Reaction Chemistry & Engineering

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5re00572h

ProcedureT5: adaptive experimental procedure prediction with data-augmented pre-training and multi-source data integration

Yuxuan Zhang,[a] Yue Fang,[a] Haifan Zhou,[a] Bowen Yu,[a] Tsz Fung Fung,[a] Qing Liu,[b] Christophe Len[b] and Hanyu Gao*[a]

Computer-aided synthesis planning (CASP) has shown strong potential to accelerate chemical research. However, a key challenge remains: the lack of effective automated techniques to translate computer-generated synthesis routes into executable experimental procedures, which still require extensive planning and evaluation by chemists. To address this gap, we introduce ProcedureT5, an approach that integrates chemistry-oriented pre-trained models with augmented multi-source datasets to enhance the prediction of experimental procedures across broader scenarios. Our method achieves state-of-the-art performance on the Pistachio dataset – a collection of reaction procedures derived from US patent literature, showing a 4-point increase in BLEU score and a 1.22% improvement in exact-match accuracy compared to existing methods. Additionally, we curate a small expert-annotated dataset, Orgsyn, consisting of verified organic synthesis procedures, to assess the model's performance in more diverse applications. Fine-tuning ProcedureT5 on the Orgsyn dataset demonstrates its adaptability, yielding a BLEU score of 40.34 and an average similarity of 49.72%. This work underscores the crucial role of ProcedureT5 in bridging the gap between computational synthesis planning and practical laboratory implementation.

Received 27th December 2025,
Accepted 7th April 2026

DOI: 10.1039/d5re00572h

rsc.li/reaction-engineering

1 Introduction

Computer-aided synthesis planning (CASP) has demonstrated its potential to assist researchers by identifying viable pathways for the synthesis of target molecules.[1–7] For moderately complex molecules, computational models have been developed to propose synthetic pathways,[2–4,7–10] recommend reaction conditions,[11–17] and assess the likelihood of success for the proposed reactions.[18–23] However, providing the aforementioned information alone is insufficient to facilitate the actual implementation of these pathways in the laboratory.[24–26] For each reaction step in a proposed pathway, chemists must develop detailed experimental procedures that include precise operational guidelines. Traditional experimental design involves iterative trial-and-error, extensive literature reviews on similar reactions, and the application of expert knowledge – processes that demand significant time and expertise. This challenge motivates researchers to develop automated experimental procedure prediction models, enabling the broader adoption of promising retrosynthesis predictions and increasing the automation of target molecule synthesis.

A primary obstacle in developing experimental procedure prediction models is the limited availability of annotated experimental procedure data. To partially address this issue, Vaucher *et al.*[27] proposed a set of annotation criteria that simplified complex experimental procedures into streamlined action sequences. Using this standardized annotation criterion, they manually annotated a small-scale dataset to develop the Paragraph2Action model specialized in automatically extracting experimental procedures from corresponding paragraphs. The model was then utilized to curate a dataset, named Pistachio dataset here, from the Pistachio database,[28] which compiles chemical reactions from literature and patents, including comprehensive metadata on catalysts, reagents, reaction conditions, and yields. With this dataset, Vaucher *et al.*[29] introduced Smiles2Actions, an AI model capable of inferring experimental procedures directly from chemical formulas, with over half of its predicted action sequences considered executable without human intervention by experienced chemists. Liu *et al.*[30] subsequently introduced ReactXT, which integrated SMILES, textual descriptions, and graph embeddings of molecules to enhance the model's understanding of chemical reactions and molecules. They applied Paragraph2Action to curate another dataset based on

[a] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, 999077, P. R. China. E-mail: hanyugao@ust.hk
[b] Institute of Chemistry for Life and Health Sciences, Chimie ParisTech, PSL Research University, 11 rue Pierre et Marie Curie, Paris, F-75005, France

the US patent literature, further refining the model for chemical experimental procedure prediction. Vaškevičius *et al.*[31] also utilized deep learning models to curate a dataset from USPTO and EPO[32] and evaluated multiple language model architectures for organic synthesis procedure prediction, demonstrating that language models can be effectively customized for chemical synthesis procedure prediction. While prior studies have shown that AI-based data extraction can generate valuable resources for training procedure prediction models, the quality and consistency of these automatically extracted datasets remain insufficiently validated, leading to limited reliability in model evaluation and deployment. At the same time, the performance of existing models for experimental procedure prediction remains suboptimal, underscoring the need for more advanced model architectures and higher-quality, expertly curated datasets to drive further progress in this field.

To address these challenges, we introduce ProcedureT5, a training framework that integrates a chemistry-oriented pre-trained model with multi-source augmented datasets to enhance and generalize experimental procedure prediction across diverse scenarios. On the model side, we build upon a domain-specific pre-trained transformer as the foundation of ProcedureT5 and train it on datasets augmented *via* a reaction precursor order-shuffling algorithm. This design leverages the chemical prior knowledge embedded in the model and the increased data diversity through augmentation, improving the model's ability to capture reaction representations and predict experimental procedures with state-of-the-art accuracy. On the dataset side, we construct Orgsyn, an expert-annotated dataset derived from the *Organic Syntheses* journal,[33] containing high-quality experimental protocols for organic compound synthesis. Further augmentation of Orgsyn using the shuffling algorithm enables effective fine-tuning and rigorous evaluation of ProcedureT5, demonstrating the model's robustness and transferability to broader synthetic contexts.

Collectively, these innovations enable ProcedureT5 to address key limitations of prior approaches, namely suboptimal model performance and the absence of rigorously curated benchmark datasets. Our findings indicate that ProcedureT5 can function as a practical tool for supporting real-world experimental design and execution. Furthermore, the experimental results suggest that ProcedureT5 exhibits favorable performance scaling with increasing model capacity and dataset size, thereby outlining a promising trajectory for future developments in experimental procedure prediction.

2 Related work

The development of artificial intelligence has represented a transformative frontier in chemical automation. This section provides a brief review of recent advances across five critical domains that collectively enable chemical reaction automation: chemical information extraction, molecular prediction, reaction condition prediction, retrosynthesis prediction, and autonomous experimentation systems. Chemical information extraction tools convert the vast body of unstructured chemical knowledge in text, tables, and figures into structured, machine-readable datasets. These datasets provide the data foundation for learning-based models of molecular property prediction, reaction condition prediction, and retrosynthetic planning, as well as for the design and operation of autonomous laboratories. Accordingly, the following subsections are organized to mirror this pipeline: we first discuss recent progress in chemical information extraction, and then summarize advances in molecular prediction, reaction condition prediction, retrosynthetic prediction, and autonomous experimentation systems.

2.1 Chemical information extraction

Efficient curation of large, machine-readable reaction datasets is a prerequisite for training data-driven models and deploying autonomous reaction platforms, yet the majority of chemical knowledge remains locked in unstructured literature and laboratory records. Recent developments in large language models (LLMs) have enabled general, end-to-end workflows for structured chemical-data extraction that significantly reduce reliance on hand-crafted rules and task-specific parsers. Schilling-Wilhelmi *et al.* provide a comprehensive tutorial review of LLM-based chemical-data-extraction pipelines, covering document acquisition and parsing, prompt and system design, constrained decoding, and chemically informed validation strategies, and highlight how multimodal and agentic LLM systems can extract reactions, conditions, and material properties at scale from heterogeneous text and document formats.[34] This work establishes methodological foundations for transforming legacy corpora into standardized reaction databases that can be readily consumed by downstream machine learning models. Building on these general frameworks, task-specific LLM systems for reaction information extraction have demonstrated high accuracy on realistic sources of synthetic chemistry data. Ai *et al.* fine-tuned an open-source LLaMA-2 model to convert free-text organic synthesis procedures from US patents into structured records following the Open Reaction Database schema, achieving over 90% accuracy at the field level for compound identifiers, quantities, reaction conditions, and workups, while also inferring context-dependent reaction roles such as reactants, solvent, and catalyst.[35] Complementing text-based extraction, Chen *et al.* introduced RxnIM, a multimodal large language model specifically designed to parse reaction images, which jointly localizes reactants, reagents, and products and interprets textual annotations of reaction conditions to produce machine-readable reaction representations, enabling the scalable construction of reaction databases directly from graphical depictions.[36] Together, these chemical information extraction tools close the loop from unstructured procedures and reaction figures to structured reaction records, providing the curated datasets required to train downstream machine-learning models and support automated reaction discovery.[37–40]

2.2 Molecular prediction

Accurate prediction of molecular properties is foundational to chemical reaction automation and serves as a prerequisite for downstream tasks in drug discovery and materials design. Recent advances in deep learning have shifted the paradigm from manually engineered molecular descriptors to automated feature extraction, enabling data-driven predictions of diverse molecular properties. Among various deep learning models, graph neural networks (GNNs) have emerged as particularly effective architectures for molecular representation learning. Sun *et al.*[41] provided a comprehensive survey on graph convolutional networks in drug discovery, demonstrating their superior ability to automatically extract task-relevant molecular features without complex manual feature engineering. More recently, Kensert *et al.*[42] implemented graph convolutional networks to predict molecular retention times across different chromatographic datasets, showing that GCNs significantly outperformed conventional benchmarks with 5–25% lower mean absolute error. Liu *et al.*[43] proposed a multi-level fusion graph neural network (MLFGNN) that combines graph attention networks with graph transformers to jointly capture both local molecular substructures and global topological dependencies. Their experimental validation across multiple benchmark datasets consistently outperformed state-of-the-art methods in both classification and regression tasks, with improvements in interpretability through attention weight analysis.

2.3 Reaction condition prediction

Predicting optimal reaction conditions—including temperature, solvent, catalyst, and reagent ratios—is essential for accelerating synthetic methodology development. Chen and Li[44] provided a comprehensive review of machine learning-guided strategies for reaction condition prediction, emphasizing the importance of acquiring and processing large, diverse datasets of chemical reactions. Gao *et al.*[11] developed a neural network-based model trained on approximately 10 million reactions from Reaxys to predict optimal reaction conditions—including catalysts, solvents, reagents, and temperature—for arbitrary organic transformations. Kwon *et al.*[12] propose a generative modeling approach using a variational autoencoder augmented with graph neural networks to predict multiple suitable reaction conditions for organic reactions, enabling sampling of diverse feasible condition sets that fully specify catalysts, solvents, and reagents. Wang *et al.*[16] introduce Reacon, a novel framework combining directed message passing neural networks, reaction templates, and a clustering algorithm to predict reaction conditions (catalysts, solvents, and reagents) for organic reactions while simultaneously ensuring chemical feasibility and providing diverse recommendations.

2.4 Retrosynthetic prediction

Retrosynthetic analysis, the inverse planning problem of identifying synthesis routes to target molecules, has undergone substantial transformation through machine learning approaches. Zhao *et al.*[45] introduced Retro-MTGR, a semi-template-based method combining reaction center prediction with leaving group inference, demonstrating superiority over 16 state-of-the-art methods. Zhong *et al.*[46] proposed root-aligned SMILES (R-SMILES), which specifies tightly aligned one-to-one mappings between product and reactant SMILES strings, substantially improving sequence-based synthesis prediction performance over general-purpose SMILES. Zhang *et al.*[47] introduced RetroDFM-R, a reasoning-driven large language model specifically designed for chemical retrosynthesis, significantly outperforming general-domain LLMs, with the double-blind human assessment validating the chemical plausibility of predictions. Wang *et al.*[48] developed RetroExplainer, employing multi-sense and multi-scale graph transformers combined with structure-aware contrastive learning and dynamic adaptive multi-task learning, enabling transparent decision-making in drug development workflows.

2.5 Autonomous experimentation

The integration of artificial intelligence with robotic platforms has enabled the development of fully autonomous laboratories capable of executing closed-loop synthesis, analysis, and optimization with minimal human intervention. Szymanski *et al.*[49] demonstrated one of the first fully autonomous synthesis platforms, A-Lab, which integrated four key components and synthesized 41 of 58 targeted inorganic materials (71% success rate) with minimal human intervention over 17 days of continuous operation. Dai *et al.*[50] developed an autonomous platform combining mobile robots with standard laboratory instruments for exploratory synthetic chemistry, exemplifying how human-like decision-making criteria, rather than single metric optimization, can enable practical autonomous chemistry. Li *et al.*[51] provided a comprehensive perspective on autonomous laboratories in China, identifying fundamental elements required for closed-loop autonomous experimentation: chemical science databases, large-scale intelligent models, automated experimental platforms, and integrated management/decision-making systems.

3 Methods

3.1 Problem definition

Chemical experimental procedure prediction aims to derive specific experimental steps for a given reaction formula. Formally, we formulate this task as a sequence-to-sequence generation problem, where the objective is to learn a mapping $\phi: R \rightarrow S$ that transforms a chemical reaction R into its corresponding experimental procedure sequence S. In this work, SMILES[52,53] is employed as the information carrier for the reaction. We assume catalysts and reagents can be retrieved through other reaction condition recommendation systems and treat them in the same manner as reactants. Then we define the reaction input as $R = \rho(r_1, r_2, \ldots, r_m; p_1, p_2, \ldots, p_n)$ the combination of reactants and products, where

r_i and p_j denote the SMILES representation of the i-th reactant and the j-th product, respectively. The function ρ denotes the SMILES concatenation protocol: individual SMILES components are joined with dot separators, while the reactant set and product set are separated by "≫". For the model output format, we adopt the same approach used by Vaucher *et al.*,[27,29] representing the experimental procedure as a concise sequence of operational steps. Specifically, the output sequence employing a concise sequence of experimental actions to represent the operational steps $S = \sigma(s_1, s_2, \ldots, s_t)$, needed for the target reaction, encodes the stepwise experimental actions, where s_i characterizes the i-th operational step and the concatenation operator σ joins individual steps with semicolon delimiters. Each operational step s_i comprises an experimental action type, Action_i, alongside associated properties, denoted as Props_i. Each step in the sequence consists of an experimental action type and its associated properties $s_i = \varepsilon(\text{Action}_i, \text{Props}_i)$, where ε denotes the method to combine the action with its properties to form an operational step. These properties fall into two categories. The first category specifies compounds involved in the experimental action, represented either as numerical tokens referring to positional indices of corresponding reactants or products in the input reaction SMILES or as their most common synonyms simplified from their original complex descriptions. The second category properties encode action-specific parameters, such as whether to retain the filtrate or precipitate for FILTER operations. Among these properties, temperature and duration for performing the experimental action are mapped to certain numerical tokens according to predefined discretization protocols. Such tokenization representation eliminates the need for the model to learn the compound names, temperature values, and exact durations of experimental actions, thereby improving model performance. Additional details regarding the data format rules are provided in the SI. Fig. 1 illustrates how the raw data for this task is manually or automatically annotated and subsequently used for model training and evaluation.

3.2 Overview of ProecdureT5

The workflow of ProcedureT5, shown in Fig. 2, consists of two stages. In the first stage, we begin by selecting several chemistry-oriented pre-trained models, previously trained on various chemistry-specific tasks such as mask filling and molecular captioning. These models are then trained using the Pistachio dataset, augmented with the reaction precursor order shuffling algorithm, to enable them to predict experimental procedures. The best-performing model from this group is designated as ProcedureT5. In the second stage, we curate a small-scale expert-annotated Orgsyn dataset from the *Organic Syntheses* database[33] and also augment it with the reaction precursor order shuffling algorithm. ProcedureT5 is then fine-tuned and evaluated on this dataset. This two-stage training workflow ensures that ProcedureT5 benefits from extensive chemical knowledge during pretraining and data augmentation, while also being evaluated in a broader range of scenarios.

3.3 Dataset

3.3.1 Pistachio dataset. Sourced from the Pistachio database,[28] the Pistachio dataset[27] contains 693 517 reaction entries, each linking a canonicalized SMILES string to experimental procedures. The dataset is divided into training, validation, and test subsets, containing 554 813, 69 352, and 69 352 entries, respectively, following a 70%, 15%, and 15% split. We then further process the dataset by removing entries that do not involve product generation operations or include

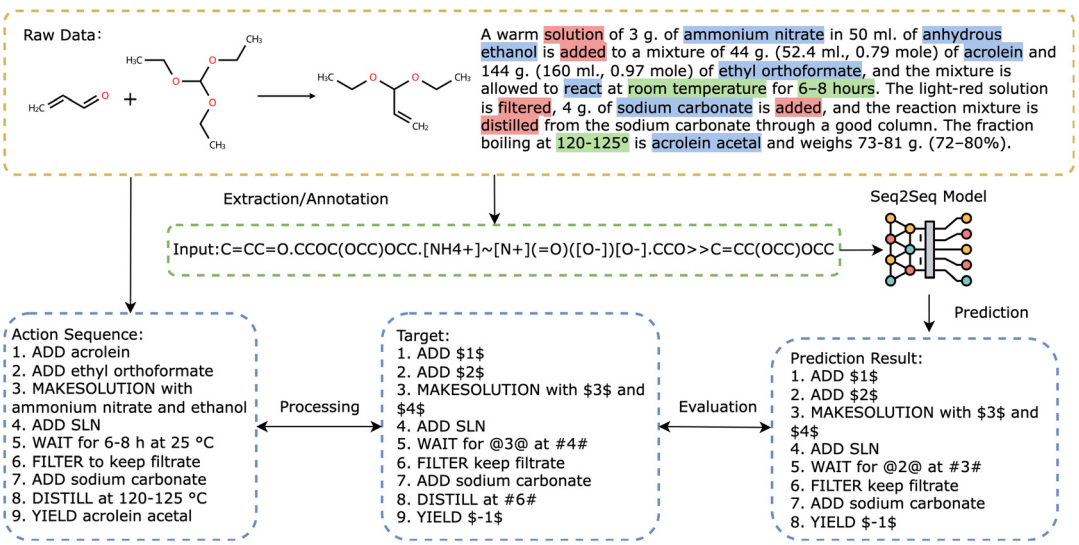


Fig. 1 Overview of the prediction task: the dataset for this task is initially generated through either automatic extraction or manual annotation, following a predefined data format. This dataset is then used to train and evaluate the predictive model. The blue, red, and green masks in the raw procedure text represent the identified actions, compounds, and action properties, respectively.

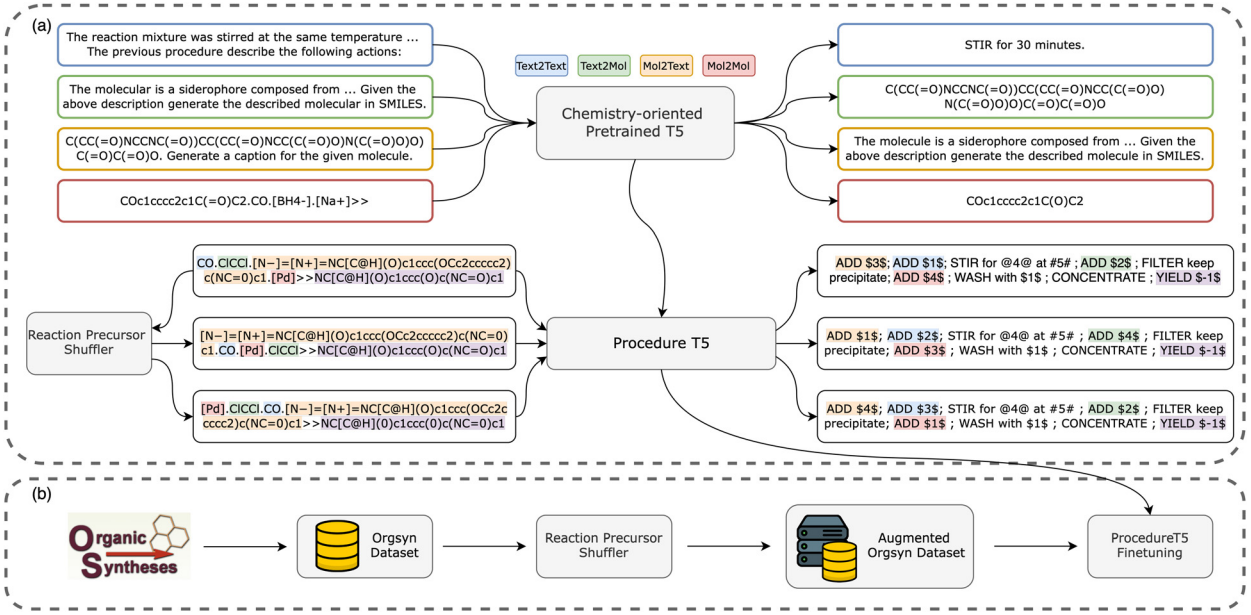


Fig. 2 The workflow of ProcedureT5. (a) Training with the Pistachio dataset: a chemistry-related pre-trained T5 model, which implicitly stores chemical knowledge in its parameters, is selected as the foundation model for ProcedureT5. The model is then trained with the Pistachio dataset, augmented with the reaction precursor order shuffling algorithm. (b) Fine-tuning with the Orgsyn dataset: ProcedureT5 is further fine-tuned using the Orgsyn dataset, which is similarly augmented by shuffling the precursor addition orders.

product addition operations. This led to the deletion of 13 118 entries, resulting in a final dataset size of 680 399 entries, with 545 010 for training, 67 751 for validation, and 67 638 for testing, as summarized in Table 1. The shuffling configuration is explained in section 3.3.4.

3.3.2 Orgsyn dataset. Exclusive reliance on the Pistachio dataset for training and evaluation may bias the evaluation of model performance due to its limitations. First, the Pistachio dataset was curated using the Paragraph2Actions model,[27] which introduces potential inaccuracies in procedure extraction. These inaccuracies can undermine the validity of model evaluations. Second, discrepancies exist between different databases of chemical experimental procedures, stemming from factors such as variations in reaction type distributions and the styles of experimental text recording. These differences make it impractical to generalize models trained on a single database for experimental step extraction or experimental procedure prediction to other databases. To validate model performance in broader contexts and establish a new benchmark in this domain, we manually curate the Orgsyn dataset from the *Organic Synthesis* journal,[33] which publishes reliable experimental procedures for synthesizing organic compounds. Fig. 3 illustrates the workflow for annotating the Orgsyn dataset. Initially, we scrape data entries from the journal's website,[33] each containing experimental texts and reaction diagrams. We then recruit annotators with at least an undergraduate-level chemistry background to annotate these entries using our custom-built data annotation platform, following specific guidelines. After annotation, the entries undergo automatic review, manual revision, and post-processing to ensure the dataset's quality. Detailed information about the annotation guidelines, annotation platform, and data processing is provided in the Notes S1 and S2. The resulting Orgsyn dataset contains 996 entries, divided into 698 training entries, 149 validation entries, and 149 test entries, with a 70%, 15%, and 15% split, respectively.

Table 1 The information of datasets

Dataset name	Training set size	Validation set size	Test set size	Shuffling configuration[a]
Pistachio	545 010	67 751	67 638	—
Orgsyn	698	149	149	—
Pistachio_Aug_1	1 579 236	67 751	67 638	$(1, 4):1, (4, 6):3, (6, +\infty):5$
Pistachio_Aug_2	2 059 396	67 751	67 638	$(1, 3):1, (3, 5):3, (5, 7):5, (7, +\infty):7$
Pistachio_Aug_3	3 028 790	67 751	67 638	$(1, 3):1, (3, 5):5, (5, 7):9, (7, +\infty):13$
Orgsyn_Aug	4045	149	149	$(1, 3):1, (3, 5):5, (5, 7):9, (7, +\infty):13$

[a] For shuffling configuration $(a, b):c$, the reaction addition order augmentation algorithm will perform c times to generate new data entries corresponding to any original entry whose precursor count lies within the interval (a, b).

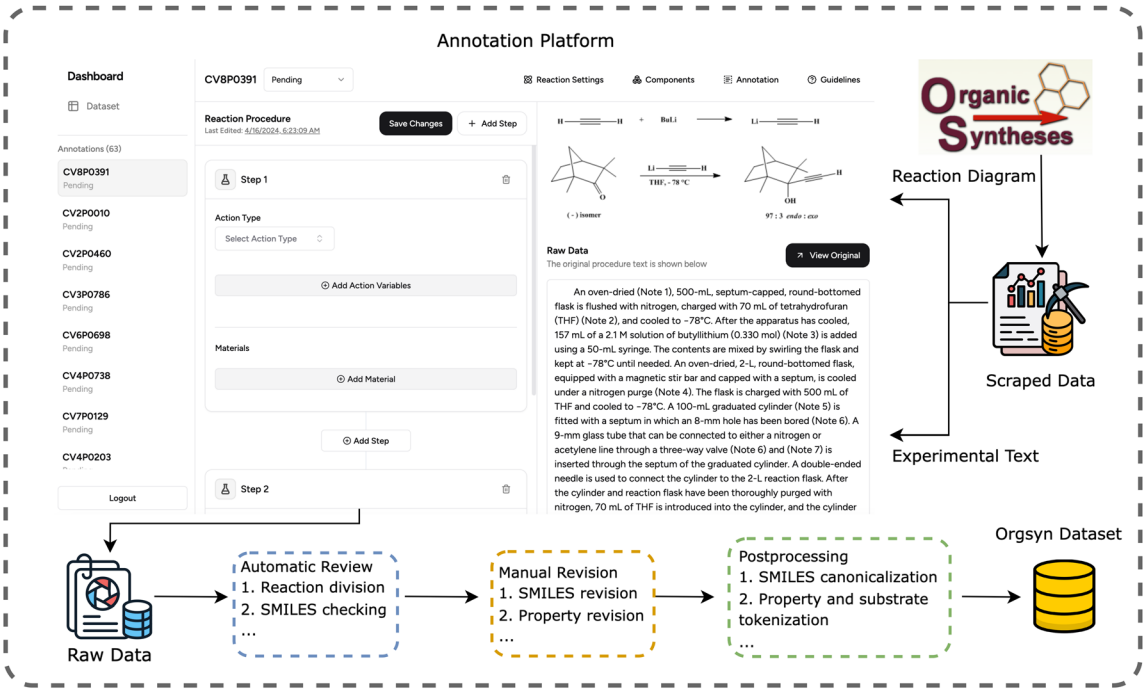


Fig. 3 Workflow of curating the Orgsyn dataset: data entries, each comprising experimental texts and reaction diagrams, are firstly scraped from the *Organic Synthesis* journal website. Annotators with a chemistry background then label these entries *via* a custom annotation platform according to the given guideline. After annotation, entries undergo automatic validation, manual expert revision, and post-processing to yield the dataset.

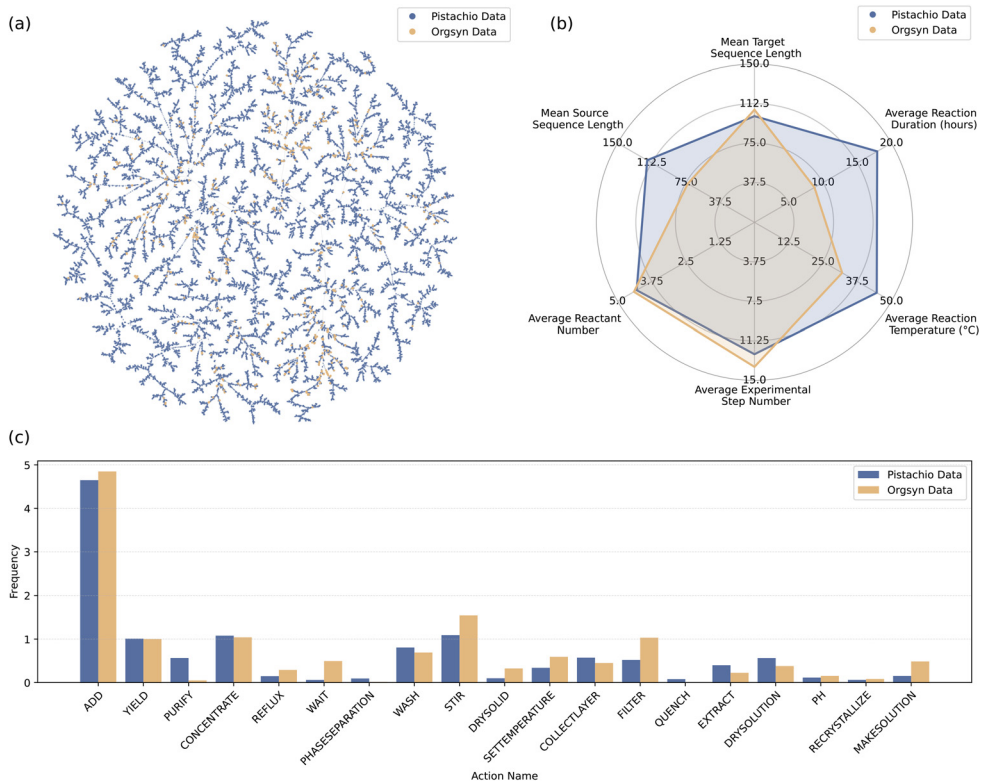


Fig. 4 Comparative dataset analysis of the Pistachio and Orgsyn datasets. (a) Reaction classes clustering atlas for the Pistachio and Orgsyn datasets. (b) Radar chart illustrating six dataset metrics for the Pistachio and Orgsyn datasets. (c) Average action usage frequencies across target procedure sequences of the Pistachio and Orgsyn datasets.

3.3.3 Dataset analysis. We characterize the Pistachio and Orgsyn datasets to elucidate their structural differences and implications for experimental procedure prediction. For source reaction data, we compute reaction fingerprints using RxnFP[54] and perform hierarchical clustering *via* *k*-nearest neighbors (kNN, scikit-learn[55]) with visualization through TMAP[56] (Fig. 4a). The Pistachio dataset exhibits broad spatial coverage across the chemical reaction space, providing diverse reaction patterns that facilitate model generalization across reaction classes. In contrast, the Orgsyn dataset occupies a restricted region due to its smaller scale; notably, Orgsyn reactions cluster proximal to Pistachio reactions, demonstrating substantial chemical space overlap.

For the target procedure sequence data, we conduct an analysis of action usage frequencies and sequence characteristics. The average action frequencies (Fig. 4c) indicate substantial disparities across datasets. While actions such as ADD, YIELD, CONCENTRATION, and RECRYSTALLIZE display comparable prevalence, others—most notably STIR and MAKESOLUTION—exhibit dataset-dependent variation. In addition, we consider multiple descriptive features, including target and source sequence lengths, the number of reactants, the number of experimental steps, reaction temperature, and reaction duration, and compare their mean values across datasets (Fig. 4b). The Orgsyn dataset closely aligns with Pistachio in terms of mean target sequence length, average reactant count, and average number of procedural steps, yet it presents a shorter mean source sequence length alongside markedly lower mean reaction temperature and duration. The comparative distributions of these features are further illustrated in Fig. S6 and Table S17. Here, the Orgsyn dataset and the Pistachio dataset exhibit similar distributions for target sequence length, reactant count, and step number, but the Orgsyn dataset demonstrates lower central tendencies for source sequence length, reaction temperature, and reaction duration. Collectively, these distinctions highlight a distributional gap between the two datasets at both the experimental procedure and reaction property levels. Such discrepancies likely arise from variations in reaction design, data extraction methodologies, and recording practices. Recognizing these differences supports the treatment of the datasets as distinct distributions requiring independent model calibration.

3.3.4 Data augmentation: reaction precursor order shuffling. Currently, all datasets for experimental procedure prediction contain one entry per reaction, yet reaction SMILES can vary by permuting precursor and product orders. Additionally, as specific tokens used in Vaucher *et al.*'s work[29] to represent the positions of precursors and products in reaction SMILES, Smiles2Actions can incorrectly predict the order of precursor addition, which is crucial in chemical experiments. To mitigate these limitations, we design a reaction precursor order shuffling algorithm that randomly shuffles the precursor SMILES for each reaction (Fig. 2a). This algorithm accepts the reaction SMILES and a shuffling configuration as inputs. As presented in Table 1, the shuffling configuration is structured as a mapping where the key represents the number range of precursors in the reaction, and the associated value indicates the number of new entries to be generated. For each reaction, the algorithm first counts the number of precursors, determines the number of new entries according to the given shuffling configuration, and then performs the shuffling to generate new entries with varied precursor orders. This augmentation method increases the dataset size and diversity, thereby enhancing the model's generalizability. For the Pistachio dataset, we apply three different shuffling configurations to generate the augmented datasets Pistachio_Aug_1, Pistachio_Aug_2, and Pistachio_Aug_3, containing 1 579 236(3×), 2 059 396(4×), and 3 028 790(6×) training entries, respectively. Given the small size of the Orgsyn dataset, we augment it with the same shuffling scheme as Pistachio_Aug_3 to generate the Orgsyn_Aug dataset, which contains 4045 training entries. All augmented datasets share the same validation and test datasets as their original counterparts during the training and evaluation process.

Utilizing the T5tokenizer function in the transformers library,[57] we convert source and target sequences of the Pistachio datasets into token lists and analyze data length distribution by regarding the length of token lists as the length of data sequences. We examine the data length distributions across the four Pistachio datasets in Fig. 5. The figure reveals that the proportion of longer sequences rises as the frequency of shuffling configuration increases, contributing to a more uniform distribution compared to the original Pistachio dataset. This more balanced distribution helps reduce the model's susceptibility to overfitting on narrow length ranges and promotes more stable training dynamics, ultimately enhancing its ability to generalize across diverse input scenarios.

3.4 Model

Predicting experimental procedures requires translating chemical information into natural language, making language models particularly suited for this challenge due to their proficiency in natural language processing. We develop ProcedureT5 based on the text-to-text transfer transformer (T5),[58] which is an encoder-decoder model from the transformer family and designed to reformulate all language tasks into a unified text-to-text format. This design allows T5 to efficiently tackle diverse tasks across multiple domains. Compared to the original transformer architecture,[59] T5 introduces two key modifications: (1) it eliminates the bias term in layer normalization and places these layers outside the residual connections, and (2) it replaces sinusoidal positional encodings with relative positional embeddings. Moreover, numerous studies have shown that pre-training models on domain-relevant tasks significantly improves their performance in downstream applications. To take advantage of this, we use pre-trained models such as Text+Chem T5

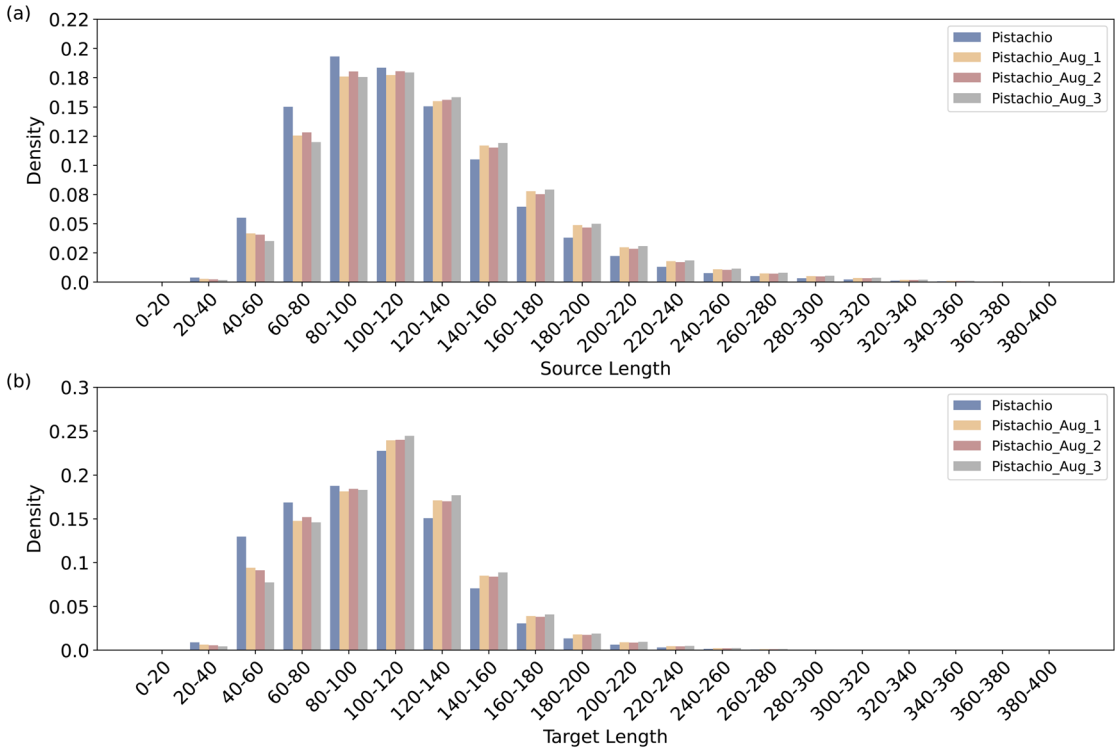


Fig. 5 Data length distributions of augmented Pistachio datasets. (a) Source sequence length distributions, wherein higher shuffling frequencies yield increasing proportions of longer sequences and more uniform profiles. (b) Target sequence length distributions, which exhibit analogous broadening and balancing effects with elevated shuffling frequency. These distributional shifts mitigate overfitting to narrow length ranges, thereby enhancing training stability and model generalization.

(ref. 60) and MolT5,[61] as foundation models for this work. Text+Chem T5 is pre-trained on four chemical task types that involve transformations between molecular representations and text, including molecule-to-molecule (mol2mol), molecule-to-text (mol2text), text-to-molecule (text2mol), and text-to-text (text2text). Meanwhile, MolT5 is pre-trained on a large corpus of unlabeled natural language text paired with molecular string representations. For computational efficiency, we adopt the small and base variants of T5 in this study. After training on the Pistachio dataset, we select the best-performing T5 model as ProcedureT5.

3.5 Evaluation metrics

We use the BLEU (bilingual evaluation understudy) score,[62] the ROUGE-L (recall-oriented understudy for gisting evaluation-longest common subsequence) score, and prediction accuracy as the primary metrics for evaluating model performance. The BLEU score measures how closely a candidate string matches a reference string by comparing n-gram overlap and is commonly used to assess machine translation models. It is formally defined as:

$$\mathrm{BLEU} = \mathrm{BP}\cdot \exp\left(\sum_{n}^{N} w_n \log p_n\right) \tag{1}$$

where N denotes the maximum n-gram level considered, which is set to 4 in this work, w_n are the weights assigned to each n-gram level (uniformly set to 0.25 for balanced contribution), p_n is the modified precision for n-grams of size n, and BP represents the brevity penalty, which is computed as follows:

$$\mathrm{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \le r \end{cases} \tag{2}$$

where c denotes the total number of words in the candidate translation corpus, and r denotes the sum of the effective reference lengths (the closest reference length for each candidate sentence) across the entire corpus.

The ROUGE-L score[63] examines the similarity between candidate and reference texts by evaluating the longest common subsequence (LCS), which represents the longest sequence of words appearing in both texts in the same order. It is formally computed using precision and recall based on the LCS length:

$$\text{ROUGE-L} = \frac{\left(1+\beta^2\right)\cdot \mathrm{Precision}_{\mathrm{LCS}}\cdot \mathrm{Recall}_{\mathrm{LCS}}}{\beta^2\cdot \mathrm{Precision}_{\mathrm{LCS}}\cdot \mathrm{Recall}_{\mathrm{LCS}}} \tag{3}$$

where $\mathrm{Precision}_{\mathrm{LCS}}$ denotes the proportion of LCS length to the candidate text while $\mathrm{Recall}_{\mathrm{LCS}}$ represents the proportion of LCS length to the reference text, and β (set to 1) controls the balance between precision and recall.

The prediction accuracy is calculated based on the Levenshtein similarity,[29] which is obtained by subtracting the

Levenshtein distance[64,65] between candidate and reference texts from 1. We report accuracy at thresholds of 100%, 90%, 75%, and 50%, representing the proportion of predictions with Levenshtein similarity exceeding these values. For the Orgsyn dataset, we adopt averaged similarity as the evaluation metric in place of the 100% accuracy and 90% accuracy, owing to the absence of predictions that perfectly match the ground truth in the test set. These metrics evaluate the overall performance of models, but lack the ability to assess a model's ability in sub-prediction tasks. To more comprehensively evaluate model performance in experimental procedure prediction, and inspired by the work of Vaškevičius *et al.*,[31] we introduce four specialized metrics: workup action and solvent coverage (WASC), reaction temperature accuracy (RTA), reaction duration accuracy (RDA), and reaction pH accuracy (RPHA). WASC quantifies the model's ability to predict workup actions—such as extraction and purification—which are critical for product isolation and yield reproducibility, and thus warrant dedicated evaluation. Specifically, WASC is defined as the fraction of target workup actions for which the associated solvents are correctly matched in the predictions, thereby revealing isolation failures due to incomplete or imprecise workup procedures. RTA measures the accuracy of the predicted reaction temperature, defined as the maximum temperature of a STIR action, which governs reaction kinetics and selectivity; accurate prediction is essential to avoid decomposition or sluggish reaction progress. RTA is computed as the relative error between predicted and target Kelvin values derived from tokenized temperature ranges to capture thermally induced deviations that compromise experimental viability. RDA evaluates the model's performance in predicting reaction duration, defined as the cumulative time of STIR and REFLUX actions, which determines whether a reaction reaches completion or proceeds toward overreaction, necessitating precise forecasting for practical synthesis. RDA is calculated as the relative error on summed token-derived durations, providing a fair penalty for timing mismatches without bias toward short or long procedures. RPHA assesses the model's ability to forecast the reaction pH profile from pH-adjustment actions that modulate reactivity in sensitive steps, where omissions can disrupt acid/base catalysis. RPHA is defined as the overlap between predicted and target pH sets, providing a quantitative measure of the model's capacity to deliver actionable pH guidance for chemical reactions. More detailed definitions, calculation methods, and illustrative examples of these metrics are provided in Note S5.

3.6 Implementation details

We optimize the transformer model (SMILES2Actions) using the same settings as Vaucher *et al.*,[29] with the modifications of employing sentence-level batching without gradient accumulation and adjusting training steps to 1 million while conducting evaluation every 50 000 steps. Training is conducted on four NVIDIA RTX 5880 GPUs, each processing 64 batches per loop. For T5 models, following the training configuration of Text+Chem T5,[59] we employ the ADAM optimizer using a weight decay of 0.01, a maximum learning rate of 3×10^{-4}, and a decay rate of 0.99 applied after each epoch. Training on the Pistachio datasets utilizes eight NVIDIA RTX 3090 GPUs with a per-GPU batch size of 8 with gradient accumulation over two steps, yielding an effective total batch size of 128. The experiments on the Orgsyn datasets are conducted using two NVIDIA RTX PRO 6000 GPUs. For fine-tuning on the Orgsyn dataset, we employ a learning rate of 1×10^{-4} and a per-GPU batch size of 16 without gradient accumulation, given the relatively small dataset size. In contrast, fine-tuning on the Orgsyn_Aug dataset uses a learning rate of 3×10^{-4} and a per-GPU batch size of 32 to accommodate its larger data volume. We optimize T5 series models with the cross-entropy loss function, which is well-suited for sequence generation tasks. Checkpoints achieving the highest validation BLEU score, which serves as the primary metric for both validation and testing, are selected for assessment.

4 Results and discussion

4.1 Model performance on the pistachio dataset

In the first stage of the ProcedureT5 workflow, we perform experiments with the Pistachio dataset by taking the transformer model from SMILES2Actions as the baseline (Table 2). Initially, we assess the impact of the reaction precursor order shuffling algorithm by training the transformer model on multiple augmented variants of the Pistachio dataset. All augmented training datasets enhanced the model performance on the original Pistachio test set. Furthermore, a distinct trend emerges: models trained on datasets with higher-frequency shuffling configurations exhibited better performance, indicating that increased shuffling frequency enhances the model's generalization ability.

Next, we evaluate the impact of model architecture and chemistry-specific pre-training using base-sized T5 models. To maximize computational efficiency, all selected T5 models are trained solely on the original Pistachio dataset and the Pistachio_Aug_3 dataset, the latter of which yields the best performance for the transformer model compared to other Pistachio datasets. Remarkably, all base-size T5 models, including those trained only on the original Pistachio dataset, outperform the transformer model trained on Pistachio_Aug_3, achieving higher BLEU scores despite marginally lower ROUGE-L scores. This result highlights the architectural advantages of the T5 framework, particularly its modified residual connections and optimized positional encoding mechanism. Furthermore, when trained on the same datasets, both MolT5 and Text+Chem T5 outperform the standard T5 model, further emphasizing the significant role of domain-specific pre-training. Model performance on experimental procedure sub-prediction tasks is evaluated using four metrics: workup action and solvent coverage

Table 2 Model performance on the Pistachio dataset. For each metric, bold denotes the best result, and underline denotes the second best

Model type	Size	Dataset	BLEU score	ROUGE-L score	100% accuracy	90% accuracy	75% accuracy	50% accuracy
SMILES2Actions	—	Pistachio	54.70	67.86	3.60	10.10	24.74	68.73
SMILES2Actions	—	Pistachio_Aug_1	55.76	68.28	3.87	10.82	25.49	69.99
SMILES2Actions	—	Pistachio_Aug_2	55.71	**68.48**	4.29	*11.27*	**26.11**	69.72
SMILES2Actions	—	Pistachio_Aug_3	55.90	68.43	3.76	10.70	25.62	*70.28*
T5	Base	Pistachio	56.45	66.57	4.36	10.08	22.04	67.30
T5	Base	Pistachio_Aug_3	57.33	67.58	4.46	10.70	23.92	69.04
MolT5	Base	Pistachio	57.34	67.30	*4.63*	10.65	23.73	68.70
MolT5	Base	Pistachio_Aug_3	**58.46**	*68.46*	**4.82**	**11.74**	*25.89*	**70.81**
Text+Chem T5	Base	Pistachio	56.92	67.18	4.50	10.39	23.39	68.57
Text+Chem T5	Base	Pistachio_Aug_3	*57.48*	67.90	4.61	11.26	24.91	69.98

(WASC), reaction temperature accuracy (RTA), reaction duration accuracy (RDA), and reaction pH accuracy (RPHA); detailed results are provided in Note S6. Across all shuffling configurations, Transformer models exhibit significant improvement in WASC scores with data augmentation, whereas RTA, RDA, and RPHA remain relatively unchanged. All models achieve comparable RTA performance (approximately 95.60%), while the Transformer models outperform their T5-based counterparts on both RDA and RPHA metrics. These findings suggest that the proposed data augmentation strategy effectively enhances model performance in workup action prediction, whereas advances in model architecture contribute only marginally to improvements in these subtasks.

To further investigate the impact of model parameter size, we compare the performance of MolT5-base with MolT5-small, as well as Text+Chem T5-base with Text+Chem T5-small. As shown in Table 3, the base-sized models consistently outperform their smaller counterparts, highlighting the advantage of a larger parameter capacity in capturing complex reaction features. Among the base-sized T5 variants, MolT5-base trained on the Pistachio_Aug_3 dataset delivers the best performance, showing an approximate 4-point improvement in BLEU score and a notable increase in 100% similarity, rising from 3.60% to 4.82% compared to the baseline transformer model. By combining the results from Tables 2 and 3, we conclude that scaling both model size and training data volume leads to performance improvements for ProcedureT5, indicating the potential for further optimization.

To illustrate model performance and evaluate the impact of the reaction precursor order shuffling augmentation algorithm, we select a representative reaction and generate two additional shuffled variants as shown in Fig. 6. The original target reaction involves the sequential addition of 3-(5-bromo-2-pyridinyl)-5-(chloromethyl)-4,5-dihydro-1,2-oxazole (3), 2-(4-pyridinyl)ethylamine (4), tetrabutylammonium iodide (1), and methyl isothiocyanate (2), followed by stirring at 100 °C (#6#) for 1 day (@4@). After adding water (5), the mixture is extracted with ethyl acetate, and the organic layer is collected, dried over sodium sulfate, concentrated, and purified to afford *N*-[[3-(5-bromo-2-pyridinyl)-4,5-dihydro-1,2-oxazol-5-yl]methyl]-2-pyridin-4-ylethanamine. Experimental procedures for the original reaction (configuration 1) and two shuffled variants (configurations 2 and 3) are predicted using two MolT5-base model variants (Table 2): MolT5-Original, trained on the original Pistachio dataset, and MolT5-Aug-3, trained on the Pistachio_Aug_3 dataset with shuffling augmentation. As illustrated in Fig. 6, MolT5-Aug-3 consistently outperforms MolT5-Original across all configurations. For precursor addition order, MolT5-Aug-3 achieves near-perfect accuracy, incorrectly reversing only the order of tetrabutylammonium iodide (1) and methyl isothiocyanate (2) in configuration 1, whereas MolT5-Original misclassifies precursor sequences in all configurations. Regarding the remaining experimental procedure steps, MolT5-Aug-3 correctly predicts all operations across configurations except for incorrectly predicting the duration of STIR action, while MolT5-Original introduces spurious actions, including MICROWAVE and FILTER. Additional representative examples are provided in Note S4. These results demonstrate that the reaction precursor order shuffling augmentation substantially improves model robustness and generalization. By increasing training data diversity through systematic reshuffling of reactant sequences, the augmented model effectively learns to predict

Table 3 Performance of chemistry-specific T5 variants on the Pistachio dataset. For each metric, bold denotes the best result, and italics denotes the second best

Model type	Size	Dataset	BLEU score	ROUGE-L score	100% accuracy	90% accuracy	75% accuracy	50% accuracy
MolT5	Small	Pistachio_Aug_3	57.31	67.28	2.32	8.10	22.85	69.34
MolT5	Base	Pistachio_Aug_3	**58.46**	**68.46**	**4.82**	**11.74**	**25.89**	**70.81**
Text+Chem T5	Small	Pistachio_Aug_3	56.98	66.92	1.85	6.92	21.07	69.55
Text+Chem T5	Base	Pistachio_Aug_3	*57.48*	*67.90*	*4.61*	*11.26*	*24.91*	*69.98*

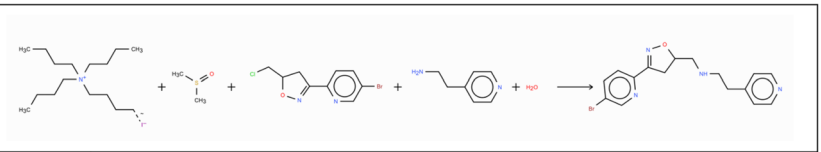
Shuffling Setting \ Reaction Diagram				
1	Source	CCCC[N+](CCCC)(CCCC)CCCC~[I-] .CS(C)=O.ClCC1CC(c2ccc(Br)cn2)=NO1.NCCc1ccncc1.O >>Brc1ccc(C2=NOC(CNCCc3ccncc3)C2)nc1		
	Target	ADD 3 ; ADD 4 ; ADD 1 ; ADD 2 ; STIR for @4@ at #6# ; ADD 5 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over Na2SO4 ; CONCENTRATE ; PURIFY ; YIELD -1		
	MolT5-Original	ADD 3 ; ADD 2 ; ADD 4 ; ADD 1 ; MICROWAVE for @2@ at #6# ; ADD 5 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; CONCENTRATE ; PURIFY ; COLLECTLAYER organic ; CONCENTRATE ; YIELD -1	Similarity	72.96%
	MolT5-Aug-3	ADD 3 ; ADD 4 ; ADD 2 ; ADD 1 ; STIR for @2@ at #6# ; ADD 5 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; CONCENTRATE ; PURIFY ; YIELD -1	Similarity	91.10%
2	Source	NCCc1ccncc1.O. CCCC[N+](CCCC)(CCCC)CCCC~[I-] .CS(C)=O.ClCC1CC(c2ccc(Br)cn2)=NO1 >>Brc1ccc(C2=NOC(CNCCc3ccncc3)C2)nc1		
	Target	ADD 3 ; ADD 1 ; ADD 5 ; ADD 4 ; STIR for @4@ at #6# ; ADD 2 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over Na2SO4 ; CONCENTRATE ; PURIFY ; YIELD -1		
	MolT5-Original	ADD 3 ; ADD 4 ; ADD 1 ; ADD 5 ; MICROWAVE for @2@ at #6# ; ADD 2 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; CONCENTRATE ; PURIFY ; COLLECTLAYER organic ; CONCENTRATE ; YIELD -1	Similarity	72.96%
	MolT5-Aug-3	ADD 3 ; ADD 1 ; ADD 5 ; ADD 4 ; STIR for @3@ at #6# ; ADD 2 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate; CONCENTRATE ; PURIFY ; YIELD -1	Similarity	92.15%
3	Source	CS(C)=O.ClCC1CC(c2ccc(Br)cn2)=NO1.CCCC[N+](CCCC)(CCCC)CCCC~[I-].O.NCCc1ccncc1 >>Brc1ccc(C2=NOC(CNCCc3ccncc3)C2)nc1		
	Target	ADD 2 ; ADD 5 ; ADD 3 ; ADD 1 ; STIR for @4@ at #6# ; ADD 4 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over Na2SO4 ; CONCENTRATE ; PURIFY ; YIELD -1		
	MolT5-Original	ADD 2 ; ADD 1 ; ADD 5 ; ADD 3 ; STIR for @4@ at #6# ; ADD 4 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; FILTER keep filtrate ; CONCENTRATE ; PURIFY ; COLLECTLAYER organic ; CONCENTRATE ; YIELD -1	Similarity	69.72%
	MolT5-Aug-3	ADD 2 ; ADD 5 ; ADD 3 ; ADD 1 ; STIR for @3@ at #6# ; ADD 4 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; CONCENTRATE ; PURIFY ; YIELD -1	Similarity	92.15%

Fig. 6 Example prediction results on the Pistachio dataset. The target reaction involves sequential addition of 3-(5-bromo-2-pyridinyl)-5-(chloromethyl)-4,5-dihydro-1,2-oxazole, 2-(4-pyridinyl)ethylamine, tetrabutylammonium iodide, and methyl isothiocyanate, followed by stirring at 100 °C for 1 day, aqueous workup with water, extraction with ethyl acetate, drying over sodium sulfate, concentration, and purification to yield *N*-[[3-(5-bromo-2-pyridinyl)-4,5-dihydro-1,2-oxazol-5-yl]methyl]-2-pyridin-4-ylethanamine.

correct precursor addition orders and procedural operations across varied input configurations, thereby reducing both sequence and action prediction errors.

4.2 Model performance on the Orgsyn dataset

We designate the MolT5-base model fine-tuned on the Pistachio_Aug_3 dataset as the best-performing baseline and refer to this model as ProcedureT5 throughout this study. To evaluate generalization across data sources, we finetune both ProcedureT5 and MolT5-base on the Orgsyn and Orgsyn_Aug datasets using 10 random seeds for each configuration. We report mean primary performance metrics and standard deviations across all runs in Table 4. The results demonstrate a significant decline in model performance across all architectures when applied to the Orgsyn series datasets relative to the Pistachio series datasets. This finding confirms that inter-dataset heterogeneity negatively impacts model generalization in predicting experimental procedures across distinct data sources. A further comparison of model performance on sub-prediction tasks across datasets helps elucidate the source of this decline. For ProcedureT5, the WASC metric on the Orgsyn dataset decreases by about 80% compared to that on the Pistachio dataset, while the RTA metric remains largely unchanged, and the RDA metric drops by approximately 35%. These results indicate that the reduction in overall model performance is not primarily driven by diminished reaction property prediction accuracy; rather, inaccuracies in predicting procedural actions likely play a more substantial role. Combined with the earlier dataset difference analysis, these observations suggest that the primary causes of performance degradation are the disparities in reaction distribution and variations in experimental action usage between the two datasets. Notably, models fine-tuned on Orgsyn_Aug consistently outperform those trained on the original Orgsyn dataset, particularly in BLEU score, ROUGE-L score, and average similarity. This improvement demonstrates the effectiveness of our reaction

Table 4 Model performance on the Orgsyn dataset. For each metric, bold denotes the best result, and italics denotes the second best

Model type	Size	Dataset	BLEU score	ROUGE-L score	Averaged similarity	75% / 50% accuracy
MolT5	Base	Orgsyn	33.50 ± 0.85	47.74 ± 1.49	44.73 ± 1.23	0.94 ± 0.61/28.99 ± 3.42
MolT5	Base	Orgsyn_Aug	34.46 ± 2.88	47.68 ± 1.06	45.24 ± 0.88	1.41 ± 0.76/28.59 ± 4.86
ProcedureT5	Base	Orgsyn	*39.88 ± 0.65*	*53.40 ± 0.61*	*49.65 ± 0.59*	**3.63 ± 0.62**/*43.89 ± 2.33*
ProcedureT5	Base	Orgsyn_Aug	**40.34 ± 0.61**	**53.47 ± 0.42**	**49.72 ± 0.59**	*2.48 ± 1.04*/**45.37 ± 2.17**

precursor order shuffling algorithm as a data augmentation strategy. Furthermore, ProcedureT5 consistently outperforms MolT5-base under equivalent fine-tuning conditions, underscoring the importance of pre-training on task-relevant datasets prior to downstream fine-tuning.

For the experimental procedure sub-prediction tasks, model performance exhibits similar declines relative to the primary evaluation metrics, except for the RTA score. The ProcedureT5 model trained on Orgsyn_Aug attains the highest WASC score of 10.43%, demonstrating its advantage over the MolT5-base model. Moreover, both models trained on the augmented dataset achieve substantially higher WASC scores than their counterparts trained on the original Orgsyn dataset, further confirming the effectiveness of the reaction precursor shuffling algorithm in enhancing model performance on the WASC metric.

Fig. 7 presents two prediction examples generated by models fine-tuned on the Orgsyn_Aug dataset. In the first case, ethyl 4-nitrobenzoate (1) is added to ethanol (2), followed by platinum dioxide (3). Hydrogen gas (4) is introduced to the reaction mixture for 10 minutes (@1@), then the mixture is stirred. After filtering to remove the catalyst, the filtrate is concentrated under reduced pressure, and the residue is recrystallized from diethyl ether to yield benzocaine. MolT5-base exhibits significant deficiencies by inaccurately predicting the precursor addition sequence and employing incorrect experimental procedure actions, resulting in low similarity to the target procedure. Conversely, ProcedureT5 accurately identifies the precursor addition order, yielding higher similarity scores, but omits the RECRYSTALLIZE action. In the second case, sodium (3) is introduced into ethanol (4), followed by sequential addition of diethyl malonate (1) and 1-bromo-3-chloropropane (2). The reaction mixture is stirred at 100 °C (#6#) for 1 hour (@2@) and subsequently heated under reflux for an additional hour (@2@). Upon concentration and aqueous workup, the organic layer is separated, washed with brine, dried over sodium sulfate, filtered, and concentrated to afford diethyl cyclobutane-1,1-dicarboxylate. MolT5 exhibits substantial errors in both precursor addition order and other experimental steps, resulting in low similarity scores. ProcedureT5 nearly perfectly predicts the precursor addition order and subsequent steps, with minor errors limited to misclassifying the sodium–ethanol addition sequence and omitting the FILTER action after DRYSOLUTION, thereby achieving a superior similarity score.

(a)

A solution of 19.5 g. of ethyl p-nitrobenzoate in 150 cc. of 95 percent alcohol is placed in the reaction bottle of the catalytic reduction apparatus, and 0.2 g. of platinum oxide catalyst is added. The mixture is shaken with hydrogen until three molecular equivalents have been absorbed. The time required is about seven minutes. The platinum is filtered off, and the alcohol is removed from the filtrate by tillation. The ethyl p-aminobenzoate, recrystallized from about 40 cc. of ether, melts at 89 °C. The yield is 15.5 g.

Source	CCOC(=O)C1=CC=C([N+](=O)[O-])C=C1.CCO.O=[Pt]=O.[HH]>>CCOC(=O)C1=CC=C(N)C=C1		
Target	ADD 1 ; ADD 2 ; ADD 3 ; ADD 4 over @1@ ; STIR ; FILTER keep filtrate ; CONCENTRATE ; RECRYSTALLIZE from ether ; YIELD -1		
MolT5-base	ADD 1 ; ADD 4 ; ADD 3 ; ADD water ; STIR ; FILTER keep precipitate ; WASH with water ; ADD 2 ; ADD 3 ; FILTER keep filtrate ; YIELD -1	Similarity	54.11%
ProcedureT5	ADD 1 ; ADD 2 ; ADD 3 ; ADD 4 ; STIR for @4@ at #5# ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1	Similarity	66.92%

(b)

A solution of sodium ethoxide is prepared by adding 138 g. of fresh-cut sodium in small pieces to 2.5 l. of absolute ethanol in a 5-l. round-bottomed flask fitted with an efficient reflux condenser capped with a calcium chloride drying tube. In a three-necked 5-l. round-bottomed flask, equipped with a reflux condenser capped with a calcium chloride tube, a rubber-sealed mechanical stirrer, and an inlet tube for the addition of the sodium ethoxide solution, are mixed with 480 g of diethyl malonate and 472 g. of trimethylene chlorobromide. The mixture is heated to 80° and vigorously stirred while the sodium ethoxide solution is slowly forced into the flask using dry air pressure. The rate of addition is regulated so that the reaction mixture refluxes smoothly. After the addition is on is complete (this requires about 1.5 hours), the mixture is refluxed, with continued stirring, for an additional 45 minutes. Upon completion of the reflux period, the alcohol is removed by distillation, 90–95% of the alcohol being recovered. The reaction mixture is cooled, and 900 ml. of cold water is added. After the sodium halides completely dissolve, the aqueous layer is extracted with three 500-ml portions of ether. The organic layer and the ether extracts are combined, shaken with 50 ml. of saturated salt solution, and dried over 100 g. of anhydrous sodium sulfate. The solution is filtered, the ether is removed by distillation on a steam bath, and the residue, which weighs 600–625 g., is distilled through a short Vigreux column. The yield of product boiling at 91–96°/4 mm is 320–330 g.

Source	CCOC(=O)CC(=O)OCC.ClCCCBr.[Na].CCO>>CCOC(=O)C1(C(=O)OCC)CCC1		
Target	ADD 3 ; ADD 4 ; ADD 1 ; ADD 2 ; STIR for @2@ at #6# ; REFLUX for @2@ ; CONCENTRATE ; ADD water ; COLLECTLAYER organic ; WASH with brine ; DRYSOLUTION over sodium sulfate ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1		
MolT5-base	ADD 1 ; ADD 4 ; ADD 2 ; ADD water ; CONCENTRATE ; ADD water ; STIR ; FILTER keep precipitate ; WASH with water ; DRYSOLID ; EXTRACT with benzene ; ADD 3 ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1	Similarity	55.07%
ProcedureT5	ADD 4 ; ADD 3 ; ADD 1 ; STIR ; ADD 2 over @2@ ; REFLUX for @3@ ; CONCENTRATE ; ADD water ; COLLECTLAYER organic ; WASH with brine ; DRYSOLUTION over sodium sulfate ; CONCENTRATE ; YIELD -1	Similarity	78.41%

Fig. 7 Example prediction results on the Orgsyn dataset. (a) Reduction of ethyl 4-nitrobenzoate in ethanol with platinum dioxide under hydrogen for 10 min, followed by stirring, catalyst filtration, concentration under reduced pressure, and recrystallization from diethyl ether to yield benzocaine. (b) Sodium in ethanol, treated sequentially with diethyl malonate and 1-bromo-3-chloropropane, stirred at 100 °C for 1 h, then refluxed for 1 h, concentrated, subjected to aqueous workup, washed with brine, dried over sodium sulfate, filtered, and concentrated to afford diethyl cyclobutane-1,1-dicarboxylate.

Overall, while the performance of ProcedureT5 on the Orgsyn dataset is lower than that on the Pistachio dataset, it still offers valuable insights that can assist chemists in developing new experimental procedures.

5 Future work

Despite these advances, several open challenges persist in this research domain, particularly concerning the establishment of standardized data formats, the exploration of performance scaling limits for language models, the development of task-specific architectures, and the identification of application-driven use cases. Regarding data formats, although the quantity and quality of available reaction procedure datasets continue to improve, the inherent complexity of chemical experimental texts produces heterogeneous extraction and post-processing methodologies. This variability hinders the integration of multi-source datasets for unified model training and evaluation. The emergence of large language models (LLMs) may offer potential solutions, leveraging their advanced text understanding capabilities to facilitate the development of standardized protocols for experimental procedure extraction. Accordingly, dataset quality and scale are expected to co-evolve with advancements in LLMs. With respect to performance scaling, although the current results indicate substantial room for improvement, computational resource constraints in this study limit the exploration of larger model variants (*e.g.*, MolT5-large) and more extensive data augmentation schemes, which should be systematically investigated in future work. Regarding model architectures, most existing approaches remain grounded in natural language processing frameworks, which often struggle to effectively capture structural information embedded in chemical reaction SMILES representations. Emerging paradigms such as graph neural networks and diffusion-based models demonstrate greater potential in modeling reaction dynamics and thus represent promising directions for future research. Beyond these technical considerations, practical adoption remains underexplored; integrating procedure prediction models with upstream and downstream components—such as target molecule design, synthesis planning, and automated robotic validation—constitutes a crucial next step toward realizing a fully integrated molecular synthesis ecosystem.

6 Conclusions

This study presents ProcedureT5, a training framework that integrates chemistry-oriented pre-trained models with data augmentation strategies for adaptive experimental procedure prediction. For model development, multiple chemistry-specialized pre-trained T5 variants are trained and assessed on augmented datasets, with the highest-performing model selected as the foundation for ProcedureT5. The results demonstrate that task-specific pre-training on chemical corpora, combined with reaction precursor shuffling augmentation, substantially enhances the model's generalization capability. Consequently, ProcedureT5 achieves a 4.0-point improvement in BLEU score and a 34% increase in exact-match accuracy on the Pistachio dataset relative to the baseline. Furthermore, the analyses indicate that additional performance gains may be attainable by scaling both model capacity and training dataset size. On the dataset side, this work contributes the Orgsyn dataset—a high-quality, expert-annotated dataset designed to facilitate model fine-tuning and comprehensive evaluation across diverse scenarios. Although the model demonstrates lower performance on the Orgsyn dataset compared with Pistachio, reflecting intrinsic differences between the two datasets, ProcedureT5 consistently surpasses baseline systems. These innovations collectively address long-standing limitations in prior research, including suboptimal model generalization and the absence of rigorously curated benchmark datasets.

Author contributions

Yuxuan Zhang formal analysis, investigation, methodology, writing – original draft, writing – review & editing; Yue Fang writing – original draft, writing – review & editing; Haifan Zhou writing – review & editing; Bowen Yu writing – review & editing; Tsz Fung Fung annotation platform development; Qing Liu writing – review & editing; Christophe Len project administration, writing – original draft, writing – review & editing; Hanyu Gao conceptualization, funding acquisition, investigation, methodology, project administration, writing – original draft, writing – review & editing.

Conflicts of interest

The authors declare no competing financial interest.

Data availability

The source code and the Orgsyn dataset associated with this work are available at https://github.com/yuxuanzhang-1024/ProcedureT5. The source code for the data annotation platform is accessible at https://github.com/yuxuanzhang-1024/data_platform_annotation.

Supplementary information (SI), including details of the dataset annotation project, examples of model predictions, definitions of the evaluation metrics, and the corresponding model performance results, is available. See DOI: https://doi.org/10.1039/d5re00572h.

Acknowledgements

This research is supported by Grant Z1269 from The Hong Kong University of Science and Technology and The Research Group of Hanyu Gao, Department of Chemical and Biological Engineering, The Hong Kong University of Science

and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, P. R. China.

References

1 Y. Jiang, Y. Yu, M. Kong, Y. Mei, L. Yuan, Z. Huang, K. Kuang, Z. Wang, H. Yao and J. Zou, *et al.*, *Engineering*, 2023, **25**, 32–50.
2 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.
3 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.
4 C. Yan, Q. Ding, P. Zhao, S. Zheng, J. Yang, Y. Yu and J. Huang, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 11248–11258.
5 M. E. Fortunato, C. W. Coley, B. C. Barnes and K. F. Jensen, *J. Chem. Inf. Model.*, 2020, **60**, 3398–3407.
6 T. Liu, Z. Cao, Y. Huang, Y. Wan, J. Wu, C.-Y. Hsieh, T. Hou and Y. Kang, *JACS Au*, 2023, **3**, 3446–3461.
7 W. Zhong, Z. Yang and C. Y.-C. Chen, *Nat. Commun.*, 2023, **14**, 3009.
8 C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers and H. Gao, *et al.*, *Science*, 2019, **365**, eaax1566.
9 J. Roh, J. F. Joung, K. Yu, Z. Tu, G. L. Bartholomew, O. A. Santiago-Reyes, M. H. Fong, R. Sarpong, S. E. Reisman and C. W. Coley, *ACS Cent. Sci.*, 2026, **12**, 345–357.
10 A. K. Hassen, M. Šícho, Y. J. van Aalst, M. C. Huizenga, D. N. Reynolds, S. Luukkonen, A. Bernatavicius, D.-A. Clevert, A. P. Janssen and G. J. van Westen, *et al.*, *J. Cheminf.*, 2025, **17**, 41.
11 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2018, **4**, 1465–1476.
12 Y. Kwon, S. Kim, Y.-S. Choi and S. Kang, *J. Chem. Inf. Model.*, 2022, **62**, 5952–5960.
13 E. Shim, J. A. Kammeraad, Z. Xu, A. Tewari, T. Cernak and P. M. Zimmerman, *Chem. Sci.*, 2022, **13**, 6655–6668.
14 M. R. Maser, A. Y. Cui, S. Ryou, T. J. DeLano, Y. Yue and S. E. Reisman, *J. Chem. Inf. Model.*, 2021, **61**, 156–166.
15 K. Chen, J. Lu, J. Li, X. Yang, Y. Du, K. Wang, Q. Shi, J. Yu, L. Li and J. Qiu, *et al.*, *arXiv*, 2023, preprint, arXiv:2311.10776, DOI: 10.48550/arXiv.2311.10776.
16 Z. Wang, K. Lin, J. Pei and L. Lai, *Chem. Sci.*, 2025, **16**, 854–866.
17 L.-Y. Chen and Y.-P. Li, *J. Cheminf.*, 2024, **16**, 11.
18 P.-X. Hua, Z. Huang, Z.-Y. Xu, Q. Zhao, C.-Y. Ye, Y.-F. Wang, Y.-H. Xu, Y. Fu and H. Ding, *Commun. Chem.*, 2025, **8**, 42.
19 R. Shi, G. Yu, X. Huo and Y. Yang, *J. Cheminf.*, 2024, **16**, 22.
20 Y. Kwon, D. Lee, Y.-S. Choi and S. Kang, *J. Cheminf.*, 2022, **14**, 2.
21 S.-W. Li, L.-C. Xu, C. Zhang, S.-Q. Zhang and X. Hong, *Nat. Commun.*, 2023, **14**, 3569.
22 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
23 A. L. Haywood, J. Redshaw, M. W. Hanson-Heine, A. Taylor, A. Brown, A. M. Mason, T. Gärtner and J. D. Hirst, *J. Chem. Inf. Model.*, 2021, **62**, 2077–2092.
24 A. Subramanian, W. Gao, R. Barzilay, J. C. Grossman, T. Jaakkola, S. Jegelka, M. Li, J. Li, W. Matusik and E. Olivetti, *et al.*, *An MIT Exploration of Generative AI*, 2024.
25 S.-Q. Zhang, L.-C. Xu, S.-W. Li, J. C. Oliveira, X. Li, L. Ackermann and X. Hong, *Chem. – Eur. J.*, 2023, **29**, e202202834.
26 J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Muller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816–9872.
27 A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller and T. Laino, *Nat. Commun.*, 2020, **11**, 3601.
28 NextMove Software, *NextMove Software*, https://www.nextmovesoftware.com/.
29 A. C. Vaucher, P. Schwaller, J. Geluykens, V. H. Nair, A. Iuliano and T. Laino, *Nat. Commun.*, 2021, **12**, 2573.
30 Z. Liu, Y. Shi, A. Zhang, S. Li, E. Zhang, X. Wang, K. Kawaguchi and T.-S. Chua, *arXiv*, 2024, preprint, arXiv:2405.14225, DOI: 10.48550/arXiv.2405.14225.
31 M. Vaškevičius and J. Kapočiūtė-Dzikienė, *Appl. Sci.*, 2024, **14**, 11526.
32 M. Vaškevičius, J. Kapočiūtė-Dzikienė, A. Vaškevičius and L. Šlepikas, *PeerJ Comput. Sci.*, 2023, **9**, e1511.
33 Organic Syntheses, Inc., *Org. Synth.*, https://www.orgsyn.org/.
34 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *Chem. Soc. Rev.*, 2025, **54**, 1125–1150.
35 Q. Ai, F. Meng, J. Shi, B. Pelkie and C. W. Coley, *Digital Discovery*, 2024, **3**, 1822–1831.
36 Y. Chen, C. T. Leung, J. Sun, Y. Huang, L. Li, H. Chen and H. Gao, *Chem. Sci.*, 2025, **16**, 21464–21474.
37 H. Zhou, Y. Fang, H. Ma, A. P. Roxas, P. Deng, G. Zhang, Y. Wang, Y. Zhang, W. Zhou and L. Li, *et al.*, *Chem. Eng. J.*, 2026, 172634.
38 Q. Wang, W. Zhang, M. Chen, X. Li, Z. Xiong, J. Xiong, Z. Fu and M. Zheng, *Chem. Sci.*, 2025, **16**, 11548–11558.
39 X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman and Y. Su, *npj Comput. Mater.*, 2025, **11**, 79.
40 C. R. Kelly and J. M. Cole, *Sci. Data*, 2025, **12**, 2000.
41 M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou and F. Wang, *Briefings Bioinf.*, 2020, **21**, 919–935.
42 A. Kensert, R. Bouwmeester, K. Efthymiadis, P. Van Broeck, G. Desmet and D. Cabooter, *Anal. Chem.*, 2021, **93**, 15633–15641.
43 X. Liu, C. Fan, Y. Liu and H.-b. Li, *J. Chem. Inf. Model.*, 2025, **65**, 9034–9048.
44 L.-Y. Chen and Y.-P. Li, *Beilstein J. Org. Chem.*, 2024, **20**, 2476–2492.
45 P.-C. Zhao, X.-X. Wei, Q. Wang, Q.-H. Wang, J.-N. Li, J. Shang, C. Lu and J.-Y. Shi, *Nat. Commun.*, 2025, **16**, 814.
46 Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M. Wu, T. Hou and M. Song, *Chem. Sci.*, 2022, **13**, 9023–9034.
47 S. Zhang, H. Li, L. Chen, Z. Zhao, X. Lin, Z. Zhu, B. Chen, X. Chen and K. Yu, *arXiv*, 2025, preprint, arXiv:2507.17448, DOI: 10.48550/arXiv.2507.17448.
48 Y. Wang, C. Pang, Y. Wang, J. Jin, J. Zhang, X. Zeng, R. Su, Q. Zou and L. Wei, *Nat. Commun.*, 2023, **14**, 6155.

49 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk and A. Merchant, *et al.*, *Nature*, 2023, **624**, 86–91.
50 T. Dai, S. Vijayakrishnan, F. T. Szczypiński, J.-F. Ayme, E. Simaei, T. Fellowes, R. Clowes, L. Kotopanov, C. E. Shields and Z. Zhou, *et al.*, *Nature*, 2024, **635**, 890–897.
51 J. Li, C. Ding, D. Liu, L. Chen and J. Jiang, *Digital Discovery*, 2025, **4**, 1672–1684.
52 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
53 D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.
54 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nat. Mach. Intell.*, 2021, **3**, 144–152.
55 L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort and J. Grobler, *et al.*, *arXiv*, 2013, preprint, arXiv:1309.0238, DOI: 10.48550/arXiv.1309.0238.
56 D. Probst and J.-L. Reymond, *J. Cheminf.*, 2020, **12**, 12.
57 HuggingFace, *Transformers Documentation: Tokenizer*, https://huggingface.co/docs/transformers/en/main_classes/tokenizer.
58 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, *J. Mach. Learn. Res.*, 2020, **21**, 1–67.
59 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *arXiv*, 2017, preprint, arXiv:1706.03762, DOI: 10.48550/arXiv.1706.03762.
60 D. Christofidellis, G. Giannone, J. Born, O. Winther, T. Laino and M. Manica, *International Conference on Machine Learning*, 2023, pp. 6140–6157.
61 C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho and H. Ji, *arXiv*, 2022, preprint, arXiv:2204.11817, DOI: 10.48550/arXiv.2204.11817.
62 K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
63 C.-Y. Lin, *Text Summarization Branches Out*, 2004, pp. 74–81.
64 V. I. Levenshtein, *Phys.-Dokl.*, 1965, **10**, 707–710.
65 Orsinium, *TextDistance: Compute Distance between the Two Texts*, https://github.com/orsinium/textdistance, accessed: 2025-04-08.