**Supplementary Section 1: Evaluation on Molecular Encoding Bias** 2
**Supplementary Section 2: Dataset Construction and Composition** 2
**Supplementary Section 3: Prompt Format of HME** 4
**Supplementary Section 4: Detailed Results of the Protein-Ligand Affinity Prediction Task** 4
**Supplementary Section 5: Detailed Results of the Molecular Property QA Task** 5
**Supplementary Section 6: Ablation Study for HME** 6
**Supplementary Section 7: Visualization of Molecular Fragments** 7
**Supplementary Section 8: Discussion about Molecular Fragment** 7

## Supplementary Section 1: Evaluation of Molecular Encoding Bias

We used 2,000 molecules from the test dataset of the captioning task to measure encoding bias. Molecular 1D, 2D, and 3D features were extracted using SciBERT, GIN, and UniMol, respectively. Global molecular features were obtained by applying mean pooling. Morgan fingerprints were extracted using RDKit, with 2,048 bits and a radius of 2. Additionally, we applied mean pooling to the first hidden layer features of HME to derive our molecular features. The statistics of molecular features are summarized in Supplementary Table. 1:

**Supplementary Table 1. Details of different molecular encodings.** There are significant differences in the dimensions and numerical types of different encodings.

| | Encoding | Dimension | Type |
|---|---|---|---|
| **1D** | SciBERT | 768 | Float Vector |
| **2D** | GIN | 300 | Float Vector |
| **3D** | UniMol | 512 | Float Vector |
| **Ours** | HME | 4096 | Float Vector |
| **Refer** | Morgan | 2048 | Int Vector |

Then we calculated the similarity matrix and analyzed the correlation coefficients between different matrices. Specifically, For 1D, 2D, 3D, and our features, we employed cosine similarity to measure the similarity between molecular pairs $[\mathcal{M}_i, \mathcal{M}_j]$ from the molecule dataset $\mathcal{M}$:

$$\text{sim}_{cos}(\mathcal{M}_i, \mathcal{M}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i|| \cdot ||\mathbf{v}_j||}, \tag{1}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ represent the feature vectors of molecules $\mathcal{M}_i$ and $\mathcal{M}_j$, respectively. For Morgan fingerprints, which are discrete integer values, we employed Tanimoto similarity instead:

$$\text{sim}_{tani}(\mathcal{M}_i, \mathcal{M}_j) = \frac{|\mathbf{f}_i \cap \mathbf{f}_j|}{|\mathbf{f}_i \cup \mathbf{f}_j|}, \tag{2}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ denote the fingerprint bit vectors of molecules $\mathcal{M}_i$ and $\mathcal{M}_j$. For each encoding method $k$, we construct a similarity matrix $\mathbf{S}^k$ for the molecule set $\mathcal{M}$:

$$\mathbf{S}^k = [s^k_{ij}]_{n \times n}, \quad s^k_{ij} = \text{sim}(\mathcal{M}_i, \mathcal{M}_j), \tag{3}$$

where $n$ is the total number of molecules in $\mathcal{M}$, and $s^k_{ij}$ represents the similarity between molecules $\mathcal{M}_i$ and $\mathcal{M}_j$ under encoding method $k$.

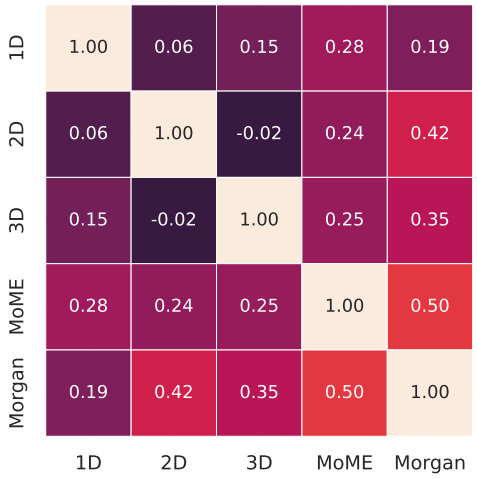
Finally, we analyzed the Pearson correlation coefficients between pairs of similarity matrices ($\mathbf{S}^p$ and $\mathbf{S}^q$) obtained from different encoding methods:

$$\rho(\mathbf{S}^p, \mathbf{S}^q) = \frac{\sum_{i,j}(s^p_{ij} - \bar{s^p})(s^q_{ij} - \bar{s^q})}{\sqrt{\sum_{i,j}(s^p_{ij} - \bar{s^p})^2}\sqrt{\sum_{i,j}(s^q_{ij} - \bar{s^q})^2}}, \tag{4}$$

where $\bar{s^p}$ and $\bar{s^q}$ denote the mean values of all elements in matrices $\mathbf{S}^p$ and $\mathbf{S}^q$, respectively. This correlation analysis helps us understand the consistency and differences among various molecular encoding strategies, as shown in Supplementary Fig. 1.

## Supplementary Section 2: Dataset Construction and Composition

In the molecular comprehension field, the 3D-MoIT dataset was adopted, which was constructed from PubChem and PubChemQC databases. 301K pairs were used for pretraining to align molecular and text modalities. For the captioning task, 12K molecule-text pairs were utilized as the training dataset and 2K as the test dataset. For molecular general QA, five QA pairs per molecule were generated using GPT-3.5 based on PubChem descriptions, focusing on physical-chemical properties, origins, and applications, resulting in 60K QA pairs for training and 10K for testing. The computational QA pairs were constructed in two parts: (1) four common properties from PubChem (molecular weight, LogP, TPSA, and complexity), yielding 46.7K QA pairs for training and 7.8K for testing; (2) four quantum chemical properties from PubChemQC (HOMO, LUMO, HOMO-LUMO Gap, and SCF Energy), resulting in 2.5M QA pairs for training and 312K for testing. More details about the dataset construction can be found in 3D-MoIT[1]. For the molecular docking score prediction task, a subset of the Tartarus dataset was adopted, which




**Supplementary Figure 1. Pearson Correlation between different molecular encodings.** Morgan: Morgan Fingerprints. Given a molecule set, we calculate the similarity matrix between molecular pairs using 5 different encodings (1D,2D,3D, our HME, and Morgan). Then we calculate the Pearson correlation among these 5 similarity matrices.

contains 100K molecules selected from the DTP database and their docking scores with protein PDB 4LDE. The dataset was split into training and test sets with a ratio of 9:1.

Molecular generation has become a cornerstone of modern drug discovery and material design, and the development of datasets that facilitate condition-driven molecule design plays a pivotal role in advancing this field. In this context, we introduce the MCMoD, a comprehensive resource tailored to support molecular design based on diverse conditioning factors, as shown in Supplementary Table .2. These conditions encompass textual descriptions, molecular properties (e.g., LogP, QED, SAS, and docking scores), and molecular fragments, providing a versatile framework for addressing a wide range of molecular design tasks. The target molecules in MCMoD span synthetic compounds, natural product-like molecules, and protein ligands, highlighting the dataset's broad applicability and relevance to both synthetic chemistry and biochemistry.

The MCMoD dataset integrates data from multiple high-quality molecular sources, each contributing unique strengths and diversity to the overall corpus. Specifically, we obtained 250K synthetic molecules from the ZINC-250K dataset. For natural product-like molecules, we incorporated 695K molecules from the COCONUT dataset, which is one of the most extensive repositories of natural compounds, enriched with property annotations. Additionally, we leveraged a curated set of 100K drug-like molecules from the Tartarus dataset, which includes docking scores obtained using QuickVina-2 for docking pose sampling and SMINA for redocking. To enable molecule design based on textual descriptions, we combined data from ChEBI-20 and PubChem, resulting in a combined set of 332K molecules paired with detailed textual descriptions. For all the obtained molecules, we used RDKit to calculate molecular properties including LogP, QED, and SAS values, which are crucial for practical applications such as drug design. We quantized LogP and SAS values with a step size of 1.0, and QED with a step size of 0.1, to help the CLMs better learn the chemical landscape implied by the magnitudes of these property values. We used our fragment generator to fragment each molecule. When fragment sequences serve as Chains of Thought (CoT), we sorted them lexicographically to mitigate the adverse effects of sequence order. When fragments serve as conditions, we deduplicated the fragments for each molecule and excluded fragments that are composed solely of carbon atoms and lack conjugated structures. Finally, we randomly selected 1 to 3 fragments to serve as the condition.

MCMoD is designed to support a diverse set of molecular design tasks, with a particular emphasis on the role of molecular fragments. Key tasks include description-based molecular generation, where models synthesize molecules conditioned on textual inputs; multi-objective molecular reverse design, where molecules are optimized to meet multiple property criteria; and affinity-based ligand generation, which involves designing ligands with high binding affinities for specific protein targets. Molecular fragments play a dual role across these tasks: as CoT, fragments are sequentially generated by models to refine the molecular design process, whereas, in prompt-based settings, fragments serve as explicit conditioning inputs, guiding the generative process toward specific structural motifs or functionalities. We formulated multiple conditions in natural language, which enhances the model's ability to generalize across diverse scenarios.

In summary, the MCMoD dataset represents a significant step forward in conditional molecular generation by integrating diverse molecular sources and advanced conditioning paradigms, offering researchers a rich and flexible resource to tackle

**Supplementary Table 2. Statistics of the MCMoD dataset.** Desc: Description. Prop: Property (LogP, QED, SAS, additional docking score for ligands in DTP). Frag: Fragment. *: Molecules that overlapped with the ChEBI test dataset were filtered out to prevent data leakage. CoT: The fragment sequences are generated by models as Chain of Thought. Prompt: The fragments are provided to models as conditions.

| Source | Size | Composition | Task | Role of Frag |
|---|---|---|---|---|
| PubChem | 299K* | Desc, Prop, Frag, SMILES | Description-based Molecular Generation | CoT |
| ChEBI | 33K | Desc, Prop, Frag, SMILES | Description-based Molecular Generation | CoT |
| ZINC | 250K | Prop, Frag, SMILES | Multi-objective Molecular Reverse Design | Prompt |
| COCONUT | 695K | Prop, Frag, SMILES | Multi-objective Molecular Reverse Design | Prompt |
| DTP | 100K | Prop, Frag, SMILES | Affinity-Based Ligand Generation | CoT |

**Supplementary Table 3. The prompt format of our HME on each task.** The content within {} represents placeholders that vary depending on each specific sample. {Features} is filled by concatenating embeddings, while others are directly filled with strings.

| Task | Prompt |
|---|---|
| Molecular Captioning | Please describe the molecule: Molecular geometric features are: {Features}. Molecular SMILES is {SMILES}. Molecular fragments are {Fragments}. |
| Molecular General QA | {Question}. Molecular geometric features are: {Features}. Molecular SMILES is {SMILES}. Molecular fragments are {Fragments}. |
| Molecular Property QA | {Question}. Molecular geometric features are: {Features}. Molecular SMILES is {SMILES}. Molecular fragments are {Fragments}. |
| Protein-Ligand Affinity Prediction | I am interested in the docking score of the molecule to Protein 4lde, could you tell me what it is? If uncertain, provide an estimate. Respond with the numerical value only. Molecular geometric features are: {Features}. Molecular SMILES is {SMILES}. Molecular fragments are {Fragments}. |
| Description-based Molecular Generation | Please give me molecular fragments based on the description. And then give me the molecular SMILES based on both the fragments and the description. The description is: {Description} |
| Multi-objective Molecular Reverse Design | There are some conditions, including logp (the hydrophobicity and solubility balance), qed (the drug-likeness), sas (the synthetic accessibility score), and the fragments (include specific fragments). Now please design a molecule under the given constraints: The molecule should have these fragments {Fragments}. The molecule should have a {Property Type} value of {Property Value}. |
| Affinity-Based Ligand Generation | Please give me molecular fragments based on the description. And then give me the molecular SMILES based on both the fragments and the description. The description is: The docking score of the molecule to Protein 4lde is {Value}. |

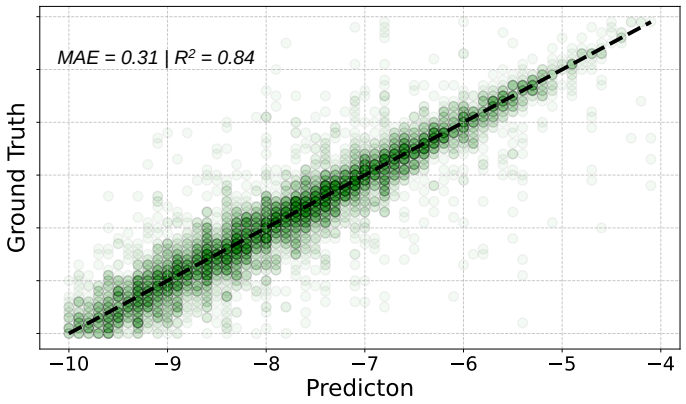
challenges in synthetic chemistry, natural product discovery, and ligand design.

## Supplementary Section 3: Prompt Format of HME

To enhance the reproducibility, we listed the prompts used for each task in Supplementary Table. 3. The content within is populated according to each sample. Notably, the Features are populated at the embedding level rather than the string level. In the Molecular General QA task, Question includes queries such as "What are some of the physical properties of 2-Phenylethylamine?" and "What are the roles of glycodihydroceramides in cellular processes?". In the Molecule Property QA task, Question encompasses queries such as "I need to know the LogP of this molecule, could you please provide it? If uncertain, provide an estimate. Respond with the numerical value only." and "Could you give me the HOMO-LUMO Gap value of this molecule? If uncertain, provide an estimate. Respond with the numerical value only."

## Supplementary Section 4: Detailed Results of the Protein-Ligand Affinity Prediction Task

This task is formulated as follows: the model receives molecular representations as input and predicts the docking score of the molecule with a given protein. We extract the numerical values of the properties using regular expressions and then compare them with the true values. As shown in Supplementary Fig. 2, most data points are distributed around the diagonal line (indicated by the dashed line), representing a perfect correlation between predictions and true values. The MAE is 0.31,




**Supplementary Figure 2. The distribution of groud-truth and HME's prediction for docking score.** The dotted line is the diagonal line, indicating that the ground truth value and the predicted value are equal. MAE denotes the mean absolute error; $R^2$ denotes the correlation coefficient.

**Supplementary Table 4. Performance for the molecular property QA task.** We propose the experimental results on our HME and the corresponding baselines including CLMs, LLMs, and Uni-Mol. MAE: Mean absolute error. Valid: The answer of the LMs contains property values and the values are within a reasonable range. $R^2$: The coefficient of determination measuring how well the predictions fit the ground truth, ranging from 0 to 1.

| Metrics | Model | Weight (g/mol) | LogP | TPSA (Å$^2$) | Complexity | HOMO (eV) | LUMO (eV) | H-L Gap (eV) | SCF ($10^4$eV) |
|---|---|---|---|---|---|---|---|---|---|
| MAE(↓) | Uni-Mol | 20.35 | 0.59 | 13.48 | 57.24 | 0.32 | 0.35 | 0.21 | 0.45 |
| | Llama2-7B | 22.10 | 1.45 | 15.87 | 69.74 | 1.24 | 1.04 | 0.88 | 0.70 |
| | 2D-MoLM | 21.48 | 0.88 | 13.52 | 55.74 | 0.92 | 0.80 | 0.67 | 0.71 |
| | 3D-MoLM | 14.79 | 0.66 | 9.71 | 44.85 | 0.26 | 0.25 | 0.28 | 0.35 |
| | MoLlama | **8.35** | **0.51** | **5.61** | **27.00** | **0.19** | **0.19** | **0.19** | **0.01** |
| Valid(↑) | Llama2-7B | 96% | 95% | 92% | 93% | 96% | 95% | 92% | 99% |
| | 2D-MoLM | 94% | 96% | 92% | 94% | 98% | 96% | 93% | 99% |
| | 3D-MoLM | 95% | 97% | 93% | 94% | 97% | 94% | 94% | 99% |
| | MoLlama | **99.90%** | **100%** | **99.90%** | **99.90%** | **100%** | **100%** | **100%** | **99.33%** |
| $R^2$(↑) | 3D-MoLM | 0.9796 | 0.9000 | 0.8583 | 0.9474 | 0.0588 | 0.5893 | 0.8366 | 0.9635 |
| | MoLlama | **0.9953** | **0.9692** | **0.9898** | **0.9871** | **0.6255** | **0.8138** | **0.9410** | **0.9994** |

**Supplementary Table 5. Value ranges for molecular properties.** We define the reasonable ranges for different molecular properties to determine the validity of model predictions. Sup and Inf represent the lower and upper bounds respectively. Numbers marked with * indicate closed intervals (inclusive bounds).

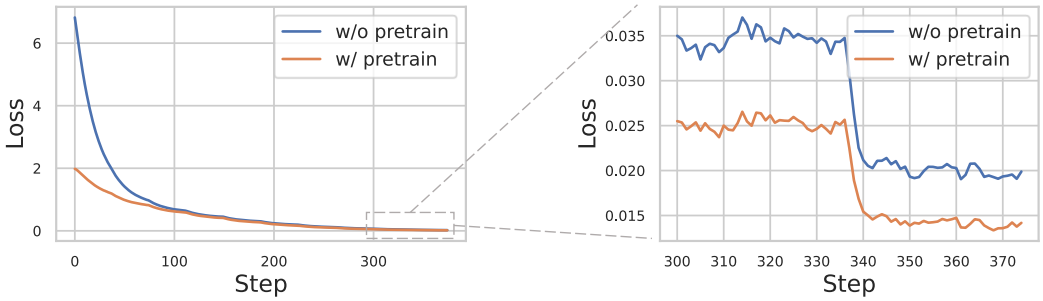
| | Weight (g/mol) | LogP | TPSA (Å$^2$) | Complexity | HOMO (eV) | LUMO (eV) | H-L Gap (eV) | SCF ($10^4$eV) |
|---|---|---|---|---|---|---|---|---|
| **Sup** | 0 | -30 | 0* | 0* | -20 | -20 | -20 | -50 |
| **Inf** | 4000 | 50 | 2000 | 10000* | 20 | 20 | 20 | 0 |

and $R^2$ is 0.84, indicating that the prediction model approximates true binding affinities well. This demonstrates that our HME has a certain understanding of protein-ligand interactions, facilitating drug screening.

## Supplementary Section 5: Detailed Results of the Molecular Property QA Task

In Supplementary Table. 4, we present the quantitative performance of our HME on the Property QA task. In addition to conventional metrics such as MAE and $R^2$, we introduce another metric, the Valid Ratio. It is necessary because LMs generate property predictions in natural language, which may deviate from expectations, such as failing to include the anticipated property value or producing values outside a reasonable range (e.g., a negative molecular weight). Specifically, a valid answer $\mathcal{A}$ is defined as follows: (1) $\mathcal{A}$ contains a property value $\mathcal{V}$. (2) $\mathcal{V}$ lies within a reasonable range, as shown in Supplementary Table. 5.

As shown in Supplementary Table. 4, HME demonstrates superior performance in answer validity compared to baseline models. It achieves a 100% validity rate for LogP, HOMO, LUMO, and H-L Gap, with nearly 100% validity for other




**Supplementary Figure 3. Comparison of the loss function curves with and without pretraining.** Two HME models are fine-tuned on the molecular captioning task; however, one underwent the first stage of pretraining, while the other did not.

**Supplementary Table 6. Performance for HME of different sizes.** HME-Small: The auto-regressive decoder is initialized by Llama-3.2-1B. HME-Medium: The auto-regressive decoder is initialized by Llama-3.2-3B. HME: The auto-regressive decoder initialized by Llama-3.0-8B.

| Task | Model | BLEU-2↑ | BLEU-4↑ | ROUGE-1↑ | ROUGE-2↑ | ROUGE-L↑ | METEOR↑ |
|---|---|---|---|---|---|---|---|
| Captioning | HME-Small | 32.23 | 23.94 | 39.98 | 24.25 | 34.31 | 35.69 |
| | HME-Medium | 32.83 | 24.38 | 41.36 | 25.54 | 35.60 | 36.99 |
| | HME | 36.26 | 27.79 | 43.01 | 27.58 | 37.27 | 39.02 |
| General QA | HME-Small | 44.43 | 35.89 | 50.76 | 34.80 | 46.44 | 53.69 |
| | HME-Medium | 51.28 | 42.42 | 55.72 | 39.82 | 51.29 | 57.75 |
| | HME | 54.94 | 46.41 | 58.94 | 43.59 | 54.90 | 60.15 |

properties. This indicates that HME successfully learns reasonable ranges of property values, reflecting its solid grasp of chemical knowledge. In terms of the Mean Absolute Error (MAE), HME achieves the lowest average error; however, MAE only reflects the overall error level and does not capture the correlation between predicted and true values. From the perspective of the $R^2$ metric, 3D-MoLM shows an $R^2$ value of merely 0.0588 for the HOMO property, suggesting that its predictions are almost uncorrelated with the ground truth and akin to random guesses. In contrast, HME achieves $R^2$ values exceeding 0.9 for six properties and above 0.6 for all eight, indicating its ability to accurately capture trends in molecular properties and establish strong linear correlations. These results highlight the critical support that HME provides for the exploration and analysis of chemical properties.

## Supplementary Section 6: Ablation Study for HME

**Different Sizes of HME.** We conduct experiments to evaluate the performance of HME variants with different model sizes on two fundamental tasks: molecular captioning and general QA. As shown in Supplementary Table. 6, we compare three versions: HME-Small, HME-Medium, and HME of the standard version. For the captioning task, HME-Small achieves BLEU-4 of 23.94 and ROUGE-L of 34.31, while HME-Medium improves these metrics to 24.38 and 35.6 respectively. HME further boosts the performance to BLEU-4 of 27.79 and ROUGE-L of 37.27. This trend is even more pronounced in the general QA task, where HME-Small obtains BLEU-4 of 35.89 and METEOR of 53.69, HME-Medium achieves 42.42 and 57.75, and HME significantly outperforms both with BLEU-4 of 46.41 and METEOR of 60.15. The performance gap between models is substantial, suggesting that a larger decoder enables HME to better capture the complex relationships between molecules and textual descriptions.

**Other Ablation Studies.** As shown in Supplementary Table. 7, the results demonstrate the importance of each component in our HME model. The baseline HME-Small achieves the best performance across all metrics, with a BLEU-4 score of 23.94 and a ROUGE-L score of 34.31. Removing different input modalities significantly impacts the model's performance, with the geometric features being the most crucial component—excluding them leads to a dramatic drop in performance (BLEU-4 decreases from 23.94 to 4.92). Similarly, removing SMILES representation and fragment tokens also results in substantial performance degradation, with BLEU-4 scores dropping to 10.92 and 19.14 respectively.

We also investigate the effect of pretraining. As shown in Supplementary Fig .3, we observe that during the first 100 steps, the model with pretraining achieves a lower loss on downstream tasks compared to the model without pretraining. Furthermore, at convergence, the model with pretraining achieves a lower loss value (0.014 vs. 0.020), which highlights the effectiveness of pretraining in molecule-text alignment. Moreover, as shown in Supplementary Table .7, the first-stage pretraining procedure

**Supplementary Table 7. Ablation studies for HME.** w/o SMILES: SMILES is not input to the model. w/o Geometry: Geometric features are not input to the model. w/o Fragment: Fragment tokens are not input to the model. w/o Pretrain: The first-stage pretraining procedure is not implemented. w/o Greedy: Probabilistic sampling is used instead of greedy sampling when decoding.

| Model | BLEU-2↑ | BLEU-4↑ | ROUGE-1↑ | ROUGE-2↑ | ROUGE-L↑ | METEOR↑ |
|---|---|---|---|---|---|---|
| HME-Small | 32.23 | 23.94 | 39.98 | 24.25 | 34.31 | 35.69 |
| HME-Small (w/o SMILES) | 18.61 | 10.92 | 29.60 | 13.96 | 24.26 | 23.81 |
| HME-Small (w/o Geometry) | 11.12 | 4.92 | 20.76 | 6.90 | 15.68 | 15.17 |
| HME-Small (w/o Fragment) | 27.72 | 19.14 | 36.55 | 20.45 | 30.87 | 31.87 |
| HME-Small (w/o Pretrain) | 29.59 | 21.57 | 38.25 | 22.78 | 32.84 | 34.01 |
| HME-Small (w/o Greedy) | 27.91 | 19.58 | 35.75 | 20.09 | 30.22 | 31.29 |

contributes to the model's performance, as evidenced by a decrease in BLEU-4 score from 23.94 to 21.57 when this stage is removed. These results demonstrate the critical role of pretraining in enhancing the model's ability to model chemical-linguistic space, thereby improving both the learning efficiency and the overall performance.

During next-token prediction, our HME employs a greedy sampling strategy. As shown in Supplementary Table. 7, we observe that when probabilistic sampling (e.g., nucleus sampling) was used instead of greedy sampling, the model's performance on text generation tasks decreased. For instance, the BLEU-4 score dropped from 23.94 to 19.58, the ROUGE-1 score decreased from 39.98 to 35.75, and the ROUGE-L score declined from 34.31 to 30.22. This indicates that greedy sampling is more effective in generating molecular description texts, enabling the model to produce more accurate textual outputs.

## Supplementary Section 7: Visualization of Molecular Fragments

As shown in Supplementary Fig .4, we randomly selected 50 fragments from our fragment vocabulary for visualization. This vocabulary was constructed using a subset of 456K molecules from the ChEMBL database and contains 800 entries. It was used in all tasks except for the multi-objective molecular reverse design task. In the case of the multi-objective molecular reverse design task, due to the significantly higher number of atoms and rings in natural product molecules compared to drug-like molecules, we constructed a separate vocabulary of size 800 based on a mixed subset of 100K natural product molecules and 100K drug-like molecules, as shown in Supplementary Fig. 5.

## Supplementary Section 8: Discussion about Molecular Fragment

Fragment-Based Drug Design (FBDD) is a pivotal strategy in drug discovery, where the core concept is to first identify favorable molecular fragments and then combine them into lead molecules[2]. A critical step in FBDD is the fragmentation of molecules. During the development of HME, we explored two types of fragmentation methods: rule-based fragmentation[3,4] and graph-based fragmentation[5,6].

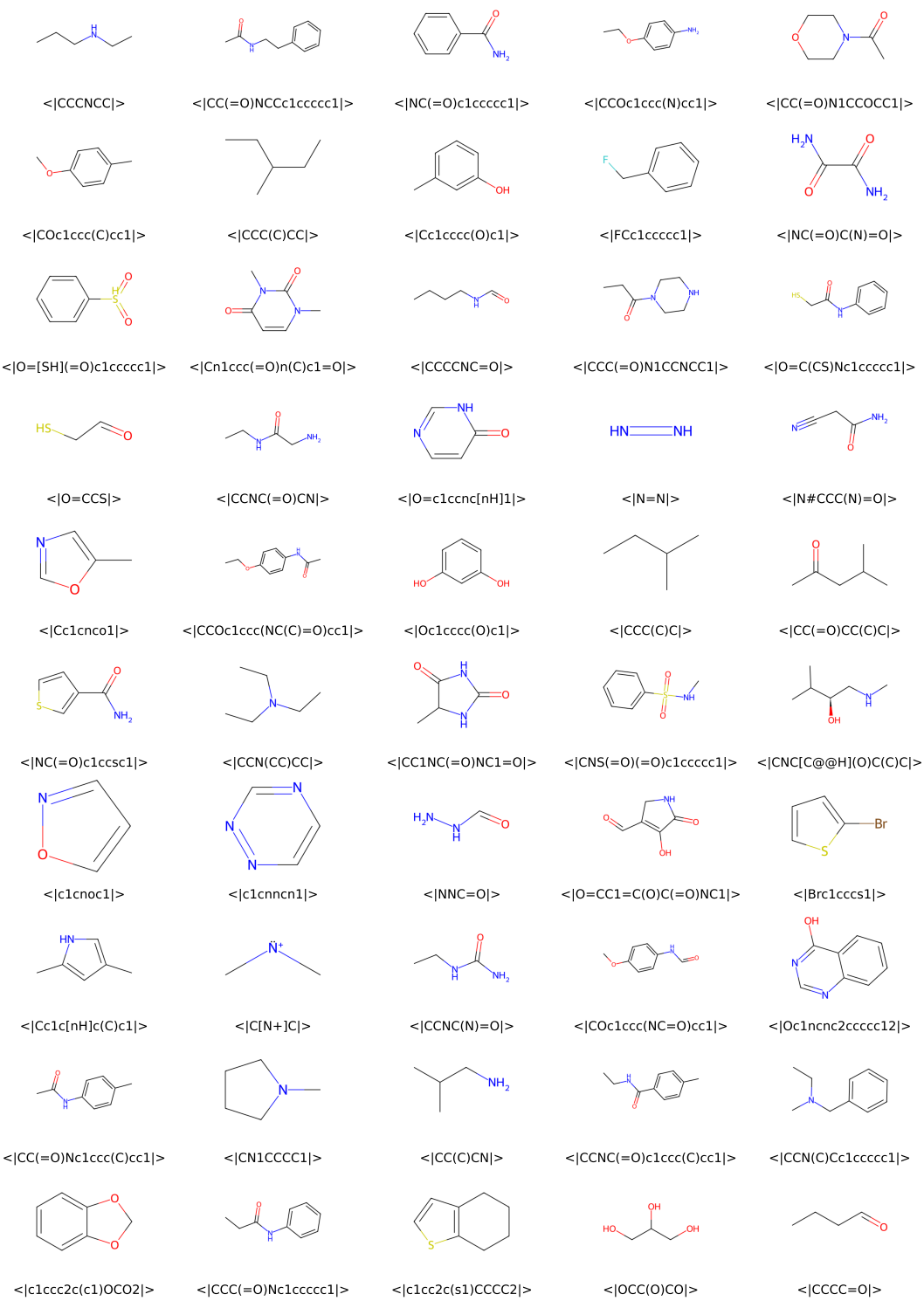
Rule-based fragmentation offers certain advantages, such as ease of interpretation and implementation, but the novelty of the resulting fragments is limited[2]. Moreover, as observed in our exploration, these methods produce an excessively large and variable-sized fragment vocabulary, which poses challenges for the application of LMs. Specifically, it becomes difficult to map a fragment to a single token within LMs. Although previous researchers have developed fragment-based molecular linear representations using these methods[7], these fragments are still not based on vocabulary IDs and the representations often rely on complex grammatical rules. We are concerned that this could lead to significant conflicts with the internal knowledge of LLMs and increase the learning difficulty. Therefore, we did not use these methods in our HME.

Graph-based fragmentation derives a fixed-size fragment vocabulary by mining high-frequency subgraphs (i.e., motifs), and the novelty of these fragments aids in the exploration of unknown chemical spaces[6]. By expanding the vocabulary of LLMs, the model gains the ability to understand and generate fragment tokens while avoiding confusion with natural language tokens. Although this approach also has the limitation of Out of Vocabulary (OOV), it can be easily addressed by adding single-atom fragments to the vocabulary. Therefore, our HME adopts this approach.

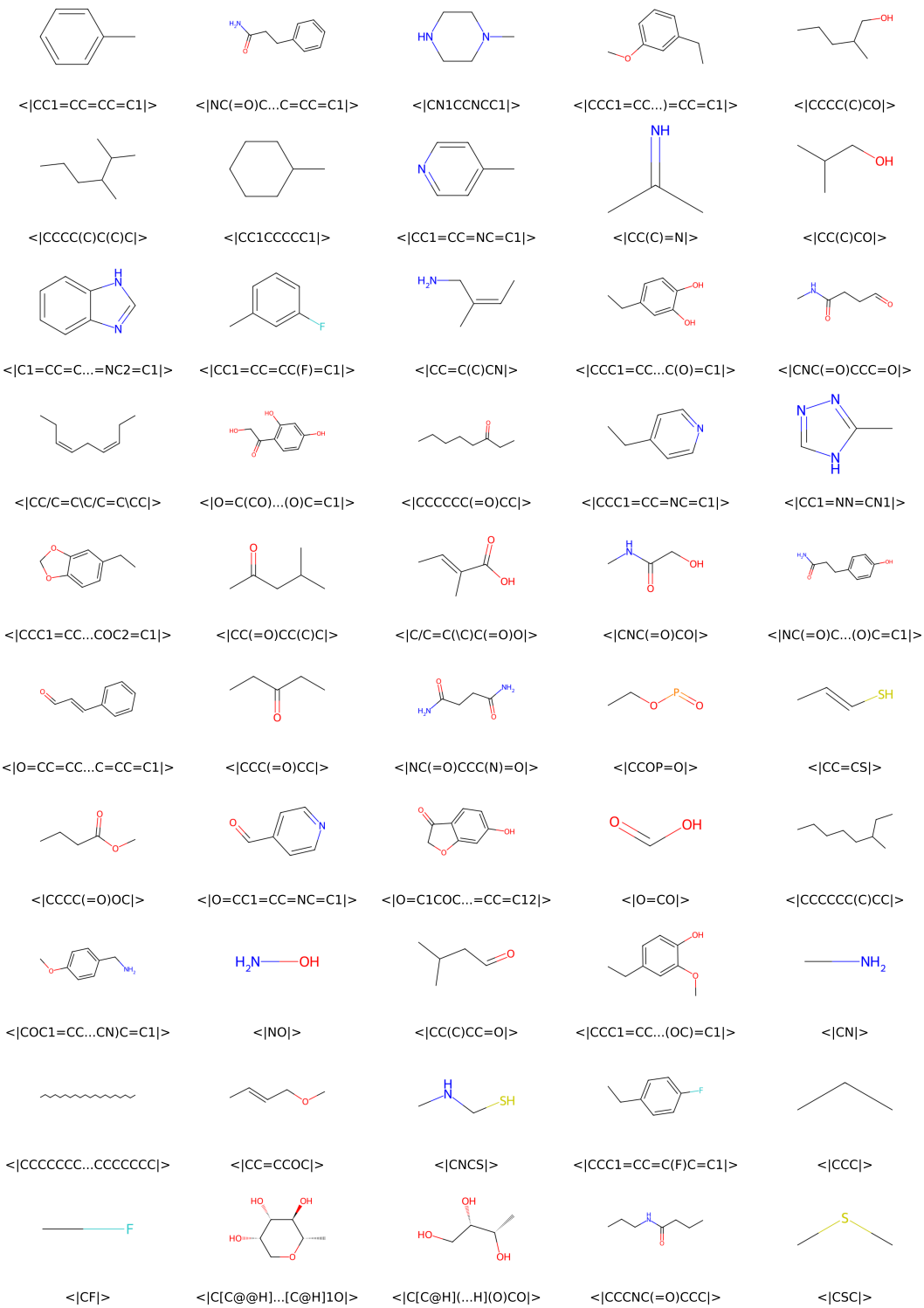
However, graph-based molecular fragmentation methods exhibit significant limitations: they focus solely on the topological structure of molecules (i.e., the connectivity between atoms) while neglecting many crucial chemical properties, such as atom types, chirality, and bond types.

Recently, the development in the field of computer vision has attracted our attention[8]. In the domain of image generation, researchers have developed a framework based on Vector Quantized Variational AutoEncoder (VQ-VAE)[9], which has demonstrated exceptional performance and flexibility. The working principle of this framework involves three stages: first, the encoder of VQ-VAE transforms an image into a finite set of discrete codes (analogous to IDs in a vocabulary); second, an auto-regressive model learns how to generate code sequences; finally, the decoder of VQ-VAE reconstructs the original image from the

discrete codes. Inspired by this technology, we sought to apply VQ-VAE to the field of molecular fragmentation, aiming to simultaneously account for both the topological structure and the chemical properties of molecules. In our experiments, we successfully transformed attributed molecular graphs into discrete codes. However, we encountered significant challenges in the attempt to reconstruct the original molecules from these discrete codes. Additionally, we found that a molecule with $n$ atoms is typically encoded by VQ-VAE into more than $n$ codes, which contradicts our initial goal of using VQ-VAE for fragmentation. In conclusion, whether VQ-VAE can be utilized effectively for molecular fragmentation remains an open question that requires collaborative efforts from the community.




**Supplementary Figure 4. Visualization of random 50 fragments in our fragment vocabulary.** The SMILES strings are also provided, with the prefix "<|" and the suffix "|>" used to indicate they are fragment tokens.




**Supplementary Figure 5. Visualization of random 50 fragments in our fragment vocabulary for the multi-objective molecular reverse design task.** The SMILES strings are also provided, with the prefix "<|" and the suffix "|>" used to indicate they are fragment tokens.

## References


1. Li, S. *et al.* Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923* (2024).
2. Yang, R., Zhou, H., Wang, F. & Yang, G. Digfrag as a digital fragmentation method used for artificial intelligence-based drug design. *Commun. Chem.* **7**, 258 (2024).
3. Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem* **3**, 1503 (2008).
4. Lewell, X. Q., Judd, D. B., Watson, S. P. & Hann, M. M. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. chemical information computer sciences* **38**, 511–522 (1998).
5. Kong, X., Huang, W., Tan, Z. & Liu, Y. Molecule generation by principal subgraph mining and assembling. *Adv. Neural Inf. Process. Syst.* **35**, 2550–2563 (2022).
6. Geng, Z. *et al.* De novo molecular generation via connection-aware motif mining. In *The Eleventh International Conference on Learning Representations* (2023).
7. Wu, J.-N. *et al.* t-smiles: a fragment-based molecular representation framework for de novo ligand design. *Nat. Commun.* **15**, 4993 (2024).
8. Ramesh, A. *et al.* Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831 (Pmlr, 2021).
9. Van Den Oord, A., Vinyals, O. *et al.* Neural discrete representation learning. *Adv. neural information processing systems* **30** (2017).