communications chemistry

Article

A Nature Portfolio journal

https://doi.org/10.1038/s42004-026-01894-y

An artificial intelligence-driven synthesis planning platform (PhotoCat) for photocatalysis

Jiangcheng Xu[1,2,7], Silong Zhai[1,3,7], Panyi Huang[1,4,7], Wenbo Yu[2], Qingyi Mao[1], Kui Du[5], Weike Su[1], Bin Sun[1] ✉, Can Jin[1,6] ✉ & An Su[6] ✉

While photocatalysis has emerged as a transformative tool in modern synthesis, AI-assisted reaction prediction faces significant challenges due to data limitations. We present PhotoCatDB - a curated, open-source database containing 26.7 K photocatalytic reactions with detailed mechanistic annotations, including 9.2 K multicomponent transformations. Leveraging this resource alongside 100 million molecular data points, we developed PhotoCat, a Transformer-based platform that achieves unprecedented accuracy in photocatalytic reaction prediction (82.6%), retrosynthesis (77.1%), and condition recommendation (88.5%). The platform's capabilities were experimentally validated through the discovery of four novel photocatalytic reactions with yields up to 75.3%. This integrated approach establishes a new paradigm for data-driven innovation in photocatalysis, bridging computational prediction with experimental validation to accelerate discovery in sustainable chemistry.

Photocatalytic reactions, which harness visible light or solar energy as a driving force, offer a green and sustainable approach to chemical synthesis[1,2]. In recent years, with advancements in photocatalyst optimization[3,4] and deeper insights into reaction mechanisms[5], photocatalysis has made significant progress in organic synthesis[2,6], particularly in radical reactions[7,8], redox transformations[9–11], and cross-coupling[12,13] processes. In this context, photocatalysis represents a breakthrough that inspires chemists to explore uncharted territories and discover elusive reaction patterns[14,15]. Despite these advances, the practical execution of photocatalysis in the lab is fraught with challenges, often necessitating years of dedicated research to discover and optimize each novel photocatalytic reaction[14].

Deep learning models have made great strides in recent years, largely due to their remarkable ability to extract knowledge from massive amounts of data[16–20]. In the field of organic synthesis, deep learning has brought about a revolution that has impacted several areas, including forward reaction prediction[21–28], retrosynthesis planning[29–33], mechanistic inference[34,35], inferring experimental procedures[36,37], reaction yield prediction[38], and new reaction development[39–41]. Specifically, deep learning models have proven to be effective in predicting specific types of reactions, including enzyme-catalyzed reactions[28,42], carbohydrate reactions[43], and electrochemical reactions[44]. In addition, the application of deep learning in the field of photocatalysis has garnered significant attention, particularly in the design and optimization of photocatalysts[45,46].

However, to the best of our knowledge, no deep learning models specifically targeting photocatalytic reactions have been reported, not only because of the challenging nature of exploring photocatalytic reactions[14,15] that leads to the scarcity of photocatalytic reaction data, but also due to the limitation of currently available reaction databases. Although acknowledged chemical reaction databases like Reaxys[47], SciFinder[48], can offer access to photocatalytic reactions via keyword searches, they suffer from incomplete reaction condition information. For instance, the water as a reactant of photocatalysis is hidden in the word "wet", which was missed by the Reaxys dataset (Reaction ID: 49168884). Another initiative to address the database format issue is the Open Reaction Database (ORD) framework proposed by Kearnes et al. [49]. ORD displays chemical reaction data in a structured format, which provides strong support for machine learning prediction of chemical reactions. However, with fewer than 300 photocatalytic reactions recorded,

[1]National Engineering Research Center for Process Development of Active Pharmaceutical Ingredients, Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals, Zhejiang University of Technology, Hangzhou, P. R. China. [2]Hangzhou Polytechnic University, Hangzhou, P. R. China. [3]Faculty of Applied Science, Macao Polytechnic University, Macau SAR, P. R. China. [4]Institute of Advanced Studies and School of Pharmaceutical Sciences, Taizhou University, Jiaojiang, P. R. China. [5]School of Chemistry and Chemical Engineering, Shaoxing University, Shaoxing, P. R. China. [6]Zhejiang Key Laboratory of Green Manufacturing Technology for Chemical Drugs, College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou, P. R. China. [7]These authors contributed equally: Jiangcheng Xu, Silong Zhai, Panyi Huang. ✉e-mail: sunbin@zjut.edu.cn; jincan@zjut.edu.cn; ansu@zjut.edu.cn

ORD lacks the volume of data required to effectively train deep learning models in this field.

In this study, a database of photocatalytic reactions, named PhotoCatDB, was established through a comprehensive literature search, and these reactions were scrutinized by human experts. To further enrich the database, we also incorporated experimentally recorded reaction data. Most importantly, we added the essential reaction conditions, such as photocatalysts, bases or acids, additives, wavelength, and solvents, to PhotoCatDB to accurately reflect the nuances of real-world laboratory settings. Building upon this robust database, we developed PhotoCat, an advanced transformer-based deep-learning platform for predicting photocatalytic reactions, conducting retrosynthesis, and recommending reaction conditions (Fig. 1). We used PhotoCat to successfully identify and experimentally validate four previously unreported photocatalytic reactions of practical significance. This study marks significant progress in predicting photocatalytic reactions and provides a powerful tool to accelerate the discovery and validation of novel photocatalytic reactions with diverse applications.

Results and discussion

PhotoCatDB, a photocatalytic reaction database

In our efforts to build a unique and valuable resource, we have gathered an extensive collection of photocatalytic reactions. Multicomponent reactions present greater chemical challenges and, from a machine learning perspective, generate larger and sparser combinatorial spaces that limit model generalization. The workflow for constructing PhotoCatDB is shown in Supplementary Fig. 1. PhotoCatDB currently contains 26.7K validated photocatalytic reactions, including 9.2K multicomponent reactions, of which 6708 are three-component, and 2455 are four-component cases (Fig. 2a). Among these multicomponent reactions, 6523 were further annotated by categorizing their conditions through manual mechanistic analysis of the primary literature. To ensure a diverse representation of photocatalytic reactions and to prevent redundancy, we have carefully managed the inclusion of similar reactions. This is demonstrated in the distribution of Tanimoto similarity[50] scores of the product molecules within the database (Fig. 2b). The data show a predominance of unique reactions, with most Tanimoto similarity scores falling below 0.2, across a range from 0 to 0.5. This skewed distribution underscores our commitment to maintaining a database that promotes the discovery of novel photocatalytic processes by minimizing overlap and maximizing the breadth of reaction scenarios covered.

Each reaction record in PhotoCatDB contains the following key components: (1) Reaction equations represented using simplified molecular input line entry system[51] (SMILES), a specialized notation for expressing chemical structures in a computer-readable format. In this notation, the symbol "»" separates reactants and products, while the symbol "." distinguishes different reactants. (2) Additional information, including reaction time, yield, and literature sources. To further capture the unique features of photocatalytic reactions, we have created a focused subset within the database, PhotoCatDB-Cond (PhotoCatDB-Condition), which includes 6523 photocatalytic reactions. Each entry also contains (3) Reaction conditions, which are categorized into five essential elements based on a manual analysis of the reaction mechanisms: *photocatalyst*, *base or acid*, *additives*, *solvent*, and *wavelength*. The detailed definitions of these categories and the corresponding classification criteria are provided in Supplementary Table 1. Some reactions, such as those involving quinoxalinone, azobenzene, or EDA complexes, do not require an external photocatalyst because the reactants themselves act as photosensitizers. The *photocatalyst* for these reactions is classified as "*autocatalysis*". The reaction conditions cover a broad range, including 59 photocatalysts, 34 bases or acids, 53 additives (of which 37 are ligands), and 42 solvents. An example of a PhotoCatDB entry is shown in Fig. 2c, with additional examples available in Supplementary Tables 2–7.

For a more insightful analysis of the dataset, TMAP[52], a tree-based unsupervised learning algorithm, was utilized to visualize the chemical reaction mapping of the PhotoCatDB (Fig. 3a). In this tree-map embedding, each point represents a reaction, and spatial proximity indicates similarity in reaction features. Clusters show chemically similar reactions, while isolated points denote unique transformations, providing an overview of the dataset's distribution in feature space. Despite the wide variety of photocatalysts, clustering of reactions catalyzed by the same photosensitizer was observed. Notably, different clustering was observed when reactions were colored according to other types of reaction conditions, such as bases or acids, additives and solvents (Supplementary Figs. 2–4), which demonstrates the multivariate effect of reaction conditions and suggests the necessity to include multiple components of reaction conditions in the database. Further discussion of dataset bias, diversity, and limitations is provided in Section 1.5 of the Supplementary Information.

To further demonstrate the advantages of the PhotoCatDB, the photocatalytic reactions collected from SciFinder were also visualized using TMAP (Fig. 3b). PhotoCatDB utilizes common names to represent complex photocatalyst structures, simplifying data reading and storage. In contrast, SciFinder-Photocatalysis uses IUPAC names, which are more challenging for chemical researchers to classify and comprehend. Meanwhile, PhotoCatDB-Cond showcases a greater diversity in reaction types, preventing the aggregation of similar reactions observed in SciFinder-Photocatalysis. PhotoCatDB was curated independently from academic literature, restricted to photocatalytic transformations, and shows no

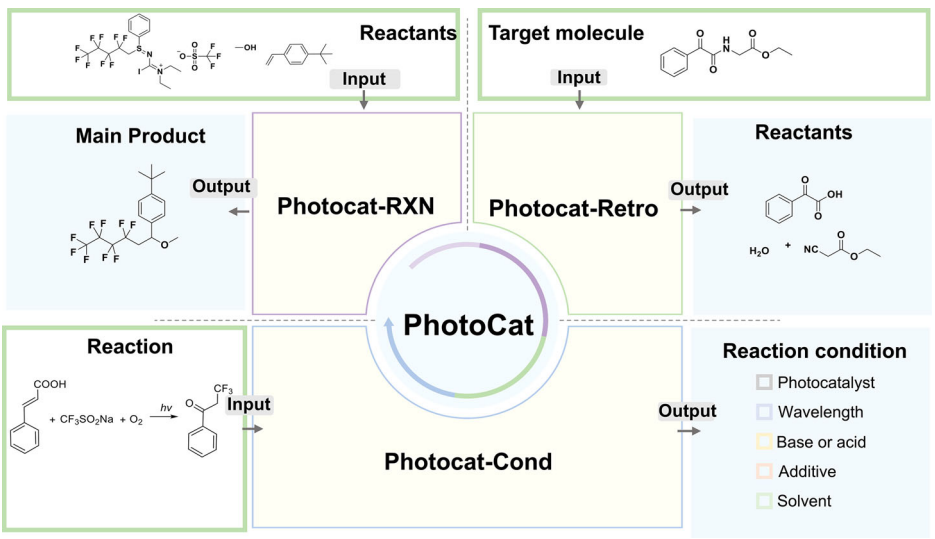
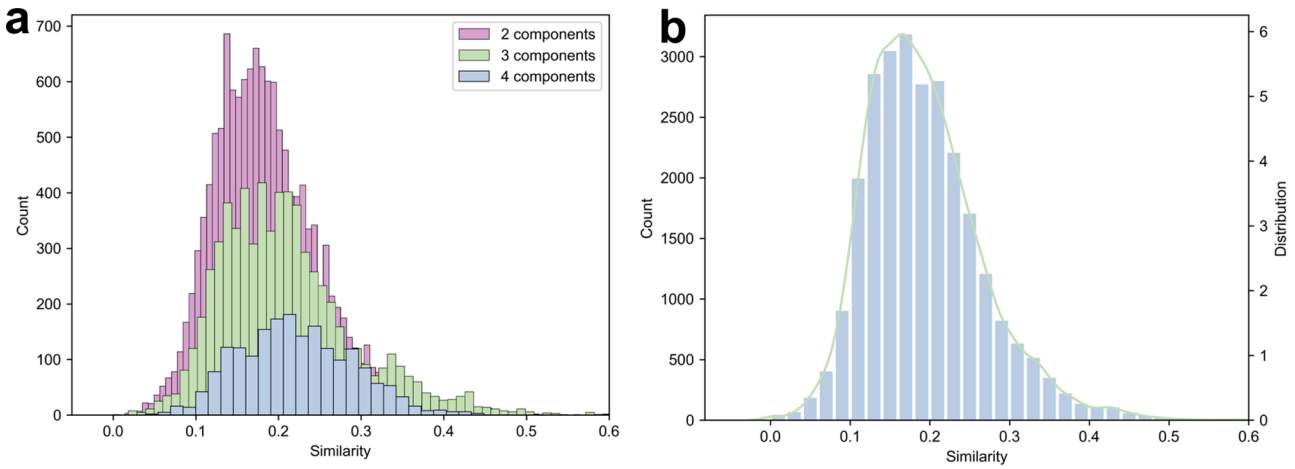


Fig. 1 | PhotoCat comprises three modules: PhotoCat-RXN for photocatalytic reaction prediction, PhotoCat-Retro for retrosynthesis, and PhotoCat-Cond for condition recommendation. Using data-driven (in-silico) approaches, it accelerates the discovery and optimization of photocatalytic reactions in wet-lab settings.



c

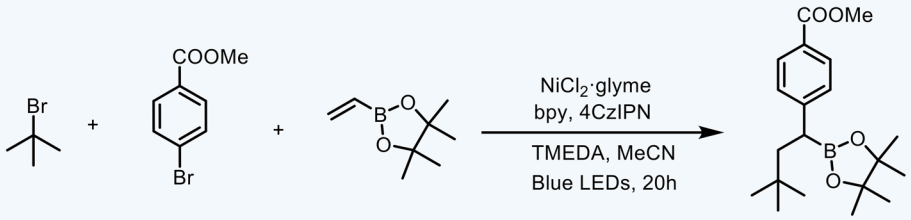
Entry	3251	Category	Three-component photocatalysis
Reaction			
Reaction equation	CC(Br)(C)C.BrC1=CC=C(C(OC)=O)C=C1.C=CB2OC(C)(C)C(C)(C)O2 >>CC(C(C)(C)O3)(C)OB3C(C4=CC=C(C(OC)=O)C=C4)CC(C)(C)C		
Reaction condition	Photocatalyst	Acid or base	Additive	Solvent
	4CzIPN	TMEDA	$Ni(bpy)Cl_2$	MeCN
	Light source	Light categories	Wavelength	Power
	LED	Blue	435nm	-
Additional information	Reaction time (h)	Yield (%)	Literature sources	
	20	73	*Angew. Chem.*, **2020**, *132*, 4400–4404.	

Fig. 2 | Analysis of the data distribution and composition of PhotoCatDB. **a** Multicomponent reactions comprise 34.46% of the dataset, with three-component and four-component reactions marked in green and blue, respectively. **b** The dataset displays a skewed distribution, with most data points falling within the lower similarity range (below 0.21). **c** An example of PhotoCatDB data entry, illustrating the format and structure of the dataset.

overlap with the USPTO reaction set (Section 1.6 of Supplementary Information). One limitation of the current study is that PhotoCatDB does not yet include fine-grained reaction class annotations; future extensions of the database will address this to enable more systematic performance analysis.

PhotoCat-RXN, a reaction-prediction model trained on USPTO and PhotoCatDB

PhotoCat-RXN uses a Transformer-based model from the OpenNMT framework[41,53] as the basis of its model architecture (Fig. 4). First, the model was pre-trained using the USPTO database containing 1 million instances of chemical reactions to gain basic knowledge of chemical reactions. After that, the model was fine-tuned using PhotoCatDB. To demonstrate the need for transfer learning, we trained the Baseline-1 model exclusively on USPTO, while the Baseline-2 model was trained solely on PhotoCatDB. It is important to note that to fairly compare the role of the reactions in USPTO and PhotoCatDB in model training, the PhotoCatDB data for training in this section does not contain any reaction conditions.

As shown in Fig. 5a and Supplementary Tables 7–10, the average Top-1 accuracy of the pre-trained and fine-tuned PhotoCat was 70.68%, and the Top-2 to Top-5 accuracies were all above 78%. In contrast, the prediction accuracy of the Baseline-1 model trained only on USPTO was significantly lower, with an average Top-1 accuracy of only 0.46% and similar low Top-2 to Top-5 accuracies. We also used a different USPTO-pretrained reaction prediction model by Zhong et al.[54] to directly predict the reactions of PhotoCatDB, and the Top-1 accuracy was only 0.69% (Supplementary Table 11). In addition, the Baseline-2 model, trained solely on PhotoCatDB, also exhibited lower prediction accuracies than PhotoCat. Figure 5b offers an alternative perspective on this comparison, demonstrating that the proportion of invalid SMILES predicted by PhotoCat-RXN (0.87%) is lower than that of the Baseline-1 and Baseline-2 models. The results above suggest that the USPTO provides the necessary information for the model to understand chemical reactions, while PhotoCatDB provides the model with the ability to predict photocatalytic reactions.

Reaction conditions improve accuracy and efficiency of reaction prediction

After further fine-tuning PhotoCat-RXN using the PhotoCatDB-Cond subset, we found that incorporating reaction conditions improved the top-1 average accuracy of PhotoCat-RXN from 78.2% to 82.3%, an increase of 4.1% (Fig. 6, Supplementary Tables 12-16). We further investigated the effect of varying the number of reaction conditions on prediction accuracy.

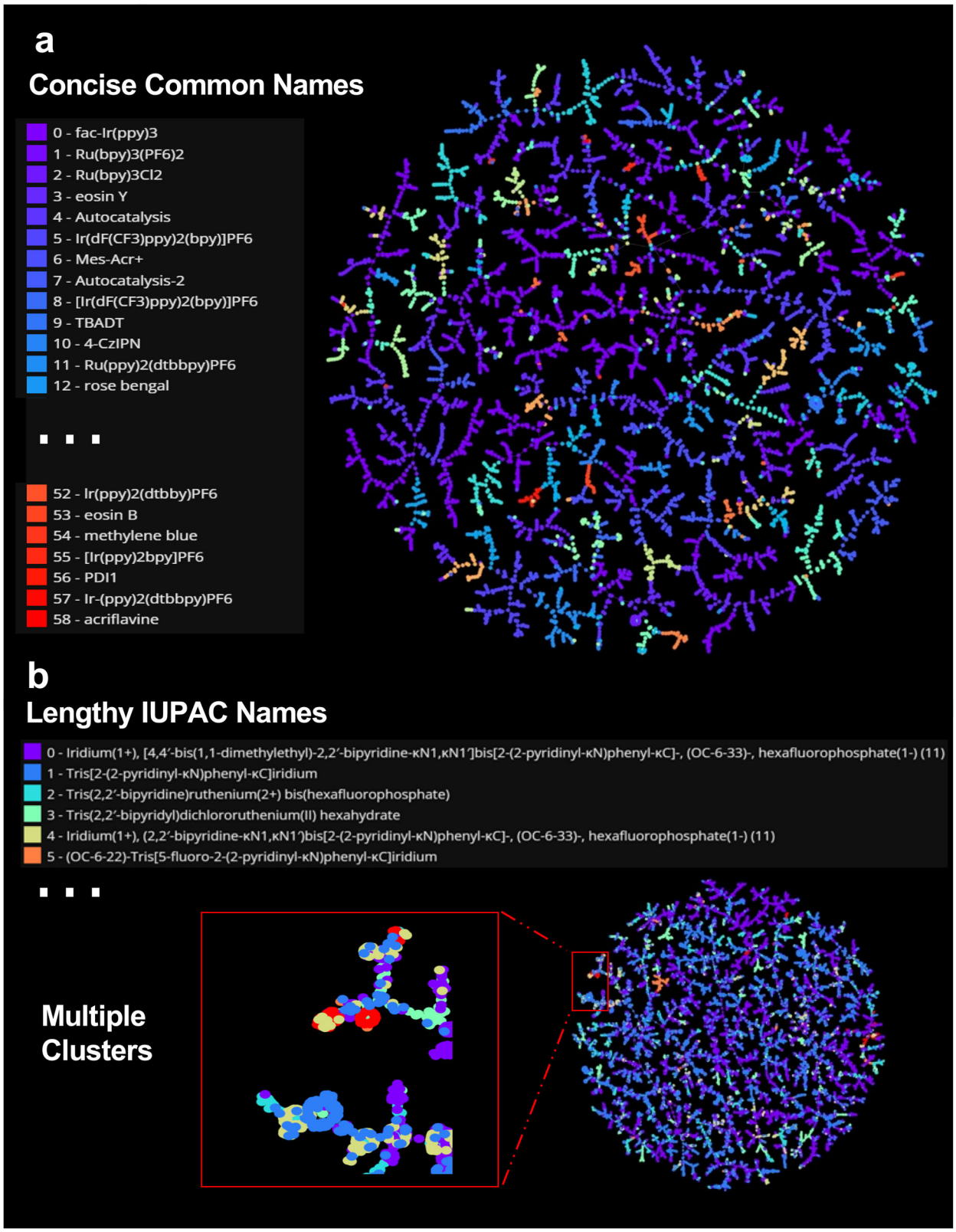


Fig. 3 | Comparison of PhotoCatDB and SciFinder-Photocatalysis (the photocatalysis dataset from SciFinder). a PhotoCatDB uses common names for photocatalysts, while **b** SciFinder-Photocatalysis employs lengthy IUPAC names. In this tree-map embedding, each point denotes a reaction, with spatial proximity indicating feature similarity. Clusters correspond to chemically related reactions, while isolated points highlight unique transformations, providing an overview of the dataset's feature space. **a** Does not show aggregation of similar reactions as "multiple clusters" as in (**b**).

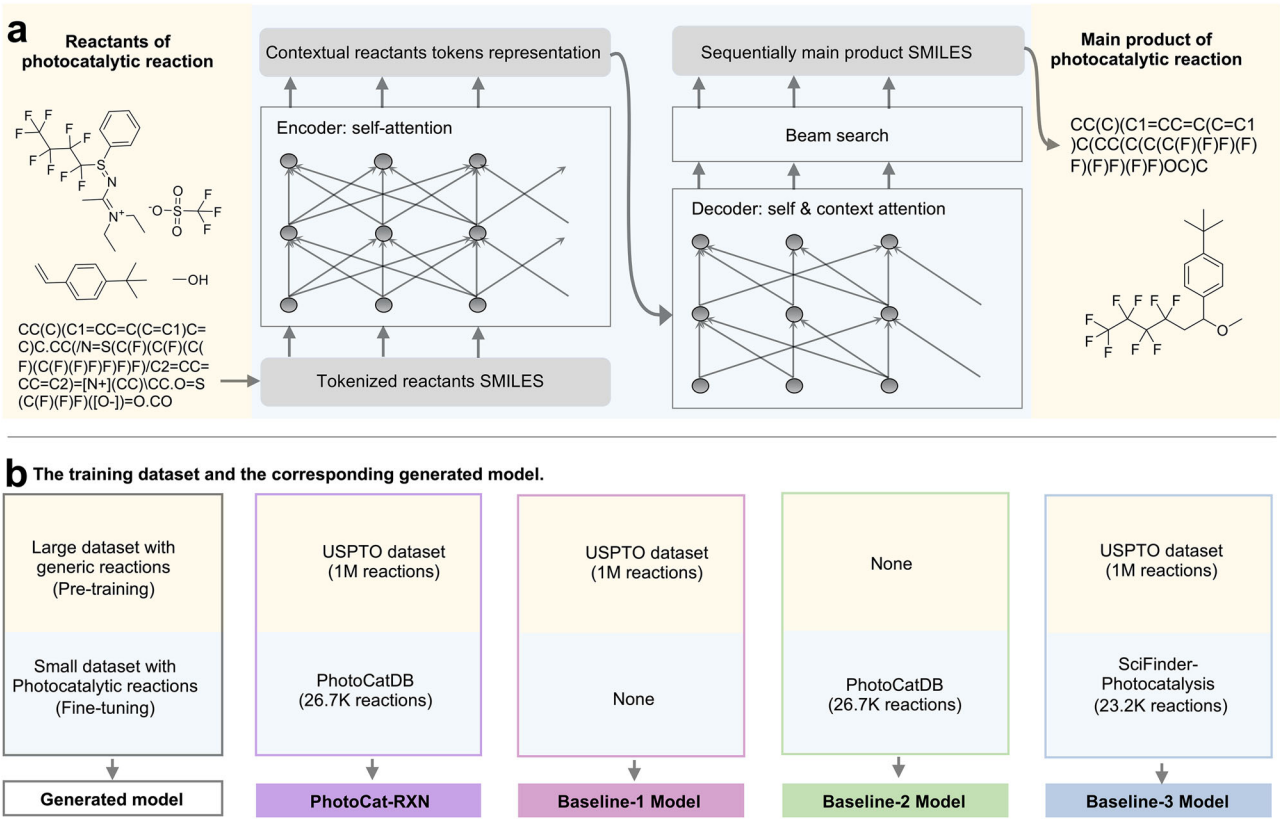


Fig. 4 | The transformer model and training schematic. a Sequence-to-sequence prediction of photocatalytic reactions and **b** the training schematic of PhotoCat and three baseline models.

The analysis began with only the *photocatalyst* and then sequentially incorporated the *base or acid*, *additives*, *wavelength*, and *solvent*. As the number of input reaction conditions increased, the accuracy of reaction predictions continued to rise (Fig. 6a). PhotoCat achieved predicted Top-1 accuracies of 78.1%, 79.8%, 81.9%, and 82.1% when training included one, two, three, and four reaction conditions, respectively. All these accuracies surpassed those of the fine-tuned model that did not utilize information about the reaction conditions. We also extended our evaluation to a graph-based architecture (Graph2SMILES[55]), which similarly benefited from incorporating reaction conditions, confirming that the improvement is not specific to the Transformer model (Supplementary Fig. 10).

Integrating reaction conditions also speeds up the training process. Figure 6b shows that the training curve grows faster for the models with more reaction conditions provided. With no reaction conditions (blue), the accuracy reaches about 76% after 40,000 training steps. In contrast, when all five reaction conditions are considered (cyan), the training accuracy grows faster, reaching over 80% after only 30,000 steps. These observations demonstrated the importance of incorporating detailed and comprehensive reaction conditions during the training and prediction phases, which promotes superior model performance and faster convergence to peak prediction accuracy.

Unlike previous studies that considered only a few commonly used reagents[21–24,56,57] or ignored reaction conditions altogether[26,58], this study shows that including reaction conditions can improve the predictive accuracy and training efficiency of chemical reaction prediction models. There are two main reasons for this improvement. First, in previous studies that considered reaction conditions, the reactants were input simultaneously as mixtures, which may increase the challenge of the model in interpreting the reaction conditions. In this study, the photocatalytic reaction conditions were methodically grouped into five different types, significantly reducing the complexity of the model in grasping the intricate chemical reaction conditions. In addition, the terminology of the reaction conditions was simplified by using concise common names or identifiers, which establishes a direct correspondence between the structure of the reagents and their names (Fig. 6c) and simplifies the model training.

The inclusion of reaction conditions improves prediction accuracy by providing critical contextual information that can fundamentally alter photocatalytic outcomes[59,60] (see Section 5.3 of the ESI). Case analysis (Fig. 7) demonstrates that incorporating classified and simplified reaction conditions into model training enhances reaction prediction accuracy while ensuring interpretability. When the reactants (styrene **1**, DMSO **2**, NHP ester **3**) are subjected to a trivalent ruthenium photocatalyst and light, ketone **4** is the main product in the absence of acid[61]. However, when a strong acid (e.g., trifluoromethanesulfonic acid) is present, the reaction shifts to alkene **5** as the main product (Fig. 7a). A high concentration of protons prevents the deprotonation of the alkoxysulfonium intermediate **6**, thus hindering the formation of ketone **4**[61] (Fig. 7b). Based on complete reaction conditions, PhotoCat precisely predicts the main product (Fig. 7c). In this case, "B0" indicates the absence of "base or acid", while "B31" represents the inclusion of trifluoromethanesulfonic acid in the reaction conditions. In the attention heatmap, the thickness of the lines represents the attention weight between the input and output. Figure 7d and e shows that PhotoCat clearly focuses on the input reaction conditions (highlighted in purple) when predicting the main product. Notably, PhotoCat gives greater attention to "B0" when outputting the key ketone carbonyl group "C = O" (Fig. 7d), and "B31" when projecting the critical alkene group "C = C" (Fig. 7e). This is consistent with human experts' approach in determining the main products of photocatalytic reactions based on reaction conditions.

PhotoCat-Retro, a photocatalytic retrosynthesis model

The photocatalytic retrosynthesis module (PhotoCat-Retro) complements the reaction prediction module (PhotoCat-RXN) by generating feasible photocatalytic pathways directly from target molecules, thereby expanding

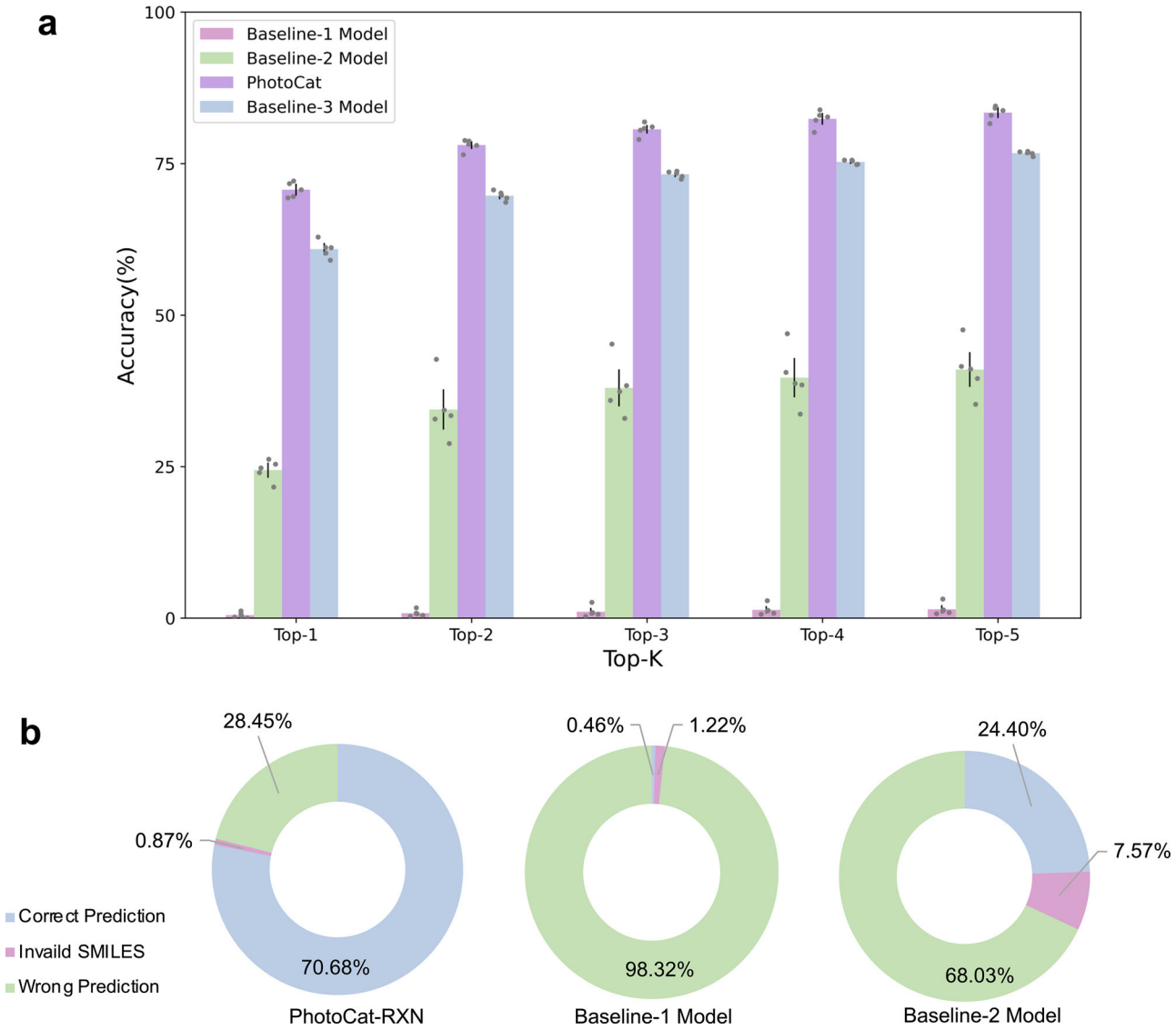


Fig. 5 | Comparison of prediction accuracies of PhotoCat and baseline models. a The comparison of Top-1 to Top-5 accuracies. **b** Percentage of invalid SMILES and incorrect predictions in Top-1 accuracies for Baseline-1 and -2 models and PhotoCat.

the space of reactions that can be explored. This capability is especially valuable for compounds without reported forward reaction data. As shown in Fig. 8a–c, traditional syntheses of aromatic ketones usually rely on Pd-catalyzed two-electron acylation using prefunctionalized aryl halides or boron reagents[62–64]. For the same target molecule, PhotoCat-Retro automatically proposed a non-obvious radical acylation disconnection in which nitrobenzene and pyruvic acid directly couple to introduce the acyl group (Fig. 8d). This reflects the model's ability to capture key mechanistic patterns in photocatalytic radical processes. After refinement by PhotoCat-Cond and confirmation by PhotoCat-RXN, the proposed pathway was successfully validated experimentally (Section 3.4 of the ESI). This example illustrates the practical value of PhotoCat-Retro in generating experimentally viable and innovative photocatalytic disconnection strategies.

PhotoCat-Retro utilizes a state-of-the-art approach, adopting a training strategy inspired by the work of Irwin et al.[65]. To evaluate optimal training strategies, we experimented with three conditions: no pretraining, pretraining at the atom level, and pretraining at the atom and reaction level. The results from 5-fold cross-validation (Fig. 9a, Supplementary Tables 17–19) show that when trained solely on PhotoCaDB for retrosynthesis analysis, the Top-1 average accuracy is 16.60%. However, after incorporating transfer learning from the ZINC database[66] and the combined ZINC + USPTO datasets, the top-1 accuracy improved to 63.28%, 69.43%, a significant increase of 52.83%. This demonstrates that the transfer learning strategy with the ZINC molecular database and the USPTO reaction database during PhotoCat-Retro training helps address the data limitations of PhotoCatDB.

In comparison, current state-of-the-art retrosynthesis models, such as NAG2G[67], EditRetro[68], and LocalRetro[69], achieve notable performance on general retrosynthesis tasks using benchmark databases like USPTO-50k, with average Top-1 accuracies ranging from 57.4% to 67.2%. However, these results remain inferior to the Top-1 accuracy attained by PhotoCat-Retro specifically for photocatalytic retrosynthesis analysis, highlighting its exceptional capability in handling complex photocatalytic reactions. On the other hand, Chemformer demonstrates Top-1 to Top-10 accuracies ranging from 54.3% to 63.0% on the benchmark USPTO-50k dataset (with known chemical reaction classification). After being fine-tuned on the PhotoCatDB, Chemformer exhibits significantly improved performance in photocatalytic organic synthesis retrosynthesis tasks, underscoring the importance and effectiveness of domain-specific training.

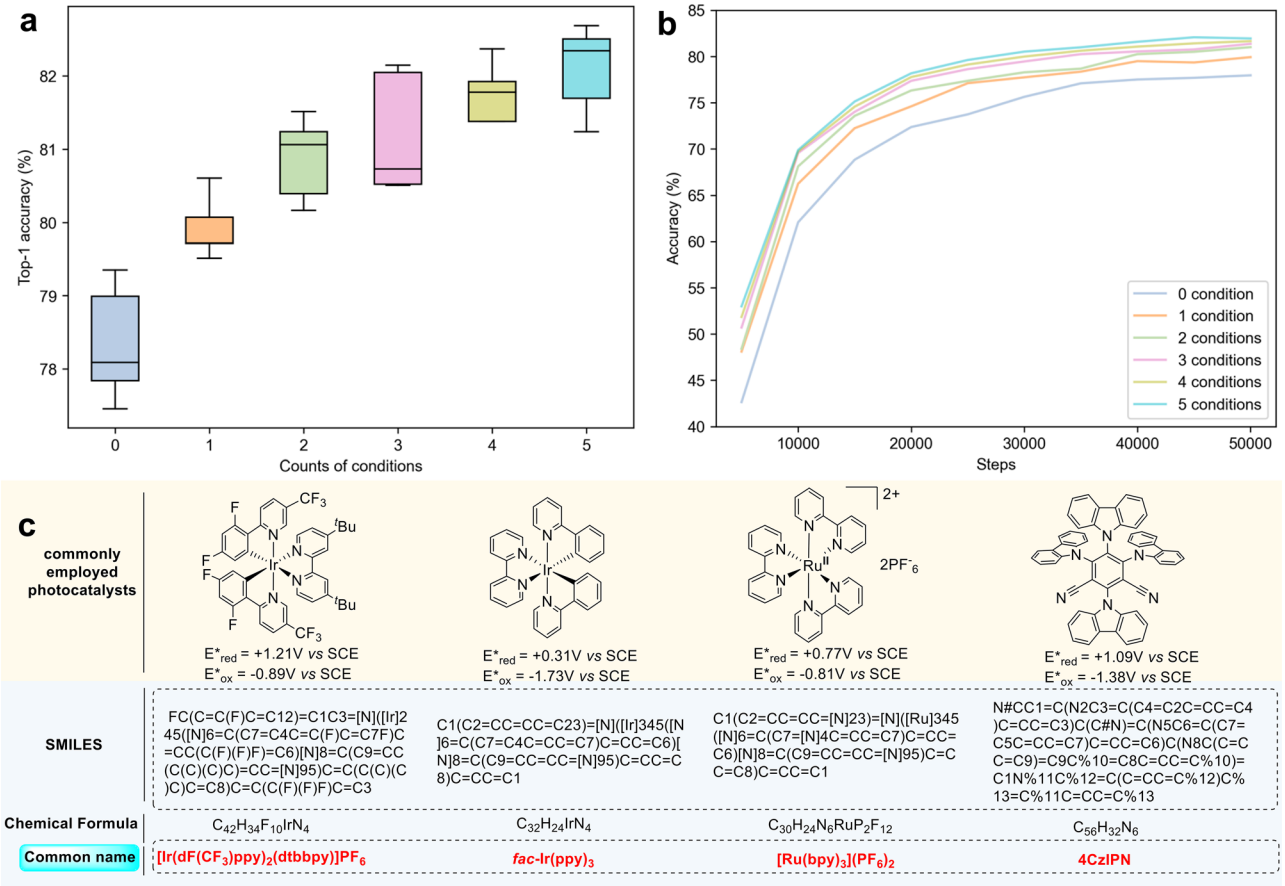


Fig. 6 | Effects of reaction condition inputs on model predictions and training efficiency. a As the quantity of input reaction conditions grows, there's a consistent ascent in prediction accuracy. **b** The adaptive inclusion of various reaction condition inputs significantly amplifies the model's training agility. **c** Simplification strategy for describing reaction conditions in PhotoCatDB-Cond.

PhotoCat-Cond, a deep learning model for recommending photocatalytic reaction conditions

Recommending reaction conditions is essential for improving reaction efficiency and reproducibility. Building on the work of Yao's research group[70] and others[71,72], this study utilizes curated reaction condition data from the PhotoCatDB to develop PhotoCat-Condition (PhotoCat-Cond), a Transformer-based model (Fig. 9b) specifically designed to recommend optimal conditions for photocatalytic reactions. A key feature of PhotoCat-Cond lies in its ability to categorize diverse photocatalytic systems and make condition-specific predictions within each category.

Results from 5-fold cross-validation (Fig. 9c, Supplementary Fig. 8, Supplementary Tables 21–25) demonstrate the robust performance of PhotoCat-Cond. The model accurately recommends the crucial light categories (wavelength), achieving an average Top-1 prediction accuracy of 94.82%. For the particularly challenging task of predicting photosensitizers with complex conjugated structures, the model, based on simplified name representations, achieved an average top-1 accuracy of 88.50%. In terms of recommending solvents, additives, and acids or bases, the model reached top-1 accuracies of 89.52%, 50.06%, and 46.12%, respectively. Notably, in 28.26% of the test cases, PhotoCat-Cond correctly predicted all five conditions of the input reaction simultaneously through an iterative process.

PhotoCat-Cond achieves accuracy comparable to state-of-the-art condition-recommendation models reported in other domains, though direct comparison is not possible due to differences in training datasets and reaction types (Supplementary Table 26). Representative models include Parrot-LM-E[70] (Top-1 accuracies of 92.5% for catalysts and 50.2% for solvents on the USPTO-cond dataset), AR-GCN[71] (64.9% for catalysts, 90.8% for ligands, and 72.2% for solvents on the Suzuki coupling dataset), and CIMG-Condition[72] (59.0% for catalysts and 93.0% for solvents on a general reaction dataset). These results indicate that PhotoCat-Cond performs competitively within its specialized photocatalytic domain, highlighting the advantages of domain-specific datasets, such as PhotoCatDB, and supporting the effectiveness of our targeted strategy for predicting complex photocatalytic conditions.

In this study, reagents and conditions were represented using standardized common names rather than SMILES. While SMILES provides an unambiguous molecular representation, we selected common names because (i) they are the dominant form used in experimental sections of synthetic reports, (ii) they yield a compact and interpretable token vocabulary that aligns with how chemists describe reaction conditions, and (iii) they facilitate direct comparison with prior condition-prediction studies. To mitigate inconsistencies in the literature, all reagent names were normalized via a curated synonym dictionary prior to tokenization. Although this representation may be less precise than SMILES and could limit extrapolation to rare or previously unseen reagents, our benchmarking indicates that generalization was not significantly impaired. Future work may benefit from integrating SMILES or hybrid encodings that combine structural fingerprints with common names to further improve robustness.

Photocatalytic synthesis planning with PhotoCat

PhotoCat provides valuable guidance for conducting wet-lab experiments through dry-lab (in silico) simulations. In this section, we demonstrate how PhotoCat assists in the synthetic planning of photocatalytic reactions (Fig. 10a). The first step is to use PhotoCat-Retro to perform retrosynthetic analysis of the target compound based on photocatalytic reaction rules, generating the corresponding reaction starting reactants. The second step

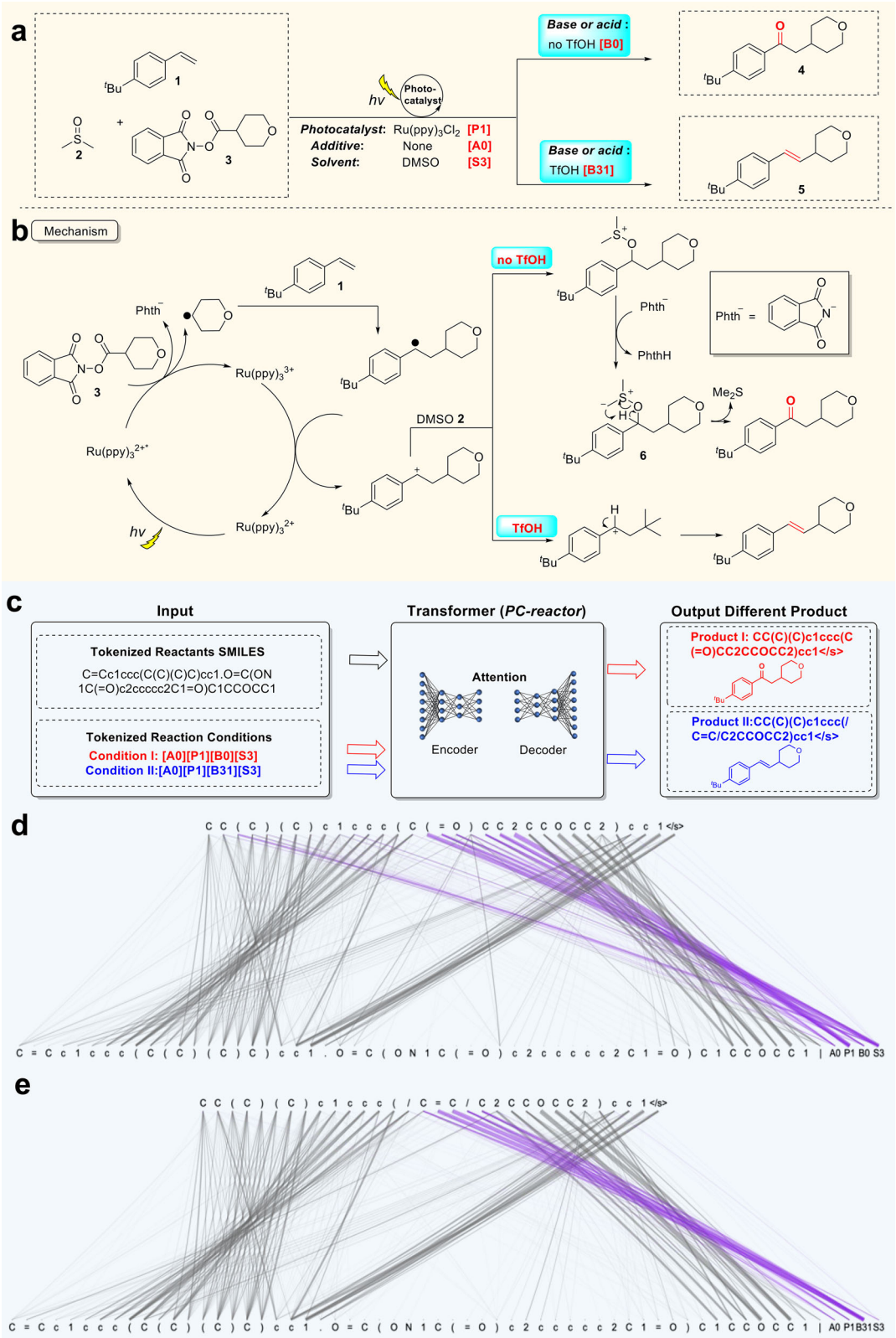


Fig. 7 | Interpretability and attentional analysis of PhotoCat-RXN. a The presence or absence of trifluoromethanesulfonic acid in the reaction system dictates the main product outcome[61], producing aldehyde **4** or alkene **5**. A plausible reaction mechanism is presented in (**b**). **c** PhotoCat accurately predicts main products when provided with corresponding reaction conditions (B0 and B31, respectively refer to "Base or acid" being "None" and "TfOH"). In the attention heatmaps **d** and **e**, the upper strings represent the SMILES of the main product from the photocatalytic reaction (output), while the lower strings denote the SMILES of the reactants and the input of the four reaction conditions. The output of the product's SMILES emphasizes the input of the four reaction conditions (highlighted in purple). Notably, when the key functional group of ketone carbonyl "C(= O)" (**d**) or alkene "C = C" (**e**) is outputted, PhotoCat pays special attention to the corresponding inputs of B0 and B31, respectively.

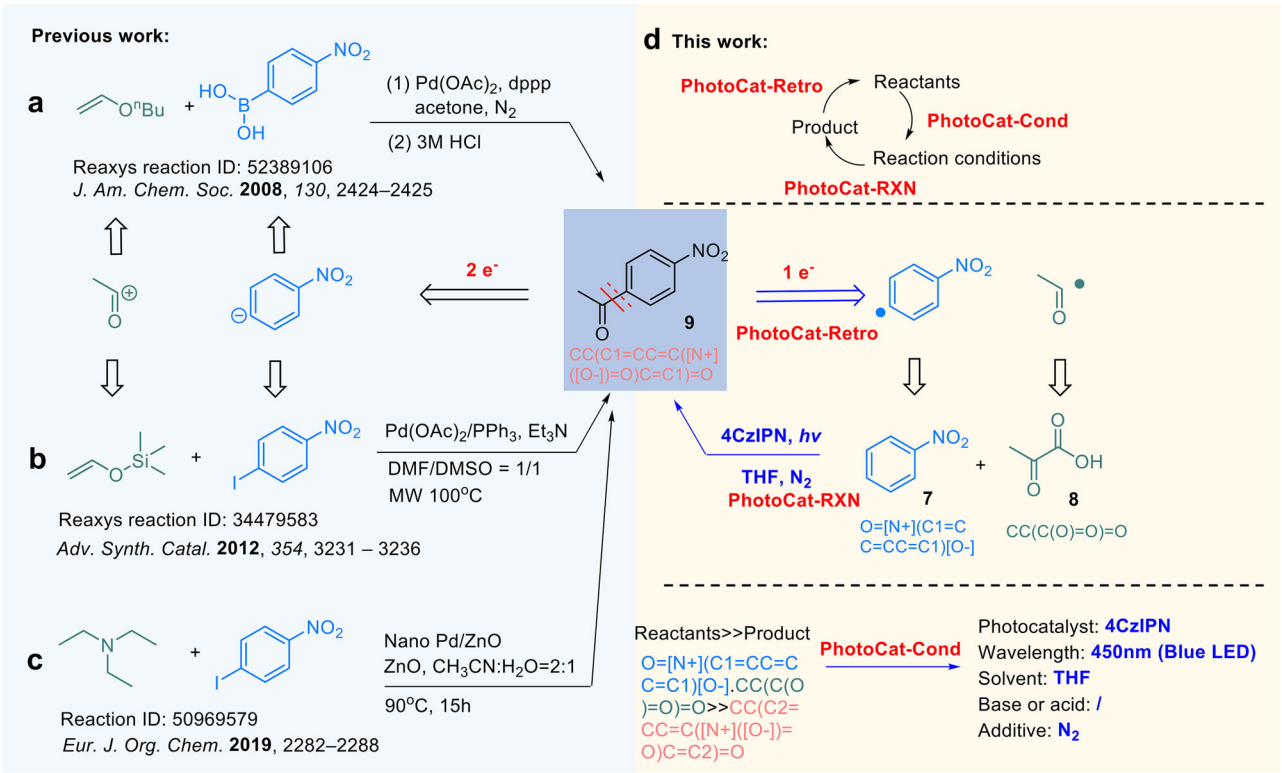


Fig. 8 | Comparison of conventional synthetic strategies and the photocatalytic retrosynthetic pathway proposed by PhotoCat-Retro. a–c Representative literature-reported approaches for aromatic ketone **9** synthesis based on Pd-catalyzed two-electron acylation strategies. **d** The retrosynthetic disconnection proposed by PhotoCat-Retro, featuring a photocatalytic radical acylation between nitrobenzene **7** and pyruvic acid **8**, which was subsequently validated experimentally.

involves using PhotoCat-Cond, where the reaction equation composed of the starting materials and reaction products is input to analyze possible reaction conditions, including photocatalysts, acids or bases, additives, solvents, and wavelengths. The third step is to use PhotoCat-RXN, inputting the reactants and conditions for reaction prediction. To demonstrate the practical utility of PhotoCat, we experimentally validated its Top-1 reaction predictions. This in silico reaction analysis quickly validates wet-lab experimental plans to accelerate the development of photocatalytic reactions and reduce trial-and-error costs. To ensure transparency, we summarize the overall prediction-validation workflow in Supplementary Table 27. PhotoCat-Retro generated 22 reaction candidates, of which 17 were excluded because they closely resembled reported literature reactions (Supplementary Table 28). The remaining 5 predictions, judged both novel and chemically feasible, were experimentally tested, resulting in 4 successful and 1 unsuccessful outcome, with full details provided in Supplementary Table 29. Detailed selection criteria and novelty assessment procedures are described in the "Methods" section.

With the assistance of PhotoCat, our team designed four novel photocatalytic reactions (Fig. 10b–e) in two hours, which had not been reported before. Subsequent wet-lab experiments confirmed that, under the recommended reaction conditions, reactions **b–e** produced the expected products efficiently. Control experiments (Supplementary Fig. 13), luminescence quenching screening studies, and electron paramagnetic resonance (EPR) analyses (Fig. 10f–i) established that these reactions proceed via a photo-induced radical initiation mechanism. Detailed comparisons of reactions **b–e** with the most closely related literature precedents are provided in Supplementary Table 30, clearly highlighting the novel features of our strategy over existing methods. Complete experimental procedures and mechanistic studies are described in Part 3 of the ESI.

In reaction **b**, pyruvate **7** and nitrobenzene **8** combine to produce aromatic ketone **9** under photocatalytic conditions. This reaction is reminiscent of the Friedel–Crafts acylation reaction. The introduction of strong electron-withdrawing groups, like the nitro group, desensitizes the benzene ring, making the execution of the Friedel–Crafts acylation reaction more challenging[73]. A notable advantage of Reaction **b** is that it obviates the need for the large amounts of Lewis acids (e.g., $AlCl_3$) typically required in traditional Friedel–Crafts acylations. Luminescence quenching screening studies (Fig. 10f) confirmed that Reaction **b** proceeds via a photocatalyzed radical pathway.

In reaction **c**, 2-aminophenol **10** and benzaldehyde **11** undergo a mild, catalyst-free photocatalytic transformation to afford 2-phenylbenzoxazole **12**. While recent reports have achieved the synthesis of product 12 through more complex strategies, such as using atomically dispersed Co/N-doped carbon catalysts obtained via high-temperature pyrolysis of ZnCo-ZIF precursors[74], or employing NaCN as a catalyst[75] with significant environmental hazards—our approach proceeds under ambient conditions without additional catalysts. This simplicity underscores the novelty of reaction **c**, providing a greener and more straightforward alternative that expands the chemist's synthetic toolbox.

In reaction **d**, the photocatalytic synthesis of α-trifluoromethyl-substituted ketone **13** is achieved using cinnamic acid **14** and CF_3SO_2Na **15**. Compared with recently reported synthetic methods that employ $AgNO_3$ and $K_2S_2O_8$ as a catalytic oxidation system[76], this approach requires only air as the oxidant and avoids the necessity of metals, external oxidants, or photocatalysts. The UV–visible absorption spectra (Fig. 10j) reveal an overlap between the absorption spectrum of cinnamic acid and the emission wavelength, indicating that compound **14** can absorb these wavelengths and form excited-state species. EPR trapping experiments using TEMP (Fig. 10h) and DMPO (Fig. 10i) confirmed the presence of singlet oxygen ($^1O_2$) and superoxide radical ($O_2^{\bullet-}$) in the reaction under purple LED irradiation. To the best of our knowledge, this represents the inaugural documentation of an oxidative decarboxylative trifluoromethylation of α,β-

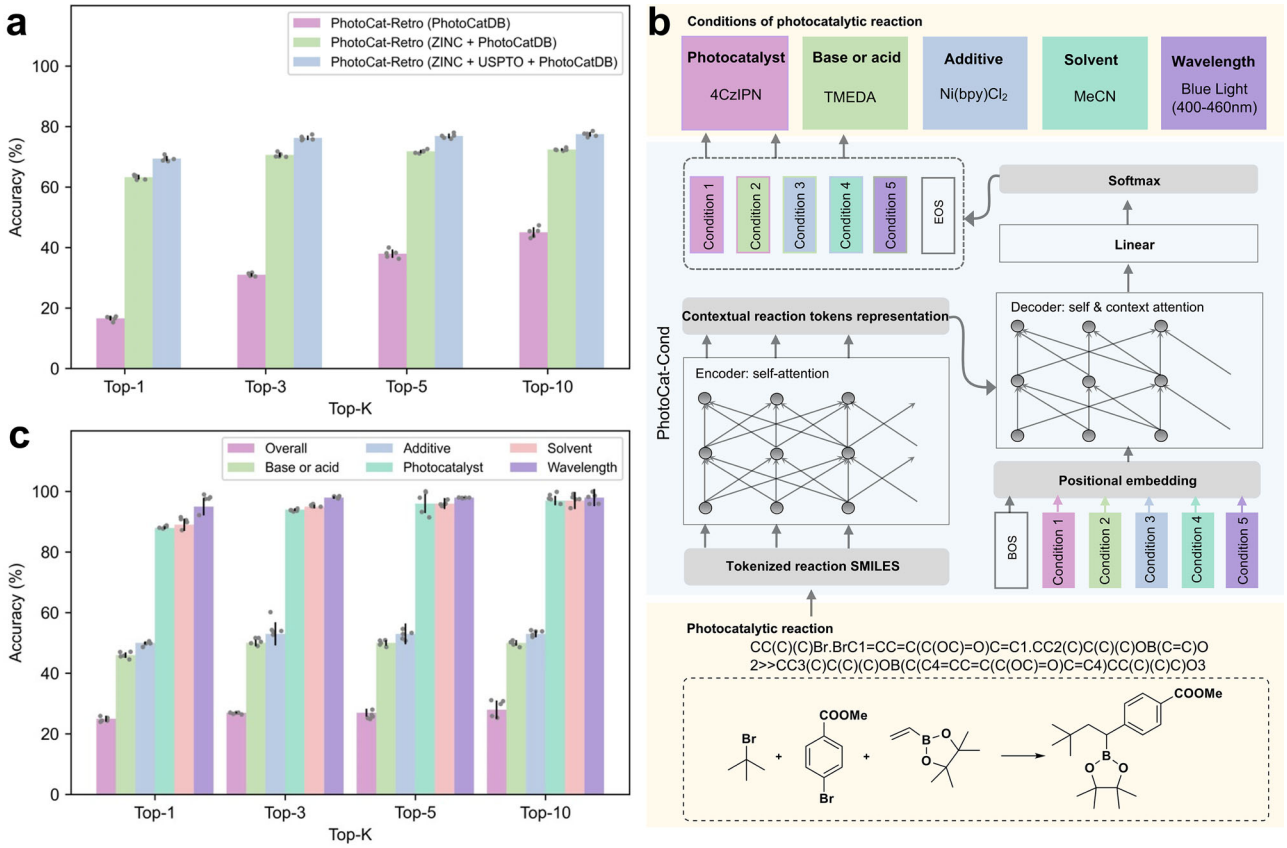


Fig. 9 | PhotoCat includes features for photocatalytic reaction retrosynthesis (PhotoCat-Retro) and reaction condition recommendation (PhotoCat-Cond). **a** With the help of transfer learning from the ZINC molecular database and the USPTO reaction database, PhotoCat-Retro shows a significant improvement in the accuracy of reaction retrosynthesis. **b** A Transformer-based model specifically developed for the recommendation of optimal conditions in photocatalytic reactions. **c** Test results of PhotoCat-Cond.

unsaturated carboxylic acid employing cost-effective CF_3SO_2Na via any photocatalytic approach[77].

In reaction **e**, a photo-triggered oxo-amination of an inactivated alkene **16** is developed, leading to the synthesis of α-amino ketones **18**[78]. This synthetic strategy resembles the recently reported photocatalytic method by the Shao group[79], but utilizes the more economical eosin Y as the photocatalyst. It highlights a vicinal heterodifunctionalization of readily available olefin feedstocks, enabling the one-step construction of the target product.

Conclusions

This study introduces PhotoCat, a deep learning platform based on Transformer architecture, designed to advance photocatalytic reaction research through artificial intelligence. We created the first specialized photocatalytic reaction database, PhotoCatDB, comprising 26,700 curated entries classified by reaction mechanisms. Using PhotoCatDB, we developed three Transformer-based modules: PhotoCat-RXN for reaction prediction, PhotoCat-Retro for retrosynthesis, and PhotoCat-Cond for reaction condition recommendation. These modules achieved top-1 accuracies of 82.3%, 69.4%, and 88.5%, respectively, demonstrating performance comparable to state-of-the-art models. Four practical photocatalytic pathways were designed and validated through wet-lab experiments, with mechanistic studies confirming that all reactions proceed through a photo-induced radical mechanism.

A significant additional finding is that classification and simplification of reaction conditions can markedly enhance the prediction accuracy of deep learning models for chemical reaction tasks. This approach addresses the common challenge that the inclusion of complex reaction conditions often reduces predictive performance. Furthermore, attention weight analysis reveals that this strategy improves model interpretability, providing insights into how reaction conditions affect prediction accuracy.

Overall, PhotoCat is a powerful tool for chemists developing photocatalytic reactions and has the potential to accelerate the broader adoption of photocatalysis as a green synthesis technology. This study also offers valuable guidance for researchers involved in scientific database construction. Future work will focus on expanding PhotoCatDB to better capture the complexity of photocatalytic reactions and further enhance PhotoCat's utility in experimental research. In parallel, integrating PhotoCat with general synthetic planning tools will ensure that photocatalysis is proposed only when truly beneficial, thereby improving its practical value for end users.

Methods

PhotoCatDB

The PhotoCatDB is developed from a comprehensive review of photocatalytic reactions and our group's expertise on photocatalytic reactions[80–85]. This database exclusively includes records from peer-reviewed published literature, excluding any pre-print versions or papers assessed as subjectively unreasonable. A team of 15 data collectors extracted and analyzed reactions from diverse sources. By analyzing the mechanisms, they organized the information into three main categories: reaction equations (expressed using SMILES; standardization of all reactions was performed using RDKit[86]), reaction conditions, and additional details. After cross-checking, the collected data was incorporated into PhotoCatDB. Reaction conditions were represented by common names mapped to unique token IDs rather than full SMILES, improving readability and reducing token complexity. Photocatalysts often contain large conjugated structures, making SMILES strings

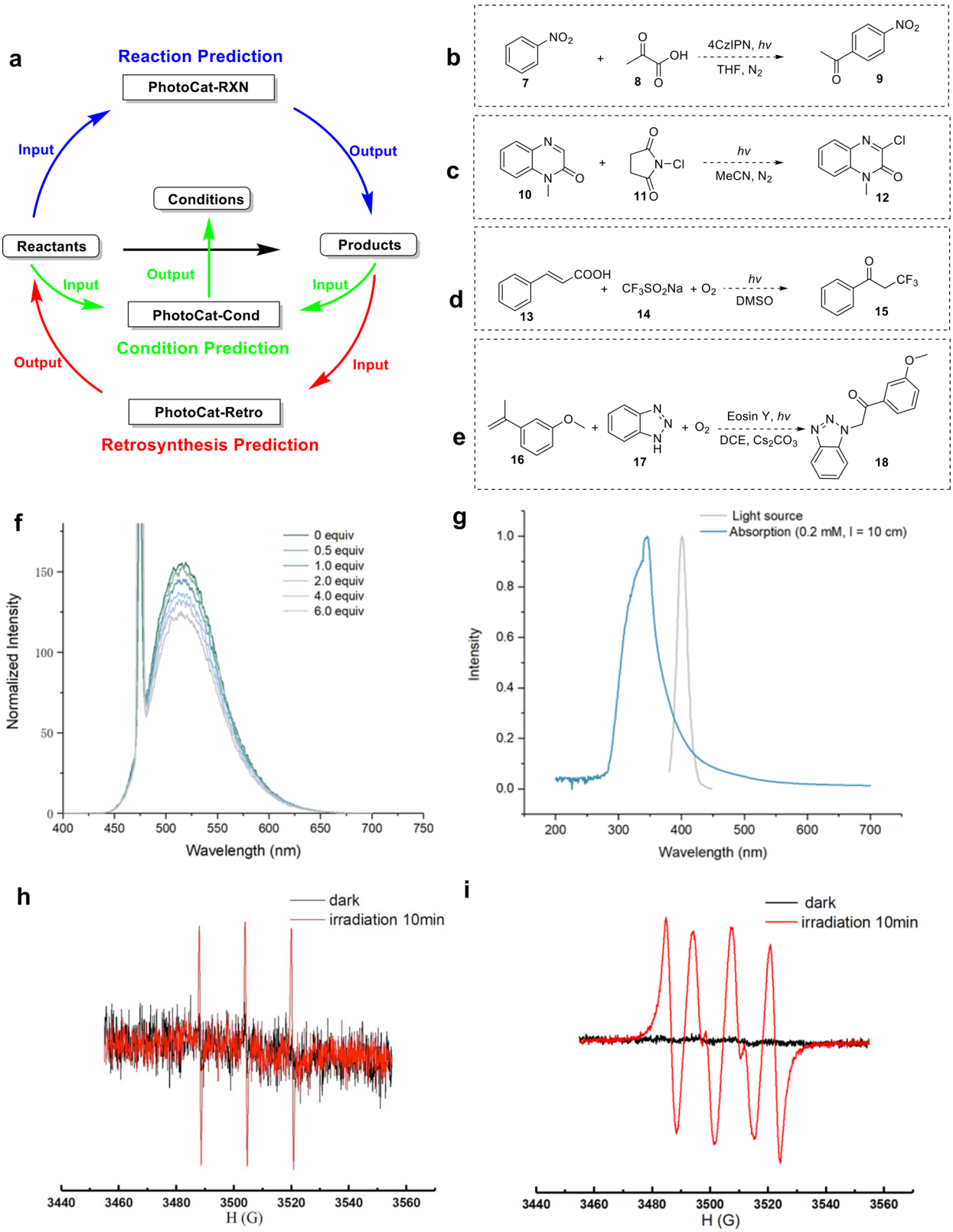


Fig. 10 | Photocatalytic synthesis planning and experimental validation guided by the PhotoCat. a Workflow of dry-lab (in silico) reaction planning using PhotoCat, integrating retrosynthesis, condition recommendation, and reaction prediction. **b–e** Four photocatalytic reactions proposed by PhotoCat and successfully validated through wet-lab experiments. **f–i** Mechanistic investigations supporting light-induced radical pathways.

long and impractical for modeling. Similar simplification strategies have been used previously[28,42]. While common names are less standardized, we mitigate this by unique ID mapping, SMILES cross-validation, and leveraging the diversity of PhotoCatDB. A current limitation of PhotoCatDB is the lack of fine-grained reaction class annotations, which prevents us from systematically evaluating prediction performance by reaction type. Future expansions of PhotoCatDB will address this by incorporating more detailed labels, thereby enabling such analyses.

TMAP generation

TMAP[87] is a dimensionality reduction algorithm designed to handle millions of data points. Its advantage over other dimensionality reduction approaches lies in its two-dimensional tree-like output, which preserves both local and global structures, with an emphasis on local neighborhoods. The algorithm consists of four key steps: (1) LSH Forest-based indexing, (2) *k*-nearest-neighbor graph generation, (3) minimum spanning tree construction using Kruskal's algorithm, and (4) creation of a tree-like layout. The resulting visualization is rendered using the interactive framework Faerun[88]. Following Schwaller et al.[52], we generated reaction fingerprints using rxnfp (https://github.com/rxn4chemistry/rxnfp) and employed the default parameters for LSH Forest indexing, *k*-nearest-neighbor graph construction, and minimum spanning tree generation, which are sufficient to reveal the global structure of the dataset distribution. To improve interpretability in large datasets, we further adjusted node size and inter-cluster spacing during layout rendering. Labels and clustering parameters used in this study are publicly available at https://github.com/su-group/PhotoCat/tree/main/cluster_tmap, ensuring full transparency and reproducibility.

PhotoCat-RXN

PhotoCat-RXN employs the Transformer model developed by Schwaller et al.[24]. The model was implemented using PyTorch as the backend deep learning framework. Both the encoder and decoder consisted of six layers. The word vectors and the recurrent neural network (RNN) had a dimensionality of 512. Gradient accumulation was performed eight times with a maximum vector norm of 0.0. The optimization process utilized the Adam optimizer with β_1 set to 0.9 and β_2 set to 0.998. The batch size was 4096, with batch type and gradient normalization method based on tokens. The learning rate was set to 2.0, and the decay method followed the Noam scheme. A dropout rate of 0.1 was applied, along with label smoothing (ε) set to 0.1. Parameter initialization was disabled, while position encoding was enabled.

PhotoCat-Retro

PhotoCat-Retro employs Chemformer[65], a state-of-the-art model specifically designed for chemical reaction retrosynthesis. Training was conducted for 100 epochs with a learning rate of 0.001, following a cyclical learning rate schedule. A batch size of 64 was used, with gradient accumulation over four batches. The retrosynthesis model training process involved three distinct stages. Initially, molecule-level pretraining was carried out on the ZINC-15[66] dataset, containing 100 million molecules. During this phase, SMILES strings were subjected to span-masking, wherein contiguous short sequences were stochastically replaced with a "<MASK>" token, thereby facilitating a deeper model comprehension of atomic connectivity and bonding patterns. Next, the pretrained molecular model underwent reaction-level pretraining using the USPTO dataset, comprising approximately 1 million chemical reactions. This step aimed to equip the model with the ability to recognize and learn diverse chemical reaction patterns. Finally, the model was fine-tuned specifically for photocatalytic reaction retrosynthesis tasks using the PhotoCatDB dataset, resulting in the specialized PhotoCat-Retro model.

PhotoCat-Cond

PhotoCat-Cond is a Transformer-based deep learning model derived from Parrot[70]. The model is specifically optimized for predicting reaction conditions in photocatalytic reactions, aiming to assist chemists in selecting optimal parameters. PhotoCat-Cond is trained on PhotoCatDB-Cond, a curated subset designed for this task, and performs simultaneous classification of photocatalysts, solvents, and reagents in a single training process, improving both efficiency and accuracy. It is fine-tuned for two epochs using a small learning rate and a 5-fold data-augmented training set.

Transfer learning and data augmentation

Transfer learning is a technique that leverages pre-trained models on a large dataset to improve performance on a related but typically smaller dataset. By transferring knowledge from a broader domain, models can achieve better generalization, especially when training data is limited. In our research, we utilized the USPTO dataset, originally derived from Lowe's dataset[89], which consists of data extracted from patents filed in the United States Patent and Trademark Office. We preprocessed this dataset by excluding reagents, solvents, temperature, and other reaction conditions, and subsequently filtered to eliminate duplicate, incorrect, and incomplete reactions. In this study, transfer learning was applied using a convex weighting scheme, where the USPTO dataset and the fine-tuning dataset were assigned weights of 9 and 1, respectively, following the approach described by Pesciullesi[43]. This strategy enables the model to retain broad chemical knowledge from USPTO while adapting to the specific characteristics of the fine-tuning dataset. Additionally, in PhotoCat-Cond, we incorporated SMILES-based data augmentation to enhance the accuracy of reaction condition predictions. Data augmentation was applied using multiple SMILES-based strategies. Specifically, we employed (i) randomized SMILES generation by altering atom order, (ii) canonical and non-canonical SMILES expansion, and (iii) span masking of subsequences with a <MASK> token. These augmentations, applied with a probability of 0.5 per sample, encourage the model to generalize across equivalent molecular encodings and mitigate overfitting. By generating augmented representations of molecular structures, this technique helps the model generalize better across diverse reaction conditions, ultimately improving its predictive performance.

Cross-validation

Cross-validation, a widely adopted machine learning evaluation technique, assesses a model's performance and generalization ability by dividing the dataset into exclusive subsets for training and validation. It effectively tackles overfitting and underfitting concerns while offering a comprehensive understanding of the model's real-world performance. In this paper, all models were constructed and evaluated using 5-fold cross-validation, ensuring the robustness of the results.

Reaction design selection and novelty assessment

For each target transformation, the integrated PhotoCat workflow (*PhotoCat-Retr*, *PhotoCat-Cond*, *PhotoCat-RXN*) was applied to generate candidate reactions. In terms of data format, taking PhotoCat-RXN as an example, the model inputs comprise reactant SMILES along with encoded condition labels (photocatalyst, solvent, base/acid, additives, and wavelength), while the output corresponds to the predicted product SMILES. Only the Top-1 prediction for each target was considered, provided that it employed readily available reagents under safe and practical conditions. For laboratory validation, concentrations and stoichiometries were guided by the molar ratios suggested by PhotoCat in combination with conventional practice for small-scale organic synthesis (0.2–0.9 mmol substrates in 2 mL solvent). Candidates overlapping with known literature or requiring rare reagents were excluded. Novelty was assessed using *Reaxys* and *SciFinder* searches with product structures, reaction types, and substrate scopes as queries. The proportion of novel reactions and the overall success rate were calculated as described in Table S26.

Chemical synthesis

Unless otherwise specified, all reagents and solvents were obtained from commercial suppliers and used without further purification. The NMR spectra were recorded on a Bruker Avance 400 spectrometer at 400 MHz in $CDCl_3$ with tetramethylsilane as the internal standard. Chemical shifts (δ) are reported in parts per million (ppm) and coupling constants (J) are reported in hertz (Hz). High-resolution mass spectra were obtained with a Bruker Impact II UHR-QTOF by electrospray ionization (ESI) on a time-of-flight (TOF) mass analyzer. Steady-state and time-resolved emission spectroscopy were conducted using an Edinburgh FLS1000. Column chromatography was performed on silica gel (200–300 mesh).

Reaction **b**: A mixture of nitrobenzene **7** (0.2 mmol), pyruvic acid **8** (0.4 mmol), and THF (2 mL) was added to a reaction tube. The tube was evacuated and backfilled with N_2 three times. The mixture was then

irradiated by 360–365 nm for 24 h. After completion of the reaction, the resulting mixture was extracted with CH_2Cl_2, and the organic phase was then removed under vacuum. The residue was purified by column chromatography using a mixture of petroleum ether and ethyl acetate as eluent to give the desired product **9** with 70.5% yield.

Reaction **c**: A mixture of 2-aminophenol **10** (0.2 mmol), benzaldehyde **11** (0.4 mmol), and DCM (2 mL) was added to a reaction tube. The reaction mixture was open to the air and stirred at room temperature under the irradiation of a 390 nm LED lamp for 48 h. After completion of the reaction, the resulting mixture was extracted with CH_2Cl_2, and the organic phase was then removed under vacuum. The residue was purified by column chromatography using a mixture of petroleum ether and ethyl acetate as eluent to give the desired product **12** with 71.1% yield.

Reaction **d**: A mixture of cinnamic acid **13** (0.2 mmol), CF_3SO_2Na **14** (0.4 mmol), and DMSO (2 mL) was added to a reaction tube. The reaction mixture was opened to the air and stirred at room temperature under the irradiation of purple light for 5 h. After completion of the reaction, the resulting mixture was extracted with CH_2Cl_2, and the organic phase was then removed under vacuum. The residue was purified by column chromatography using a mixture of petroleum ether and ethyl acetate as eluent to give the desired product **15** with 75.3% yield.

Reaction **e**: In an oven-dried reaction tube equipped with a magnetic stirrer bar was charged with α-methylstyrene **16** (0.9 mmol), benzotriazole **17** (0.3 mmol), cesium carbonate (0.9 mmol), Eosin Y (3.0 mol%), and DCE (2.0 mL). The tube was then exposed to blue LED irradiation at room temperature under an O_2 atmosphere with stirring for 36 h. After completion of the reaction, the resulting mixture was extracted with CH_2Cl_2, and the organic phase was then removed under vacuum. The residue was purified by column chromatography using a mixture of petroleum ether and ethyl acetate as eluent to give the desired product **18** with a yield of 63.7%.

Data availability

The PhotoCatDB and supplementary datasets used in this study are available at https://doi.org/10.6084/m9.figshare.24532918.

Code availability

The source code for PhotoCat, along with the associated Python scripts for data preparation, is publicly available at https://github.com/su-group/PhotoCat and archived at https://doi.org/10.5281/zenodo.18149429.

Received: 23 June 2025; Accepted: 7 January 2026;
Published online: 21 January 2026

References

1. Melchiorre, P. Introduction: photochemical catalytic processes. *Chem. Rev.* **122**, 1483–1484 (2022).
2. Huang, H., Steiniger, K. A. & Lambert, T. H. Electrophotocatalysis: combining light and electricity to catalyze reactions. *J. Am. Chem. Soc.* **144**, 12567–12583 (2022).
3. Lin, H. et al. 3D-Printed photocatalysts for revolutionizing catalytic conversion of solar to chemical energy. *Prog. Mater. Sci.* **151**, 101427 (2025).
4. Jing, L., Li, P., Li, Z., Ma, D. & Hu, J. Influence of π–π interactions on organic photocatalytic materials and their performance. *Chem. Soc. Rev.* **54**, 2054–2090 (2025).
5. He, S., Chen, Y., Fang, J., Liu, Y. & Lin, Z. Optimizing photocatalysis via electron spin control. *Chem. Soc. Rev.* **54**, 2154–2187 (2025).
6. Lamb, M. C. et al. Electrophotocatalysis for organic synthesis. *Chem. Rev.* **124**, 12264–12304 (2024).
7. Zubkov, M. O. & Dilman, A. D. Radical reactions enabled by polyfluoroaryl fragments: photocatalysis and beyond. *Chem. Soc. Rev.* **53**, 4741–4785 (2024).
8. Zhang, Y., Zhou, G., Liu, S. & Shen, X. Radical Brook rearrangement: past, present, and future. *Chem. Soc. Rev.* **54**, 1870–1904 (2025).
9. Cheng, Y. et al. Outstanding photocatalytic degradation performance under low light intensity via mitigating the quenching reaction of oxygen-centered organic radicals and oxygen. *Appl. Catal. B: Environ. Energy* **355**, 124166 (2024).
10. Saini, P. et al. Photocatalytic single electron reduction of CO_2 into carbon dioxide radical anion ($CO_2^{\cdot-}$): generation, detection and chemical utilization. *J. Energy Chem*. https://doi.org/10.1016/j.jechem.2025.02.013 (2025).
11. Grover, J., Sebastian, A. T., Maiti, S., Bissember, A. C. & Maiti, D. Unified approaches in transition metal catalyzed C(sp^3)–H functionalization: recent advances and mechanistic aspects. *Chem. Soc. Rev.* **54**, 2006–2053 (2025).
12. Wang, P.-Z., Zhang, B., Xiao, W.-J. & Chen, J.-R. Photocatalysis meets copper catalysis: a new opportunity for asymmetric multicomponent radical cross-coupling reactions. *Acc. Chem. Res.* **57**, 3433–3448 (2024).
13. Xu, Q. et al. Renewable ultrathin carbon nitride nanosheets and its practical utilization for photocatalytic decarboxylation free radical coupling reaction. *Chem. Eng. J.* **466**, 142990 (2023).
14. Coppola, G. A., Pillitteri, S., Van der Eycken, E. V., You, S.-L. & Sharma, U. K. Multicomponent reactions and photo/electrochemistry join forces: atom economy meets energy efficiency. *Chem. Soc. Rev.* **51**, 2313–2382 (2022).
15. Garbarino, S., Ravelli, D., Protti, S. & Basso, A. Photoinduced multicomponent reactions. *Angew. Chem. Int. Ed.* **55**, 15476–15484 (2016).
16. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
17. Zhang, X. et al. π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nat. Commun.* **16**, 267 (2025).
18. Marchand, A. et al. Targeting protein–ligand neosurfaces with a generalizable deep learning tool. *Nature* https://doi.org/10.1038/s41586-024-08435-4 (2025).
19. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
20. Keith, J. A. et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
21. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
22. Gale, E. M. & Durand, D. J. Improving reaction prediction. *Nat. Chem.* **12**, 509–510 (2020).
23. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
24. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
25. Hoque, A., Surve, M., Kalyanakrishnan, S. & Sunoj, R. B. Reinforcement learning for improving chemical reaction performance. *J. Am. Chem. Soc.* **146**, 28250–28267 (2024).
26. Wang, L., Zhang, C., Bai, R., Li, J. & Duan, H. Heck reaction prediction using a transformer model based on a transfer learning strategy. *Chem. Commun.* **56**, 9368–9371 (2020).
27. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
28. Probst, D. et al. Biocatalysed synthesis planning using data-driven learning. *Nat. Commun.* **13**, 964 (2022).
29. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
30. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).

31. Wang, X. et al. RetroPrime: a diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chem. Eng. J.* **420**, 129845 (2021).
32. 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).
33. Chen, Z., Ayinde, O. R., Fuchs, J. R., Sun, H. & Ning, X. G2Retro as a two-step graph generative models for retrosynthesis prediction. *Commun. Chem.* **6**, 102 (2023).
34. Xu, J. et al. Providing direction for mechanistic inferences in radical cascade cyclization using a Transformer model. *Org. Chem. Front.* **9**, 2498–2508 (2022).
35. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
36. Vaucher, A. C. et al. Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **12**, 2573 (2021).
37. Tu, Z., Stuyver, T. & Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **14**, 226–244 (2023).
38. Hua, P.-X. et al. An active representation learning method for reaction yield prediction with small-scale data. *Commun. Chem.* **8**, 42 (2025).
39. de Almeida, A. F., Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **3**, 589–604 (2019).
40. Caramelli, D. et al. Discovering new chemistry with an autonomous robotic platform driven by a reactivity-seeking neural network. *ACS Cent. Sci.* **7**, 1821–1830 (2021).
41. Su, A. et al. Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions. *Phys. Chem. Chem. Phys.* **24**, 10280–10291 (2022).
42. Kreutter, D., Schwaller, P. & Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **12**, 8648–8659 (2021).
43. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
44. Zahrt, A. F. et al. Machine-learning-guided discovery of electrochemical reactions. *J. Am. Chem. Soc.* **144**, 22599–22610 (2022).
45. Mai, H., Le, T. C., Chen, D., Winkler, D. A. & Caruso, R. A. Machine Learning for electrocatalyst and photocatalyst design and discovery. *Chem. Rev.* **122**, 13478–13515 (2022).
46. Su, A. et al. Deep transfer learning for predicting frontier orbital energies of organic materials using small data and its application to porphyrin photocatalysts. *Phys. Chem. Chem. Phys.* **25**, 10536–10549 (2023).
47. Radestock, S. Optimising chemical information workflows: integrating Reaxys—use cases and applications. *J. Cheminform.* **5**, P39 (2013).
48. Somerville, A. N. SciFinder Scholar (by Chemical Abstracts Service). *J. Chem. Educ.* **75**, 959 (1998).
49. Kearnes, S. M. et al. The Open Reaction Database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
50. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *J. Cheminform.* **7**, 20 (2015).
51. Skinnider, M. A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* **6**, 437–448 (2024).
52. Schwaller, P. et al. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).
53. Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. M. *OpenNMT: Open-Source*. *Toolkit for Neural Machine Translation* (Harvard NLP group, SYSTRAN, 2017).
54. Zhong, Z. et al. Root-aligned SMILES: a tight representation for chemical reaction prediction. *Chem. Sci.* **13**, 9023–9034 (2022).
55. Tu, Z. & Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *J. Chem. Inf. Model.* **62**, 3503–3513 (2022).
56. Zhang, Y. et al. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Org. Chem. Front.* **8**, 1415–1423 (2021).
57. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
58. Wu, Y., Zhang, C., Wang, L. & Duan, H. A graph-convolutional neural network for addressing small-scale reaction prediction. *Chem. Commun.* **57**, 4114–4117 (2021).
59. Chapman, S. J. et al. Cooperative stereoinduction in asymmetric photocatalysis. *J. Am. Chem. Soc.* **144**, 4206–4213 (2022).
60. Qin, Q., Han, Y.-Y., Jiao, Y.-Y., He, Y. & Yu, S. Photoredox-catalyzed diamidation and oxidative amidation of alkenes: solvent-enabled synthesis of 1,2-diamides and α-amino ketones. *Org. Lett.* **19**, 2909–2912 (2017).
61. Xia, Z.-H., Zhang, C.-L., Gao, Z.-H. & Ye, S. Switchable Decarboxylative Heck-type reaction and oxo-alkylation of styrenes with N-hydroxyphthalimide esters under photocatalysis. *Org. Lett.* **20**, 3496–3499 (2018).
62. Bazyar, Z. & Hosseini-Sarvari, M. Visible-light-driven direct oxidative coupling reaction leading to alkyl aryl ketones, catalyzed by nano Pd/ZnO. *Eur. J. Org. Chem.* **2019**, 2282–2288 (2019).
63. Ruan, J., Li, X., Saidi, O. & Xiao, J. Oxygen and base-free oxidative Heck reactions of arylboronic acids with olefins. *J. Am. Chem. Soc.* **130**, 2424–2425 (2008).
64. Qian, W., Zhang, L., Sun, H., Jiang, H. & Liu, H. Microwave-assisted one-step synthesis of acetophenones via palladium-catalyzed regioselective arylation of vinyloxytrimethylsilane. *Adv. Synth. Catal.* **354**, 3231–3236 (2012).
65. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn.: Sci. Technol.* **3**, 015022 (2022).
66. Tingle, B. I. et al. ZINC-22—a free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* **63**, 1166–1176 (2023).
67. Yao, L. et al. Node-aligned graph-to-graph: elevating template-free deep learning approaches in single-step retrosynthesis. *JACS Au* https://doi.org/10.1021/jacsau.3c00737(2024).
68. Han, Y. et al. Retrosynthesis prediction with an iterative string editing model. *Nat. Commun.* **15**, 6404 (2024).
69. Chen, S. & Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **1**, 1612–1620 (2021).
70. Wang, X. et al. Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research* **6**, 0231.
71. Maser, M. R. et al. Multilabel classification models for the prediction of cross-coupling reaction conditions. *J. Chem. Inf. Model.* **61**, 156–166 (2021).
72. Zhang, B. et al. Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning. *Proc. Natl. Acad. Sci. USA* **119**, e2212711119 (2022).
73. Sartori, G. & Maggi, R. Use of solid catalysts in Friedel−Crafts acylation reactions. *Chem. Rev.* **106**, 1077–1104 (2006).
74. Chen, J.-Y. et al. An atomically dispersed Co catalyst for efficient oxidative fabrication of benzoheterocycles under ambient oxygen conditions. *Green Chem.* **26**, 4834–4843 (2024).
75. Cho, Y. H., Lee, C.-Y., Ha, D.-C. & Cheon, C.-H. Cyanide as a powerful catalyst for facile preparation of 2-substituted benzoxazoles via aerobic oxidation. *Adv. Synth. Catal.* **354**, 2992–2996 (2012).

76. Deb, A. et al. Oxidative trifluoromethylation of unactivated olefins: an efficient and practical synthesis of α-trifluoromethyl-substituted ketones. *Angew. Chem. Int. Ed.* **52**, 9747–9750 (2013).
77. Muralirajan, K. et al. Exploring the structure and performance of Cd–chalcogenide photocatalysts in selective trifluoromethylation. *ACS Catal.* **11**, 14772–14780 (2021).
78. Nguyen, Q. H., Hwang, H. S., Cho, E. J. & Shin, S. Energy transfer photolysis of N-enoxybenzotriazoles into benzotriazolyl and α-carbonyl radicals. *ACS Catal.* **12**, 8833–8840 (2022).
79. Wang, J. et al. Visible-light-induced regioselective radical oxo-amination of alkenes with O_2 as the oxygen source. *Org. Lett.* **25**, 5333–5338 (2023).
80. Zhuang, X. et al. Photoinduced cascade C–N/C═O bond formation from bromodifluoroalkyl reagents, amines, and H_2O via a triple-cleavage process. *Org. Lett.* **24**, 1668–1672 (2022).
81. Zhuang, X. et al. Light-fuelled nitro-reduction via cascaded electron donor–acceptor complexes in aqueous media. *Green Chem.* **26**, 9682–9689 (2024).
82. Huang, P. et al. Selective C(sp^3)–H bond aerobic oxidation enabled by a π-conjugated small molecule-oxygen charge transfer state. *Green Chem.* **26**, 9241–9249 (2024).
83. Huang, P. et al. An in situ generated proton initiated aromatic fluoroalkylation via electron donor–acceptor complex photoactivation. *Green Chem.* **26**, 7198–7205 (2024).
84. Huang, P. et al. Catalyst-free intramolecular radical cyclization cascades initiated by the direct homolysis of Csp^3–Br under visible light. *Green Chem.* **25**, 3989–3994 (2023).
85. Sun, B. et al. Decatungstate/cobalt dual catalyzed dehydrogenation of ketones enabled by polarity-matched site-selective activation. *ACS Catal.* **14**, 11138–11146 (2024).
86. Landrum, G. et al. *RDKit: Open-Source Cheminformatics Software, Release 2019-3-4* (2019). https://doi.org/10.5281/zenodo.3366468.
87. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).
88. Probst, D. & Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **34**, 1433–1435 (2018).
89. Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature* (University of Cambridge, 2012).

Acknowledgements

This research was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMZ23B060001 and the National Natural Science Foundation of China under Grant Nos. 22108252, 22078299, and 22478356. Many thanks to Lu Jiejia, Hu Bing, Zhang Zhenyang, Zhuang Yu, Liu Rongwu, Zhou Na, Zhou Shengqi, Huang Xiao, Lei Rong, Peng Jiehai, Shen Shoutao, Xu Zhengcheng, Zhang Xinyi, Li Yue, and Fan Zhidan in Hangzhou Polytechnic University for their contributions to the data collection and verification in PhotoCatDB.

Author contributions

Jiangcheng Xu: Writing—original draft, validation, methodology, data curation, and conceptualization; Silong Zhai: Writing—original draft, software and validation; Panyi Huang: Writing—original draft and validation; Wenbo Yu: Data curation; Qingyi Mao: Software; Kui Du: Validation; Weike Su: Conceptualization; Bin Sun: Supervision and project administration; Can Jin: Supervision and funding acquisition; An Su: Writing—review and editing, supervision, project administration, and funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-026-01894-y.

Correspondence and requests for materials should be addressed to Bin Sun, Can Jin or An Su.

Peer review information *Communications Chemistry* thanks Jolene Reid and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.