MolGen-Transformer: An open-source self-supervised model for Molecular Generation and Latent Space Exploration

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present the MolGen-Transformer, a generative AI model achieving 100% reconstruction accuracy through self-supervised training using a large, curated meta-dataset of organic molecules with less than 168 atoms. MolGen-Transformer produces valid molecular structures using the SELF-referencing Embedded Strings (SELFIES) representation. Our training dataset comprises 198 million organic molecules, selected to encompass a wide range of organic structures. We illustrate the generative capability of this model in three ways: (a) *Generating chemically similar molecules*, where the model creates structurally similar valid molecules to a given prompt molecule; (b) *Producing Diverse Molecules*, where the model creates structurally diverse valid molecules given a random latent seed, and (c) *Identifying Chemical Intermediates*, where the model creates a sequence of valid molecules connecting two given molecules. MoleGen-Transformer allows the generation and exploration of structurally similar molecules and provides insights into structural pathways between molecules. The model weights and inference methods are publicly available to support community use. We also provide an easy-to-use website for exploration.

1 Introduction

The integration of generative Artificial Intelligence (AI) into computational chemistry has significantly advanced the field, yielding promising developments that extend from theoretical frameworks to practical applications. An emphasis on molecule representation and generation has produced rapid advances across broad chemical research areas such as drug development, materials discovery, and chemical synthesis [1, 2, 3, 4, 5, 6, 7].

We focus on developing a molecular generation framework that ensures the generation of 100% valid molecular structures, which is crucial for advancing chemical research. This guarantees that all produced molecules are chemically plausible and syntactically correct. This level of reliability is essential for practical applications in drug development, materials science, and other fields, as it reduces the need for extensive post-generation validation and correction [8, 1, 3].

Several notable works in the field include various representations and learning techniques. For instance, Zeng et al. [9] developed a self-supervised image representation learning framework for predicting molecular properties and drug targets, utilizing an image processing framework combined with molecular chemistry knowledge to capture structural characteristics. Xu et al. [10] introduced a triple generative self-supervised learning method for molecular property prediction, leveraging variational autoencoders (VAEs) and incorporating BiLSTM, Transformer, and GAT. Chen et al. [11] focused on extracting predictive representations from hundreds of millions of molecules using a bidirectional encoder transformer (BET). Wu et al. [12] explored self-supervised learning on graphs using contrastive, generative, or predictive techniques, employing graph convolutional networks (GCNs) and graph attention networks (GATs). However, none of these approaches ensure 100% molecular validity. Achieving 100% validity in the generation of diverse molecules remains a challenge [13, 8, 14, 15, 16, 17].

In contrast, our work, MolGen-Transformer, employs the SELFIES (SELF-referencing Embedded Strings) representation introduced by Krenn et al. [8]. SELFIES overcomes the limitations of SMILES, ensuring both syntactic and semantic validity of the generated molecular graphs. This 2D representation is computationally efficient and guarantees 100% valid molecular structures, addressing a critical gap in current methodologies. Recent studies have explored its application across various domains by leveraging the SELFIES representation. The SELFormer model, proposed by Atakan Yüksel [14], utilizes SELFIES for predicting aqueous solubility and adverse drug reactions, demonstrating its superiority over both traditional graph-based methods and SMILES-based chemical language models (CLMs). Furthermore, research conducted by Shengmin Piao et al. [15] introduced SELF-EdiT, a molecular structure editing model that employs SELFIES alongside Levenshtein transformer models.

We aim to develop a generative model for organic molecules that caters to a broad chemical research audience, ensuring the generation of 100% valid molecules. This model is versatile across various datasets, free from constraints tied to specific pre-trained datasets' distributions, and features an embedding space capable of containing an extensive dataset. This enables the generation of diverse organic molecules and new molecules structurally akin to given target molecules. We utilize a meta-dataset encompassing 198 million public and in-house organic molecules. This dataset is chosen to cover an extensive range of organic structures and applications, distinctively positioning our work to transcend specific distribution learning models. Our MolGen-Transformer, a Transformer model paired with an Auto-Encoder (AE) framework, including a bidirectional encoder and an autoregressive decoder, leverages the datasets structural and application diversity.

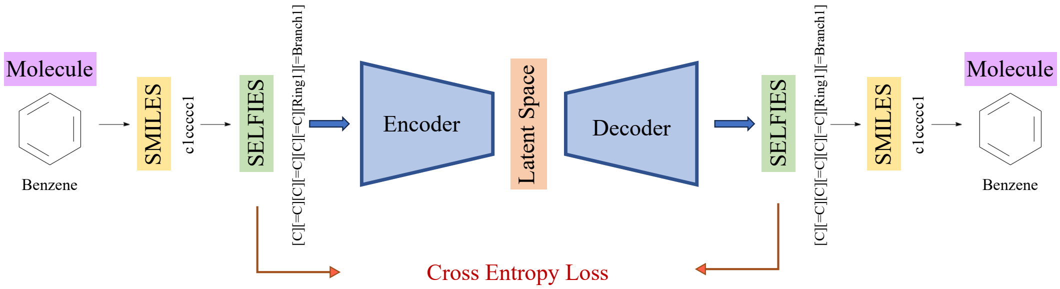


Figure 1: **Self-Supervised Auto-Encoder Training:** The process involves converting a molecule to its SMILES and SELFIES representations, encoding SELFIES into latent space using a bidirectional encoder, and decoding with an autoregressive decoder to reconstruct the molecule, minimizing cross-entropy loss between the input and reconstructed SELFIES strings.

2 Contributions

Meta Dataset Training. The MolGen-Transformer was trained on an extensive meta-dataset comprising 198 million organic molecules.

High Reconstruction Accuracy. Achieving 100% reconstruction accuracy is a significant milestone. This ability to encode and decode molecular structures ensures the generation of chemically valid molecules.

Inference Methods. We illustrate three inference methods:

- **Generating Chemically Similar Molecules:** This method generates molecules from an initial molecule of interest. The results demonstrate our algorithm's ability to generate molecules that are structurally similar to user-defined molecules, preserving the integrity of rings and bonds while ensuring validity.

- **Producing Diverse Molecules:** This method samples from the latent space on a normal distribution and then decodes these latent vectors into molecules. We demonstrate a Tanimoto diversity [18] score of 0.93, indicating a high degree of diversity in the generated molecules.
- **Identifying Chemical Intermediates:** Given two input molecules, this method generates intermediate molecules along the segments in the latent space.

Open Source Model and Package. The model weights and some random batches of testing datasets have been made publicly available to enhance accessibility and impact within the research community. The package can be easily installed and tested via `pip install`.

3 Method

3.1 Meta-dataset description

The meta dataset comprises approximately 198 million organic molecules, combining proprietary and publicly available sources. This dataset was curated to cover a broad range of molecular diversity, ensuring the generative model can capture intricate representations of organic molecules. The selection of subsets was strategically made to include a variety of molecular sizes and structures. Table 1 provides a summary of the key datasets included in the Meta dataset.

Dataset	**Dataset Size**	**Data Brief Description**
Zinc [19]	250k	Contains commercially available molecules with at most 38 heavy atoms.
GDB-17 [20]	50 M	Synthetic dataset of molecules with at most 17 heavy atoms, consisting of halogens, along with C, N, O, and S.
OCELOT + [21]	33 M	Synthetic dataset created from the combinatorial generation of the largest connected group of fused rings of scaffolds from OCELOT.
ORNL	10 M	Synthetic dataset which is a subset of 10 million molecules from the Enamine REAL database.
PubChem [22]	106 M	Real molecules spanning many fields of chemistry.
HCEP [23]	2 M	Includes π-conjugated organic molecules with properties such as HOMO and LUMO gap calculated via DFT.
D^{3}TaLES [24]	43 K	Contains redox-active small molecules tailored for applications in non-aqueous redox-flow batteries, with various properties calculated via DFT.
OCELOT [21]	24 K	Contains unique small molecules. This dataset represents the chemical space of structures for OSC applications.

Table 1: **Summary of Datasets Used in the Study**. The Meta dataset spans from simple carbon chains to complex molecules containing a multitude of rings and over 100 atoms, ensuring broad coverage of the chemical space.

Initially, all datasets were collected in their SMILES representations, which were then converted into the Kekulé form to standardize molecular representations. This conversion explicitly represented single and double bonds in aromatic molecules. We then filtered out non-organic molecules to ensure the dataset's focus on organic compounds. After this, the Kekulé SMILES were translated into SELFIES representations to ensure data uniformity and robustness.

Further details on the statistical distribution of key molecular features in the Meta dataset can be found in Supporting Information Section 1.1. This rigorous preprocessing ensures the dataset's fidelity and uniformity, providing a robust foundation for model training and evaluation.

3.2 Self-Supervised Transformer

Our self-supervised auto-encoder training method encodes the meta-dataset of 198 million organic molecules into a latent space.

The SELFIES representation is first tokenized, and each token is embedded into a vector space with an embedding size of 30. The vocabulary consists of 121 unique SELFIES symbols, each representing distinct molecular fragments, atoms, bonds, or rings. These embeddings are then fed into a bidirectional encoder, which consists of 2 layers and 3 attention heads. This configuration allows the encoder to process the input sequence from both directions, capturing both forward and backward contextual information. The hidden size of the model is set to 100, ensuring sufficient capacity to model complex molecular relationships while maintaining computational efficiency. In total, the model comprises 54,821 trainable parameters. Next, the latent vector is passed through an autoregressive decoder, which operates similarly to next-word prediction models used in NLP [25].

The reconstructed SELFIES string is converted back to its SMILES representation and finally to the molecular structure. The training process aims to minimize the cross-entropy loss between the input and the reconstructed SELFIES representations. This ensures that the encoder-decoder pair learns to accurately reconstruct the input molecules, achieving high reconstruction accuracy (Figure 1).

Loss Function We employed a Negative Log Likelihood (NLL) loss function for training, defined as:

$$\mathcal{L}(\mathbf{y}, \mathbf{p}) = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \cdot \log(p_{i,y_i}), \tag{1}$$

where N is the number of samples, w_{y_i} denotes the class weight for the true class y_i, and p_{i,y_i} is the predicted probability of the true class y_i for sample i. The class weights w_{y_i} are computed as follows:

$$w_{y_i} = 10^{-4} \cdot \min\left(\frac{0.01}{\frac{\log(N_{y_i})}{\sum_{s\in S}\log(N_s)} + \epsilon}, 0.90\right), \tag{2}$$

where N_{y_i} is the frequency of the true class y_i in the dataset, and $\epsilon = 10^{-6}$ is a small constant to avoid division by zero. This weighting scheme ensures that less frequent classes are given higher importance during training.

Other Training Details Key components of our training strategy included:

Data Handling: To manage memory usage effectively and maintain data randomness, the dataset was randomly divided into 20 sequential chunks, with an 80% split designated for training and the remaining 20% for testing. Each chunk was independently shuffled before training to ensure randomness across the iterations.

Training Epochs and Duration: The model was trained on each of the 20 chunks of the training data for two epochs per chunk. This approach resulted in a total of 244 training iterations across the entire dataset.

Optimizer and Learning Rate: The Adam optimizer with a dynamically adjusted learning rate.

Training Hardware Configuration: The training was conducted on the NOVA High-Performance Computing (HPC) system, utilizing nodes equipped with Intel 8358 processors, 369GB of memory, and four Nvidia A100 GPUs per node. Each job was allocated a wall time limit of 120 hours, with 16 processor cores per node dedicated to the task.

This hardware configuration ensured that the MolGen-Transformer could efficiently learn complex patterns within the molecular datasets while maximizing computational performance.

3.3 Application details

3.3.1 Generating Chemically Similar Molecules via Latent Space Exploration of Initial Molecule

We leverage the trained MolGen-Transformer to develop an inference method for molecule generation by exploring the latent space around an initial molecule. In the latent space, a set of n random normalized vectors is generated around the initial molecule's latent vector, representing potential new molecules nearby. A binary search mechanism within the decoder identifies the closest neighbors and reconstructs their SELFIES representations. These SELFIES are then converted back to SMILES and finally to molecular structures. The generated molecules undergo several filtering and sorting steps to ensure quality and novelty. First, duplicate molecules are removed. Then, molecules are {sorted by a Pareto frontier algorithm. Finally, molecules are filtered based on synthetic accessibility scores, ensuring they are practically synthesizable.

Synthetic Accessibility (SA) Consideration Although the global efficacy of measuring synthetic accessibility (SA) is still debated among scientists [26, 27], we consider SA crucial for real-world engineering applications. We utilize the SA score from Ertl et al.'s study [28]. This method combines fragment contributions and a complexity penalty. The molecular complexity score accounts for non-standard structural features, such as large rings, non-standard ring fusions, stereocomplexity, and molecule size. The method has been validated by comparing calculated SA scores with the ease of synthesis estimated by experienced medicinal chemists, showing a high agreement ($r^2 = 0.89$). While the SA threshold is user-defined in our package, our demonstrations use a threshold of 6. According to Ertl et al. [28], molecules with an SA score above 6 are difficult to synthesize, whereas those with lower scores are more easily synthesizable. This threshold helps identify molecules that are likely to be practical for synthesis and real-world applications.

Pareto Frontier Algorithm We employ a Pareto frontier approach to adjust α based on user needs, toggling between the L2 norm distance in latent space and the Tanimoto similarity. Generally, we found that smaller inverted distances indicate higher atom-wise similarity, while higher Tanimoto similarity reflects greater structural similarity. The choice of α is left to the user, enabling a flexible and customizable approach to molecule generation.

$$\text{pareto\_frontier}_i = \alpha \cdot \text{distances\_inverted}_i + (1 - \alpha) \cdot \text{similarities\_norm}_i \tag{3}$$

where: α is a parameter balancing the importance of similarity matrices.distances_inverted$_i$ is the inverted normalized latent space distance of the i-th molecule, and similarities_norm$_i$ is the normalized Tanimoto similarity of the i-th molecule.

3.3.2 Producing Diverse Molecules from Latent Space Sampling

Molecules can also be generated by sampling the latent space using a normal distribution. This method involves decoding a sample of n latent vectors into SELFIES, converting them to SMILES, and ultimately to molecular structures. To evaluate the generated molecules, we analyze the Tanimoto diversity score, uniqueness ratio, distribution of atom types, and distribution of atom counts. The Tanimoto diversity score and uniqueness ratio reflect the structural diversity among the molecules, while the atom type and atom count distributions provide insights into the generated molecules' diverse sizes and compositions. The Tanimoto diversity score is simply $1 -$ Tanimoto Similarity

The uniqueness ratio is computed as:

$$\text{Uniqueness Ratio} = \frac{\text{Number of Unique Standard InChI}}{n} \tag{4}$$

where n is the total number of InChi strings. InChI (International Chemical Identifier) is a structure-based chemical identifier developed by IUPAC and the InChI Trust [29], serving as a standard for chemical databases and facilitating effective information management. Each molecule can only have one standard InChI. Here they are decoded from SMILES string.

3.3.3 Identifying Chemical Intermediates in Latent Space

To enable the exploration of molecular intermediates, provide insights into the chemical nature of the latent space, and facilitate the discovery of new molecules, we developed an inference method for identifying chemical intermediates in latent space. The process begins by encoding two initial molecules, referred to as the start and end molecules, into their respective latent space representations using the encoder. A line segment is created in the latent space between the latent vectors of

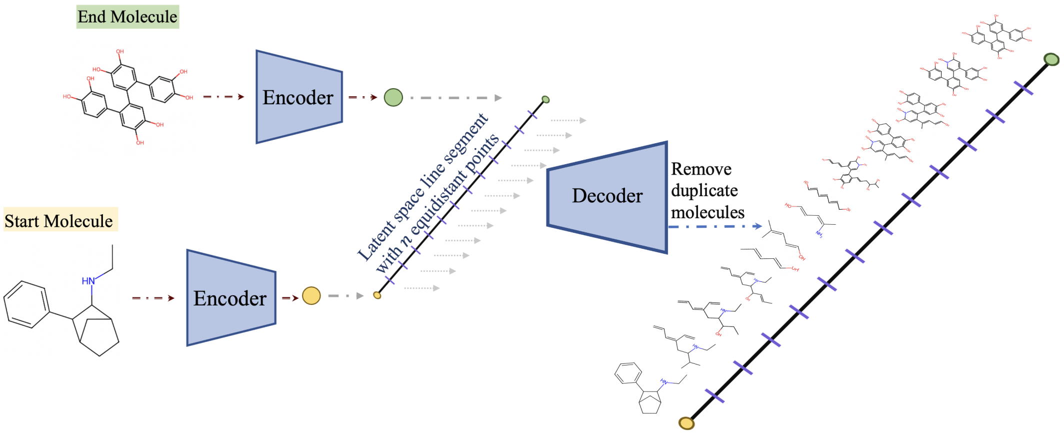
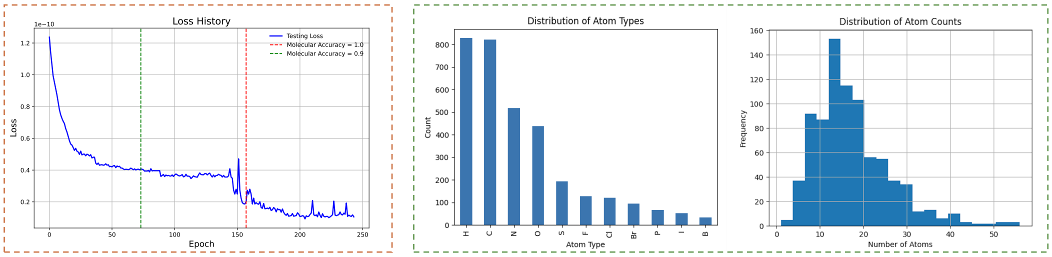

End Molecule
Start Molecule
Encoder
Encoder
Latent space line segment with n equidistant points
Decoder
Remove duplicate molecules

Figure 2: Identifying Chemical Intermediates: The process starts with encoding both the start and end molecules (the two initial molecules) into latent space representations. A line segment in the latent space is created between these representations, with interpolation points along the segment. These points are decoded into molecular structures.

these two molecules, with multiple interpolation points generated along this segment. These points represent potential intermediate molecular structures between the start and end molecules. Each interpolation point is then decoded into its corresponding SELFIES representation using the decoder, converted back to SMILES, and finally into the molecular structure. The generated molecules are filtered to remove duplicates, ensuring each molecule in the series is unique (Figure 2).

4 Results and Discussion

4.1 Self-Supervised Transformer



(a)

(b)

Figure 3: (a) Loss History of MolGen-Transformer: The plot shows the testing loss across 244 training iterations, with each iteration representing two epochs of training on a shuffled subset of the Meta dataset. The model achieves 100% molecular and symbolic accuracy from iteration 157 onwards, indicating the model's convergence and robustness in reconstructing molecular structures. (b) Distribution of atom types and atom counts for $n = 1000$ generated molecules: The left panel illustrates the distribution of various atom types present in the generated molecules, while the right panel displays the distribution of the number of atoms per molecule. The Molecular Uniqueness Ratio is 0.83, and the Tanimoto Diversity score is 0.93, highlighting the diversity and uniqueness of the generated molecules.

Figure 3 depicts the corresponding testing loss. Notably, at iteration 73, where the testing loss reaches 4.0734e-11, the model's molecular reconstruction accuracy reaches 90%. Molecular reconstruction accuracy refers to the model's ability to encode a SELFIES string into the latent space and then decode it back to the correct SELFIES string. At the same iteration, the symbolic accuracy, which measures the accuracy of individual SELFIES symbols, was observed to be 99.8%. From iteration 157 onwards, the model achieved perfect performance, with both molecular reconstruction accuracy and symbolic accuracy consistently reaching 100%. This indicates that the model not only learned to accurately reconstruct the molecular structures but also generalized well across the diverse molecular representations in the dataset.

Although the MolGen-Transformer might be considered small compared to typical NLP transformers, with its embedding size of 30, the model's size is well-suited to the problem at hand. Unlike human languages, which have a vast vocabulary, the vocabulary in our problem definition is naturally small, consisting of 121 unique SELFIES symbols. Consequently, a large embedding space, such as 512 or 1024 dimensions, is unnecessary. The number of trainable parameters in transformers scales as the square of the embedding size, so a larger model would result in significantly more parameters without providing additional benefits for our specific task. Our findings demonstrate that the current model size is sufficient to effectively learn and represent the latent space, as evidenced by the 100% accuracy in both molecular and symbolic testing from iteration 157 onwards.

4.2 Generating Chemically Similar Molecules via Latent Space Exploration

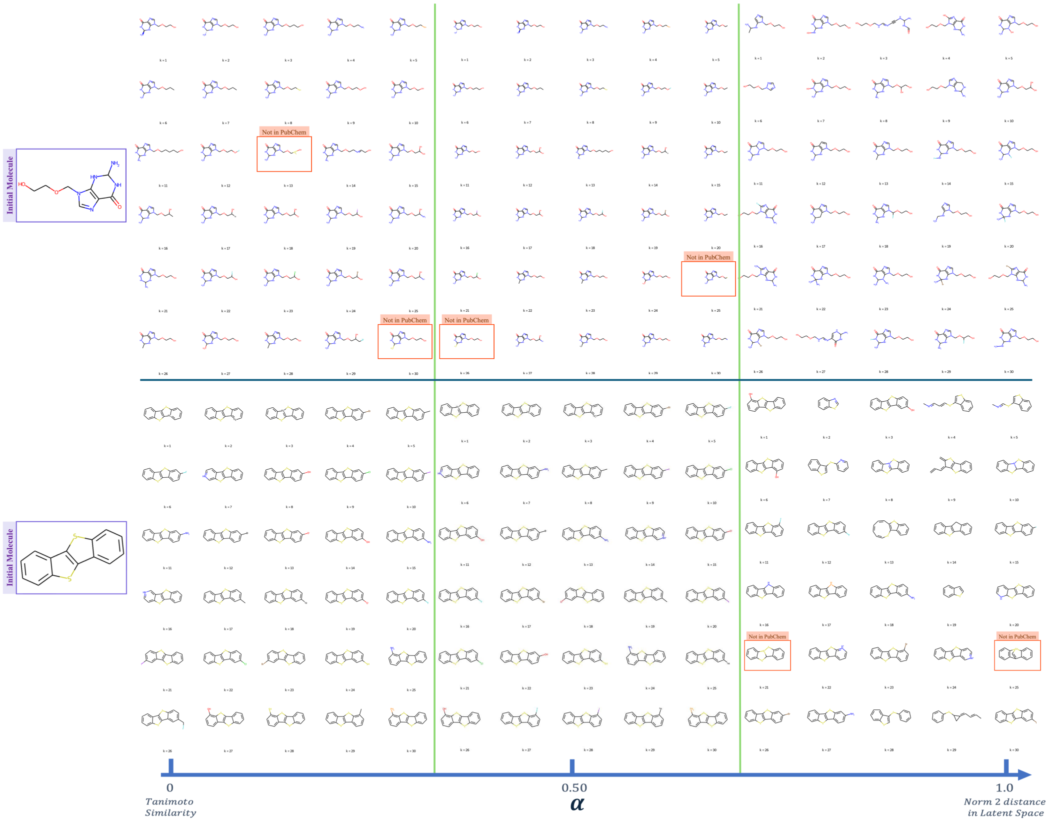


Figure 4: Results of Generating Chemically Similar Molecules: Initial molecules are in the purple box in the top left. Generated molecules not found in PubChem are highlighted in red.

We applied the Generating Chemically Similar Molecules method (Section 3.3) to 6 representative molecules from the dataset to test the model's ability to generate novel molecules in a given chemical space. Recall that the parameter α ranges from 0 to 1; higher values prioritize L2 norm distance in the latent space, while lower values emphasize Tanimoto Similarity. Here, we test $\alpha = 0$, $\alpha = 0.5$, and $\alpha = 1$. Figure 4 shows the top $k = 30$ neighbors using each α value for two of the six initial molecules: [1]benzothieno[3,2-*b*][1]benzothiophene (BTBT) and the ethylene glycol substituted guanine, 2-Amino-9-[(2-hydroxyethoxy)methyl]-6-oxo-2,3,6,7-tetrahydro-1H-purin-9-ium (MEG-G). Inspection of all generated molecules reveals structural and compositional similarity to the initial molecule. However, greater considerations of Tanimoto similarity ($\alpha = 0$) produce more structural similarity and more diverse atom types. For both BTBT and MEG-G, the $\alpha = 0$ generation produced structures containing F, Br, Cl, and (for MEG-G) I, while the $\alpha = 1$ generation produced only F and Br halogens in the generated structures. On the other hand, greater considerations of the L2 Norm distance ($\alpha = 1$) produce more diverse structures and similar atom types, with several generated molecules containing broken or altered ring structures. Notably, the Generating Chemically Similar Molecules produced molecules not found in the PubChem database (the largest collection of freely accessible chemical information[30]), suggesting the generation of novel chemical structures. Although we show ^{13}C in the results, users have the option to filter out ^{13}C from the top k generated neighbors if desired.

We show that the MolGen-Transformer can generate molecules structurally similar to user-defined molecules, preserving the structural integrity of rings and bonds while ensuring the validity of the generated molecules. Moreover, when testing $\alpha = 0, 0.5$, and 1 with $k = 30$ for six initial molecules, we generated twelve potentially novel (not found in PubChem) structures (Figure 5). This feature enhances the model's applicability in real-world scenarios, providing a valuable tool for generating novel yet structurally relevant molecules. For additional example results, please refer to the Supporting Information Section 1.3.

4.3 Producing Diverse Molecules from Latent Space

To measure the ability of the latent space to generate diverse molecules, we generated molecules (Section 3.3.2) by normal sampling $n = 1000$ vectors in the latent space. The results presented here are averages from experiments repeated 10 times. The uniqueness ratio may be less than 1 because multiple SMILES can map to the same standard InChI, and multiple SELFIES can map to the same SMILES. These results demonstrate the diverse generation capability of the MolGen-Transformer inference method, achieving a Tanimoto diversity score of 0.93. This corresponds to an average Tanimoto similarity of 0.07, which is considered low in the context of chemistry, where two structures are typically deemed similar if $T > 0.85$ [31].

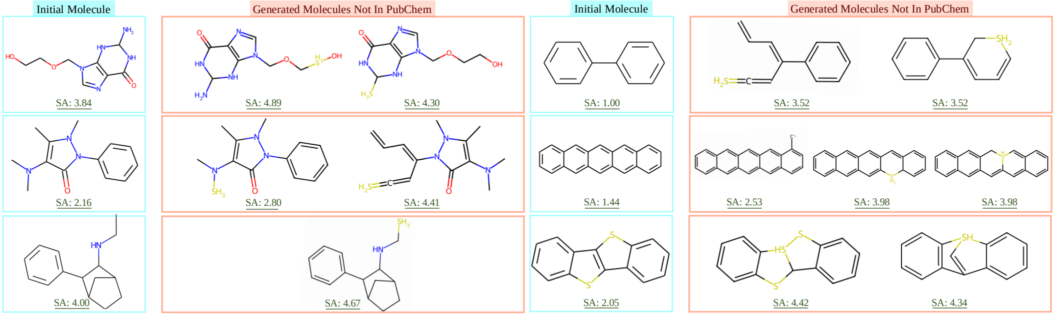


Figure 5: Newly Generated Molecules Not in PubChem: This figure shows the generated molecules not found in PubChem when testing $\alpha = 0, 0.5$, and 1 with $k = 30$ for six initial molecules. The SA scores are indicated for each molecule, with all molecules passing the SA score filter (threshold of 6). Higher SA scores indicate greater synthetic difficulty.

4.4 Identifying Chemical Intermediates

To gain more insight into the chemical nature of the latent space, we generate molecules along the line segment between the latent spaces of two initial molecules, referred to as the start molecule and the end molecule (Section 3.3.3). Here, we select two pairs of distinct start and end molecules, to highlight the evolutionary process through the latent space from one molecular structure to a distinctively different one.

First, as the model evolves benzene to BTBT, we observe the model opening the ring and adding sulfur. It then adds more sulfur and carbon chains/rings before closing all rings to produce BTBT. The evolution of biphenyl to MEG-G shows a ring opening, then iterative additions of $-OH$ and $-NH_3$ groups (Figure 6). While imperfect, these molecular evolution examples align with a general sense of chemical intuition. For more example results of Identifying Chemical Intermediates, refer to the Supporting Information Section 1.4.

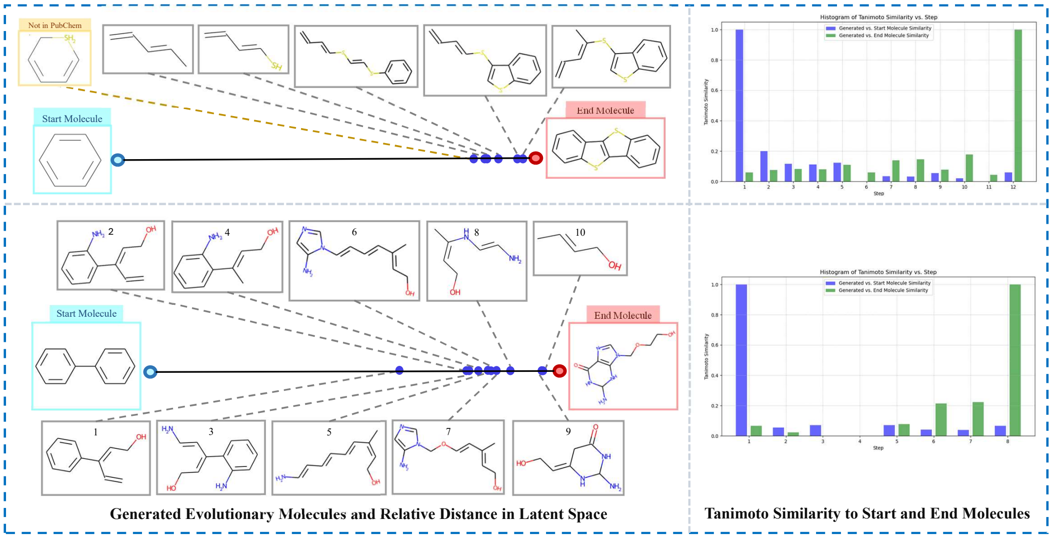


Figure 6: Identifying Chemical Intermediates results for (a) benzene to BTBT and (b) biphenyl to MEG-G. The top panels in each section depict the structural intermediates between each structure pair, while the bottom panels present the Tanimoto similarities for each intermediate molecule relative to the start molecule (blue) and the end molecule (green). As expected, the leftmost blue bar and the rightmost green bar always equal 1.0, indicating the similarity of the start molecule to itself and the end molecule to itself, respectively.

5 Conclusion

This study presents the MolGen-Transformer, a generative AI model achieving 100% reconstruction accuracy and generating 100% valid molecular structures using the SELF-referencing Embedded Strings (SELFIES) representation. The model was trained on a curated meta dataset of 198 million organic molecules, selected to cover a wide range of organic structures. This comprehensive training ensures the model's applicability across diverse chemical research domains, from drug development to materials science.

In addition to creating the MolGen-Transformer model, we develop three inference methods: Generating Chemically Similar Molecules, Producing Diverse Molecules, and Identifying Chemical Intermediates. These methods showcase the model's versatility in generating diverse and structurally relevant molecules and provide insights into molecular transitions to facilitate the discovery of new molecules. Detailed results in the paper validate these capabilities.

By making the model weights and inference methods publicly available, we aim to foster further advancements in the field and support the broader chemical research community. The MolGen-Transformer represents a significant step towards more universally applicable and reliable molecular generative models, offering valuable tools for scientific investigations and practical applications in chemical research.

References

[1] A. Author. Advances in machine learning for molecular design. *Nature Chemistry*, 13(7):577–587, 2021.

[2] B. Author. Deep learning for molecular design—a review of the state of the art. *ACS Central Science*, 4(11):1460–1468, 2018.

[3] C. Author. Machine learning in chemistry: data-driven algorithms, learning systems, and applications. *Chemical Science*, 10(1):205–219, 2019.

[4] D. Author. Recent advances in machine learning for drug discovery. *Chemical Reviews*, 121(11):7420–7498, 2021.

[5] E. Author. Applications of machine learning in chemical engineering. *Journal of the American Chemical Society*, 142(9):4090–4106, 2020.

[6] F. Author. A review of applications of machine learning in chemistry. *Journal of Chemical Physics*, 153(2):020902, 2020.

[7] G. Author. Generative models for molecular design. *Nature Machine Intelligence*, 4(1):10–21, 2022.

[8] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[9] Xiangxiang Zeng, Xin Gao, Jie Liu, and Ying Zhang. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.

[10] Lei Xu, Jing Wang, Hongyang Li, and Wei Zhang. Triple generative self-supervised learning method for molecular property prediction. *International Journal of Molecular Sciences*, 25(7):3794, 2024.

[11] Dong Chen, Hao Wang, Jian Li, and Yan Zhang. Extracting predictive representations from hundreds of millions of molecules. *The Journal of Physical Chemistry Letters*, 12(44):10793–10801, 2021.

[12] Lirong Wu, Yehong Rao, Zhenxing Dai, and Philip S Yu. Self-supervised learning on graphs: Contrastive, generative, or predictive. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):4216–4235, 2021.

[13] Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics*, 10(1):1–9, 2018.

[14] Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. Selformer: Molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 2023.

[15] Shengmin Piao, Jonghwan Choi, Sangmin Seo, and Sanghyun Park. Self-edit: Structure-constrained molecular optimisation using selfies editing transformer. *Applied Intelligence*, 53(21):25868–25880, 2023.

[16] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[17] Shuang Wang, Tao Song, Shugang Zhang, Mingjian Jiang, Zhiqiang Wei, and Zhen Li. Molecular substructure tree generative model for de novo drug design. *Briefings in bioinformatics*, 23(2):bbab592, 2022.

[18] Jeffrey W Godden, Ling Xue, and Jürgen Bajorath. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. *Journal of Chemical Information and Computer Sciences*, 40(1):163–166, 2000.

[19] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012. PMID: 22587354.

[20] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. PMID: 23088335.

[21] Qianxiang Ai, Vinayak Bhat, Sean M. Ryno, Karol Jarolimek, Parker Sornberger, Andrew Smith, Michael M. Haley, John E. Anthony, and Chad Risko. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *The Journal of Chemical Physics*, 154(17):174705, 05 2021.

[22] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022.

[23] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alán Aspuru-Guzik. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.

[24] Rebekah Duke, Vinayak Bhat, Parker Sornberger, Susan A. Odom, and Chad Risko. Towards a comprehensive data infrastructure for redox-active organic molecules targeting non-aqueous redox flow batteries. *Digital Discovery*, 2:1152–1162, 2023.

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

[26] Connor W. Coley, Liam Rogers, William H. Green, and Klavs F. Jensen. Scscore: Synthetic complexity learned from a reaction corpus. *Chemical Science*, 11(2):566–572, 2020.

[27] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *European Journal of Medicinal Chemistry*, 45(6):2606–2615, 2010.

[28] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.

[29] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5:1–9, 2013.

[30] Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl_2):W623–W633, 2009.

[31] David E Patterson, Richard D Cramer, Allan M Ferguson, Robert D Clark, and Laurence E Weinberger. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of medicinal chemistry*, 39(16):3049–3059, 1996.

article neurips$_2$024

super,sortcompress,commanatbib [version=3]mhchem graphicx lastpage hyperref epstopdf verbatim caption subcaption

6 Supporting Information

This Supporting Information document provides supplementary analysis and additional results to support the findings presented in the main text. It includes detailed examinations of the Meta dataset used in training the MolGen-Transformer, such as distribution and atom count analyses, as well as further examples of local molecular generation and molecular evolution. These additional insights are intended to offer a more comprehensive understanding of the dataset and the model's capabilities in generating and evolving molecular structures.

6.1 Statistical and Distribution Analysis of the Meta Dataset

This section provides a comprehensive analysis of the Meta dataset, focusing on key molecular properties such as atom count per molecule and the frequency of each atom type. Understanding these properties is essential for evaluating the model's ability to generalize across diverse molecular structures.

In addition to the statistical overview, Figure 7 visualizes the distribution of key molecular features from a random sample of 2 million molecules within the Meta dataset's testing set. The left panel illustrates the distribution of atom counts per molecule, the middle panel shows the distribution of ring counts, and the right panel depicts the distribution of atom types. This visualization provides a clear summary of the dataset's diversity, which is fundamental to the model's robust performance.

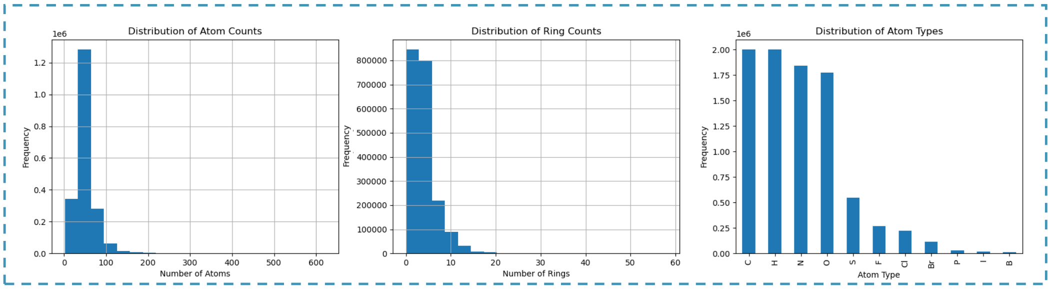


Figure 7: Distribution Analysis of the Meta Dataset Testing Set: The figure presents detailed distributions from a random sample of 2 million molecules within the testing set of the Meta dataset. The left panel shows the distribution of atom counts, indicating the frequency of molecules with varying numbers of atoms. The middle panel illustrates the distribution of ring counts, showing the frequency of molecules with different numbers of rings. The right panel displays the distribution of atom types, highlighting the prevalence of different elements, including carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and others within the sampled molecules.

6.2 Model Capability for Atom Count

This section provides an analysis of the MolGen-Transformer's ability to handle molecules of varying sizes, specifically focusing on the number of atoms per molecule. The results demonstrate the model's versatility in processing a wide range of atom counts, making it suitable for diverse chemical applications.

Figure 8 presents a detailed examination of the SELFIES representation and corresponding atom counts within a random sample of 2 million molecules from the Meta dataset's testing set. The figure is divided into three parts: (a) the distribution of SELFIES string lengths across the dataset, offering insights into the complexity of molecular representations; (b) the atom count distribution for molecules with SELFIES lengths greater than 400 symbols, highlighting the model's ability to handle larger molecules, where the minimum number of atoms in this category is 168, with 8,763 such molecules present; and (c) the atom count distribution for molecules with SELFIES lengths less than 400 symbols. This analysis provides a comprehensive understanding of the SELFIES representation within the dataset and helps estimate the range of molecular sizes that the MolGen-Transformer can effectively capture without capping the SELFIES representation, which covers approximately 99.56% of the molecules in the dataset.

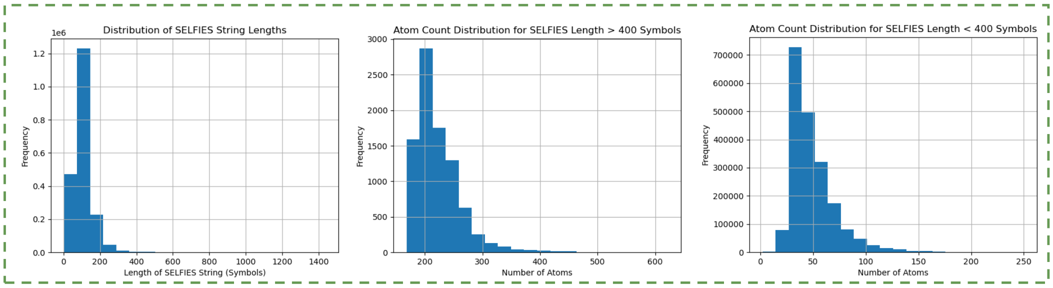


Figure 8: SELFIES Representation and Atom Count Analysis: This figure presents the distribution of SELFIES string lengths and corresponding atom counts within a random sample of 2 million molecules from the Meta dataset testing set. (a) Distribution of SELFIES string lengths, providing insights into the complexity of molecular representations. (b) Atom count distribution for molecules with SELFIES lengths greater than 400 symbols, indicating the model's capability to handle larger molecules. Molecules in this category have a minimum of 168 atoms, with 8,763 such molecules present in the dataset. (c) Atom count distribution for molecules with SELFIES lengths less than 400 symbols. This analysis helps estimate the size of molecules that the MolGen-Transformer can fully capture without capping the SELFIES representation, covering approximately 99.56% of the molecules in the dataset.

6.3 Additional Results of Local Molecular Generation Results

Figure 9 provides additional examples of local molecular generation, illustrating the MolGen-Transformer's capability to generate novel molecules that are structurally similar to a given input molecule. The generated molecules maintain the integrity of molecular rings and bonds while introducing variations, demonstrating the model's effectiveness in producing chemically relevant structures.

6.4 Additional Results of Molecular Evolution and Generation

Figures 10 provide additional examples of molecular evolution, illustrating the MolGen-Transformer's ability to generate intermediate molecules as it interpolates between two input molecules in the latent space. These results further demonstrate the model's capability to explore and navigate the latent chemical space, producing a continuum of molecular structures.

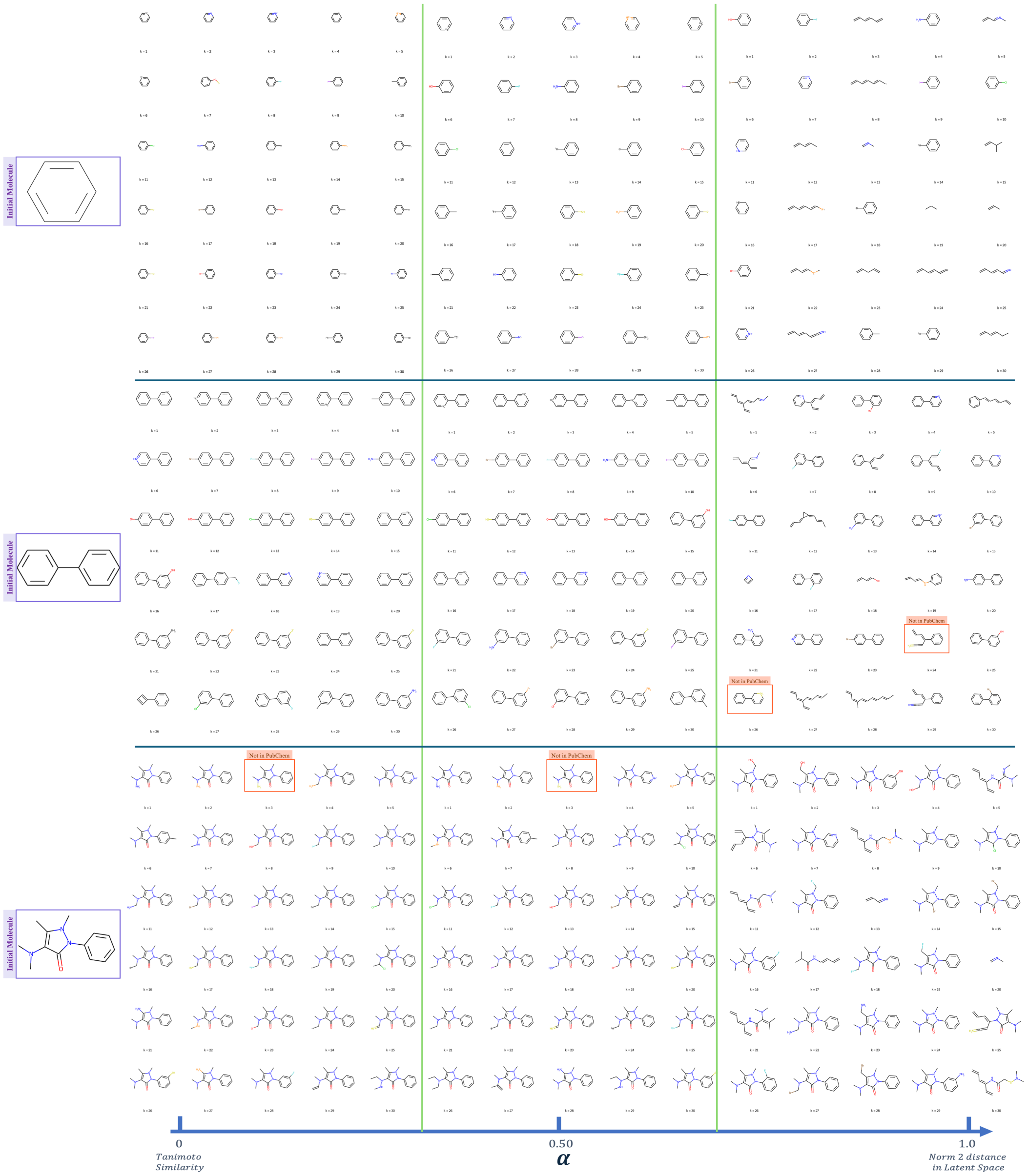


Figure 9: Additional results of local molecular generation, showing the MolGen-Transformer's ability to generate novel molecules similar to a given input, preserving structural features while introducing variations.

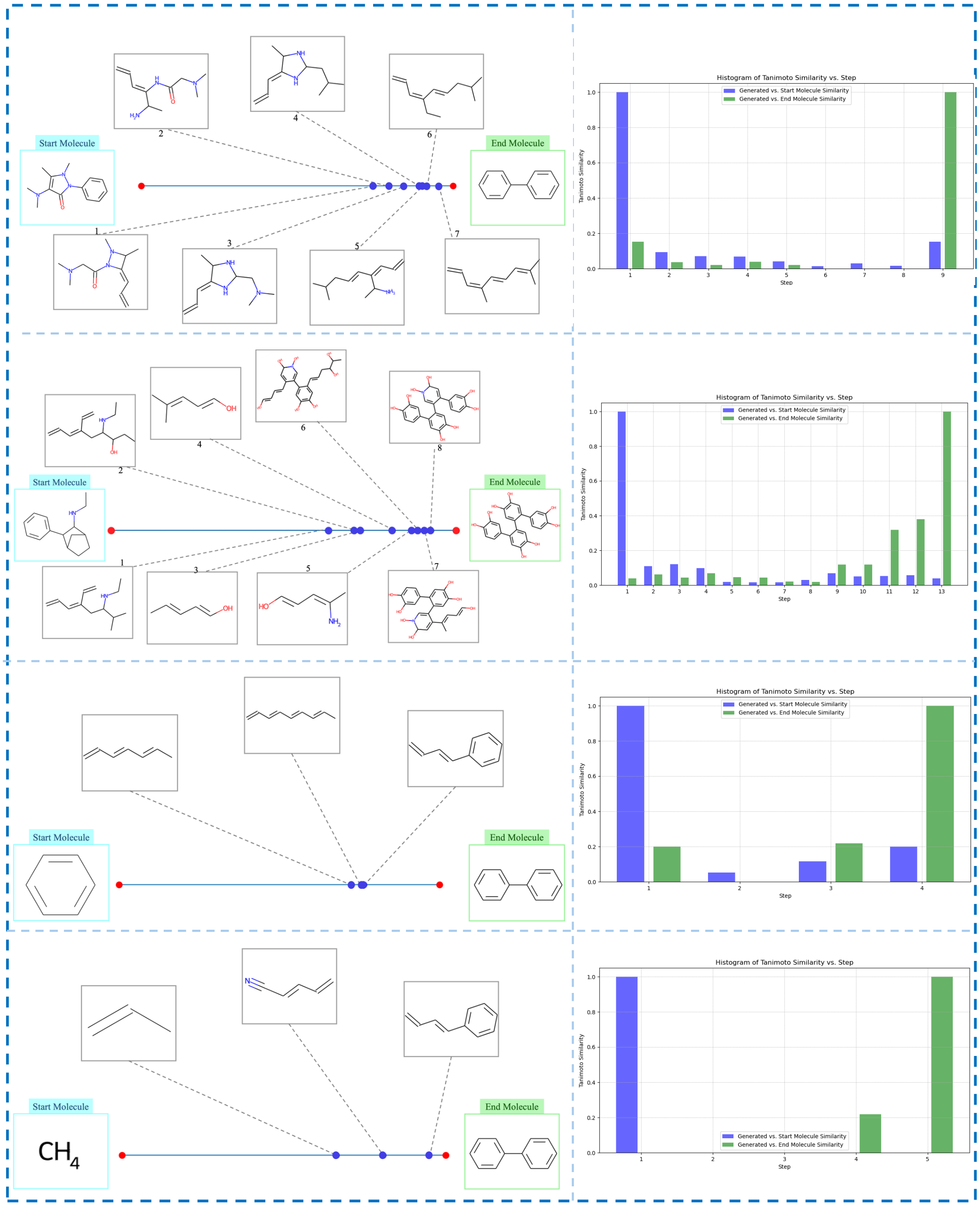


Figure 10: Additional results of Evolution: The figure illustrates the molecular evolution process, where the MolGen-Transformer generates intermediate molecules between two input molecules, showcasing the model's exploration of the latent chemical space.