex and novel molecules. This same approach, rebranded as retro-biosynthesis, is also used in biocatalysis, where reactions are catalyzed by enzymes. Compared with conventional chemical synthesis, enzymatic processes can catalyze chemical reactions in a specific, highly efficient manner, requiring less energy and generating minimal waste. Biochemical reactions can be carried out *in vitro*, as with an enzymatic cascade, or *in vivo*, as in synthetic biology via metabolic engineering.[2] Distinct challenges arise in biochemistry and synthetic biology, such as enzyme isolation and managing the trade-off between cell growth and molecular production, respectively.[3]

Identifying a set of building block molecules, also referred to as precursors, readily available in the commercial market or naturally existing in the environment is essential for synthesizing a desired target product through retro(-bio)synthesis. The technological disruption brought about by artificial intelligence (AI) paves the way for new possibilities across each of the key components required for this task.

Retrosynthesis proceeds by iteratively applying single-step processes, which consists of finding all the possible reactants of a given product. Methods have been proposed for single-step retrosynthesis that can be grouped into three categories: template-based, template-free, and semitemplate-based. Template-based methods rely on a library of reaction templates constructed from chemical reaction data sets to match them against a target molecule and extract the reactants from the selected templates. Here, AI techniques have been developed for selecting the most promising templates. Distinctively, template-free approaches use AI generative models for translating products directly into candidate reactants, whereas semitemplate-based methods forecast reactants by iteratively manipulating the bonds within the product.

In all the aforementioned cases, the single-steps are iterated via route-planning algorithms to produce pathways of reactions and identify available precursors. Predicting multistep retrosynthesis pathways is inherently challenging due to the extensive search space for synthesis routes, the subjective determination of selecting suitable candidates, which motivates the use of AI-based combinatorial graph search methods. To guide the route planning and rank-predicted solutions, AI strategies are applied to suggest the best choice according to dedicated algorithms, scoring functions, or the availability of enzymes in retro-biosynthesis.[4]

Because of the fast-paced developments in retrosynthesis for chemical synthesis planning, there is an evident requirement for a thorough summary of pertinent literature. The

Received: February 9, 2024
Revised: June 14, 2024
Accepted: June 14, 2024

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

compilation of scientific articles was inspired by the preferred reporting items for systematic reviews and meta-analyses extension for scoping reviews (PRISMA-ScR)[5] guidelines (Note S1). A comprehensive literature search was conducted across four academic search engines, covering a broad spectrum of disciplines including biology and computer science, both pertinent to the interdisciplinary field of retro(-bio)synthesis. While this search was successful in identifying papers published in academic journals, expanding it to include conference proceeding papers further extended the research scope, though it may not have fully captured all pertinent research presented at academic conferences. Outlined in Figure 1, this review summarizes retrosynthesis methods which can be utilized in several application domains (chemistry, biocatalysis, and synthetic biology) even when originally developed for synthesis planning in organic chemistry. In the following sections, we explore the diversity of AI methods and models (explained in a glossary in Note S2) tailored for both single-step and multistep processes, including the types of data and predictors used and data sets and evaluation metrics employed. We then reviewed popular databases and data set preparation. Finally, we assess the limitations of AI methods and highlight the distinctions across application domains. Our review also aims to identify gaps in knowledge, emphasizing areas that require additional research to advance the application of AI in retro-biosynthesis.

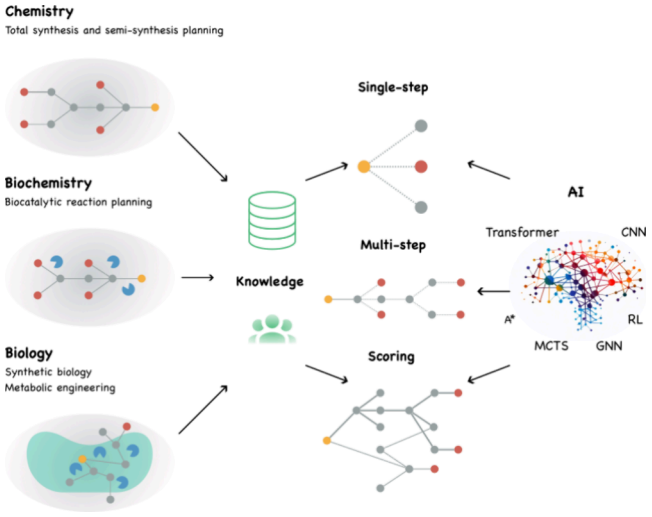


Figure 1. Retrosynthesis principles and its applications. Retrosynthesis is a computer-aided method that uses data sets and user expertise. Current algorithms for retro(-bio)synthesis are applied across several domains. In chemical synthesis, target molecules are crafted through organic chemistry reactions from commercially available building blocks. In biocatalysis, enzymes are employed to catalyze the reactions. Synthetic biology and metabolic engineering go a step further, using living cells such as bacteria, fungi, or plants to facilitate bioproduction pathways and supply the necessary building blocks. The retro(-bio)synthesis process unfolds in three key stages, leveraging molecular databases. The single-step stage consists of predicting reactants (red and gray nodes) necessary for producing a given product (yellow). The multistep stage determines possible routes linking the desirable product (yellow node) to available building blocks (red nodes) using sequences of single-step moves. Completed predictions are shown by a solid line, whereas future predictions are shown by a hatched line. Finally, route scoring helps in finding the best strategy to produce a molecule as well as ranking completed routes. AI techniques now play a critical role in each phase of the retro(-bio)synthesis process. A*, A* search; CNN, convolutional neural network; GNN, graph neural network; MCTS, Monte Carlo tree search; and RL, reinforcement learning.

SINGLE-STEP RETROSYNTHESIS

As AI and its applications continue to advance, computational approaches have been proposed to predict the outcomes of chemical reactions. Chemical reactions involve the transformation of one set of chemical substances into another, which results in chemical changes. Two key challenges exist: predicting the products that result from given reactants (i.e., substrates) and solving the reverse problem of identifying the reactants when the products are known.[6] One approach is the template-based method that extracts information from a reaction chemical database to generalize the application of existing reactions through reaction templates.[7] A template corresponds to subgraph patterns that describe changes in the connectivity between a product and its reactants. A second approach is the template-free method that uses the ability of generative models to predict target molecules.[8] Finally, semitemplate-based methods aim to reconcile the use of dedicated rules and the ability to generalize via AI.[9] The general principle associated with single-step retrosynthesis is shown in Figure 2.

Molecular and Reaction Representations. To utilize molecules and reactions as inputs to AI models, employing a representation or vectorization is essential. The widely adopted approach is to use SMILES and SMARTS notations to encode molecules and reaction templates into strings for models predicting substrates from the product(s).[10,11] Representing molecules as character strings is helpful to use tools from natural language processing, such as transformers and generative models that use text sequences as inputs. Molecular fingerprints form a family of representations of molecules into vectors. In particular, circular molecular fingerprints capture local features around atoms, such as topology, atom types, bond types, and connectivity patterns within a specified radius. This type of representation is mostly used to predict properties associated with reactions, such as reaction template[7] or molecular similarity.[12] A molecule has a natural graph structure that considers atoms to be nodes and bonds to be edges. Consequently, molecular graphs have also been extensively used as inputs in models such as graph neural networks.[11,13] Other less common representations include SELFIES,[14] molecular signatures,[15] and atom environments[16] that focus on local information to characterize molecules. Table 1 and Figure 3A provide a comprehensive overview of various molecular representations, highlighting the proportion of their utilization in single-step retrosynthesis. Surveys[17,18] are available for more details about the different types of molecular representations, their respective advantages, and limitations.

Template-Based. Template-based single-step methods require templates that are derivatized either by human experts or extracted automatically from reaction databases in the form of reactants and major products. Template reactions, sometimes referred to as reaction rules or generalized reactions, are usually represented by atom-mapped SMARTS strings which can handle stereochemistry.[21] Examples of templates used in chemistry are found in Szymkuć et al.[61] and in biocatalysis and synthetic biology in Finnigan et al.[62] and the RetroRules database.[63] Template-based methods select templates that can be applied to a given product. Then, the main challenge is to apply the best template to the product to obtain the reactants.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

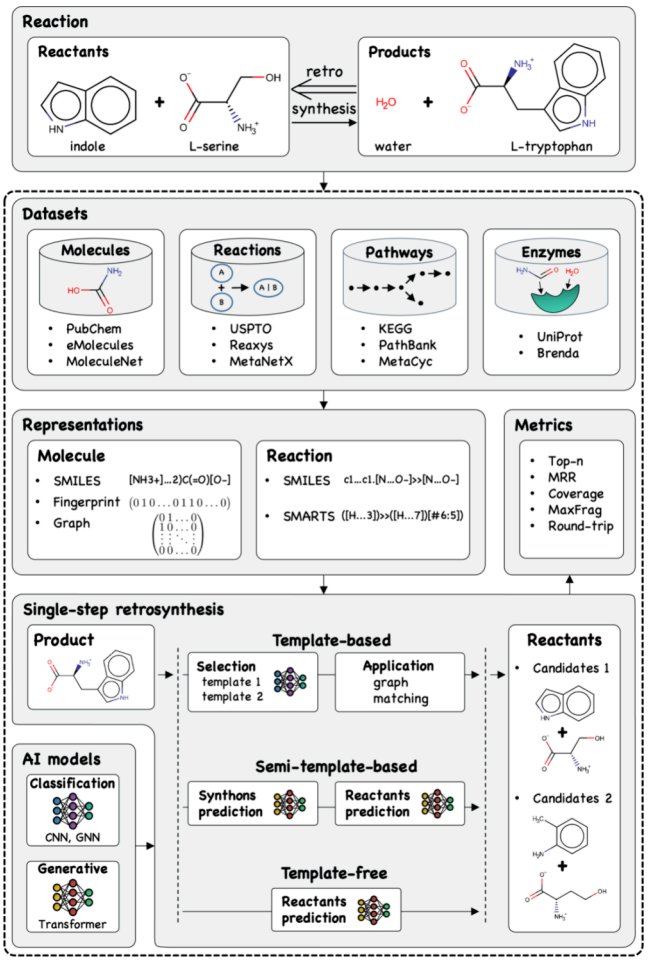


Figure 2. General principle of single-step retrosynthesis is represented by the example of a L-serine hydro-lyase reaction (indole + L-serine → water + L-tryptophan). To implement and train a single-step retrosynthesis model, databases of molecules, reactions, pathways, and enzymes are required. A molecular representation is then used as input to the retrosynthesis model. Then, either a template-based, semitemplate-based, or template-free method is chosen with an AI model used to perform the retrosynthesis. Finally, metrics, detailed in the single-step models evaluation section, measure the performance of the framework. CNN, convolutional neural network; GNN, graph neural network; and MRR: mean reciprocal rank.

Neural networks (NN) are used to learn patterns from molecular functional groups or fingerprints to select templates belonging to the same type of reaction rules for the products or substrates.[54] This strategy was applied to both reaction prediction and retrosynthesis tasks using deep highway networks[26] and Hopfield networks.[47] NN inputs are fed with the SMILES notation of the product, molecular fingerprints, or both.[29] Another strategy is based on graph neural networks (GNN) which use the graph structure of the molecule to select reaction templates, providing explainability through the model.[55,64] To this end, reaction templates are embedded in a conditional graphical model built using GNN[56] or partially encoded using the reaction center.[57]

Template-based methods have the advantage of mimicking the bond rearrangements of existing reactions. However, these models suffer from several limitations, such as the duplication and overlap of templates describing the same chemical transformation in databases. Some studies have been conducted to optimize and palliate these limitations through canonicalizing templates[11] or using dedicated NN models.[7] Among other limitations, template-based methods infer reactions derived from template databases but cannot suggest new mechanisms. To alleviate this inconvenience, Yan et al.[36] employed a template composition strategy to create new reactions and templates.

In retro-biosynthesis, the point of view is slightly different. Molecules may be more complex than those usually considered for retrosynthesis in chemistry,[29] and reactions involve enzymes which are highly specific. To manage this situation, strategies such as identifying disconnections in ring systems[52] or evaluating the similarity between products and substrates have been suggested for selecting templates in both chemistry[12] and retro-biosynthesis with the aim of retaining enzymatic activity.[40,43] To the best of our knowledge, AI models have not been coupled to template-based single-step methods in the context of synthetic biology.

Template-Free. Template-free methods avoid relying on reaction templates and use generative models to directly predict reactants. To carry out this task, various models such as encoder-decoder built with long short-term memory cells[33,49] and Weisfeiler-Lehman networks were first employed.[13] More recently, deep generative models such as transformers have extensively been used and improved upon. Transformer models are a type of NN architecture composed of an encoder and a decoder, particularly suited for natural language processing. Sequences of tokens, which represent single or multiple characters, are processed by the encoder to generate a set of hidden representations and decoded by the decoder part to generate the output sequence. This model embeds an attention mechanism to focus on the most informative parts of the sequence. Thus, the translation task from the SMILES product to the SMILES reactants has been conducted using a model built for forward prediction[31,65] and by integrating a reaction prediction model into a retrosynthesis process.[66]

Table 1. List of Molecular Representations Used in Single-Step Retrosynthesis

Representation	Description	Related works
SMILES (canonical, rooted, ...)	String representation of a molecule	8, 10, and 19−46
Circular fingerprints (ECFP, FCFP, HSFP, ...)	Vectorization of a molecule by hashing local molecular features	7, 11, 12, 25, 26, 29, 34, 43, and 46−56
Molecular graph	Graph representation of a molecule where atoms are nodes and bonds are edges	8, 9, 11, 13, 19, 24, 28, 36, 41, 42, 48, and 55−59
Atom environment	Circular atom-centered topological neighborhood fragments of a molecule in SMARTS format	46 and 50
Signature descriptor	Representation of a molecule encoding all atom environments in its molecular graph up to a predefined radius in SMILES format	60
SELFIES	Extended SMILES string representation of a molecule	46

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

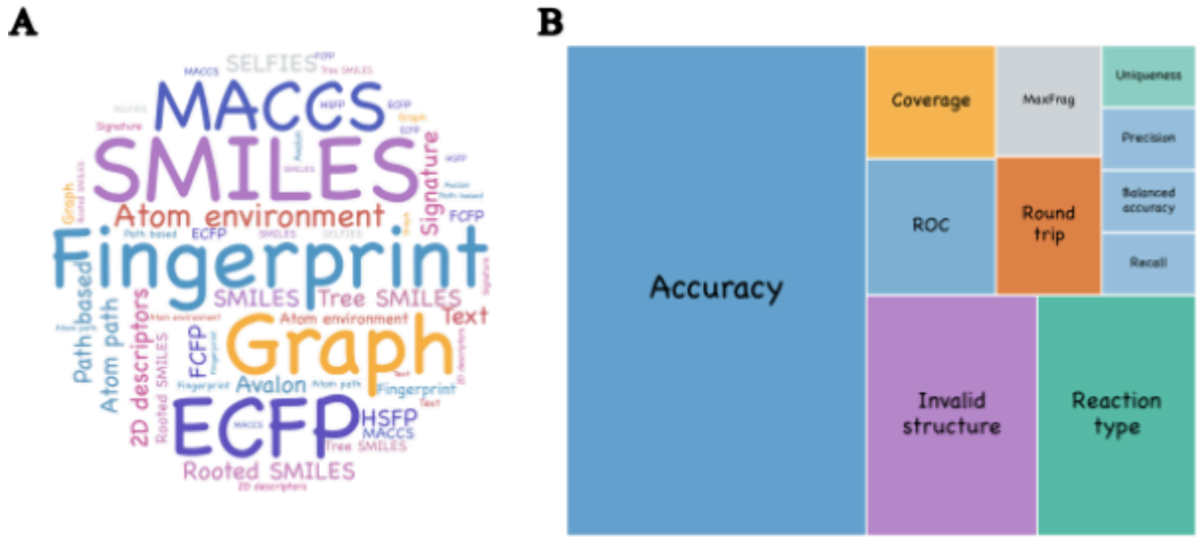


Figure 3. Analysis of the use of molecular representations and metrics by single-step algorithms. (A) Word cloud of types of molecular representations used as input in single-step AI models. (B) Proportions of different metrics used to evaluate models for single-step retro(-bio)synthesis. These metrics are further described in the single-step models evaluation section. Metrics mentioned only once in articles are not included in the analysis. Blue squares and green squares reflect the usage of metrics quantifying observational error and the diversity of reaction types, respectively. MaxFrag, maximum fragment and ROC, receiver operating characteristic.

Instead of using SMILES representations as input, transformer models have considered other types of representations of molecules and combine them. For example, models such as Graph2SMILES,[67] BiG2S,[68] DVMP,[28] graph enhanced transformer,[24] graph truncated attention,[41] and G2GT[59] combine graph representations of molecules with SMILES sequence representations. The transformer model retroformer[8] associates local and global attention heads to identify the reactive region in molecules. Another possibility is to customize the transformer model with a tree representation of the SMILES,[35] use a set of predefined compounds,[58] or add information to the reaction using byproducts[39] or a reaction graph.[69] Transformer architectures have also been used to predict atom environments of the reactants knowing the products[50] and to translate the reactants back.[46]

During the reconstruction of the representation of the molecule, transformer models are liable to add or omit characters, therefore producing grammatical errors in the SMILES notation. To alleviate this inconvenience, a neural-network-based syntax checker called the SCROP[30] framework was plugged into the transformer architecture, leading to a decrease in the number of invalid outputs. In order to increase the diversity of predictions, avoid invalid SMILES,[20] and predict the type of the reaction,[42] a latent variable has been integrated into the model. Moreover, different strategies have been employed to increase prediction diversity. Adding reaction types to SMILES strings provides additional context to the transformer model, leading to an increase in the diversity of suggestions.[45] Achieving this objective without altering the data set, predictions were enhanced by employing a GNN[19] or an energy-based model[70,71] to rerank predictions that the single-step model deemed less confident. Reranking methods harness additional chemical feature information from molecular graphs to refine the results.

Some methods have been tested to improve learning capacities of models, such as data augmentation and transfer learning. Data augmentation enriches data sets by applying diverse transformations on the original data SMILES. These transformations are basic, such as the swapping of reactants and products,[38] or more elaborate, such as the SMILES enumeration method, which randomly selects a starting atom to generate different SMILES for the same molecule.[27,44] On the other hand, transfer learning is a technique in which a pretrained model is used as a starting point for a new task to take advantage of the knowledge and features learned by the pretrained model. Learning on a data set of 380 K molecules, a transfer done on USPTO-50K[25,37] and on a data set of around 2200 Baeyer−Villiger reactions,[23] demonstrated better performance than on the data set alone. In contrast, multitask transformers trained on text and molecular representations show interesting results without the need for transfer learning.[72]

To the best of our knowledge, template-free methods have not yet been used in the context of synthetic biology, while few methods are available for biocatalysis. BioNavi-NP[73] employed a transformer with a data set representing natural products to predict biosynthesis pathways. In the context of reaction predictions, Kreutter et al.[74] added a textual representation of enzymes to SMILES, whereas in retro-biosynthesis, Probst et al.[75] used EC numbers to enrich SMILES, thus predicting both molecules and the enzymes, and the EC number respectively associated with the reactions.

Semitemplate-Based. Semitemplate-based methods aim to imitate the reasoning of chemists. First, the reaction site within the product structure is detected, followed by the disconnection of bonds to create intermediate molecules known as synthons. Substrates are then retrieved from the synthons. Unlike template-based methods, this approach does not rely on a database of predefined chemical reaction templates.

After producing synthons, reactants are predicted with a generative model such as the graph-to-graph translation models found in G2Gs,[76] a transformer model such as RetroXpert,[77] RetroPrime,[32] and RetroSub,[34] or based on a precomputed vocabulary using GraphRetro.[78] Another possibility is to optimize the molecular input; for instance, the hot-

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Table 2. Single-Step Retro-Biosynthesis Methods

Single-step	Framework	Description	Focused on stereochemistry	Enzyme information	Data set
Template-based	EHreact[40]	Templates are stored in a Hasse diagram and ranked according to their similarity	Yes	Name, EC number	Brenda, RetroRules, and Rhea
	retrosim_enz[43]	Template ranking based on similarity and evolution scoring	Yes	Name	Rhea
	RingBreaker[52]	Template ranking based on synthesis of ring systems	Yes	No	Reaxys and USPTO
Template-free	BioNavi-NP[73]	Transformer trained on organic and biosynthetic reactions	Yes	No	MetaNetX and USPTO
	Kreutter et al.[74]	Enrich SMILES with enzyme name using Transformer (forward prediction)	Yes	Name	Reaxys and USPTO
	Probst et al.[75]	Enrich SMILES with EC number using Transformer	Yes	EC number	Brenda, MetaNetX, PathBank, and Rhea
Semitemplate-based	Thakkar et al.[80]	Enrich SMILES with EC number using a prompt-based method	No	EC number	Pistachio and USPTO

spot fingerprint (HSFP) has been employed by Hasic et al.[53] to identify the reaction site and generate synthons. Additionally, RetroExplainer[79] and Graph2Edits[9] frameworks rely on a set of actions, such as deleting bonds or attaching groups of atoms to retrieve reactants.

Interestingly, a semitemplate method has been developed in the context of biocatalysis, using a prompt-based paradigm that enables the inclusion of additional information into the inputted SMILES, such as the EC number of the reactions.[80] To the best of our knowledge, semitemplate-methods have not yet been used in the context of synthetic biology.

Single-Step Models Evaluation. When comparing methods, it is important to establish standards for measuring their effectiveness. Using a metric facilitates fair and consistent comparisons, helping in the identification of the strengths and weaknesses of each model. Figure 3B shows the adoption of different metrics by the community.

Many of the models utilized for single-step prediction suggest multiple candidates, where "candidates" refer to either a collection of reaction templates or some sets of molecules envisioned as credible reactants. Accuracy metrics indicate the proximity of predictions to the true values and are typically evaluated using the top-n accuracy. While the top-n metric is widely used for reaction prediction, its relevance has been questioned,[22] as many molecules can be built from more than one set of reactants, i.e., there are several "true" answers for a given product. Less common metrics in this context include fractional accuracy,[35] balanced accuracy,[26,55] weighted precision,[54] and ROC curve.[38,43]

Metrics have also been used to assess the quality of ranking. The mean reciprocal rank (MRR) calculates the average rank of the first relevant prediction across multiple runs.[54] An alternative, called the coverage, counts the number of samples for which one or more valid predictions are made.[20,22,42,45] It is crucial to evaluate reactants belonging to multiple types of reactions to obtain a wide range of available synthetic routes. The class diversity metric measures the range of reaction types predicted by the single-step models,[22,32,45] while the Jensen−Shannon divergence quantifies the similarity between likelihood distributions of predicted reactions belonging within a fixed number of reaction types.[22] Moreover, rather than evaluating richness, the presence of repeated predictions, which indicates a lack of variety produced by the models, has been estimated in Kim et al.[20] and in Yan et al.[39]

Other ad-hoc metrics have been developed specifically for one-step retrosynthesis. The round-trip accuracy reflects the validity of retrosynthetic suggestions with a forward transformer that predicts the product molecule from the predicted precursor,[22] but it has been shown that this transformer can produce bias.[81] Due to the nature of the template-free algorithms, some predictions are grammatically invalid. More than one-third of the research papers utilizing template-free methods quantify the occurrence of invalid SMILES (Figure 3B). Finally, in an original approach specific to semitemplate-based methods, a dedicated metric was employed to measure the success of disconnection.[80]

In retro(-bio)synthesis, one or more reactants are involved in a reaction, and the largest molecule better reflects the reaction type and is more susceptible to having an important reaction site, avoiding unambiguous reactions. To that end, the maximum fragment (MaxFrag) accuracy was applied to the predictions[39,44] to consider the similarity between molecules to reflect their bioactivity.[35]

Perspectives in Retro-Biosynthesis. Although many approaches have been developed for retrosynthesis, only a few have been specifically designed for retro-biosynthesis. These methods are summarized in Table 2. Indeed, one of the prerequisites for a reaction to be biocatalyzed is, at the very least, to determine if the reaction could be catalyzed by an enzyme.[80] Additionally, enzymatic stability is achieved at narrow ranges of temperature, pH, and pressure values. It is therefore essential to accurately characterize these reaction elements to ensure the feasibility of reactions.[82] The reaction solvent also plays a crucial role in enhancing reaction yield and in making chemical reactions more sustainable. Considering all these factors could lead to the development of a dedicated score for retro-biosynthesis to compare the suggestions made by the algorithm. Before opting for a retro-biosynthesis algorithm, users must ascertain whether the results obtained are highly exploratory and meant for theoretical contemplation or if they are intended for practical application. Indeed, if the user is open to novel reactions predicting a hypothetical substrate, the use of generative models is suitable.[31] Conversely, employing SELFIES enables the generation of reactions not present in the data set, while ensuring the production of coherent molecules.[46] However, for the *in vivo* implementation of reactions, it is recommended to utilize known reaction mechanisms and therefore prefer template-based methods. We believe that the numerous approaches developed in retrosynthesis can serve as a source of inspiration to adapt them for retro-biosynthesis.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

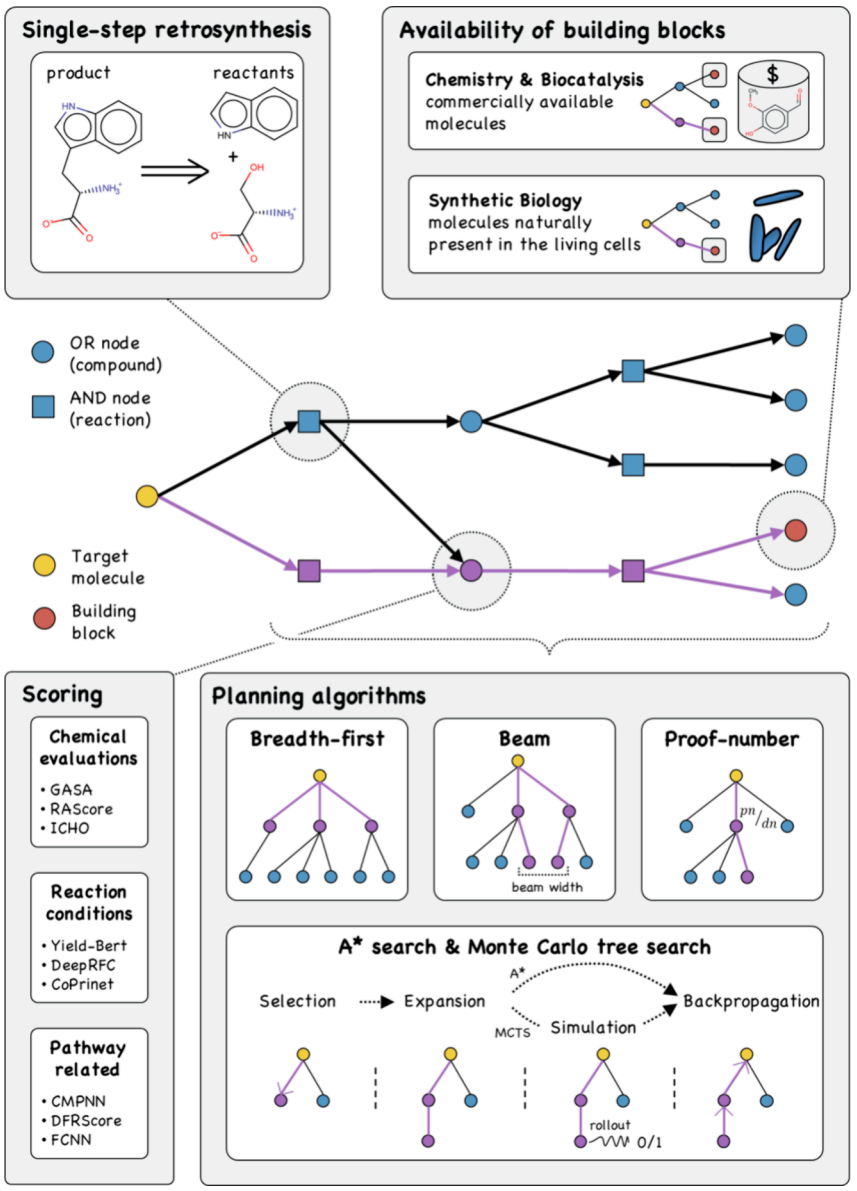


Figure 4. General principle of multistep retrosynthesis is represented as an AND/OR tree where circles represent molecules with an OR node, because multiple reactions can synthesize this product, and squares represent reactions with an AND node, indicating that all reactants are necessary to produce the product. Starting from a target molecule (yellow ●), single-step retrosynthesis is used at each step of the retrosynthesis to reach building blocks (red ●). In chemistry and biocatalysis, the building blocks are commercially available molecules, while in synthetic biology, the building blocks are molecules present in living cells. A scoring function, as explained in the next section, and a search algorithm guide the navigation over synthetic possibilities during retrosynthetic planning due to the vast search space of possible synthesis routes. For simplicity, the planning algorithms are explained with graphs instead AND/OR trees.

■ MULTISTEP RETROSYNTHESIS

Multistep retrosynthesis, as introduced in Corey's seminal work, is a strategic framework for synthesizing complex molecules by reversing known chemical reactions to break down a target molecule into simpler predecessors. The process involves multiple steps, varying in number based on the molecule's complexity, available reactions, and starting materials. The ultimate aim is to develop efficient synthetic routes using these building blocks, employing available reactions for practical implementation. Originally focused on simplifying molecules, the concept has evolved to accommodate the synthesis of intermediates that may be more complex than the target, as seen in natural biosynthesis processes involving complex structures enhanced by cofactors like phosphorylated intermediates or coenzymes such as coenzyme A. The principles of multistep retrosynthesis now also apply to biocatalysis and synthetic biology, leveraging biological databases to guide reaction selection.

While algorithms suggesting viable single-step retrosynthesis pathways provide a critical foundation, forecasting multistep retrosynthesis pathways introduces substantial challenges. The complexity arises from the enormous potential synthesis routes and the subjective nature of determining a "good" synthesis route.

Chemists and biochemists grapple with these complexities, faced with a broad spectrum of potential intermediates and differing views on what constitutes an optimal retrosynthetic pathway. Key components of a comprehensive multistep

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

retrosynthesis search, namely the planning algorithm, the selection of starting materials, and the single-step reaction predictor, are illustrated in Figure 4.

Availability of Building Blocks. The set of available building blocks, sometimes referred to as precursors or sinks, that could be used as starting material is a critical element influencing an algorithm's capabilities to predict synthetic routes. Indeed, it is intuitive that a more diverse and extensive collection of building blocks will provide a broader "landing pad" for algorithms for ending retrosynthesis explorations.

In chemistry and biocatalysis, the building blocks often consist of chemicals commercially available, as in Probst et al.,[75] indexed from online portals like eMolecules or chemical providers such as the Sigma-Aldrich catalog. Additional information available, such as the chemical price, is eventually taken into account for guiding and ranking the retrosynthesis planning, as in Zhang et al.[83]

The planning of biosynthetic pathways within living organisms poses a challenge in selecting exploitable building blocks because the entry of molecules into cells is highly selective. The building blocks are more specific and usually encompass sets of molecules naturally present in organisms, as in Koch et al.,[84] where the available precursors were extracted from a genome-scale metabolic model of the *Escherichia coli* bacterium.

Planning and Search Algorithms. Predicting potential synthesis routes relies heavily on search algorithms that navigate through various possibilities within a chemical space, as shown in Figure 4. These algorithms are crucial for effectively mapping out multistep pathways. Broadly, these search algorithms fall into two categories: uninformed searches and informed searches. Uninformed searches, such as depth-first and breadth-first searches, operate without relying on additional information to guide the exploration toward a specific area of the solution space. Conversely, informed searches incorporate heuristic functions that assess "how good" compounds are to be expanded. These heuristics may solely focus on the compounds discovered thus far, as in the beam searches mentioned in this review. Alternatively, they also estimate the proximity to a solution using a method such as rollout simulations in Monte Carlo Tree Search (MCTS) or value function estimators, as in A*-related algorithms. Overall, these heuristics guide the search process, ensuring a more efficient exploration. While informed search approaches do not guarantee an optimal solution, they significantly enhance the likelihood of finding a good solution within a reasonable time frame, balancing solution quality with search efficiency. In essence, these algorithms play a pivotal role in optimizing retrosynthesis planning by navigating through complex possibilities to derive viable synthetic routes. The retrosynthesis graph is generally represented as an AND/OR tree, where an OR node corresponds to a molecule since several reactions are possible to synthesize this product, while an AND node corresponds to a reaction, since all the reactants are required to synthesize the product. This graph could also be considered as a hypergraph where an edge can connect several nodes, thus representing the link between products and substrates.[22] We now briefly present the main types of search algorithms recently used for multistep retro(-bio)synthesis.

Breadth-First Search. The breadth-first search algorithm is a textbook case for graph traversal algorithms. It explores a graph by visiting all its neighbor nodes at the present depth before moving on to nodes at the next depth level. This process continues until all reachable nodes have been visited or, in a pragmatic retrosynthesis context, until a certain depth is reached. Although being quite slow, this algorithm is useful for finding short routes. In the domain of retro(-bio)synthesis, a filter is classically applied between each level expansion, aiming to exclude nodes that are unlikely to participate in plausible solutions, thereby limiting combinatorics. Breadth-first search has been used in chemistry to predict thermodynamically feasible pathways, as in the analytic tool RetroSynX.[85] Here, reactions that are unlikely to occur are filtered out before advancing to the next depth level. Similarly, in the field of biocatalysis, Liu et al.[86] employ thermodynamic estimations as a filtering criterion to mitigate combinatorial explosion during breadth-first searches.

Beam Search. Beam search is nowadays an algorithm used in many AI generative methods. It constructs a collection of possible routes in a breadth-first manner, but it imposes a predefined limit on the number of nodes to be expanded at each depth level, named the beam width, selecting nodes to keep from heuristic evaluations. Often presented as an enhancement of the breadth-first algorithm, beam-search offers improved efficiency in finding solutions, especially in large solution spaces. This algorithm has been used in chemistry, such as in the work of Schwaller et al.[22] where promising chemicals are chosen for further expansion based on a synthetic complexity score, SCScore, and generative single-step log-probabilities. In addition to these metrics, Kreutter et al.[87] incorporated a route penalty score, RPScore, in the heuristic evaluations. In retro-biosynthesis, the classification and availability of enzymes are crucial factors. For example, Probst et al.[75] rely on scores that consider EC number annotations and SCScore to select chemicals for expansion. Conversely, in synthetic biology, the RetroPath2.0[60] software integrates the beam search with RetroRules scores to prioritize reactants associated with high confidence in enzyme availability.

Depth-First Proof Number Search (DFPN). The proof number search algorithm is a tree search method primarily used in game tree solving. It evaluates game positions by assigning proof numbers (values indicating a win) and disproof numbers (values indicating a loss) to nodes in the game tree. DFPN is a variant that conducts a depth-first search while updating these numbers, aiming to prove or disprove the existence of a forced win within a specific depth limit. It prunes branches of the search tree based on these proof and disproof numbers to focus on the most promising lines of play. This search algorithm is used in chemistry in DFPN-E,[88] where authors couple DFPN with a heuristic edge initialization method to address the imbalance between the numbers of OR versus AND moves, later improved to output multiple solutions.[89] The CompRet framework[90] proposes a comprehensive tool to enumerate and rank possible routes to synthesize compounds, using metrics derived from the SCScore to recommend most promising routes.

Monte Carlo Tree Search (MCTS). Monte Carlo tree search is an informed heuristic search algorithm often used in decision-making processes. It builds a search tree by repeatedly simulating random sequences of moves, which are single-step transformations in the context of retrosynthesis, from selected leaves (e.g., compounds) of the tree. The algorithm prioritizes exploring promising paths by balancing between exploitation (focusing on most promising nodes) and exploration (focusing on alternative potential routes) to make informed decisions.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Within the past few years, this algorithm has been widely used for retrosynthetic route planning in chemistry.[66,91−97] Segler et al.[91] pioneered MCTS for retrosynthesis by combining rule-based single-step transformations and three NNs to assist the exploration of the chemical space. In this study, AI is notably utilized to preselect templates to apply, limiting combinatorial explosion while directing searches toward the most plausible routes. This advantageous integration of AI has been implemented in several other works combining NNs and template-based approaches in MCTS, such as ASKCOS[98] and AiZynthFinder,[96] as well as the work of Zhang et al.,[92] where authors propose to combine efficiently five GNNs.[92] The data source for building templates is an important factor, as highlighted by Thakkar et al.,[94] wherein the impact of four data sets (AiZynthFinder, Pistachio, Reaxys, and USPTO) on MCTS performance is investigated. Furthermore, the configuration of the model significantly influences the success of route discovery.[99] Template reactions model a finite number of transformations, which may lack exhaustiveness. As an alternative to template-based formalism, template-free single-step approaches have also been employed in MCTS. For instance, Lin et al. implement single-steps using Transformers in the AutoSynRoute.[95] Studies falling within the field of biocatalysis are by far fewer. However, reuses of ASKCOS have been proposed to exploit template-based transformations coming from both chemical and metabolic reaction databases.[100,101] As far as we know, RetroPath RL[84] stands as the sole implementation of the MCTS algorithm for synthetic biology where rule-based transformations have been extracted from metabolic databases and the list of available building blocks extracted from genome-scale metabolic models.

A* Search. A* search is an informed graph traversal algorithm used for pathfinding and optimization in graphs or search problems. It determines the priority of nodes for exploration by considering both the cost-so-far (i.e., historical cost from starting node) and an estimated cost-to-go (i.e., future cost to reach a goal node). By utilizing this evaluation function, occasionally designated as the value function, A* proposes to find the optimal path from the start to the goal node while minimizing the total cost. Compared to MCTS, this algorithm does not have a rollout phase and therefore does not depend on randomness and is faster. This search algorithm gained in popularity in the last years and has been used in several chemical planning software such as ASICS,[102] Retro*,[103] RetroGraph,[104] GNN-Retro[105] for chemistry application and BioNavi-NP[73] for biosynthetic pathways predictions. While the A* algorithm serves as the multistep engine for guiding route discovery, it can be combined with different flavors of single-step moves, such as template-based transformations as in Retro* where a NN selects a template to be applied depending on the product molecule, or template-free moves as in BioNavi-NP, which relies on a Transformer for predicting the reactants given the product. The A* search is one variant of best-first search algorithms. Other nondata-driven AI-based approaches have been developed, such as greedy best-first search, that prioritize exploration based solely on the cost incurred so far (historical cost) without attempting to predict the future cost to reach a goal node. One example is SynRoute, where authors demonstrated this greedy approach to be the most effective among four planning algorithms tested.[106] Another prime example is Synthia[107] (previously known as Chematica), a well-known commercial software for synthesis planning. This expert system utilizes a comprehensive set of rules for reaction-template selection and application, alongside a dual scoring function for choosing chemicals for the next retrosynthesis iteration. While not described as such in the literature, Synthia's exploration strategy can be classified as a best-first search.[61]

Other Reinforcement Learning Related Search. Similar to MCTS and A* search algorithms, other types of reinforcement learning methods can be found in the literature. Through iterative exploration and learning, the computer's agent powered by NN[108] refines its decision-making process, effectively navigating the graph of possible reactions to identify the most efficient pathways for retrosynthesis. This approach has been used in chemistry[109,110] and in biocatalysis.[83]

Table 3 summarizes various search algorithms used for multistep retrosynthesis, along with their respective scope of application in chemistry, biocatalysis, and synthetic biology. Within retro-biosynthesis, it is important to acknowledge the existence of foundational methods like novoStoic,[111] XTMS,[112] or BNICE.ch.[113] These methods offer distinct approaches to metabolic pathway design, diverging from the AI-based algorithms discussed in this review. NovoStoic utilizes a template-based reaction formalization within a stoichiometric modeling framework, coupled with Mixed Integer Linear Programming (MILP) to efficiently enumerate metabolic pathways. XTMS constructs a retrosynthetic network using template-based reactions named *reaction signatures*. This network then serves as a starting point to extract all possible biosynthetic pathways connecting a preset of known compounds to the *E. coli* chassis organism. Similar to XTMS, BNICE.ch uses template-based reactions to build a comprehensive biochemical network, referred to as ATLAS,[114] which is then inspected to enumerate in an exhaustive manner linear routes (i.e., no branched pathways). However, in both approaches, the prebuilt networks limit users from exploring entirely new compounds, hindering generalizability. Readers seeking more details on these techniques can find comprehensive reviews elsewhere.[2,115]

Perspectives in Retro-Biosynthesis. Various algorithms have been developed, yet none has demonstrated superiority in terms of results output. Depending on the application, it is important to pay attention to model parametrization for algorithms such as MCTS[99] or A* and to define an evaluation function that is tailored to the data.[73] Also, reducing the production cost of a molecule is often a goal in the search for new synthesis pathways.[117] Consequently, the cost of building blocks and, more broadly, atom economy become significant criteria in the quest for innovative synthesis routes.[83] In this context, the use of enzymes that do not depend on cofactors is particularly appealing. For *in vitro* applications, although the addition of cofactors can regulate catalytic activity, this represents a cost factor that must be considered. To reduce this cost, cofactor recycling that utilizes photosensitization, electrochemical activation,[118] or the creation of enzymatic cascades are promising strategies. To the best of our knowledge, retro-biosynthesis tools do not currently support the generation of pathways recycling cofactors. For *in vivo* applications, although implementing molecular cascades in a cellular host allows for the use of compounds that would be unstable if isolated,[119] the use of cofactors can lead to competition between cell growth and the production of the desired chemical species, which are not considered by retro-biosynthesis tools.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Table 3. Search Algorithms Commonly Used in Multi-Step Retrosynthesis[a]

Multi-step	Scope of application	Framework	Important feature(s)/Highlights	Building block source	Single-step	Code availability
Breadth-first	Chemistry	RetroSynX[85]	Intermediate chemicals filtered using thermodynamics estimation	aladdin-e.com	Template-based	No
	Biocatalysis	Liu et al.[86]	Intermediate chemicals filtered using thermodynamics estimation	aladdin-e.com	Template-based	No
Beam	Chemistry	Schwaller et al.[22]	Most promising chemicals are selected for further expansion based on SCScore and generative single-step log-probabilities	eMolecules	Template-free	No
		Kreutter et al.[87]	Further extends beam selection from Schwaller et al.[22] by using RPScore	Enamine, Molport		Yes
	Biocatalysis	Probst et al.[75]	Beam selection considering enzyme classification and SCScore	eMolecules	Template-free	Yes
	Synthetic Biology	RetroPath2.0[60]	Beam selection based on enzyme availability estimations	Metabolic model	Template-based	Yes
DF Proof Number	Chemistry	CompRet[90]	Proof and disproof numbers according to reaching of building blocks	Enamine	Template-based	Yes
		DFPN-E[88]	Use of an estimator to assess the difficulty of finding a proof + attainment of building blocks	USPTO		No
MCTS	Chemistry	Gao et al.[97]	Two NNs used for template selection and reaction filtering	NA	Template-based	Yes
		Segler et al.[91]	3N-MCTS method: three NNs in use for template selection (expansion), feasibility, and rollouts	AlfaAesar, Acros, Reaxys, Sigma-Aldrich, ZINC		No
		AiZynthFinder[96]	NN guided template selection	ZINC		Yes
		ASKCOS[98]	Two NNs for template selection and reaction filtering	eMolecules, Sigma-Aldrich		Yes
		Wang et al.[93]	Reinforcement learning network used instead of the MCTS rollout step	eMolecules, Sigma-Aldrich		No
		Zhang et al.[92]	Five GNNs used for selecting templates, infer reaction solvent and catalyzer, filter reaction, and efficiently evaluate the rollout step	Molport		Yes
	Chemistry	AutoSynRoute[95]	Rollouts guided by a heuristic based on log probabilities from Transformer output	Sigma-Aldrich, USPTO, and ZINC	template-free	Yes
	Biocatalysis	Sankaranarayanan et al.[100]	Two NNs for template selection and reaction filtering applied to both	eMolecules, LabNetwork, Sigma-Aldrich	Template-based	No
		Levin et al.[101]	ASKCOS reimplementation using NN for template prioritization and balancing between chemical vs enzymatic templates	eMolecules, Sigma-Aldrich		Yes
	Synthetic Biology	RetroPath RL[84]	Template selection according to reaction feasibility and enzyme confidence scores	Metabolic model	Template-based	Yes
A*	Chemistry	Retro*[103]	Template selection using NN, cost of current paths estimated from the cost of current reactions, cost of future paths learned from NN trained from knowledge database	eMolecules	Template-based	Yes
		Retro*+[116]	Template selection is coupled during the search with the actual already predicted reactions, using self-improving learning	eMolecules		No
		RetroGraph[104]	Cost of future routes estimated using an offline trained GNN, multitarget search using a graph search instead of a tree	eMolecules		No
		GNN-Retro[105]	Cost of future routes estimated using an offline trained GNN	NA		No
		ASICS[102]	Combines known reactions extracted from knowledge databases with template-based prediction, SAScore is used to estimate the cost of route toward the goal	eMolecules		Yes
	Biocatalysis	BioNavi-NP[73]	Combination of chemical and biochemical data sets with the use of transfer learning to set up the single-step transformer, cost of future estimated with a NN	Custom list, main precursors of natural products	Template-free	Yes
Best-first	Chemistry	SynRoute[106]	NN used to evaluate a feasibility score of predicted transformations	eMolecules	Template-based	No
	Chemistry	Synthia[61]	Chemical and reaction function scores are combined for selecting the next graph expansion	Sigma-Aldrich	Template-based	No
	Biocatalysis	RetroBioCat[62]	SCScore used to guide the best-first search	eMolecules, Molport, ZINC	Template-based	Yes

Table 3. continued

Multi-step	Scope of application	Framework	Important feature(s)/Highlights	Building block source	Single-step	Code availability
Reinforcement learning	Chemistry	Schreck et al.[109]	Improvement of node selection (policy) by training a NN using simulated experience	eMolecules, Sigma-Aldrich, LabNetwork	Template-based	Yes
		GRASP[110]	GRASP follows an MCTS-like approach, where the vanilla online roll-out is replaced by the reinforcement learning agent	eMolecules	Template-free	No
	Biocatalysis	Zhang et al.[83]	A "decision maker" NN is trained to bias exploration, with a random component that promotes exploration of dissimilar routes	ChemSpace	Known reactions	No

[a]NA, not applicable.

SCORING FUNCTION

Selecting the most promising reactions and pathways to synthesize the target molecule is a crucial component for guiding the retrosynthesis planning process. The scoring functions outlined in Table 4 assist in navigating through the multiple synthetic possibilities encountered during retrosynthetic planning and route enumeration. These functions rely on various criteria, including factors such as chemical cost, structural considerations, and insights from enzyme knowledge.

Synthetic Accessibility Scores. Synthetic accessibility scores can discriminate feasible molecules from infeasible ones and are a helper retrosynthesis planning tool to identify viable synthetic routes from impractical ones.[120] The graph attention-based assessment of synthetic accessibility (GASA) score evaluates the synthetic accessibility of small molecules by labeling compounds as "easy" or "hard" to synthesize.[48] Contrary to relying solely on the structure of the molecule, RAScore[121] evaluates the feasibility of synthesis, incorporating reaction information into the assessment, and similar strategies predict the probability of finding molecules involved in a reaction included in a database.[51] Also, to surrogate synthesis accessibility, the number of steps required to produce a compound is estimated in chemistry,[122] in drug-relevant,[123] and in biological[124] applications or has been integrated into a composite score.[125] Moreover, the scores provide estimates for data concerning the compound, including its price[126] or its thermodynamic properties,[127] and the reaction, its yield,[128] or its feasibility.[129−132]

Routes Ranking. Regardless of the specific multistep algorithm employed, numerous retrosynthetic routes are typically generated, necessitating strategies to identify the most promising ones. To this end, route ranking strategies have been devised, integrating discriminative criteria related to chemicals (e.g., SAScore,[133] SCScore,[134] or cost of building blocks), to reactions (e.g., RAScore,[121] reaction yield and thermodynamics, or enzyme availability), and overall properties of route (e.g., route diversifiability, number of reaction steps, or theoretical production flux). For instance, the SynRoute[106] framework evaluates routes based on the length, cost of building blocks, and reaction yield estimations to select an optimal pathway. In biocatalysis, RetroBioCat prioritizes pathways by considering the number of reaction steps, change in chemical complexity, the proportion of commercially available chemicals, and the linkage of steps to literature references. Interestingly, a diversity score is added to penalize pathways making use of reactions already ranked in top routes.[62] Meanwhile, in synthetic biology, the Galaxy-SynBioCAD platform combines multiple criteria including enzyme availability, theoretical product flux (via Flux Balance Analysis[135]), reaction thermodynamics (via eQuilibrator[136]), and step count to train a classifier model for pathway scoring and ranking.[137] Additional notable efforts include a "route diversifiability" score that facilitates comparison of synthetic pathways[138] based on the potential production of analogous chemicals, and the RPScore,[87] which assesses routes by step count and molecular synthetic accessibility. More broadly, developments that predict Gibbs free energy (ΔG), such as eQuilibrator[136] and dGPredictor,[139] are valuable tools for both filtering single-step predictions and assessing the thermodynamic feasibility of overall pathways.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Table 4. List of Scores Used to Evaluate the Routes

Classification	Purpose	Scope of application	Score
Evaluate molecule	Structure availability	Chemistry	GASA[48] and SCScore[134]
	Reaction availability	Chemistry	RAscore[121] and ICHO[51]
Reaction condition	Experiments in continuous reactor	Chemistry	InFlow[129]
	Liquid−liquid extraction	Chemistry	ExtractionScore[132]
	Reaction yield	Chemistry	Yield-Bert[128]
	Enzyme availability	Biocatalysis	DeepRFC[145] and EHReact[40]
	Compound price	Chemistry	CoPriNet[126]
	Thermodynamic	Multidisciplinary	eQuilibrator,[136] dGPredictor,[139] and GC-NORM-based[127]
Pathway related	Predict the number of steps	Chemistry	CMPNN[122]
		Drug relevant application	RetroGNN[152] and DFRScore[123]
		Synthetic biology	FCNN[124]

Enzyme Search. Transitioning from synthesis planning to biocatalysis and synthetic biology implementations requires identifying catalysts for predicted reactions, a crucial step in biosynthesis where enzymes play a predominant role. While numerous methods for enzyme selection exist,[140] only a few specifically tackle the challenges posed by retro-biosynthesis. The primary task in this context is taking a reaction defined by the chemical structures of its reactants and products and outputting a collection of candidate amino acid sequences that could catalyze the reaction.

Enzyme searches typically fall into two main categories: (i) referencing reactions and their catalyzing enzymes in metabolic databases such as KEGG and MetaCyc, and (ii) addressing *de novo* reactions not found in these databases, which necessitates predictive algorithms. E-zyme,[140] Selenzyme,[141,142] and BridgIT[143] are three methods instrumental for retrieving enzyme sequences for *de novo* reactions. In brief, their strategies involve a two-step process: first, use the *de novo* "query" reaction to identify similar known reactions in a reference database; second, retrieve the sequences associated with the best-hit known reactions. Sequence-to-reaction associations depend on the reference database. E-zyme and BridgIT use the KEGG Ortholog and Reaction databases, whereas Selenzyme uses the BIOCHEM4J, which integrates KEGG and Rhea databases for linking sequences to reactions. Interestingly, Selenzyme extends its ranking of sequences by incorporating factors like phylogenetic distance and sequence properties such as solubility and transmembrane regions. Meanwhile, the RetroBioCat database[144] allows users to search for enzymes from among the RetroBioCat tool predictions. However, it does not support direct querying using specific reactions SMILES.

In addition to sequence retrieval systems, computational methods such as Deep-RFC[145] use AI to evaluate whether a chemical reaction, given its substrate and product structures, is likely to occur. Furthermore, the EnzRank[146] tool aims to predict enzyme activity using as input a reactant and an enzyme sequence, effectively aiding in the selection of suitable enzymes for *de novo* reactions. Other methods exist that focus on different aspects of enzyme characterization, such as predicting EC numbers,[143,147−151] further refining the selection process for appropriate enzymatic catalysts.

Perspectives in Retro-Biosynthesis. Pathway evaluation should include the thermodynamics of the reactions, the use of cofactors, the solubility of substrates, and the goals regarding the cost and sustainability of the building blocks. Estimating the viability of pathways within a cellular host ensures the reliability of predictions, but several elements must be considered, such as the choice of the host, the thermodynamics of all the reactions in the pathway, kinetic feasibility, or the presence and the accumulation of toxic compounds. It is also preferable to aim for the shortest possible pathway length and favor reactions that have already been characterized.[153] A significant number of sequential or tandem transformations in a one-pot process have been successfully carried out by combining enzyme-catalyzed reactions and metallic ions. However, it is crucial to ensure that all reactions within the pathways can share the same reaction conditions: solvent, temperature, pH, and that the catalysts can coexist with the enzymes. Indeed, some enzymes are less tolerant of the presence of metallic ions.[154] To enhance enzyme specificity, stability, and resistance, adaptation strategies can be pursued, and we refer the reader to these reviews for recent advances in enzyme engineering.[155,156]

DATA SETS

Common Databases. The use of AI for retrosynthesis relies on the quality of data and their diversity of information about molecules, reactions, pathways, and enzymes. In chemistry, the most popular benchmark data set of reactions is the US Patent and Trademark Office (USPTO) open-source database. It is composed of 3.7 million chemical reactions, and its subsets commonly used in AI models are the USPTO-50k, USPTO-full, and USPTO-MIT. Information about molecules and their properties is frequently extracted from the ChEMBL, MoleculeNet, and eMolecules data sets. Reactions have also been extracted from Reaxys and Pistachio. In biocatalysis and synthetic biology, biology databases of reactions include Rhea, RetroRules, and MetaNetX, which are open-source data sets. Databases of pathways such as the KEGG, MetaCyc, and PathBank databases are also used, along with databases of enzymes such as Brenda and UniProt. In Table 5, we summarize data sets used for retro(-bio)synthesis, their characteristics, and their scope of application in chemistry, biocatalysis, and synthetic biology.

Data Preparation. Each algorithm dedicated to retro-biosynthesis selects reactions from databases to create reaction templates or a data set feeding AI algorithms. This selection results from either manual curation[62] or automated extraction from one or more databases, applying specific filters to isolate relevant reactions and molecules. Some combined chemical and biological databases to increase the data set and to handle biochemistry reactions.[73,83] Initially, reactions are decomposed to isolate a single product per reaction.[73] Then, depending on their role in these reactions, some molecules are filtered out: molecules found as products in many reactions are identified as

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Table 5. List of Datasets Commonly Used in AI Applied to Retro(-bio)synthesis

Data set	Classification	Availability	Description	Scope of application
USPTO	Reactions	Public access (CC0 license)	A data set of organic reactions extracted from US Patent and Trademark Office-granted patents; it has been refined in several subsets, ranging from 50k reactions for the USPTO-50k data set to 1 M reactions for the USPTO-FULL data set	Chemistry[7−13,19−28,30−37,39,41,42,44,45,47,49,50,52,53,55−59,66−71,76−79,87,88,94−96,102−105,116,122,132,157−161] Biocatalysis[29,73,80]
Reaxys	Reactions	Commercial	Database that provides experimentally validated chemical data, including chemical structures, reactions, and properties	Chemistry[23,38,52,54,59,90,91,93,94,97,109,129] Biocatalysis[83]
Pistachio	Reactions	Commercial	Data set consisting of about 2,500,000 unique reactions	Chemistry[22,45,45,94,106,110,122,162] Biocatalysis[80]
Brenda	Enzymes, reactions	Public access (CC-BY license)	Database of enzyme catalyzed reactions, with information on alternative substrates, kinetic parameters, and protein sequences	Biocatalysis[40,75,146] Synthetic biology[111]
Rhea	Reactions	Public access (CC-BY license)	Database of biochemical reactions that uses the chemical ontology ChEBI covering enzymatic reactions and transport reactions	Biocatalysis[40,43,75] Synthetic Biology[111]
RetroRules	Reactions	Public access (CC-BY license)	Database of reaction template modeling enzyme catalyzed reactions for metabolic pathway discovery and metabolic engineering	Biocatalysis[40,145] Synthetic biology[84,124,137]
MetaNetX	Reactions	Public access (CC-BY license)	Collection of metabolites and biochemical pathways compiling more than 10 different biological databases (BiGG, ChEBI, Rhea, enviPath, HMDB, KEGG, MetaCyc, ...)	Biocatalysis[29,73,75] Synthetic Biology[60,84,137]
KEGG	Pathways	Public access with paywall for complete access	Comprehensive database integrating biological pathways, genomic, chemical, and disease information to facilitate understanding of biological systems and their functions	Biocatalysis[73,83,145] Synthetic biology[111,113]
PathBank	Pathways	Public access (open database license)	About 110,000 pathways found in 10 model organisms providing a pathway for every protein and a map for every metabolite	Biocatalysis[75]
MetaCyc	Pathways	Commercial	Metabolic pathways of about 3000 pathways and 19,000 reactions and metabolites across diverse life forms, encompassing primary and secondary metabolism	Biocatalysis[73] Synthetic biology[111]
UniProt	Enzymes	Public access (CC-BY license)	Database of protein sequences	Chemistry[163] Biocatalysis[43,146] Synthetic biology[60,84,137]

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

coproducts,[43] while cosubstrates, essential for enzymatic catalysis, are spotted from a predefined list and are partially[73] or entirely excluded.[75] Due to the enzymes' ability to be stereospecific, the use of stereochemistry is central in retro-biosynthesis.[40] Noninformative reactions for this process, such as transport reactions or those without substrates, are eliminated.[43] Finally, the way molecules and reactions are represented is tailored to the specific needs of the algorithm.

DISCUSSION AND OUTLOOK

While AI models in retrosynthesis and retro-biosynthesis have made significant strides, several challenges persist, requiring focused attention to enhance the models and effectively navigate the constraints inherent in their application in biocatalysis and synthetic biology. Below, we examine specific aspects, including molecular representations, model improvements, and evaluations, while also pointing out successful applications.

Single-step algorithms use multiple molecular representations or a combination thereof to gather complementary features from each representation. However, commonly used molecular representations exhibit notable limitations, such as the possibility of producing a SMILES which does not represent a molecule or certain fingerprints or atom environments that hinder accurate reconstruction.[18] Alternatives have been introduced to address these shortcomings, like SELFIES or molecular signatures,[60] and require further development or a wider adoption. One avenue yet to be further explored is to prevent retrosynthesis exploration toward intermediate chemicals that may have undesirable properties, as in RetroPath RL, where toxic chemicals are avoided during the multistep exploration. Since properties and activities are generally well predicted using fingerprints such as ECFP,[164] and the same fingerprints are used in many parts of the retrosynthesis process, including single-step,[7] multistep,[12] and scoring function,[51] one could envision developing methods to perform retro(-bio)synthesis starting not from a targeted molecule but from a targeted fingerprint.

As mentioned earlier, there is a clear need for advancing template-free and semitemplate methods tailored for retro-biosynthesis, deserving focused investigation. A promising direction for future research involves the use of multimodal models that handle different kinds of data, like prompt-based methods, enhancing prediction performance by integrating additional information, such as the EC number, in addition to the molecular representation.[80] While large language models have recently showcased impressive capabilities in natural language processing and have shown promise in aggregating various types of data and leveraging automation,[165] their performance in retrosynthesis still remains behind that of state-of-the-art models.[166]

The performance of AI models is significantly influenced by the availability and quality of data, while the creation of high-quality data sets is often both challenging and costly. The USPTO data set is widely used to train and evaluate single-step models. This data set, constructed from some patented syntheses not validated by experiments, suffers from unbalanced reaction classes, includes reactions with missing side products,[167] lacks reliable atom-mappings, and contains noisy stereochemical data.[168] Therefore, we believe that using data sets from well-maintained databases, although not perfect,[22] is a preferable practice. Nevertheless, retrosynthetic data sets are frequently segmented into distinct subsets based on specific attributes, such as USPTO-50k or USPTO-STEREO, complicating model comparison even when assessed using the top-n metric. Accordingly, the performance of several single-step methods was aggregated by the data set.[169] Moreover, efforts are underway to better organize existing data, particularly through the use of AI models.[170] Concurrently, initiatives to improve access to biocatalytic information[144,171] have emerged. Such long-term commitments should be encouraged and promoted in the context of retro-biosynthesis to build reliable data sets.

Retro-(bio)synthesis have proven to be valuable across several applications. In chemistry, it has been used to conduct lead optimization of drug molecules[172] and to synthesize alkaloid molecules[173] and natural products.[174] In the realm of synthetic biology, the Galaxy-SynBioCAD portal offers an all-in-one solution for designing metabolic pathways.[137] For instance, using RetroPath2.0, reactions were identified to produce lycopene in *E. coli* cells, and the pathway implementation was assessed *in vivo* using robotic equipment. The same platform has also been used to identify reactions crucial for producing biosensing intermediate molecules in cell-free systems.[175] Biocatalysis has been recognized as a method for advancing green and sustainable chemistry.[3] In this regard, the potential of biofoundries has been showcased to produce material monomers assisted by human expertise and retro-biosynthesis tools.[176] Utilizing retro-biosynthetic tools, Zhang et al.[177] successfully produced aliphatic diamines without hazardous hydrogen cyanide, and Liu et al.[178] engineered cells to produce 3-phenylpropanol, circumventing petroleum-based processes. Similarly, Yiakoumetti et al.[179] synthesized flavonoids, avoiding extraction from plant sources, and Brito et al.[180] leveraged methanol as a sustainable alternative for producing 5-aminovalerate molecules. Once the pathway to bioproduction of a molecule is established, subsequent optimizations of the implemented pathways using retro-biosynthesis tools could enhance production levels. For example, the production of eugenol by Hanko et al.[181] nearly tripled compared to previous reports when produced in small quantities.

In conclusion, this review extensively examines the latest advancements in AI-driven methods for both retrosynthesis and retro-biosynthesis paving the way for potential subsequent systematic reviews. The favorable outcomes observed in chemistry hold promise for its application in the field of retro-biosynthesis. As developments continue, we anticipate notable breakthroughs and increased incorporation of AI models in retro-biosynthesis, unlocking its full potential to catalyze innovation.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.4c00091.

(Note S1) Literature review, (Note S2) glossary of terms and definitions, (Figure S1) results of the literature search and selection process, and (Table S1) research queries utilized for selecting articles from academic search engines (PDF)

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

Jean-Loup Faulon − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France; The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, U.K.;* orcid.org/0000-0003-4274-2953; Email: jean-loup.faulon@inrae.fr

Authors

Guillaume Gricourt − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France;* orcid.org/0000-0003-0143-5535

Philippe Meyer − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France;* orcid.org/0000-0002-0618-2947

Thomas Duigou − *Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France;* orcid.org/0000-0002-2649-2950

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.4c00091

Author Contributions

G.G., P.M., T.D., and J.L.F. conceived the study. G.G. created the collection of papers. J.L.F. acquired the funding. G.G., P.M., and T.D. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-22-PEBB-0008. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

A*, A* search; AI, artificial intelligence; CNN, convolutional neural network; DFPN, depth-first proof number search; GNN, graph neural network; HSFP, hot-spot fingerprint; MaxFrag, maximum fragment; MCTS, Monte Carlo tree search; MILP, mixed integer linear programming; MRR, mean reciprocal rank; NN, neural network; RL, Reinforcement learning; ROC, receiver operating characteristic

REFERENCES

(1) Corey, E. J. General Methods for the Construction of Complex Molecules. In *The Chemistry of Natural Products*; Elsevier, 1967; pp 19−37.

(2) Lin, G.-M.; Warden-Rothman, R.; Voigt, C. A. Retrosynthetic Design of Metabolic Pathways to Chemicals Not Found in Nature. *Curr. Opin. Syst. Biol.* **2019**, *14*, 82−107.

(3) Sheldon, R. A.; Woodley, J. M. Role of Biocatalysis in Sustainable Chemistry. *Chem. Rev.* **2018**, *118* (2), 801−838.

(4) Yu, T.; Boob, A. G.; Volk, M. J.; Liu, X.; Cui, H.; Zhao, H. Machine Learning-Enabled Retrobiosynthesis of Molecules. *Nat. Catal.* **2023**, *6* (2), 137−151.

(5) Tricco, A. C.; Lillie, E.; Zarin, W.; O'Brien, K. K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M. D. J.; Horsley, T.; Weeks, L.; Hempel, S.; Akl, E. A.; Chang, C.; McGowan, J.; Stewart, L.; Hartling, L.; Aldcroft, A.; Wilson, M. G.; Garritty, C.; Lewin, S.; Godfrey, C. M.; Macdonald, M. T.; Langlois, E. V.; Soares-Weiser, K.; Moriarty, J.; Clifford, T.; Tunçalp, Ö.; Straus, S. E. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Int. Med.* **2018**, *169* (7), 467−473.

(6) Aal E Ali, R. S.; Meng, J.; Khan, M. E. I.; Jiang, X. Machine Learning Advancements in Organic Synthesis: A Focused Exploration of Artificial Intelligence Applications in Chemistry. *Artif. Intell. Chem.* **2024**, *2*, 100049.

(7) Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. Machine Learned Prediction of Reaction Template Applicability for Data-Driven Retrosynthetic Predictions of Energetic Materials; *AIP Conf. Proc.*; AIP Publishing, Portland, OR, USA, 2020; Vol. *2272*, p 070014.

(8) Wan, Y.; Liao, B.; Hsieh, C.-Y.; Zhang, S. Retroformer: Pushing the Limits of Interpretable End-to-End Retrosynthesis Transformer. In *Proceedings of the 39th International Conference on Machine Learning*; Proceedings of Machine Learning Research; PMLR, 2022; Vol. *162*, pp 22475−22490.

(9) Zhong, W.; Yang, Z.; Chen, C. Y.-C. Retrosynthesis Prediction Using an End-to-End Graph Generative Architecture for Molecular Graph Editing. *Nat. Commun.* **2023**, *14* (1), 3009.

(10) Karpov, P.; Godin, G.; Tetko, I. V. A Transformer Model for Retrosynthesis. In *Artificial Neural Networks and Machine Learning - ICANN 2019: Workshop and Special Sessions*; Tetko, I. V., Kůrková, V., Karpov, P., Theis, F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2019; Vol. *11731*, pp 817−830.

(11) Heid, E.; Liu, J.; Aude, A.; Green, W. H. Influence of Template Size, Canonicalization, and Exclusivity for Retrosynthesis and Reaction Prediction Applications. *J. Chem. Inf. Model.* **2022**, *62* (1), 16−26.

(12) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3* (12), 1237−1245.

(13) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370−377.

(14) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn. Sci. Technol.* **2020**, *1* (4), 045024.

(15) Carbonell, P.; Carlsson, L.; Faulon, J.-L. Stereo Signature Molecular Descriptor. *J. Chem. Inf. Model.* **2013**, *53* (4), 887−897.

(16) Hähnke, V. D.; Bolton, E. E.; Bryant, S. H. PubChem Atom Environments. *J. Cheminformatics* **2015**, *7* (1), 41.

(17) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), e1603.

(18) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminformatics* **2020**, *12* (1), 56.

(19) Li, J.; Fang, L.; Lou, J.-G. RetroRanker: Leveraging Reaction Changes to Improve Retrosynthesis Prediction through Re-Ranking. *J. Cheminformatics* **2023**, *15* (1), 58.

(20) Kim, E.; Lee, D.; Kwon, Y.; Park, M. S.; Choi, Y.-S. Valid, Plausible, and Diverse Retrosynthesis Using Tied Two-Way Transformers with Latent Variables. *J. Chem. Inf. Model.* **2021**, *61* (1), 123−133.

(21) 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* (6), 2529−2537.

(22) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* **2020**, *11* (12), 3316−3325.

(23) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data Augmentation and Transfer Learning Strategies for Reaction Prediction in Low Chemical Data Regimes. *Org. Chem. Front.* **2021**, *8* (7), 1415−1423.

(24) Mao, K.; Xiao, X.; Xu, T.; Rong, Y.; Huang, J.; Zhao, P. Molecular Graph Enhanced Transformer for Retrosynthesis Prediction. *Neurocomputing* **2021**, *457*, 193−202.
(25) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: A Pre-Trained Transformer for Computational Chemistry. *Mach. Learn. Sci. Technol.* **2022**, *3* (1), 015022.
(26) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **2019**, *59* (2), 673−688.
(27) Zhang, B.; Lin, J.; Du, L.; Zhang, L. Harnessing Data Augmentation and Normalization Preprocessing to Improve the Performance of Chemical Reaction Predictions of Data-Driven Model. *Polymers* **2023**, *15* (9), 2224.
(28) Zhu, J.; Xia, Y.; Wu, L.; Xie, S.; Zhou, W.; Qin, T.; Li, H.; Liu, T.-Y. Dual-View Molecular Pre-Training. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: Long Beach, CA, USA, 2023; pp 3615−3627.
(29) Yang, F.; Liu, J.; Zhang, Q.; Yang, Z.; Zhang, X. CNN-Based Two-Branch Multi-Scale Feature Extraction Network for Retrosynthesis Prediction. *BMC Bioinformatics* **2022**, *23* (1), 362.
(30) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (1), 47−55.
(31) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152−12155.
(32) Wang, X.; Li, Y.; Qiu, J.; Chen, G.; Liu, H.; Liao, B.; Hsieh, C.-Y.; Yao, X. RetroPrime: A Diverse, Plausible and Transformer-Based Method for Single-Step Retrosynthesis Predictions. *Chem. Eng. J.* **2021**, *420*, 129845.
(33) 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* (10), 1103−1113.
(34) Fang, L.; Li, J.; Zhao, M.; Tan, L.; Lou, J.-G. Single-Step Retrosynthesis Prediction by Leveraging Commonly Preserved Substructures. *Nat. Commun.* **2023**, *14* (1), 2446.
(35) Zhang, K.; Mann, V.; Venkatasubramanian, V. G MATT: Single step Retrosynthesis Prediction Using Molecular Grammar Tree Transformer. *AIChE J.* **2024**, *70*, e18244.
(36) Yan, C.; Zhao, P.; Lu, C.; Yu, Y.; Huang, J. RetroComposer: Composing Templates for Template-Based Retrosynthesis Prediction. *Biomolecules* **2022**, *12* (9), 1325.
(37) Bai, R.; Zhang, C.; Wang, L.; Yao, C.; Ge, J.; Duan, H. Transfer Learning: Making Retrosynthetic Predictions Based on a Small Chemical Reaction Dataset Scale to a New Level. *Molecules* **2020**, *25* (10), 2357.
(38) Qiao, H.; Wu, Y.; Zhang, Y.; Zhang, C.; Wu, X.; Wu, Z.; Zhao, Q.; Wang, X.; Li, H.; Duan, H. Transformer-Based Multitask Learning for Reaction Prediction under Low-Resource Circumstances. *RSC Adv.* **2022**, *12* (49), 32020−32026.
(39) Yan, Y.; Zhao, Y.; Yao, H.; Feng, J.; Liang, L.; Han, W.; Xu, X.; Pu, C.; Zang, C.; Chen, L.; Li, Y.; Liu, H.; Lu, T.; Chen, Y.; Zhang, Y. RPBP: Deep Retrosynthesis Reaction Prediction Based on Byproducts. *J. Chem. Inf. Model.* **2023**, *63* (19), 5956−5970.
(40) Heid, E.; Goldman, S.; Sankaranarayanan, K.; Coley, C. W.; Flamm, C.; Green, W. H. EHreact: Extended Hasse Diagrams for the Extraction and Scoring of Enzymatic Reaction Templates. *J. Chem. Inf. Model.* **2021**, *61* (10), 4949−4961.
(41) Seo, S.-W.; Song, Y. Y.; Yang, J. Y.; Bae, S.; Lee, H.; Shin, J.; Hwang, S. J.; Yang, E. GTA: Graph Truncated Attention for Retrosynthesis. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35* (1), 531−539.
(42) He, H.-R.; Wang, J.; Liu, Y.; Wu, F. Modeling Diverse Chemical Reactions for Single-Step Retrosynthesis via Discrete Latent Variables. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*; ACM: Atlanta, GA, USA, 2022; pp 717−726.
(43) Sankaranarayanan, K.; Heid, E.; Coley, C. W.; Verma, D.; Green, W. H.; Jensen, K. F. Similarity Based Enzymatic Retrosynthesis. *Chem. Sci.* **2022**, *13* (20), 6039−6053.
(44) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-Art Augmented NLP Transformer Models for Direct and Single-Step Retrosynthesis. *Nat. Commun.* **2020**, *11* (1), 5575.
(45) Toniato, A.; Vaucher, A. C.; Schwaller, P.; Laino, T. Enhancing Diversity in Language Based Models for Single-Step Retrosynthesis. *Digit. Discovery* **2023**, *2* (2), 489−501.
(46) Ucak, U. V.; Ashyrmamatov, I.; Lee, J. Reconstruction of Lossless Molecular Representations from Fingerprints. *J. Cheminformatics* **2023**, *15* (1), 26.
(47) Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; Klambauer, G. Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. *J. Chem. Inf. Model.* **2022**, *62* (9), 2111−2120.
(48) Yu, J.; Wang, J.; Zhao, H.; Gao, J.; Kang, Y.; Cao, D.; Wang, Z.; Hou, T. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. *J. Chem. Inf. Model.* **2022**, *62* (12), 2973−2986.
(49) Ucak, U. V.; Kang, T.; Ko, J.; Lee, J. Substructure-Based Neural Machine Translation for Retrosynthetic Prediction. *J. Cheminformatics* **2021**, *13* (1), 4.
(50) Ucak, U. V.; Ashyrmamatov, I.; Ko, J.; Lee, J. Retrosynthetic Reaction Pathway Prediction through Neural Machine Translation of Atomic Environments. *Nat. Commun.* **2022**, *13* (1), 1186.
(51) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem., Int. Ed.* **2020**, *59* (2), 725−730.
(52) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Ring Breaker": Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **2020**, *63* (16), 8791−8808.
(53) Hasic, H.; Ishida, T. Single-Step Retrosynthesis Prediction Based on the Identification of Potential Disconnection Sites Using Molecular Substructure Fingerprints. *J. Chem. Inf. Model.* **2021**, *61* (2), 641−652.
(54) Segler, M. H. S.; Waller, M. P. Neural Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23* (25), 5966−5971.
(55) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J. Chem. Inf. Model.* **2019**, *59* (12), 5026−5033.
(56) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. In *Advances in Neural Information Processing Systems 32*; NeurIPS 2019; Curran Associates, Inc., 2019; Vol. 32.
(57) Chen, S.; Jung, Y. Deep Retrosynthetic Reaction Prediction Using Local Reactivity and Global Attention. *JACS Au* **2021**, *1* (10), 1612−1620.
(58) Lee, H.; Ahn, S.; Seo, S.-W.; Song, Y. Y.; Yang, E.; Hwang, S.-J.; Shin, J. RetCL: A Selection-Based Approach for Retrosynthesis via Contrastive Learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*; IJCAI-21; International Joint Conferences on Artificial Intelligence Organization, 2021; pp 2673−2679.
(59) Lin, Z.; Yin, S.; Shi, L.; Zhou, W.; Zhang, Y. J. G2GT: Retrosynthesis Prediction with Graph-to-Graph Attention Neural Network and Self-Training. *J. Chem. Inf. Model.* **2023**, *63* (7), 1894−1905.
(60) Delépine, B.; Duigou, T.; Carbonell, P.; Faulon, J.-L. RetroPath2.0: A Retrosynthesis Workflow for Metabolic Engineers. *Metab. Eng.* **2018**, *45*, 158−170.
(61) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55* (20), 5904−5937.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(62) 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* (2), 98−104.
(63) Duigou, T.; du Lac, M.; Carbonell, P.; Faulon, J.-L. RetroRules: A Database of Reaction Rules for Engineering Biology. *Nucleic Acids Res.* **2019**, *47* (D1), D1229−D1235.
(64) Dong, Z.; Chen, Z.; Wang, Q. Retrosynthesis Prediction Based on Graph Relation Network. In *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*; IEEE: Beijing, China, 2022; pp 1−5.
(65) 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* (9), 1572−1583.
(66) Guo, Z.; Wu, S.; Ohno, M.; Yoshida, R. Bayesian Algorithm for Retrosynthesis. *J. Chem. Inf. Model.* **2020**, *60* (10), 4474−4486.
(67) Tu, Z.; Coley, C. W. Permutation Invariant Graph-to-Sequence Model for Template-Free Retrosynthesis and Reaction Prediction. *J. Chem. Inf. Model.* **2022**, *62* (15), 3503−3513.
(68) Hu, H.; Jiang, Y.; Yang, Y.; Chen, J. X. BiG2S: A Dual Task Graph-to-Sequence Model for the End-to-End Template-Free Reaction Prediction. *Appl. Intell* **2023**, *53*, 29620.
(69) Liu, S.; Tu, Z.; Xu, M.; Zhang, Z.; Lin, L.; Ying, R.; Tang, J.; Zhao, P.; Wu, D. FusionRetro: Molecule Representation Fusion via In-Context Learning for Retrosynthetic Planning. In *Proceedings of the 40th International Conference on Machine Learning*; ICML'23; JMLR.org: Honolulu, Hawaii, USA, 2023.
(70) Lin, M. H.; Tu, Z.; Coley, C. W. Improving the Performance of Models for One-Step Retrosynthesis through Re-Ranking. *J. Cheminformatics* **2022**, *14* (1), 15.
(71) Sun, R.; Dai, H.; Li, L.; Kearnes, S.; Dai, B. Towards Understanding Retrosynthesis by Energy-Based Models. In *Advances in Neural Information Processing Systems 34*; NeurIPS 2021; Curran Associates, Inc., 2021.
(72) Christofidellis, D.; Giannone, G.; Born, J.; Winther, O.; Laino, T.; Manica, M. Unifying Molecular and Textual Representations via Multi-Task Language Modelling. In *Proceedings of the 40th International Conference on Machine Learning*; ICML'23; JMLR.org: Honolulu, Hawaii, USA, 2023.
(73) 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* (1), 3342.
(74) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* **2021**, *12* (25), 8648−8659.
(75) Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed Synthesis Planning Using Data-Driven Learning. *Nat. Commun.* **2022**, *13* (1), 964.
(76) 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*; ICML'20; JMLR.org, 2020.
(77) Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; Huang, J. RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; NIPS'20; Curran Associates Inc.: Red Hook, NY, USA, 2020.
(78) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. Learning Graph Models for Retrosynthesis Prediction. In *Advances in Neural Information Processing Systems 34*; NeurIPS 2021; Curran Associates, Inc., 2021.
(79) 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* (1), 6155.
(80) Thakkar, A.; Vaucher, A. C.; Byekwaso, A.; Schwaller, P.; Toniato, A.; Laino, T. Unbiasing Retrosynthesis Language Models with Disconnection Prompts. *ACS Cent. Sci.* **2023**, *9* (7), 1488−1498.
(81) Jaume-Santero, F.; Bornet, A.; Valery, A.; Naderi, N.; Vicente Alvarez, D.; Proios, D.; Yazdani, A.; Bournez, C.; Fessard, T.; Teodoro, D. Transformer Performance for Chemical Reactions: Analysis of Different Predictive and Evaluation Scenarios. *J. Chem. Inf. Model.* **2023**, *63* (7), 1914−1924.
(82) Wang, X.; Hsieh, C.-Y.; Yin, X.; Wang, J.; Li, Y.; Deng, Y.; Jiang, D.; Wu, Z.; Du, H.; Chen, H.; Li, Y.; Liu, H.; Wang, Y.; Luo, P.; Hou, T.; Yao, X. Generic Interpretable Reaction Condition Predictions with Open Reaction Condition Datasets and Unsupervised Learning of Reaction Center. *Research* **2023**, *6*, 0231.
(83) Zhang, C.; Lapkin, A. A. Reinforcement Learning Optimization of Reaction Routes on the Basis of Large, Hybrid Organic Chemistry-Synthetic Biological, Reaction Network Data. *React. Chem. Eng.* **2023**, *8* (10), 2491−2504.
(84) Koch, M.; Duigou, T.; Faulon, J.-L. Reinforcement Learning for Bioretrosynthesis. *ACS Synth. Biol.* **2020**, *9* (1), 157−168.
(85) Wang, W.; Liu, Q.; Zhang, L.; Dong, Y.; Du, J. RetroSynX: A Retrosynthetic Analysis Framework Using Hybrid Reaction Templates and Group Contribution-Based Thermodynamic Models. *Chem. Eng. Sci.* **2022**, *248*, 117208.
(86) Liu, Q.; Tang, K.; Zhang, L.; Du, J.; Meng, Q. Computer assisted Synthetic Planning Considering Reaction Kinetics Based on Transition State Automated Generation Method. *AIChE J.* **2023**, *69* (7), e18092.
(87) Kreutter, D.; Reymond, J.-L. Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search. *Chem. Sci.* **2023**, *14* (36), 9959−9969.
(88) Kishimoto, A.; Buesser, B.; Chen, B.; Botea, A. Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning. In *Advances in Neural Information Processing Systems 32*; NeurIPS 2019; Curran Associates, Inc., 2019.
(89) Franz, C.; Mogk, G.; Mrziglod, T.; Schewior, K. Completeness and Diversity in Depth-First Proof-Number Search with Applications to Retrosynthesis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization: Vienna, Austria, 2022; pp 4747−4753.
(90) Shibukawa, R.; Ishida, S.; Yoshizoe, K.; Wasa, K.; Takasu, K.; Okuno, Y.; Terayama, K.; Tsuda, K. CompRet: A Comprehensive Recommendation Framework for Chemical Synthesis Planning with Algorithmic Enumeration. *J. Cheminformatics* **2020**, *12* (1), 52.
(91) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604−610.
(92) Zhang, B.; Zhang, X.; Du, W.; Song, Z.; Zhang, G.; Zhang, G.; Wang, Y.; Chen, X.; Jiang, J.; Luo, Y. Chemistry-Informed Molecular Graph as Reaction Descriptor for Machine-Learned Retrosynthesis Planning. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (41), e2212711119.
(93) Wang, X.; Qian, Y.; Gao, H.; Coley, C. W.; Mo, Y.; Barzilay, R.; Jensen, K. F. Towards Efficient Discovery of Green Synthetic Pathways with Monte Carlo Tree Search and Reinforcement Learning. *Chem. Sci.* **2020**, *11* (40), 10959−10972.
(94) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *Chem. Sci.* **2020**, *11* (1), 154−168.
(95) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic Retrosynthetic Route Planning Using Template-Free Models. *Chem. Sci.* **2020**, *11* (12), 3355−3364.
(96) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A Fast, Robust and Flexible Open-Source Software for Retrosynthetic Planning. *J. Cheminformatics* **2020**, *12* (1), 70.
(97) Gao, H.; Coley, C. W.; Struble, T. J.; Li, L.; Qian, Y.; Green, W. H.; Jensen, K. F. Combining Retrosynthesis and Mixed-Integer Optimization for Minimizing the Chemical Inventory Needed to Realize a WHO Essential Medicines List. *React. Chem. Eng.* **2020**, *5* (2), 367−376.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(98) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), eaax1566.
(99) Westerlund, A. M.; Barge, B.; Mervin, L.; Genheden, S. Data driven Approaches for Identifying Hyperparameters in Multi step Retrosynthesis. *Mol. Inform.* **2023**, *42*, 2300128.
(100) Sankaranarayanan, K.; Jensen, K. F. Computer-Assisted Multistep Chemoenzymatic Retrosynthesis Using a Chemical Synthesis Planner. *Chem. Sci.* **2023**, *14* (23), 6467−6475.
(101) Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W. Merging Enzymatic and Synthetic Chemistry with Computational Synthesis Planning. *Nat. Commun.* **2022**, *13*, 7747.
(102) Jeong, J.; Lee, N.; Shin, Y.; Shin, D. Intelligent Generation of Optimal Synthetic Pathways Based on Knowledge Graph Inference and Retrosynthetic Predictions Using Reaction Big Data. *J. Taiwan Inst. Chem. Eng.* **2022**, *130*, 103982.
(103) Chen, B.; Li, C.; Dai, H.; Song, L. Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search. In *Proceedings of the 37th International Conference on Machine Learning*; Proceedings of Machine Learning Research (PMLR), 2020; Vol. *119*, pp 1608−1616.
(104) Xie, S.; Yan, R.; Han, P.; Xia, Y.; Wu, L.; Guo, C.; Yang, B.; Qin, T. RetroGraph: Retrosynthetic Planning with Graph Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: Washington DC, USA, 2022; pp 2120−2129.
(105) Han, P.; Zhao, P.; Lu, C.; Huang, J.; Wu, J.; Shang, S.; Yao, B.; Zhang, X. GNN-Retro: Retrosynthetic Planning with Graph Neural Networks. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36* (4), 4014−4021.
(106) Latendresse, M.; Malerich, J. P.; Herson, J.; Krummenacker, M.; Szeto, J.; Vu, V.-A.; Collins, N.; Madrid, P. B. SynRoute: A Retrosynthetic Planning Software. *J. Chem. Inf. Model.* **2023**, *63* (17), 5484−5495.
(107) Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; Molga, K.; Dittwald, P.; Wołos, A.; Klucznik, T. Chematica: A Story of Computer Code That Started to Think like a Chemist. *Chem.* **2018**, *4* (3), 390−398.
(108) Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed.; Pearson series in artificial intelligence; Pearson: Hoboken, 2021.
(109) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **2019**, *5* (6), 970−981.
(110) Yu, Y.; Wei, Y.; Kuang, K.; Huang, Z.; Yao, H.; Wu, F. GRASP: Navigating Retrosynthetic Planning with Goal-Driven Policy. In *Advances in Neural Information Processing Systems 35*; NeurIPS 2022; Curran Associates, Inc., 2022.
(111) Kumar, A.; Wang, L.; Ng, C. Y.; Maranas, C. D. Pathway Design Using de Novo Steps through Uncharted Biochemical Spaces. *Nat. Commun.* **2018**, *9* (1), 184.
(112) Carbonell, P.; Parutto, P.; Herisson, J.; Pandit, S. B.; Faulon, J.-L. XTMS: Pathway Design in an eXTended Metabolic Space. *Nucleic Acids Res.* **2014**, *42* (W1), W389−W394.
(113) Tokic, M.; Hadadi, N.; Ataman, M.; Neves, D.; Ebert, B. E.; Blank, L. M.; Miskovic, L.; Hatzimanikatis, V. Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors. *ACS Synth. Biol.* **2018**, *7* (8), 1858−1873.
(114) Hadadi, N.; Hafner, J.; Shajkofci, A.; Zisaki, A.; Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **2016**, *5* (10), 1155−1166.
(115) Otero-Muras, I.; Carbonell, P. Automated Engineering of Synthetic Metabolic Pathways for Efficient Biomanufacturing. *Metab. Eng.* **2021**, *63*, 61−80.
(116) Kim, J.; Ahn, S.; Lee, H.; Shin, J. Self-Improved Retrosynthetic Planning. In *Proceedings of the 38th International Conference on Machine Learning*, Virtual, July 18−24, 2021; Proceedings of Machine Learning Research (PMLR), 2021.
(117) Gao, D.; Song, W.; Wu, J.; Guo, L.; Gao, C.; Liu, J.; Chen, X.; Liu, L. Efficient Production of L-Homophenylalanine by Enzymatic-Chemical Cascade Catalysis. *Angew. Chem., Int. Ed.* **2022**, *61* (36), e202207077.
(118) 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* (1), 12−22.
(119) Finnigan, W.; Flitsch, S. L.; Hepworth, L. J.; Turner, N. J. Enzyme Cascade Design: Retrosynthesis Approach. In *Enzyme Cascade Design and Modelling*; Kara, S., Rudroff, F., Eds.; Springer International Publishing: Cham, 2021; pp 7−30.
(120) Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. Critical Assessment of Synthetic Accessibility Scores in Computer-Assisted Synthesis Planning. *J. Cheminformatics* **2023**, *15* (1), 6.
(121) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) - Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. *Chem. Sci.* **2021**, *12* (9), 3339−3349.
(122) Li, B.; Chen, H. Prediction of Compound Synthesis Accessibility Based on Reaction Knowledge Graph. *Molecules* **2022**, *27* (3), 1039.
(123) Kim, H.; Lee, K.; Kim, C.; Lim, J.; Kim, W. Y. DFRscore: Deep Learning-Based Scoring of Synthetic Complexity with Drug-Focused Retrosynthetic Analysis for High-Throughput Virtual Screening. *J. Chem. Inf. Model* **2024**, *64*, 2432.
(124) Correia, J.; Carreira, R.; Pereira, V.; Rocha, M. Predicting the Number of Biochemical Transformations Needed to Synthesize a Compound. In *2022 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Padua, Italy, 2022; pp 1−8.
(125) Parrot, M.; Tajmouati, H.; Da Silva, V. B. R.; Atwood, B. R.; Fourcade, R.; Gaston-Mathé, Y.; Do Huu, N.; Perron, Q. Integrating Synthetic Accessibility with AI-Based Generative Drug Design. *J. Cheminformatics* **2023**, *15* (1), 83.
(126) Sanchez-Garcia, R.; Havasi, D.; Takács, G.; Robinson, M. C.; Lee, A.; Von Delft, F.; Deane, C. M. CoPriNet: Graph Neural Networks Provide Accurate and Rapid Compound Price Prediction for Molecule Prioritisation. *Digit. Discovery* **2023**, *2* (1), 103−111.
(127) Tang, K.; Zhuang, Y.; Wang, W.; Liu, Q.; Zhang, L.; Du, J.; Meng, Q. GC-NORM-Based Thermodynamic Framework for Evaluations of Organic Reactions Involving Carbon Dioxide Utilization. *Chem. Eng. Sci.* **2023**, *278*, 118913.
(128) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1), 015016.
(129) Plehiers, P. P.; Coley, C. W.; Gao, H.; Vermeire, F. H.; Dobbelaere, M. R.; Stevens, C. V.; Van Geem, K. M.; Green, W. H. Artificial Intelligence for Computer-Aided Synthesis In Flow: Analysis and Selection of Reaction Components. *Front. Chem. Eng.* **2020**, *2*, 5.
(130) Toniato, A.; Unsleber, J. P.; Vaucher, A. C.; Weymuth, T.; Probst, D.; Laino, T.; Reiher, M. Quantum Chemical Data Generation as Fill-in for Reliability Enhancement of Machine-Learning Reaction and Retrosynthesis Planning. *Digit. Discovery* **2023**, *2* (3), 663−673.
(131) Genheden, S.; Engkvist, O.; Bjerrum, E. Fast Prediction of Distances between Synthetic Routes with Deep Learning. *Mach. Learn. Sci. Technol.* **2022**, *3* (1), 015018.
(132) Kuznetsov, A.; Sahinidis, N. V. ExtractionScore: A Quantitative Framework for Evaluating Synthetic Routes on Predicted Liquid-Liquid Extraction Performance. *J. Chem. Inf. Model.* **2021**, *61* (5), 2274−2282.
(133) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(134) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252−261.
(135) Ebrahim, A.; Lerman, J. A.; Palsson, B. O.; Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **2013**, *7* (1), 74.
(136) Beber, M. E.; Gollub, M. G.; Mozaffari, D.; Shebek, K. M.; Flamholz, A. I.; Milo, R.; Noor, E. eQuilibrator 3.0: A Database Solution for Thermodynamic Constant Estimation. *Nucleic Acids Res.* **2022**, *50*, D603−D609.
(137) Hérisson, J.; Duigou, T.; Du Lac, M.; Bazi-Kabbaj, K.; Sabeti Azad, M.; Buldum, G.; Telle, O.; El Moubayed, Y.; Carbonell, P.; Swainston, N.; Zulkower, V.; Kushwaha, M.; Baldwin, G. S.; Faulon, J.-L. The Automated Galaxy-SynBioCAD Pipeline for Synthetic Biology Design and Engineering. *Nat. Commun.* **2022**, *13* (1), 5082.
(138) Levin, I.; Fortunato, M. E.; Tan, K. L.; Coley, C. W. Computer aided Evaluation and Exploration of Chemical Spaces Constrained by Reaction Pathways. *AIChE J.* **2023**, *69*, e18234.
(139) Wang, L.; Upadhyay, V.; Maranas, C. D. dGPredictor: Automated Fragmentation Method for Metabolic Reaction Free Energy Prediction and de Novo Pathway Design. *PLOS Comput. Biol.* **2021**, *17* (9), e1009448.
(140) Feehan, R.; Montezano, D.; Slusky, J. S. G. Machine Learning for Enzyme Engineering, Selection and Design. *Protein Eng. Des. Sel.* **2021**, *34*, gzab019.
(141) Stoney, R. A.; Hanko, E. K. R.; Carbonell, P.; Breitling, R. SelenzymeRF: Updated Enzyme Suggestion Software for Unbalanced Biochemical Reactions. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 5868−5876.
(142) 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. *Bioinforma. Oxf. Engl.* **2018**, *34* (12), 2153−2154.
(143) Hadadi, N.; MohammadiPeyhani, H.; Miskovic, L.; Seijo, M.; Hatzimanikatis, V. Enzyme Annotation for Orphan and Novel Reactions Using Knowledge of Substrate Reactive Sites. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (15), 7298−7307.
(144) Finnigan, W.; Lubberink, M.; Hepworth, L. J.; Citoler, J.; Mattey, A. P.; Ford, G. J.; Sangster, J.; Cosgrove, S. C.; da Costa, B. Z.; Heath, R. S.; Thorpe, T. W.; Yu, Y.; Flitsch, S. L.; Turner, N. J. RetroBioCat Database: A Platform for Collaborative Curation and Automated Meta-Analysis of Biocatalysis Data. *ACS Catal.* **2023**, *13* (17), 11771−11780.
(145) Kim, Y.; Ryu, J. Y.; Kim, H. U.; Jang, W. D.; Lee, S. Y. A Deep Learning Approach to Evaluate the Feasibility of Enzymatic Reactions Generated by Retrobiosynthesis. *Biotechnol. J.* **2021**, *16* (5), 2000605.
(146) Upadhyay, V.; Boorla, V. S.; Maranas, C. D. Rank-Ordering of Known Enzymes as Starting Points for Re-Engineering Novel Substrate Activity Using a Convolutional Neural Network. *Metab. Eng.* **2023**, *78*, 171−182.
(147) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.* **2004**, *126* (50), 16487−16498.
(148) Rahman, S. A.; Cuesta, S. M.; Furnham, N.; Holliday, G. L.; Thornton, J. M. EC-BLAST: A Tool to Automatically Search and Compare Enzyme Reactions. *Nat. Methods* **2014**, *11* (2), 171−174.
(149) Egelhofer, V.; Schomburg, I.; Schomburg, D. Automatic Assignment of EC Numbers. *PLoS Comput. Biol.* **2010**, *6* (1), e1000661.
(150) Hu, Q.-N.; Zhu, H.; Li, X.; Zhang, M.; Deng, Z.; Yang, X.; Deng, Z. Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. *PLoS One* **2012**, *7* (12), e52901.
(151) Probst, D. An Explainability Framework for Deep Learning on Chemical Reactions Exemplified by Enzyme-Catalysed Reaction Classification. *J. Cheminformatics* **2023**, *15* (1), 113.
(152) Liu, C.-H.; Korablyov, M.; Jastrzębski, S.; Włodarczyk-Pruszyński, P.; Bengio, Y.; Segler, M. RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software. *J. Chem. Inf. Model.* **2022**, *62* (10), 2293−2300.
(153) Hafner, J.; Mohammadi-Peyhani, H.; Hatzimanikatis, V. Pathway Design. In *Metabolic Engineering*; John Wiley & Sons, Ltd., 2021; pp 237−257.
(154) de Souza, R. O. M. A.; Miranda, L. S. M.; Bornscheuer, U. T. A Retrosynthesis Approach for Biocatalysis in Organic Synthesis. *Chem. - Eur. J.* **2017**, *23* (50), 12040−12063.
(155) Song, Z.; Zhang, Q.; Wu, W.; Pu, Z.; Yu, H. Rational Design of Enzyme Activity and Enantioselectivity. *Front. Bioeng. Biotechnol* **2023**, *11*, 1129149.
(156) Ribeiro, A. J. M.; Riziotis, I. G.; Borkakoti, N.; Thornton, J. M. Enzyme Function and Evolution through the Lens of Bioinformatics. *Biochem. J.* **2023**, *480* (22), 1845−1863.
(157) Beaudoin, C.; Kundu, S.; Topaloglu, R. O.; Ghosh, S. Quantum Machine Learning for Material Synthesis and Hardware Security (Invited Paper). In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*; ACM: San Diego, California, 2022; pp 1−7.
(158) Fan, Y.; Xia, Y.; Zhu, J.; Wu, L.; Xie, S.; Qin, T. Back Translation for Molecule Generation. *Bioinformatics* **2022**, *38* (5), 1244−1251.
(159) Zahoránszky-Kőhalmi, G.; Lysov, N.; Vorontcov, I.; Wang, J.; Soundararajan, J.; Metaxotos, D.; Mathew, B.; Sarosh, R.; Michael, S. G.; Godfrey, A. G. Algorithm for the Pruning of Synthesis Graphs. *J. Chem. Inf. Model.* **2022**, *62* (9), 2226−2238.
(160) Chen, Z.; Ayinde, O. R.; Fuchs, J. R.; Sun, H.; Ning, X. G2Retro as a Two-Step Graph Generative Models for Retrosynthesis Prediction. *Commun. Chem.* **2023**, *6* (1), 102.
(161) Genheden, S.; Norrby, P.-O.; Engkvist, O. AiZynthTrain: Robust, Reproducible, and Extensible Pipelines for Training Synthesis Prediction Models. *J. Chem. Inf. Model.* **2023**, *63* (7), 1841−1846.
(162) Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. Evaluating and Clustering Retrosynthesis Pathways with Learned Strategy. *Chem. Sci.* **2021**, *12* (4), 1469−1478.
(163) Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakarajan, N.; Cardinale, A.; Laino, T.; Rodríguez Martínez, M. Data-Driven Molecular Design for Discovery and Synthesis of Novel Ligands: A Case Study on SARS-CoV-2. *Mach. Learn. Sci. Technol.* **2021**, *2* (2), 025024.
(164) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(165) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570−578.
(166) Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; Zhang, X. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks. http://arxiv.org/abs/2305.18365 (accessed 2023-11-27).
(167) Meng, Z.; Zhao, P.; Yu, Y.; King, I. A Unified View of Deep Learning for Reaction and Retrosynthesis Prediction: Current Status and Future Challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization: Macau, SAR China, 2023; pp 6723−6731.
(168) Hu, W.; Liu, Y.; Chen, X.; Chai, W.; Chen, H.; Wang, H.; Wang, G. Deep Learning Methods for Small Molecule Drug Discovery: A Survey. *IEEE Trans. Artif. Intell.* **2024**, *5*, 459.
(169) Jiang, Y.; Yu, Y.; Kong, M.; Mei, Y.; Yuan, L.; Huang, Z.; Kuang, K.; Wang, Z.; Yao, H.; Zou, J.; Coley, C. W.; Wei, Y. Artificial Intelligence for Retrosynthesis Prediction. *Engineering* **2023**, *25*, 32−50.
(170) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143* (45), 18820−18826.
(171) Heid, E.; Probst, D.; Green, W. H.; Madsen, G. K. H. EnzymeMap: Curation, Validation and Data-Driven Prediction of Enzymatic Reactions. *Chem. Sci.* **2023**, *14* (48), 14229−14242.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(172) Seierstad, M.; Tichenor, M. S.; DesJarlais, R. L.; Na, J.; Bacani, G. M.; Chung, D. M.; Mercado-Marin, E. V.; Steffens, H. C.; Mirzadegan, T. Novel Reagent Space: Identifying Unorderable but Readily Synthesizable Building Blocks. *ACS Med. Chem. Lett.* **2021**, *12* (11), 1853−1860.
(173) Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. Computer-Aided Key Step Generation in Alkaloid Total Synthesis. *Science* **2023**, *379* (6631), 453−457.
(174) Hardy, M. A.; Nan, B.; Wiest, O.; Sarpong, R. Strategic Elements in Computer-Assisted Retrosynthesis: A Case Study of the Pupukeanane Natural Products. *Tetrahedron* **2022**, *104*, 132584.
(175) Soudier, P.; Zúñiga, A.; Duigou, T.; Voyvodic, P. L.; Bazi-Kabbaj, K.; Kushwaha, M.; Vendrell, J. A.; Solassol, J.; Bonnet, J.; Faulon, J.-L. PeroxiHUB: A Modular Cell-Free Biosensing Platform Using H_2O_2 as Signal Integrator. *ACS Synth. Biol.* **2022**, *11* (8), 2578−2588.
(176) Robinson, C. J.; Carbonell, P.; Jervis, A. J.; Yan, C.; Hollywood, K. A.; Dunstan, M. S.; Currin, A.; Swainston, N.; Spiess, R.; Taylor, S.; Mulherin, P.; Parker, S.; Rowe, W.; Matthews, N. E.; Malone, K. J.; Le Feuvre, R.; Shapira, P.; Barran, P.; Turner, N. J.; Micklefield, J.; Breitling, R.; Takano, E.; Scrutton, N. S. Rapid Prototyping of Microbial Production Strains for the Biomanufacture of Potential Materials Monomers. *Metab. Eng.* **2020**, *60*, 168−182.
(177) Zhang, Z.; Fang, L.; Wang, F.; Deng, Y.; Jiang, Z.; Li, A. Transforming Inert Cycloalkanes into α,ω-Diamines by Designed Enzymatic Cascade Catalysis. *Angew. Chem., Int. Ed.* **2023**, *62* (16), e202215935.
(178) Liu, Z.; Zhang, X.; Lei, D.; Qiao, B.; Zhao, G.-R. Metabolic Engineering of Escherichia Coli for de Novo Production of 3-Phenylpropanol via Retrobiosynthesis Approach. *Microb. Cell Factories* **2021**, *20* (1), 121.
(179) Yiakoumetti, A.; Hanko, E. K. R.; Zou, Y.; Chua, J.; Chromy, J.; Stoney, R. A.; Valdehuesa, K. N. G.; Connolly, J. A.; Yan, C.; Hollywood, K. A.; Takano, E.; Breitling, R. Expanding Flavone and Flavonol Production Capabilities in Escherichia Coli. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1275651.
(180) Brito, L. F.; Irla, M.; Nærdal, I.; Le, S. B.; Delépine, B.; Heux, S.; Brautaset, T. Evaluation of Heterologous Biosynthetic Pathways for Methanol-Based 5-Aminovalerate Production by Thermophilic Bacillus Methanolicus. *Front. Bioeng. Biotechnol* **2021**, *9*, 1.
(181) Hanko, E. K. R.; Valdehuesa, K. N. G.; Verhagen, K. J. A.; Chromy, J.; Stoney, R. A.; Chua, J.; Yan, C.; Roubos, J. A.; Schmitz, J.; Breitling, R. Carboxylic Acid Reductase-Dependent Biosynthesis of Eugenol and Related Allylphenols. *Microb. Cell Factories* **2023**, *22* (1), 238.

https://doi.org/10.1021/acssynbio.4c00091
ACS Synth. Biol. XXXX, XXX, XXX−XXX