1 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.

1.1 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 1 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.

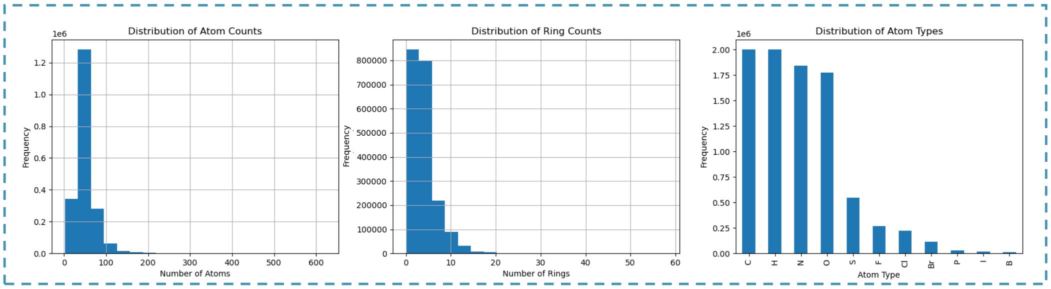

Figure 1: Distribution Analysis of the Meta Dataset Testing Set: The figure presents detailed distributions from a random sample of 2 million molecules within the testing set of the Meta dataset. The left panel shows the distribution of atom counts, indicating the frequency of molecules with varying 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.

1.2 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 2 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.

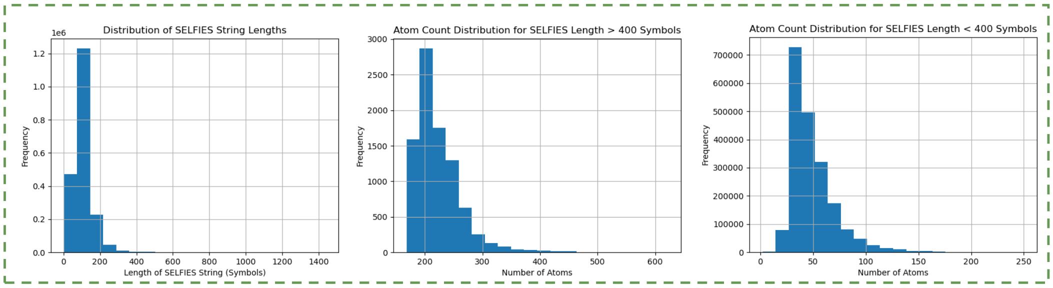

Figure 2: SELFIES Representation and Atom Count Analysis: This figure presents the distribution of SELFIES string lengths and corresponding atom counts within a random sample of 2 million molecules from the Meta dataset testing set. (a) Distribution of SELFIES string lengths, providing insights into 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.

1.3 Additional Results of Local Molecular Generation Results

Figure 3 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.

1.4 Additional Results of Molecular Evolution and Generation

Figures 4 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.

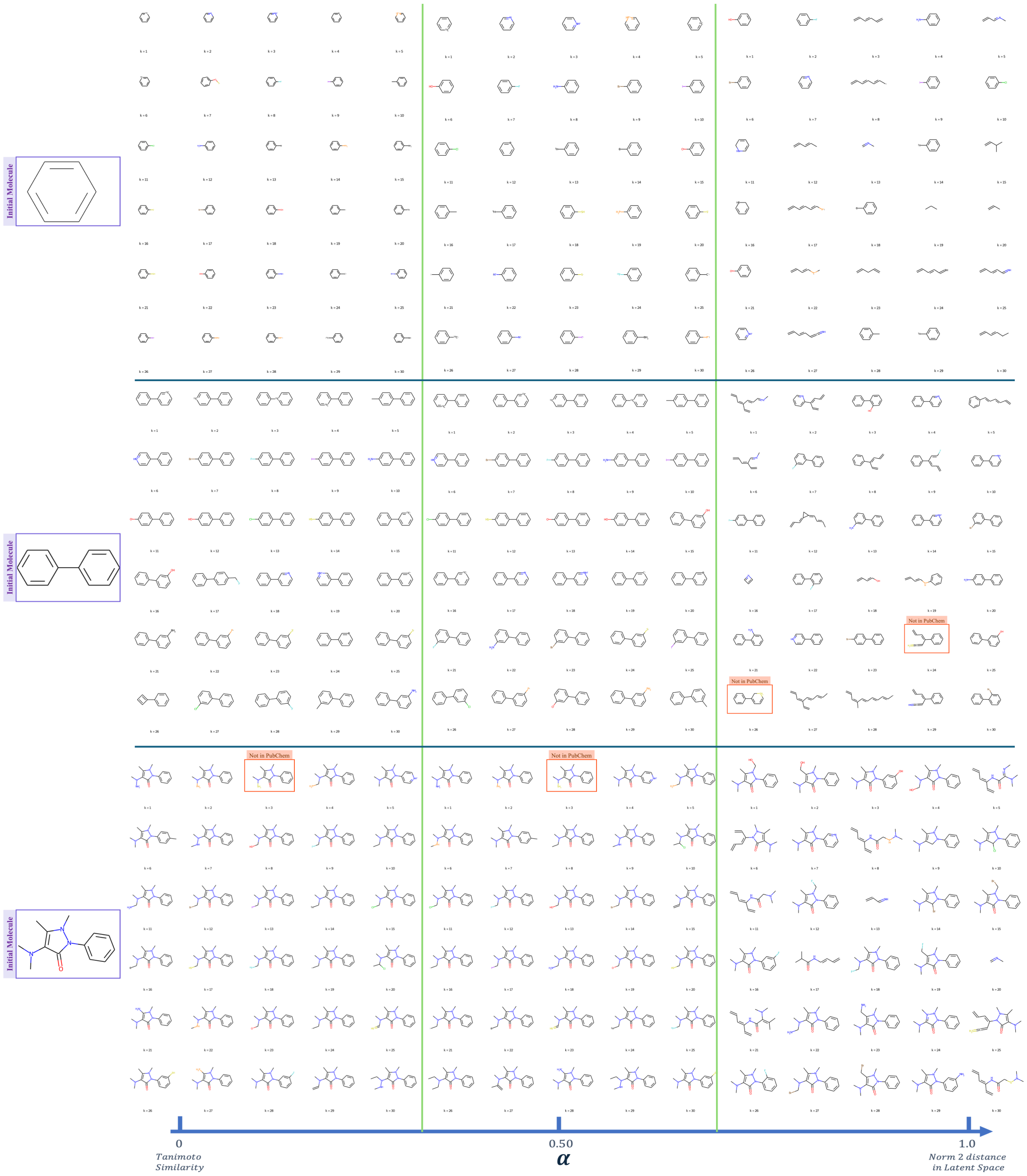

Figure 3: Additional results of local molecular generation, showing the MolGen-Transformer's ability to generate novel molecules similar to a given input, preserving structural features while introducing variations.

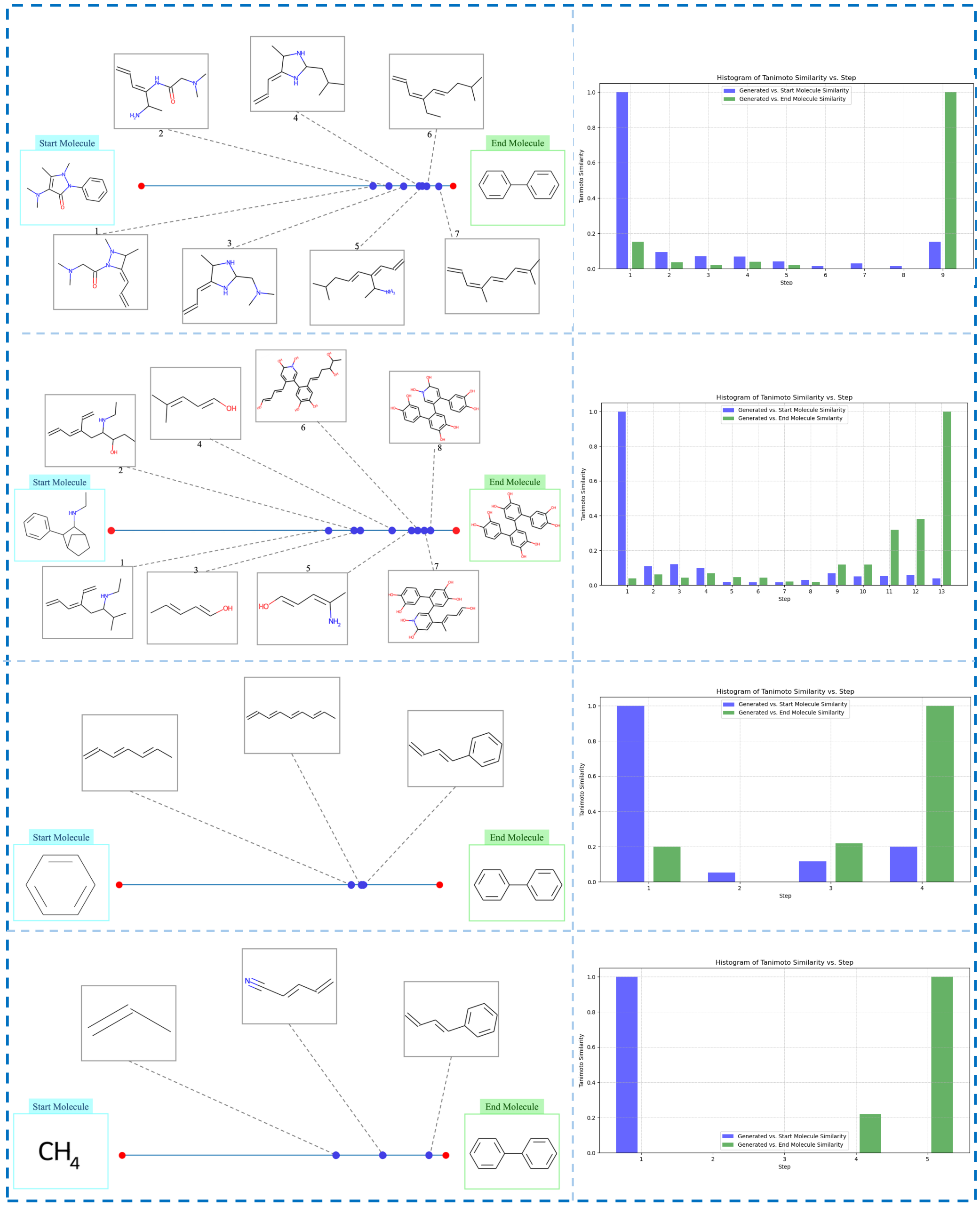

Figure 4: Additional results of Evolution: The figure illustrates the molecular evolution process, where the MolGen-Transformer generates intermediate molecules between two input molecules, showcasing the model's exploration of the latent chemical space.