Hierarchical Structure-Property Alignment for Data-Efficient Molecular Generation and Editing

Ziyu Fan[1], Zhijian Huang[1], Yahan Li[1], Xiaowen Hu[1], Siyuan Shen[1], Yunliang Wang[2], Zeyu Zhong[1], Shuhong Liu[1], Shuning Yang[1], Shangqian Wu[1], Min Wu[3*], Lei Deng[1*]

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]School of Software, Xinjiang University, Wulumuqi 830000, China
[3]Institute for Infocomm Research, Agency for Science, Technology and Research (A* STAR), 138632, Singapore
fzychina@csu.edu.cn, wumin@a-star.edu.sg, leideng@csu.edu.cn

Abstract

Property-constrained molecular generation and editing are crucial in AI-driven drug discovery but remain hindered by two factors: (i) capturing the complex relationships between molecular structures and multiple properties remains challenging, and (ii) the narrow coverage and incomplete annotations of molecular properties weaken the effectiveness of property-based models. To tackle these limitations, we propose HSPAG, a data-efficient framework featuring hierarchical structure–property alignment. By treating SMILES and molecular properties as complementary modalities, the model learns their relationships at atom, substructure, and whole-molecule levels. Moreover, we select representative samples through scaffold clustering and hard samples via an auxiliary variational auto-encoder (VAE), substantially reducing the required pre-training data. In addition, we incorporate a property relevance-aware masking mechanism and diversified perturbation strategies to enhance generation quality under sparse annotations. Experiments demonstrate that HSPAG captures fine-grained structure–property relationships and supports controllable generation under multiple property constraints. Two real-world case studies further validate the editing capabilities of HSPAG.

Code — https://github.com/ZiyuFanCSU/HSPAG
Extended version — https://arxiv.org/abs/2511.08080

Introduction

Molecular generation has become a cornerstone of Artificial Intelligence for Drug Discovery (AIDD), enabling the direct design of novel compounds (Jones et al. 2024; Sadybekov and Katritch 2023). These models explore vast chemical spaces, offering clear advantages over traditional virtual screening (Gottipati et al. 2020). Notably, rather than merely producing chemically valid molecules, drug discovery increasingly demands compounds that satisfy diverse property requirements (Yang et al. 2022; Pang et al. 2023).

Existing conditional generation methods employ diverse architectures (Jin, Barzilay, and Jaakkola 2018, 2020; Zhu et al. 2023; Hou et al. 2022; Wang et al. 2022; Dobberstein, Maass, and Hamaekers 2024), typically incorporating properties either as auxiliary inputs or as optimization targets (Zhang et al. 2025). For example, LIMO (Eckmann et al. 2022) integrates QED and LogP into the decoder to guide molecule generation toward desired properties, while MOLGEN (Fang et al. 2024) employs a feedback-driven framework that iteratively refines generated molecules using chemical evaluations, enabling broad property optimization. Besides, inspired by recent efforts that seek to build unified molecular representations across multiple modalities (Shen et al. 2024; Chen et al. 2023; Cao et al. 2023; Liu et al. 2024), approaches such as SPMM (Chang and Ye 2024) treat molecular properties as an additional modality and align them with molecules to guide generation. However, existing methods often align different modalities at a global level (Huang et al. 2024; Radford et al. 2021). Such coarse-grained alignment overlooks intricate correspondences at the atom and substructure levels, which are critical for accurately capturing structure–property relationships.

Moreover, current property-constrained molecular generation often depends on a limited set of predefined property features, restricting the exploration of diverse chemical properties, especially for ADMET (Dowden and Munro 2019; Pammolli, Magazzini, and Riccaboni 2011). However, acquiring large-scale datasets with property annotations remains challenging due to high annotation costs. To alleviate the data scarcity challenge, active learning (Settles 2009) has emerged as a data-efficient paradigm, iteratively selecting and annotating the most informative samples. Recent works have applied it to molecular tasks, such as KDBNet (Luo, Liu, and Peng 2023) for kinase–drug interaction discovery and BayesianGeoGNN (Subedi et al. 2024) for 3D molecular graphs. These studies demonstrate the potential of active learning to improve data efficiency in specific molecular tasks. However, to our knowledge, no existing methods have been designed to integrate active learning with property-guided molecular representation learning and generation tasks.

To this end, we propose **HSPAG**, a data-efficient framework for **H**ierarchical **S**tructure–**P**roperty **A**lignment in molecular **G**eneration and editing. We propose a hierarchi-

*Corresponding authors

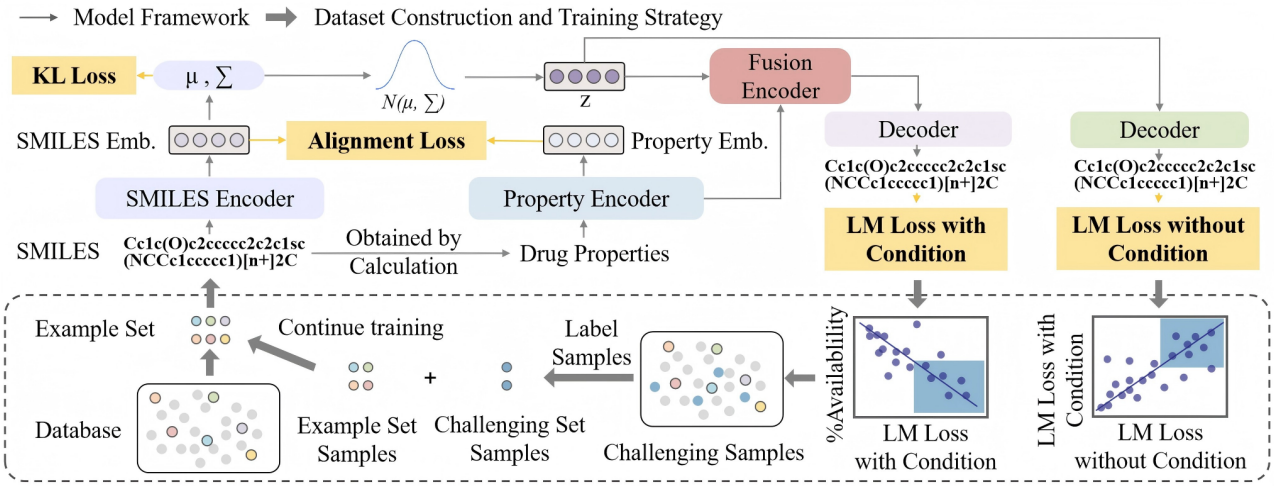

Figure 1: Overview of the HSPAG. The upper part illustrates the model architecture, comprising five modules: SMILES encoder, property encoder, substructure-property alignment module, conditional decoder, and unconditional decoder. The unconditional decoder is used for data sampling. The lower part presents the representative subset construction and training strategy.

cal contrastive learning module to more effectively model fine-grained structure–property dependencies. To improve data efficiency in settings with limited property annotations, we sample the data into example and challenging sets for pre-training. This allows the model to focus on learning from representative samples, while significantly reducing the amount of data. Our key contributions are summarized as follows:

- We propose hierarchical alignment spanning atoms, fragments, and molecules to capture fine-grained structure–property relationships and address strongly coupled multi-property scenarios.
- We introduce an active-learning-inspired subset selection strategy coupled with progressive curriculum training, which explicitly balances structural diversity and information density under limited annotations.
- We develop a correlation-aware property masking module with structured regularization, which exploits inter-property correlations to mitigate information leakage and biases introduced by random masking.
- Our method achieves state-of-the-art performance, with interpretability analysis confirming that the learned representations balance structural and property similarity and transfer well to tasks like drug repurposing.

Methodology

This section describes the components of HSPAG. An overview of the framework is shown in Fig. 1, and the algorithmic details are provided in the Appendix B.

Encoders for SMILES and Molecular Properties

We adopt a unified Transformer-based architecture to encode both SMILES and property values. For SMILES, the input is tokenized at the character level (e.g., atom symbols, bonds), embedded into fixed-dimensional vectors, and processed with positional encodings before passing through the Transformer encoder to obtain contextualized molecular representations. For properties, each real-valued attribute is first normalized, linearly projected into an embedding space, and augmented with positional encodings before being encoded by a Transformer to model dependencies among properties.

Hierarchical Modality Alignment with Data Augmentation

We introduce a cross-modal alignment module that maps SMILES and properties into a shared latent space, enhanced by a hierarchical contrastive mechanism to strengthen their alignment. As shown in Fig. 2 (b), each SMILES is represented at three levels: global, substructure, and atom, while each property is represented at global and local levels. Using $\mathcal{V}_S = \{S^{\text{global}}, S^{\text{sub}}, S^{\text{atom}}\}$ and $\mathcal{V}_P = \{P^{\text{global}}, P^{\text{local}}\}$ to represent. We define the cross-level similarity from SMILES i to property j as:

$$S_{\text{S}\to\text{P}}(i,j) = \frac{1}{|v^S(i)|} \sum_{t=1}^{|v^S(i)|} \max_{k=1}^{|v^P(j)|} \cos\left(v_t^S(i), v_k^P(j)\right), \tag{1}$$

where $v^S \in \mathcal{V}_S, v^P \in \mathcal{V}_P$, $v_t^S(i)$ denotes the t-th vector dimension of SMILES i's representation, and $v_k^P(j)$ denotes k-th vector dimension of property j's representation. Similarly, the cross-level similarity from property i to SMILES j can be obtained:

$$S_{\text{P}\to\text{S}}(i,j) = \frac{1}{|v^P(i)|} \sum_{k=1}^{|v^P(i)|} \max_{t=1}^{|v^S(j)|} \cos\left(v_k^P(i), v_t^S(j)\right). \tag{2}$$

These similarities are used in the InfoNCE loss to perform bidirectional contrastive alignment over all pairwise combi-

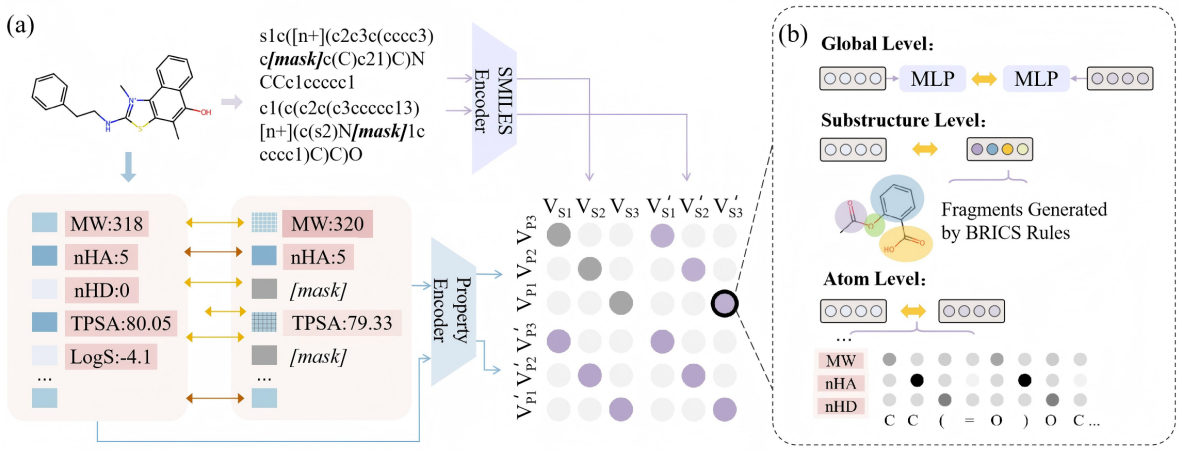

Figure 2: Cross-modal contrastive alignment of SMILES and property representations. (a) Data augmentation strategies applied to both SMILES and property modalities. (b) Hierarchical alignment module that captures hierarchical structure-property correspondences at the atom, substructure, and global levels.

nations across hierarchical levels:

$$\mathcal{L}_{\text{InfoNCE}}^{\text{S}\to\text{P}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(S_{\text{S}\to\text{P}}(i,i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(S_{\text{S}\to\text{P}}(i,j)/\tau\right)}, \quad (3)$$

$$\mathcal{L}_{\text{InfoNCE}}^{\text{P}\to\text{S}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(S_{\text{P}\to\text{S}}(i,i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(S_{\text{P}\to\text{S}}(i,j)/\tau\right)}, \quad (4)$$

where N is the batch size. Details of the InfoNCE loss are provided in Appendix B. The Hierarchical Contrastive Alignment Loss can be formulated as the following:

$$\mathcal{L}_{\text{MLA}}^{\mathcal{V}_S\leftrightarrow\mathcal{V}_P} = \mathcal{L}_{\text{InfoNCE}}^{S^{\text{global}}\leftrightarrow P^{\text{global}}} + \mathcal{L}_{\text{InfoNCE}}^{S^{\text{atom}}\leftrightarrow P^{\text{local}}} + \mathcal{L}_{\text{InfoNCE}}^{S^{\text{sub}}\leftrightarrow P^{\text{local}}}. \quad (5)$$

To improve data efficiency under limited annotations, we introduce tailored augmentation strategies for both molecular modalities, as shown in Fig. 2 (a). For SMILES, we generate multiple syntactically valid variants by enumerating alternative traversal orders within the canonical grammar, and apply token-level masking. For properties, we enhance robustness and mimic real-world sparsity by (i) injecting small perturbations and (ii) randomly masking property dimensions. To avoid shortcut learning from redundant correlations, we further mask all strongly correlated properties together, encouraging more discriminative learning from incomplete inputs.

Let $\mathcal{V}_S^{'}$ and $\mathcal{V}_P^{'}$ denote the two augmentations from the SMILES and property. The total contrastive loss is given by:

$$\mathcal{L}_{\text{align}} = \text{mean}\big(\mathcal{L}_{\text{MLA}}^{\mathcal{V}_S\leftrightarrow\mathcal{V}_P} + \mathcal{L}_{\text{MLA}}^{\mathcal{V}_S^{'}\leftrightarrow\mathcal{V}_P} + \mathcal{L}_{\text{MLA}}^{\mathcal{V}_S\leftrightarrow\mathcal{V}_P^{'}} + \mathcal{L}_{\text{MLA}}^{\mathcal{V}_S^{'}\leftrightarrow\mathcal{V}_P^{'}}\big). \quad (6)$$

Decoder with Bidirectional Cross Attention Fusion

To generate molecular sequences conditioned on target properties, we employ a conditional variational autoencoder (CVAE) framework, which includes two losses: a KL divergence regularization and a reconstruction loss. Detailed variational inference formulation is provided in Appendix B.

To regularize the latent representation, we compute the KL divergence between the approximate posterior $q_\phi(z|S,P) = \mathcal{N}(\mu_q,\sigma_q^2)$ and the conditional prior $p(z|P) = \mathcal{N}(\mu_p,\sigma_p^2)$, which is given by:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2}\sum_{i=1}^{d}\Big(\log\frac{\sigma_p^2(i)}{\sigma_q^2(i)} - 1 + \frac{\sigma_q^2(i)+(\mu_q(i)-\mu_p(i))^2}{\sigma_p^2(i)}\Big), \quad (7)$$

where d is the latent dimension, $\mu_q(i)$ and $\sigma_q^2(i)$ are the i-th elements of the approximate posterior mean and variance, and $\mu_p(i)$ and $\sigma_p^2(i)$ are those of the conditional prior.

Prior to decoding, we introduce a bidirectional cross-attention fusion mechanism to enhance the interaction between the latent representation and property embeddings. Specifically, given the latent representation $Z \in \mathbb{R}^{l_z\times d}$, obtained from the variational inference process, and property representation $V^P \in \mathbb{R}^{l_p\times d}$, we treat Z and V^P as both queries and contexts:

$$H_S = \text{Attn}\left(ZW_Q^z, V^PW_K^p, V^PW_V^p\right), \quad (8)$$

$$H_P = \text{Attn}\left(V^PW_Q^p, ZW_K^z, ZW_V^z\right), \quad (9)$$

where $W_Q^z, W_K^z, W_V^z \in \mathbb{R}^{d\times d_h}$ and $W_Q^p, W_K^p, W_V^p \in \mathbb{R}^{d\times d_h}$ are the projection matrices for queries, keys, and values. H_S and H_P are concatenated and refined by a Transformer encoder into Z_{fused}.

Given the fused representation Z_{fused}, the decoder autoregressively models the conditional likelihood of the SMILES sequence $S = \{s_1,\ldots,s_L\}$. The reconstruction loss is defined as the cross-entropy between the predicted token distribution and the ground-truth sequence:

$$\mathcal{L}_{\text{LM}}^{\text{CVAE}} = -\sum_{j=1}^{L}\sum_{s} p_{\text{true}}(s|S_{<j}) \log p_\theta(s|Z_{\text{fused}}, S_{<j}; \theta), \tag{10}$$

where p_{true} is the one-hot distribution over s_j, and p_θ is the decoder's prediction.

Dataset Construction and Training Strategy

Conditional molecular generation faces high annotation costs and often restricted by limited tool access. Coupled with the resource demands of large-scale model training, this underscores the need for efficient strategies to reduce labeling overhead and enhance data utilization. To reduce annotation costs, we first construct a structurally diverse "example set" via scaffold clustering and redundancy filtering, which ensures broad structural coverage.

To further supplement the dataset, we construct a "**challenging set**" based on reconstruction difficulty. We observe molecules with high CVAE LM loss tend to yield poor generation performance (Pearson $r = 0.608$). Since CVAE requires labeled data, we instead use an auxiliary VAE without property input, whose LM loss closely correlates with that of the CVAE (Pearson $r = 0.856$). Based on this, we select high-loss molecules as challenging set. This strategy analogous to uncertainty-based sampling in active learning.

Given that VAE loss correlates with structural complexity (e.g., SMILES length, molecular weight), overloading these samples early disrupts training. We adopt a progressive sampling strategy: early epochs focus on diverse, moderately complex molecules; challenging samples are gradually introduced mid-training; and their ratio is reduced later to restore distribution balance. This approach improves robustness to complex structures while maintaining overall stability and generalization.

Training Objective Summary

Following the above analysis, the overall loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \beta \cdot \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{LM}}^{\text{CVAE}} + \mathcal{L}_{\text{LM}}^{\text{VAE}}. \tag{11}$$

Here, $\mathcal{L}_{\text{align}}$ is the contrastive loss for SMILES-property alignment. $\mathcal{L}_{\text{LM}}^{\text{CVAE}}$ and $\mathcal{L}_{\text{KL}}$ provides conditional generation losses, while $\mathcal{L}_{\text{LM}}^{\text{VAE}}$ adds reconstruction losses to support difficulty-aware sampling. All modules are trained end-to-end, with β controlling KL regularization strength.

Experiments

Data and Experimental Setup

Our training dataset is based on ChEMBL 24 (Zhu et al. 2023), and properties are evaluated using ADMETlab 3.0 (Fu et al. 2024) and RDKit. Finally, only about one-fifth of the original dataset was used. More details in Appendix A. The detailed description of the experimental setup, including dataset preprocessing, model configurations, and training protocols, can be found in Appendix C.

Real-World Distribution Capturing

A molecular generator should be close to the real distribution and produce effective, diverse, and realistic structures. Although HSPAG is conditional, we emulate unconditional generation by sampling property vectors from the training dataset (Zhu et al. 2023). The generation process begins with the [CLS] token and proceeds by greedily selecting the most probable token at each step.

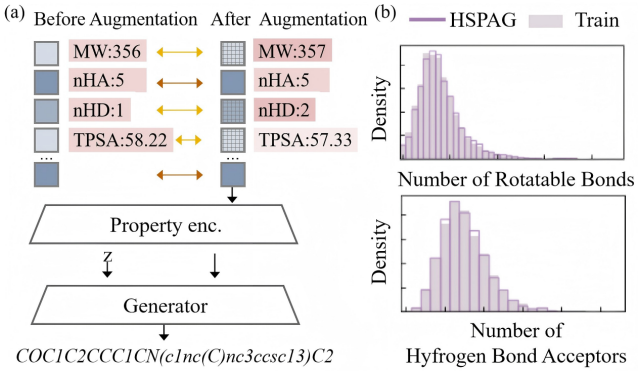

Figure 3: (a) Property-conditioned generation under perturbation. (b) Comparison of the molecular property distributions generated by HSPAG and the training set.

We introduce slight perturbations to the conditioning vectors during inference (e.g., changing molecular weight from 356 to 357), as shown in Fig 3(a), to enable broader chemical space exploration. Keep_ratio means how many properties remain unchanged, and by measures, the value was determined to be 90% (Appendix D). We generate 10,000 molecules and use eight well-established metrics(Liu et al. 2018). All baselines follow consistent training and sampling protocols (Appendix E).

As shown in Tab. 1 (left), HSPAG achieves the best Novelty and Availablility, with the latter improving by +4.26% over SPMM, highlighting its superior capability in generating structurally and chemically feasible molecules. Furthermore, HSPAG consistently delivers competitive results on distribution and diversity metrics, such as IntDiv and FCD, suggesting its effectiveness in capturing dataset statistics. The property distributions of the generated and training molecules, as shown in Fig. 3(b), provide additional evidence supporting the effectiveness of our method. In contrast, LIMO, despite perfect validity, shows poor alignment—supporting prior findings on the trade-off between syntactic robustness and semantic fidelity (Skinnider 2024).

Multi-Constraint Molecular Generation

In real-world scenarios such as drug discovery, generating diverse candidates that satisfy property constraints is crucial (Jin, Barzilay, and Jaakkola 2018). HSPAG promote diversity by sampling multiple latent vectors from the Gaussian prior and decoding tokens stochastically rather than greedily. The decoding strategies and metrics detailed in Appendix F. Results are reported in Tab. 1 (right).

Method	Molecular Distribution Learning							Conditional Molecule Generation				
	Val.↑	Nov.↑	Avail.↑	SNN↑	Frag.↑	FCD↓	IntDiv.↑	Val.↑	Nov.↑	Avail.↑	RMSE↓	Scaf.↑
MD-VAE (Kwon et al. 2023)	.9081	.9414	.8548	.3845	.9977	.5480	.8776	.8402	.9984	.7025	.2704	589.16
CProMG (Li et al. 2023)	.8455	.9447	.7978	.4708	.9954	.6970	.8593	.4984	.9967	.3670	.3316	355.77
AAE (Makhzani et al. 2015)	.8691	.9492	.8248	.3499	.9847	.5911	.8768	.7627	.9987	.7589	.3248	602.06
SPMM (Chang and Ye 2024)	.9211	.9363	.8614	.4584	.9980	.3780	.8094	.8963	.9991	.3720	.2989	309.53
LIMO (Eckmann et al. 2022)	**1.000**	.8170	.8100	.4485	.9754	.7925	**.8907**	**1.000**	**.9998**	.3183	.8202	225.03
Chemformer (Irwin et al. 2022)	.8768	.9218	.8075	.4835	.9887	.5746	.8768	.5326	.9979	.4156	.3072	371.44
HSPAG	.9483	**.9536**	**.9040**	**.4855**	**.9999**	**.3074**	.8843	.9055	**.9998**	**.7825**	**.2630**	**629.38**

Table 1: Performance comparison on property-conditioned molecular generation. The best results are highlighted in bold, and the second best results are highlighted in underlined bold. Due to the high cost involved, RMSE is reported only for properties that can be efficiently computed with RDKit, and a comprehensive analysis over all properties is provided in the Appendix F.

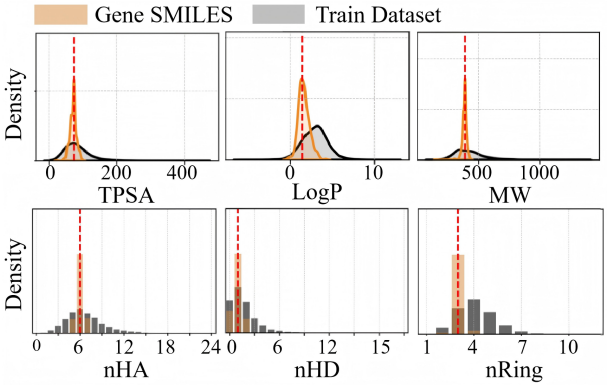

Figure 4: Property distribution of the generated molecules (orange) and training dataset (gray).

In conditional generation, HSPAG maintains a strong overall balance across multiple metrics, while still outperforming baselines in Novelty and Availablility. Its superior scaffold diversity highlights strong scaffold hopping ability, crucial for lead optimization. Additionally, low normalized RMSE scores confirm robust property alignment under complex constraints. In contrast, models such as LIMO and AAE tend to overfit or focus narrowly on specific objectives, often sacrificing either validity or controllability.

Figure 4 shows that the generated molecules of HSPAG (orange) exhibit focused property distributions centered around the target values (red dashed line), in contrast to the broader distributions observed in the training dataset (gray), demonstrating our model's ability to effectively generate molecules under property constraints.

Outlier Region Targeting

To further assess the model's extrapolation capability in outlier regions of the property distribution, we evaluate HSPAG on conditional generation tasks with sparsely represented target property ranges. Following the setup from LIMO, we define three challenging property conditions. As shown in Tab. 2, HSPAG achieves a strong balance between constraint satisfaction and novelty, demonstrating robustness in extrapolative generation. While GCPN yields the highest success rate under the low logP constraint, its low diversity suggests mode collapse. See Appendix G for details.

Method	$-2.5 \leq \log P \leq -2$		$5 \leq \log P \leq 5.5$		$150 \leq \text{MW} \leq 200$	
	Success	**Diversity**	**Success**	**Diversity**	**Success**	**Diversity**
JT-VAE	11.3%	0.846	7.6%	0.907	1.7%	0.938
ORGAN	0	–	0.2%	0.909	15.1%	0.759
GCPN	**85.5%**	0.392	**54.7%**	0.855	76.1%	0.921
MOLDQN	9.66%	0.854	1.44%	0.734	2.40%	0.804
LIMO	10.4%	0.914	6.4%	0.866	13.7%	0.873
HSPAG	33.3%	**0.917**	**55.6%**	**0.911**	**79.5%**	0.925

Table 2: Performance of outlier region targeting. **Success (%)**: Ratio of molecules within the target range; **Diversity**: One minus average Tanimoto similarity.

Molecular Editing Capability

Molecular editing aims to optimize target properties while preserving core structures. We modify the target properties based on the original properties of the reference molecule, in order to preserve the structure of the reference molecule. To avoid conflicting inputs, we mask properties strongly correlated with the target property using a threshold μ, and apply random masking to non-target dimensions for diversity.

Model	KMO Inhibitor				NMT Inhibitor			
	1st	2nd	3rd	Ratio	1st	2nd	3rd	Ratio
CProMG	-8.82	-8.80	-8.36	0.053	-9.21	-8.66	-8.16	0.067
SPMM	**-10.33**	-9.75	-9.61	0.107	-10.20	-8.96	-8.86	0.106
Chemformer	-9.77	-9.69	-8.39	0.164	-9.98	-9.96	-8.86	0.227
HSPAG	-10.31	-10.24	-9.582	**0.372**	**-10.75**	-10.46	-10.33	**0.324**

Table 3: Top three candidates with lowest BFE after property-threshold and fragment-overlap filtering. **Ratio (%)**: fraction of generated molecules retaining the specified fragment.

To further validate the practical utility of our proposed molecular editing framework in real-world drug optimization scenarios, we conducted case studies on two compounds with clearly defined optimization objectives (Xiong et al. 2021). Detailed experimental settings are provided in Appendix I. We filter the generated molecules by requiring

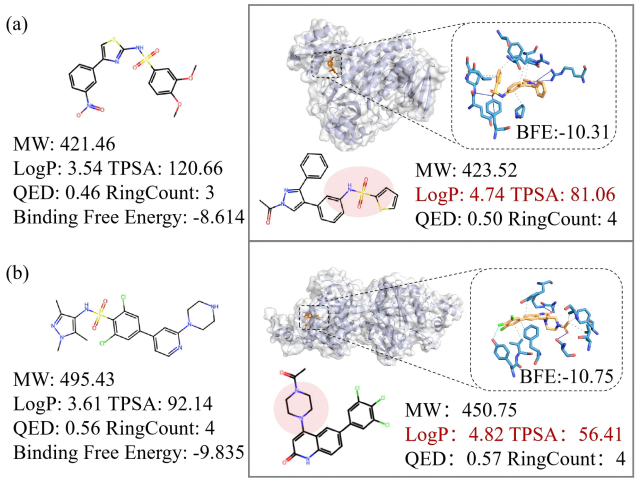

Figure 5: Optimization results of KMO and NMT inhibitors.

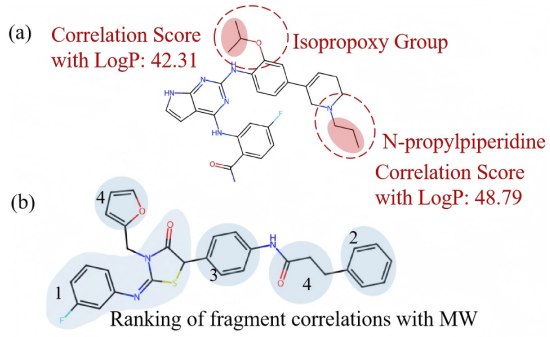

Figure 6: Analysis of substructure alignment.

property improvements beyond predefined thresholds and partial fragment overlap with the reference compounds, and report the top three candidates with the lowest binding free energy (BFE) in Tab. 3. As shown, our model demonstrates clear advantages in molecular editing, with the resulting optimized molecules visualized in Fig. 5.

Structure–Property Insight Modeling

To probe the model's capture of nuanced structure–property relationships, we analyzed substructure embeddings and their correlations with physicochemical properties. As shown in Figure 6(a), although both substructures contain three carbons, the isopropoxy group exhibits a lower correlation with LogP than the N-propylpiperidine, reflecting that their contributions to lipophilicity depend not only on current substructure but also on the local chemical environment, such as polarity contributions from ether oxygen. The correlation between MW and substructures follows the ranking of substructure molecular mass (Figure 6 (b)).

Building upon these insights from substructure analysis, we further examine molecular representations at the whole-molecule level. We fine-tuned the pre-trained encoder together with a projection head on molecular property prediction datasets from MoleculeNet. Results are reported in Table 4 and Appendix Section H. To avoid data leakage, downstream property labels and their correlated ones were masked during pretraining. HSPAG performs strongly on molecular property benchmarks, showing good transferability and enabling cross-property reasoning through its learned structure–property distributions.

	BBBP↑	Tox21↑	ToxCast↑	ClinTox↑	BACE↑
Number of Molecules	2039	7831	8575	1478	1513
Number of Tasks	1	12	617	2	1
Hu et al.	$70.8_{\pm 1.5}$	$78.7_{\pm 0.4}$	$65.7_{\pm 0.6}$	$72.6_{\pm 1.5}$	$84.5_{\pm 0.7}$
GEM	$88.8_{\pm 0.4}$	$78.1_{\pm 0.4}$	$68.6_{\pm 0.2}$	$90.3_{\pm 0.7}$	$87.9_{\pm 1.1}$
GROVER	$86.8_{\pm 2.2}$	$80.3_{\pm 2.0}$	$56.8_{\pm 3.4}$	$70.3_{\pm 13.7}$	$82.4_{\pm 3.6}$
Mole-Bert	$91.8_{\pm 1.4}$	$\mathbf{82.7}_{\pm 0.7}$	$69.5_{\pm 3.5}$	$74.7_{\pm 4.3}$	$87.4_{\pm 1.0}$
MolCLR	$73.3_{\pm 1.0}$	$74.1_{\pm 5.3}$	$65.9_{\pm 2.1}$	$82.9_{\pm 1.2}$	$82.6_{\pm 0.7}$
MolMVC	$92.6_{\pm 2.6}$	$77.4_{\pm 2.3}$	$68.5_{\pm 2.8}$	$68.7_{\pm 3.8}$	$85.0_{\pm 1.3}$
BartSmiles	$76.2_{\pm 4.0}$	$75.4_{\pm 0.9}$	$65.6_{\pm 3.8}$	$93.8_{\pm 5.6}$	$76.2_{\pm 1.1}$
SPMM	$88.5_{\pm 2.8}$	$80.1_{\pm 0.7}$	$68.3_{\pm 4.0}$	$88.1_{\pm 2.4}$	$87.1_{\pm 2.9}$
HSPAG-w/o pretrain	$84.8_{\pm 3.2}$	$78.5_{\pm 0.8}$	$63.6_{\pm 3.7}$	$90.2_{\pm 3.5}$	$80.2_{\pm 1.2}$
HSPAG	$\mathbf{94.8}_{\pm 1.8}$	$80.9_{\pm 0.6}$	$\mathbf{70.7}_{\pm 2.1}$	$\mathbf{97.3}_{\pm 2.7}$	$\mathbf{88.1}_{\pm 0.8}$

Table 4: Performance comparison on molecular property prediction tasks. We perform on three random-seeded scaffold splitting for all datasets, with a train/validation/test ratio of 8:1:1.

We further compared the pre-trained embeddings from different models and observed that HSPAG captures key substructures pertinent to molecular properties more effectively (Tab. 15 in Appendix H), which also explains why the pre-trained molecular embeddings from HSPAG achieve a strong balance between structural and property similarity. Leveraging this, we apply HSPAG to a **drug repurposing** task targeting the immuno-oncology protein HPK1, and successfully retrieve Abemaciclib, which has been experimentally validated as a potential HPK1 inhibitor and the top three candidate compounds shown in Fig. 7.

Hierarchical Alignment Boosts Generation

We conducted ablation studies to assess the impact of hierarchical alignment on controllable molecule generation. As summarized in Tab. 5 (left), the results show that removing alignment components degrades conditional generation performance, underscoring the importance of alignment prior to generation. Furthermore, incorporating hierarchical hierarchical alignment yields additional improvements, as reflected by lower RMSD scores. This improvement arises from the model's ability to capture relationships between molecular properties and substructures.

In addition, we observe that removing augmentation has a pronounced impact on generation: while the uniqueness of generated molecules increases substantially, RMSD also rises. This suggests that HSPAG tends to map distinct SMILES representations of the same molecule closer in the latent space, which benefits constraint enforcement during controllable generation but may reduce diversity under comparable computational budgets (e.g., same runtime). We further assess the impact of different modules on property prediction and molecular retrieval tasks, as shown in Fig. 8 and Appendix J, which reveal the contributions of each component to overall performance.

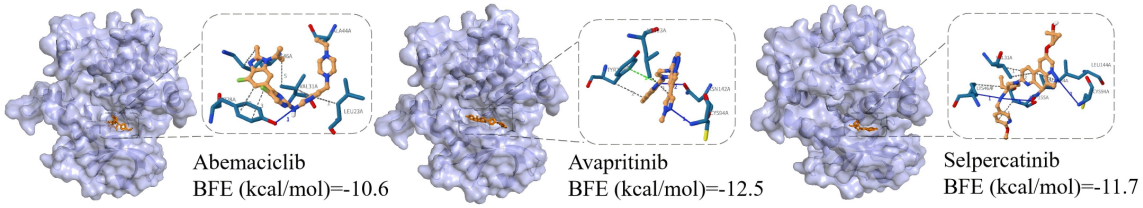

Figure 7: The protein-ligand structure (PDB ID: 7SIU46) was utilized as a reference for the identification of the binding pocket. The 3D structures of Abemaciclib, Selpercatinib, and Avapritinib were downloaded from PubChem. The interactions between molecules and HPK1 were profiled by PLIP.

	15.0% (Lev=3)				9.3% (Lev=5)				4.0% (Lev=10)			
	Property Pred.	Conditional Gen.			Property Pred.	Conditional Gen.			Property Pred.	Conditional Gen.		
Ablation Settings	Avg. AUC↑	Uniq.↑	Avail.↑	RMSE↓	Avg. AUC↑	Uniq.↑	Avail.↑	RMSE↓	Avg. AUC↑	Uniq.↑	Avail.↑	RMSE↓
HSPAG-NCA	82.98	**0.985**	**0.844**	0.419	82.90	**0.972**	**0.846**	0.432	81.96	**0.975**	**0.852**	0.467
HSPAG-NMC	83.38	0.877	0.771	0.285	83.62	0.837	0.741	0.295	83.36	0.811	0.741	0.303
HSPAG-NC	82.44	0.855	0.759	0.290	83.26	0.831	0.736	0.304	82.44	0.807	0.722	0.313
HSPAG-ND	85.12	-	-	-	83.32	-	-	-	82.77	-	-	-
HSPAG	**86.36**	0.864	0.783	**0.263**	**85.46**	0.855	0.764	**0.279**	**84.81**	0.820	0.753	**0.288**

Table 5: Ablation results on molecular property prediction and generation tasks across different data scales. Evaluated variants include HSPAG-NCA (without clip-based augmentation), HSPAG-NMC (without hierarchical clip alignment only molecular level), HSPAG-ND (without decoder) and HSPAG-NC (without any clip alignment).

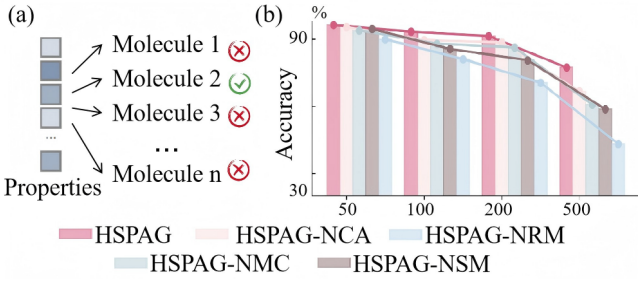

Figure 8: Ablation results on molecular retrieval task (GDB database). Two Additional Ablation Variants: HSPAG-NSM (without property similarity mask) and HSPAG-NRM (without property random mask).

Data Efficiency via Sampling and Augmentation

To construct example sets of varying sizes, we adjust the Levenshtein distance thresholds to 10, 5, and 3, yielding 63,539 (4.0%), 146,244 (9.3%), and 236,920 (15.0%) SMILES, respectively, from the full training dataset of 1,574,222 molecules. As shown in Table 5, smaller data scales exhibit a pronounced drop in both property prediction and conditional generation metrics when data augmentation is removed, underscoring that augmentation is particularly critical under limited data conditions. Even when the dataset size is reduced to one-third of its original scale, our model still outperforms baseline methods, demonstrating the data efficiency of our sampling–augmentation framework. These results highlight the robustness of HSPAG in maintaining controllable generation performance even under data scarcity.

Method	Val.↑	Nov.↑	Avail.↑	RMSE↓
SPMM	0.6579	0.9971	0.2836	0.4271
HSPAG-RO	0.3644	0.9998	0.3232	0.5207
HSPAG-EO	0.6503	**0.9999**	0.6157	0.4662
HSPAG	**0.6881**	**0.9999**	**0.6845**	**0.3906**

Table 6: Performance of molecular generation on structurally complex molecules.

We construct a dataset by selecting 100 molecules with high LM loss to evaluate generation on structurally complex molecules. We also introduce HSPAG-RO (trained on random set only) and HSPAG-EO (trained on example set only) for comparison. Results in Figure 6 show that incorporating the challenging set yields clear performance gains, confirming the benefit of progressively integrating difficult samples.

Conclusion And Future Work

We present HSPAG, a data-efficient framework for property-constrained molecular generation through hierarchical structure–property alignment. By combining multi-level features with representative sample selection, HSPAG achieves strong representations and controllable generation, validated by extensive experiments and real-world cases. In future work, we plan to incorporate 3D structural information to further enrich the molecular representation (Guan et al. 2023; Morehead and Cheng 2024). Moreover, as the molecular property data used in this study are primarily obtained from computational tools, we aim to explore more comprehensive and experimentally validated datasets to enhance the realism and reliability of property conditioning.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. U23A20321 and 62272490); the Natural Science Foundation of Hunan Province of China (Grant No. 2025JJ20062); and A*STAR's Decentralised Gap Funding (I23D1AG081) and AIDD Catalyst Grant (H25A1N0003).

References

Cao, H.; Bao, C.; Liu, C.; Chen, H.; Yin, K.; Liu, H.; Liu, Y.; Jiang, D.; and Sun, X. 2023. Attention where it matters: Rethinking visual document understanding with selective region concentration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19517–19527.

Chang, J.; and Ye, J. C. 2024. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15(1): 2323.

Chen, J.; Zhang, X.; Ma, Z.-M.; Liu, S.; et al. 2023. Molecule joint auto-encoding: Trajectory pretraining with 2d and 3d diffusion. *Advances in Neural Information Processing Systems*, 36: 55077–55096.

Dobberstein, N.; Maass, A.; and Hamaekers, J. 2024. Llamol: a dynamic multi-conditional generative transformer for de novo molecular design. *Journal of Cheminformatics*, 16(1): 73.

Dowden, H.; and Munro, J. 2019. Trends in clinical success rates and therapeutic focus. *Nat Rev Drug Discov*, 18(7): 495–496.

Eckmann, P.; Sun, K.; Zhao, B.; Feng, M.; Gilson, M. K.; and Yu, R. 2022. Limo: Latent inceptionism for targeted molecule generation. *Proceedings of machine learning research*, 162: 5777.

Fang, Y.; Zhang, N.; Chen, Z.; Guo, L.; Fan, X.; and Chen, H. 2024. Domain-Agnostic Molecular Generation with Chemical Feedback. In *The Twelfth International Conference on Learning Representations*.

Fu, L.; Shi, S.; Yi, J.; Wang, N.; He, Y.; Wu, Z.; Peng, J.; Deng, Y.; Wang, W.; Wu, C.; et al. 2024. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic acids research*, 52(W1): W422–W431.

Gottipati, S. K.; Sattarov, B.; Niu, S.; Pathak, Y.; Wei, H.; Liu, S.; Blackburn, S.; Thomas, K.; Coley, C.; Tang, J.; et al. 2020. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International conference on machine learning*, 3668–3679. PMLR.

Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; and Ma, J. 2023. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In *The Eleventh International Conference on Learning Representations*.

Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; and Tang, J. 2022. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 594–604.

Huang, Z.; Fan, Z.; Shen, S.; Wu, M.; and Deng, L. 2024. MolMVC: Enhancing molecular representations for drug-related tasks through multi-view contrastive learning. *Bioinformatics*, 40(Supplement_2): ii190–ii197.

Irwin, R.; Dimitriadis, S.; He, J.; and Bjerrum, E. J. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, 2323–2332. PMLR.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2020. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, 4839–4848. PMLR.

Jones, J.; Clark, R. D.; Lawless, M. S.; Miller, D. W.; and Waldman, M. 2024. The AI-driven Drug Design (AIDD) platform: an interactive multi-parameter optimization system integrating molecular evolution with physiologically based pharmacokinetic simulations. *Journal of Computer-Aided Molecular Design*, 38(1): 14.

Kwon, K.; Jeong, K.; Park, J.; Na, H.; and Shin, J. 2023. String-Based Molecule Generation Via Multi-Decoder VAE. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Li, J.-N.; Yang, G.; Zhao, P.-C.; Wei, X.-X.; and Shi, J.-Y. 2023. CProMG: controllable protein-oriented molecule generation with desired binding affinity and drug-like properties. *Bioinformatics*, 39(Supplement_1): i326–i336.

Liu, C.; Yin, K.; Cao, H.; Jiang, X.; Li, X.; Liu, Y.; Jiang, D.; Sun, X.; and Xu, L. 2024. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15534–15545.

Liu, Q.; Allamanis, M.; Brockschmidt, M.; and Gaunt, A. 2018. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31.

Luo, Y.; Liu, Y.; and Peng, J. 2023. Calibrated geometric deep learning improves kinase–drug binding predictions. *Nature machine intelligence*, 5(12): 1390–1401.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Morehead, A.; and Cheng, J. 2024. Geometry-complete diffusion for 3D molecule generation and optimization. *Communications Chemistry*, 7(1): 150.

Pammolli, F.; Magazzini, L.; and Riccaboni, M. 2011. The productivity crisis in pharmaceutical R&D. *Nature reviews Drug discovery*, 10(6): 428–438.

Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; and Wei, L. 2023. Deep generative models in de novo drug molecule generation. *Journal of Chemical Information and Modeling*, 64(7): 2174–2194.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sadybekov, A. V.; and Katritch, V. 2023. Computational approaches streamlining drug discovery. *Nature*, 616(7958): 673–685.

Settles, B. 2009. Active learning literature survey.

Shen, A.; Yuan, M.; Ma, Y.; Du, J.; and Wang, M. 2024. Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction. *Briefings in Bioinformatics*, 25(4): bbae256.

Skinnider, M. A. 2024. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nature Machine Intelligence*, 6(4): 437–448.

Subedi, R.; Wei, L.; Gao, W.; Chakraborty, S.; and Liu, Y. 2024. Empowering Active Learning for 3D Molecular Graphs with Geometric Graph Isomorphism. *Advances in Neural Information Processing Systems*, 37: 55507–55537.

Wang, S.; Guo, X.; Lin, X.; Pan, B.; Du, Y.; Wang, Y.; Ye, Y.; Petersen, A.; Leitgeb, A.; AlKhalifa, S.; et al. 2022. Multi-objective deep data generation with correlated property control. *Advances in neural information processing systems*, 35: 28889–28901.

Xiong, B.; Wang, Y.; Chen, Y.; Xing, S.; Liao, Q.; Chen, Y.; Li, Q.; Li, W.; and Sun, H. 2021. Strategies for structural modification of small molecules to improve blood–brain barrier penetration: a recent perspective. *Journal of Medicinal Chemistry*, 64(18): 13152–13173.

Yang, Y.; Wu, Z.; Yao, X.; Kang, Y.; Hou, T.; Hsieh, C.-Y.; and Liu, H. 2022. Exploring low-toxicity chemical space with deep learning for molecular generation. *Journal of Chemical Information and Modeling*, 62(13): 3191–3199.

Zhang, K.; Yang, X.; Wang, Y.; Yu, Y.; Huang, N.; Li, G.; Li, X.; Wu, J. C.; and Yang, S. 2025. Artificial intelligence in drug development. *Nature medicine*, 31(1): 45–59.

Zhu, H.; Zhou, R.; Cao, D.; Tang, J.; and Li, M. 2023. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications*, 14(1): 6234.