Automating Deep Learning-based Generation and Evaluation of De Novo Chemical Reaction with ChemRxnSAGE

Anis Ismail,[*,†,¶] Joe Tekli,[*,†] and Brigitte Wex[*,‡]

†*Electrical and Computer Engineering Department, Lebanese American University, 36 Byblos, Lebanon*

‡*Department of Physical Sciences, Lebanese American University, 36 Byblos, Lebanon*

¶*Laboratory of Multi-omic Integrative Bioinformatics, Department of Human Genetics, Faculty of Medicine, KU Leuven, 3000 Leuven, Belgium*

E-mail: anis.ismail@kuleuven.edu; joe.tekli@lau.edu.lb; brigitte.wex@lau.edu.lb

Abstract

The generation and evaluation of chemical reactions remain challenging, with limited comprehensive studies addressing these issues. We introduce the **Chem**ical Reaction (**Rxn**) **S**ystematic **A**ssessment of **G**eneration and **E**valuation (**ChemRxnSAGE**) framework, an adaptable end-to-end approach for evaluating the quality, validity, and diversity of machine-generated chemical reactions. Combining automated validity filters with quality metrics and expert insights, ChemRxnSAGE systematically eliminates invalid reactions. We test its robustness using generative models, including Recurrent Neural Networks and Variational Autoencoders, followed by validation using a chemical "Turing test" with domain experts. Additionally, we assess reaction feasibility through thermodynamic analysis and compare the generated reactions against existing literature to ensure relevance and novelty. By combining computational tools with expert-driven metrics, ChemRxnSAGE offers a comprehensive and extendable solution that advances the state of chemical reaction generation and evaluation.

Introduction

Chemical reaction modeling is a critical component of de novo drug design (DNDD), enabling the design and planning of feasible chemical transformations that support the development of novel molecules. While advances in machine learning (ML) have accelerated molecular generation, a key bottleneck remains in ensuring the synthetic feasibility and diversity of reactions proposed by these models. In recent years, ML-powered approaches have made significant strides in generating valid and diverse molecular structures;[1–4] however, their application to reaction synthesis remains comparatively underexplored.

A successful de novo drug design method should generate molecules that are synthetically accessible.[5] Various tools such as RAscore,[6] DFRscore,[7] and BR-SAScore[8] have been developed to estimate synthetic feasibility, yet these focus largely on molecular structures rather than full reaction pathways. Bridging this gap requires robust frameworks that can evaluate the chemical validity and diversity of reactions, which is an essential step toward the development of viable synthetic routes for ML-generated candidates.

While deep learning can enhance chemical pathways for drug design and augment small datasets to improve retrosynthesis models, its adoption in de novo chemical reaction synthesis has so far been limited.[9–12] This can be attributed to the following challenges:

- **Challenge 1**: Difficulty in codifying comprehensive chemical rules, coupled with the scarcity of negative reactions, i.e. reactions that do not occur chemically. Prior work has highlighted the importance of incorporating negative data to improve model performance,[13] with recent studies demonstrating enhanced accuracy in state-of-the-art reaction outcome prediction through the use of such data.[14] However, integrating negative reactions into the training process remains nontrivial. This complicates the evaluation of machine-generated chemical reactions, particularly in the absence of negative data. Assessing the validity of these reactions without relying on computationally expensive simulations or expert annotation represents a substantial barrier.

- **Challenge 2**: Limited diversity of generated reactions, restricting the discovery of novel synthetic pathways. Ensuring both the generation of a wide variety of reactions while maintaining validity is essential for broadening the space of possible chemical reactions. To overcome the limitations of chemical synthesis planning based on existing transformation rules,[15] which restricts reaction diversity, prior studies have advocated for template-free approaches capable of predicting reaction outcomes[16–18] and even entirely novel reactions.[9–12]

- **Challenge 3**: Lack of standardized evaluation protocols, with existing works[9–12] developing their own approach to assess the quality of generated reactions, and complicating the reproducibility and comparison of de-novo chemical reaction synthesis.

This study introduces **Chem**ical Reaction (**Rxn**) **S**ystematic **A**ssessment of **G**eneration and **E**valuation (ChemRxnSAGE), an adaptable end-to-end framework for evaluating the quality, validity, and diversity of machine-generated chemical reactions. ChemRxnSAGE addresses the key challenges in de novo chemical reaction synthesis by integrating automated validity checks and expert-driven evaluation to assess machine-generated reactions without heavily relying on costly simulations. It ensures reaction diversity and variety through quantitative metrics and filters grounded in chemical rules and enriched by the expertise of domain experts. By providing a standardized evaluation framework with common metrics, ChemRxnSAGE facilitates direct comparisons between models, establishing a consistent foundation for future research while enhancing the assessment of generated reactions.

Literature Review

Chemical Reactions Synthesis using Deep Learning

Conventional SMILES representation for reactions[19] presents significant challenges due to its non-sequential format for molecular structures and the need for lengthy sequences to accurately depict a reaction, making it difficult to identify relationships between substructures (see Supplementary Section S1 for background on the SMILES format). Additionally, reaction center identification requires atom-to-atom mapping, which further complicates its use. Consequently, a language model must learn not only the semantics and syntax of SMILES but also the intricate atom-to-atom mapping rules.[9,20] The introduction of the Condensed Graph of Reaction (CGR) addresses these challenges by encoding complex reaction information, including reactants, products, bond formation, and bond breaking, into a more concise and SMILES-like string representation.[10]

Despite the extensive literature on DL-powered chemical compounds generation, few paper tackled the de-novo generation of chemical reactions. The authors in[9] and[12] propose the use of Variational Autoencoders (VAEs)[21] for chemical reaction generation, employing SMILES strings to encode Condensed Graphs of Reactions (CGRs) and SMILES reaction strings, respectively.[22] They train a sequence-to-sequence autoencoder with bidirectional Long short-term memory (LSTM) layers[23] using reactions from the USPTO database and NIST chemical kinetics database[24] respectively (see Supplementary section S2 for background on VAEs and LSTMs). The models' latent spaces were sampled for new points around the latent areas of trained reactions which are then decoded into novel chemical reactions. Additionally, these two studies highlight the importance of expert analysis by validating the generated reactions with thermodynamic calculations. Similar work[10] also relies on CGR-encoded SMILES to train on autoregressive LSTM model and a Temporal Convolutional Network (TCN). Compared with training a LSTM model alone, training a combination of LSTM and TCN models improved the quality of generated reactions. The authors in[10] also highlight that different fine-tuning protocols significantly affect the generative capabilities of the trained model, which is particularly important when applying the model to small datasets through transfer learning. Recent work of[11] uses transformers, particularly the Transformer-XL architecture[25] to generate Heck coupling reactions.[11] The transformer model is trained on a dataset of Heck reactions, generating a total of 4717 reactions. Out of these, 2253 novel Heck reactions were validated by chemists. Authors in[26] rely on an autoregressive transformer-based models to represent the space of chemical reactions highlighting the ability of transformers to learn reactions types while being trained with unannotated chemical reactions. Although the trained model's auto-regressive capability in[26] is not evaluated (since chemical reactions generation is not discussed in the paper), yet a reaction fingerprint is extracted from the model's learned representation, which is more informative than traditional molecular fingerprints.

Evaluating the Quality of Machine-Generated Chemical Reactions

Evaluating chemical reactions generated by machine learning models is a persisting challenge with several approaches proposed to assess the validity, feasibility, novelty, and uniqueness of generated reactions. There is no standardized protocol to evaluate machine-generated reaction, with different approaches being proposed, each with their unique set of metrics and strategies to vet the experimental feasibility of these reactions.[9–12] An overview of the proposed techniques across recent works can be found in Table 1.

Validity

The authors in[9] employ two levels of validity assessment in their pipeline. First, on the generative model's level, they evaluate the accuracy of their model in reconstructing both training and validation data. To filter out unfeasible reactions, they develop checks for stoichiometric balancing and chemically infeasible transformations based on heuristics, such as preventing C-C bond cleavages or unbalanced hydrogens. Similarly, the authors in[12] also filter out unbalanced reactions. Both authors in[9] and[10] rely on RDKit[27] and CGRtools[28] toolkit to check for valence and aromaticity, ensuring that invalid CGR/SMILES or SMILES strings are discarded. In contrast, the authors in[11,12] consider a generated reaction to be valid if its reactants and products meet the criteria of the RDKit molecular parser.[27]

Feasibility

To assess the chemical feasibility of the generated reactions, the authors in[9,12] incorporate reaction enthalpy as a proxy for thermodynamic factors. In contrast, the authors in[11] adopt a more human-centric approach by involving 12 experimental chemists to provide feedback on the validity and feasibility of the reactions. Additionally, the authors in[11] conduct an in-depth analysis of the chemical reactions, focusing on several intra- and intermolecular factors such as regioselectivity, stereoselectivity, and chemoselectivity. Using these factors, they identify common issues, including chirality errors, carbon number errors, and reaction type errors. They also experimentally validated eight representative generated reactions to confirm the consistency of the experimental results with the proposed reactions.

Diversity, Uniqueness, and Novelty

To assess the diversity of generated chemical reactions, the authors in[10] calculate inter-similarity scores among generated reactions and compare them to those in the original dataset. The authors in[10–12] measure uniqueness by computing the fraction of unique reactions in their datasets, using CGR/SMILES encodings and reaction SMILES strings. The authors in[9] evaluate novelty by comparing hashed reaction signatures of reaction centers and their environments to known reaction databases, identifying five novel transformations among 13 that are not in the training data. Similarly, the authors in[10] categorize reaction centers based on their closest neighbors using hash codes, enabling the detection of novel reaction centers not categorized in the training data. In contrast, authors in[11,12] quantify novelty as the fraction of unique generated reaction SMILES strings absent from the training set.

To sum up, recent advancements in chemical reaction modeling leveraging deep learning have shown promising results. However, several key challenges remain unaddressed. First, existing solutions utilize different variations of chemical rule codifications, which presents a major challenge to evaluating machine-generated chemical reactions. Assessing the validity of these reactions without reverting to costly simulations or expert knowledge presents a significant obstacle. Second, the literature does not provide a streamlined procedure for reaction filtering and evaluation. Depending on the input format (e.g., SMILES, CGR), defining and implementing evaluation metrics becomes inconsistent. Most methods[9–12] filter out only the reactions whose compounds do not follow valid SMILES syntax, focusing on molecular fragments rather than overall reaction. While additional heuristics are proposed,[9] these are not incorporated into a standardized and reproducible process, and were only applied for CGR strings. Similarly, measuring novelty in CGR strings differs significantly from in SMILES, as novel reaction centers are easier to identify in CGR strings. This highlights the need for data-type-agnostic metrics that can be applied whether the model was trained with reaction SMILES or CGR strings. Third, existing methods lack standardized evaluation protocols: where assessing the quality of generated reactions ranges from using fully automated techniques[10] to requiring varying levels of expert involvement (e.g., analysis of reaction centers,[11] or reaction enthalpies,[9,12] or experimental validation[11]). This diversity complicates the benchmarking and comparison of different models for de-novo chemical reaction synthesis.

Methods

To address the limitations mentioned earlier, we introduce the **ChemRxnSAGE** framework: a novel framework for the **Chem**ical Reaction (**Rxn**) **S**ystematic **A**ssessment of **G**eneration and **E**valuation. The ChemRxnSAGE framework generates chemical reactions using an integrated DL module combining Long-Short Term Memory (LSTMs) and Variational Autoencoders (VAEs). It combines the reactants and products SMILES into one reaction SMILES format, adding the necessary tokens to delineate chemical sequences. It assesses the quality and diversity of chemical reactions through a set of metrics and filters developed in collaboration with domain experts, providing a comprehensive view of the model's capability to generate reactions, and capturing their inherent diversity, similarity, and faithfulness to a reference dataset. ChemRxnSAGE is highly extensible, supporting any generative models, reaction data formats, and new evaluation metrics in a generalized workflow. ChemRxnSAGE's overall architecture is shown in Figure 1. It consists of four main modules: i) Data Preprocessing, ii) Deep Learning-based Generation, iii) Chemical Validity Filtering and iv) Quality Evaluation.

Table 1: Comparison of Chemical Reaction Evaluation Frameworks

Framework	**Chemical Feasibility Checks**	**Reaction Enthalpy Evaluation**	**Expert Feedback**	**Diversity**	**Uniqueness**	**Novelty**	**Type of Data**
Bort et al.[9]	✓	✓				✓	CGR/SMILES
Buin et al.[10]	✓			✓	✓	✓	CGR/SMILES
Wang et al.[11]	✓		✓		✓	✓	Reaction SMILES Strings
Tempke & Musho[12]	✓	✓			✓	✓	Reaction SMILES Strings
Our Method	✓	✓	✓	✓	✓	✓	Reaction SMILES Strings

Data Preprocessing Module

We use the dataset prepared for retrosynthesis in.[29] It is based on a filtered patent data set, derived from an open-source patent database first coined in,[30] containing 50,000 reactions in SMILES format and classified into 10 reactions classes. The dataset is divided into train, validation, and test sets, following a 80-10-10 split. We then preprocess all three datasets to eliminate all reagents while keeping the reactants and products in the reactions canonicalized. We also split the reactions with more than one product into separate reactions each with only one product. To prepare the dataset for training, we combine the reactants and products SMILES into one reaction SMILES format. We add tokens at both the start and end of the sequences to indicate the beginning (BOS) and ending (EOS) of a sentence, as shown in Figure 2. Next, we remove any sequences longer than 200 tokens. The remaining sequences are then tokenized and padded, using a vocabulary size of 56 (see Supplementary Table 1 for the overview of parameters used in database generation and model training, with corresponding rationales.). The reaction classes distribution in the training dataset is displayed in Figure 3. We utilize RDKit[27] to transform the chemical reactions into atom-pair fingerprints,[31] capturing their structural and chemical properties. To analyze and compare the distribution of these reactions to the generated reactions, we apply Uniform Manifold Approximation and Projection (UMAP) to project them in a lower-dimensional space. The resulting 2D visualization is presented in Figure 4.

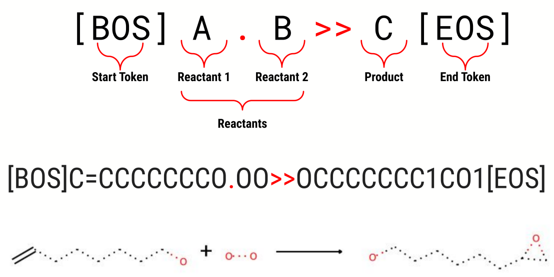


Figure 2: Reaction format with Example Reaction Sequence

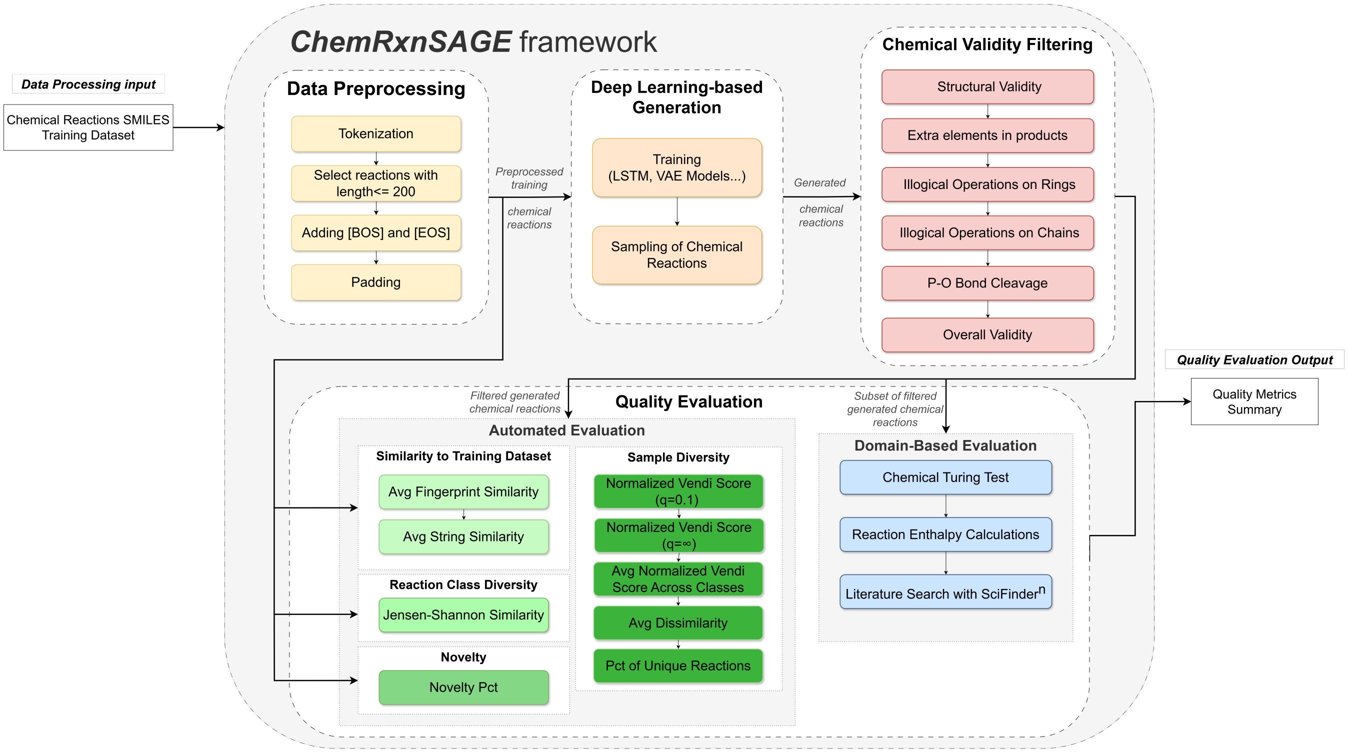

Data Processing input
Chemical Reactions SMILES Training Dataset
ChemRxnSAGE framework
Data Preprocessing
Tokenization
Select reactions with length<= 200
Adding [BOS] and [EOS]
Padding
Preprocessed training
chemical reactions
Deep Learning-based Generation
Training (LSTM, VAE Models...)
Sampling of Chemical Reactions
Generated
chemical reactions
Chemical Validity Filtering
Structural Validity
Extra elements in products
Illogical Operations on Rings
Illogical Operations on Chains
P-O Bond Cleavage
Overall Validity
Quality Evaluation Output
Quality Metrics Summary
Filtered generated chemical reactions
Quality Evaluation
Subset of filtered generated chemical reactions
Automated Evaluation
Similarity to Training Dataset
Avg Fingerprint Similarity
Avg String Similarity
Reaction Class Diversity
Jensen-Shannon Similarity
Novelty
Novelty Pct
Sample Diversity
Normalized Vendi Score (q=0.1)
Normalized Vendi Score (q=∞)
Avg Normalized Vendi Score Across Classes
Avg Dissimilarity
Pct of Unique Reactions
Domain-Based Evaluation
Chemical Turing Test
Reaction Enthalpy Calculations
Literature Search with SciFinderⁿ

Figure 1: Overview of the ChemRxnSAGE Framework

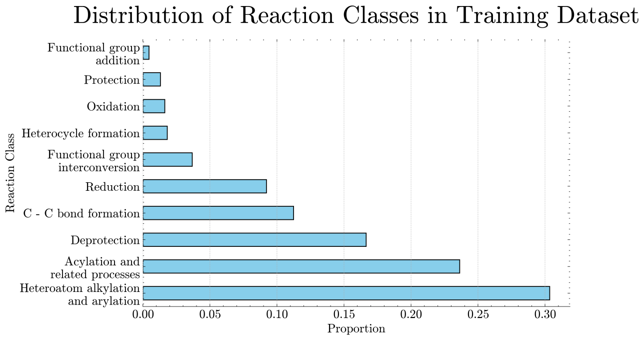


Figure 3: Training Dataset Distribution

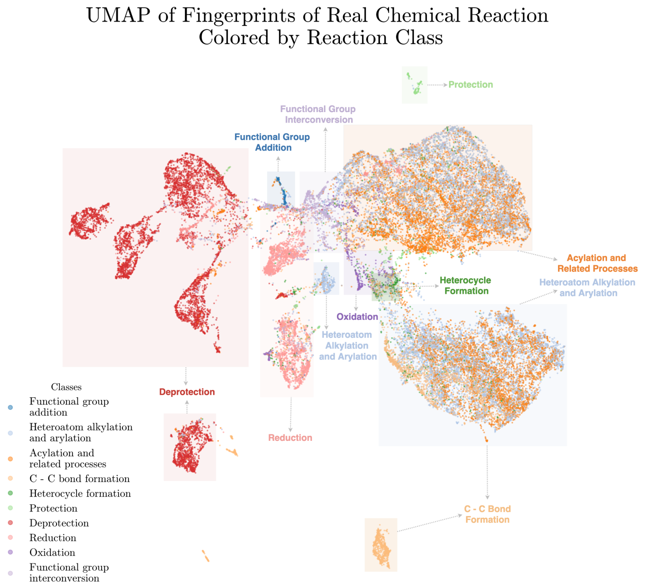


Figure 4: UMAP of Fingerprints of Training Dataset Colored by Reaction Classes

Deep Learning-based Generation Module

In this work, we will use two DL architectures for our experiments: LSTM models and VAEs (with an LSTM backbone). We rely on these two architectures due to their proven effectiveness in low-data settings and computational efficiency, making them well-suited for rapid experimentation. While diffusion models have demonstrated strong performance in molecular generation,[32,33] they typically require larger datasets and greater computational resources.[34] Importantly, ChemRxnSAGE is model-agnostic: the framework is designed to readily integrate diffusion-based or transformer-based architectures in future work, and these extensions are part of our planned roadmap.

We train a Long-Short Term Memory (LSTM) auto-regressive model to learn chemical reaction sequences. The model is optimized using Maximum-Likelihood Estimation (MLE) with a categorical cross-entropy loss function. We present four variations of the LSTM model, each with an increasing number of layers, 256 hidden units per layer, and a dropout rate of 0.5. The hyperparameters for these models are listed in Table 2. We train all four variations for 200 epochs using the Adam optimizer,[35] recording both training and validation losses throughout the process. We present also two variations of the VAE model: a vanilla VAE model ($\text{VAE}_{\text{vanilla}}$) and a VAE with KL cost annealing/ linear β warmup ($\text{VAE}_{\beta\ \text{warmup}}$) both inspired by work of[36] (see detailed overview of hyperparameters used in Table 2). Both models use an LSTM encoder-decoder architecture, and are trained for 200 epochs using Maximum Likelihood Estimation (MLE) with an SGD optimizer for cross-entropy loss. For the second model, a variable weight β is added to the KL loss term and is increased linearly from 0 to 1 over the first 50 epochs and then it is kept constant over the remaining training epochs. Ramping β up from 0 to 1 over the first 50 of 200 epochs allows the model to first learn meaningful latent codes before strong KL regularization sets in, therefore avoiding posterior collapse and aligning with common warm-up heuristics in literature.[36] To stabilize models' optimization, the learning rate decays with each iteration if the current loss exceeds the best loss achieved so far, continuing until the learning rate reaches its minimum value. For both the LSTM and VAE models, the loss function is configured to ignores padding tokens, as these are just placeholders and not real data. Ignoring padding tokens in the loss function is a standard practice in sequence modeling to ensure learning focuses only on meaningful data, improving training efficiency and accuracy.

Chemical Validity Filtering Module

Verifying the chemical validity of the generated chemical reactions is a challenging task. To our knowledge, there are no set rules that chemists follow to ensure that a generated reaction at hand can occur naturally. Instead, organic and synthetic chemists usually rely on their experience, previous works, and chemical databases to look for similar reactions that might occur with similar functional groups. Trying to frame chemical validity as a supervised ML problem is equally hard since the literature only covers "positive" or chemically valid reactions and does not consider "negative" or chemically invalid reactions meaningful to publish. Therefore, in this paper we propose an alternative way to circumvent these issues by proposing several heuristics designed with the help of a domain expert in synthetic and organic chemistry, with only the reactions that pass all filters being selected for the next evaluation steps:

- **Structural Validity ($\mathbf{F_{structural}}$)**: Developed to filter out chemical reactions that do not follow the format defined in Figure 2. It will then evaluate the SMILES syntax of the reactants and products using the RDKit molecular parser and will eliminate compounds not meeting the RDKit criteria.[27] Only the reactions meeting the structural validity criteria will be evaluated with the following filters.

- **Illogical usage of elements in products ($\mathbf{F_{illogical\text{-}use}}$)**: Eliminates all reactions that introduce elements that do not exist in the reactants to the products.

Table 2: Hyperparameters for training the LSTM and VAE models

Hyper-parameters	**$\mathrm{LSTM}_{\mathrm{L=1}}$**	**$\mathrm{LSTM}_{\mathrm{L=2}}$**	**$\mathrm{LSTM}_{\mathrm{L=3}}$**	**$\mathrm{LSTM}_{\mathrm{L=4}}$**	**$\mathrm{VAE}_{\mathrm{vanilla}}$ (enc/dec)**	**$\mathrm{VAE}_{\beta\ \mathrm{warmup}}$ (enc/dec)**
Batch size	64					
Sequence length	200					
Vocabulary size	56					
Number of LSTM layers (L)	1	2	3	4	1/1	1/1
Dropout probability	0.5	0.5	0.5	0.5	0.5/0.5	0.5/0.5
Hidden layer size	256	256	256	256	1024/1024	1024/1024
LSTM embedding size	64	64	64	64	512/512	512/512
Nb of β warm-up epochs	-	-	-	-	0	50

- **Illogical Operations on Rings ($\mathbf{F}_{\mathbf{rings}}$)**: Eliminate all reactions that contribute to:
  - Addition of an atom into a ring,
  - Replacement of an atom in a ring while the ring stays of the same size,
  - Transformations that lead to addition/removal of atom(s) in a ring,
  - Transformations that lead to addition/removal of carbon atom(s) in a ring.
- **Illogical Operations on Chains ($\mathbf{F}_{\mathbf{chains}}$)**: Eliminates reactions that contribute to:
  - Transformations that lead to addition/removal of atom(s) in a chain
  - Transformations that lead to addition/removal carbon(s) in a chain
- **Eliminating reactions allowing Phosphorus-Oxygen bond cleavage ($\mathbf{F}_{\mathbf{PO\text{-}bond}}$)**: Eliminates all reactions that allows the cleavage of bonds between Phosphorus and Oxygen in the reactants.

Quality Evaluation Module

The evaluation of the quality of the generated data is divided into two parts: i) an automated validation process using in-house developed metrics and heuristics, and ii) an expert evaluation process in the form of a ”chemical” Turing test. The quality evaluation module's overall architecture is shown in Figure 1.

Automated Evaluation

We introduce several metrics that allow the comparison of models and the quality and diversity of the generated reactions, namely: i) similarity to the training dataset, ii) reaction class diversity, and iii) sample diversity. The framework is modular, extensible and can integrate further metrics as needed.

Table 3: Metrics Equations

$\mathrm{FpSim}(\hat{Y}_{\mathrm{fp}}, Y_{\mathrm{fp}}) = 1 - \frac{1}{n} \sum_{\hat{y}_{\mathrm{fp}} \in \hat{Y}_{\mathrm{fp}}} \min_{y_{\mathrm{fp}} \in Y_{\mathrm{fp}}} \mathrm{Dist}_{\mathrm{Jaccard}}(y_{\mathrm{fp}}, \hat{y}_{\mathrm{fp}}) \quad \in [0,1]$ where $\mathrm{Dist}_{\mathrm{Jaccard}}(y_{\mathrm{fp}}, \hat{y}_{\mathrm{fp}}) = 1 - \frac{\lvert y_{\mathrm{fp}} \cap \hat{y}_{\mathrm{fp}} \rvert}{\lvert y_{\mathrm{fp}} \cup \hat{y}_{\mathrm{fp}} \rvert}$ and $\hat{Y}_{\mathrm{fp}}$ and Y_{fp} are the sets of generated and training reactions fingerprints respectively and n is the number of generated chemical reactions. **a.** Average Similarity of Generated Reaction Fingerprints to Training Dataset Representatives (FpSim)
$\mathrm{StrSim}(\hat{Y}, Y) = 1 - \frac{1}{n} \sum_{\hat{y} \in \hat{Y}} \min_{y \in Y} \mathrm{Dist}_{\mathrm{Cosine}}(y, \hat{y}) \quad \in [0,1]$ where $\mathrm{Dist}_{\mathrm{Cosine}}(y, \hat{y}) = 1 - \frac{y \cdot \hat{y}}{\lvert y \rvert \lvert \hat{y} \rvert}$ and $\hat{Y}$ and Y are the sets of generated and training reactions SMILES respectively and n is the number of generated chemical reactions. **b.** Average Similarity of Generated Reaction SMILES to Training Dataset Representatives (StrSim)
$\mathrm{JSS}(P \Vert \hat{P}) = 1 - \sqrt{\mathrm{JSD}(P \Vert \hat{P})} \in [0,1]$ where $\mathrm{JSD}(P \Vert \hat{P}) = \frac{D(P \Vert M) + D(\hat{P} \Vert M)}{2}$ Given that $C = \{c_1, c_2, \ldots, c_{10}\}$ is the set of reaction classes, P is a vector expressing the probability distribution of reactions belonging to each class, where each element $P(c_i)$ denotes the probability of a sample belonging to class c_i in the training dataset. For the generated set of reactions, the distribution $\hat{P}$ is defined similarly. D is the Kullback–Leibler divergence and M is a mixture distribution of P and $\hat{P}$. **c.** Jensen-Shannon Similarity Between Generated and Training Reaction Class Distributions (JSS)
$\mathrm{IntDiv}(\hat{y}_1 ... \hat{y}_n) = 1 - \frac{1}{n^2} \sum_{i,j \in \hat{Y}} \mathrm{Jaccard}(\hat{y}_i, \hat{y}_j) \in [0,1]$ **d.** Average Dissimilarity between Generated Reaction Fingerprints (IntDiv)
$NVS_{\mathrm{class}}(\hat{Y}_{\mathrm{fp}}, Y_{\mathrm{fp}}) = \frac{1}{\lvert C \rvert} \sum_{c \in C} \frac{\mathrm{VS}(\hat{y}_{\mathrm{fp},c}, y_{\mathrm{fp},c})}{n_c} \in [0,1]$ where VS is the Vendi Score with order $q = 0.1$ and n_c is the number of reactions in a reaction class c **e.** Average Normalized Vendi Score Across Reaction Classes ($\mathrm{NVS}_{\mathrm{class}}$)

Similarity to the Training Dataset To understand the similarity of the generated reactions with respect to the original dataset on both syntactical and structural levels, we develop two similarity metrics: i) StrSim and ii) FpSim. The former measures the average similarity of the generated reactions to those in the training dataset based on their SMILES strings (StrSim), while the latter measures the average similarity between their corresponding reaction fingerprints (FpSim). Reactions are encoded as *difference fingerprints* of length 2048, computed by subtracting the fingerprint of the reactants from that of the products. Molecules are represented using atom-pair fingerprints,[31] in which atoms are characterized by atomic number, number of π electrons, atom degree, and, optionally, chirality. To perform the similarity assessment, representative reactions are selected from each of the ten reaction classes. The top 5% representatives of the ten reaction classes are selected using the k-medoids algorithm,[37] a proportion empirically chosen to balance coverage of chemical diversity within each class and computational efficiency during similarity evaluation.

To compute StrSim, we calculate the average cosine similarity between the generated reactions and the selected representatives based on their SMILES encoding (cf. Table 3). Cosine similarity is a widely used metric in natural language processing and information retrieval for comparing text representations due to its effectiveness in measuring angular similarity in high-dimensional vector spaces, independent of vector magnitude.[38] This property is valuable for comparing tokenized SMILES strings, which represent chemical reactions as sequences of discrete symbols. Moreover, cosine similarity has been successfully applied to several chemical contexts in attention-based models including reaction classification and similarity assessment,[39,40] supporting its relevance for this task. For computing FpSim, the average Jaccard (Tanimoto) similarity is used (cf. Table 3). The Jaccard index is widely recognized and validated in cheminformatics[41–43] for its effectiveness in capturing substructural similarity by quantifying the overlap between molecular fingerprints. A higher StrSim or FpSim indicates stronger resemblance between generated reactions and the representative reactions from the training data. A model can however achieve FpSim or StrSim scores of 1 by generating reactions very similar to these representatives, indicating the model possibly overfitting on the representative set and consequently failing to produce diverse reactions within each reaction class.

Reaction Class Diversity The training dataset, as previously shown in Figure 3, includes a diverse yet imbalanced set of reactions from ten reaction classes. It is crucial to understand whether the model is able of generating a similarly diverse set of reactions covering all the ten reaction classes. However, the generated reactions are not labeled with a reaction class, therefore an auxiliary neural classifier was trained to classify the reactions based on their classes. To solve this problem, we trained a fully-connected neural networks model on the reaction fingerprints using the Adam optimizer[35] and Cross-Entropy loss, with overfitting monitored on a validation set (See Supplementary Table 2 for details about the model design and training hyperparameters). The weights of the neural classifier achieving the lowest validation loss and least overfitting are saved and used to predict the reaction classes of all generated reactions. To measure the diversity across all classes, we use the Jensen-Shannon Divergence (JSD) to compare the reaction classes distribution of the generated dataset to the training dataset.[44] The JSD between two probability distributions P and $\hat{P}$ is shown in Equation c in Table 3.The JSD cannot be used as a distance metric since it does not satisfy the triangle inequality, while its square root does.[45] We then convert the square root of the JSD into the Jensen-Shannon Similarity (JSS) index. The JSS is bounded between $[0, 1]$, with higher values indicating greater diversity in the generated reactions and a distribution more similar to the training dataset.

Sample Diversity Measuring the diversity of the reactions without relying on a reference might provide more insights on the model capacity to generalize beyond the training dataset. A model that produces samples highly similar to those seen during training may be at risk memorization, either by generating the same reactions in the training dataset or duplicates of few reactions (mode collapse). To measure the intrinsic diversity of the generated reactions and identify mode collapse, we use the Vendi Score,[46,47] a metric inspired from quantum mechanism and ecology that provides a reference-free measure of the diversity of a generated dataset. It is calculated as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix. It estimates the number of effective unique samples in a dataset and is more sensitive to the increases in the number of modes compared to average similarity of samples in the generated dataset.

We next propose five metrics capable together of capturing an overview of the generated samples diversity:

- **Percentage of Unique Reactions (UniqPct)** The simplest approach quantifies the proportion of distinct reactions. A higher uniqueness score indicates that the model generates fewer duplicate and can produce a varied dataset with less redundancy.

- **Average Dissimilarity/Diversity between Generated Reaction Fingerprints (IntDiv)**: Another common method to measure the average dissimilarity (or diversity) between compounds using their molecular fingerprints.[48] For all generated reactions, their corresponding atom-pair fingerprints are generated and the average of the Jaccard distance of all samples to each other are calculated (see Equation d in Table 3). A higher IntDiv score indicates that the generated reactions fingerprints are less similar to each other and therefore these reactions are more diverse.

- **Normalized Vendi Score q=0.1 ($\mathbf{NVS_{q=0.1}}$)**: We used the Vendi Score with a small order $q = 0.1$ to measure diversity with high sensitivity to rarer classes. A Vendi Score with order of $q = 0.1$ would have a higher value when less common classes of the generated dataset are more diverse. We also normalize the obtained Vendi Score based on the size of the generated dataset to get a better overview on the percentage of effective unique samples across the dataset.

- **Normalized Vendi Score q=∞ ($\mathbf{NVS_{q=\infty}}$)**: We use the Vendi Score with a infinite order to measure duplication and memorization indirectly. A high Vendi Score of order q=∞ would suggest that the model might be generating large groups of very similar reactions (reflecting the possibility that it might be memorizing the dataset[47]). We also normalize the obtained Vendi Score based on the size of the generated dataset to get a better overview on the percentage of effective unique samples across the dataset.

- **Average Normalized Vendi Score Across Classes ($\mathbf{NVS_{class}}$)**: The Vendi Score ($q = 1$) is reported to fail in datasets with strong class imbalance,[49] a common phenomenon in many ML settings including de-novo reaction synthesis. We use the Vendi Score with order $q = 1$ instead as a proxy for diversity within each reaction class. We normalize the Vendi Score calculated for every reaction class and then average all scores to get a more granular overview of the diversity within each reaction class. The average Normalized Vendi Score per Class ($\mathrm{NVS_{class}}$) would represent the average percentage of unique elements across all reaction classes (Equation e in Table 3). A higher score would mean that the model is generating diverse reactions across all reaction classes.

Novelty To evaluate the model's ability to produce new and previously unseen reactions during the training procedure, we propose the **Novelty Percentage (NovPct)**, which is the percentage of reactions in the generated dataset that are not found in the training dataset.

Domain-Based Evaluation

After training all models, we sample them to generate a dataset of chemical reactions matching the size of the original training dataset. We then evaluate these generated datasets using the previously mentioned metrics and filters. We choose the top three models for human evaluation based on their performance metrics and the highest percentage of reactions passing the filters .

Chemical Turing Test We develop a "chemical" Turing test to understand the quality of the generated reactions as assessed by domain experts versus real reactions inspired by previous works.[9,11] Originally, the Turing test is used to test whether machine-generated language can fool a human reader as if it is a human-written one. In this work, we adapt the Turing test to chemical reactions generation to understand i) if the machine-generated reactions are confused to be human-created, and ii) if these reactions are actually chemically valid and can occur naturally. To keep the evaluation manageable, we select only 10 reactions each representing our reaction type per model, in addition to 10 reactions from the training dataset, which are then shuffled, resulting in a total of 40 chemical reactions (See Supplementary Information for the complete chemical Turing test). Several domain experts were invited to fill out the survey. They have no previous knowledge of the differences between the reactions and were encouraged to analyze the reactions thoroughly. Understanding whether the chemists can be fooled by the machine-generated reactions would provide insights into the quality of the machine-generated reactions, possibly showcasing the innovative power of ML algorithm in coming up with novel chemical reactions. This limited sample size is not intended to provide statistically conclusive results but rather offers a complementary, human-centered evaluation of generation quality and helps identify representative reactions for subsequent chemical analysis.

Reaction Enthalpy Calculations Out of the reactions selected for the chemical Turing test, six reactions, two per model, that were rated as valid by the majority of experts are shortlisted for further analysis. The structures in the reactions are updated and the equations are balanced for atom stoichiometry. The feasibility of the reactions is determined by thermodynamic analysis, particularly reaction enthalpy (ΔH). ΔH is tackled using density functional theory (DFT) computations. DFT computations are carried out using GAUSSIAN09 and GAUSSIAN16.[50] The proposed reactants and products of the new reactions are subjected to energy minimized using MM2 force field.[51,52] Then, the geometries are optimized using B3LYP/6-311G(d,p) level of theory. The lowest energy stationary point is verified with frequency calculations. A single point calculation for all starting materials and products at the same level of theory is used to compute ΔH for the reactions.[9]

Literature Search with Scifinder$^{\mathbf{n}}$[53] To check whether the reactions have been previously reported in the literature, all reaction are searched against the SciFinder$^{\mathrm{n}}$ reaction database.[53]

Results

Experimental Data and Set-up

All experiments using the ChemRxnSAGE Framwork, spanning training, generation, and evaluation, were performed on a system equipped with an NVIDIA A40 GPU, dual-socket Intel Xeon Gold 6438Y+ CPUs (32 cores per socket, 2 threads per core), and 1 TB of system memory. The dataset comprised 50,000 chemical reactions, split into training, validation, and test sets following an 80-10-10 ratio, resulting in 40,000, 5,000, and 5,000 reactions, respectively. After preprocessing, the reactions were reduced to 39,555 (train), 4,935 (validation), and 4,956 (test) (see Supplementary Table 1). For reaction generation, each model was run across five random seeds (0, 42, 250, 350, and 1000), and the results were averaged across these runs. All metrics were very stable with low variation over the five runs (see Supplementary Tables 4,5 and 6 for standard deviation measurements for all metrics). Model losses and validity metrics were tracked throughout training, with model weights saved at the end of each epoch (See Supplementary Figures 3, 4, 5, 6, 7 and 8). The weights corresponding to the epoch with the highest overall validity were used for evaluation. Benchmarking of chemical reactions generation and evaluation was conducted using the hardware and software environment detailed earlier (See Supplementary Figure 9). Code for model training, evaluation and visualization of results is available online[a].

Automated Evaluation

The classifier trained to predict the chemical reaction was evaluated on a separate test set and achieved a 97% accuracy (see Supplementary Table 3). This high performance aligns with the observation that most reaction classes are well separated in the latent space, however classes such as Acylation and related processes, Heteroatom Alkylation and Arylation, and C–C bond formation exhibit greater overlap (Figure 4). The classifier was then used to predict the reaction classes of the reactions generated by the models. The generative models were trained and evaluated using previously proposed metrics and filters, with results presented in Tables 4, 5 and 6. As shown in Table 4, the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ model emerged as the best-performing model, achieving a $\mathrm{F}_{\text{overall}}$ score of 37.39%, closely followed by the $\mathrm{LSTM}_{\mathrm{L}=3}$ model at 36.45% and the $\mathrm{VAE}_{\text{vanilla}}$ model at 32.81%. The $\mathrm{VAE}_{\beta\ \text{warm-up}}$ model demonstrated robustness in generating chemical reactions that passed $\mathrm{F}_{\text{rings}}$ and $\mathrm{F}_{\text{PO-bond}}$ with success rates of 100% and 99.65%, respectively, effectively avoiding illogical operations on rings and preserving P–O bonds.

The $\mathrm{VAE}_{\beta\ \text{warm-up}}$ model, unlike the other models, struggled the least with generating reactions that passed the $\mathrm{F}_{\text{illogical-use}}$ filter (86.46%) but performed poorly on $\mathrm{F}_{\text{chains}}$ (48.92%).

When comparing the generated chemical reactions of the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ model and $\mathrm{LSTM}_{\mathrm{L}=1}$ before and after applying the filters to the training dataset, as visualized in Figures 6a and 6b, we note that the filters remove many outlier reactions that are not similar to the training dataset. Across all reaction classes in Figure 5, the filtering procedure of the generated reactions by $\mathrm{VAE}_{\beta\ \text{warm-up}}$ resulted in a reduction of 40% to 70% of the reactions (Figure 5c), with no clear correlation with the reaction class abundance. The filtering procedure made the reaction distribution farther from the original reaction class distributions, except for the Deprotection reaction class, where the filtering helped refine the reactions to follow the training dataset distribution.

Regarding similarity and diversity results (Table 5), the $\mathrm{LSTM}_{\mathrm{L}=3}$ model achieved the highest score for JSS, followed closely by the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ model. All the models exhibited similarly high values for StrSim around 0.86 and very high NovPct around 99%, while for FpSim, the $\mathrm{LSTM}_{\mathrm{L}=3}$ model demonstrated the highest score (0.1721), with the $\mathrm{LSTM}_{\mathrm{L}=2}$ model coming second (0.1685) and $\mathrm{VAE}_{\text{vanilla}}$ third (0.1665). Regarding the diversity results (Table 6), all models had very high scores for IntDiv around 0.97 as well a high number of unique reactions (UniqPct) with values all around 99%. For $\mathrm{NVS}_{\mathrm{q}=0.1}$, $\mathrm{LSTM}_{\mathrm{L}=1}$ ranked the highest score (95.96%) followed by $\mathrm{LSTM}_{\mathrm{L}=4}$ and $\mathrm{VAE}_{\beta\ \text{warm-up}}$ (95.02% and 94.61% respectively). Similarly, for class diversity, $\mathrm{LSTM}_{\mathrm{L}=1}$ ranked the highest score for $\mathrm{NVS}_{\text{class}}$ (81.18%) followed by $\mathrm{LSTM}_{\mathrm{L}=4}$ (76.30%) and $\mathrm{VAE}_{\text{vanilla}}$ model (73.68%). Finally, for measuring duplication and memorization, $\mathrm{NVS}_{\mathrm{q}=\infty}$ highlights that $\mathrm{VAE}_{\beta\ \text{warm-up}}$ and $\mathrm{LSTM}_{\mathrm{L}=3}$ have the lowest scores (0.23% and 0.24% respectively) and therefore the lowest levels of memorization.

Compared with the $\mathrm{LSTM}_{\mathrm{L}=1}$ model, the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ model generated more chemical reactions that passed the filtering procedure (Figures 6a and 6b). Additionally, it produced

[a] `https://github.com/anisdismail/ChemRxnSAGE`

Table 4: Summary of filter metrics across generative models

Metrics Models	$\mathbf{F_{structural}}\uparrow$	$\mathbf{F_{illogical\text{-}use}}\uparrow$	$\mathbf{F_{rings}}\uparrow$	$\mathbf{F_{chains}}\uparrow$	$\mathbf{F_{PO\text{-}bond}}\uparrow$	$\mathbf{F_{overall}}\uparrow$
$\mathbf{LSTM_{L=1}}$	0.6720	0.5792	1.0000	0.4617	0.9934	0.1785
$\mathbf{LSTM_{L=2}}$	0.7298	0.7784	1.0000	0.5550	0.9957	0.3139
$\mathbf{LSTM_{L=3}}$	0.7843	0.7814	1.0000	0.5966	0.9968	0.3645
$\mathbf{LSTM_{L=4}}$	0.7923	0.6623	1.0000	0.4929	0.9953	0.2574
$\mathbf{VAE_{vanilla}}$	0.8793	0.7834	1.0000	0.4783	0.9959	0.3281
$\mathbf{VAE_{\beta\ warm\text{-}up}}$	0.8871	0.8646	1.0000	0.4892	0.9965	**0.3739**

Table 5: Summary of similarity, novelty and reaction classes diversity metrics across generative models

Metrics Models	**JSS**↑	**FpSim**↑	**StrSim**↑	**NovPct**↑
$\mathbf{LSTM_{L=1}}$	0.8128	0.1445	0.8693	0.9998
$\mathbf{LSTM_{L=2}}$	0.8728	0.1685	0.8653	0.9987
$\mathbf{LSTM_{L=3}}$	0.9019	0.1721	0.8650	0.9984
$\mathbf{LSTM_{L=4}}$	0.8518	0.1567	0.8646	0.9990
$\mathbf{VAE_{vanilla}}$	0.8668	0.1665	0.8646	0.9953
$\mathbf{VAE_{\beta\ warm\text{-}up}}$	0.8791	0.1652	0.8628	0.9959

Table 6: Summary of samples diversity metrics across generative models

Metrics Models	**IntDiv**↑	$\mathbf{NVS_{q=0.1}}\uparrow$	$\mathbf{NVS_{q=\infty}}\downarrow$	$\mathbf{NVS_{class}}\uparrow$	**UniqPct**↑
$\mathbf{LSTM_{L=1}}$	0.9699	0.9596	0.0044	0.8118	0.9993
$\mathbf{LSTM_{L=2}}$	0.9729	0.9406	0.0028	0.7168	0.9979
$\mathbf{LSTM_{L=3}}$	0.9735	0.9393	0.0024	0.6992	0.9980
$\mathbf{LSTM_{L=4}}$	0.9722	0.9502	0.0033	0.7630	0.9988
$\mathbf{VAE_{vanilla}}$	0.9723	0.9382	0.0026	0.7368	0.9978
$\mathbf{VAE_{\beta\ warm\text{-}up}}$	0.9727	0.9461	0.0023	0.7166	0.9981

a more diverse set of reactions, spanning a wider range of reaction classes and covering more areas of the training dataset's latent space (Figure 7).

When comparing the diversity of chemical reactions after filtering (Figure 8), we note while VAE$_{\beta \text{ warm-up}}$ had a higher JSS score than LSTM$_{\mathrm{L=1}}$, LSTM$_{\mathrm{L=1}}$ seems to have higher or equal proportions of its generated reactions belonging to rarer classes than VAE$_{\beta \text{ warm-up}}$, and generates lower proportions for the more dominant classes. Interestingly, LSTM$_{\mathrm{L=1}}$ records higher diversity scores across all classes than VAE$_{\beta \text{ warm-up}}$ (see Figure 9 and Supplementary Table 7).

Table 7: Percentages of reactions as reported in the survey for models and reference dataset

Source	**Invalid↓**	**Valid↑**	**Human↑**	**Machine↓**
Reference	31.67%	68.33%	63.33%	36.67%
VAE$_{\mathbf{vanilla}}$	75.00%	25.00%	43.33%	56.67%
LSTM$_{\mathbf{L=3}}$	58.33%	41.67%	**46.67%**	53.33%
VAE$_{\beta \text{ warm-up}}$	53.33%	**46.67%**	**46.67%**	53.33%

Table 8: Computation of reaction enthalpies (ΔH^{DFT}) at B3LYP/6-311G(d,p) level of theory calculations.

Reaction Nb	ΔH^{DFT} **kcal/mol** (298.15K, 1.00 Atm)
1	-19.01
2	-3.45
3	-24.24
4	-27.53
5	-8.86
6	-23.56

Domain-Based Evaluation

Both VAE models and the LSTM$_{\mathrm{L=3}}$ were selected for domain evaluation due to their superior scores for F$_{\mathrm{overall}}$. Six domain experts having a doctoral or master's degree in either Organic or Synthetic Chemistry were invited to

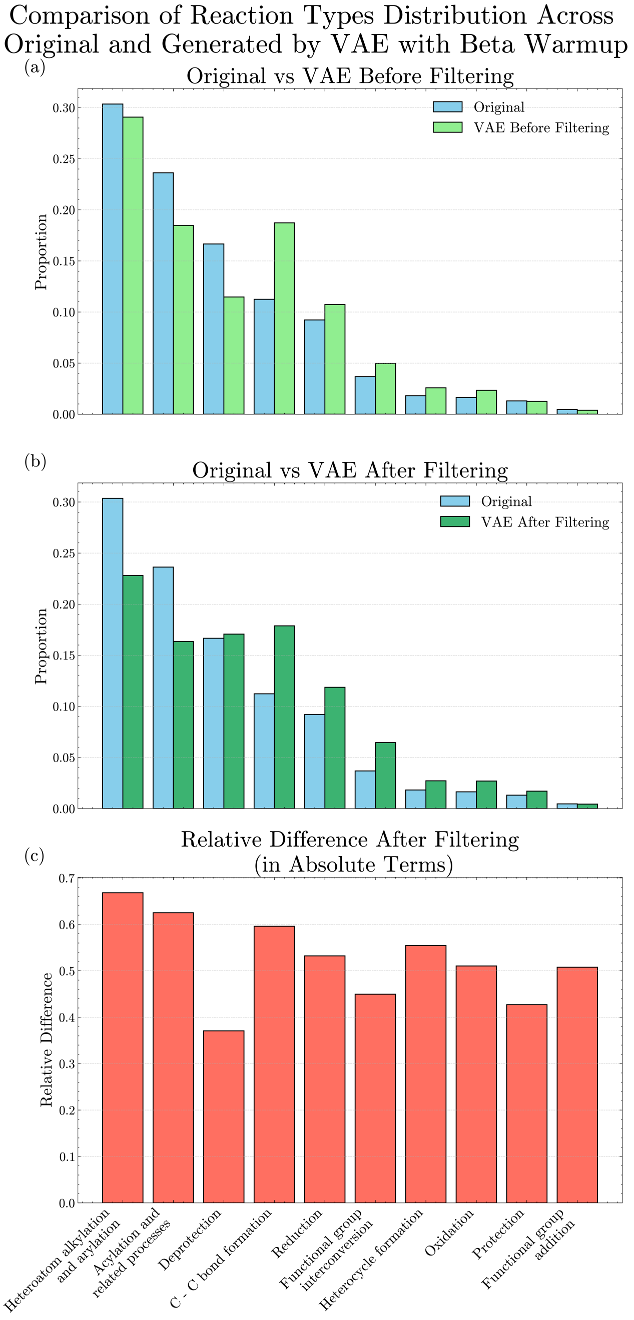
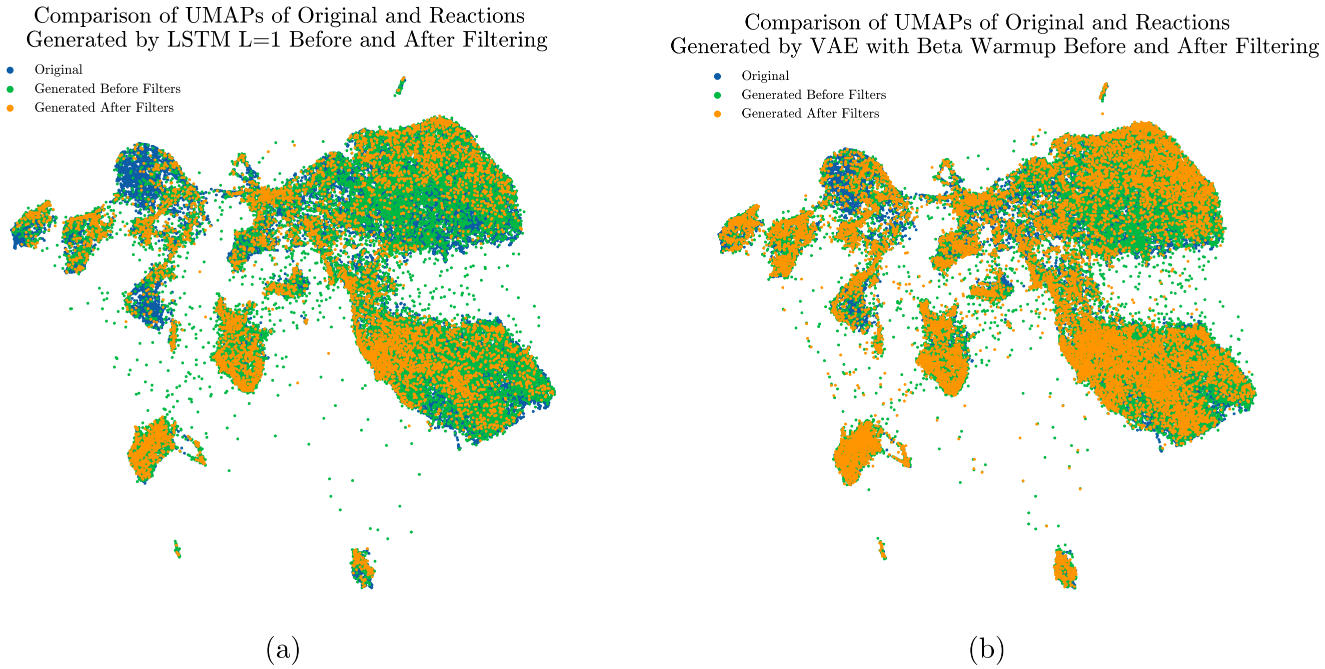


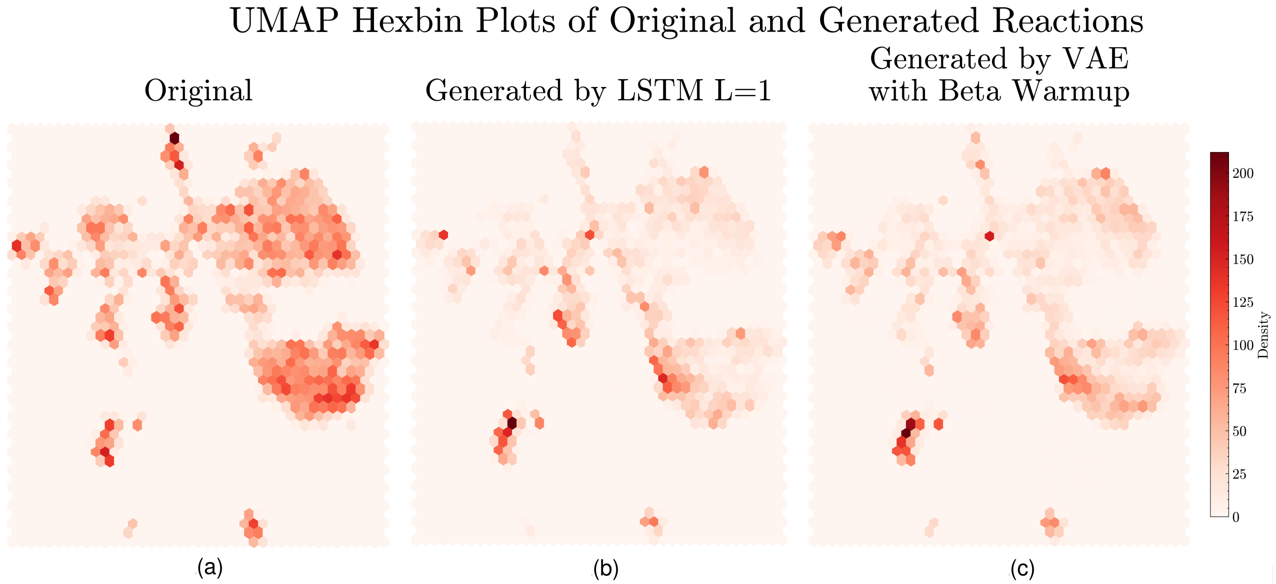
Figure 5: Distribution of Reaction Classes in Real and Generated Data. a) Original vs. VAE$_{\beta \text{ warm-up}}$ before filtering; b) Original vs. VAE$_{\beta \text{ warm-up}}$ after filtering; c) Absolute relative differences.



(a)

(b)

Figure 6: UMAP of reaction fingerprints before and after filtering. (a) $\text{LSTM}_{\text{L=1}}$, (b) $\text{VAE}_{\beta\text{ warm-up}}$.



(a)

(b)

(c)

Figure 7: Comparison of UMAP of Reaction Fingerprints Across a) Original, b) LSTM $_{\text{L=1}}$ and c) $\text{VAE}_{\beta\text{ warm-up}}$

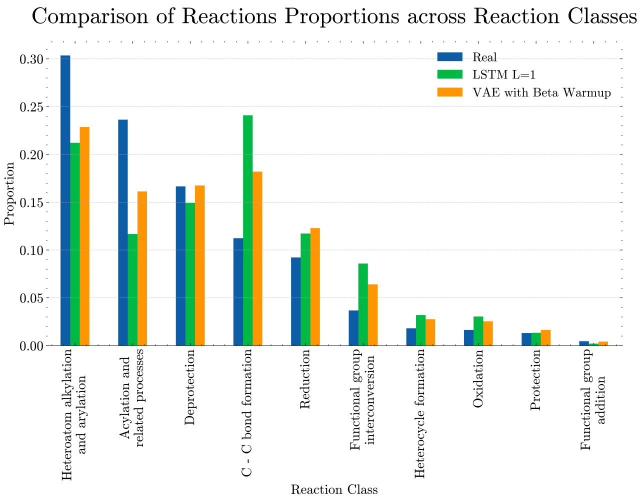


Figure 8: Proportion of reaction classes for $\mathrm{LSTM}_{\mathrm{L}=1}$ and $\mathrm{VAE}_{\beta\ \text{warm-up}}$, averaged across seeds.

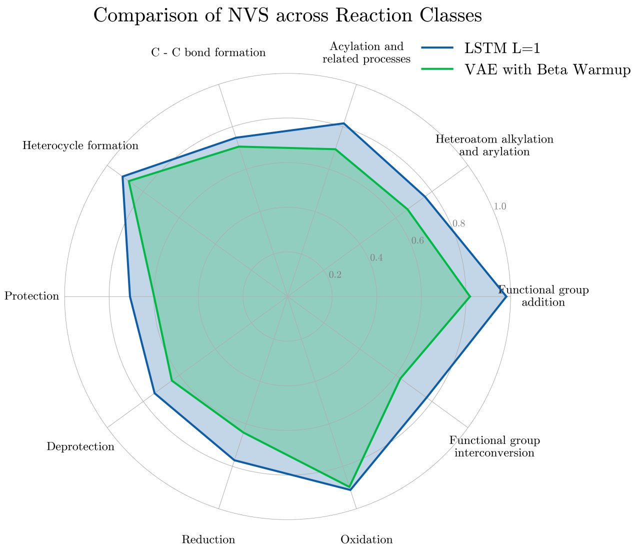


Figure 9: NVS comparison for $\mathrm{LSTM}_{\mathrm{L}=1}$ and $\mathrm{VAE}_{\beta\ \text{warm-up}}$, averaged across seeds.

complete the survey. The survey results (Table 7) revealed that the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ achieved the highest percentage of valid reactions (46.67%) and the closest resemblance to the human-generated dataset (46.67%). The survey results reinforced our previous observations that the $\mathrm{VAE}_{\text{vanilla}}$ achieved the best performance according to the surveyed chemists. 46.67% of the reactions generated by the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ are claimed to be chemically valid by the surveyed chemists, which showcases that the generative power of the $\mathrm{VAE}_{\beta\ \text{warm-up}}$. Not only the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ was able to generate valid reactions, but it also is the model that was able to fool the respondents to consider its reactions as human-generated. Although none of the models was able to beat the human score and pass the Turing test, it becomes clear that the $\mathrm{VAE}_{\beta\ \text{warm-up}}$ is the most promising candidate achieving 46.67% versus 63.33%, which is the smallest difference compared to the other models.

The reactions which were voted as valid by the majority of the respondents were selected for further analysis (Table 9). Regarding the feasibility of the reactions, a thermodynamic analysis, particularly reaction enthalpy (ΔH) was performed and the results are shown in Table 8. It is noted that all reactions 1-6 were found to be exothermic. Using SciFinder^{n} [53] search (Table 10), experimental procedures were found for the reactions 1 and 3. Only similar reactions were found in literature for reactions 2,4,5, and 6.

Discussion

In this work, we introduced the ChemRxn-SAGE framework, which enabled the analysis of several generative models for chemical reaction modeling across multiple dimensions. We used the framework to evaluate all models' performance in generating valid reactions and then assessed their novelty, diversity, and similarity relative to a reference dataset.

The VAE models demonstrated their robustness in generating chemical reactions that successfully passed filters with a high number of reactions overall, excelling in avoiding illogical operations on rings and the breaking of P-O bonds. The VAE models along with the $\mathrm{LSTM}_{\mathrm{L}=3}$ had the highest overall scores, which is expected as these models had higher capacities compared to the other tested models. However, challenges were observed in generating reactions that met filters related to atom and chain consistency. Notably, some generated reactions included products with atoms absent from the reactants, and chain operations frequently led to errors, such as the addition or removal of atoms. These issues can be at-

Table 9: The generated chemical reaction equations are non-balanced (”Before”) and balanced (”After”) to account for reactant(s) and product(s) of both sides and for the atom stoichiometry. Additional reagents or side products were included.

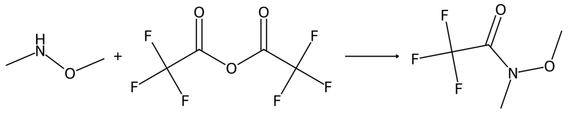
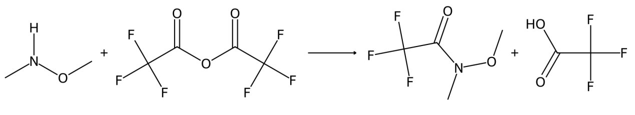
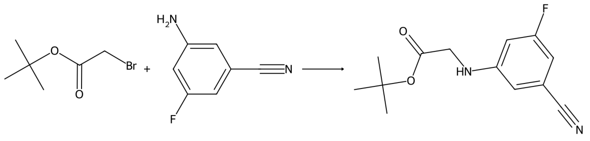
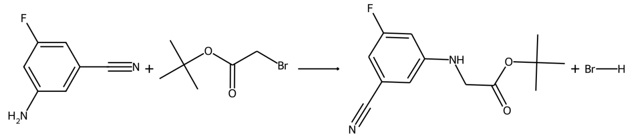
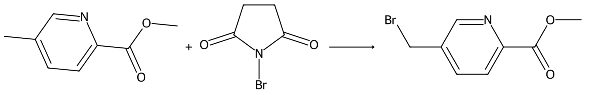
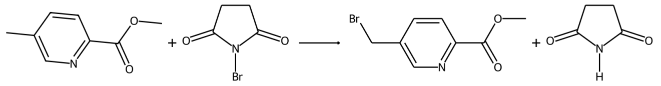
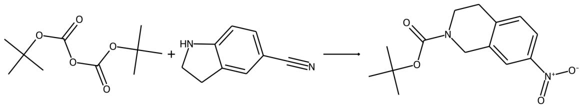
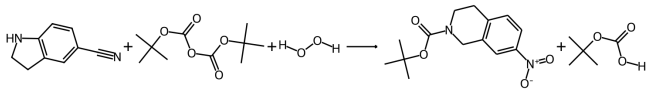
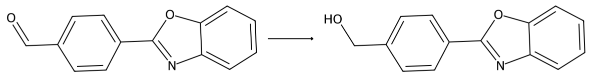
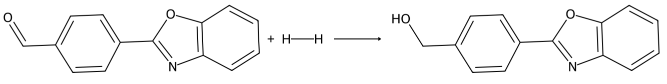
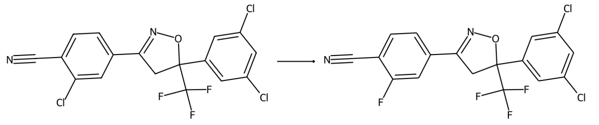
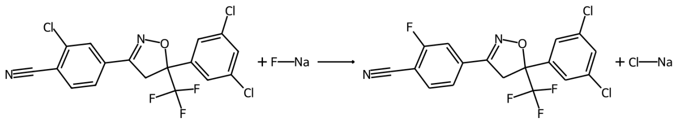
Reaction Nb	**Model**		**Reaction**
1	LSTM$_{L=4}$	Before	
		After	
2	LSTM$_{L=4}$	Before	
		After	
3	VAE$_{\text{vanilla}}$	Before	
		After	
4	VAE$_{\text{vanilla}}$	Before	
		After	
5	VAE$_{\beta\ \text{warm-up}}$	Before	
		After	
6	VAE$_{\beta\ \text{warm-up}}$	Before	
		After	

Table 10: SciFinder Search to identify experimentally reported reactions.

Reaction Number	SciFinder^{n} [53] Results
1	This acylation reaction was reported at a temperature of 0 °C in 95% yield.[54]
2	A reaction on this particular amine with a tert-butyl bromoacetate as starting materials is not found, however, the alkylation of primary arylamines with ethyl bromoacetate is reported at room temperature in 100% yield.[55]
3	The benzylic bromination of this particular substrate was reported in literature using azobisisobutyronitrile catalyst in carbon tetrachloride at 48 °C in 66% yield[56]
4	This transformation requires several steps and is not as such found as a one-step process in literature. Specifically, the nitrile needs to be removed and replaced by a nitro group on the adjacent carbon. The 5-membered ring of the indoline unit is expanded to a six-membered ring by insertion of a methylene (-CH2-) group to form 1,2,3,4-tetrahydroisoquinoline. The nitrogen of the indoline group is converted into a carbamate by an acylation reaction. This last step is reported in literature for related molecules in 87% yield at room temperature.[57]
5	The reaction is a reduction of an aromatic aldehyde, specifically benzaldehyde. This particular reaction on this particular substrate is not found in SciFinder^{n}.[53] The reduction can be accomplished by a range of reducing agents. An example of a reported reaction with H2 and catalyst is carried out at room temperature in near quantitative yield[58]
6	A reaction on this particular aryl halide as starting material is not found in literature, however, the nucleophilic aromatic substitution of an aryl chloride with anhydrous tetramethylammonium fluoride to prepare an aryl fluoride at 80 °C in 98% yield is reported using anhydrous tetramethylammonium fluoride.[59]

tributed to the limitations of the LSTM network used as a backbone for the VAE models, as its performance degrades with increased sequence length. For example, an LSTM network may struggle to "remember" all tokens in reactants, leading to inaccuracies in products. Similarly, tracking atoms within long chains, particularly those with numerous methyl and functional groups, presents difficulties. Rings, typically represented by shorter SMILES sequences, are less affected by these issues. Additionally, the inherent nature of SMILES strings being very long and lacking geometric and structural context makes training a model to meet such constraints even harder. Another challenge we faced is that the dataset we used lacked atom-to-atom mapping making training models that can identify reaction centers and evaluating the unmapped generated reactions even more challenging without analyzing their reaction centers. To address these limitations, authors in[10] and[9] incorporate CGR strings merging reactants and products into a single graph with atom-to-atom mapping and use dedicated SMILES encodings for CGRs.

Analyzing the validity of generated chemical reactions proved challenging, yet the procedure followed in this paper effectively assessed reaction quality. The initial distribution of generated reactions suffered from numerous outliers and potentially invalid reactions. Applying the developed filters effectively reduced outliers, aligning the filtered generated dataset's distribution more closely with the original dataset. Furthermore, the reaction class distribution of the filtered generated set highlights that the $\text{VAE}_{\beta \text{ warm-up}}$ struggled in generating reactions belonging to the *heteroatom alkylation and arylation*, *acylation and related processes*, and *C-C bond formation*, which required extensive filtering and shifting the reaction class distribution away from the original dataset distribution. These reaction classes typically involve bond cleavages around reaction centers that may occur adjacent to long carbon chains, consistent with our previous finding that models often struggle to generate reactions while maintaining chain consistency.

Our framework also helped shed some light on the diversity and similarity of the generated reactions to the training dataset. The $\mathrm{LSTM}_{\mathrm{L=1}}$ surprisingly achieved the lowest similarity to the training dataset's top representatives compared to the other models. Its inability to generate reactions similar to the top representatives of the training dataset could be related to the fact that this model could be possibly underparametrized. The $\mathrm{LSTM}_{\mathrm{L=3}}$ and VAE models had the highest FpSim scores and the highest similarity to the reaction class distribution of the training dataset. The trend generally increased as the capacity of the LSTM models increased, with the scores peaking for LSTM with 3 layers instead of 4, potentially highlighting that $\mathrm{LSTM}_{\mathrm{L=4}}$ is overfitting and at risk of higher memorization. On the other hand, The β warm-up seems to helped reduce posterior collapse allowing the VAE to learn informative latent representations of reactions and capture better chemical grammar than other models, as evidenced by the higher validity and similarity scores, and lower levels of memorization.

Due to the random sampling nature of reaction generation, we see that all models score very high for novelty percentages. Very high similarity scores were observed when comparing the string similarity of the models. We hypothesize that this may be attributed to the syntactical nature of reaction SMILES, as they inherently share numerous similar components, such as carbon bonds, ring structures, representation syntax, and padding. These shared elements could contribute to inflated similarity scores, potentially skewing the scores.

Different NVS metrics helped in debugging further the differences in diversity between all models. For instance, at $q = 0.1$, $\mathrm{LSTM}_{\mathrm{L=1}}$ surprisingly achieved the highest results, and not any of the VAE models. This trend may arise due to $\mathrm{LSTM}_{\mathrm{L=1}}$'s possibly generating reactions with higher intra-class variance, making the $\mathrm{NVS}_{\mathrm{q=0.1}}$ which is more sensitive to rarer classes score higher. $\mathrm{LSTM}_{\mathrm{L=1}}$'s higher dataset proportions for rarer classes explain its higher NVS at $q = 0.1$, while still retaining higher overall NVS scores than the $\mathrm{VAE}_{\beta\ \mathrm{warm\text{-}up}}$ across all classes. The $\mathrm{VAE}_{\beta\ \mathrm{warm\text{-}up}}$ might be overfitting to larger classes, struggling to generate diverse set of reactions across all classes especially rarer classes. Another possibility is that the learned latent space of the VAE may not be expressive enough, allowing it to succeed in generating diverse reactions with simpler transformations, but limiting its ability to handle more complex reactions.

The survey results indicate that respondents found it challenging to differentiate between machine-generated and human-generated reactions. Moreover, the machine-generated reactions selected for further analysis were not only thermodynamically feasible but contained potentially novel ones that were not previously reported in the literature. These findings align with our objective to promote a deeper understanding of reaction differences and, with the aid of generative models, enable more innovative approaches to analyzing patterns in existing reactions and designing novel ones. While the results of the survey confirm that the models of higher capacity achieve higher scores approaching those of human-generated reactions, a significant gap remains between machine and human generation. This highlights the need for further research to develop models better suited for generating de novo chemical reactions.

Several limitations and opportunities for future work emerged during this study. Our current models exhibit limitations in capturing long-range dependencies and subtle structural nuances in chemical reactions, leading to inaccuracies in reaction generation and reduced diversity. Our current models rely on SMILES representations, which do not retain 2D or 3D coordinates of individual atoms, contributing to errors in reaction mapping and generation. Filters applied in the current workflow can also produce false positives, which reduces their robustness and specificity. To address these issues, future work could explore expanding beyond SMILES to incorporate CGR representations and atom-to-atom mapping, enhancing structural and geometric understanding. Refining filters through improved atom mapping could then reduce false positives and improve the pipeline's reliability. Additionally, adopting more advanced model architectures, such as transformers[25] or diffusion models,[60] would increase the capacity to capture long-range dependencies and complex structural patterns. Leveraging larger datasets with more granular reaction type labels would further support generalization and diversity in generated reactions. Addressing these limitations would enable our framework to evolve and tackle more complex challenges, driving innovation in chemical reaction generation and evaluation.

Conclusion

Generating and evaluating chemical reactions remains a significant challenge, with limited research addressing this area. In this paper, we introduce **ChemRxnSAGE**, a standardized and extensible end-to-end DL-based framework for evaluating the quality, validity, and diversity of machine-generated chemical reactions. ChemRxnSAGE combines LSTMs and VAEs to generate new reactions, and provides an extensible battery of automated validity filters with quality metrics to systematically eliminate invalid reactions. The framework was tested with multiple LSTM and VAE models, considering expert feedback to assess the diversity, novelty, and similarity of generated reactions to reference datasets. A chemical "Turing test" involving domain experts, along with enthalpy calculations and literature comparisons via SciFinder$^{\text{n}}$,[53] validated the generated reactions. Results demonstrated that the VAE model with linear β warm-up training followed by the LSTM model with three layers consistently produced the most valid and diverse reactions across reaction classes, closely resembling real chemical reactions. The superior performance of VAEs over LSTMs can be attributed to their greater capacity and ability to learn a structured latent space, enabling better generalization to novel reactions. By combining automated analysis with expert evaluation, the framework bridges computational tools with domain-specific knowledge, facilitating both discovery and reproducibility. We hope that this work paves the way for the development of new algorithms, fostering innovation in chemical reaction generation and evaluation.

Acknowledgement The authors appreciate the participation of all chemists in the experimental assessment of the generated reactions. Specifically, we thank Dr. Bilal Kaafarani, Dr. Areej Merhi, Anthony Sokhn, Chadi Massoud, and Farah Abi Tayeh. The authors also thank Ms. Nour Ismail (Bachelor of Science in Biology from the Lebanese American University) for her support in drawing and preparing the chemical reactions for reaction enthalpy analysis, and Professor Alejandro Sifrim for providing computing resources to run the framework.

Data and Software Availability

We freely share all code developed in this study, along with the generated outputs and Jupyter Notebooks used to reproduce our results and figures at `https://github.com/anisdismail/ChemRxnSAGE`.

Author Information

Corresponding Author

Joe Tekli - Electrical and Computer Engineering Department, Lebanese American University, 36 Byblos, Lebanon; Orcid: `https://orcid.org/0000-0003-3441-7974` ; Email: joe.tekli@lau.edu.lb

Authors

Anis Ismail - Laboratory of Multi-omic Integrative Bioinformatics, Department of Human Genetics, Faculty of Medicine, KU Leuven, 3000 Leuven, Belgium; Orcid: `https://orcid.org/0009-0006-6990-0893` ; Email: anis.ismail@kuleuven.edu

Brigitte Wex - Department of Physical Sciences, Lebanese American University, 36 Byblos, Lebanon; Orcid https://orcid.org/0000-0002-9339-5472 ;Email: brigitte.wex@lau.edu.lb

Author Contributions

Anis Ismail: Conceptualization, Methodology, Formal Analysis, Software, Investigation, Data Curation, Writing – Original Draft, Visualization.
Joe Tekli: Supervision, Validation, Writing – Review & Editing.
Brigitte Wex: Investigation, Resources, Writing – Review & Editing.

Notes

The authors declare no competing financial interest.

Supporting Information Available

The following supporting files are available free of charge. Overview of Sequential Data Generation Techniques Using Deep Learning (Section S1), SMILES Chemical Notation Format (Section S2), Key parameters used in database generation and model training, with corresponding rationales (Supplementary Table 1), Hyperparameters for training the reaction classifier (Supplementary Table 2), Performance of Reaction Classifier on Test Set (Supplementary Table 3), Summary of standard deviation of filter metrics across five runs (Supplementary Table 4), Summary of standard deviation of similarity, novelty and reaction classes diversity metrics across five runs (Supplementary Table 5), Summary of standard deviation of samples diversity metrics across five runs (Supplementary Table 6), Evaluation Metrics for all models over training epochs (Supplementary Figures 3-8), Comparison of NVS across Reaction Classes (Supplementary Table 7), Time scaling of chemical reaction generation and evaluation (Supplementary Figure 9), Survey Questions (Supplementary Section 9) and Anonymized Survey Responses (Supplementary Section 10).

References

(1) Özçelik, R.; de Ruiter, S.; Criscuolo, E.; Grisoni, F. Chemical language modeling with structured state space sequence models. *Nature Communications* **2024**, *15*, 6176, Publisher: Nature Publishing Group.

(2) Guo, H.; Zhang, C.; Shang, J.; Zhang, D.; Guo, Y.; Gao, K.; Yang, K.; Gao, X.; Yao, D.; Chen, W.; Yan, M.; Wu, G. Drug-Target Affinity Prediction Based on Topological Enhanced Graph Neural Networks. *Journal of Chemical Information and Modeling* **2025**, *65*, 3749–3760, Publisher: American Chemical Society.

(3) Wang, Y.; Guo, M.; Chen, X.; Ai, D. Screening of multi deep learning-based de novo molecular generation models and their application for specific target molecular generation. *Scientific Reports* **2025**, *15*, 4419, Publisher: Nature Publishing Group.

(4) Wang, Y.; Wang, B.; Zou, J.; Wu, A.; Liu, Y.; Wan, Y.; Luo, J.; Wu, J. Capsule neural network and its applications in drug discovery. *iScience* **2025**, *28*, Publisher: Elsevier.

(5) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery* **2005**, *4*, 649–663.

(6) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science* **2021**, *12*, 3339–3349, Publisher: The Royal Society of Chemistry.

(7) Kim, H.; Lee, K.; Kim, C.; Lim, J.; Kim, W. Y. DFRscore: Deep Learning-Based Scoring of Synthetic Complexity with Drug-Focused Retrosynthetic Analysis for High-Throughput Virtual Screening. *Journal of Chemical Information and*

Modeling **2024**, *64*, 2432–2444, Publisher: American Chemical Society.

(8) Chen, S.; Jung, Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *Journal of Cheminformatics* **2024**, *16*, 83.

(9) Bort, W.; Baskin, I. I.; Gimadiev, T.; Mukanov, A.; Nugmanov, R.; Sidorov, P.; Marcou, G.; Horvath, D.; Klimchuk, O.; Madzhidov, T.; Varnek, A. Discovery of novel chemical reactions by deep generative recurrent neural network. *Scientific Reports* **2021**, *11*, 3178.

(10) Buin, A.; Chiang, H. Y.; Gadsden, S. A.; Alderson, F. A. De-novo Chemical Reaction Generation by Means of Temporal Convolutional Neural Networks. 2023; `http://arxiv.org/abs/2310.17341`, arXiv:2310.17341 [cs].

(11) Wang, X.; Yao, C.; Zhang, Y.; Yu, J.; Qiao, H.; Zhang, C.; Wu, Y.; Bai, R.; Duan, H. From theory to experiment: transformer-based generation enables rapid discovery of novel reactions. *Journal of Cheminformatics* **2022**, *14*, 60.

(12) Tempke, R.; Musho, T. Autonomous design of new chemical reactions using a variational autoencoder. *Communications Chemistry* **2022**, *5*, 40, Publisher: Nature Publishing Group.

(13) King-Smith, E.; Berritt, S.; Bernier, L.; Hou, X.; Klug-McLeod, J. L.; Mustakis, J.; Sach, N. W.; Tucker, J. W.; Yang, Q.; Howard, R. M.; Lee, A. A. Probing the chemical 'reactome' with high-throughput experimentation data. *Nature Chemistry* **2024**, *16*, 633–643, Publisher: Nature Publishing Group.

(14) Toniato, A.; Vaucher, A. C.; Laino, T.; Graziani, M. Negative chemical data boosts language models in reaction outcome prediction. *Science Advances* **2025**, *11*, eadt5578, Publisher: American Association for the Advancement of Science.

(15) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610, Publisher: Nature Publishing Group.

(16) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* **2018**, *9*, 6091–6098, Publisher: The Royal Society of Chemistry.

(17) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science* **2019**, *10*, 370–377, Publisher: The Royal Society of Chemistry.

(18) Shim, E.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Machine Learning Strategies for Reaction Development: Toward the Low-Data Limit. *Journal of Chemical Information and Modeling* **2023**, *63*, 3659–3668, Publisher: American Chemical Society.

(19) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(20) Noutahi, E.; Gabellini, C.; Craig, M.; C. Lim, J. S.; Tossou, P. Gotta be SAFE: a new framework for molecular design. *Digital Discovery* **2024**, *3*, 796–804, Publisher: Royal Society of Chemistry.

(21) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. arXiv (stat.ML), 2013; `https://arxiv.org/abs/1312.6114`, Preprint, submitted December 2013 (accessed 2025-09-14).

(22) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. A REPRESENTATION

TO APPLY USUAL DATA MINING TECHNIQUES TO CHEMICAL REACTIONS—ILLUSTRATION ON THE RATE CONSTANT OF SN 2 REACTIONS IN WATER. *International Journal on Artificial Intelligence Tools* **2011**, *20*, 253–270.

(23) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.

(24) Manion, J. A. et al. NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.8, Data version 2015.09. `https://kinetics.nist.gov/`, 2015; Accessed: 2025-07-12.

(25) Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019; pp 2978–2988.

(26) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **2021**, *3*, 144–152, Publisher: Nature Publishing Group.

(27) Landrum, G. RDKit: Open-Source Cheminformatics Software. **2016**,

(28) Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A. CGRtools: Python library for molecule, reaction, and condensed graph of reaction processing. *Journal of chemical information and modeling* **2019**, *59*, 2516–2521.

(29) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Central Science* **2017**, *3*, 1103–1113, PMID: 29104927.

(30) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, Apollo - University of Cambridge Repository, 2012.

(31) Smith, D. H.; Carhart, R. E.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **2002**, *25*, 64–73.

(32) Yuan, Y.; Pan, X.; Li, X.; Zhang, R.; Su, W. A 3D generation framework using diffusion model and reinforcement learning to generate multi-target compounds with desired properties. *Journal of Cheminformatics* **2025**, *17*, 93.

(33) Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K.-C.; Zhang, H. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nature Communications* **2024**, *15*, 2657, Publisher: Nature Publishing Group.

(34) Ajagekar, A.; Decardi-Nelson, B.; Shang, C.; You, F. Computer-aided molecular design by aligning generative diffusion models: Perspectives and challenges. *Computers & Chemical Engineering* **2025**, *194*, 108989.

(35) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014; `http://arxiv.org/abs/1412.6980`, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

(36) Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; Bengio, S. Generating Sentences from a Continuous Space. Proceedings of The 20th SIGNLL Conference on Computational Natural Language

Learning. Berlin, Germany, 2016; pp 10–21.

(37) Jin, X.; Han, J. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G. I., Eds.; Springer US: Boston, MA, 2010; pp 564–565.

(38) Salton, G.; McGill, M. J. *Introduction to Modern Information Retrieval*; McGraw-Hill, 1983; Google-Books-ID: 7f5TAAAAMAAJ.

(39) Zhang, X.; Li, Y.; Li, C.; Zhu, J.; Gan, Z.; Wang, L.; Sun, X.; You, H. A chemical reaction entity recognition method based on a natural language data augmentation strategy. *Chemical Communications* **2024**, *60*, 9610–9613, Publisher: The Royal Society of Chemistry.

(40) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions using Attention-Based Neural Networks. 2020; https://chemrxiv.org/engage/chemrxiv/article-details/60c753a0bdbb89acf8a3a4b5.

(41) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 163–166, Publisher: American Chemical Society.

(42) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7*, 20.

(43) Rácz, A.; Bajusz, D.; Héberger, K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of Cheminformatics* **2018**, *10*, 48.

(44) Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **1991**, *37*, 145–151.

(45) Endres, D.; Schindelin, J. A new metric for probability distributions. *IEEE Transactions on Information Theory* **2003**, *49*, 1858–1860, Conference Name: IEEE Transactions on Information Theory.

(46) Friedman, D.; Dieng, A. B. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research* **2023**,

(47) Pasarkar, A. P.; Dieng, A. B. Cousins Of The Vendi Score: A Family Of Similarity-Based Diversity Metrics For Science And Machine Learning. 2024; http://arxiv.org/abs/2310.12952, arXiv:2310.12952 [physics, q-bio].

(48) Benhenda, M. ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? 2017; http://arxiv.org/abs/1708.08227, arXiv:1708.08227 [cs, stat].

(49) Stein, G.; Cresswell, J.; Hosseinzadeh, R.; Sui, Y.; Ross, B.; Villecroze, V.; Liu, Z.; Caterini, A. L.; Taylor, E.; Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems* **2023**, *36*, 3732–3784.

(50) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(51) Allinger, N. L. In *Theoretical and Computational Models for Organic Chemistry*; Formosinho, S. J., Csizmadia, I. G., Arnaut, L. G., Eds.; Springer Netherlands: Dordrecht, 1991; pp 125–135.

(52) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society, 1982.

(53) SciFinder, Chemical Abstracts Service: Columbus, OH. https://scifinder.cas.org.

(54) Jiang, C. J.; Cheng, C. L.; Yuan, S. F. Economical and Practical Strategies for

Synthesis of a-Trifluoromethylated Amines. *Asian Journal of Chemistry* **2015**, *27*, 2406–2408, Publisher: https://asianpubs.org/.

(55) Teno, N.; Gohda, K.; Wanaka, K.; Tsuda, Y.; Akagawa, M.; Akiduki, E.; Araki, M.; Masuda, A.; Otsubo, T.; Yamashita, Y. Novel type of plasmin inhibitors: Providing insight into P4 moiety and alternative scaffold to pyrrolopyrimidine. *Bioorganic & Medicinal Chemistry* **2015**, *23*, 3696–3704.

(56) Pagare, P. P.; Ghatge, M. S.; Chen, Q.; Musayev, F. N.; Venitz, J.; Abdulmalik, O.; Zhang, Y.; Safo, M. K. Exploration of Structure–Activity Relationship of Aromatic Aldehydes Bearing Pyridinylmethoxy-Methyl Esters as Novel Antisickling Agents. *Journal of Medicinal Chemistry* **2020**, *63*, 14724–14739, Publisher: American Chemical Society.

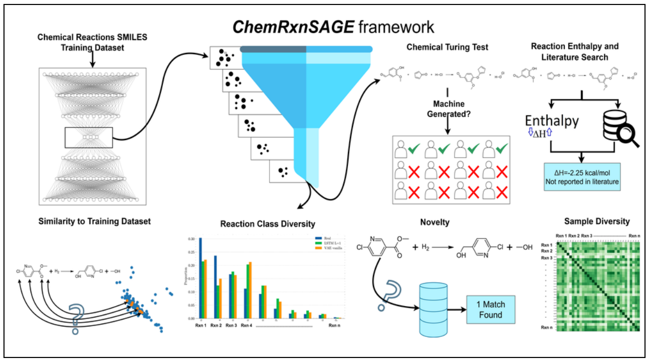
(57) Han, Y.-Y.; Chen, W.-B.; Han, W.-Y.; Wu, Z.-J.; Zhang, X.-M.; Yuan, W.-C. Highly Efficient and Stereoselective Construction of Dispiro-[oxazolidine-2-thione]bisoxindoles and Dispiro[imidazolidine-2-thione]bisoxindoles. *Organic Letters* **2012**, *14*, 490–493, Publisher: American Chemical Society.

(58) Liu, C.; Bao, H.; Wang, D.; Wang, X.; Li, Y.; Hu, Y. Highly chemoselective hydrogenation of active benzaldehydes to benzyl alcohols catalyzed by bimetallic nanoparticles. *Tetrahedron Letters* **2015**, *56*, 6460–6462.

(59) Schimler, S. D.; Ryan, S. J.; Bland, D. C.; Anderson, J. E.; Sanford, M. S. Anhydrous Tetramethylammonium Fluoride for Room-Temperature SNAr Fluorination. *The Journal of Organic Chemistry* **2015**, *80*, 12137–12145, Publisher: American Chemical Society.

(60) Lou, A.; Meng, C.; Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834* **2024**,

For Table of Contents Only


ChemRxnSAGE framework
Chemical Reactions SMILES Training Dataset
Chemical Turing Test
Machine Generated?
Reaction Enthalpy and Literature Search
Enthalpy
ΔH
ΔH=-2.25 kcal/mol
Not reported in literature
Similarity to Training Dataset
Reaction Class Diversity
Real
LSTM L=1
VAE vanilla
Proportion
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Rxn 1
Rxn 2
Rxn 3
Rxn 4
Rxn n
Novelty
Cl
N
O
+ H₂
OH
+ —OH
1 Match Found
Sample Diversity
Rxn 1
Rxn 2
Rxn 3
Rxn n