nature machine intelligence

Supplementary information

https://doi.org/10.1038/s42256-025-01066-y

Large language models to accelerate organic chemistry synthesis

In the format provided by the authors and unedited

Contents

Supplementary Notes

1 Open-source chemical reaction data for Chemma training and evaluation

1.1 Data description

- **The United States Patent and Trademark Office (USPTO)** (*1*), the largest open-source chemical reaction dataset, has various versions to operate on different chemical tasks. In our work, we utilize USPTO-50k, which is a filtered patent dataset with 50,000 atom-mapped reactions. It covers the 10 most common reaction types, among which the 5 most frequent categories are presented in Extended Data Fig. 1A. We explore the inner distribution characteristics of USPTO-50k in Extended Data Fig. 1B. It can be confirmed that the occurrence frequency of substrate scopes in the datasets follows a power-law distribution. It illustrates that the USPTO-50k dataset has a property of diversity and imbalance, which emphasizes the difficulty of LLMs to learn chemical knowledge.

- **The Open Reaction Database (ORD)**, launched by Kearnes et al. (*2*), is an open-access database with structured and standardized organic reaction data. The dataset encompasses about 1.7 million entries, encapsulating a wide variety of reaction types. It is designed to accommodate information about organic reactions spanning many distinct types. Among them, the four most common categories are illustrated in Extended Data Fig. 1A. We also explore the inner distribution characteristics of ORD datasets. In Extended Data Fig. 1C and 1E, we confirm that the occurrence frequency of chemical conditions e.g. catalysts, in the datasets follow a power-law distribution. The power-law distributions demonstrate the long-tail characteristics and a small number of categories account for the majority of the whole dataset. Such phenomenon is in light with the distribution of words in natural language, indicating that there is potential for tackling chemical tasks

with natural language models, like generative pre-trained transformers (GPTs).

- **High-throughput experimentation (HTE) reaction dataset**

  (1) **Pd-catalysed Buchward-Hartwig [HTE] C-N coupling reaction** dataset is released by Ahneman et al. in 2018 (*3*). In Extended Data Fig. 2A, the substrates are performed with varying combinations of aryl halide, Pd precatalyst, additive and base. Before utilizing the data in our organic synthesis tasks, we process the raw reaction data with a preliminary processing procedure, proposed in Ahneman et al.'s work (*3*). Initially, control reactions run without substrates were removed from valid reactions. Any additives that lacked corresponding reaction yields were subsequently and entirely omitted from the study. Finally, substrates and additives are labeled as "s#" and "a#", where "#" represents sequentially attributed numbers. Upon completion of the data processing phase, a residual of 3,600 reaction entries was retained for analysis (comprising 15 aryl chloride substrates, 20 additives, 3 based, and 4 ligands). Structures for all reaction components are shown in Extended Data Fig. 2A. Subsequently, a comprehensive yield assessment of the performed reactions are conducted.

  (2) **Pd-catalysed Suzuki–Miyaura [HTE] reaction** dataset is extracted from work by Perera et al. in 2018 (*4*), where the reaction yield is measured as a function of boronic acid derivative, aryl halide, ligand, base, and solvent (Extended Data Fig. 2B). By exploring a diverse range of reaction variables in the Suzuki-Miyaura coupling on nanomole scale at elevated temperatures, Perera et al. obtained 5,760 effective reactions with aryl halides and aryl N-methyliminodiacetic acid (MIDA) boronates.

  (3) **Imidazole C–H arylation reaction** dataset is extracted from the work by Shields et

al. in 2021 (*5*), where substrate scope contains 8 imidazoles and 8 aryl bromides. In Extended Data Fig. 2C we can see that this reaction generates 64 cross-coupled products labeled as $< imidazole >< arylbromide >$ (e.g. A1, B2). It should be noted that the original substrate scope is designed with 10 imidazoles and 10 aryl bromides, but the imidazole H, J, and aryl bromides 6, and 8 are later removed from the scope. Subsequently, a comprehensive yield assessment of the performed reactions was conducted.

(4) **Regioselectivie radical C–H functionalization of heterocycles reaction** dataset is extracted from Li et al. in 2020 (*6*). As a representative case study, this dataset consists of reactions about radical C–H functionalization over heterocycles. It provides the reaction SMILES including different arene scaffolds, substitution patterns and radicals, along with the free energy barriers (ΔG) calculated by DFT descriptors. Finally, the designed sample space presents a collection of 3406 radical C–H functionalization reactions, and 5174 regioisomeric competitions.

(5) **Chiral phosphoric acid–catalyzed thiol addition reaction** the dataset is extracted from Zahrt et al. in 2019 (*7*). In this reaction, both yield and enantioselectivity predictions are tested using Pd-catalyzed C–N cross-coupling reactions (*3*) between 4-methylaniline and aryl halides as well as chiral phosphoric acid (CPA)-catalyzed thiol addition to *N*-acylimines (*7*). These high-quality datasets provide valuable statistics that map the complete synthetic space given the studied reaction components.

- **Pd-catalysed Buchward-Hartwig [ELN] reaction** dataset is extracted from the work by Saebi et al. in 2023 (*8*). Generally, ELNs datasets have very different characteristics. Buchward-Hartwig [ELN] covers a much wider chemical space, with 340 aryl halides,

260 amines, 24 ligands, 15 bases and 15 solvents. With 1,000 examples to cover 4.7×108 possible combinations of reactants, ligands, bases and solvents, the dataset is much sparser. As a result, there are only a very small number (35 of 781 data entries) of cases where reactions with identical conditions and substrates were run multiple times.

1.2 Data usage for each chemical task

For each specific chemical organic synthesis task, we indicate the dataset employed for both training and testing, which can be seen as follows:

- **Forward prediction & retrosynthesis** For the forward prediction and retrosynthesis task, we utilize the benchmark United States Patent and Trademark Office (USPTO-50k) dataset, extracted from the US patent literature (*1*). The USPTO-50k is a filtered patent dataset with 50,000 atom-mapped reactions, where 40,000 reactions as training sets, 5,000 reactions for validation and 5,000 reactions as test sets. It covers the ten most common reaction types, among which the five most frequent categories are presented in Extended Data Fig. 1A. USPTO-50k contains canonized reactants, reagents, and products that are split by a separator token.

- **Condition generation** For the condition generation task, we utilize ORD datasets for training and imidazole C–H arylation [HTE] reaction for ligand recommendation evaluation. Further, USPTO datasets contain reagent compounds in condition variables, we apply USPTO for reagent recommendation tasks. We construct pair-wise Q&A instruction prompts using SMILES of the entire chemical compounds, subsequently designating the ligands as corresponding responses.

- **Yield prediction** For yield prediction tasks, we leverage both open reaction datasets (ORD) and USPTO-50k for stage-1 fine-tuning (see Fig. 2 in the main text). Subse-

quently, we evaluate the performance of yield prediction on four efficient experiment datasets, Pd-catalysed Suzuki–Miyaura reaction [HTE] (*4*), Pd-catalysed Buchward-Hartwig [ELN]and [HTE] reaction (*3, 8*), and imidazole C–H arylation reaction [HTE] (*5*). For the Suzuki–Miyaura and C–H arylation reactions, we allocate 30% reactions of the full capacity whose substrates are not included in the training set as an out-of-sample test set; For the Buchward-Hartwig [ELN] reaction, we randomly split 30% of the entire dataset for testing. For the Buchward-Hartwig [HTE] reaction, we also examine the performances of models via different training/testing split strategies, including different fractions of substrates, splitting randomly and by additives which can be seen in Extended Data Fig. 4.

- **Selectivity prediction** For regioselectivity, we focus on the dataset created by Baxter et al. (*9*), which consists of reactions about radical C–H functionalization over heterocycles. The dataset provides the reaction SMILES including different arene scaffolds, substitution patterns, and radicals, along with the free energy barriers (ΔG) calculated by DFT annotations. We randomly select 30% of the original dataset as the out-of-sample test set. For enantioselectivity, we focus on chiral phosphoric acid–catalyzed thiol addition to *N*-acylimines from Zahrt et al. (*7*). The dataset includes 25 reactants combinations of 5 imines and 5 thiols, and 43 catalysts, making up 1,075 reactions. We randomly select 600 reactions for training and 475 reactions for testing.

2 Prompt design and training strategies

2.1 Prompt design

Prompts serve as important instructional inputs in LLMs. They provide a useful means of controlling model output and setting the context and intention for the model's response. Carefully designed prompts can guide the model to generate more accurate, relevant, and detailed re-

sponses. Often, prompts can be tuned and specialized to specific tasks, benefiting from the vast knowledge encoded during the pre-trained models' unsupervised learning phase.

In this work, we design a prompt management consisting of four parts: role identification, task description, examples, and output instruction, which can be seen in Supplementary Fig. 1. '**Role identification**' aims to help LLMs identify the chemist (quizzer) and assistant (answerer). By establishing roles, LLMs are encouraged to generate responses that align with the expertise expected in the chemistry domain. '**Task description**' provides a comprehensive definition of the task's contents, ensuring that LLMs have a clear understanding of the specific tasks. It also includes critical definitions to clarify terms or concepts that are specialized in chemical tasks such as SMILES explanation of reactants, products, and conditions for a chemical reaction. The next component '**Examples**' is designed to define the user input including the illustration of reaction components, which serve as the evidence for all chemical tasks, allowing LLMs to leverage the given information to generate reasonable responses. Finally, '**Output Instruction**' specifies the desired format for the response, which is specified directly through a natural language instruction (e.g. *Could you suggest potential solvents that could have been used in the given chemical reaction?*) but could also be indirectly through either few-shot examples (e.g. *We already obtain the yield of this reaction is 11.94%, could you please give me some new ligands that potential get higher yields?*). Here, we restrict the output to a simplified molecular-input line-entry system (SMILES) format on forward prediction, retrosynthesis, and condition generation tasks, and numbers within 0-100 on yield prediction.

For several organic synthesis tasks, we construct different supervised fine-tuning datasets for training, which are composed of instruction contents and answer pairs. The instruction contents can be split into two components: instruction templates and input reaction data. Here, we depict the demonstration prompts for retrosynthesis and condition generation, which can be seen in Supplementary Fig. 1. "Chemist" in the instruction template refers to role identification.

"*Considering a chemical reaction, SMILES is a sequenced-based string used to encode the molecular structure. A chemical reaction includes reactants, conditions, and products*" is the task description, and input data for the question is the products of a specific reaction, then we leverage GPT-4 to generate question templates like "*What product might be formed from the reaction involving the given reactants?*". The corresponding answer is the SMILES of reactants used in this reaction. The difference in prompts for the two tasks lies in the "examples". From the instruction prompts for condition generation, we can see that the complete SMILES of reaction is required to allow the Chemma to generate conditions (e.g., catalysts) used in the reaction.

2.2 Training strategies

We incorporate three crucial stages for training. In stage 1, we construct multi-task Q&A datasets through the amalgamation of demonstrative prompts and reaction data to fully fine-tune the original LLaMA-2-7b model, termed as Chemma-SFT. In stage 2, we construct pair-wise Q&A ranking datasets from HTE. These datasets contain reactions under varying conditions, and we expect Chemma to learn the performance difference between the two reaction conditions based on experiment results. Next, we train a reward model, Chemma-RM, to generate refined reaction conditions. Following the training strategy of reinforcement learning from human feedback (RLHF) (*10*), we train Chemma-RM with the Proximal Policy Optimization (PPO) algorithm (*11*).

In experiments, we fine-tune from the pre-trained weights of the LLaMA-2-7b model. We set batch sizes as 64 for Chemma-SFT and 16 for Chemma-RM. The learning rate is set as 9.65e-6, and we train for 4 epochs for all in approximately one week using 48 $\times$ NVIDIA A800 GPUs. For multi-GPU training, we employ the distributed data parallel (DDP) strategy. DDP is chosen due to its efficiency in synchronizing gradients across multiple GPUs, minimizing

communication overhead, and ensuring reproducibility. We do not use data parallel (DP) or fully sharded data parallel (FSDP) strategies, as DDP provides a balance between scalability and implementation complexity for our specific use case. All reported performances of Chemma are evaluated on the final weights after training.

3 Retrosynthesis

The existing single-step retrosynthesis prediction methods can be categorized into three classes: template-based, semi-template-based, and template-free with the dependency on reaction templates (*12, 13*). Within this knowledge, Chemma generates reactants directly from products, meaning that it is clearly a template-free method.

Since the test case prompt we illustrated represents a common amine bond formation reaction, it is not surprising that GPT-4o, having been pre-trained on a very large-scale corpus (definitely including some chemical corpus), can generate the correct answer. However, our extensive experiments have confirmed that a general-purpose LLM like GPT-4o is not capable of handling most retrosynthesis problems. Thus, we test the performance of several LLMs for the retrosynthesis task on the USPTO-50k dataset, including GPT-4, GPT-3.5, GPT-4o, and Llama2-13B. The performance of LLM is usually influenced by two key strategies: retrieval methods (Random and Scaffold) and a grid search over the hyperparameter k with values of 5 and 20, both of which are the configuration during the generation process. For GPT-4o, we randomly select 500 samples from the USPTO-50k test set for evaluation. The results can be seen in Supplementary Table 2, and the statements can be found in Supplementary Note 3. We observe that all general-purpose LLMs underperform compared to existing machine learning baselines that have been specifically trained on large-scale reaction datasets. Specifically, the top-1 accuracy of the general-purpose LLM, GPT-4o, reaches only 21.9%, whereas the baseline method Chemformer achieves a markedly higher top-1 accuracy of 53.6%. This suggests

that while LLMs exhibit strong generalization capabilities, they may struggle with domain-specific optimization tasks that require extensive chemical knowledge and structured learning from reaction data. Next, we list the performance of several domain-specific models in Supplementary Table Supplementary Table 2. These encompass a broad range of methodologies, including template-based, template-free, and semi-template-based approaches, examining both top-1, top-3, and top-5 predictive results. Notably, we can see that Chemma demonstrates the most favorable performance in the realm of retrosynthesis tasks, where achieves 72.2% top-1 accuracy when compared with other baseline methods such as the template-free Retroformer (51.4%), the template-based LocalRetro (53.4%), and semi-template-based method G2Retro (54.1%). Furthermore, Chemma also outperforms the current state-of-the-art method, G2G, with a significant exceed of 25.1% (55.1% v.s. 72.2%). The reason why Chemma achieves much better performance is that LLMs have a strong representation capability of molecules and reactions by previous fine-tuning on large datasets. Furthermore, sequence-to-sequence generation makes it more likely to identify functional group incompatibilities in every reaction center so that it will not be confused by common substructures between products and reactants.

To assess the capability of designing synthetic routes for organic molecules, we select three drug molecules as the target product and predict reactants using Chemma. The designed routes can be seen in Extended Data Fig. 3. Osimertinib, is known for its effectiveness in treating patients with non-small cell lung cancer (*14*). Chemma successfully delineates a four-step synthesis, tracing the synthetic pathway from commercially available pyrimidines to the final product Osimertinib. The initial step of the reverse synthesis is an acylation reaction, ranking the first in order of likelihood, followed by another rank-1 reduction of the nitro group. It correctly identifies two consecutive nucleophilic aromatic substitution steps as the top-1 choice. Notably, even though predicted intermediate reactants in step 3 are not exactly matched with those from the literature, we still confirm that this process is reasonable, which is verified by experts'

knowledge. In the final step, the model's highest probability prediction is the aryl substitution reaction, which is also the rank-1 choice.

For the second case study, we test Vonoprazan, a novel potassium-competitive acid blocker used for the treatment and prevention of acid-related disease (*15*). Chemma predicts the three-step synthetic route. In the first step, the intermediate product is formed by a reductive amination reaction, which does not hit the reference. Next, Chemma predicts 5-(2-fluorophenyl)-1H-pyrrole-3-carbaldehyde reacted with pyridine-3-sulfonyl chloride with rank-1 score, which performs sulfonylation reaction. Subsequently, Chemma identifies that the aldehyde group could be used to reduce the ester group for step 3.

For the third case study, we test Mitapivat, a small molecule allosteric activator of pyruvate kinase R (PKR). Chemma successfully predicts the three-step core synthetic route reported in (*16*). The first step entails an amide formation reaction, which Chemma places at rank-1, yielding the reactants 1-(cyclopropylmethyl)piperazine and 4-(quinoline-8-sulfonamido)benzoate. Next, for the synthesis of 4-(quinoline-8-sulfonamido)benzoic acid, Chemma predicts ester hydrolysis process with rank-1 choice, which is a functional group protection strategy. Subsequently, formation of the sulfonamide, the last step of retrosynthesis effectively deconstructs the target molecule into readily available precursors.

For the more complex case, we test Pirtobrutinib, an anticancer medication that is used to treat mantle cell lymphoma. It inhibits B cell lymphocyte proliferation and survival by binding and inhibiting Bruton's tyrosine kinase (BTK). Chemma predicts a six-step synthetic route. Notably, the referenece (*17*) also employs six steps to synthesize the target molecule. In the first step, the carboxylic acid is converted to an acid chloride named 4-((5-fluoro-2-methoxybenzamido)methyl)benzoic acid, which does not hit the synthetic route reported in the literature. However, we think this step activates the acid for subsequent coupling reactions. Next, Chemma predicts the amide formation via nucleophilic acyl substitution reaction. The im-

portant synthesis step is the fifth step, that Chemma predicts a condensation reaction with rank-1 choice. The ketone or aldehyde reacts with a hydrazine derivative (such as CF_3NHNH_2) to form (S)-N-(4-(5-amino-4-cyano-1-(1,1,1-trifluoropropan-2-yl)-1H-pyrazol-3-yl)benzyl)-5-fluoro-2-methoxybenzamide.

We select Ritlecitinib as the second complex case for evaluation. Ritlecitinib is a kinase inhibitor which inhibits Janus kinase 3 and tyrosine kinase. Chemma predicts a six-step synthesis for this molecule, whereas the reference method (*18*) proposes an eight-step route. In the first step, Chemma predicts amide formation from amine and diketene. The amine reacts with diketene, which opens the diketene ring, forming an acetoacetate amide. Then we consider the formation of imidazopyridine (step-4) predicted by Chemma which has rank-1 probability. The acyl chloride reacts with the secondary amine in the presence of a base to form an amide bond. Then the amide reacts with the chloroimidazopyridine through nucleophilic aromatic substitution. The chlorine on the imidazopyridine is displaced by the amine, forming a C-N bond.

4 Yield prediction

4.1 HTE datasets

Extended Data Fig. 4 offers an insightful comparison of yield prediction capabilities between the random forest (RF) algorithm and Chemma for two pivotal Pd-catalyzed chemical reactions: Suzuki-Miyaura and Buchward-Hartwig. Extended Data Figs. 4D-E illustrate a notable decrease in predictive accuracy when the training dataset is reduced from 90% to just 5%. This trend highlights the data dependency of both models, with their performance deteriorating in conditions of sparse data. Noteworthy is Chemma's consistent superiority over RF across varying data proportions, emphasizing its more effective use of limited datasets.

In Extended Data Figs. 4D-E, we examine the models' performance on reactions isolated

by random data splits. We compare the performance of yield prediction with varying fractions of training data on Suzuki–Miyaura and Buchward-Hartwig reactions, respectively. It is clear that RF performs better than Chemma when the fraction of training data is lower than 60%. Notably, Chemma achieves an R^2 value of 0.94 (0.94) on the Suzuki–Miyaura (Buchward-Hartwig) reactions when utilizing 90% data for training, which performs better than RF (R^2 0.9 and 0.93).

In Extended Data Figs. 4F-G, we examine the models' performance on reactions isolated by substrate scopes. For the Suzuki-Miyaura and Buchward-Hartwig reaction, the entire reaction encompasses 12 substrates and 15 aryl chloride substrates, respectively. We defines four scenarios with distinct training and testing substrate sets, verified through ten-fold cross-validation to ensure comparability. The performance of prediction is robust, maintaining stability across different substrates, with the maximum R^2 value of 0.9. While RF model exhibits significant fluctuations, particularly in the Buchward-Hartwig reaction, primarily due to the diversity of substrates.

The analysis is further extended in Extended Data Figs. 4H-I, which analyzes the models' performance with regard to ligand and additive scopes, respectively. Chemma sustains a commendable prediction performance (from R^2 value of 0.56 to 0.67) for the Suzuki-Miyaura reaction when divided by eleven ligands. Conversely, the performance of RF drops precipitously, with the R^2 value of 0.43, implying a deficiency in its capacity to generalize across ligand variations. In the Buchward-Hartwig reaction, we test four additive cases for evaluation, following the configuration reported by Ahneman et al. (*3*). From the results, we can see that Chemma shows competitive predictive results in every scenario, even if the test additive set is toxic.

This comprehensive examination, inclusive of randomized data splits, substrate scopes, and divisions by ligands or additives, affords a thorough appraisal of the Chemma's proficiency.

Chemma's uniform excellence suggests a resilient model less affected by training condition variability, likely attributable to its advanced capacity for representation learning and sequence-to-sequence generation, as suggested in the manuscript. Conversely, RF depends more heavily on the DFT descriptors and size of the dataset, with its efficacy more impacted by chemical context variations.

4.2 Literature datasets

The application of machine learning (ML) for reaction yield prediction has demonstrated remarkable accuracy when applied to high-throughput computational and experimental data. However, generating such high-throughput data often involves substantial experimental or computational efforts and typically restricts the exploration of chemical reactions to a limited range of variables. The chemistry community has turned to extensive repositories of reaction data available in the literature and databases to overcome these constraints. Unfortunately, ML-based methods using literature-derived data exhibit lower predictive performance, exposing a significant gap in the capabilities of current ML models compared to their performance on high-throughput sources. To address this, we evaluate the performance of Chemma for yield prediction using literature-derived data. Using the dataset compiled by Li et al. (*19*), we focus on Pd-catalyzed carbonylation reactions. This dataset comprises 2,512 records of homogeneous carbonylation reactions involving Pd catalysts, manually compiled from 113 literature sources identified via the Web of Science. Each record includes detailed information about the substrate, product, carbon source, catalyst precursor, ligand, base, oxidant, additive, and solvent (Extended Data Fig. 5A). The yield distribution of this reaction data shows a bimodal pattern, with peaks in the low (0–10%) and medium (70–80%) yield intervals, as illustrated in the Extended Data Fig. 5B. To assess Chemma's generalization capability, we employed an out-of-sample testing strategy. The training set included data from literature sources with IDs ≤ 100,

while the testing set exclusively comprised data from literature sources with IDs >100. This setup challenged Chemma to predict reaction performance in unexplored chemical spaces. As shown in Extended Data Fig. 5C, Chemma achieves an RMSE of 16.16 and a coefficient of determination R^2 value of 0.74, demonstrating acceptable performance. Furthermore, compared to the R^2 value of 0.51 in their study (*19*), Chemma achieves a significant improvement with a 47.1% gain. In a word, through challenges in prediction with literature-derived data, Chemma still offers acceptable performance results, highlighting its effectiveness for tackling complex chemistry tasks.

5 Selectivity prediction

5.1 Regioselectivity prediction for the radical C–H functionalization over heterocycles

In our work, we examine the performance of regioselectivity prediction. Extended Data Fig. 6 validates the test set performance by Chemma. Each training set with different proportions is selected randomly from the original full data set, and 30% data of the entire datasets are randomly selected as test sets. From the results, we find that Chemma performs well even when the fractions of real data are as low as a mere 5% for chiral phosphoric acid-catalyzed thiol addition reaction. Consequently, we select the task of regioselectivity prediction as a particular case study to elucidate our insights. We select a radical C–H functionalization over heterocycles reaction created by Li et al. (*6*), with free energy barriers (ΔG) calculated by DFT descriptors. It also provides the SMILES of reactions including different arene scaffolds, substitution patterns, and radicals. Given a reaction, we apply Chemma to predict the maximum probability of multiple reactive sites. The predicted results are visualized in Extended Data Fig. 6. Also, from the results we can see that Chemma depicts comparable performance of predicting the correct reactive sites (Extended Data Figs. 6A-C). However, Chemma does not adequately address all

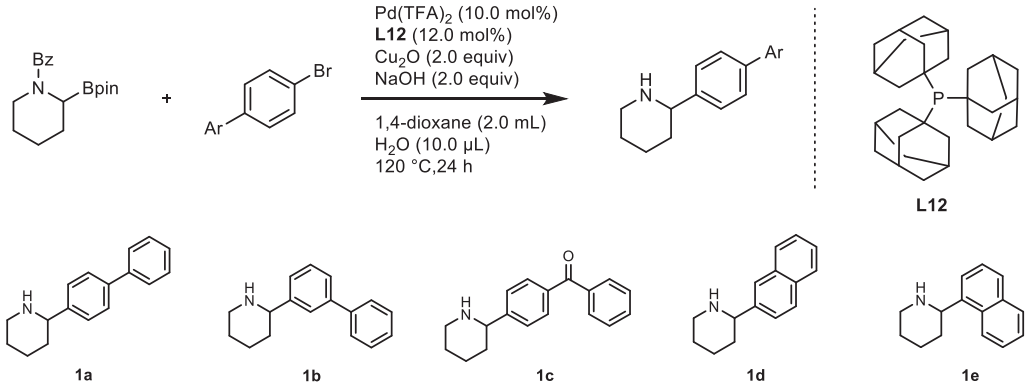
potential scenarios. For instance, incorrect predictions of reactive sites may occur when multiple potential reactive sites are in close proximity, indicating an area needing further refinement.

5.2 Enantioselectivity prediction for the Cu-catalysed enantioselective addition reaction

High-throughput screening (HTS) of chiral catalysts for enantioselective reactions is a powerful promising approach to the discovery of new catalysts. Combinatorial libraries are particularly useful for the optimization of catalytic systems. However, the number of experiments that can be screened within reasonable resource constraints is typically limited. Here, we also illustrate Chemma's capability for the virtual screening of catalysts by predicting the enantiomeric excess given the molecular structure. The reaction data is obtained from Long et. al (*20*) for the enantioselective addition of diethyl zinc to benzaldehyde in the presence of a racemic catalyst (RC) and an enantiopure chiral additive (CA) (*21*). The 65 objects were randomly split into three subsets of the same size. Each subset covers the whole range of observed ee values (-71% to +16%). The results can be seen in Supplementary Fig. 4. Chemma outperforms all baseline models with R^2 values from 0.896 to 0.96, and achieves the lowest RMSE values, indicating superior accuracy in predictions. This suggests that Chemma is well-suited for tasks requiring high accuracy in predicting enantioselective outcomes, making it a valuable tool in computational chemistry.

6 Wet-experimental details

Procedure: In the glove box, (2-(1-benzoylpiperidin-2-yl)-4,5,5-trimethyl-1,3,2-dioxaborolan-4-yl)methylium (1.00 equiv), Phenyl bromide (2.00 equiv), $Pd(TFA)_2$ (10.0 mol%), L12 (12.0 mol%), Cu_2O (2.0 equiv), NaOH (2.0 equiv) were added to a 4.0 mL vial with a screw-top septum. Anhydrous 1,4-dioxane (2.0 mL) and H_2O (10.0 μL) were added, and the reaction



mixture was stirred at 120 °C for 24 h. The reaction solution was then cooled to rt, quenched by H_2O and diluted with DCM. The organic layer was dried over Na_2SO_4, filtered, evaporated under reduced pressure and concentrated *in vacuo*. The crude product was purified by column chromatography on silica gel.

2-([1,1'-Biphenyl]-4-yl)piperidine (1a). In the glove box, (2-(1-benzoylpiperidin-2-yl)-4,5,5-trimethyl-1,3,2-dioxaborolan-4-yl)methylium (1.00 equiv, 0.2 mmol, 63.0 mg), 4-bromo-1,1'-biphenyl (2.00 equiv, 0.4 mmol, 93.0 mg), $Pd(TFA)_2$ (10.0 mol%, 0.02 mmol, 6.7 mg), L12 (12.0 mol%, 0.024 mmol, 10.5 mg), Cu_2O (2.0 equiv, 0.4 mmol, 58.0 mg), NaOH (2.0 equiv, 0.4 mmol, 24.0 mg) were added to a 4.0 mL vial with a screw-top septum. Anhydrous 1,4-dioxane (2.0 mL) and H_2O (10.0 μL) were added, and the reaction mixture was stirred at 120 °C for 24 h. The reaction solution was then cooled to rt, quenched by H_2O and diluted with DCM. The organic layer was dried over Na_2SO_4, filtered, evaporated under reduced pressure and concentrated *in vacuo*. The crude material was purified by column chromatography on SiO_2 (Methanol/ Dichloromethane = 1:8) to give 1a (14.9 mg, 67%) as a colorless foam: ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.67 (d, J = 7.9 Hz, 1H), 7.54 (d, J = 7.9 Hz, 2H), 7.48 (d, J = 7.6 Hz, 2H), 7.38 (t, J = 7.5, 7.5 Hz, 2H), 7.33 (t, J = 7.2, 7.2 Hz, 1H), 4.02 (dd, J = 12.4, 2.8 Hz, 1H), 3.11 (d, J = 12.6 Hz, 1H), 2.81 (td, J = 13.2, 13.1, 3.2 Hz, 1H), 2.18 (dd, J = 13.3, 9.8

Hz, 1H), 2.08 – 1.88 (m, 2H), 1.65 (d, J = 14.4 Hz, 1H), 1.54 (dd, J = 15.5, 11.9 Hz, 1H), 1.24 (d, J = 10.6 Hz, 1H); ^{13}C NMR (126 MHz, Chloroform-d) δ 141.9, 140.1, 135.4, 128.8 (d, J = 32.5 Hz), 127.6 (d, J = 13.0 Hz), 127.0, 61.1, 45.8, 30.1, 23.2, 21.6; HRMS (ESI) m/z calcd for $C_{17}H_{20}N$ $[M + H]^+$ 238.1594, found 238.1596.

2-([1,1'-Biphenyl]-3-yl)piperidine (1b). In the glove box, (2-(1-benzoylpiperidin-2-yl)-4,5,5-trimethyl-1,3,2-dioxaborolan-4-yl)methylium (1.00 equiv, 0.2 mmol, 63.0 mg), 3-bromo-1,1'-biphenyl (2.00 equiv, 0.4 mmol, 93.0 mg), $Pd(TFA)_2$ (10.0 mol%, 0.02 mmol, 6.7 mg), L12 (12.0 mol%, 0.024 mmol, 10.5 mg), Cu_2O (2.0 equiv, 0.4 mmol, 58.0 mg), NaOH (2.0 equiv, 0.4 mmol, 24.0 mg) were added to a 4.0 mL vial with a screw-top septum. Anhydrous 1,4-dioxane (2.0 mL) and H_2O (10.0 μL) were added, and the reaction mixture was stirred at 120 °C for 24 h. The reaction solution was then cooled to rt, quenched by H_2O and diluted with DCM. The organic layer was dried over Na_2SO_4, filtered, evaporated under reduced pressure and concentrated *in vacuo*. The crude material was purified by column chromatography on SiO_2 (Methanol/ Dichloromethane = 1:8) to give 1b (13.7 mg, 56%) as a yellow foam: ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.84 (s, 1H), 7.68 (d, J = 7.6 Hz, 2H), 7.52 (t, J = 5.9, 5.9 Hz, 2H), 7.41 (t, J = 7.5, 7.5 Hz, 2H), 7.33 (dt, J = 14.2, 7.5, 7.5 Hz, 2H), 4.08 – 3.87 (m, 1H), 3.04 (d, J = 12.1 Hz, 1H), 2.72 (t, J = 13.0, 13.0 Hz, 1H), 2.16 (q, J = 12.6, 12.2, 12.2 Hz, 1H), 1.92 (t, J = 16.9, 16.9 Hz, 3H), 1.58 (d, J = 14.2 Hz, 1H), 1.53 – 1.44 (m, 1H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 141.6, 140.2, 129.5, 127.7 (d, J = 11.9 Hz), 127.3, 126.9 (d, J = 2.8 Hz), 61.4, 45.7, 30.1, 23.2, 21.7; HRMS (ESI) m/z calcd for $C_{17}H_{20}N$ $[M + H]^+$ 238.1593, found 238.1596.

Phenyl(4-(piperidin-2-yl)phenyl)methanone (1c). In the glove box, (2-(1-benzoylpiperidin-2-yl)-4,5,5-trimethyl-1,3,2-dioxaborolan-4-yl)methylium (1.00 equiv, 0.2 mmol, 63.0 mg), (4-bromophenyl)(phenyl)methanone (2.00 equiv, 0.4 mmol, 104.4 mg), $Pd(TFA)_2$ (10.0 mol%, 0.02 mmol, 6.7 mg), L12 (12.0 mol%, 0.024 mmol, 10.5 mg), Cu_2O (2.0 equiv, 0.4 mmol, 58.0 mg), NaOH (2.0 equiv, 0.4 mmol, 24.0 mg) were added to a 4.0 mL vial with a screw-top

septum. Anhydrous 1,4-dioxane (2.0 mL) and H_2O (10.0 μL) were added, and the reaction mixture was stirred at 120 °C for 24 h. The reaction solution was then cooled to rt, quenched by H_2O and diluted with DCM. The organic layer was dried over Na_2SO_4, filtered, evaporated under reduced pressure and concentrated *in vacuo*. The crude material was purified by column chromatography on SiO_2 (Methanol/ Dichloromethane = 1:8) to give 1c (16.7 mg, 63%) as a colorless foam: ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.79 – 7.75 (m, 6H), 7.59 (t, J = 7.4, 7.4 Hz, 1H), 7.47 (t, J = 7.6, 7.6 Hz, 2H), 4.06 (d, J = 12.1 Hz, 1H), 3.19 (d, J = 12.7 Hz, 1H), 2.86 (t, J = 12.5, 12.5 Hz, 1H), 2.24 – 2.15 (m, 1H), 2.05 (dd, J = 30.0, 13.4 Hz, 4H), 1.81 (d, J = 14.6 Hz, 1H), 1.60 (d, J = 11.0 Hz, 1H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 195.8, 140.1, 138.4, 137.0, 132.7, 130.5, 130.0, 128.4, 128.2, 61.1, 45.7, 30.1, 22.9, 21.3; HRMS (ESI) m/z calcd for $C_{18}H_{20}NO$ $[M + H]^+$ 266.1545, found 266.1543.

2-(Naphthalen-2-yl)piperidine (1d). In the glove box, (2-(1-benzoylpiperidin-2-yl)-4,5,5-trimethyl-1,3,2-dioxaborolan-4-yl)methylium (1.00 equiv, 0.2 mmol, 63.0 mg), 2-bromonaphthalene and 1-bromonaphthalene(2.00 equiv, 0.4 mmol, 83.1 mg), $Pd(TFA)_2$ (10.0 mol%, 0.02 mmol, 6.7 mg), L12 (12.0 mol%, 0.024 mmol, 10.5 mg), Cu_2O (2.0 equiv, 0.4 mmol, 58.0 mg), NaOH (2.0 equiv, 0.4 mmol, 24.0 mg) were added to a 4.0 mL vial with a screw-top septum. Anhydrous 1,4-dioxane (2.0 mL) and H_2O (10.0 μL) were added, and the reaction mixture was stirred at 120 °C for 24 h. The reaction solution was then cooled to rt, quenched by H_2O and diluted with DCM. The organic layer was dried over Na_2SO_4, filtered, evaporated under reduced pressure and concentrated *in vacuo*. The crude material was purified by column chromatography on SiO_2 (Methanol/ Dichloromethane = 1:8) to give 1d (11.7 mg, 55%) as a colorless foam: ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.06 (s, 1H), 7.83 (d, J = 7.8 Hz, 1H), 7.73 (dd, J = 10.9, 8.2 Hz, 2H), 7.65 (d, J = 8.4 Hz, 1H), 7.44 (dt, J = 14.2, 7.1, 7.1 Hz, 2H), 4.04 (dd, J = 12.5, 2.9 Hz, 1H), 2.93 – 2.82 (m, 1H), 2.62 – 2.51 (m, 1H), 2.24 – 2.13 (m, 1H), 1.88 (dd, J = 45.6, 14.1 Hz, 3H), 1.46 – 1.31 (m, 2H); ^{13}C NMR (126 MHz, Chloroform-d) δ 133.6, 133.4, 133.0,

128.9, 128.5, 127.8, 126.8, 126.5, 125.2, 61.4, 45.6, 30.0, 23.0, 21.3; HRMS (ESI) m/z calcd for $C_{15}H_{18}N$ $[M + H]^+$ 212.1439, found 212.1435.

2-(Naphthalen-1-yl)piperidine (1e). In the glove box, (2-(1-benzoylpiperidin-2-yl)-4,5,5-trimethyl-1,3,2-dioxaborolan-4-yl)methylium (1.00 equiv, 0.2 mmol, 63.0 mg), 1-bromonaphthalene (2.00 equiv, 0.4 mmol, 83.1 mg), $Pd(TFA)_2$ (10.0 mol%, 0.02 mmol, 6.7 mg), L12 (12.0 mol%, 0.024 mmol, 10.5 mg), Cu_2O (2.0 equiv, 0.4 mmol, 58.0 mg), NaOH (2.0 equiv, 0.4 mmol, 24.0 mg) were added to a 4.0 mL vial with a screw-top septum. Anhydrous 1,4-dioxane (2.0 mL) and H_2O (10.0 μL) were added, and the reaction mixture was stirred at 120 °C for 24 h. The reaction solution was then cooled to rt, quenched by H_2O and diluted with DCM. The organic layer was dried over Na_2SO_4, filtered, evaporated under reduced pressure and concentrated *in vacuo*. The crude material was purified by column chromatography on SiO_2 (Methanol/ Dichloromethane = 1:8) to give 1e (9.5 mg, 45%) as a colorless foam: ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.96 (d, J = 8.4 Hz, 1H), 7.83 (t, J = 9.1, 9.1 Hz, 2H), 7.75 (d, J = 8.2 Hz, 1H), 7.51 (dt, J = 22.5, 7.4, 7.4 Hz, 2H), 7.36 (t, J = 8.0, 8.0 Hz, 1H), 4.68 (d, J = 11.5 Hz, 1H), 3.19 (d, J = 12.1 Hz, 1H), 2.89 (t, J = 12.8, 12.8 Hz, 1H), 2.04 (d, J = 13.8 Hz, 1H), 1.91 (dd, J = 26.1, 11.3 Hz, 3H), 1.67 (d, J = 16.3 Hz, 2H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 133.8, 130.1, 129.1, 128.8, 126.0, 125.7, 124.0, 122.3, 57.1, 46.9, 32.1, 24.1, 22.9; HRMS (ESI) m/z calcd for $C_{17}H_{19}N_5$ $[M + H]^+$ 212.1439, found 212.1434.

Prompts templates for retrosynthesis

Instruction Template

Chemist: Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecular structure. A chemical reaction includes reactants, conditions and products. Thus, SMILES for products of reactions are ***Cc1ccc(Nc2ccc(C(F)(F)F)cc2)cc1***, Can you decipher which reactants would lead to this resulting product?, Assistant:

Input

Cc1ccc(Nc2ccc(C(F)(F)F)cc2)cc1

Can you decipher which reactants would lead to this resulting product?

Answer

Cc1ccc(Cl)cc1.Nc1ccc(C(F)(F)F)cc1

Prompts templates for condition generation

Instruction Template

Chemist: Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecular structure. A chemical reaction includes reactants, conditions and products. Thus,reactants for this reaction are FC(F)(F)c1ccc(Cl)cc1.Cc1ccc(N)cc1, SMILES for products of reactions are Cc1ccc(Nc2ccc(C(F)(F)F)cc2)cc1, then the reaction can be described as FC(F)(F)c1ccc(Cl)cc1.Cc1ccc(N)cc1>>Cc1ccc(Nc2ccc(C(F)(F)F)cc2)cc1. Could you offer catalysts for this specific chemical reaction? Assistant:

Input

FC(F)(F)c1ccc(Cl)cc1.Cc1ccc(N)cc1>>Cc1ccc(Nc2ccc(C(F)(F)F)cc2)cc1

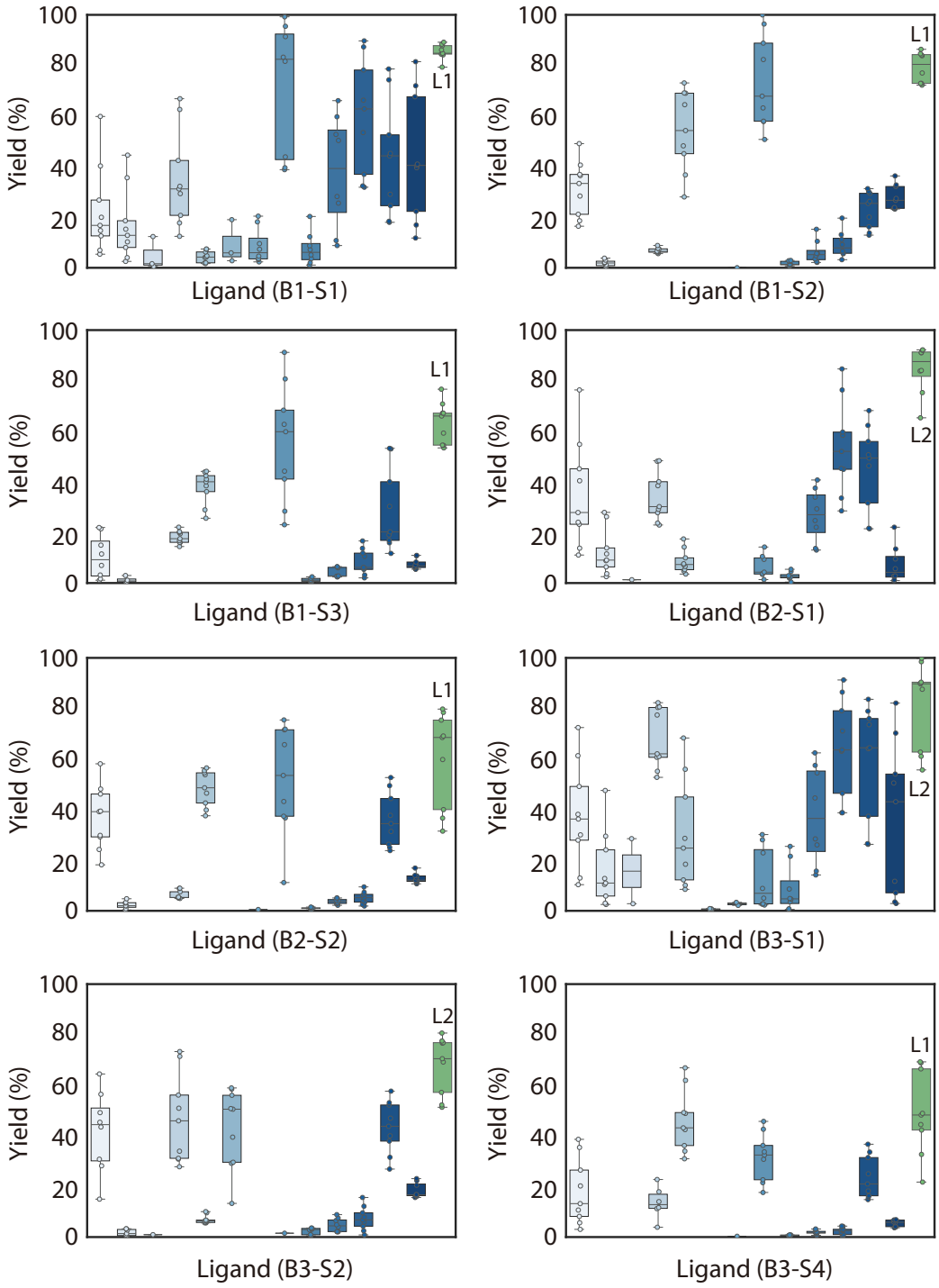
Could you offer catalysts for this specific chemical reaction?

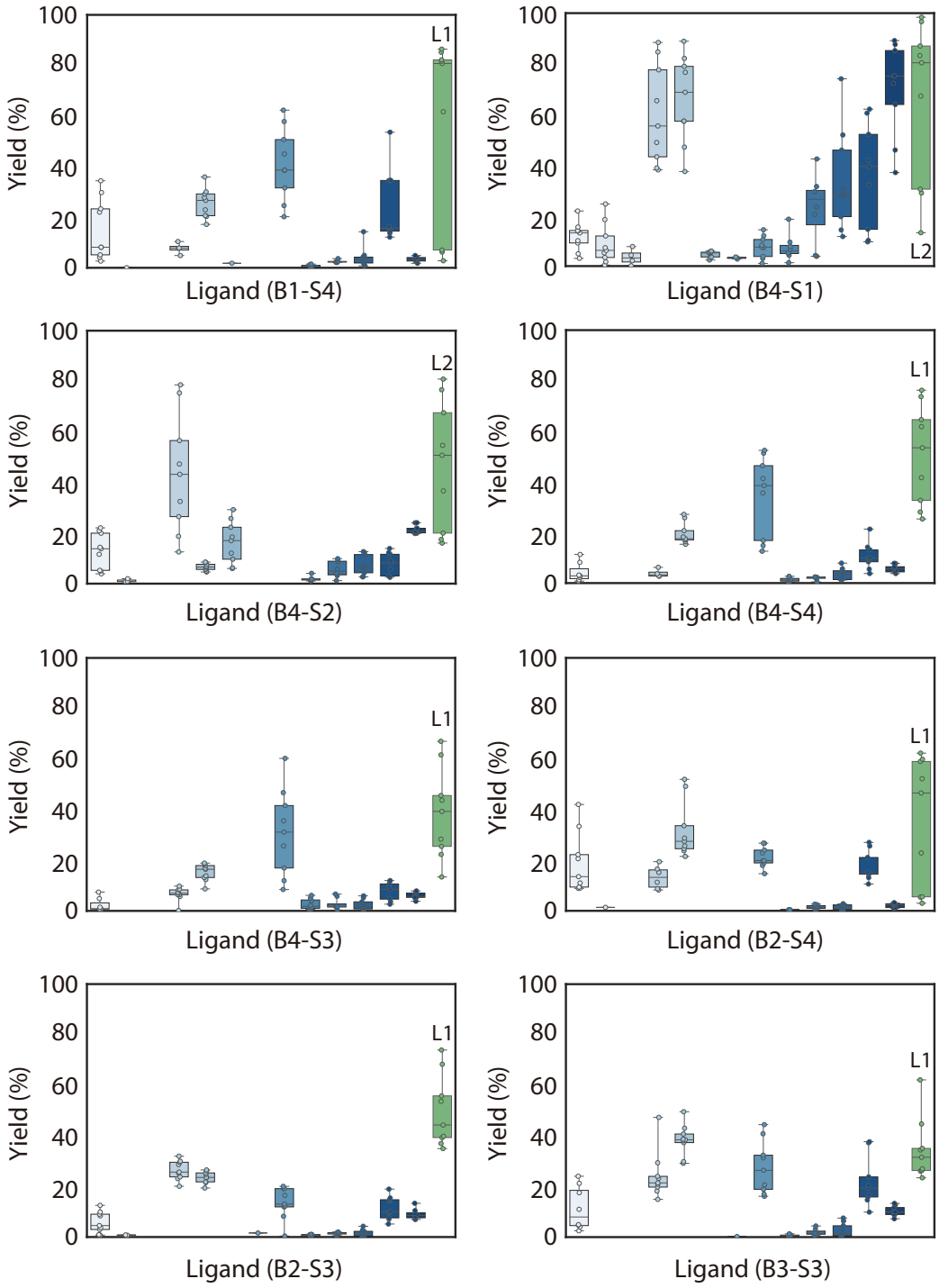
Answer

[Pd].O=C(/C=C/c1ccccc1)/C=C/c1ccccc1.CC1(C)c2cccc(P(c3ccccc3)c3ccccc3)c2Oc2c(P(c3ccccc3)c3ccccc3)cccc21

Supplementary Fig. 1. Instruction template prompts for retrosynthesis and condition generation, respectively. Instruction prompts consist of instruction question contents and corre-

sponding answers. The instruction contents can be split into two components: instruction templates and input reaction data.


Yield (%)
100
80
60
40
20
0
L1
Ligand (B1-S1)
Ligand (B1-S2)
Ligand (B1-S3)
L2
Ligand (B2-S1)
Ligand (B2-S2)
Ligand (B3-S1)
Ligand (B3-S2)
Ligand (B3-S4)


Yield (%)
100
80
60
40
20
0
L1
Ligand (B1-S4)
L2
Ligand (B4-S1)
Ligand (B4-S2)
Ligand (B4-S4)
Ligand (B4-S3)
Ligand (B2-S4)
Ligand (B2-S3)
Ligand (B3-S3)

Supplementary Fig. 2. Performance of ligand recommendation under different combinations of solvents and bases. For Pd-catalysed C–H arylation reaction, given its solvent and reagent, Chemma recommends a suitable ligand. Considering varying solvents and bases on C–H arylation datasets, we depict yield distribution under different combinations of solvent and bases. The molecule structure of B, S, and L are shown in Extended Data Fig. 2.

A *Chat log for C-H arylation reaction*

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Cn1cnc(C#N)c1. Fc1ccccc1Br**, SMILES for products of reactions are **N#CC1=C(C2=CC=CC=C2F)N(C)C=N1**, the reaction can be described as **Cn1cnc(C#N)c1.Fc1ccccc1Br>>N#CC1=C(C2=CC=C-C=C2F)N(C)C=N**, base for this reaction is **[Cs+].CC(=O)[O-]**, solvents for this reaction is **CCCC#N,** Could you offer any recommendations for catalysts that could have been used in this specific chemical reaction?

CC(C)c1cc(C(C)C)c(-c2ccccc2P(C2CCCCC2)C2CCCCC2)c(C(C)C)c1

Yield: 76.63%

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Cn1cnc(C#N)c1.Fc1ccccc1Br**, SMILES for products of reactions are **N#CC1=C(C2=CC=CC=C2F)N(C)C=N1**, the reaction can be described as **Cn1cnc(C#N)c1.Fc1ccccc1Br>>N#CC1=C(C2=C-C=C-C=C2F)N(C)C=N1**, reagents for this reaction is **[Cs+].CC(=O)[O-]**, solvents for this reaction is **CCCC#N,** catalyst for this reaction is **CC(C)c1cc(C(C)C)c(-c2ccccc2P(C2CCCCC2)C2CCCCC2)c(C(C)C)c1**, product yield for this reaction is 76.63%, could you offer any recommendations for another catalysts that could have been used to get high yield in this specific chemical reaction?

C[C@]12C[C@@]3(C)O[C@](C)(C[C@@](C)(O1)P3c1ccccc1)O2

Yield: 91.19%

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Cn1cnc(C#N)c1. Fc1ccccc1Br**, SMILES for products of reactions are **N#CC1=C(C2=CC=CC=C2F)N(C)C=N1**, the reaction can be described as **Cn1cnc(C#N)c1.Fc1ccccc1Br>>N#CC1=C(C2=CC=C-C=C2F)N(C)C=N**, base for this reaction is **[K+].CC(C)(C)C(=O)[O-]**, solvents for this reaction is **Cc1ccc(C)cc1** , Could you offer any recommendations for catalysts that could have been used in this specific chemical reaction?

CC(C)c1cc(C(C)C)c(-c2ccccc2P(C2CCCCC2)C2CCCCC2)c(C(C)C)c1

Yield: 48.34%

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Cn1cnc(C#N)c1. Fc1ccccc1Br**, SMILES for products of reactions are **N#CC1=C(C2=CC=CC=C2F)N(C)C=N1**, the reaction can be described as **Cn1cnc(C#N)c1.Fc1ccccc1Br>>N#CC1=C(C2=CC=C-C=C2F)N(C)C=N**, base for this reaction is **[K+].CC(C)(C)C(=O)[O-]**, solvents for this reaction is **Cc1ccc(C)cc1** , catalyst for this reaction is **CC(C)c1cc(C(C)C)c(-c2ccccc2P(C2CCCCC2)C2CCCCC2)c(C(C)C)c1**, product yield for this reaction is 48.34%, could you offer any recommendations for another catalysts that could have been used to get high yield in this specific chemical reaction?

C[C@]12C[C@@]3(C)O[C@](C)(C[C@@](C)(O1)P3c1ccccc1)O2

Yield: 100%

B

Chat log for Buchward-Hartwig reaction

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Brc1ccccn1.NC1=CC=C(C)C=C1**, SMILES for products of reactions are **Cc1ccc(Nc2ccccn2)cc1**, the reaction can be described as **Brc1ccccn1.NC1=CC=C(C)C=C1>>Cc1ccc(Nc2ccccn2)cc1**, base for this reaction is **CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C**, additive for this reaction is **c1ccc2nocc2c1,** Could you offer any recommendations for catalysts that could have been used in this specific chemical reaction?

COc1ccc(OC)c(P(C23CC4CC(CC(C4)C2)C3)C23CC4CC(CC(C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C

Yield: 4.95%

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Brc1ccccn1.NC1=CC=C(C)C=C1**, SMILES for products of reactions are **Cc1ccc(Nc2ccccn2)cc1**, the reaction can be described as **Brc1ccccn1.NC1=CC=C(C)C=C1>>Cc1ccc(Nc2ccccn2)cc1**, base for this reaction is **CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C**, additive for this reaction is **c1ccc2nocc2c1,** catalyst for this reaction is COc1ccc(OC)c(P(C23CC4CC(C-C(C4)C2)C3)C23CC4CC(CC(C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C, product yield for this reaction is 4.95%, could you offer any recommendations for another catalysts that could have been used to get high yield in this specific chemical reaction?

CC(C)c1cc(C(C)C)c(-c2ccccc2P(C(C)(C)C)C(C)(C)C)c(C(C)C)c1

Yield: 21.66%

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Brc1ccccn1.NC1=CC=C(C)C=C1**, SMILES for products of reactions are **Cc1ccc(Nc2ccccn2)cc1**, the reaction can be described as **Brc1ccccn1.NC1=CC=C(C)C=C1>>Cc1ccc(Nc2ccccn2)cc1**, base for this reaction is **CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C**, additive for this reaction is **c1ccc2oncc2c1,** Could you offer any recommendations for catalysts that could have been used in this specific chemical reaction?

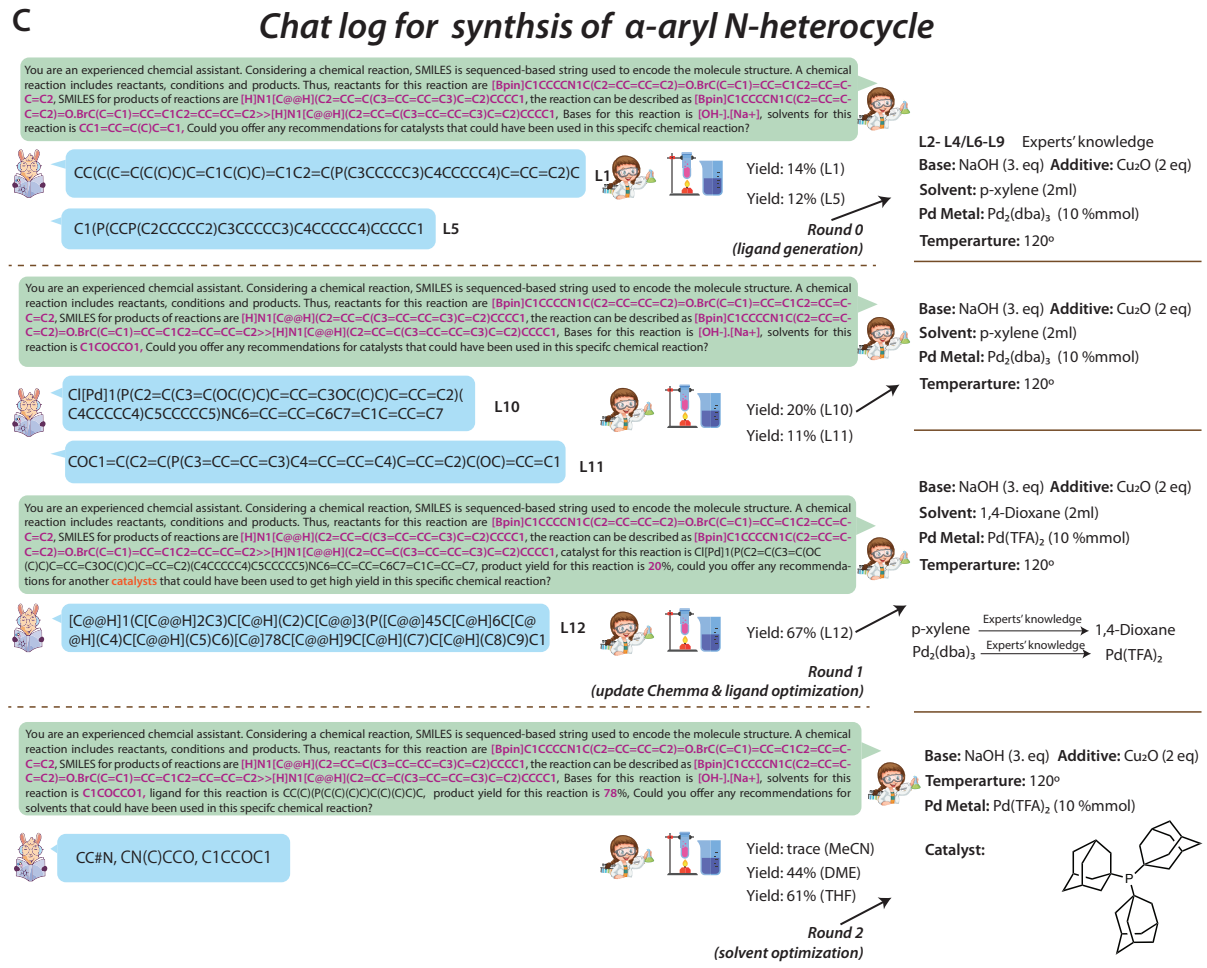
COc1ccc(OC)c(P(C23CC4CC(CC(C4)C2)C3)C23CC4CC(CC(C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C

Yield: 25.99%

You are an experienced chemical assistant. Considering a chemical reaction, SMILES is sequenced-based string used to encode the molecule structure. A chemical reaction includes reactants, conditions and products. Thus, reactants for this reaction are **Brc1ccccn1.NC1=CC=C(C)C=C1**, SMILES for products of reactions are **Cc1ccc(Nc2ccccn2)cc1**, the reaction can be described as **Brc1ccccn1.NC1=CC=C(C)C=C1>>Cc1ccc(Nc2ccccn2)cc1**, base for this reaction is **CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C**, additive for this reaction is **c1ccc2oncc2c1**, catalyst for this reaction is COc1ccc(OC)c(P(C23CC4CC(C-C(C4)C2)C3)C23CC4CC(CC(C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C, product yield for this reaction is 4.95%, could you offer any recommendations for another catalysts that could have been used to get high yield in this specific chemical reaction?

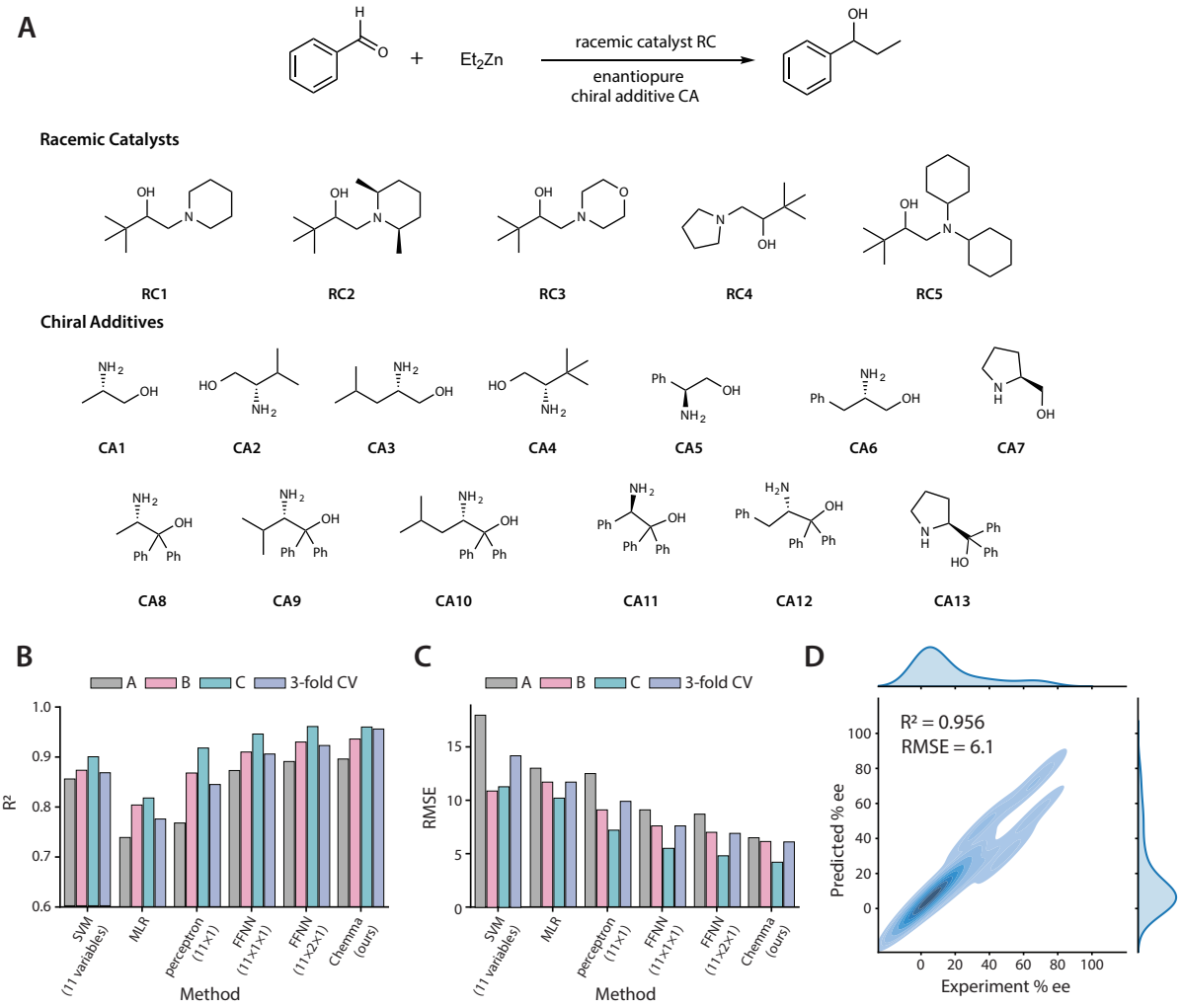
CC(C)c1cc(C(C)C)c(-c2ccccc2P(C(C)(C)C)C(C)(C)C)c(C(C)C)c1

Yield: 65.48%



Supplementary Fig. 3. The chat log of C–H arylation, Buchward-Hartwig reaction, and unreported synthesis of α-aryl N-heterocycles, respectively. (**A-B**) The chat log for optimizing a C–H arylation and Buchward-Hartwig reaction. Within the predefined reaction space, we first construct zero-shot prompts and ask Chemma to recommend a suitable ligand. Other condition variables including solvents and base are randomly selected. For the next, we control reaction condition variables except ligands unchanged and ask to suggest a 'higher-yield' ligand using ICL instruction prompts. If the suggested ligand has been tested in previous runs, we randomly change the reaction variables except for ligands for the next round of zero-shot interac-

tion experiments. (**C**) The chat log for exploring an unexplored Suzuki-Miyaura cross-coupling reaction of cyclic aminoboronates and aryl halides for the synthesis of α-Aryl N-heterocycles.



Supplementary Fig. 4. Performance evaluation of Chemma for enantiomeric excess prediction. (**A**) Reaction of the enantioselective addition of diethyl zinc to benzaldehyde. (**B-C**) Bar plot of prediction performance. Comparison of R^2 and RMSE values across different models for datasets A, B, C, and 3-fold cross-validation, respectively. (**D**) Density plot of prediction performance for enantiomeric excess (% ee).

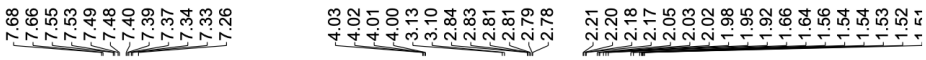
Supplementary Table 1: The performance of LLMs and baseline in retrosynthesis task. The best baseline is underlined.

Method	Top-1 Accuracy (↑)	Invalid SMILES (↓)
Chemformer (*22*)	<u>0.536</u>	<u>0%</u>
GPT-4 (zero-shot)	0.006 ± 0.005	$20.6\% \pm 4.7\%$
GPT-4 (Scaffold, k=20)	0.096 ± 0.013	$10.4\% \pm 3.4\%$
GPT-4 (Scaffold, k=5)	0.114 ± 0.013	$11.0\% \pm 1.2\%$
GPT-4 (Random, k=20)	0.012 ± 0.011	$18.2\% \pm 4.2\%$
GPT-4o* (500 samples)	0.219 ± 0.018	$12.9\% \pm 4.6\%$
GPT-3.5 (Scaffold, k=20)	0.022 ± 0.004	$6.4\% \pm 1.3\%$
LLaMA-2-13B (Scaffold, k=20)	0	$27.2\% \pm 1.5\%$
Chemma	$\mathbf{0.72 \pm 0.012}$	$\mathbf{3.2\% \pm 1.5\%}$

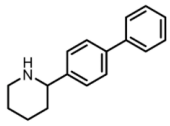
* indicates that we randomly select 500 testing samples for performance evaluation.

Supplementary Table 2: Top-k accuracy for retrosynthesis prediction tasks on USPTO-50k (reaction class unknown). The best performance in each scenario is in bold font.

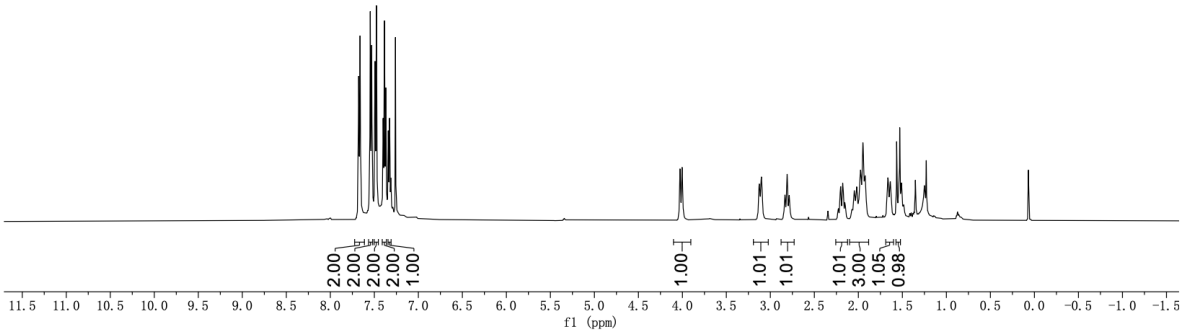
Model	**Top-k accuracy (%)**		
	1	3	5
Template-Based			
GLN (*23*)	52.5	69.0	75.6
LocalRetro (*24*)	53.4	77.5	85.9
RetroKNN (*12*)	**57.2**	**78.9**	**86.4**
Semi-Template-Based			
RetroXpert (*25*)	50.4	61.1	62.3
GraphRetro (*26*)	53.7	68.3	72.2
RetroPrime (*27*)	51.4	70.8	74.0
MEGAN (*28*)	48.1	70.7	78.4
G2Retro (*29*)	**54.1**	**74.1**	**81.2**
Template-Free			
Transformer (*30*)	42.4	58.6	63.8
GTA (*31*)	51.1	67.6	74.8
Graph2SMILES (*32*)	52.0	70.2	75.2
Retroformer (*33*)	53.2	71.1	76.6
NAG2G (*34*)	55.1	76.9	83.4
Chemma	**72.2**	**78.9**	**84.2**

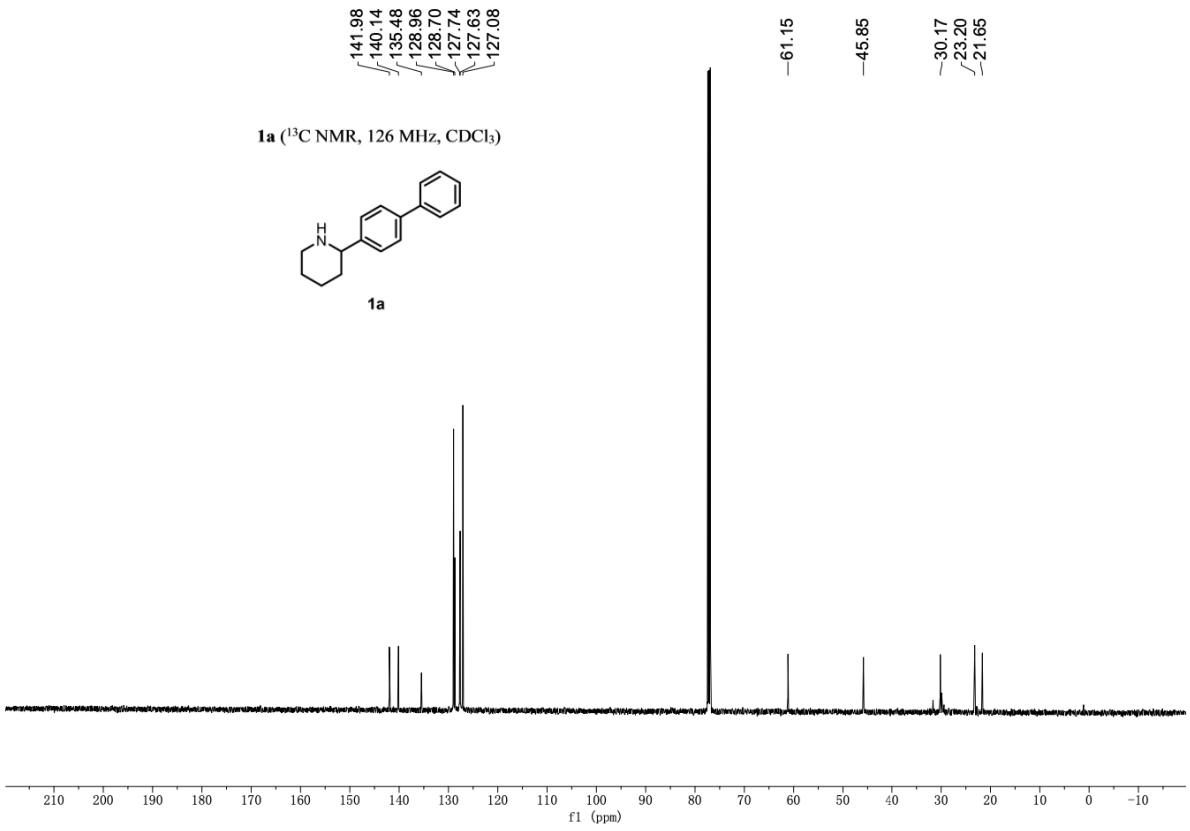


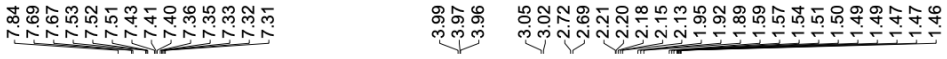
1a (^{1}H NMR, 500 MHz, $CDCl_3$)



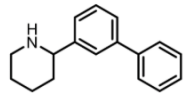
1a



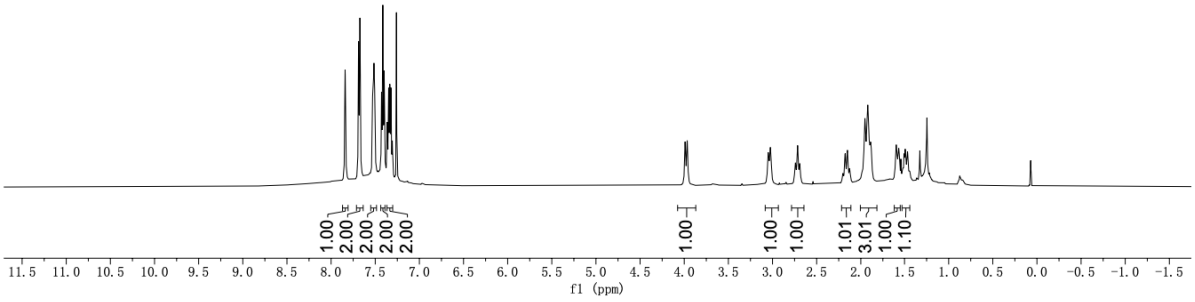

141.98
140.14
135.48
128.96
128.70
127.74
127.63
127.08
61.15
45.85
30.17
23.20
21.65
1a (13C NMR, 126 MHz, CDCl3)
1a
f1 (ppm)

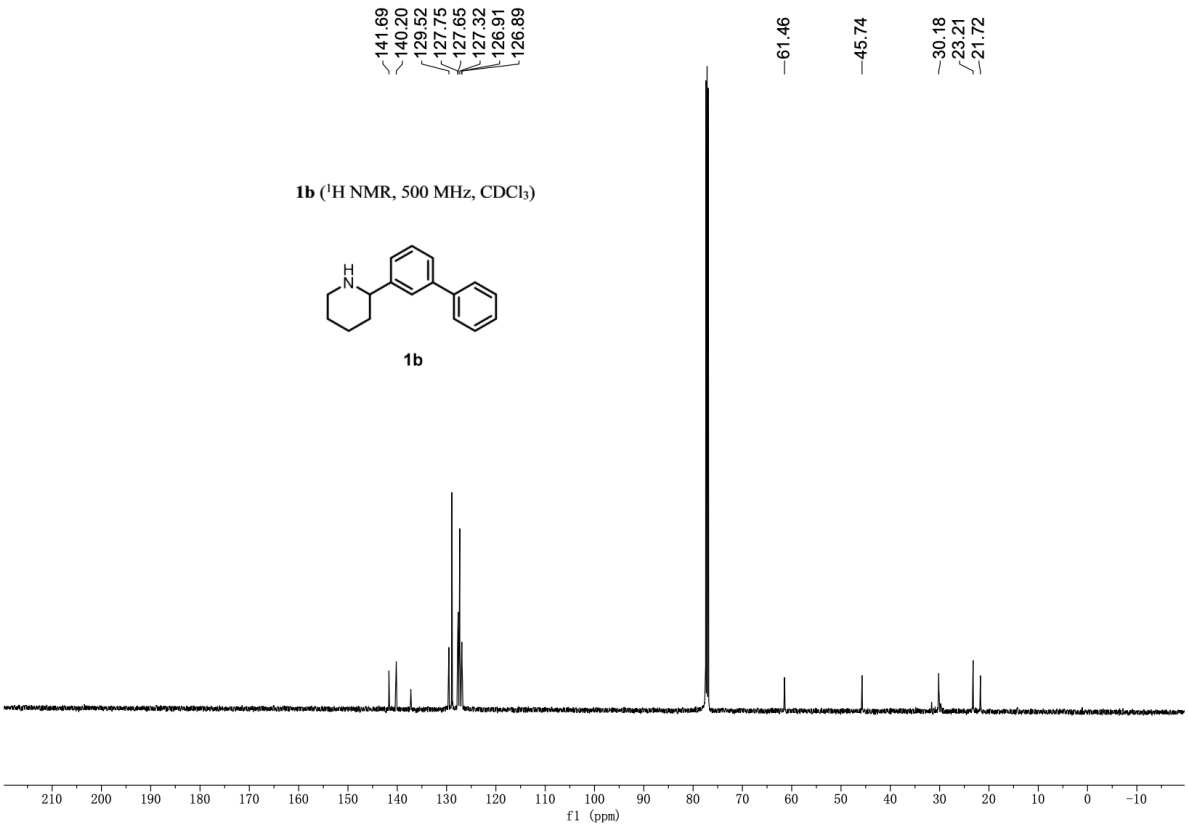


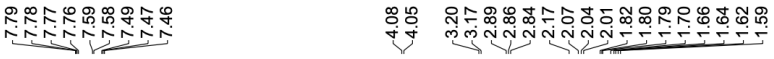
1b (^{1}H NMR, 500 MHz, $CDCl_3$)



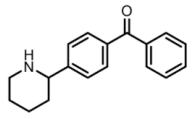
1b



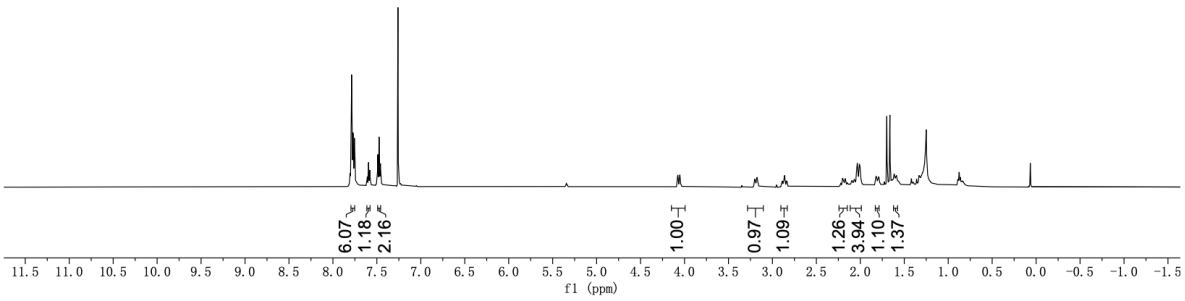

141.69
140.20
129.52
127.75
127.65
127.32
126.91
126.89
61.46
45.74
30.18
23.21
21.72
1b (1H NMR, 500 MHz, CDCl3)
H
N
1b
210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 −10
f1 (ppm)

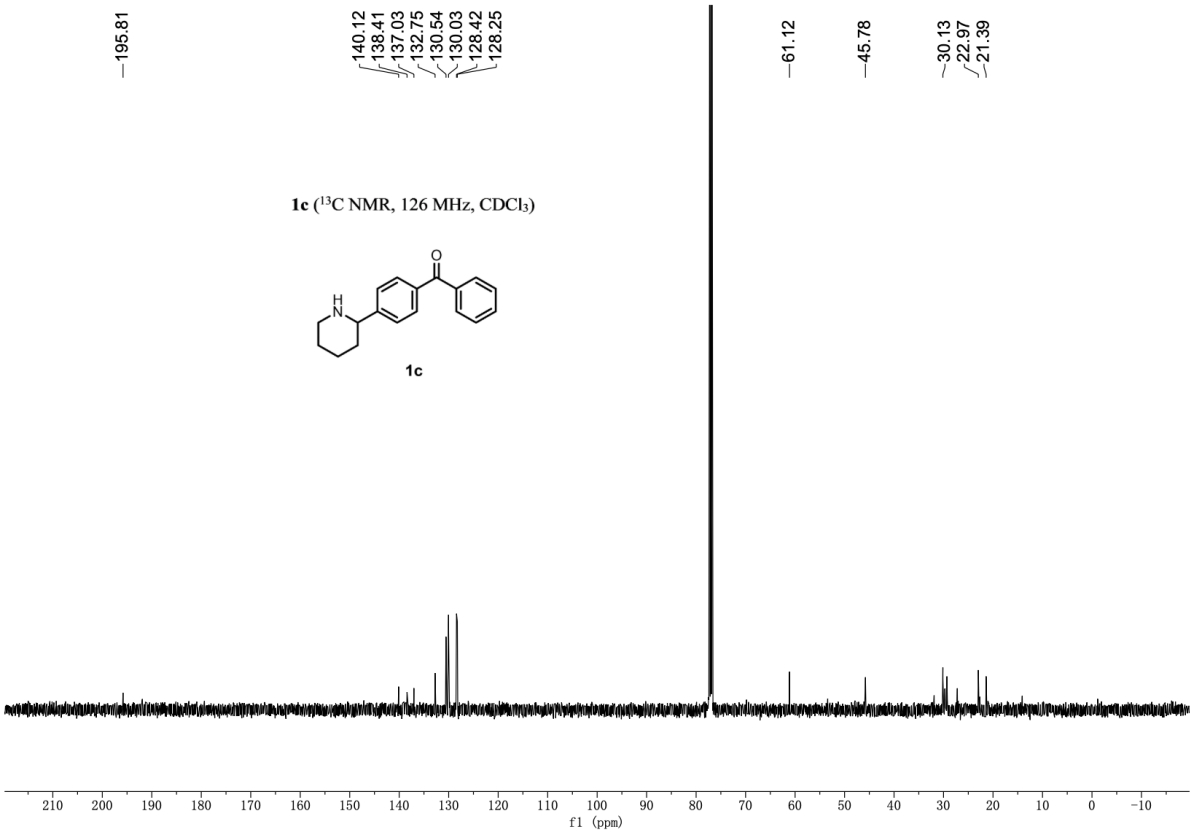


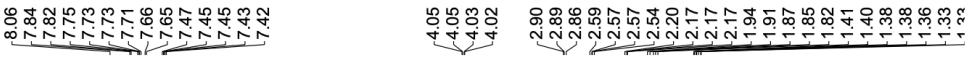
1c (^{1}H NMR, 500 MHz, $CDCl_3$)



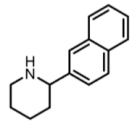
1c



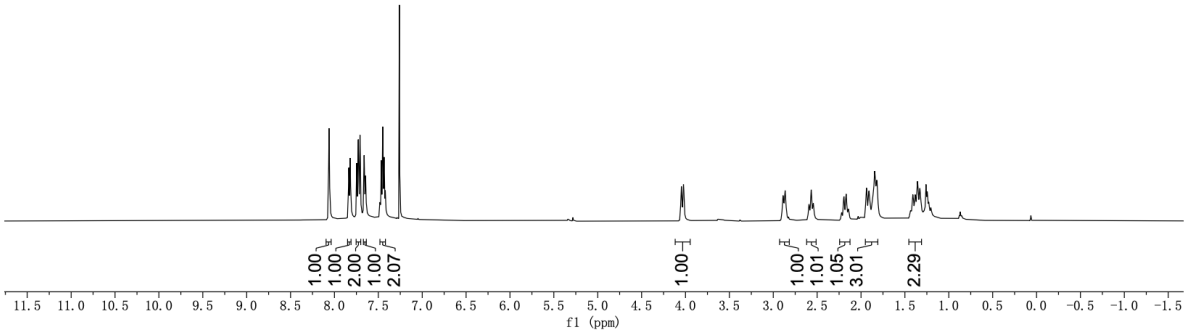

195.81
140.12
138.41
137.03
132.75
130.54
130.03
128.42
128.25
61.12
45.78
30.13
22.97
21.39
1c (^{13}C NMR, 126 MHz, $CDCl_3$)
O
N
H
1c
210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 −10
f1 (ppm)

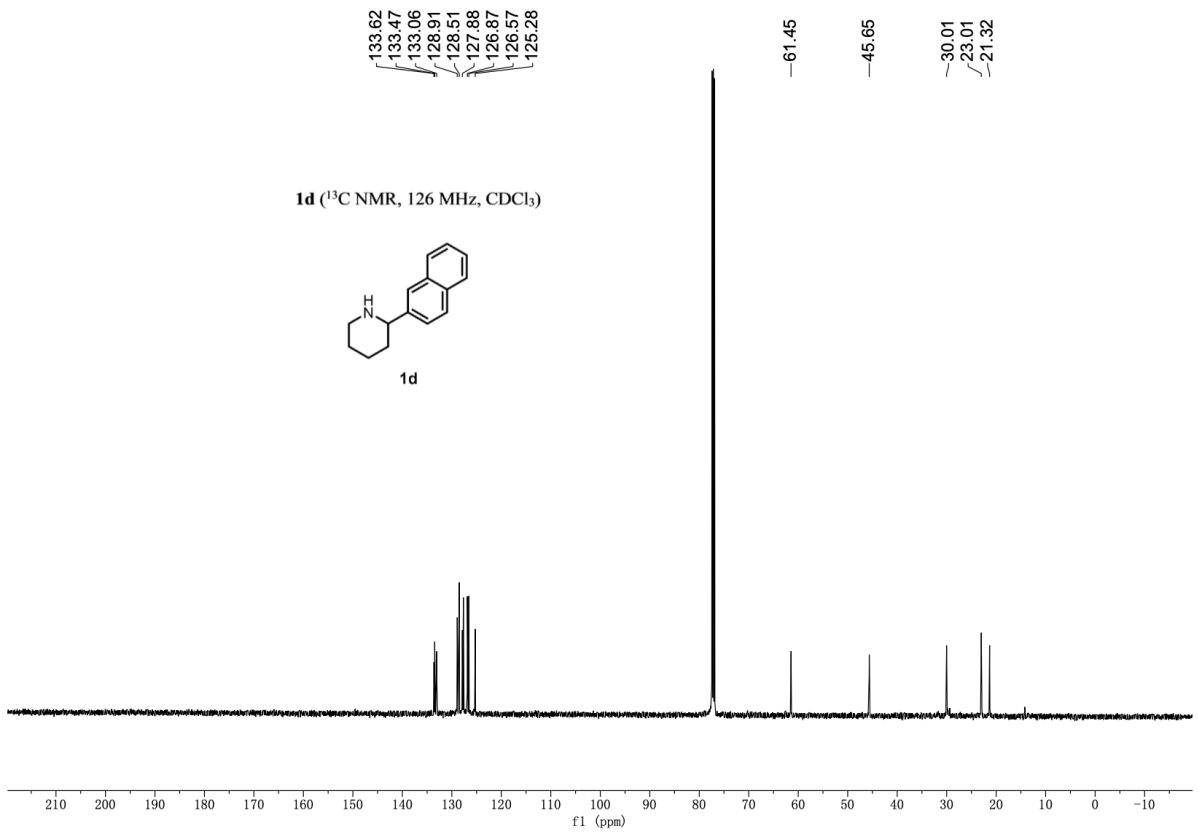


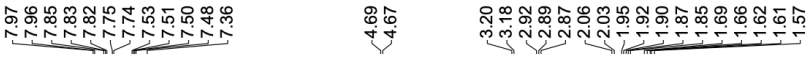
1d (^{1}H NMR, 500 MHz, $CDCl_3$)



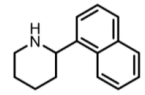
1d



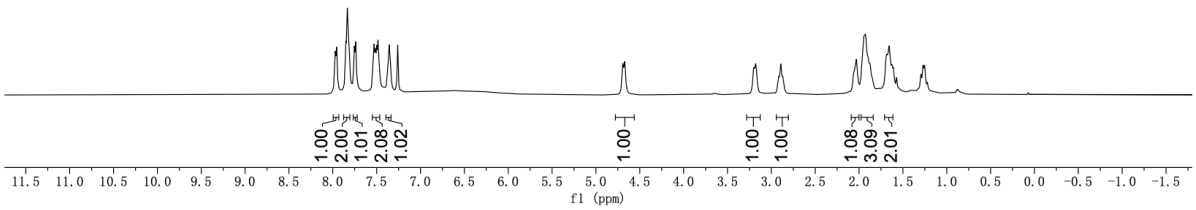

133.62
133.47
133.06
128.91
128.51
127.88
126.87
126.57
125.28
61.45
45.65
30.01
23.01
21.32
1d (^{13}C NMR, 126 MHz, $CDCl_3$)
1d
f1 (ppm)

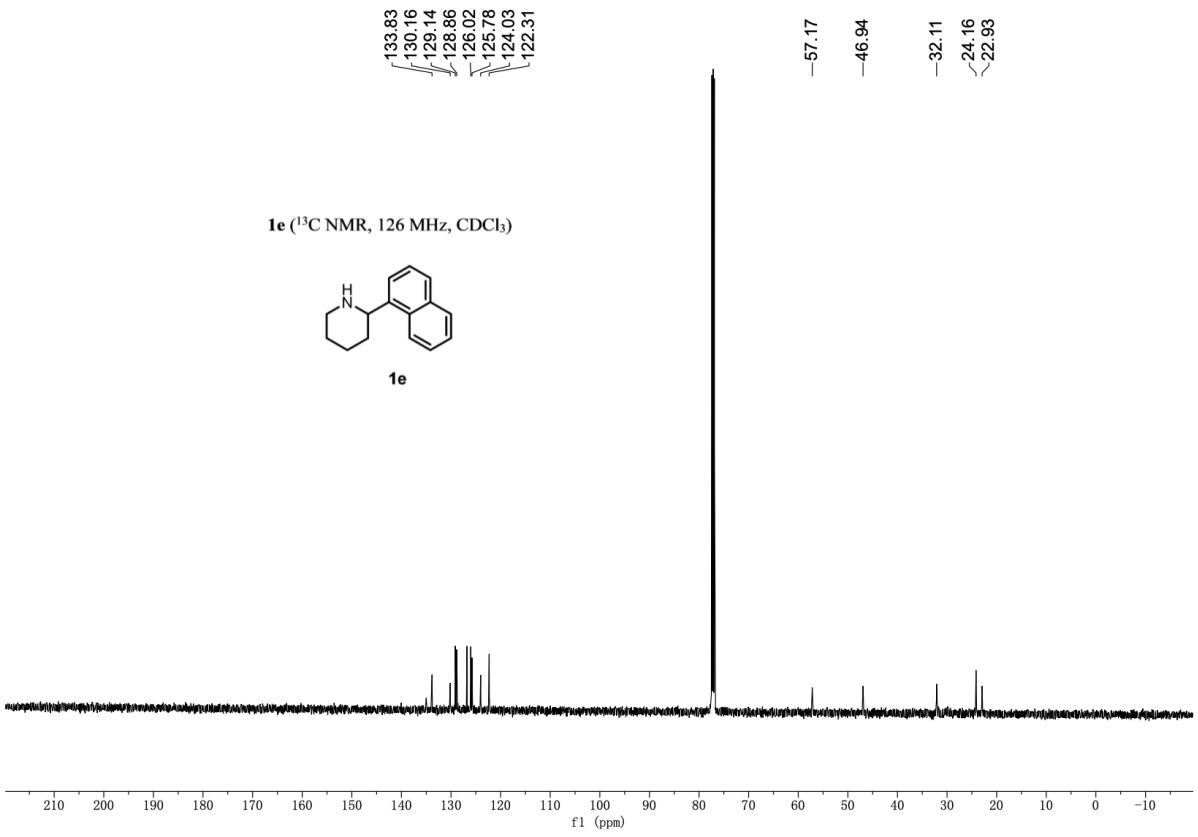


1e (^{1}H NMR, 500 MHz, $CDCl_3$)



1e





References

1. D. M. Lowe, Extraction of chemical structures and reactions from the literature .

2. S. M. Kearnes, *et al.*, The open reaction database, *Journal of the American Chemical Society* **143**, 18820–18826 (2021).

3. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, Predicting reaction performance in c–n cross-coupling using machine learning, *Science* **360**, 186–190 (2018).

4. D. Perera, *et al.*, A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow, *Science* **359**, 429–434 (2018).

5. B. J. Shields, *et al.*, Bayesian reaction optimization as a tool for chemical synthesis, *Nature* **590**, 89–96 (2021).

6. X. Li, S.-Q. Zhang, L.-C. Xu, X. Hong, Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning, *Angewandte Chemie International Edition* **59**, 13253–13259 (2020).

7. A. F. Zahrt, *et al.*, Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning, *Science* **363**, eaau5631 (2019).

8. M. Saebi, *et al.*, On the use of real-world datasets for reaction yield prediction, *Chemical science* **14**, 4997–5005 (2023).

9. R. D. Baxter, *et al.*, Mechanistic insights into two-phase radical C–H arylations, *ACS central science* **1**, 456–462 (2015).

10. L. Ouyang, *et al.*, Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022).

11. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* (2017).

12. S. Xie, *et al.*, Retrosynthesis prediction with local template retrieval, *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, 5330–5338.

13. S. Chen, Y. Jung, Deep retrosynthetic reaction prediction using local reactivity and global attention, *JACS Au* **1**, 1612–1620 (2021).

14. M. R. V. Finlay, *et al.*, Discovery of a potent and selective EGFR inhibitor (AZD9291) of both sensitizing and t790m resistance mutations that spares the wild type form of the receptor (2014).

15. Q.-Y. Yu, H. Zeng, K. Yao, J.-Q. Li, Y. Liu, Novel and practical synthesis of vonoprazan fumarate, *Synthetic Communications* **47**, 1169–1174 (2017).

16. D. Benedetto Tiz, *et al.*, FDA-approved small molecules in 2022: Clinical uses and their synthesis, *Pharmaceutics* **14**, 2538 (2022).

17. M. Barreca, F. Bertoni, P. Barraja, New strategies to hit hematological cancers., *European Journal of Medicinal Chemistry* 116350–116350 (2024).

18. Y.-T. Wang, P.-C. Yang, Y.-F. Zhang, J.-F. Sun, Synthesis and clinical application of new drugs approved by fda in 2023, *European Journal of Medicinal Chemistry* 116124 (2024).

19. D.-Z. Li, X.-Q. Gong, Challenges with literature-derived data in machine learning for yield prediction: A case study on pd-catalyzed carbonylation reactions, *The Journal of Physical Chemistry A* **128**, 10423–10430 (2024).

20. J. Long, K. Ding, Engineering catalysts for enantioselective addition of diethylzinc to aldehydes with racemic amino alcohols: nonlinear effects in asymmetric deactivation of racemic catalysts, *Angewandte Chemie* **113**, 562–565 (2001).

21. J. Aires-de Sousa, J. Gasteiger, Prediction of enantiomeric excess in a combinatorial library of catalytic enantioselective reactions, *Journal of Combinatorial Chemistry* **7**, 298–301 (2005).

22. R. Irwin, S. Dimitriadis, J. He, E. J. Bjerrum, Chemformer: a pre-trained transformer for computational chemistry, *Machine Learning: Science and Technology* **3**, 015022 (2022).

23. H. Dai, C. Li, C. Coley, B. Dai, L. Song, Retrosynthesis prediction with conditional graph logic network, *Advances in Neural Information Processing Systems* **32** (2019).

24. S. Chen, Y. Jung, Deep retrosynthetic reaction prediction using local reactivity and global attention, *JACS Au* **1**, 1612–1620 (2021).

25. C. Yan, *et al.*, Retroxpert: Decompose retrosynthesis prediction like a chemist, *Advances in Neural Information Processing Systems* **33**, 11248–11258 (2020).

26. V. R. Somnath, C. Bunne, C. Coley, A. Krause, R. Barzilay, Learning graph models for retrosynthesis prediction, *Advances in Neural Information Processing Systems* **34**, 9405–9415 (2021).

27. X. Wang, *et al.*, Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions, *Chemical Engineering Journal* **420** (2021).

28. M. Sacha, *et al.*, Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits, *Journal of Chemical Information and Modeling* **61**, 3273–3284 (2021).

29. Z. Chen, O. R. Ayinde, J. R. Fuchs, H. Sun, X. Ning, G2retro as a two-step graph generative models for retrosynthesis prediction, *Communications Chemistry* **6**, 102 (2023).

30. A. Vaswani, *et al.*, Attention is all you need, *Advances in neural information processing systems* **30** (2017).

31. S.-W. Seo, *et al.*, Gta: Graph truncated attention for retrosynthesis, *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), vol. 35, 531–539.

32. Z. Tu, C. W. Coley, Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction, *Journal of Chemical Information and Modeling* **62**, 3503–3513 (2022).

33. Y. Wan, C.-Y. Hsieh, B. Liao, S. Zhang, Retroformer: Pushing the limits of end-to-end retrosynthesis transformer, *International Conference on Machine Learning* (PMLR, 2022), 22475–22490.

34. L. Yao, *et al.*, Node-aligned graph-to-graph: Elevating template-free deep learning approaches in single-step retrosynthesis, *JACS Au* (2024).