Leveraging Domain Motif Assembler for Multi-objective, Multi-domain and Explainable Molecular Design

Zijun Chen[1*], Yu Wang[1,2*], Liuzhenghao Lv[1,2], Hao Li[1,2,3], Zongying Lin[1], Li Yuan[1,3†], and Yonghong Tian[1,2,3†]

[1] *School of Electronic and Computer Engineering, Peking University, China*
[2] *School of Computer Science, Peking University, China*
[3] *Peng Cheng Laboratory, China*
[*] These authors contributed equally to this work
[†] Corresponding authors: yuanli-ece@pku.edu.cn, yhtian@pku.edu.cn

Contents

Supplementary Section 1: Dataset Specifications and Characteristics

Our experiments employ four distinct molecular datasets, each with unique characteristics and scope, as summarized in Supplementary Table 1. These datasets encompass diverse domains and structural complexities:

Supplementary Table 1. Comprehensive statistical characteristics of the HCE, SNB-60K, GDB-13, and DTP datasets employed in this study.

Datasets	Number of graphs	Number of nodes	Number of node types
HCE	24,953	$13 \leq \mid \mathcal{V} \mid \leq 35$	6
SNB-60K	60,828	$13 \leq \mid \mathcal{V} \mid \leq 45$	8
GDB-13	398,453	$5 \leq \mid \mathcal{V} \mid \leq 13$	4
DTP	105,338	$3 \leq \mid \mathcal{V} \mid \leq 38$	20

Supplementary Section 2: Hyperparameters

Supplementary Section 2.1: Coarse-grained Score-based Generation Hyperparameters

The hyperparameters used in the coarse-grained score-based generation model across the four datasets are summarized in Supplementary Table 2, including the hyperparameters of the data definitions determined by the frequency-based motif mining process, the neural networks ε_θ and ε_ϕ used for score matching, the diffusion processes (i.e., the SDEs for $\boldsymbol{F}$ and $\boldsymbol{C}$), the SDE solver, and the training procedure.

Supplementary Table 2. Hyperparameters for coarse-grained score-based generation.

	Param	HCE	GDB-13	SNB-60K	DTP
Data	Maximum number of Motifs	11	8	10	17
	Motif feature dimension	100	200	100	200
ε_θ	Number of GCN layers	2	2	2	2
	Hidden dimension	16	16	16	16
ε_ϕ	Number of attention heads	4	4	4	4
	Number of initial channels	2	2	2	2
	Number of hidden channels	8	8	8	8
	Number of final channels	4	4	4	4
	Number of GCN layers	2	6	2	6
	Hidden dimension	16	16	16	16
SDE for $\boldsymbol{F}$	Type	VP	VP	VP	VP
	Number of sampling steps	1000	1000	1000	1000
	β_{min}	0.1	0.1	0.1	0.1
	β_{max}	1.0	1.0	1.0	1.0
SDE for $\boldsymbol{C}$	Type	VE	VE	VE	VE
	Number of sampling steps	1000	1000	1000	1000
	β_{min}	0.1	0.2	0.1	0.2
	β_{max}	1.0	1.0	1.0	1.0
Solver	Predictor	Reverse	Reverse	Reverse	Reverse
	Corrector	Langevin	Langevin	Langevin	Langevin
	SNR	0.2	0.2	0.2	0.2
	Scale coefficient	0.5	0.9	0.5	0.9
Train	Optimizer	Adam	Adam	Adam	Adam
	Learning rate	5e-3	5e-3	5e-3	5e-3
	Weight decay	1e-4	1e-4	1e-4	1e-4
	Batch size	2048	8192	1024	2048
	Number of epochs	300	500	300	500
	EMA	0.999	0.999	0.999	0.999
	Learning rate decay	0.999	0.999	0.999	0.999

Supplementary Section 2.2: Fine-grained Bond Scoring Model Hyperparameters

Supplementary Table 3 provides the model and training hyperparameters for the fine-grained bond scoring phase. In this model, each node feature is represented by a combination of atom and motif embeddings.

Supplementary Table 3. Hyperparameters for the fine-grained bond scoring model.

	Param	HCE	GDB-13	SNB-60K	DTP
Model	Dimension of atom embeddings	50	50	50	50
	Dimension of motif embeddings	100	100	100	100
	Dimension of node representations	300	300	300	300
	Dimension of graph embeddings	400	400	400	400
	Number of iterations of GINE	4	4	4	4
Train	Optimizer	Adam	Adam	Adam	Adam
	Learning rate	1e-3	1e-3	1e-3	1e-3
	Number of epochs	10	10	10	10
	Batch size	32	32	32	32

Supplementary Section 3: Baseline Settings

To comprehensively assess the performance of our model across the four aforementioned datasets, we establish comparisons with eight state-of-the-art baseline models: SMILES-VAE[1], SMILES-LSTM-HC[2], MoFlow[3], REINVENT[4], GB-GA[5], GDSS[6], MiCaM[7], and MolT5[8]. For each model, we conduct five independent experiments and report the mean and standard deviation of their optimal objective values. The implementation details are as follows: for SMILES-VAE, SMILES-LSTM-HC, MoFlow, REINVENT, and GB-GA, we adopt the configurations outlined in TARTARUS[9]. For GDSS, MiCaM, and MolT5, we implement the models based on the settings proposed in their respective original works.

Supplementary Section 4: Details of Benchmarks

This section provides comprehensive details of the benchmark design across four distinct domains: organic photovoltaics, chemical reaction substrates, organic emissive materials, and protein ligands.

Supplementary Section 4.1: Design of Organic Photovoltaics

For organic photovoltaic (OPV) materials, we formulate two design objectives by combining the power conversion efficiencies (PCEs) with synthetic accessibility (SA) scores[10]. The objectives are defined as:

- Maximizing $PCE_{PCBM} - SAscore$;
- Maximizing $PCE_{PCDTBT} - SAscore$.

The simulation workflow for calculating PCEs begins by accepting a molecular input in the form of a SMILES string[11]. Initial Cartesian coordinates are generated using Open Babel[12], which are then subjected to a conformer search conducted by CREST[13]. After conformer selection, geometry optimization is carried out utilizing the XTB method[14]. Following optimization, a single-point energy calculation at the GFN2-xTB level[14] is performed, which yields key electronic properties such as the energies of the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), the HOMO-LUMO energy gap, and the molecular dipole moment. These calculated properties are subsequently employed in the Scharber model[15] to estimate the PCE of the organic photovoltaic device. This workflow integrates both quantum chemical calculations and theoretical performance prediction models to streamline the evaluation of OPV candidates.

Supplementary Section 4.2: Design of Chemical Reaction Substrates

To evaluate chemical reactivity in this benchmark, we identify activation energy and reaction energy as the key thermodynamic parameters governing reaction kinetics and energetics. These fundamental descriptors directly influence reaction feasibility and efficiency. The benchmark objectives for chemical reactivity are formulated as follows:

- Minimizing the activation energy ΔE_a ;
- Minimizing the reaction energy ΔE_r ;
- Minimizing the sum $\Delta E_a + \Delta E_r$;

- Minimizing the difference $-\Delta E_a + \Delta E_r$.

We perform this benchmark on the SNB-60K dataset, where each molecule contains a syn-sesquinorbornene structural unit. The simulation workflow begins with verification of hard constraints[9,16] in the SMILES string of the proposed substrate, ensuring that all generated molecules preserve the syn-sesquinorbornene core and remain stable. For the two objectives involving combinations of target properties, an additional constraint is imposed: the SAscore must not exceed six. Upon satisfying these constraints, initial Cartesian coordinates for the reactant and product are generated using the CREST conformer search[13], initiated through Open Babel[12]. Following this, the SEAM optimization process is conducted, where an initial geometry for the transition state is obtained by interpolating between the optimized geometries of the reactant and product (via polanyi optimization). The SEAM optimization then refines the guessed transition state structure, leading to the identification of the transition state, which is further optimized through constrained conformational sampling using CREST and finalized with polanyi optimization. In the final stage, the energy of the reactant (E_R), transition state (E_{TS}), and product (E_P) are calculated. From these, the reaction energy ($\Delta E_r = E_P - E_R$) and the approximate SEAM activation energy ($\Delta E_a = E_{TS} - E_R$) are extracted, providing critical insights into the energetic profile of the reaction.

Supplementary Section 4.3: Design of Organic Emitters

For organic light-emitting materials, we establish three design objectives that target key photophysical properties essential for high-performance organic emitters. These objectives encompass critical parameters that determine emission efficiency, color purity, and device stability. The specific design criteria are formulated as follows:

- Minimizing the singlet-triplet gap (ST) : $\Delta E(S_1\text{-}T_1)$;
- Maximizing the oscillator strength (OSC) for the transition between S_1 and S_0 ;
- Maximizing the combined objective : $+OSC - ST - |\Delta E(S_0\text{-}S_1) - 3.2\ eV|$.

In this benchmark, high-fitness molecules must exhibit a SAscore of 4.5 or lower. The workflow initiates with the generation of a molecule, where the initial geometry is obtained through Open Babel[12] and RDKit[17]. This geometry undergoes a conformer search using CREST[13] to explore possible low-energy conformations. Following the conformer search, the geometry is further optimized using the XTB method[14,18] to refine the molecular structure. Subsequently, excited state properties are calculated via TD-DFT single-point calculations[19], performed using the PySCF package[20]. This stage yields key electronic properties, including the singlet-triplet energy gap ($\Delta E(S_1\text{-}T_1)$), oscillator strength, and vertical excitation energy ($\Delta E(S_0\text{-}S_1)$). These properties are critical for evaluating the photophysical behavior of the organic emitter[9].

Supplementary Section 4.4: Design of Protein Ligands

This benchmark focuses on two primary objectives in protein-ligand design:

- Minimizing docking scores (DS) ;
- Maximizing the success rate (SR) for sampled molecules passing structure filters.

Notably, the objectives of this benchmark are not solely determined by docking scores but also incorporate stringent structural constraints[9]. If these constraints are not met, an exceedingly unfavorable score of 10,000 is assigned in place of the actual docking score. The list of these filters includes:

- Absence of reactive groups ;
- Absence of formal charges ;
- Absence of radicals ;
- At most 2 bridgehead atoms ;
- No rings larger than 8-membered ;
- Fulfils Lipinski's Rule of Five ;
- $SAscore < 4.5$;
- $qed > 0.3$;

- $TPSA > 140$;
- Molecule passes the PAINS, WEHI and MCF filters ;
- Molecule does not contain Si and Sn atoms.

This benchmark is conducted using the DTP dataset. The simulation workflow is initiated with the SMILES string of the proposed molecule, followed by the creation of an initial Cartesian coordinate guess using Open Babel[12], which serves as the starting structure for the docking procedure. The structure is then subjected to molecular docking using QuickVina2[21] and Smina[22], both of which are molecular docking software tools. QuickVina2 focuses on speed and sampling efficiency, whereas Smina prioritizes accurate scoring and pose refinement. The final output is the docking score, which quantifies the molecule's binding energy to the target site, providing critical insights into the molecule's potential efficacy as a ligand for the specified protein target.

Supplementary Section 5: Pseudo Code of Motif Vocabulary Construction

Our goal is to construct a frequency-based molecular motif vocabulary for a given dataset, drawing inspiration from prior works[23,24]. The pseudo code outlining this process is provided in Algorithm 1. Merge$(\cdot)$ represents combining neighboring motifs to generate new substructures and then converting them into corresponding SMILES representations. Update$(\cdot)$ combines the motifs corresponding to a given SMILES string representing a potential new motif formed by merging two existing motifs. It modifies the internal representation of the molecule by uniting the atom indices of the merged motifs, updating their SMILES, and removing the individual motifs that have been merged.

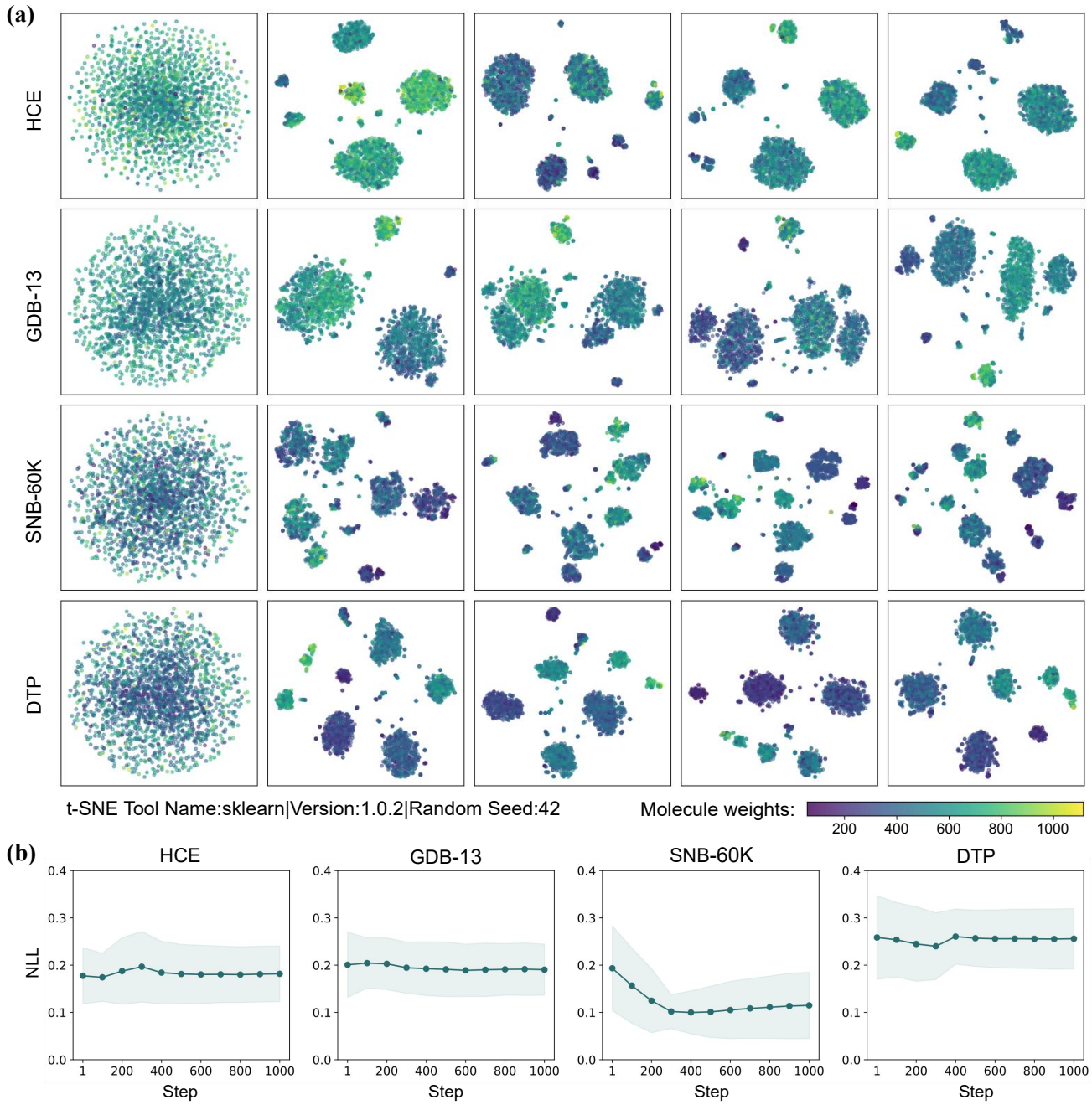
Algorithm 1 Motif Vocabulary Construction

Input: A molecule set $\mathcal{D} = \{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)\}_{i=1}^{m}$, desired vocabulary size K
Output: A molecular motif vocabulary $\mathbb{V}$

$\mathbb{V} \leftarrow \{v\}$; ▷ Initialize a vocabulary to all atoms v in $\mathcal{D}$
for $k = 1$ to $K - |\mathbb{V}|$ **do**
  $\mathcal{C} \leftarrow \{\}$; ▷ Initialize an empty frequency counter
  for $\mathcal{G}_d$ in $\mathcal{D}$ **do**
    for $(\mathcal{F}_i, \mathcal{F}_j, \mathcal{E}_{ij})$ in $\mathcal{G}_d$ **do**
      $\mathcal{F} \leftarrow \text{Merge}(\mathcal{F}_i, \mathcal{F}_j, \mathcal{E}_{ij})$; ▷ Merge neighboring motifs into a novel motif
      $\mathcal{C}[\mathcal{F}] \leftarrow \mathcal{C}[\mathcal{F}] + 1$; ▷ Update the frequency of the motif (initial value is 0)
    end for
  end for
  $\mathcal{F}_t \leftarrow \arg\max_{(\mathcal{F}, f) \in \mathcal{C}} f$; ▷ Identify the most frequent merged motif
  $\mathbb{V} \leftarrow \mathbb{V} \cup \{\mathcal{F}_t\}$
  $\mathcal{D}' \leftarrow \{\}$
  for $\mathcal{G}_d$ in $\mathcal{D}$ **do**
    $\mathcal{G}'_d \leftarrow \text{Update}(\mathcal{G}_d, \mathcal{F}_t)$; ▷ Update the molecules by merging all $\mathcal{F}_t$ motifs
    $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{\mathcal{G}'_d\}$
  end for
  $\mathcal{D} \leftarrow \mathcal{D}'$
end for
return $\mathbb{V}$

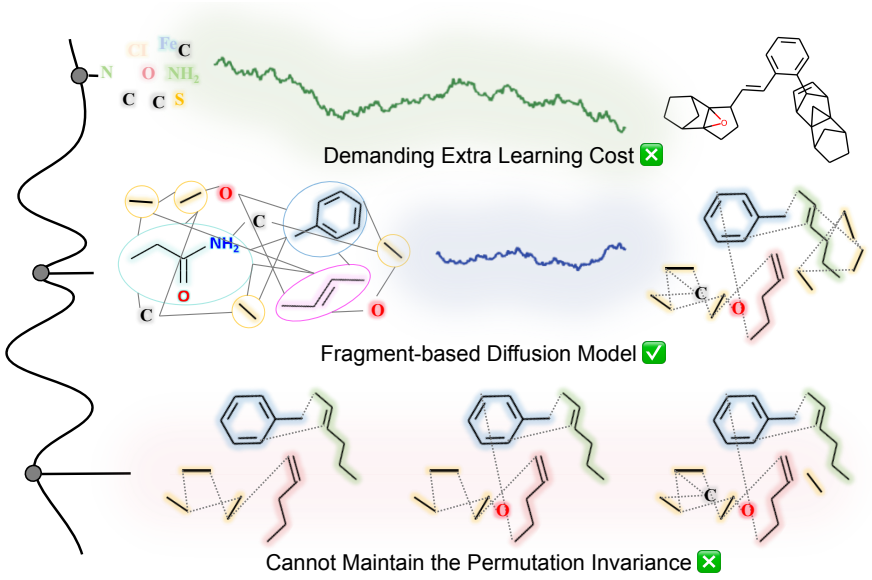
Supplementary Section 6: Organized Latent Spaces and Reliable Bond Formation in Score-based Diffusion Sampling

In this section, we analyze the progression and underlying mechanisms of molecular generation within our score-based diffusion sampling process. Supplementary Figure 1(a) visualizes the latent space embeddings of hidden layers during the diffusion sampling steps from 1 to 1000 via t-SNE for the four datasets. The t-SNE projection reveals that the generated molecules are clearly grouped based on their molecular weights. As the diffusion sampling progresses, molecules begin to form stable clusters, indicating that DM-Assembler successfully organizes molecules within the latent space. This organization reflects the model's ability to distinguish molecules with different characteristics and ensures that molecular weight variations are captured effectively during the generation process.



Supplementary Figure 1. (**a**) 2D visualization of hidden layer embeddings from diffusion sampling (steps 1 to 1000) using t-SNE, colored by molecule weights. (**b**) NLL average of atom-level bonds in diffusion sampling (steps 1 to 1000).

Supplementary Figure 1(b) depicts the average negative log-likelihood (NLL) of atom-level bond formation during the fine-grained bond scoring stage across diffusion sampling steps, which measures the confidence of forming chemically valid bonds during generation. The NLL values remain consistently low and stable across datasets, indicating the reliability of our scoring mechanism in enforcing probabilistic bond validity throughout the generation process. Notably, even at intermediate steps, the NLL remains low, implying that DM-Assembler can generate chemically plausible molecules at various stages of the sampling process. A specific observation can be made regarding the SNB-60K dataset, where the NLL reaches its lowest value around step 200. This suggests that the intermediate stages of the diffusion process align most closely with the characteristics of the training data, leading to higher confidence in bond formation. Similar trends are observed across other datasets, where NLL stabilization further confirms the robustness of the generative process. These results highlight two key strengths of DM-Assembler: (1) its ability to organize the latent molecular space into interpretable clusters aligned with



Supplementary Figure 2. Conceptual framework and advantages of our motif-based diffusion approach.

molecular properties like molecular weight, and (2) its stable and reliable probabilistic scoring mechanism, which ensures chemically valid bond formation throughout the generative process. These features contribute to the model's effectiveness in generating molecules that exhibit both structural diversity and chemical validity with high confidence, even under multi-step generative constraints.

Supplementary Section 7: Motivation

Supplementary Figure 2 illustrates the fundamental rationale behind our approach. Our motif-based diffusion framework effectively addresses two major limitations prevalent in existing methods: the computational overhead associated with atom-level diffusion models and the inability of structure-by-structure models to preserve permutation invariance. By operating at the motif level, our framework achieves both computational efficiency and structural consistency.

Supplementary Section 8: Coarse-grained Score-based Generative Modeling

In this section, we present a comprehensive description of our proposed coarse-grained score-based generative model.

Supplementary Section 8.1: VE and VP SDEs

We provide two types of stochastic differential equations (SDEs), i.e., the Variance Exploding (VE) SDE and Variance Preserving (VP) SDE, whose discretizations cause noise perturbations of our coarse-grained score-based generative model[6].

The formulation of the VE SDE is given by:

$$\mathrm{d}\mathbf{x} = \sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^{t}\sqrt{2\log\frac{\sigma_{max}}{\sigma_{min}}}\mathrm{d}\mathbf{w},\ t\in(0,1], \tag{1}$$

where σ_{min} and σ_{max} are predefined hyperparameters (detailed in Supplementary Section 2.1). The corresponding perturbation kernel is formulated as follows:

$$p_{0t}(\mathbf{x}(t)\mid\mathbf{x}(0)) = \mathcal{N}\left(\mathbf{x}(t)\mid\mathbf{x}(0),\sigma_{min}^{2}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^{2t}\mathbf{I}\right),\ t\in(0,1]. \tag{2}$$

The process of VP SDE is as follows:

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta_t\mathbf{x}\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\mathbf{w},\ t\in(0,1], \tag{3}$$

where $\beta_t = \beta_{min} + t(\beta_{max}-\beta_{min})$ with both β_{max} and β_{min} serve as hyperparameters (detailed in Supplementary Section 2.1). Accordingly, the perturbation kernel is expressed as:

$$p_{0t}(\mathbf{x}(t)\mid\mathbf{x}(0)) = \mathcal{N}\left(\mathbf{x}(t)\mid e^{-\frac{1}{4}t^2(\beta_{max}-\beta_{min})-\frac{1}{2}t\beta_{min}}\mathbf{x}(0),\mathbf{I}-\mathbf{I}e^{-\frac{1}{2}t^2(\beta_{max}-\beta_{min})-t\beta_{min}}\right),\ t\in(0,1]. \tag{4}$$

Supplementary Section 8.2: Reverse Diffusion-driven Domain Motif Generation and Quantization

To generate domain motifs through the reverse diffusion process, we first sample N, representing the maximum number of motifs in a molecule, according to the empirical distribution of motif counts observed in the training dataset. Subsequently, we sample noise with a batch size B from the prior distribution. In this context, $\boldsymbol{F}_T \in \mathbb{R}^{N\times K\times B}$ corresponds to motif features, while $\boldsymbol{C}_T \in \mathbb{R}^{N\times N\times B}$ captures inter-motif connections, where F is the molecular motif vocabulary size. The reverse-time SDE process is then simulated to produce the final motif features, $\boldsymbol{F}_0$, and their corresponding connections, $\boldsymbol{C}_0$. These outputs are then quantized to yield discrete motifs and their associated connections. Specifically, we determine the index of the maximum value along the second dimension of $\boldsymbol{F}_0$ as the corresponding motif. Furthermore, the entries of $\boldsymbol{C}_0$ are quantized to $\{0,1\}$ with values in $(0,0.5)$ set to 0, and those in $[0.5,1)$ set to 1, indicating the absence or existence of a connection, respectively. The hyperparameters related to this process are detailed in Supplementary Section 2.

Supplementary Section 9: Fine-grained Molecular Structure Design via Bond Scoring

Algorithm 2 provides the pseudo code for our fine-grained motif assembly strategy based on bond scoring. Initially, we employ RDKit[17] to add the atoms and bonds of each molecular motif to the molecule being constructed and record the edges within the motif. Subsequently, for pairs of nodes residing in distinct motifs where a connection between the motifs is known, we include the corresponding edges to the candidate inter-motif bond set and calculate their associated scores. The candidate edges are then sorted in descending order according to their scores. During iteration, if an edge has a score greater than the threshold and passes the chemical-rule check, it is added to the molecule. The chemical-rule check ensures adherence to valence rules and prevents the formation of unstable cycles consisting of fewer than five or more than six nodes. Given the possibility of generating disconnected graphs during this procedure, we select the largest connected component as the final unconditionally generated molecular structure.

Algorithm 2 Fine-grained Molecular Structure Design via Bond Scoring

Input: A molecular motif set $\{\mathcal{F}_i = \{\tilde{\mathcal{V}}_i, \tilde{\mathcal{E}}_i\}\}_{i=1}^{m}$, a motif-level adjacency matrix $\boldsymbol{C} \in \mathbb{R}^{m\times m}$, the bond scoring model $\boldsymbol{p}$, a mapping from each atom to its motif ω, and a score threshold Ψ_{th}
Output: A complete, valid molecular graph $\mathcal{G}$

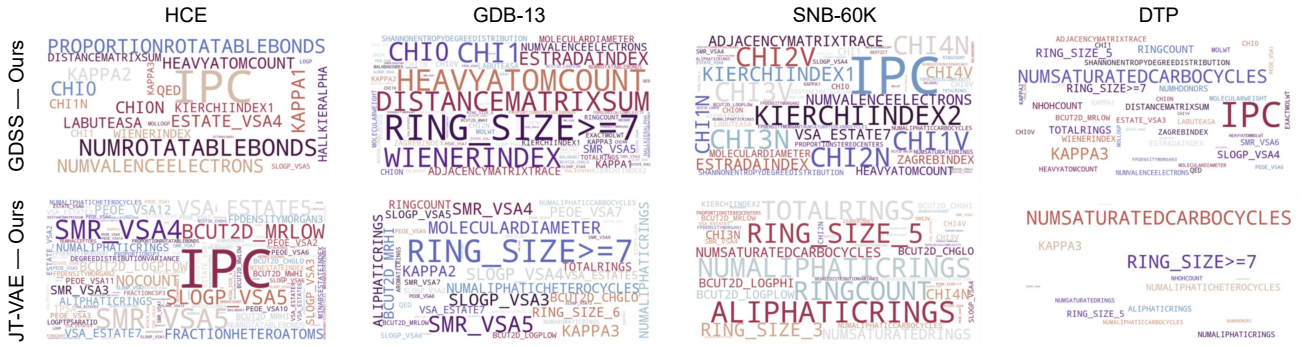
$\mathcal{G} \leftarrow$ Chem.RWMol() ▷ *Create an empty molecular graph*
$\mathcal{E}^{intra} \leftarrow \{\}$ ▷ *Create the intra-motif bond set*
$\mathcal{B} \leftarrow \{\}$ ▷ *Create the candidate inter-motif bond set*
for $\mathcal{F}_i$ in $\{\mathcal{F}_i\}_{i=1}^{m}$ **do**
    $\mathcal{G} \leftarrow$ AddAtom$(\tilde{\mathcal{V}}_i)$ ▷ *Add intra-motif atoms to* $\mathcal{G}$
    $\mathcal{G} \leftarrow$ AddBond$(\tilde{\mathcal{E}}_i)$ ▷ *Add intra-motif bonds to* $\mathcal{G}$
    $\mathcal{E}^{intra} \leftarrow \mathcal{E}^{intra} \cup \tilde{\mathcal{E}}_i$ ▷ *Update the intra-motif bond set*
end for
for (u,v) where $u \in \mathcal{F}_i$ and $v \in \mathcal{F}_j$ **do**
    if $(u,v) \notin \mathcal{E}^{intra}$ and $\boldsymbol{C}(\omega(u),\omega(v)) = 1$ **then**
        Compute the bond score $\mathcal{J}(u,v)$ via $\boldsymbol{p}$
        $\mathcal{B} \leftarrow \mathcal{B} \cup \{(u,v,\mathcal{J}(u,v))\}$ ▷ *Add bond and score to the inter-motif bond set*
    end if
end for
$\mathcal{B} \leftarrow$ SortByScore$(\mathcal{B})$ ▷ *Sort the candidate bonds based on their scores*
for b in $\mathcal{B}$ **do**
    if $\mathcal{J}(b) > \Psi_{th}$ and ChemicalRuleCheck(b) **then**
        $\mathcal{G} \leftarrow$ AddBond(b) ▷ Incorporate a checked inter-motif bond into $\mathcal{G}$
    end if
end for
$\mathcal{G} \leftarrow$ LargestConnectedComponent$(\mathcal{G})$ ▷ *Determine the largest component as final molecule*
return $\mathcal{G}$

Supplementary Section 10: Ablation Studies

To verify the influence of different modules of our model on the quality of the generated molecules, we make variants of the original design. In the first variant, motif-level connections $\boldsymbol{C}$ are removed, limiting the fine-grained atom-level bond prediction to only consider nodes that do not belong to the same motif. The second variant excludes the bond scoring model, which

Supplementary Table 4. Comparative results of DM-Assembler and its variants.

	GDB-13			SNB-60K		
	Uniqueness(%)	FCD	Novelty(%)	Uniqueness(%)	FCD	Novelty(%)
Ours	**99.36**	**8.90**	**99.14**	**84.72**	**5.19**	**99.43**
Ours w/o ***C***	96.77	9.20	98.77	77.21	6.65	99.40
Ours w/o bond scoring	1.92	22.32	89.58	0.96	14.72	82.25



Supplementary Figure 3. Word cloud visualization comparing property-specific KL divergence metrics across four datasets. The visualization highlights the molecular properties where GDSS and JT-VAE exhibit performance advantages relative to DM-Assembler.

means randomly connecting the generated motifs at the atomic level. We use **Uniqueness**, **FCD**, and **Novelty** as the evaluation metrics. High uniqueness indicates greater diversity in the generated molecules, while high novelty reflects the generation of new, previously unseen molecules. A lower FCD score signifies that the distribution of the generated molecules more closely aligns with that of the training set in chemical space. Supplementary Table 4 presents a comparison of our model's performance against these two variants, with the results representing the mean values derived from 3 independent runs.

We can conclude from the Supplementary Table 4 that removing the motif-level connections significantly reduces the diversity of generated molecules, as indicated by the drop in uniqueness, and slightly worsens the alignment with the training set distribution, as reflected by the increase in FCD. However, the most substantial impact is observed when the fine-grained bond scoring model is removed. Without it, the model's ability to generate diverse, novel, and chemically realistic molecules is severely compromised, leading to a drastic drop in uniqueness and a significant increase in FCD. These results clearly demonstrate that both components are essential for maintaining the high quality of molecule generation, with the fine-grained bond scoring being particularly critical. Therefore, the design of our model, which integrates these two components, is necessary to ensure optimal performance across different datasets.

Supplementary Section 11: Word Cloud Visualization of KL Divergence for Properties

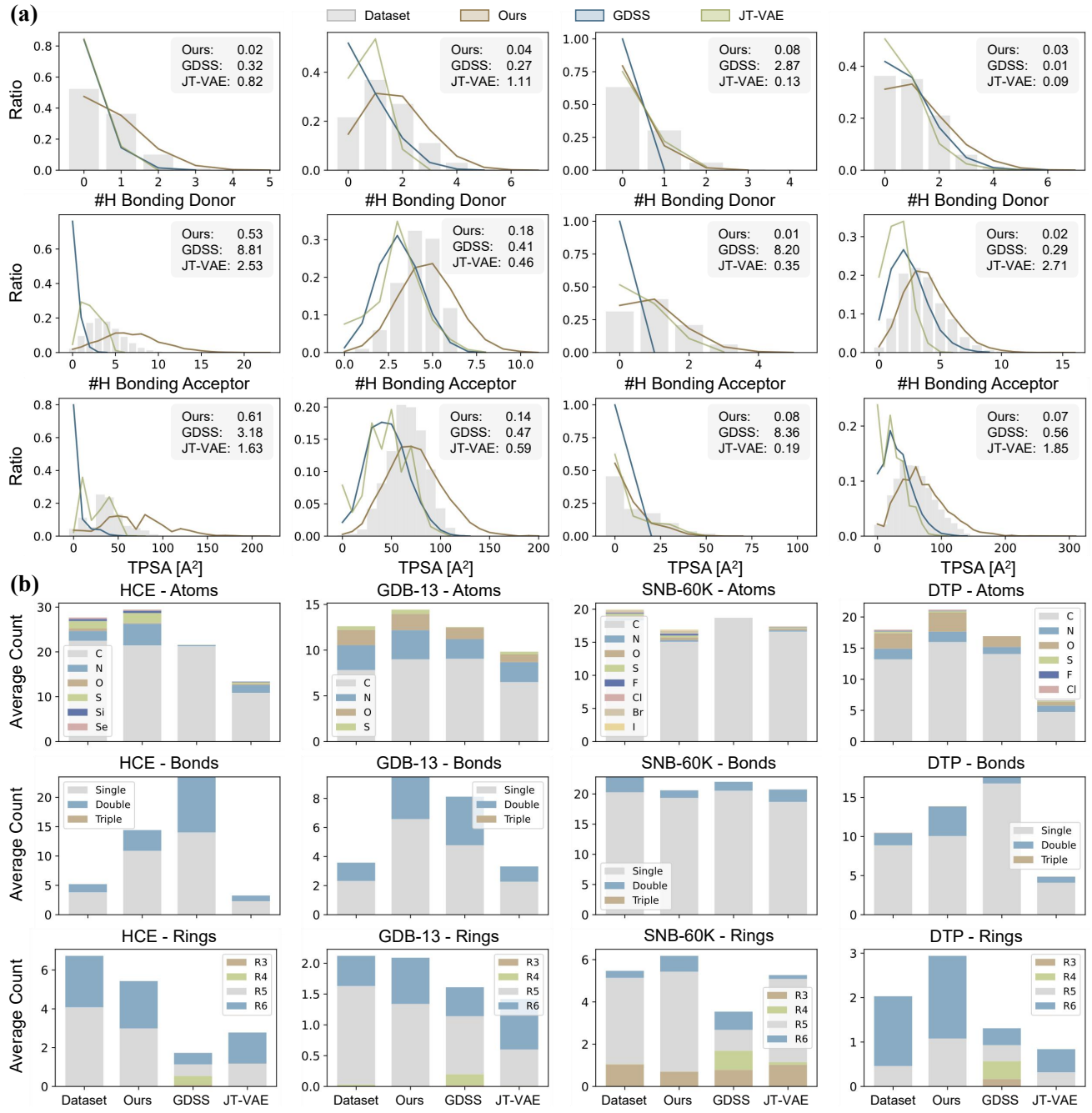
In this section, we visualize the word clouds indicating the degree to which GDSS and JT-VAE are superior to DM-Assembler in capturing specific molecular properties, as shown in Supplementary Figure 3.

Supplementary Section 12: Distribution of Partial Properties and Topological Structures

To facilitate a better understanding, we examine the distributions of several fundamental biochemical properties and topological structural characteristics across different models. The comparison, illustrated in Supplementary Figure 4, evaluates our approach against two established baselines: GDSS and JT-VAE.

Supplementary Section 13: Listing and Explanations of Properties

This section provides a detailed enumeration and description of all molecular properties examined in our experimental analysis. The complete listing and corresponding explanations are presented in Supplementary Table 5.



Supplementary Figure 4. **(a)** Comparative distribution of molecular properties #H Bonding Donor, #H Bonding Acceptor, and TPSA, with corresponding KL divergence values relative to the datasets shown in the top right corner. **(b)** Topological structure distributions comparing molecules from reference datasets with those generated by DM-Assembler, GDSS, and JT-VAE.

Supplementary Table 5. Molecular properties and corresponding explanations.

Molecular properties	Explanations
MolecularWeight	The total mass of the molecule, measured in Daltons (Da), which is the sum of the atomic weights of all atoms in the molecule.

Continued on next page

Continued from previous page

Molecular properties	Explanations
LogP	The logarithm of the partition coefficient between octanol and water, used to measure the hydrophobicity of the molecule.
TopologicalComplexity	Describes the complexity of the molecule's topological structure, typically referring to the complexity of how atoms and bonds are connected.
TPSA	The surface area of the molecule contributed by polar atoms (such as nitrogen and oxygen), often used to predict solubility and permeability.
ProportionSP3	The ratio of carbon atoms in SP3 hybridization (typically found in saturated bonds) relative to the total number of carbon atoms in the molecule.
ProportionRotatableBonds	The ratio of bonds in the molecule that allow free rotation, which affects the flexibility and conformational changes of the molecule.
ProportionStereocenters	The ratio of stereocenters (chiral centers) in the molecule, which influence the stereochemical properties and biological activity.
FractionHeteroatoms	The proportion of heteroatoms (such as nitrogen, oxygen, sulfur) relative to the total number of atoms in the molecule.
AliphaticRings	The number of saturated carbon rings (alkyl or cycloalkyl) in the molecule.
AromaticRings	The number of aromatic rings in the molecule, typically containing conjugated π-electrons.
TotalRings	The total number of rings in the molecule, including both aromatic and saturated rings.
NumHDonors	The number of hydrogen atoms in the molecule that can participate in hydrogen bonding by donating a hydrogen atom to another electronegative atom.
NumHAcceptors	The number of electronegative atoms (such as oxygen, nitrogen, or fluorine) that can accept a hydrogen bond from a donor.
BalabanIndex	A topological index that measures the complexity of the molecule's structure, where a higher value typically indicates more complex molecular connectivity.
WienerIndex	A topological index that represents the sum of the distances between all pairs of atoms in the molecule, often used to predict chemical reactivity.
ZagrebIndex	A topological index based on the connectivity of atoms in the molecule, used to describe molecular complexity.
ShannonEntropyDegreeDistribution	The entropy value of the degree distribution of the atoms in the molecule, which reflects the uniformity or diversity of atomic connections.
KierChiIndex1	A topological index based on atomic distances and connectivity, used to describe the overall shape and interatomic interactions in the molecule.
KierChiIndex2	A variant of the Kier Chi Index 1 that takes into account more complex interatomic relationships.
AdjacencyMatrixTrace	The trace (sum of diagonal elements) of the adjacency matrix, which describes the connectivity of atoms in the molecule.
DistanceMatrixSum	The sum of all elements in the distance matrix, representing the overall compactness of the molecule.
LogPTPSARatio	The ratio of LogP (lipophilicity) to TPSA (polar surface area), used to predict the solubility and permeability of molecules.
DegreeDistributionVariance	The variance of the degree distribution of the atoms in the molecule, which reflects the diversity of atomic connections.
MolecularDiameter	The diameter of the molecule, typically determined by the spatial distribution of atoms and bonds.
EstradaIndex	A topological index based on the matrix eigenvalues of the molecular connectivity matrix, used to measure the complexity of the molecular structure.
MaxEStateIndex	The highest electronic state index, representing the distribution of electron density within the molecule.
MinEStateIndex	The lowest electronic state index, which corresponds to the least favorable electronic state in the molecule.
MaxAbsEStateIndex	The largest absolute value of the electronic state index, representing the most significant electronic feature of the molecule.

Continued on next page

Continued from previous page

Molecular properties	Explanations
MinAbsEStateIndex	The smallest absolute value of the electronic state index, reflecting the least pronounced electronic feature.
qed	A measure of how "drug-like" a molecule is, with higher values indicating greater similarity to known drugs in terms of molecular properties.
MolWt	Same as Molecular Weight, representing the mass of the molecule.
HeavyAtomMolWt	The total molecular weight excluding hydrogen atoms.
ExactMolWt	The precise molecular weight, taking into account isotopic variations.
NumValenceElectrons	The total number of valence electrons in the molecule, which determines its bonding and reactivity.
NumRadicalElectrons	The number of unpaired electrons in the molecule, which are associated with free radical reactivity.
MaxPartialCharge	The highest partial charge on any atom in the molecule, indicating regions of high electronic density.
MinPartialCharge	The lowest partial charge on any atom in the molecule, indicating regions of low electronic density.
MaxAbsPartialCharge	The highest absolute value of partial charge, reflecting the most significant charge disparity in the molecule.
MinAbsPartialCharge	The lowest absolute value of partial charge, indicating the least pronounced charge disparity.
FpDensityMorgan1 to 3	The density of the first to third order Morgan fingerprints.
BCUT2D_MWHI	The high molecular weight portion of the 2D BCUT index, reflecting larger molecular components.
BCUT2D_MWLOW	The low molecular weight portion of the 2D BCUT index, indicating smaller molecular components.
BCUT2D_CHGHI	The high charge portion of the 2D BCUT index, reflecting regions with significant electron density.
BCUT2D_CHGLO	The low charge portion of the 2D BCUT index, indicating areas with lower electron density.
BCUT2D_LOGPHI	The portion of the 2D BCUT index associated with high LogP values, reflecting hydrophobic regions.
BCUT2D_LOGPLOW	The portion of the 2D BCUT index associated with low LogP values, indicating hydrophilic regions.
BCUT2D_MRHI	The high refractivity portion of the 2D BCUT index, related to the polarizability of the molecule.
BCUT2D_MRLOW	The low refractivity portion of the 2D BCUT index, indicating less polarizable regions.
BalabanJ	A measure of molecular complexity based on the connectivity of atoms in the molecule.
BertzCT	A measure of the molecular complexity based on graph theory, reflecting the diversity of atomic connections.
Chi0	A topological index based on the number of bonds between atoms.
Chi0n	A normalized version of Chi Index 0, adjusting for molecular size.
Chi0v	A variant of Chi Index 0 that focuses on atomic vertices in the molecular graph.
Chi1	A topological index that accounts for the bond connectivity between atoms.
Chi1n	A normalized version of Chi Index 1.
Chi1v	A variant of Chi Index 1 that focuses on atomic vertices in the molecular graph.
Chi2n	A normalized version of Chi Index 2, which is based on molecular branching.
Chi2v	A variant of Chi Index 2 based on atomic vertices.
Chi3n	A normalized version of Chi Index 3, which takes into account the branching complexity.
Chi3v	A vertex-based version of Chi Index 3, reflecting atomic connectivity.
Chi4n	A normalized version of Chi Index 4, representing higher-level branching.
Chi4v	A vertex-based version of Chi Index 4.

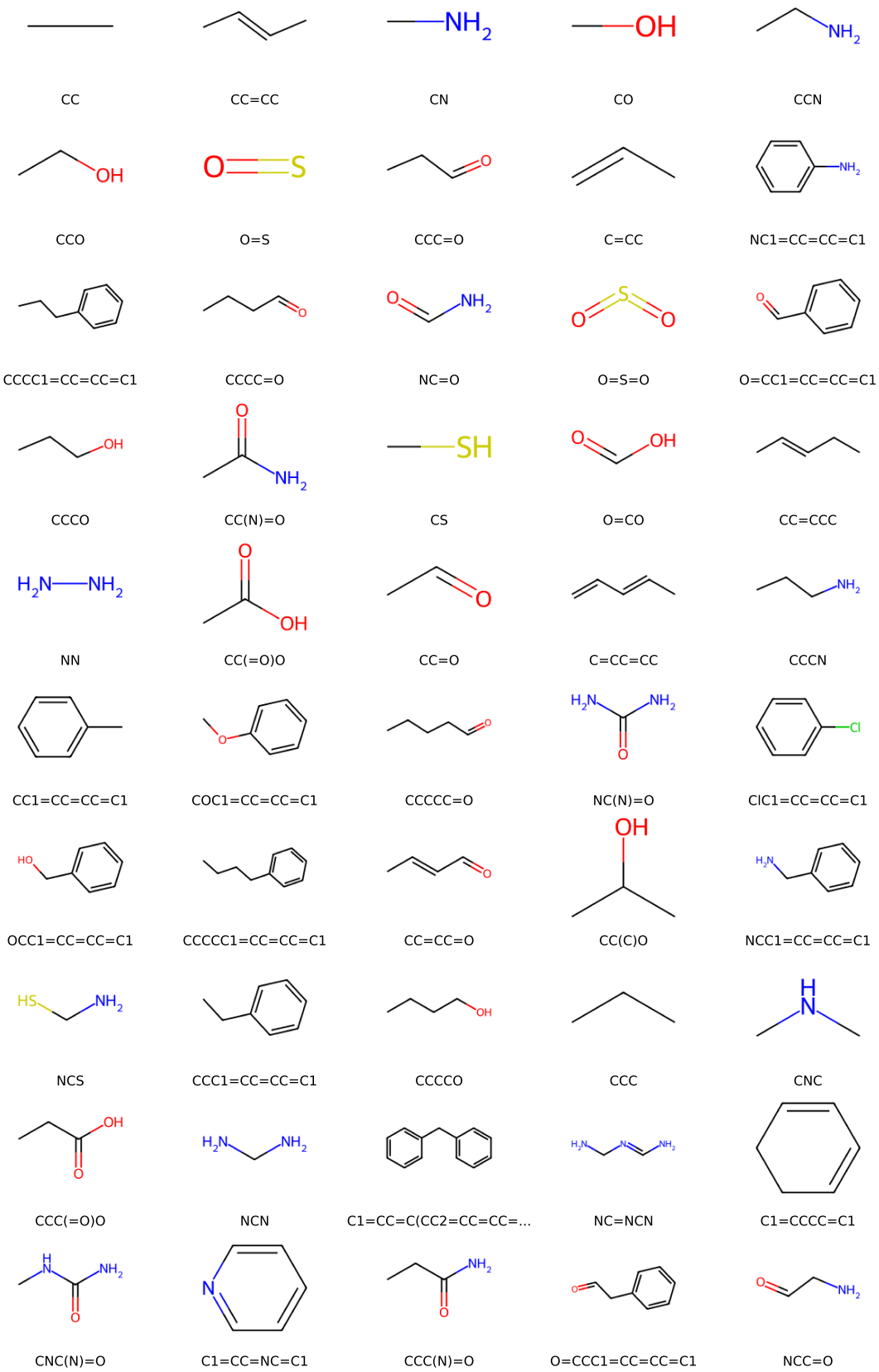
Continued on next page

Continued from previous page

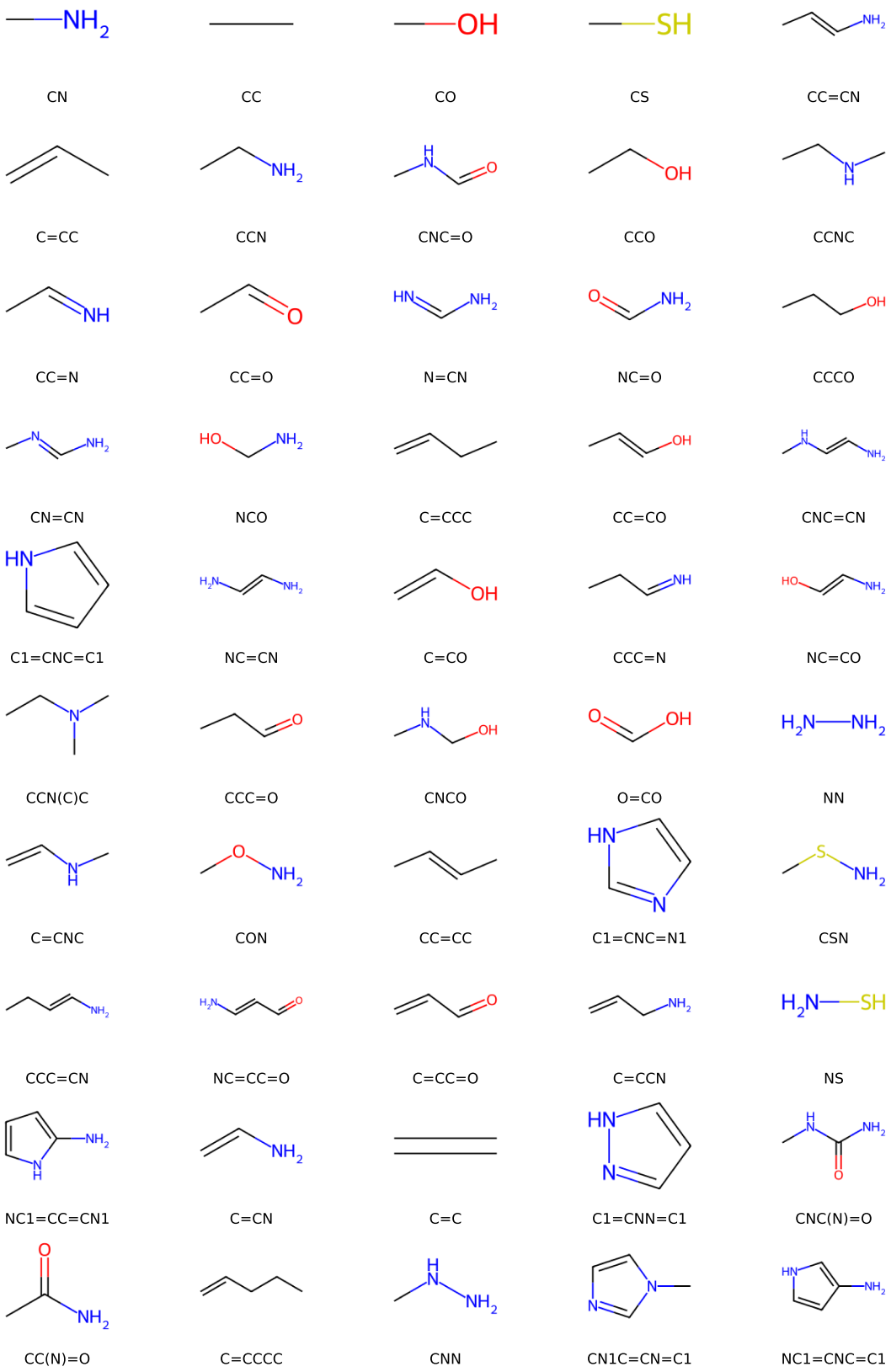
Molecular properties	Explanations
HallKierAlpha	A topological index that represents the overall shape of the molecule, considering atomic distances.
Ipc	The energy required to ionize the molecule, representing its electronic properties.
Kappa1	A topological index that measures molecular shape and complexity.
Kappa2	A topological index that measures molecular shape and complexity.
Kappa3	A third-order Kappa index that considers more complex topological features.
LabuteASA	The accessible surface area of the molecule, which can be used to predict its interaction with solvents or other molecules.
PEOE_VSA1 to PEOE_VSA14	Various surface area descriptors based on the Partial Equalized Energy of Atoms (PEOE) model, representing different polarity and electron density regions on the molecular surface.
SMR_VSA1 to SMR_VSA10	Various surface area descriptors based on the SMR (Symmetry of Molecular Representation) model, representing symmetry-related features of the molecule's surface.
SlogP_VSA1 to SlogP_VSA12	Surface area descriptors based on the SlogP model, representing different levels of hydrophobicity and polarity on the molecular surface.
EState_VSA1 to EState_VSA11	Surface area descriptors based on the electronic state (EState) model, reflecting electron density distributions on the molecule's surface.
VSA_EState1 to VSA_EState10	Descriptors based on the electronic state (EState) model, capturing electron density and molecular surface interactions.
FractionCSP3	The proportion of carbon atoms in SP3 hybridization, which influences the molecule's rigidity and stability.
HeavyAtomCount	The total number of non-hydrogen atoms in the molecule.
NHOHCount	The number of functional groups containing hydroxyl or amino functionalities in the molecule.
NOCount	The number of bonds between nitrogen and oxygen atoms, often seen in polar functional groups.
NumAliphaticCarbocycles	The number of saturated carbon rings (alkyl or cycloalkyl) in the molecule.
NumAliphaticHeterocycles	The number of saturated heterocyclic rings in the molecule.
NumAliphaticRings	The total number of saturated rings in the molecule.
NumAromaticCarbocycles	The number of aromatic carbon rings in the molecule.
NumAromaticHeterocycles	The number of aromatic heterocyclic rings in the molecule.
NumAromaticRings	The total number of aromatic rings in the molecule.
NumHeteroatoms	The total number of heteroatoms (atoms other than carbon and hydrogen, such as nitrogen, oxygen, sulfur, etc.) in the molecule.
NumRotatableBonds	The number of bonds in the molecule that allow free rotation, influencing molecular flexibility.
NumSaturatedCarbocycles	The number of saturated carbon rings in the molecule.
NumSaturatedHeterocycles	The number of saturated heterocyclic rings in the molecule.
NumSaturatedRings	The total number of saturated rings in the molecule.
RingCount	The total number of rings, including both aromatic and saturated rings.
MolLogP	The LogP value of the molecule, which reflects its lipophilicity or hydrophobicity.
Ring_size_3	The number of ring structures containing three atoms in a molecule.
Ring_size_4	The number of ring structures containing four atoms in a molecule.
Ring_size_5	The number of ring structures containing five atoms in a molecule.
Ring_size_6	The number of ring structures containing six atoms in a molecule.
Ring_size$\geq$7	The number of ring structures containing seven or more atoms in a molecule.

Supplementary Section 14: Visualization of Molecular Motifs

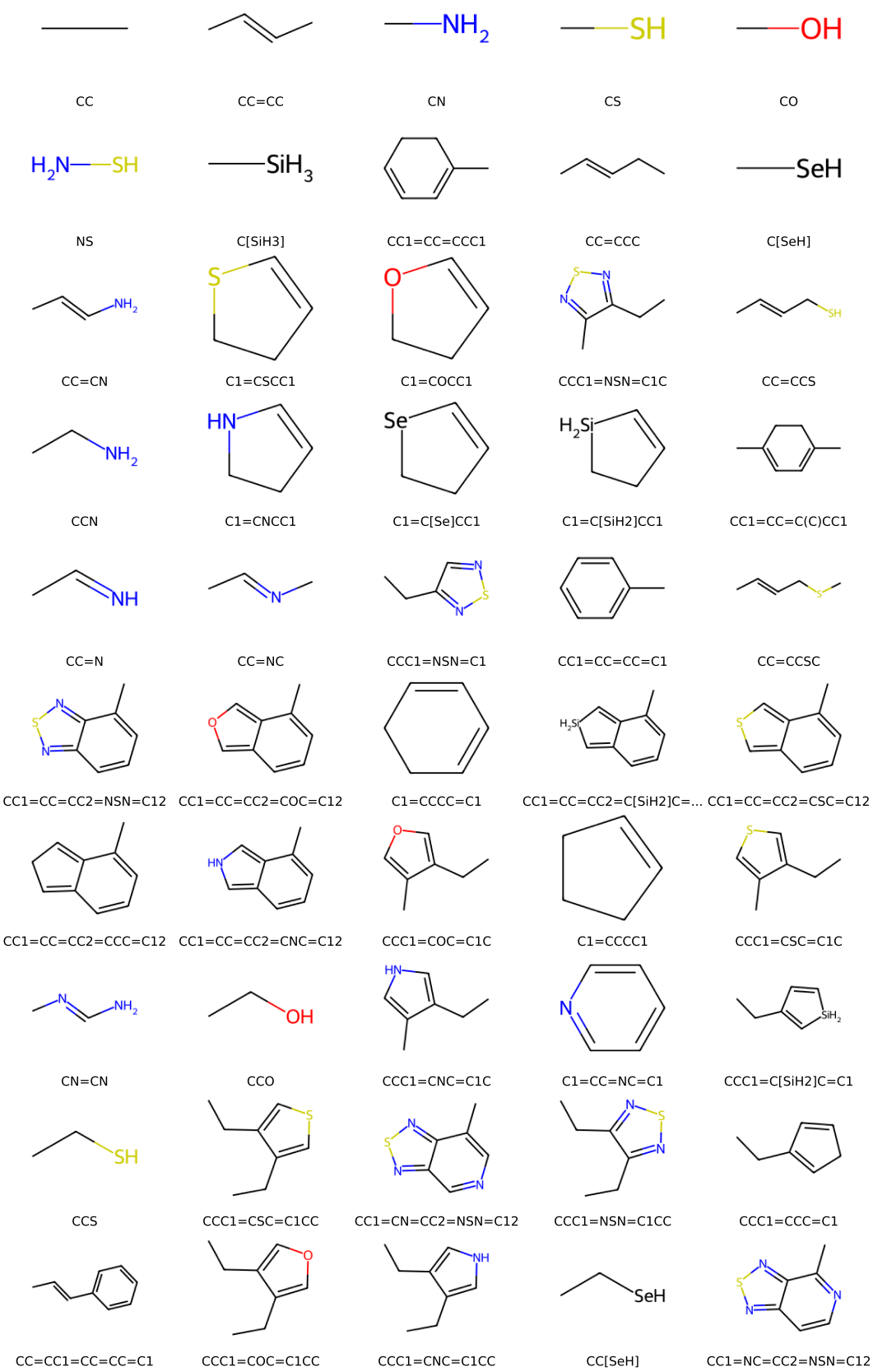
We visualize the top 50 frequency-ranked motifs extracted from the DTP, GDB-13, HCE, and SNB-60K datasets using our motif mining algorithm in Supplementary Figures 5-8, respectively, which offer insight into the characteristic structural patterns within each dataset.



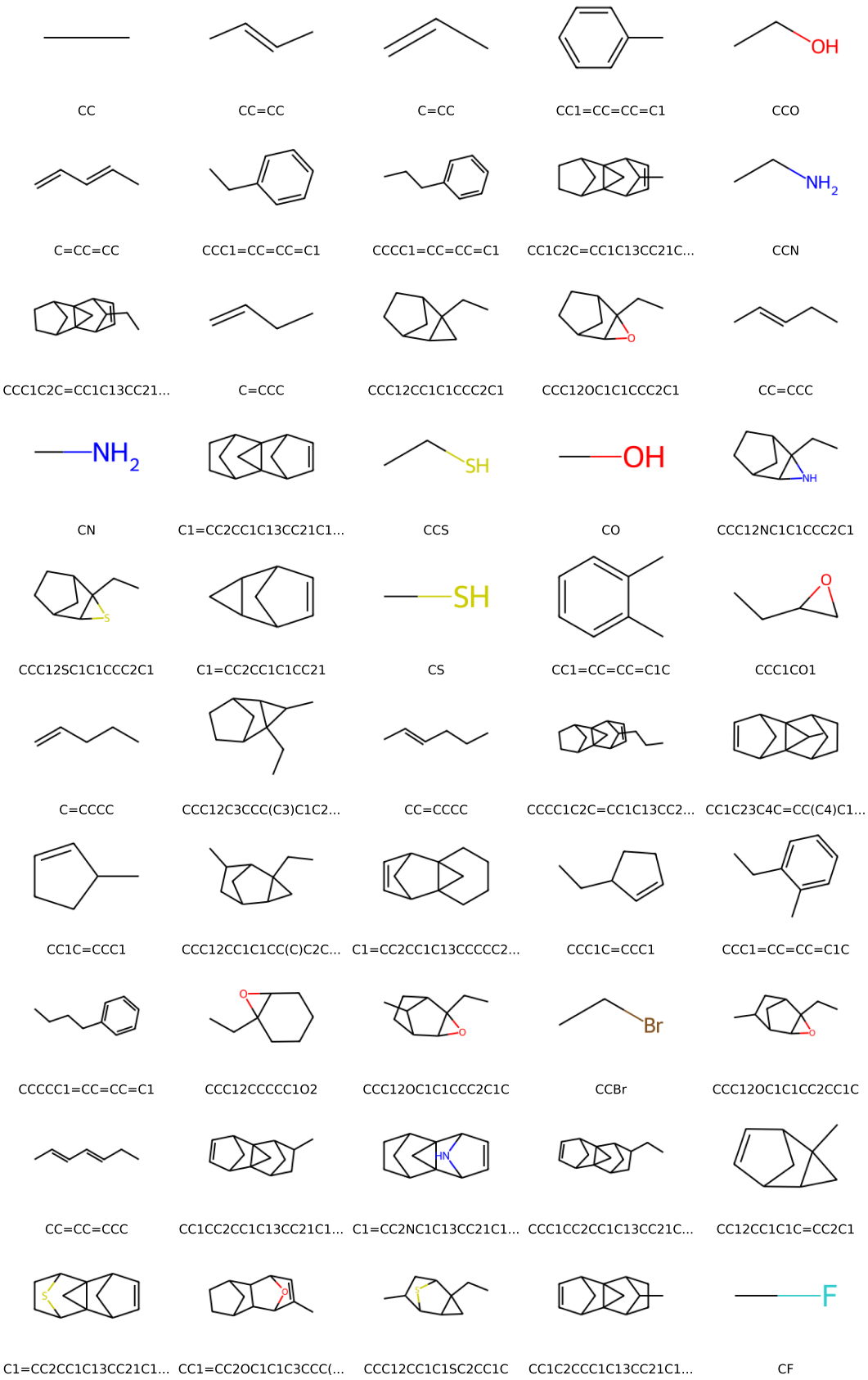
Supplementary Figure 5. The top-50 motifs of the DTP dataset.



Supplementary Figure 6. **The top-50 motifs of the GDB-13 dataset.**



Supplementary Figure 7. **The top-50 motifs of the HCE dataset.**



Supplementary Figure 8. **The top-50 motifs of the SNB-60K dataset.**

References

1. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276, DOI: 10.1021/acscentsci.7b00572 (2018). PMID: 29532027, https://doi.org/10.1021/acscentsci.7b00572.

2. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108, DOI: 10.1021/acs.jcim.8b00839 (2019). PMID: 30887799, https://doi.org/10.1021/acs.jcim.8b00839.

3. Zang, C. & Wang, F. Moflow: An invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 617–626, DOI: 10.1145/3394486.3403104 (Association for Computing Machinery, New York, NY, USA, 2020).

4. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminformatics* **9** (2017).

5. Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chem. Sci.* **10**, 3567–3572, DOI: 10.1039/C8SC05372C (2019).

6. Jo, J., Lee, S. & Hwang, S. J. Score-based generative modeling of graphs via the system of stochastic differential equations. *arXiv:2202.02514* (2022).

7. Geng, Z. *et al.* De novo molecular generation via connection-aware motif mining. In *International Conference on Learning Representations* (2023).

8. Edwards, C. *et al.* Translation between molecules and natural language. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022* (2022).

9. Nigam, A. *et al.* Tartarus: A benchmarking platform for realistic and practical inverse molecular design. *Adv. Neural Inf. Process. Syst.* **36**, 3263–3306 (2023).

10. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. cheminformatics* **1**, 8 (2009).

11. Weininger, D., Weininger, A. & Weininger, J. L. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101, DOI: 10.1021/ci00062a008 (1989).

12. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33, DOI: 10.1186/1758-2946-3-33 (2011).

13. Pracht, P. & Grimme, S. Conformer-Rotamer Ensemble Sampling Tool. https://github.com/grimme-lab/crest (2020).

14. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671, DOI: 10.1021/acs.jctc.8b01176 (2019).

15. Ameri, T., Dennler, G., Lungenschmied, C. & Brabec, C. J. Organic tandem solar cells: A review. *Energy Environ. Sci.* **2**, 347–363, DOI: 10.1039/B817952B (2009).

16. Li, H. *et al.* Decoupled peak property learning for efficient and interpretable ecd spectra prediction. In *Nature Computational Science* (2024).

17. Landrum, G. Rdkit documentation. *Release* **1**, 4 (2013).

18. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *J. Chem. Theory Comput.* **13**, 1989–2009, DOI: 10.1021/acs.jctc.7b00118 (2017). PMID: 28418654, https://doi.org/10.1021/acs.jctc.7b00118.

19. Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. chimica acta* **28**, 213–222 (1973).

20. Sun, Q. *et al.* Pyscf: the python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.* **8**, e1340, DOI: https://doi.org/10.1002/wcms.1340 (2018). https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1340.

21. Alhossary, A., Handoko, S. D., Mu, Y. & Kwoh, C.-K. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics* **31**, 2214–2216 (2015).

22. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *J. chemical information modeling* **53**, 1893–1904 (2013).

23. Gage, P. A new algorithm for data compression. *C Users J.* **12**, 23–38 (1994).

24. 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).