Rethinking *NLP* for Chemistry: A Critical Look at the *USPTO* Benchmark

Derin Ozer[1], Nicolas Gutowski[1], Benoit Da Mota[1], Thomas Cauchy[2], Sylvain Lamprier[1]
[1]Univ Angers, LERIA, SFR MATHSTIC, F-49000 Angers, France,
[2]Univ Angers, CNRS, MOLTECH-ANJOU, SFR MATRIX, F-49000 Angers, France
Correspondence: derin.ozer@univ-angers.fr

Abstract

Natural Language Processing (*NLP*) has catalyzed a paradigm shift in Computer-Aided Synthesis Planning (*CASP*), reframing chemical synthesis prediction as a sequence-to-sequence modeling problem over molecular string representations like *SMILES*. This framing has enabled the direct application of language models to chemistry, yielding impressive benchmark scores on the *USPTO* dataset, a large text corpus of reactions extracted from US patents. However, we show that *USPTO*'s patent-derived data are both industrially biased and incomplete. They omit many fundamental transformations essential for practical real-world synthesis. Consequently, models trained exclusively on *USPTO* perform poorly on simple, pharmaceutically relevant reactions despite high benchmark scores. Our findings highlight a broader concern in applying standard *NLP* pipelines to scientific domains without rethinking data and evaluation: models may learn dataset artifacts rather than domain reasoning. We argue for the development of chemically meaningful benchmarks, greater data diversity, and interdisciplinary dialogue between the *NLP* community and domain experts to ensure real-world applicability.

1 Introduction

Recent advances in Natural Language Processing (*NLP*) have had a transformative impact well beyond linguistics, enabling breakthroughs across diverse domains. Among these, computational chemistry has emerged as a particularly promising field, where tools and models developed for human language are increasingly repurposed to reason over structured, sequential chemical representations (Chithrananda et al., 2020; Bagal et al., 2021).

A key challenge in this domain is Computer-Aided Synthesis Planning (*CASP*), which aims to automatically design viable synthesis routes to obtain target molecules. In practice, this involves identifying a series of chemical reactions that transform simple and/or commercially available starting materials into a desired target compound (Blakemore et al., 2018). This series of reactions is known as a synthesis route. Traditionally, designing such routes often requires expert intuition and extensive domain knowledge (Coley et al., 2019). *CASP* models aim to automate this reasoning process, providing chemists with viable synthetic strategies.

Finding a synthesis route is crucial in modern drug discovery, materials design, and green chemistry. Drawing on structural parallels with language modeling, researchers have increasingly reformulated synthesis planning as a sequence-to-sequence task, encoding molecules as linear string representations (Schwaller et al., 2019). This approach represents a direct recontextualization of *NLP* methods: language models are applied to a domain that shares some structural properties with natural language, yet differs fundamentally in semantics, evaluation, and goals.

Contemporary approaches today in the field of *CASP* are predominantly data-driven. A widely held argument in the field is that rule-based systems are costly to maintain and difficult to scale. From this perspective, models that bypass such rules are seen as more scalable and are often assumed to generalize better to previously unseen reactions (Wei et al., 2024).

However, this assumption does not always hold, especially when the training data is limited in scope, unrepresentative of the real distribution, or exhibits systematic biases. One dataset dominates the literature on single and multi-step chemical synthesis, *USPTO* (Lowe, 2012), which remains the only large-scale and open-source reaction dataset. It is derived from patents published by the United States Patent and Trademark Office. Beyond its accessibility, the structure of the dataset, where reactions are represented as SMILES strings in an input–output format, played a pivotal role in shifting the field

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22813–22825
November 4-9, 2025

toward data-driven, *NLP*-style modeling.

The SMILES notation (Weininger, 1988) offers a compact, linear representation of molecular graphs by encoding atoms and bonds as a sequence of characters. This linearization is achieved through a canonical traversal of the molecular graph, ensuring that each molecule is mapped to a unique and reproducible string while retaining sufficient information for reconstruction. Such a representation makes molecules inherently compatible with tokenization and sequence-based modeling. Consequently, the *USPTO*'s adoption of the SMILES format naturally aligned with sequence-to-sequence learning paradigms, motivating researchers to apply language models to chemical synthesis. In doing so, the *USPTO* dataset not only served as a foundational resource but also influenced the methodological direction of synthesis prediction, reinforcing the dominance of data-driven *NLP* approaches over symbolic or rule-based alternatives.

While different versions of the original *USPTO* dataset have become a standard benchmark in synthesis prediction, their scope is inherently constrained by their origin: patent literature. As such, they primarily capture reactions that are considered novel, industrially relevant, or commercially valuable enough to warrant patent protection. Despite repeated efforts to clean and curate the *USPTO* dataset (Schneider et al., 2016; Jin et al., 2017), and despite growing critiques regarding its efficiency (Yu et al., 2024), diversity (Torren-Peraire et al., 2024), and labeling bias (Griffiths et al., 2021), it remains the most used benchmark in the field of single-step and multi-step synthesis prediction. However, contrary to widespread assumptions, we argue that relying exclusively on this benchmark, and thus on patent-derived data, introduces significant bias during model training. This reliance systematically overlooks a broad portion of the chemical space, particularly foundational and non-patented reactions, thereby limiting the generalizability and real-world utility of current models. The lack of real-world adoption or application of models trained exclusively on these datasets further reinforces our claim: these models often fail to generalize to diverse chemical structures and fall short of the requirements for practical synthesis planning. This reveals a deeper issue: the adoption of *NLP*-style benchmarking practices in scientific domains can create the illusion of progress, even when models fail to achieve meaningful domain-level impact.

In this study, we critically examine the prevailing benchmark of *USPTO* that has been optimized for synthesis prediction research for a decade. We highlight that the problem of synthesis planning, originally rooted in assisting chemists, has drifted into a purely informatics challenge, disconnected from the experimental realities and needs of chemistry. To ground this critique, we conduct an empirical investigation that tests the generalization capacity of a state-of-the-art model trained on *USPTO*, *Chemformer* (Irwin et al., 2022). To ensure that our findings are not an artifact of model choice, we also include as a sanity check a *T5* model, *ProPreT5* (Ozer et al., 2025), we previously trained for the same task, which performs considerably worse than *Chemformer*. By evaluating these models on a set of foundational, pharmaceutically relevant reactions from the Hartenfeller dataset (Hartenfeller et al., 2011) absent from the *USPTO* benchmark, we demonstrate that high accuracy on *USPTO* does not necessarily translate into chemical reasoning or practical utility. This disconnect, we argue, calls for a reassessment of current benchmarks and evaluation practices within the field.

2 Background & Related Work

2.1 Computer-Aided Synthesis Planning

Chemical synthesis is the process of assembling small or readily available molecules (reactants) through a series of chemically feasible reactions to obtain desired compounds (products). The inverse of chemical synthesis, known as retrosynthesis, involves deconstructing a target molecule into simpler or commercially accessible precursors. This concept was formally introduced in the 1960s by (Corey, 1967), who laid the foundations for systematic synthetic planning. Chemical Synthesis can be roughly divided into two sub-problems: Single-step and multi-step synthesis. Single-step synthesis involves predicting the immediate product of a given set of reactants, while multi-step synthesis requires generating an entire series of reactions that iteratively construct a target molecule from simpler, purchasable precursors.

CASP can naturally be formalized as a computational problem, as it requires the exploration of a search space to identify a sequence of reactions that transform an initial state into a defined goal state.

Early implementation efforts led to rule-based expert systems such as *LHASA* (Corey et al., 1985)

and later *Chematica* (Szymkuć et al., 2016). These systems encoded chemical knowledge through an extensive set of manually curated reaction rules, based on decades of experimental findings from the literature. By recursively applying these rules, they could suggest valid reaction sequences, effectively emulating the decision-making process of experienced synthetic chemists. Although these approaches demonstrated the feasibility of *CASP*, they were limited by their reliance on static knowledge bases and their inability to easily adapt to new chemistry.

The recent surge of freely available reaction data (Lowe, 2012) coupled with *NLP* techniques has led to a shift from rule-based systems toward data-driven models.

2.2 *NLP* Models Applied to *CASP*

Since (Lowe, 2012) proposed his reaction dataset, data-driven approaches have gained immense popularity, leading to the increasing application of *NLP* techniques to *CASP* in recent years. This development is largely driven by the ability to represent molecules as linear strings through notations such as *SMILES* (Weininger, 1988), allowing the direct application of models initially designed for language tasks.

In the context of *CASP*, chemical reactions can be modeled as transformations from one sequence of tokens (representing the reactants and reagents) to another (the products). This creates a natural analogy with neural machine translation, where the objective is to "translate" a source sequence into a target sequence. As a result, both single-step and multi-step synthesis planning problems are increasingly formalized as machine translation tasks, allowing a wide range of *NLP* methods to be adapted for synthesis/retrosynthesis (Schwaller et al., 2019).

Before the advent of Transformer architectures, early approaches in this space relied heavily on sequence-to-sequence models with recurrent neural networks (RNNs). For instance, (Nam and Kim, 2016) first demonstrated the potential of neural machine translation for single-step synthesis, using encoder-decoder models to translate *SMILES* strings of reactants and reagents into products. These models were later improved with the use of different attention mechanisms (Schwaller et al., 2018), which improved the model's ability to focus on the relevant parts of the input during prediction.

However, it was the introduction of the Transformer architecture (Vaswani et al., 2017) that truly transformed the field. The *Molecular Transformer* (Schwaller et al., 2019) demonstrated that the Transformer architecture, originally developed for text generation, could achieve state-of-the-art performance in single-step synthesis when trained purely on *SMILES* strings, without requiring domain-specific knowledge. This architecture quickly became the foundation for subsequent innovations. Through data augmentation in the *Augmented Transformer* (Tetko et al., 2020), advanced pretraining strategies in *Chemformer* (Irwin et al., 2022), and the integration of structural information, models have achieved improved benchmark performance. For instance, Hybrid architectures like *MEGAN* (Sacha et al., 2021) combined structural information with language models to retain the strengths of both paradigms.

Yet, despite architectural advances, these *NLP* models are trained and evaluated on the same benchmark dataset. The improvements in benchmark scores give the illusion of progress, while these methods, due to the lacks in the dataset on which they are trained, fail to generalize to the chemical challenges they aim to solve.

2.3 *USPTO* Benchmark

The *USPTO* dataset traces its origins to the work of Daniel Lowe, who developed an *NLP* pipeline to extract chemical reactions from the full text of US patent documents published between 1976 and 2013. Lowe processed millions of patent documents to create a dataset of approximately 1.8 million reactions (Lowe, 2012). Extracted from unstructured patent text, the *USPTO* dataset can be regarded as a large text-derived corpus, lending itself naturally to NLP-based modeling approaches. However, this early dataset contained significant noise, e.g., redundant entries, poorly parsed *SMILES* strings, and missing conditions. Importantly, it lacked any formal reaction classification or quality control, as the extraction goal was broad coverage rather than curated benchmark construction.

Recognizing these issues, subsequent efforts attempted to curate Lowe's dataset for machine learning applications. One of the most prominent curated versions is the *USPTO-MIT* dataset (Jin et al., 2017), which removed duplicates and erroneous reactions, resulting in approximately 479,000 reactions. The reactions were cleaned, but no effort was made to balance reaction classes or correct deeper

biases inherited from patent literature. Notably, the *USPTO-MIT* dataset is available in two versions, *Mixed* and *Separate*. In the *Mixed* version, reactants (the molecules that undergo transformation) and reagents (such as solvents, catalysts, or additives that facilitate the reaction but are not transformed) are combined into a single input string. In contrast, the *Separate* version keeps reactants and reagents distinct. Although the *Separate* version allows for more chemically accurate modeling, many NLP-based models use the *Mixed* version, as it simplifies tokenization and fits more naturally into a sequence-to-sequence framework.

To facilitate model benchmarking, the *USPTO-50K* subset was later introduced, manually selecting 50,000 reactions grouped into 10 major reaction classes of *USPTO* (Schneider et al., 2016). While *USPTO-50K* provides a standardized testbed for reaction prediction tasks, it represents an even narrower view of chemical reactivity, favoring only high-frequency transformations, while underrepresenting many other classes of chemical reactions.

Throughout the evolution of the *USPTO* benchmark, the biases inherent in how the dataset was generated, due to its data acquisition strategy, have remained unchanged.

3 Benchmark Bias

3.1 Dataset Limitations

The *USPTO* benchmark, although widely used, exhibits several structural biases that limit its suitability to evaluate general-purpose synthesis models applicable to real-world scenarios.

1. Patent bias: Reactions included in the dataset are extracted exclusively from U.S. patent filings, which, by definition, focus on novelty and reactions innovative enough to warrant protection. As a result, the dataset overrepresents reaction types that are deemed patentable, such as those related to pharmaceutical or agrochemical innovations.

2. Lack of basic, textbook chemistry: Many simple and foundational transformations commonly encountered in introductory organic chemistry courses are absent from *USPTO*-based corpora. Despite their simplicity, these reactions serve as essential building blocks in the synthesis of a wide range of organic molecules. This is largely because such routine transformations are rarely featured in patents, which tend to focus on novel or proprietary chemistry. Consequently, foundational reactions, despite being central to many synthesis routes, are often underrepresented or entirely missing from the *USPTO* dataset. We argue that models trained exclusively on *USPTO* therefore struggle to predict even well-established synthesis routes for simple or well-known molecules, due to a lack of exposure to such fundamental chemistry. The following sections provide a more detailed examination of this observation.

3. Imbalanced distribution: Apart from *USPTO-50K*, which represents an even narrower subset of chemical space, the broader *USPTO* datasets are highly imbalanced, as illustrated in Figure 1 and Figure 2. Certain reaction classes appear tens of thousands of times, whereas many others are underrepresented. This imbalance biases model learning and evaluation toward dominant patterns, potentially inflating benchmark performance without corresponding gains in general chemical reasoning.

The *USPTO* benchmark is shaped by biases stemming from its data acquisition strategy, which has created a deep structural bias that manifests as a long-tailed distribution of reactions and poor out-of-distribution (OOD) generalization. This limitation can be illustrated with an analogy to language. A language model trained exclusively on legal contracts would acquire detailed knowledge of technical vocabulary and rigid syntactic patterns specific to that domain. Within legal text, such a model might appear highly competent, predicting clauses with accuracy and reproducing stylistic conventions with ease. Yet its competence would break down in broader linguistic settings. Confronted with an everyday expression such as “break the ice”, the model would fail to capture the intended meaning, even though it has encountered the words break and ice many times, because that usage never occurs in the training distribution. The situation is directly parallel in chemistry: models trained on patent-derived corpora like *USPTO* learn associations shaped by the narrow domain of patent literature but fail to generalize to foundational textbook reactions that are absent from the dataset, despite being composed of the same atoms and following regular chemical logic. Importantly, this limitation cannot be remedied with standard techniques for

handling long-tailed data or OOD generalization. Such methods may mitigate distributional imbalance when examples are at least present in training, but they cannot compensate for entire classes of reactions that are missing altogether.

OOD generalization itself remains a notoriously difficult challenge with no universal solution despite extensive study (Hendrycks and Dietterich, 2019; Arjovsky et al., 2019; Gulrajani and Lopez-Paz, 2020). As a result, when the training distribution is structurally incomplete by design, as with *USPTO*, even the most sophisticated architectures and pretraining strategies are fundamentally constrained. Addressing these limitations requires rethinking benchmark design rather than relying on model innovation or data-driven adjustments alone.

3.2 Experimental Setup

To move beyond theoretical critique and obtain a concrete understanding of how the *USPTO* dataset influences model behavior, we conduct an empirical analysis focused on its impact on generalization performance.

3.2.1 Model Selection

To carry out our empirical investigation, we selected *Chemformer* (Irwin et al., 2022) as a representative model. *Chemformer* is a sequence-to-sequence Transformer model specifically designed for different molecular downstream tasks. Architecturally, it builds upon the BART model (Lewis et al., 2019), employing both an encoder and a decoder stack. This makes it well-suited for SMILES-to-SMILES prediction tasks. What distinguishes *Chemformer* from other models in synthesis prediction is its use of transfer learning and supervised pretraining. The model is pretrained on 100 million unlabelled SMILES strings from the ZINC-15 database (Sterling and Irwin, 2015) using three complementary tasks: 1. Span masking (Lewis et al., 2019), where short sequences of tokens are replaced with a mask token. 2. SMILES augmentation (Bjerrum, 2017), in which non-canonical forms are used as input, a chemistry-specific analogue of paraphrasing. 3. A combined task that integrates both strategies.

During pretraining, the model receives these corrupted sequences as input and is trained to reconstruct the original SMILES strings. This denoising objective provides a strong inductive bias for reaction modeling, where output products are often a transformed version of the input reactants. The result is a general-purpose chemical language model that can be fine-tuned efficiently on a variety of downstream tasks.

Chemformer, when trained on the *USPTO-MIT Mixed* dataset for synthesis prediction on top of its pretraining, achieves state-of-the-art performance on *USPTO*-based benchmarks. It reports a top-1 accuracy of 90.9% on the dataset, making it a strong and credible baseline for evaluation.

We selected the base version of *Chemformer*, which is a pure *NLP* model, because its extensive pretraining on large molecular corpora endows it with broad chemical knowledge, and its readily available trained weights facilitate reproducible and efficient experimentation. While other models also demonstrate high performance on *USPTO*, as summarized in Table 1, our goal is not to conduct a comprehensive benchmark study. Instead, we aim to critically assess how heavy reliance on a single benchmark like *USPTO*, despite impressive reported metrics, can mask certain limitations in model generalization and chemical reasoning. To ensure that these limitations do not simply reflect our choice of model, we additionally include results from *ProPreT5* (Ozer et al., 2025), a classic *T5* (Raffel et al., 2020) model that we trained on the *USPTO-MIT* dataset for single-step synthesis in prior work. Although *ProPreT5* performs worse than *Chemformer*, it serves as a useful sanity check, confirming that the observed failures stem from the benchmark's limitations rather than from a particular architecture.

This investigation is not intended as a critique of *Chemformer*'s technical design, but rather as a critical assessment of the dataset on which it is trained and evaluated, as well as the broader implications of relying on *USPTO* as the primary benchmark for synthesis modeling.

3.2.2 Constructing a Test Set for Generalization Assessment

In this section, we base our comparison on *USPTO-MIT Mixed*, as it is the most commonly used version of the *USPTO* dataset and remains closest to the original, with only duplicate and erroneous reactions removed.

To complement the investigation into the limitations of *USPTO*, we employed the *Hartenfeller* reaction rules (Hartenfeller et al., 2011) to generate a test dataset. This curated collection of 58 reaction templates was specifically designed for the generation of drug-like molecules. In contrast to *USPTO*,

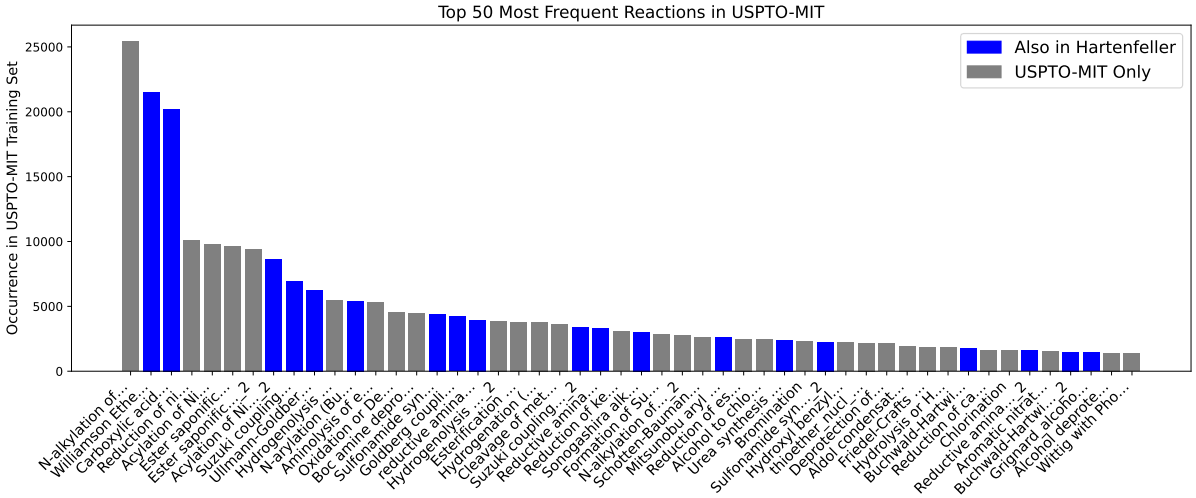


Figure 1: Top 50 most frequent reaction names in the *USPTO-MIT* dataset. Reactions also present in the *Hartenfeller* dataset are shown in blue. Reaction names are obtained using the *Rxn-INSIGHT* (Dobbelaere et al., 2024) tool.

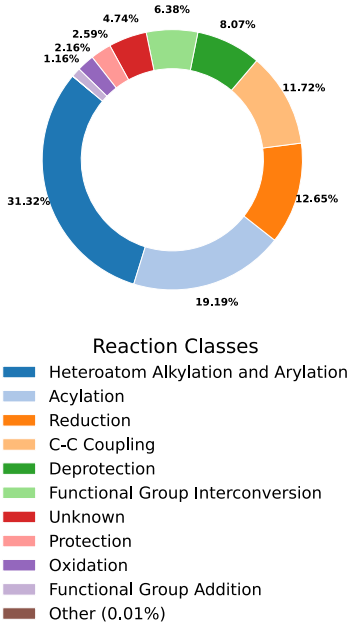


Figure 2: Distribution of reaction classes in the *USPTO-MIT* dataset. Reaction names are obtained using the *Rxn-INSIGHT* (Dobbelaere et al., 2024) tool.

which is derived from industrial patents and thus biased toward novel chemistry, the *Hartenfeller* dataset emphasizes reliable, fundamental transformations that are routinely used in medicinal chemistry. Each reaction was selected based on its practical utility, robustness, and compatibility with a wide range of starting materials, making it particularly well-suited for probing the generalization capabilities of synthesis models trained on patent-centric data.

We selected the *Hartenfeller* reaction templates for comparison precisely because they contain transformations that are underrepresented in or entirely absent from *USPTO-MIT Mixed*, yet are crucial for real-world compound design. This choice explicitly allows us to probe the absence of textbook chemistry in *USPTO* and its impact on model generalization. By evaluating *Chemformer* on reactions from this set, we aim to assess whether high benchmark performance on *USPTO-MIT Mixed* (see Table 1) reflects genuine generalization to practically important chemical transformations.

Table 1: Top-k test accuracy for single-step synthesis on the *USPTO-MIT Mixed* dataset. Results taken directly from the references.

Model	Top-1	Top-10
MEGAN (Sacha et al., 2021)	86.3%	95.4%
ProPreT5 (Ozer et al., 2025)	87.9%	-
Molecular Transformer (Schwaller et al., 2019)	88.6%	-
Graph2SMILES (Tu and Coley, 2022)	90.3%	95.2%
Augmented Transformer (Tetko et al., 2020)	90.4%	96.5%
Chemformer (Irwin et al., 2022)	90.9%	94.1%

One important distinction between the *Hartenfeller* dataset and *USPTO-MIT Mixed* is their structure and intended use. The *USPTO-MIT Mixed* dataset is a synthesis prediction corpus, where each example consists of input molecule strings(reactants) and an output molecule string(the product). In contrast, the *Hartenfeller* dataset provides reaction templates, encoded as SMARTS (Inc.) strings, which define generic chemical transformations in a rule-based form. SMARTS is a language for specifying molecular patterns, much like regular expressions in text processing, and is

commonly used to define substructure patterns in molecules.

To use the *Hartenfeller* templates for evaluation, we needed to convert them into a format compatible with synthesis prediction models, that is, to generate explicit input–output pairs like those in *USPTO-MIT Mixed*. This required identifying real molecules that could be used as inputs (reactants) for each reaction template.

To do this, we performed substructure matching using *RDKit* (RDKit, 2025), a widely used open-source cheminformatics toolkit. Substructure matching is analogous to pattern matching in NLP: just as a parser might check if a certain syntactic pattern exists in a sentence, here we check if a molecular fragment described by the SMARTS template exists within a given molecule. We applied this matching process to molecules from the *USPTO-MIT Mixed* training set to ensure compatibility with *Chemformer*'s tokenizer and avoid issues with out-of-vocabulary tokens.

Each *Hartenfeller* template was matched against *USPTO-MIT Mixed* molecules to find valid reactant combinations. These combinations were then used to simulate the reaction using *RDKit*'s reaction engine. If the reaction produced a valid product, the input–output pair was retained. This process was repeated until up to 50 valid examples per reaction were collected. For completeness, we retained all *Hartenfeller* reactions in this process, despite the expected overlap with reaction types in *USPTO-MIT Mixed*.

3.2.3 Criteria for Determining Seen Reactions

To identify overlapping reaction types between the *USPTO-MIT Mixed* training set and the *Hartenfeller* test set, we needed a consistent naming scheme for reactions in both corpora. For this purpose, we employed *Rxn-INSIGHT* (Dobbelaere et al., 2024), an open-source reaction classification tool that automatically assigns standardized names based on the underlying chemical transformation. *Rxn-INSIGHT* analyzes the structural changes between reactants and products and assigns reaction names from a large, curated taxonomy.

We applied *Rxn-INSIGHT* to both the training data from *USPTO-MIT Mixed* and the *Hartenfeller*-generated examples. This allowed us to align naming conventions and systematically determine which *Hartenfeller* reaction types were absent from the training data. Ensuring consistent reaction labels across datasets was critical for performing a controlled generalization study and fairly isolating the impact of unseen reaction types. As a result, if a reaction from the *Hartenfeller* set is assigned the same reaction name by *Rxn-INSIGHT* as one found in *USPTO-MIT Mixed*, we consider it a "seen" reaction.

3.3 Empirical Evaluation

To empirically assess the impact of benchmark-driven bias, we evaluated the generalization capabilities of *Chemformer* on the generated test set described in the previous section. The model weights used were those pretrained and fine-tuned on the *USPTO-MIT Mixed* dataset, which are publicly available.

3.3.1 Frequency-Driven Bias

Data-driven approaches are often favored in chemical synthesis prediction for their presumed ability to generalize beyond explicitly seen examples (Wei et al., 2024). However, our results challenge this assumption. Figure 3 visualizes model performance by plotting accuracy for each reaction type in the generated test set against its frequency in the *USPTO-MIT Mixed* training data. In addition to *Chemformer*'s Top-1 accuracy, we also report its Top-10 accuracy as well as the Top-1 accuracy of *ProPreT5*. Across both models and metrics, the same trend is observed: reactions frequently represented in the training corpus achieve moderate to high predictive accuracy, whereas reactions that are rare or absent from training yield near-zero performance. Accordingly, the lack of generalization persists under both a more permissive evaluation metric (Top-10 accuracy with *Chemformer*) and an alternative architecture (*ProPreT5*), confirming that the limitation stems from benchmark-driven bias rather than Top-1 variability or model choice.

This performance pattern confirms that, regardless of architecture or pretraining, models trained on *USPTO-MIT Mixed* remain constrained by the biases present in the benchmark. *Chemformer*, despite its sophisticated design and extensive pretraining, achieves high accuracy primarily when test examples closely resemble patterns frequently seen during training. Thus, the observed limitation is not specific to a single model but inherent to the dataset, dooming any *USPTO*-trained model to fail at true generalization.

Interestingly, some frequent reactions from *USPTO-MIT Mixed* are poorly predicted, while some infrequent ones achieve non-zero accuracy.

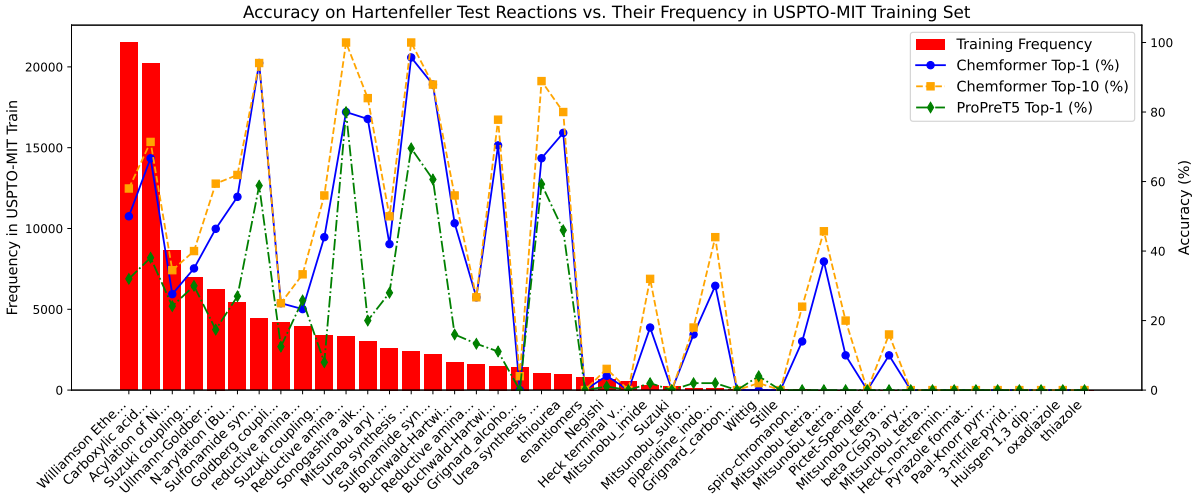


Figure 3: Top-1 accuracy of *Chemformer* on transformations from the *Hartenfeller* test set, plotted against the frequency of each reaction name in the *USPTO-MIT Mixed* training set.

This suggests factors beyond frequency, such as reaction complexity, template consistency, or similarity to other reactions, also affect performance. For example, different variants of the *Mitsunobu* reaction are correctly predicted despite being underrepresented, likely due to shared mechanisms with the frequent variants the model can learn. We revisit this in the next section using other mechanistically similar reactions. Nonetheless, the trend reflects a reliance on memorized patterns rather than generalized reasoning.

Table 2: Top-*k* prediction accuracy of *Chemformer* on different subsets of the generated *Hartenfeller* test set for single-step synthesis. The subsets are grouped by the number of times each reaction type appears in the *USPTO-MIT Mixed* training set.

***Hartenfeller* Subset**	**Top-1**	**Top-5**	**Top-10**
Seen < 1000× in training	8.15%	10.81%	11.04%
Seen < 100× in training	7.07%	10.05%	10.30%
Seen < 10× in training	1.09%	1.53%	1.75%

To examine performance as a function of reaction frequency in the training data, we grouped the evaluation set by rarity. Table 2 presents Top-1, Top-5, and Top-10 accuracy across reactions that occur fewer than 1000, 100, and 10 times in *USPTO-MIT Mixed*. The results are stark: for reactions seen fewer than 1000 times, Top-1 accuracy drops to 8.15%. For fewer than 100, it falls to 7.07%, and for reactions appearing fewer than 10 times, performance collapses to 1.09%. These findings underscore that simply including a reaction type in the training set is not sufficient for accurate prediction, as high-frequency transformations overshadow rarer ones. While this imbalance clearly hinders learning, it is only part of the problem: many foundational reactions are entirely absent from *USPTO*, and such structural gaps cannot be resolved through balancing or reweighting.

3.3.2 Analogical Generalization

To better understand what is missing from the current *USPTO-MIT Mixed* benchmark in enabling robust generalization, we designed a controlled experiment targeting analogical inference. Specifically, we selected two mechanistically related reactions from the *Hartenfeller* dataset that are absent from *USPTO-MIT Mixed*. For clarity, we refer to these reactions as A: *1,2,4-triazole_acetohydrazide* and B: *1,2,4-triazole_carboxylicacid/ester*. The mechanisms of both transformations are visualized in Appendix C using a *SMARTS*-based tool (Ehrt et al., 2020). Although they proceed via different mechanisms, both reactions converge on the formation of the same cyclic structure in the product. This setup enables us to test whether a model trained on *USPTO-MIT Mixed* can generalize based on the underlying reaction logic, rather than relying solely on memorized templates.

As shown in Table 3, before any fine-tuning, *Chemformer* achieved 0.0% Top-1 accuracy on both reactions. We then, using the fine-tuning setup proposed in (Irwin et al., 2022), fine-tuned the model on just 1,000 examples of Reaction A and re-evaluated it on both Reaction A and the unseen Reaction B. After fine-tuning, the model achieved 97.4% Top-1 accuracy on Reaction A and 18.8%

on Reaction B (Top-5: 36.6%, Top-10: 38.0%). These results demonstrate that the model is capable of transferring knowledge between mechanistically related transformations but only when the training data provides a sufficiently clear and similar mechanistic signal.

Table 3: Top-*k* accuracy (%) of *Chemformer* on Reaction A and Reaction B before and after fine-tuning. Fine-tuning was done on Reaction A using 1000 examples and training for 20 epochs.

Setting	Top-1	Top-5	Top-10
Reaction A (before)	0.0%	0.0%	0.0%
Reaction B (before)	0.0%	0.0%	0.0%
Reaction A (after)	97.4%	97.8%	98.0%
Reaction B (after)	18.8%	36.6%	38.0%

We also monitored how this fine-tuning affected performance on the original *USPTO MIT Mixed* test set. A minor drop was observed (Top-1: 88.6% vs. 90.9%), suggesting some degree of overfitting to the fine-tuned class. Nonetheless, the broader implication is clear: For data-driven *NLP* models to generalize effectively in chemistry, the training data must be as diverse as possible. Only then can these models extend beyond memorized templates to handle unseen but mechanistically related chemical transformations.

Our findings in this section echo those of (Su et al., 2022), who observed a similar effect in the context of *Chan–Lam coupling*. Historically, the *Chan–Lam* reaction was developed by combining elements of the *Suzuki* and *Barton* couplings. That study aimed to simulate this intuition-driven discovery process using *NLP* models. When *Chan–Lam*, *Suzuki*, and *Barton* reactions were removed from training, the model's Top-1 accuracy on *Chan–Lam coupling* dropped to 4.4%. When *Suzuki* and *Barton* reactions were reinstated, accuracy increased to 24.8%, highlighting the importance of mechanistically related reactions in enabling model inference. Our work confirms that this is not an isolated case and extends the analysis by systematically probing the limitations of this widely used *NLP* benchmark, providing a broader empirical perspective on its lack of diversity and limited support for generalization.

Taken together, these results reveal what the *USPTO* benchmark currently lacks: sufficient diversity, broad distributional coverage, and representation of mechanistically related transformations. For synthesis prediction to reach real-world applicability, benchmarks must evolve to reflect the structure of chemical knowledge itself, enabling models to move from memorization to inference.

4 Conclusion

This study critically re-evaluates the *USPTO* benchmark, a cornerstone resource for *NLP*-based synthesis prediction. While the dataset and its subsets have advanced the field, their widespread adoption has also introduced systemic biases. Our findings demonstrate that models trained on *USPTO*, despite strong benchmark performance, fail to generalize beyond its narrow scope. The current benchmark provides an incomplete view of the chemical transformation space, omitting much of its diversity.

The aim of this work is not to propose new benchmarks or corrective methods, but to draw attention to the structural issues in the current standard. Highlighting these limitations is, in itself, a timely and necessary contribution: without recognizing what is missing, progress risks being evaluated against an incomplete benchmark. More broadly, as *NLP*-inspired methods expand into scientific domains, benchmarks must be critically assessed to ensure they capture genuine scientific reasoning rather than artifacts of data collection. Only by confronting these limitations can interdisciplinary research move toward models that meaningfully support discovery.

Code and Data Availability

Code and data are available at: https://github.com/DerinOzer/benchmark-bias

Acknowledgments

This work was supported by the University of Angers, and the French Ministry of Education and Research (JL PhD grant). This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011014840R1 and AD011014032R2).

AI assistance was used to improve the clarity, flow, and phrasing of the manuscript.

Limitations

Our study faces several limitations that we are currently unable to overcome. Most significantly, we do not propose an immediate alternative to the widely adopted *USPTO* datasets. While we have effectively demonstrated the limitations of *USPTO*

and raised concerns about its suitability as a benchmark, the creation of a more representative and generalizable dataset remains an open challenge. Constructing such a corpus would require the collaborative effort of researchers across the globe, given the breadth and diversity of synthetic chemistry. If language models are to become competent agents for synthesis planning, they must be exposed to the full landscape of chemical transformations known to human experts.

A second limitation stems from the naming consistency and reaction classification across datasets. Although we employed *Rxn-INSIGHT* to standardize reaction names between the *USPTO-MIT Mixed* and generated test data using *Hartenfeller* reaction set, the tool itself is not infallible. Errors in classification or missed equivalences between similar reactions could have affected the exclusion or inclusion of certain reactions in our test set. Consequently, our definition of "unseen" reaction types, while methodologically principled, may still suffer from edge cases or subtle misalignments.

Finally, our reliance on *SMARTS*-based templates to generate the evaluation set introduces its own limitations. While templates offer control and interpretability, they encode simplified views of reactivity and do not capture all the nuances of context-dependent reactivity in real-world chemical systems. The evaluation set, while diverse, may thus underrepresent complex reaction conditions or fail to reflect the full spectrum of transformation types relevant in synthesis.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. 2021. Molgpt: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*, 62(9):2064–2076.

Esben Jannik Bjerrum. 2017. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*.

David C Blakemore, Luis Castro, Ian Churcher, David C Rees, Andrew W Thomas, David M Wilson, and Anthony Wood. 2018. Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry*, 10(4):383–394.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377.

Elias James Corey. 1967. General methods for the construction of complex molecules. *Pure and Applied chemistry*, 14(1):19–38.

Elias James Corey, Alan K Long, and Stewart D Rubenstein. 1985. Computer-assisted analysis in organic synthesis. *Science*, 228(4698):408–418.

Maarten R Dobbelaere, István Lengyel, Christian V Stevens, and Kevin M Van Geem. 2024. Rxn-insight: fast chemical reaction analysis using bond-electron matrices. *Journal of Cheminformatics*, 16(1):37.

Christiane Ehrt, Bennet Krause, Robert Schmidt, Emanuel SR Ehmki, and Matthias Rarey. 2020. Smarts. plus–a toolbox for chemical pattern design. *Molecular informatics*, 39(12):2000216.

Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha A Lee. 2021. Dataset bias in the natural sciences: a case study in chemical reaction prediction and synthesis design. *arXiv preprint arXiv:2105.02637*.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Markus Hartenfeller, Martin Eberle, Peter Meier, Cristina Nieto-Oberhuber, Karl-Heinz Altmann, Gisbert Schneider, Edgar Jacoby, and Steffen Renner. 2011. A collection of robust organic synthesis reactions for in silico molecule design. *Journal of chemical information and modeling*, 51(12):3093–3098.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Daylight Chemical Information Systems Inc. Smarts – a language for describing molecular patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed: 2025-05-02.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. 2017. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Daniel Mark Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis.

Juno Nam and Jurae Kim. 2016. Linking the neural machine translation and the prediction of organic chemistry reactions. *arXiv preprint arXiv:1612.09529*.

Derin Ozer, Sylvain Lamprier, Thomas Cauchy, Nicolas Gutowski, and Benoit Da Mota. 2025. A transformer model for predicting chemical reaction products from generic templates. *arXiv preprint arXiv:2503.05810*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

RDKit. 2025. RDKit: Open-source cheminformatics. https://rdkit.org/. Accessed 10 May 2025.

Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. 2021. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. 2016. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. 2018. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

Teague Sterling and John J Irwin. 2015. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337.

An Su, Xinqiao Wang, Ling Wang, Chengyun Zhang, Yejian Wu, Xinyi Wu, Qingjie Zhao, and Hongliang Duan. 2022. Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions. *Physical Chemistry Chemical Physics*, 24(17):10280–10291.

Sara Szymkuć, Ewa P Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A Grzybowski. 2016. Computer-assisted synthetic planning: the end of the beginning. *Angewandte Chemie International Edition*, 55(20):5904–5937.

Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. 2020. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575.

Paula Torren-Peraire, Alan Kai Hassen, Samuel Genheden, Jonas Verhoeven, Djork-Arné Clevert, Mike Preuss, and Igor V Tetko. 2024. Models matter: The impact of single-step retrosynthesis on synthesis planning. *Digital Discovery*, 3(3):558–572.

Zhengkai Tu and Connor W Coley. 2022. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling*, 62(15):3503–3513.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yixin Wei, Leyu Shan, Tong Qiu, Diannan Lu, and Zheng Liu. 2024. Machine learning-assisted retrosynthesis planning: current status and future prospects. *Chinese Journal of Chemical Engineering*.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Kevin Yu, Jihye Roh, Ziang Li, Wenhao Gao, Runzhong Wang, and Connor Coley. 2024. Double-ended synthesis planning with goal-constrained bidirectional search. *Advances in Neural Information Processing Systems*, 37:112919–112949.

A License and Terms for Use and Distribution of Artifacts

Transparency around data, models, and tools is critical in interdisciplinary research, both to ensure reproducibility and to clarify the scope and limitations of our findings. This study relies exclusively on publicly available datasets, pretrained models, and open-source software. All artifacts are used in accordance with their respective licenses and strictly within a research context. We also release our own code and generated datasets under the *MIT License* to facilitate reproducibility. None of the datasets used contain demographic or sensitive personal information, and there is no risk of re-identification or inference of protected attributes.

Artifact	Description / Usage	License
USPTO-MIT Mixed (Jin et al., 2017)	Patent-derived dataset	MIT
Hartenfeller reaction set (Hartenfeller et al., 2011)	58 *SMARTS* templates for drug-like synthesis	Public research use
Chemformer (Irwin et al., 2022)	Pretrained seq2seq model for reaction prediction	Apache 2.0
ProPreT5(Ozer et al., 2025)	*T5*-based model trained on *USPTO-MIT Mixed*	MIT
SMILES (Weininger, 1988)	Molecular line notation	Public domain
RDKit (RDKit, 2025)	Cheminformatics toolkit	BSD License 2.0
Rxn-INSIGHT (Dobbelaere et al., 2024)	Reaction classification and standardization	MIT

Table 4: Summary of datasets, models, notations, and software used in this study.

A.1 Overview of Artifacts

Table 4 summarizes the datasets, models, chemical notations, and software libraries used in this work, together with their intended role and license.

A.2 Datasets

1. ***USPTO-MIT Mixed*** (Jin et al., 2017): Derived from U.S. patent literature, containing ~480K chemical reactions in *SMILES* notation. Licensed under the MIT License and widely adopted in synthesis prediction research. We use the standard split (410,000 Train, 30,000 Validation, 40,000 Test).

2. ***Hartenfeller reaction set***(Hartenfeller et al., 2011): A curated collection of 58 reaction templates, expressed in *SMARTS* notation. Designed for drug-like molecule generation and made available for academic research. We adapted these templates into explicit input–output examples for our test set.

A.3 Models

1. ***Chemformer*** (Irwin et al., 2022): A transformer-based sequence-to-sequence model pretrained on large molecular corpora, released under the Apache License 2.0. We fine-tuned the base version using the provided pretrained weights.

2. ***ProPreT5***: Chemical language model, trained on *USPTO-MIT Mixed* for single-step synthesis. *ProPreT5* is released under the MIT License.

A.4 Chemical Representations

1. ***SMILES*** (Weininger, 1988): A line notation for representing molecules. It is in the public domain and widely used in cheminformatics.

2. ***SMARTS*** (Inc.): An extension of *SMILES* that encodes molecular substructure patterns and generic transformations. Publicly available for academic research through Daylight Chemical Information Systems.

A.5 Software Libraries

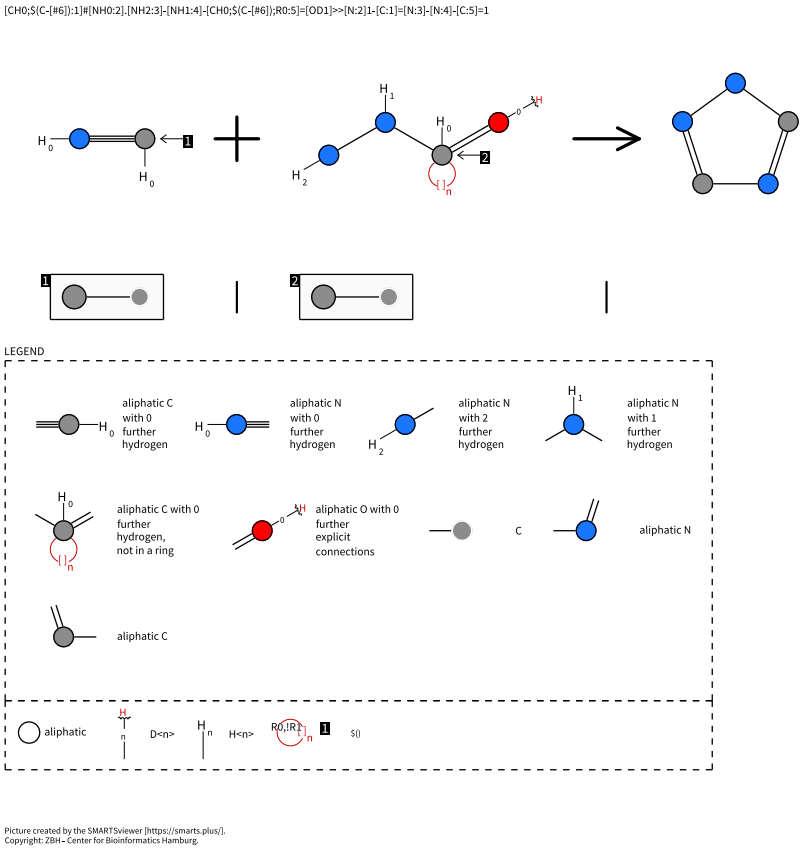
1. ***RDKit*** (RDKit, 2025): An open-source cheminformatics toolkit (BSD License 2.0) used for substructure matching, reaction application, and molecule manipulation.

2. ***Rxn-INSIGHT*** (Dobbelaere et al., 2024): A reaction classification tool that assigns standardized reaction labels based on structural changes between reactants and products. Released under the MIT License.

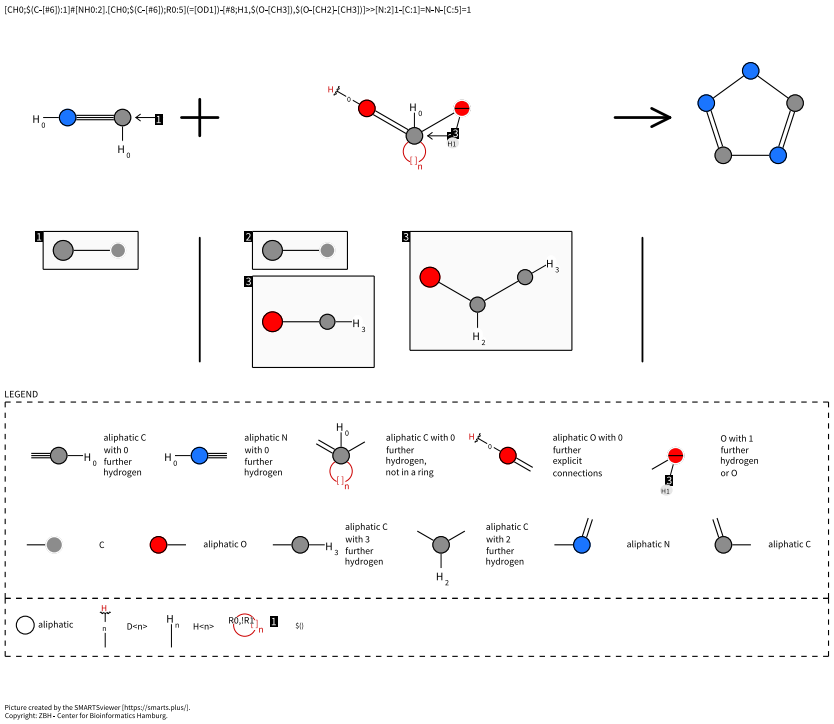
B Computational Resources and Training Details

We fine-tuned *Chemformer* using the setup described in the original paper (Irwin et al., 2022). Training was performed on a single NVIDIA V100 GPU for 20 epochs, with all other hyperparameters kept identical to those reported by (Irwin et al., 2022).

C Reaction Mechanisms Used in the Analogical Experiment



(a) Transformation mechanism of reaction A used in fine-tuning: *1,2,4-triazole_acetohydrazide*



(b) Transformation mechanism of reaction B used in test: *1,2,4-triazole_carboxylic-acid/ester*

Figure 4: Visualisation of transformation mechanisms using *SMARTS.plus* (Ehrt et al., 2020).