nature communications

Article

https://doi.org/10.1038/s41467-024-52048-4

Site-specific template generative approach for retrosynthetic planning

Received: 20 March 2024

Accepted: 26 August 2024

Published online: 06 September 2024

Check for updates

Yu Shee[1,3], Haote Li[1,3], Pengpeng Zhang[1], Andrea M. Nikolic[1], Wenxin Lu[1], H. Ray Kelly[2], Vidhyadhar Manee[2], Sanil Sreekumar[2], Frederic G. Buono[2], Jinhua J. Song[2], Timothy R. Newhouse[1] & Victor S. Batista[1]

Retrosynthesis, the strategy of devising laboratory pathways by working backwards from the target compound, is crucial yet challenging. Enhancing retrosynthetic efficiency requires overcoming the vast complexity of chemical space, the limited known interconversions between molecules, and the challenges posed by limited experimental datasets. This study introduces generative machine learning methods for retrosynthetic planning. The approach features three innovations: generating reaction templates instead of reactants or synthons to create novel chemical transformations, allowing user selection of specific bonds to change for human-influenced synthesis, and employing a conditional kernel-elastic autoencoder (CKAE) to measure the similarity between generated and known reactions for chemical viability insights. These features form a coherent retrosynthetic framework, validated experimentally by designing a 3-step synthetic pathway for a challenging small molecule, demonstrating a significant improvement over previous 5-9 step approaches. This work highlights the utility and robustness of generative machine learning in addressing complex challenges in chemical synthesis.

Retrosynthesis is the design of deconstructing complex molecules into simpler building blocks, a concept originally developed by Corey as a means to educate students to conduct multistep synthesis[1]. This intellectual framework laid the foundation for the development of ComputerAided Synthesis Planning (CASP), a field that emerged to assist chemists in navigating various paths of synthesis, playing a pivotal role in augmenting human capabilities for refining a synthetic approach[2,3]. In the earlier stages, systems based on expert rules provided valuable insights for chemists[4–7]. As organic chemistry advanced, encompassing broader chemical space and synthetic methodologies, recent advancements in CASP have shifted from rule-based to precedent-based approaches[8]. This shift was facilitated by large-scale extraction of reaction rules[9]. The process progressed from manual creation to automated extraction from extensive chemical datasets. Several extraordinary software packages have emerged due to this transition which empowered CASP tools to tap into repositories of historical reaction data[8,10,11]. Grzybowski and others[12,13] further introduced user-purpose-driven tools for route optimization, demonstrating remarkable success through experimental validations[14–18]. Furthermore, the integration of machine learning (ML) methods has marked the latest chapter in the ongoing evolution of CASP[19,20]. ML models offer promising alternatives and can be broadly categorized as selection-based, semi-template, or generation-based methods[21] (see Fig. 1a).

Selection-based methods, such as reactant selection and template selection methods, aim to choose appropriate molecules or reaction rules from the given sets. Reactant selection methods[22,23] involve ranking molecules from a collection of candidates based on the target compounds. While reactant selection methods have the advantage of ensuring the chosen molecules are valid, their effectiveness relies on the availability of reactants in the candidate sets. Template selection methods[24–30] rank the reaction templates in terms of their applicability to the target molecules. These templates capture subgraph patterns representing the change in atoms and bonds during a reaction.

[1]Department of Chemistry, Yale University, New Haven, CT, USA. [2]Chemical Development, Boehringer Ingelheim Pharmaceuticals Inc, Ridgefield, CT, USA. [3]These authors contributed equally: Yu Shee, Haote Li. e-mail: timothy.newhouse@yale.edu; victor.batista@yale.edu

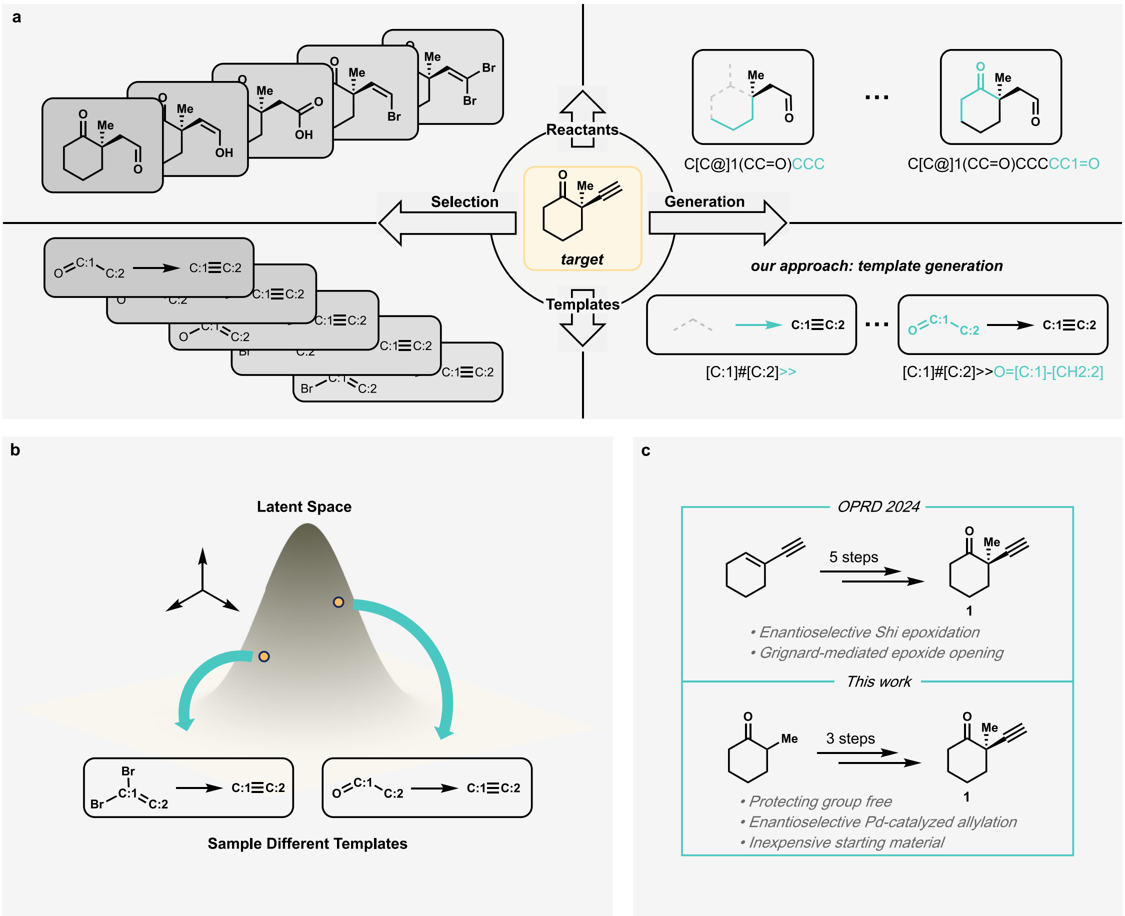
a
Reactants
Selection
Generation
target
Templates
C[C@]1(CC=O)CCC
C[C@]1(CC=O)CCCCC1=O
our approach: template generation
[C:1]#[C:2]>>
[C:1]#[C:2]>>O=[C:1]-[CH2:2]
b
Latent Space
Sample Different Templates
c
OPRD 2024
5 steps
1
• Enantioselective Shi epoxidation
• Grignard-mediated epoxide opening
This work
3 steps
1
• Protecting group free
• Enantioselective Pd-catalyzed allylation
• Inexpensive starting material

Fig. 1 | Common machine learning methods for retrosynthesis and our approach. a Reactants and templates can be selected or generated based on a target compound using different machine learning models. Template generation is used in this work. **b** A structured latent space is incorporated in one of the models in this work. Sampling in the latent space can give different reaction templates for given products. **c** Reduction of synthetic steps for a key intermediate for active pharmaceutical ingredients (API). OPRD 2024 refers to ref. 62.

Notably, the RDChiral repository by Coley et al.[31] offers template extraction methods and a collection of reaction templates in the form of SMARTS strings. Template selection methods simplify the reaction representation to a single template instead of multiple reactants. In addition, the same template can be applied to different target compounds instead of having multiple sets of reactants for the target compounds, thereby providing a higher coverage of reaction space. However, like reactant selection methods, template selection methods are dependent on the coverage and diversity of available templates within predefined reaction rules.

Semi-template methods[32–37] involve the identification of reaction centers, synthons, or leaving groups, followed by the prediction of corresponding reactants based on these rules. Some semi-template methods[34,36,37] are akin to selection-based methods, where reactants are obtained by predicting reaction centers and selecting from a collection of leaving groups, or by selecting necessary edits on molecular graphs. Other semi-template methods adopt generation components, in which reactants are generated from products and identified synthons or rules.

Generation-based methods are not bound by the sets of available reactants or templates and hold promise to map wider areas of chemical space. These include template-free methods[38–53] that treat reactant generation as a translation task, aiming to predict the reactants directly from the given products without having in-dataset reaction rules. They therefore bear the potential to explore a wider range of possible reactions.

In this study, we introduce template generation which represents a new distinct category of generation-based methods for retrosynthetic planning. Template generation models employ the Sequence-to-Sequence (S2S) architecture trained to translate product information into reaction templates, as opposed to generating reactants. The capability of template generation thus extends beyond the available templates or predefined reaction rules of template selection-based approaches, enabling the discovery of novel reaction templates that expand the scope of retrosynthetic planning. The combination of generated reaction templates and the “RunReactants” function from RDKit, offer an efficient means to swiftly identify templates that yield grammatically coherent reactants from given products. This facilitates the exploration of previously uncharted chemical reactions and pathways.

One of the major benefits of using template generation is the ease of checking the reaction validity. During the transformation of a reaction template, the product is guaranteed to be converted to the reactant with exact matching of atoms indices and relevant functional groups from the description of the template. In comparison to reactant generative models, this feature greatly reduces the uncertainty in the produced reactants which might not correspond to any known reactions or have key atom mismatches due to problems during decoding.

Our template generation method introduces a design where site-specific templates (SST) are generated along with target compounds with labeled reaction centers (i.e., center-labeled products, CLP) that specify the reaction centers. This results in the generation of concise and informative sets of templates that are different from the templates available in the RDChiral repository[31]. Through benchmarking with a public dataset, the performance of our approach is demonstrated.

The second design is a sampling generative model (sampling model) for template generation conditioned by target compounds. S2S models, such as those employed in the template-free methods, predict pathways deterministically and do not have a sampling process or definition of latent space. In contrast, our sampling model has a latent space, enabling the generation, interpolation, extrapolation, and distance measurement of various templates (Fig. 1b). Deterministic models that take target compounds as inputs and generate templates are also developed in this work. Importantly, the encoder of the model can incorporate positional embedding for reaction centers, enabling users to specify specific reacting sites during prediction. Results are benchmarked on the USPTO-FULL dataset.

Our sampling model, based on the conditional kernel elastic autoencoder (CKAE)[54], is the first of its kind in the field of retrosynthesis. This model conditions on corresponding products during training, allowing interpolation and extrapolation of reaction templates in latent space to generate new reaction templates during the sampling process. The latent space also provides a measure of distances between reaction templates, allowing us to identify the closest reaction reference within the dataset, or determine the similarity between two chemical reactions. Previous works on assessing reaction similarity and reaction classification use physicochemical properties[55–57], molecular fingerprints[58,59], or reaction SMILES strings as input[60]. In this work, SSTs and CLPs are used to evaluate the similarity of reactions. Schwaller et al.[60] include reaction conditions such as catalysts and solvents in the reaction SMILES strings, while the fingerprint methods from Schneider et al.[58] and Ghiandoni et al.[59] require that reactants are separated from reagents. Our method is similar to Schwaller et al.[60] in terms of using strings as input, but SSTs and CLPs provide a more concise way to represent reactions (without reagents) and carry additional information about atom mapping, like the method from ref. 59.

With SSTs and generation methods in place, our approach is validated through the practical application of synthesis. A library of potent anti-cancer agents was recently reported by Boehringer Ingelheim[61]. One of the key intermediates for the synthesis of these anti-cancer compounds is compound **1**, a cyclohexanone with a quaternary stereogenic center in the α-position containing an alkyne moiety (Fig. 1c)[62,63]. Our objective was to develop a more step-efficient route to synthesize compound **1**. The route proceeds over 3 steps, compared to prior approaches that required 5–9 steps, including a recent process involving Grignard-mediated epoxide opening as a key step in a 5-step route starting from commercial starting materials[62,63]. Reducing the number of steps in a synthetic process is enabling to develop scalable and more sustainable approaches, while also reducing the amount of time necessary for each batch[64,65]. Our experimental validation demonstrates the practicality and reliability of the retrosynthetic predictions, suggesting their underlying promise to address a wide spectrum of synthetic challenges.

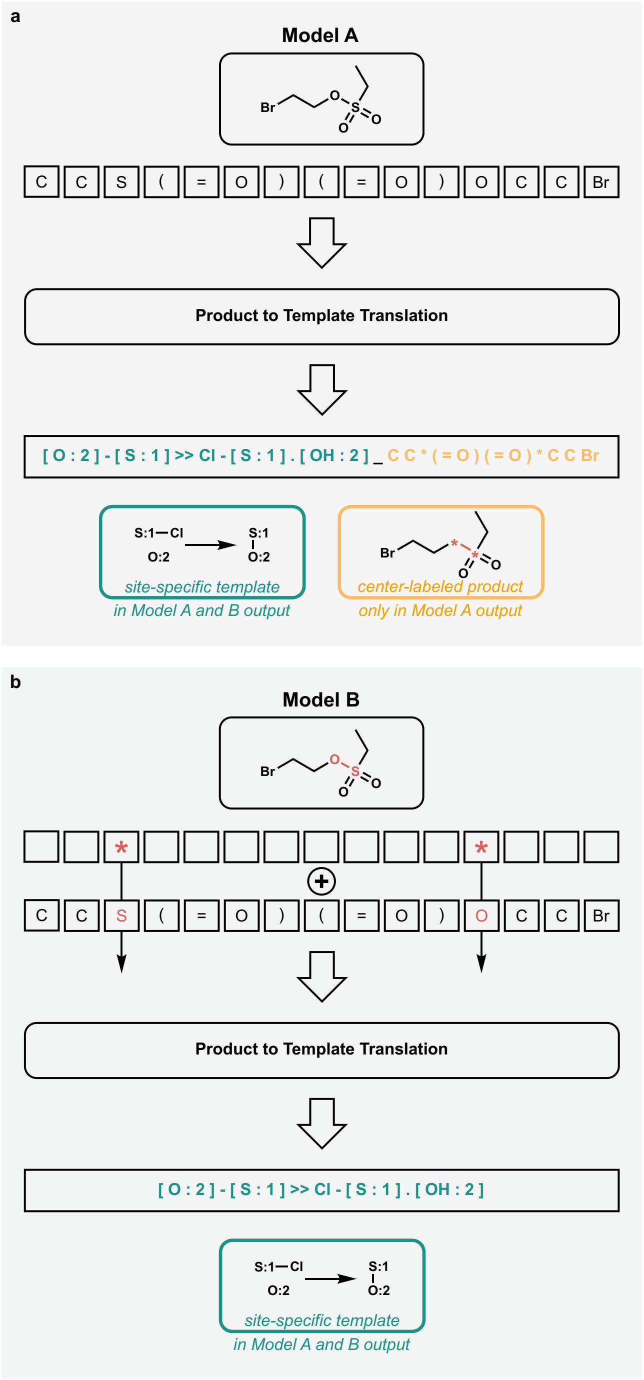

Fig. 2 | Schematic retrosynthetic workflow for Models A and B. a Workflow of Model A. **b** Workflow of Model B. Model B has reaction center embeddings and does not have center-labeled products in the output. Detailed descriptions of the models are provided in the Supplementary Information (Sec. 3 and Sec. 4).

Results

Site-specific templates and center-labeled products

Reaction templates that only apply to reaction centers within the target compounds are referred to as site-specific templates (SST). These are different from RDChiral templates which involve a broader structural context[31] since SSTs do not differentiate neighboring atoms or special functional groups when matching substructures within the target compounds. The presence of center-labeled products (CLP) is a pre-requisite for the effective use of SSTs. Such labeling is essential to avoid ambiguity when a SST can be applied to multiple sites within a target compound. Examples of SST and CLP are shown in Fig. 2a where the "*" symbol represents the reaction centers. To prepare SSTs, the radius parameter in RDChiral is set to 0 (while RDChiral normally sets radius to 1 which captures 1 bond away from the reaction centers) and special functional groups are removed. Therefore, neighboring atoms and distal functional groups are not included in SSTs. Also, explicit degrees and explicit numbers of hydrogens are not included in the SSTs. To prepare CLPs, RDChiral also has implementations to capture the changed atoms, so the centers can be labeled for target compounds. Further explanations and examples are provided in the Supplementary Information (Sec. 1 and Sec. 2).

Deterministic model performance

Figure 2 shows a schematic representation of the deterministic models, Model A and Model B. Model A takes a target compound as an input and translates it into SSTs and CLPs. CLPs specify how the SSTs should be applied to the target compound. Model B takes the target and the specific reaction centers and generates templates corresponding to those specific sites (see Supplementary Information Sec. 3 for more information and Supplementary Fig. 3 for a comparison of the models).

Figure 3 shows the comparative analysis of the performance for Models A and B (highlighted in red), in terms of Top-*K* accuracy, as compared to state-of-the-art methods.

Top-*K* accuracy measures the percentage of top-*K* predictions containing reactants that precisely match the ground truth reactants in the testing set. The Top-*K* results are derived from the beam search method, where the product of the next token probabilities is used to rank the output templates and their corresponding precursors. This ranking is referred to as the beam score in this work. Figure 3 includes results for the original USPTO-Full testing dataset as well as for a cleaned testing set to address errors related to atom mappings such as solvent and reagent atoms erroneously considered as part of the reactions. The cleaned testing set is prepared by removing reactions containing reactants that are the 50 most frequently observed spectators in the USPTO-FULL dataset. The size of the cleaned testing set is 90.7% of the original set of 95k reactions.

Model A, which does not use reaction centers, performs comparably well to other methods. The cleaned set allows for higher accuracy although it may inadvertently exclude some reactions where the common spectators actually participate as reactants. Model B leverages reaction center information. On the cleaned set, Model B reaches a performance milestone, achieving an accuracy rate as high as 80% for Top-10 predictions (see Supplementary Information Sec. 9 for details).

RetroExplainer[36], with semi-template components, demonstrates remarkable prediction accuracy owing to its data modeling approach and the utilization of a set of leaving groups. However, this approach may experience variations in performance when handling uncommon scenarios or leaving groups not explicitly represented in the dataset. R-SMILES[52], a template-free generation-based method, introduced the root-aligned SMILES representation to ensure minimal edit distances between product and reactant SMILES. Through this customized string representation and data augmentation, they achieved the highest accuracy among template-free methods. Nonetheless, data augmentation is not utilized in this work, leaving room for potential improvements in accuracy for future endeavors.

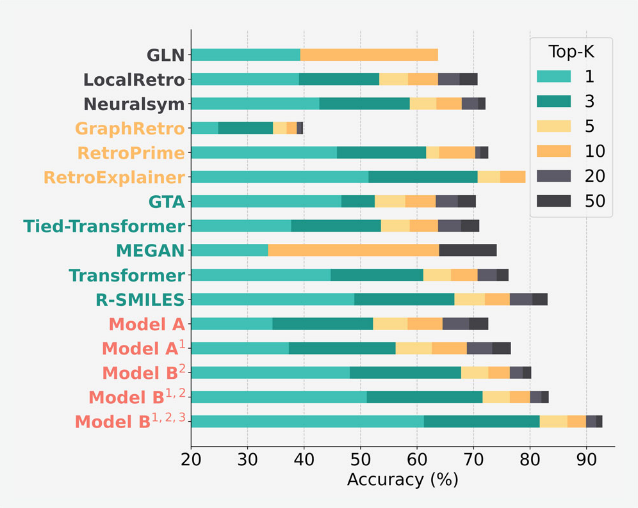

Fig. 3 | USPTO-Full Top-*K* accuracy of retrosynthesis models. GLN[28], LocalRetro[29], and Neuralsym[24] in black are template-based selection methods. GraphRetro[34], RetroPrime[35], and RetroExplainer[36] in yellow are semi-template methods. GTA[45], Tied-Transformer[50], MEGAN[47], Transformer[44], and R-SMILES[52] in green are template-free generation methods. This work (in red) uses a template generation method. Reactant-based selection methods are not included due to out-of-memory for the USPTO-FULL dataset[21]. [1]Indicates that if the correct reactants contain one of the 50 most commonly seen spectators in the USPTO-Full dataset, the reaction is removed from the test set. [2]Indicates that reaction centers are provided. [3]Indicates that the maximum number of reaction centers is two.

Top-*K* accuracy is not the sole criterion for evaluating retrosynthetic methods; explainability and inference time are equally important factors. The explainability of the template generation approach is facilitated by atom mappings from templates. Neuralsym[24] is widely used for benchmarking multistep tree search methods due to its fast inference time. Our template generative approach has a similar order of magnitude of inference time as Neuralsym (10^1 s), while other methods operate at an order of 10^2 s or above, as shown in ref. 21. In this reference[21], batch size optimization or multi-process multi-GPU acceleration are not implemented, so a batch size of 1 is used for comparison. R-SMILES, for instance, has an inference time at the order of 10^3 s. Although both R-SMILES and our approach are generation methods, the template generation approach using SSTs has a much shorter string representation than reactants and does not require augmented SMILES inputs to reach high accuracy, resulting in significantly shorter inference times.

In addition, an analysis of the Top-*K* accuracy considering different numbers of reaction centers for Model B is shown. Over half of the test reactions possess one or two reaction centers, following the same distribution of reaction center counts of the training set. Consequently, for test reactions with a maximum of two reaction centers, Model B achieved the highest Top-*K* accuracy compared to other center counts, with the Top-10 accuracy reached 90% (see last row of Fig. 3 and Supplementary Information Sec. 9), showcasing exceptional predictive capabilities in scenarios characterized by a limited number of reaction centers. This also aligns with the precursor selection process illustrated in “Experimental validation”, where two reaction centers are consistently utilized for Model B. The high Top-*K* accuracy achieved by Model B for reactions with few reaction centers is particularly significant, as it corresponds to real-world applications where a majority of reactions feature a low number of reaction centers. For instance, 90% of the dataset comprises reactions with no more than four reaction centers (see Supplementary Information Sec. 9).

Sampling model with latent space

A sampling generative model, which exploits a sampling process with a latent space, is different from the deterministic approach. To the best of our knowledge, the application of a sampling model for retrosynthetic planning has not been explored. Model C is built upon the architecture of Conditional Kernel-Elastic Autoencoder (CKAE)[54]. In Model C, both the input and output consist of combinations of SSTs and CLPs. The goal of Model C, akin to a variational autoencoder, is to reconstruct the input with latent space compression. Comparing to previous CKAE molecular generation models where conditions are represented by specific values or molecular properties, the CKAE model as applied to Model C utilizes the SMILES representation of target molecules as conditions. During the sampling process, a target compound is provided as the condition and latent vectors are sampled, different SSTs and CLPs, which correspond to the same target compound condition, can then be generated for different latent vectors.

In addition to generative sampling, the encoder of Model C offers a valuable referencing feature. It maps the input into a latent space with a distance regularized by a modified maximum mean discrepancy loss (m-MMD)[54]. This distance furnishes a quantifiable metric for assessing the similarity between reactions, aiding in evaluating and

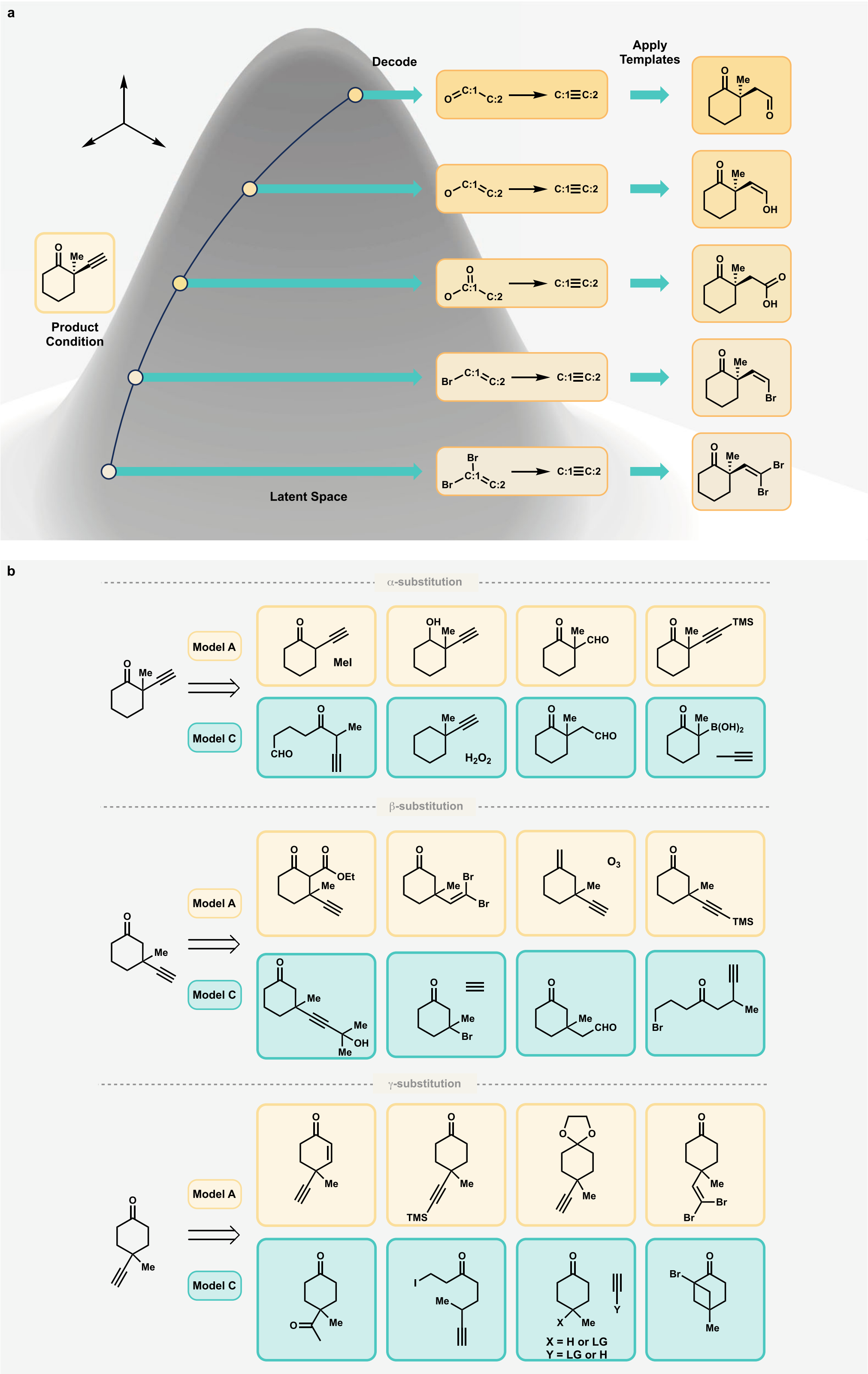

Fig. 4 | Interpolation of templates in the latent space of Model C and reactants from Model A and Model C outputs. a The intermediates of the top and bottom latent representations are decoded. **b** Selected reactants for 2-, 3-, 4-substituted cyclohexanone derivatives as target compounds.

understanding the differences between chemical transformations. Such capability enables the identification of similar reactions within the dataset.

The distance between various chemical transformations in latent space can be used to interpolate between chemical reactions. This can be useful when searching for a reaction that could be an intermediate between two known chemical reactions. In Fig. 4a, an interpolation process is visualized. Initially, two reaction templates are selected, represented by the top and bottom templates and the latent vectors in the latent space. These templates serve as the starting points to explore the intermediates. This interpolation allowed the discovery of the templates corresponding to each of the latent vectors along the path between the two originals. It can be observed that the middle templates and reactants form a blending of the starting templates and reactants. This observation provides evidence that the latent space captures chemical information, showing the distance measure between various chemical transformations.

To illustrate the differences between Model A (deterministic) and Model C (generative sampling), the single-step predictions of 2-, 3-, and 4-substituted cyclohexanone derivatives are examined. Model A and Model C are compared because, unlike Model B, they both do not take reaction center information as input, and this comparison highlights the effects of latent sampling on model outputs. Based on the acquired results, representative precursors are selected for all three target molecules. As shown in Fig. 4b, Model A suggestions are primarily based on functional group interconversions and protection reactions. While Model C also proposes these transformations, diverse precursors and reactions are also proposed. These examples complement the intuitive bias of many synthetic chemists and point to areas of opportunity for the creative development of novel chemical transformations. Please refer to Supplementary Information Sec. 7 to see experiments and examples of how the models can generate reactions that extend beyond the available templates.

Regarding the usage of each model: Model A should be used when high-accuracy predictions are needed without explicit reaction center information. Model B is suitable when there are specific insights or constraints regarding reaction centers, requiring precise control over disconnections. Model C is ideal for seeking greater diversity and potential for unconventional transformations, as it leverages generative sampling to explore a broader chemical space without predefined reaction centers.

Experimental validation

Developing inexpensive, rapid, and robust methods for the synthesis of bioactive molecules is one of the key goals in pharmaceutical chemistry[66]. Herein, we utilized our Model B, chosen because of its high accuracy and reaction center embedding, for establishing the shortest route for the synthesis of a target compound. Figure 5 highlights how Model B can be used to navigate multiple options for retrosynthesis (see Supplementary Information Sec. 8 for more pharmaceutical examples). The top five ranked precursors, based on beam scores or synthetic accessibility (SA) scores[67] as implemented in RDKit[68], are shown. Each level corresponds to a new prediction by the single-step model to reach an intermediate. This tool highlights the interactive nature of the model with a human expert who selects intermediates for further analysis. Both ketone and aldehyde

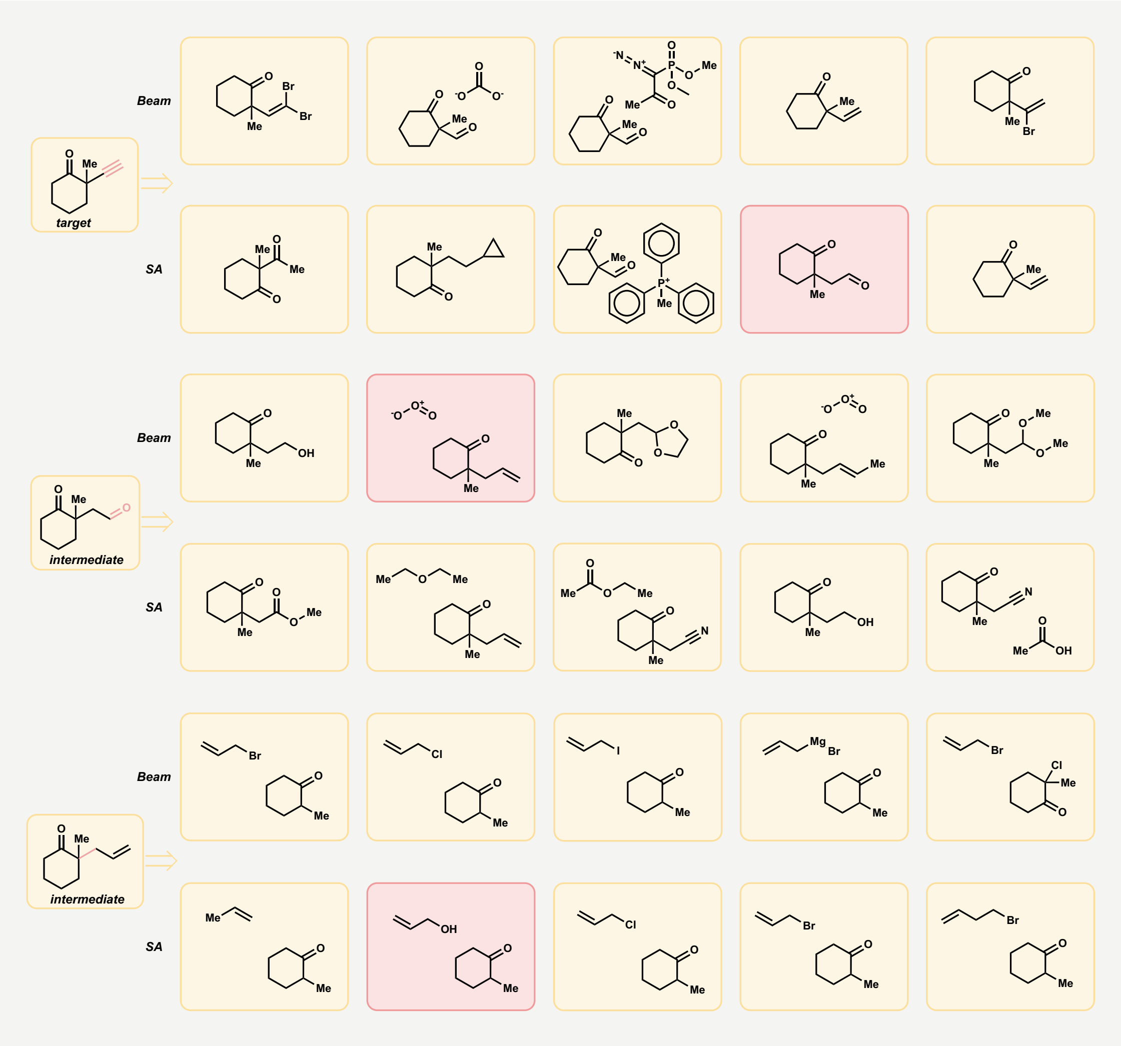

Fig. 5 | Retrosynthetic planning for compound 1 by Model B. The top five precursors, ranked by beam scores or SA scores, are displayed for each compound in left-to-right order. The red boxes highlight the selected precursors.

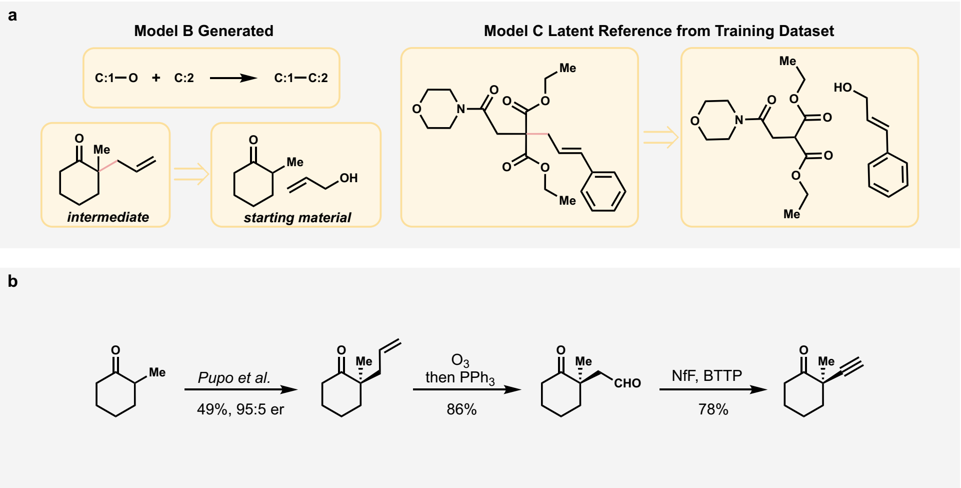

Fig. 6 | Synthesis of compound 1. a Reference found with Model C for the allylation step. **b** Experimental procedure for the selected route: (i) 2-methylcyclohexanone (1 equiv.), allyl methyl carbonate (3 equiv.), $Pd_2(dba)_3$ (5 mol % Pd), *t*-BuXPhos (11 mol %), *R*-TRIP (10 mol %), 3 Å MS, CyH, 45 °C, 5 days, 49%; (ii) O_3, CH_2Cl_2, − 78 °C, then PPh_3 (2 equiv.), −78 °C → rt, 16 h, 86%; iii) NfF (1.05 equiv.), BTTP (6 equiv.), DMF, −30 °C → rt, 19 h, 78%. The enantiomeric ratio is reported by Pupo et al.[69].

precursors for the target compound are highly ranked. Several intermediates and subsequent retrosynthetic steps were examined. The aldehyde was chosen due to subsequent retrosynthetic evaluation showing it to be a highly enabling retron. The next retrosynthetic step from the aldehyde included the alpha-allyl cyclohexanone, which facilitates the application of the highly robust Tsuji–Trost allylation.

Figure 6a serves as a reference point derived from Model C. The left-hand side illustrates the allylation step that we employed in our synthesis. On the right-hand side, the reference is obtained by encoding the allylation template and the product labeled with the reaction center into Model C's latent space. This process allows us to identify the closest latent vectors from the training dataset, and that closest reference corresponds to the reaction shown on the right-hand side of Fig. 6a. Interestingly, the exact chemical transformation that was suggested had previously been conducted, but is not in the USPTO-FULL dataset. This highlights how our approach compliments other synthetic planning tools, such as Reaxys and SciFinder.

In order to synthesize the enantiomerically enriched target molecule, we applied the enantioselective Pd-catalyzed Tsuji–Trost allylation of a ketone and applied conditions recently reported by Pupo et al.[69]. The prior literature protocol for this substrate reported an enantiomeric ratio of 95.5:4.5. The allylated intermediate was treated with ozone in order to obtain the ketoaldehyde derivative in good yield (Fig. 6b). For the final step, a modified procedure by Boltukhina et al[70]. was applied to form the alkyne in 78% yield. The overall yield of our 3-step route is 33%, despite our route not having undergone process optimization. It should be stated that further process optimization is expected to improve the efficiency of this approach, although this proof-of-concept demonstrates the ability to develop step-efficient routes. This experimental procedure serves as evidence that the newly developed ML models can facilitate the development of synthetic routes for pharmaceutically significant molecules and enhance existing routes.

An alternative to the route presented in Fig. 6b, an even shorter route to compound **1**, could be one entailing direct α-alkynylation of 2-methylcyclohexanone. Methods for direct introduction of an alkyne moiety next to a ketone are scarce and rely on substitution with electrophilic alkyne species (selected examples[71–76]). Most commonly used in modern organic chemistry are hypervalent iodine reagents such as Waser's or Ochiai's reagent[77]. While this method would furnish the target molecule in fewer synthetic steps, it would have to be followed by the separation of two enantiomers since enantioselective α-alkynylation of ketones has not yet been reported.

Discussion

In this work, a string-based approach for retrosynthesis planning is introduced, utilizing generative models to address the challenges posed by the vast chemical space and synthesis complexity. Specifically, this work introduces template generation as a new category in machine learning methods for computer-aided synthesis planning. Two types of generative models are developed, including deterministic generative models (Model A and Model B) and a sampling generative model that utilizes CKAE (Model C).

Model A and Model B are benchmarked on the USPTO-FULL dataset. Particularly, Model B can incorporate reaction center information, enabling the generation of templates that apply to the specified reacting sites. On the other hand, Model C represents a pioneering application of sampling method from latent space, capable of generating diverse reactions. The design of Model C defines distances between reactions, which allows Model C to identify the closest reference from the dataset for newly generated templates, making it a suitable tool for generating and validating a wide range of potential reactions.

This work presents two approaches for single-step synthetic planning, high-accuracy deterministic models and high-diversity sampling models. The capability of specifying reacting sites, the availability of relevant reaction references, and the successful results of experimental validations on an important pharmaceutically relevant intermediate make the models valuable tools in guiding retrosynthetic analysis.

Methods

Training details

In total, 10% dropout was applied to all attention matrices and embedding vectors. ADAM optimizer[78] was used with a learning rate of 5×10^{-5}. Gradient normalization[79] was set to 1.0. During training, each token in the input to the encoders is replaced by a mask token for Model A and Model B with the probability of 0.15.

Model architecture

Models A, B, and C each has 6 layers of transformer encoders and decoders as implemented in ref. 80. For Models A and B, 8 attention heads and an embedding size of 256 are used. For Model C, 16 attention heads and an embedding size of 512 are used.

The reaction center embeddings for Model B are achieved by adding the embedding of the reaction center token "*" at the specific position of the atoms similar to the concept of positional embeddings.

Model C is constructed based on the conditional kernel elastic autoencoder model[54], with a 5120-dimensional latent space. The conditions are embeddings of target compounds and are also achieved by 6 layers of transformer encoders and 16 heads with an embedding size of 512[80]. These embeddings are then compressed into 10 embedding vectors by a linear layer and concatenated with the input embedding and the latent space. See Supplementary Information Sec. 4 and Supplementary Fig. 5 for more details and visualization of the architecture.

Beam search

To derive multiple possible predictions, beam search[44] is used across all models. During decoding, the transformer decoder attends to the encoder output and the sequence that had been generated. The decoder outputs probabilities of all possible tokens for the next position in the sequence. Beam search maintains a fixed-size set of candidate sequences, the number that the method keeps is called the beam size *B*. The top *B* most probable sequences at each decoding step are selected to proceed to the next step of decoding until the stopping criteria of maximum allowed length are reached or an End Of Sequence (<EOS>) token is output.

For the Top-*K* accuracy test, beam search with a beam size of 50 is used during all decoding processes. At each decoding step, the model outputs the 50 most probable candidate tokens and continues the sequence until the stopping criteria are met.

The diversity of deterministic models is solely derived from the beam search process, as this type of model lacks a latent space for sampling. Consequently, generating novel reactions using a deterministic model through beam search can be challenging. In contrast, the sampling model, equipped with a latent space, can generate diverse and novel reactions more effectively.

Synthesis

Details of the synthesis, such as reaction conditions, purification, and NMR spectra, are provided in the Supplementary Information.

Data availability

The 50 most commonly seen spectators are obtained from the USPTO-Full reaction file in RDChiral GitHub Repository[31]. While the train-validation-test split of the USPTO-Full dataset is obtained from the GitHub repository of ref. 44. Experimental data, such as the NMR spectra, are provided in the Supplementary Information. All data are available from the corresponding authors upon request.

Code availability

A user-friendly interface was developed, and all pre-trained models from this work can be accessed on models.batistalab.com.

References

1. James Corey, E. & Todd Wipke, W. Computer-assisted design of complex organic syntheses: pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science* **166**, 178–192 (1969).
2. Thakkar, A. J. The coming of the computer age to organic chemistry: recent approaches to systematic synthesis analysis. *Computers Chem*. 3–18 (2006).
3. Alan, R. et al. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev*. **110**, 5714–5789 (2010).
4. Wipke, W. T. & Howe, W. J. *Computerassisted Organic Synthesis* (ACS Publications, 1977).
5. Gelernter, H. L. et al. Empirical explorations of SYNCHEM: the methods of artificial intelligence are applied to the problem of organic synthesis route discovery. *Science* **197**, 1041–1049 (1977).
6. Bauer, J., Fontain, E., Forstmeyer, D. & Ugi, I. Interactive generation of organic reactions by igor 2 and the PC-assisted discovery of a new reaction. *Tetrahedron Comput. Methodol*. **1**, 129–132 (1988).
7. Hanessian, S., Franco, J. & Larouche, B. The psychobiological basis of heuristic synthesis planning-man, machine and the Chiron approach. *Pure Appl. Chem*. **62**, 1887–1910 (1990).
8. Ravitz, O. Data-driven computer aided synthesis design. *Drug Discov. Today.: Technol*. **10**, e443–e449 (2013).
9. Cook, A. et al. Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev.: Comput. Mol. Sci*. **2**, 79–107 (2012).
10. Bøgevig, A. et al. Route design in the 21st century: the ic synth software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev*. **19**, 357–368 (2015).
11. Davies, I. W. The digitization of organic synthesis. *Nature* **570**, 175–181 (2019).
12. Gothard, C. M. et al. Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angew. Chem. Int. Ed*. **51**, 7922–7927 (2012).
13. Kowalik, M. et al. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem. Int. Ed*. **51**, 7928–7932 (2012).
14. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
15. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
16. Lin, Y. et al. Reinforcing the supply chain of umifenovir and other antiviral drugs with retrosynthetic software. *Nat. Commun*. **12**, 7327 (2021).
17. Hardy, M. A., Nan, B., Wiest, Olaf & Sarpong, R. Strategic elements in computer-assisted retrosynthesis: a case study of the pupukeanane natural products. *Tetrahedron* **104**, 132584 (2022).
18. Lin, Y., Zhang, R., Wang, D. & Cernak, T. Computer-aided key step generation in alkaloid total synthesis. *Science* **379**, 453–457 (2023).
19. Filipa de Almeida, A., Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem*. **3**, 589–604 (2019).
20. Struble, T. J. et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem*. **63**, 8667–8682 (2020).
21. Zhong, Z. et al. Recent advances in deep learning for retrosynthesis. *Wiley Interdiscip. Rev.: Comput. Mol. Sci*. **14**, e1694 (2024).
22. Guo, Z., Wu, S., Ohno, M. & Yoshida, R. Bayesian algorithm for retrosynthesis. *J. Chem. Inf. Model*. **60**, 4474–4486 (2020).
23. Lee, H. et al. Retcl: a selection-based approach for retrosynthesis via contrastive learning. Preprint at. https://arxiv.org/abs/2105.00795 (2021).
24. Segler, M. H. & Waller, M. P. Neuralsymbolic machine learning for retrosynthesis and reaction prediction. *Chem. A Eur. J*. **23**, 5966–5971 (2017).
25. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci*. **3**, 1237–1245 (2017).
26. 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*. **59**, 5026–5033 (2019).
27. Fortunato, M. E., Coley, C. W., Barnes, B. C. & Jensen, K. F. Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. *J. Chem. Inf. Model*. **60**, 3398–3407 (2020).

28. Dai, H., Li, C., Coley, C., Dai, B. & Song, L Retrosynthesis prediction with conditional graph logic network. *Adv. Neural Inf. Process. Syst.* **32** (2019).
29. Chen, S. & Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **1**, 1612–1620 (2021).
30. Seidl, P. et al. Improving few-and zero-shot reaction template prediction using modern Hopfield networks. *J. Chem. Inf. Model.* **62**, 2111–2120 (2022).
31. 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.* **59**, 2529–2537 (2019).
32. Yan, C. et al. Retroxpert: decompose retrosynthesis prediction like a chemist. *Adv. Neural Inf. Process. Syst.* **33**, 11248–11258 (2020).
33. Shi, C., Xu, M., Guo, H., Zhang, M. & Tang, J. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning* 8818–8827 (PMLR, 2020).
34. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. *Adv. Neural Inf. Process. Syst.* **34**, 9405–9415 (2021).
35. Wang, X. et al. A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chem. Eng. J.* **420**, 129845 (2021).
36. Wang, Y. et al. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks. *Nat. Commun.* **14**, 6155 (2023).
37. Zhong, W., Yang, Z. & Chen, C. Y. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nat. Commun.* **14**, 3009 (2023).
38. Liu, B. et al. reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
39. Karpov, P., Godin, G. & Tetko, I. V. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*, 817–830 (Springer, 2019).
40. Chen, B., Shen, T., Jaakkola, T. S. & Barzilay, R. Learning to make generalizable and diverse predictions for retrosynthesis. Preprint at. https://arxiv.org/abs/1910.09688 (2019).
41. Lee, A. A. et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.* **55**, 12152–12155 (2019).
42. Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **11**, 3355–3364 (2020).
43. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J. Chem. Inf. Model.* **60**, 47–55 (2019).
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.* **11**, 5575 (2020).
45. Seo, S. W. et al. Gta: graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, 531–539 (Association for the Advancement of Artificial Intelligence (AAAI), 2021).
46. Mao, K., Xiao, X., Xu, T., Rong, Y., Huang, J. & Zhao, P. Molecular graph enhanced transformer for retrosynthesis prediction. *Neurocomputing* **457**, 193–202 (2021).
47. Sacha, M. et al. edit graph attention network: modeling chemical reactions as sequences of graph edits. *J. Chem. Inf. Model.* **61**, 3273–3284 (2021).
48. Mann, V. & Venkatasubramanian, V. Retrosynthesis prediction using grammar-based neural machine translation: an information-theoretic approach. *Computers Chem. Eng.* **155**, 107533 (2021).
49. Ucak, U. V., Kang, T., Ko, J. & Lee, J. Substructure-based neural machine translation for retrosynthetic prediction. *J. Cheminformatics* **13**, 4 (2021).
50. 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.* **61**, 123–133 (2021).
51. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pretrained transformer for computational chemistry. *Mach. Learn.: Sci. Technol.* **3**, 015022 (2022).
52. Zhong, Z. et al. Root-aligned smiles: a tight representation for chemical reaction prediction. *Chem. Sci.* **13**, 9023–9034 (2022).
53. Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **13**, 1186 (2022).
54. Li, H. et al. Kernel-elastic autoencoder for molecular design. *PNAS Nexus* **3**, 168 (2024).
55. Chen, L. & Gasteiger, J. Organic reactions classified by neural networks: Michael additions, Friedel–Crafts alkylations by alkenes, and related reactions. *Angew. Chem. Int. Ed. Engl.* **35**, 763–765 (1996).
56. Chen, L. & Gasteiger, J. Knowledge discovery in reaction databases: landscaping organic reactions by a self-organizing neural network. *J. Am. Chem. Soc.* **119**, 4033–4042 (1997).
57. Satoh, H. et al. Classification of organic reactions: similarity of reactions based on changes in the electronic features of oxygen atoms at the reaction sites. *J. Chem. Inf. Computer Sci.* **38**, 210–219 (1998).
58. Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **55**, 39–53 (2015).
59. Ghiandoni, G. M. et al. Development and application of a data-driven reaction classification model: comparison of an electronic lab notebook and medicinal chemistry literature. *J. Chem. Inf. Model.* **59**, 4167–4187 (2019).
60. Schwaller, P. et al. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).
61. Abbott, J. et al. Annulated 2-amino-3cyano thiophenes and derivatives for the treatment of cancer. US Patent 11,945,812 (2024).
62. Tan, Z. et al. Development of a scalable synthesis toward a kras g12c inhibitor building block bearing an all-carbon quaternary stereocenter, part 2: asymmetric synthesis via shi epoxidation. *Org. Process Res. Dev.* **28**, 78–91 (2024).
63. Leung, J. C. et al. Development of a scalable synthesis toward a kras g12c inhibitor building block bearing an all-carbon quaternary stereocenter, part 1: from discovery route to kilogram-scale production. *Org. Process Res. Dev.* **28**, 67–77 (2024).
64. Newhouse, T., Baran, P. S. & Hoffmann, R. W. The economies of synthesis. *Chem. Soc. Rev.* **38**, 3010–3021 (2009).
65. Colberg, J., K(Mimi) Hii, K. & Koenig, S. G. Importance of green and sustainable chemistry in the chemical industry: a joint virtual issue between acs sustainable chemistry & engineering and organic process research & development. *Org. Process Res. Dev.* **26**, 2176–2178 (2022).
66. Eastgate, M. D., Schmidt, M. A. & Fandrick, K. R. On the design of complex drug candidate syntheses in the pharmaceutical industry. *Nat. Rev. Chem.* **1**, 0016 (2017).
67. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics* **1**, 1–11 (2009).
68. Landrum, G. et al. *Rdkit: Open-source Cheminformatics* https://scholar.google.com/citations?view_op=view_citation&hl=zh-TW&user=xr9paY0AAAAJ&citation_for_view=xr9paY0AAAAJ:J_g5lzvAfSwC (2006).
69. Pupo, G., Properzi, R. & List, B. Asymmetric catalysis with CO_2: the direct α-allylation of ketones. *Angew. Chem. Int. Ed.* **55**, 6099–6102 (2016).

70. Boltukhina, E. V., Sheshenev, A. E. & Lyapkalo, I. M. Convenient synthesis of nonconjugated alkynyl ketones from keto aldehydes by a chemoselective one-pot nonaflation—base catalyzed elimination sequence. *Tetrahedron* **67**, 5382–5388 (2011).
71. Kende, A. S. & Fludzinski, P. Chloroacetylenes as Michael acceptors. ii. direct ethynylation and vinylation of tertiary enolates. *Tetrahedron Lett.* **23**, 2373–2376 (1982).
72. Nishimura, Y., Amemiya, R. & Yamaguchi, M. α-ethynylation reaction of ketones using catalytic amounts of trialkylgallium base. *Tetrahedron Lett.* **47**, 1839–1843 (2006).
73. Utaka, A., Cavalcanti, L. N. & Silva, L. F. Electrophilic alkynylation of ketones using hypervalent iodine. *Chem. Commun.* **50**, 3810–3813 (2014).
74. Wegener, M. & Kirsch, S. F. The reactivity of 4-hydroxy-and 4-silyloxy-1, 5-allenynes with homogeneous gold (i) catalysts. *Org. Lett.* **17**, 1465–1468 (2015).
75. Wang, J. et al. Protecting-group-free syntheses of ent-kaurane diterpenoids:[3+ 2+ 1] cycloaddition/cycloalkenylation approach. *J. Am. Chem. Soc.* **142**, 2238–2243 (2020).
76. Jang, D., Choi, M., Chen, J. & Lee, C. Enantioselective total synthesis of (+)garsubellin A. *Angew. Chem.* **133**, 22917–22921 (2021).
77. Hari, D. P., Caramenti, P. & Waser, J. Cyclic hypervalent iodine reagents: enabling tools for bond disconnection via reactivity umpolung. *Acc. Chem. Res.* **51**, 3212–3225 (2018).
78. Kingma, D. P. & Ba, J. Adam: a method stochastic optimization. Preprint at. https://arxiv.org/abs/1412.6980 (2014).
79. Chen, Z., Badrinarayanan, V., Lee, C. Y. & Rabinovich, A. Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, 794–803 (PMLR, 2018).
80. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).

Acknowledgements

V.S.B. acknowledges a generous allocation of high-performance computing time from the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC0205CH11231 using NERSC award BES-ERCAP0024372. The development of the methodology was supported by the NSF CCI grant (VSB, Award Number 2124511). The applications and experiments were supported by Boehringer Ingelheim.

Author contributions

The machine learning methods are developed by Y.S. and H.L., with equal contributions, under the guidance of V.S.B. The experimental validations are conducted by A.M.N. and P.Z., with equal contribution, and W.L., under the guidance of T.R.N. The experimental design and execution were advised and supervised by H.R.K., V.M., S.S., F.B., J.J.S., and T.R.N. The initial draft of the manuscript was primarily written by Y.S., with contributions from all authors during the final draft preparation.

Competing interests

V.S.B., H.L., and Y.S. have filed a patent application related to the work described in this manuscript. The patent is assigned to Yale University. The authors confirm that the patent filing does not affect the integrity or objectivity of the research presented. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52048-4.

Correspondence and requests for materials should be addressed to Timothy R. Newhouse or Victor S. Batista.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints