

Supplementary Information for

Semantic-level multimodal molecular learning inspired by biological concept formation via soft matching

Lin Feng et al.

Contents

Supplementary Notes 1-8

Supplementary Note 1 | Dataset statistics

Supplementary Note 2 | Data preprocessing and feature extraction

Supplementary Note 3 | Data partitioning and evaluation strategy

Supplementary Note 4 | Modality contribution analysis

Supplementary Note 5 | Raw data for Fig. 3d of the manuscript

Supplementary Note 6 | Clustering analysis of molecular representations and examples of intra-cluster molecular structures

Supplementary Note 7 | Classification distribution and distance heatmap to centers for all molecules in the BBBP dataset

Supplementary Note 8 | Metric calculation

Supplementary Tables 1-5

Supplementary Table 1 | Details of nine molecular property benchmark datasets.

Supplementary Table 2 | Raw data corresponding to Figs. 3a and 3b of the manuscript.

Supplementary Table 3 | Raw data corresponding to Fig. 3d of the manuscript: average SHAP values of the most important modality features for the top 30 molecules with the highest prediction scores on the BBBP dataset.

Supplementary Table 4 | Raw data corresponding to Figs. 3 and 5 of the manuscript: SMARTS patterns of the functional groups involved.

Supplementary Table 5 | Detailed parameter configurations.

Supplementary Figures 1-10

Supplementary Fig. 1 | Modality importance analysis on the BACE, ClinTox, ESOL, FreeSolv, Lipophilicity, SIDER, Tox21, and ToxCast datasets. The top 30 molecules were used as statistical units, and the average SHAP values of their features are reported. Feature prefixes are defined as follows: C denotes 1D SMILES, G denotes 2D graph structures, and 3D denotes three-dimensional geometry. The numbers on the left side of each subfigure represent the indices of the corresponding molecules in the dataset.

Supplementary Fig. 2 | a, Overall contribution proportions of molecular graphs, SMILES sequences, and 3D descriptors across the BACE, ClinTox, ESOL, FreeSolv, Lipophilicity, SIDER, Tox21, and ToxCast datasets. b, Interpretability of SemMol in molecular intervention experiments. For each dataset, the top 20 molecules and their top 50 features are visualized with SHAP heatmaps (red/blue = positive/negative contribution, darker/lighter = stronger/weaker intensity), covering the same eight datasets.

Supplementary Fig. 3 | Ablation study of the Dynamic Central Library (DCL) within the SemMol framework, with statistical significance tests validating its effectiveness. Performance comparison of the full model (ALL), Abl-1, Abl-2, and Abl-3 across a six MoleculeNet classification and b three regression datasets, respectively. “Abl-1” removes the entire DCL module and replaces it with random centers. “Abl-2” removes the periodic update mechanism of dynamic centers, fixing initial clustering centers. “Abl-3” weakens the update mechanism by setting the EMA momentum coefficient β in Eq. 4 of the manuscript to a value close to 1 (e.g., 0.999). The first six datasets are classification tasks, and the last three are regression tasks. Error bars indicate standard deviation across 10 independent runs.

Supplementary Fig. 4 | Ablation study of the Dynamic Central Library (DCL) within the SemMol framework, with statistical significance tests validating its effectiveness. Statistical significance analysis comparing ALL, Abl-1, Abl-2, and Abl-3 across a six MoleculeNet classification and b three regression datasets, respectively. “Abl-1” removes the entire DCL module and replaces it with random centers. “Abl-2” removes the periodic update mechanism of dynamic centers, fixing initial clustering centers. “Abl-3” weakens the update mechanism by setting the EMA momentum coefficient β in Eq. 4 of the manuscript to a value close to 1 (e.g., 0.999). The first six

datasets are classification tasks, and the last three are regression tasks. Error bars indicate standard deviation across 10 independent runs.

Supplementary Fig. 5 | Ablation study on the effectiveness of the ACSM mechanism in the SemMol framework. Synergistic effects of ACSM and unified spatial projection. Performance comparison of the full model (ALL), “Abl-D,” “Abl-E,” and “Abl-DE” across a six MoleculeNet classification and b three regression datasets, respectively. “Abl-D” removes the unified spatial projection within DCL component; “Abl-E” removes the ACSM component (Fig. 2b of the manuscript); “Abl-DE” removes both and falls back to instance-level contrastive learning (Figs. 2b and 2c of the manuscript). Bars show mean performance across 10 independent runs, with error bars indicating standard deviation.

Supplementary Fig. 6 | Ablation study on the effectiveness of the soft matching strategies in ACSM mechanism. Effectiveness of positive sample generation strategies in the soft matching mechanism across a six MoleculeNet classification and b three regression datasets, respectively.. Under unified spatial projection, performance of “ALL,” “W/U,” “W/T,” and “W/I” is compared via significance testing. “ALL” (full model) uses similarity-weighted fusion of retrieved centers; “W/U” applies uniform weighting; “W/T” retrieves only the Top-1 center; “W/I” applies traditional instance-level contrastive learning. Bilateral paired t-tests assess performance differences, with Cohen’s d measuring effect size. The significance level is $\alpha=0.05$, where $d\approx 0.2$, 0.5, and 0.8 indicate weak, moderate, and strong effects, respectively.

Supplementary Fig. 7 | Clustering analysis of molecular representations learned by the model on the BBBP, BACE, and ESOL datasets. Representations were projected into two dimensions using t-SNE for visualization, followed by clustering to evaluate distributional patterns across molecules.

Supplementary Fig. 8 | a, Visualization of molecules in Cluster 3 of the BBBP dataset; b, Visualization of molecules in Cluster 1 of the BACE dataset; c, Visualization of molecules in Cluster 3 of the ESOL dataset.

Supplementary Fig. 9 | Three collections a, b and c of representative molecular structures corresponding to clusters b, c, and d in Fig. 5 of the manuscript on the BBBP dataset.

Supplementary Fig. 10 | For all molecules in the BBBP dataset shown in Figure 6 of the manuscript: a, distribution of classes; and b, heatmap of their distances to the

cluster centers. Red indicates positive samples, blue indicates negative samples, and darker colors correspond to shorter distances.

Supplementary Notes

Supplementary Note 1 | Dataset statistics

To comprehensively evaluate the performance and generalization of the SemMol model in molecular representation learning, we adopted representative benchmark datasets from public sources such as MoleculeNet, covering both classification and regression tasks. Classification datasets include Tox21, ToxCast, BBBP, ClinTox, and SIDER; regression datasets include ESOL, FreeSolv, and Lipophilicity. These datasets span diverse challenges such as toxicity prediction, physicochemical property estimation, and quantum chemical property prediction, with varying scale and complexity. Detailed statistics (e.g., compound counts, average atom numbers, and task distributions) are provided in **Supplementary Table 1**.

(1) BACE: Contains quantitative (IC₅₀) and qualitative (binary) binding results for human β -secretase 1 (BACE-1) inhibitors.

(2) BBBP: Derived from studies modeling and predicting blood–brain barrier permeability.

(3) ClinTox: Compares FDA-approved drugs with those that failed clinical trials due to toxicity.

(4) Tox21: Includes toxicity measurements for ~8,000 compounds across 12 targets, including nuclear receptors and stress response pathways.

(5) ToxCast: Expanded dataset from the Tox21 initiative, providing high-throughput in vitro screening data from >600 assays on ~8,000 compounds.

(6) SIDER: Catalogues marketed drugs and their adverse drug reactions (ADRs), grouped into 27 system organ classes.

(7) FreeSolv: Reports experimental and computed hydration free energies for small molecules in water, with computational values derived from molecular dynamics simulations.

(8) ESOL: A regression dataset of 1,128 compounds, pairing molecular structures with water solubility data; widely used for benchmarking solubility prediction models.

(9) Lipophilicity (Lipophilicity/“Lipo”): Extracted from ChEMBL, containing experimental octanol–water partition coefficients (logD at pH 7.4), a critical property for permeability and solubility prediction.

Supplementary Note 2 | Data preprocessing and feature extraction

We systematically processed SMILES strings from each dataset to construct multimodal molecular representations:

1D Modality: Raw SMILES strings were directly encoded using the ChemBERTa model.

2D Modality: Molecular graphs were generated from SMILES strings using RDKit, then encoded with a pre-trained GROVER model. GROVER captures both local and global structural information via contrastive and generative tasks.

3D Modality: Conformations were generated using RDKit's ETKDGv2 algorithm, followed by minimization with the MMFF94 force field. The lowest-energy conformation was selected and encoded using DimeNet. This model integrates radial and angular basis functions to capture spatial features such as bond lengths, angles, and dihedrals. Node features in the 3D graph were aligned with atomic features in the 2D graph to support cross-modal alignment.

Feature extraction was primarily performed with ChemBERTa, GROVER, and DimeNet. Molecules that failed to generate valid 3D conformations were excluded.

1D Sequence Feature Extraction

For SMILES sequence encoding, we adopt the ChemBERTa-77M-MLM pre-trained model. This model is a 12-layer Transformer encoder, where each layer combines a multi-head self-attention mechanism and a feedforward network: the former aggregates information from multiple subspaces, while the latter enhances representational capacity via nonlinear transformations.

Given a SMILES string s , tokenization is first performed, followed by the TokenEmbedding function to generate initial embeddings. The sequence is then passed through the Transformer encoder, where the representation at layer $l+1$ is denoted as:

$$z^{(0)} = \text{TokenEmbedding}(s), \quad (1)$$

$$z^{(l+1)} = \text{TransformerLayer}^{(l)}(z^{(l)}), \quad l = 0, 1, \dots, L-1, \quad (2)$$

where L is the sequence length. The final token sequence $z^{(L)}$ is aggregated using a global pooling strategy, and the representation corresponding to the special [CLS] token serves as the overall molecular sequence embedding:

$$z_{1D} = z_{[\text{CLS}]}^{(12)}. \quad (3)$$

2D Graph Structure Feature Extraction

To capture structural and topological information, we employ the GROVER model, a graph Transformer pre-trained in a self-supervised manner. GROVER jointly optimizes node-, edge-, and graph-level objectives (e.g., masked atom prediction, substructure recognition), enabling effective encoding of chemical semantics.

Molecules are represented as graphs $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where \mathbf{V} is the set of atoms (nodes) and \mathbf{E} is the set of chemical bonds (edges). Node features are iteratively updated using a Gated Graph Convolutional Network (GatedGCN):

$$\mathbf{z}_v^{(l+1)} = \sigma(\mathbf{W}_1 \mathbf{z}_v^{(l)} + \sum_{u \in \text{Neigh}(v)} \text{Gate}(\mathbf{z}_u^{(l)}, \mathbf{e}_{uv})), \quad (4)$$

where $\mathbf{z}_v^{(l+1)}$ is the embedding of node v at layer $l+1$, $\text{Neigh}(v)$ is the neighbor set of v , \mathbf{e}_{uv} is the edge feature between nodes u and v , and $\text{Gate}(\cdot)$ denotes the gating function that computes adaptive weights based on node and edge features.

Finally, a graph-level attention pooling module (**AttentiveReadout**) aggregates node representations into a molecular graph embedding \mathbf{z}_{2D} , which preserves molecular topology, adjacency relationships, and atomic semantics.

$$\mathbf{z}_{2D} = \text{AttentiveReadout}(\{\mathbf{z}_v^{(L)} \mid v \in \mathbf{V}\}) \in \mathbb{R}^d. \quad (5)$$

3D Geometric Feature Extraction

To model geometric information, we adopt DimeNet, a directional message passing network designed for 3D molecular conformations $\{(\mathbf{p}_i, \mathbf{t}_i)\}_{i=1}^{\mathbf{N}_{\text{atom}}}$, where \mathbf{N}_{atom} denotes the number of atoms. Each molecule is represented by atomic coordinates \mathbf{p}_i and atomic indexes \mathbf{t}_i . Based on this, a molecular graph is constructed where both distances and angles are explicitly encoded:

$$\mathbf{r}_{ij}^l = \sum_{k \in \text{Neigh}(i) \setminus j} \text{MLP}(d_{ij}, d_{ik}, \angle_{ijk}, \mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{h}_k^l), \quad (6)$$

where **MLP** is a multilayer perceptron that jointly models distances, angles, and atomic features. And d_{ij} represents the distance between atoms i and j , \angle_{ijk} denotes the angle involving atoms i , j , and common neighbor k . Atomic representations are then updated:

$$\mathbf{z}_i^{(l+1)} = \text{MLP}\left(\mathbf{z}_i^{(l)} + \sum_{j \in \text{Neigh}(i)} \mathbf{r}_{ij}^l\right). \quad (7)$$

After \mathbf{H} layers of message passing, a global pooling operation yields the final 3D-aware molecular embedding \mathbf{z}_{3D} , which effectively captures geometric interactions and directional dependencies:

$$\mathbf{z}_{3D} = \text{Pool}(\{\mathbf{z}_i^{(H)} \mid i=1, \dots, \mathbf{N}\}) \in \mathbb{R}^d. \quad (8)$$

Supplementary Note 3 | Data partitioning and evaluation strategy

To ensure objectivity in evaluation and assess the model's generalization capability, particularly its performance on structurally novel compounds, we employ a rigorous molecular scaffold splitting strategy. Specifically, the Bemis–Murcko skeleton for each molecule was extracted using RDKit. Molecules sharing the same skeleton were grouped together, then randomly shuffled and split into training, validation, and test sets at an 8:1:1 ratio. This ensures that molecules with the same scaffold only appear in one dataset, preventing data leakage due to structural similarity. This partitioning simulates “de novo design” or “novel scaffold discovery” scenarios in drug discovery, providing a more robust benchmark for SemMol's generalization in practical applications.

Supplementary Note 4 | Modality contribution analysis

The modality contribution analysis was performed for eight datasets excluding BBBP (BACE, ClinTox, ESOL, FreeSolv, Lipophilicity, SIDER, Tox21, ToxCast), as shown in **Supplementary Figs. 1 and 2**. **Supplementary Fig. 1** presents the average SHAP values of features, computed using the top-30 molecules as statistical units. Feature prefixes denote: C (1D SMILES), G (2D graph structure), 3D (3D geometry). **Supplementary Fig. 2a** shows the contribution ratio of each modality across all datasets, while **Supplementary Fig. 2b** presents the modality feature contribution ratio for the top-20 molecules. Overall, SemMol demonstrates complementary effects in the relative contributions of each modality to model predictions.

Supplementary Note 5 | Raw data for Fig. 3d of the manuscript

Supplementary Table 3 provides the raw data for **Fig. 3d** of the manuscript, presenting the average SHAP values of the most important feature vectors for the top-30 molecules with the highest prediction scores on the BBBP dataset. The table includes the following data: molecular indices within the BBBP dataset; modalities (1D_SMILES, 2D_Graph, 3D_Geo); feature indices and names for each modality; corresponding average SHAP values.

Supplementary Note 6 | Clustering analysis of molecular representations and examples of intra-cluster molecular structures

Supplementary Fig. 7 displays the clustering results of molecular representations obtained from the BBBP, BACE, and ESOL datasets. The clustering was evaluated using the Davies-Bouldin index (DB index), which measures the ratio of intra-class compactness to inter-class separation (lower values indicate better clustering). Each subfigure title includes the corresponding DB index, with red asterisks marking the cluster centers. Results demonstrate that SemMol's learned molecular representations form tightly clustered, well-separated structures. Notably, the ESOL dataset achieved the lowest DB index of 0.57, with the clearest cluster boundaries, indicating the model effectively captures latent semantic features related to physicochemical properties.

Additionally, one cluster each from the BBBP, BACE, and ESOL datasets is shown in **Supplementary Figs. 8(a-c)**, respectively. Within these clusters, molecules exhibit overall structural similarity, which is not solely dependent on specific functional groups, but rather arises from combinations of similar structural motifs. This characteristic leads to less overfitting under the anchor-multicenter matching mechanism, as compared to instance-level alignment models.

Supplementary Note 7 | Classification distribution and distance heatmap to centers for all molecules in the BBBP dataset

The distance relationships between all samples and centers on the BBB dataset were quantified, with darker colors indicating closer distances. The heatmap shows that samples are enriched around specific centers, suggesting these centers represent molecules with similar structural characteristics.

In **Supplementary Fig. 10b**, the distance distributions of all samples to 2D-C4 and 3D-C1 align with the classification distributions in **Supplementary Fig. 10a**, indicating that the molecular structures in these clusters are critical for positive category classification. Conversely, clusters such as 2D-C3, 2D-C6, and 3D-C3 correspond to negative class features.

Positive samples are concentrated on the left (darker colors), indicating they are effectively pulled closer to the anchor. Negative samples are on the right (lighter colors), showing they are repelled. This separation demonstrates that SemMol successfully achieves positive – negative sample discrimination through target optimization. The anchor–center soft matching mechanism ensures anchors maintain balanced similarity with multiple centers, enhancing the model’s generalization capability.

Supplementary Note 8 | Metric calculation

(1) Noise robustness in *Fig. 4 of the manuscript* in classification tasks

The decision boundary in a classification model is defined by the conditional probability distribution $P(y|x)$. Noise distorts the feature space, causing misclassification of samples near the boundary. For binary classification tasks, noise sensitivity measures the model's robustness to feature perturbations. The label distribution under noise can be expressed as:

$$P(y|x) = (1-\epsilon)P(y=1|x) + \epsilon P(y=0|x), \quad (9)$$

where ϵ is the noise rate and $y=1$ denotes the noisy case.

The dynamic center library reduces the impact of noise ϵ_{eff} through feature aggregation:

$$\epsilon_{\text{eff}} = \epsilon \cdot \frac{\text{Var}(x)}{\text{Var}(c_k)}, \quad (10)$$

where c_k is the cluster center with smaller variance than the original features. And $\text{Var}(\cdot)$ denotes the empirical variance of molecular representations in the latent feature space and serves as a measure of feature dispersion and robustness to noise perturbations.

(2) Inter-modal gradient conflict rate (*Fig. 4 of the manuscript*)

The inter-modal gradient conflict rate evaluates the frequency of inconsistent gradient directions between different modalities during multi-modal model training. This metric assesses the model's ability to coordinate gradients from different modalities, ensuring stable parameter updates. A low conflict rate indicates that multi-modal signals are effectively guiding parameter updates collaboratively.

For the multimodal shared representation space in the dynamic center library (DCL), the shared parameters are denoted as θ , which encode modality-agnostic semantic centers and are dynamically updated through joint multimodal optimization. If gradients from modalities m_i and m_j satisfy $\nabla_{\theta} L_{m_i} \cdot \nabla_{\theta} L_{m_j} < 0$, this is counted as one conflict. The conflict rate formula is:

$$\text{Conflict Rate} = \frac{\sum_{i=1}^T \sum_{i < j} \mathbf{I}(\nabla_{\theta} L_{m_i} \cdot \nabla_{\theta} L_{m_j} < 0)}{T \cdot \binom{|M|}{2}}, \quad (11)$$

where $\nabla_{\theta} L_{m_i}$ denote the gradient of the modality-specific loss L_{m_i} with respect to the shared parameters θ . The inner product $\nabla_{\theta} L_{m_i} \cdot \nabla_{\theta} L_{m_j}$ between the gradients of modality i and modality j is used to assess their directional consistency: a positive value indicates aligned gradient directions, whereas a negative value indicates

opposing gradient directions. $\mathbf{I}(\cdot)$ denotes the indicator function, which takes the value 1 when the condition inside the parentheses is satisfied and 0 otherwise. $\mathbf{M}=\{\mathbf{1D}, \mathbf{2D}, \mathbf{3D}\}$ denotes the set of modalities, and \mathbf{T} denotes the total number of training steps.

(3) *Outlier error in regression tasks (Fig. 4 of the manuscript)*

In regression tasks, risk in extreme value regions is quantified as:

$$\mathbf{R}_{\text{ext}} = \int_{|y|>t} (f(\mathbf{x}) - y)^2 \mathbf{P}(\mathbf{x}) d\mathbf{x}, \quad (12)$$

where \mathbf{R}_{ext} denotes the outlier error, t is the threshold, $f(\mathbf{x})$ represents the regression prediction of the model for input \mathbf{x} , and $\mathbf{P}(\mathbf{x})$ denotes the probability density function of \mathbf{x} , which is used to weight the relative importance of different samples in the population. Prediction variance is then reduced through multimodal constraints:

$$\text{Var}(f(\mathbf{x})) \approx \sum_{m=1}^{|\mathbf{M}|} w_m \text{Var}(f_m(\mathbf{x})), \quad (13)$$

where $f_m(\mathbf{x})$ denotes the variance of the prediction results from the m -th modality, and w_m denotes the weight coefficient of the m -th modality in multimodal fusion.

(4) *Modal consistency in Fig. 4 of the manuscript*

Modal consistency measures how closely features from different modalities align in a unified metric space. It quantifies whether similar samples exhibit close feature representations across modalities and whether conflicts or complementarity exist. It is computed as the average cosine similarity:

$$\text{Consistency} = \frac{1}{|\mathbf{M}|} \sum_{i \neq j} \frac{1}{N} \sum_{n=1}^N \frac{z_{m_i}^{(n)} \cdot z_{m_j}^{(n)}}{\|z_{m_i}^{(n)}\| \|z_{m_j}^{(n)}\|}, \quad (14)$$

where $z_{m_i}^{(n)}$ denotes the feature representation vector obtained for the n -th sample under the modality m_i (see Equations 3, 5 and 8), and N is the number of samples.

(5) *Calculation formula for POR@Scaffold (Fig. 3 of the manuscript)*

POR@Scaffold measures the structural similarity between a perturbed molecule and its original form. Higher values indicate a greater degree of structural similarity. The change in model predictions after perturbing the structure is represented as $\Delta\hat{\mathbf{y}}$. For each scaffold group s , the true and predicted values are as follows:

\mathbf{G}_s : Index set of samples in the s -th scaffold group.

$N_s=|\mathbf{G}_s|$: Number of samples in the s -th group.

$\mathbf{y}_i \in \{0,1\}$: True label of sample i .

$\hat{\mathbf{y}}_i \in [0,1]$: Predicted score of sample i .

k : Top- k size after sorting samples by $\hat{\mathbf{y}}$.

$\mathbf{T}_{s@k}$: Index set of the top- k samples within the s -th group.

The formula for the true and predicted values in the (s)-th group are:

$$\mathbf{p}_s = \frac{1}{N_s} \sum_{i \in G_s} \mathbf{y}_i, \hat{\mathbf{p}}_s = \frac{1}{N_s} \sum_{i \in G_s} \hat{\mathbf{y}}_i, \quad (15)$$

Intra-group POR quantifies the improvement in the Top- k hit rate relative to a random baseline:

$$\mathbf{POR}_{s@k} = \frac{1}{k \cdot p_s} \sum_{i \in \mathbf{T}_{s@k}} \mathbf{y}_i \approx \frac{1}{k \cdot p_s} \sum_{i \in \mathbf{T}_{s@k}} \hat{\mathbf{y}}_i, \quad (16)$$

A value greater than 1 outperforms random, 1 is equivalent to random, and less than 1 is worse than random. Macro-average POR across all scaffold groups:

$$\mathbf{POR@ Scaffold} = \frac{1}{|\mathbf{S}|} \sum_{s \in \mathbf{S}} \mathbf{POR}_{s@k}, \quad (17)$$

For a single sample i , For individual samples, the change in prediction score:

$$\Delta \hat{\mathbf{y}}_i = \widehat{\mathbf{y}}_i^{\text{after}} - \widehat{\mathbf{y}}_i^{\text{before}}, \quad (18)$$

The change in group (s) prediction mean due to perturbations:

$$\Delta \mathbf{p}_s = \frac{1}{N_s} \sum_{i \in G_s} \Delta \hat{\mathbf{y}}_i, \quad (19)$$

$$\Delta \mathbf{POR}_{s@k} \approx \frac{1}{k \cdot p_s} \left[\sum_{i \in \mathbf{T}_{s@k}} \Delta \hat{\mathbf{y}}_i - \left(\sum_{i \in \mathbf{T}_{s@k}} \hat{\mathbf{y}}_i \right) \frac{\Delta p_s}{p_s} \right]. \quad (20)$$

Supplementary Tables

Supplementary Table 1 Details of nine molecular property benchmark datasets.

Datasets	Number of molecules	Number of tasks	Category	Metrics
BACE	1513	1	Biophysics	ROC-AUC
BBBP	2039	1	Physiology	
ClinTox	1478	2		
Tox21	7831	12		
ToxCast	8575	617		
SIDER	1427	27	Physical chemistry	RMSE
FreeSolv	642	1		
ESOL	1128	1		
Liop	4200	1		

Supplementary Table 2 Raw data corresponding to **Figs. 3a and 3b** of the manuscript.

Classification/Regression		ROC-AUC					RMSE		
Datasets	BACE	BBBP	ClinTox	Tox21	ToxCast	SIDER	FreeSolv	ESOL	Lipo
A-Smiles	0.782	0.709	0.824	0.738	0.631	0.608	1.923	0.982	0.762
	(0.013)	(0.003)	(0.018)	(0.009)	(0.004)	(0.011)	(0.06)	(0.005)	(0.015)
A-Graph	0.085	0.674	0.856	0.761	0.654	0.635	1.857	0.894	0.698
	(0.007)	(0.01)	(0.015)	(0.008)	(0.003)	(0.009)	(0.012)	(0.009)	(0.013)
A-Geo	0.736	0.711	0.792	0.712	0.602	0.581	2.014	1.076	0.834
	(0.011)	(0.005)	(0.021)	(0.011)	(0.005)	(0.013)	(0.003)	(0.002)	(0.018)
B-Smiles	0.831	0.718	0.881	0.768	0.662	0.644	1.651	0.792	0.651
	(0.003)	(0.009)	(0.014)	(0.007)	(0.003)	(0.009)	(0.004)	(0.007)	(0.012)
B-Graph	0.842	0.725	0.872	0.754	0.648	0.632	1.618	0.758	0.672
	(0.001)	(0.003)	(0.016)	(0.008)	(0.003)	(0.010)	(0.009)	(0.005)	(0.014)
B-Geo	0.854	0.741	0.895	0.776	0.671	0.656	1.573	0.723	0.623
	(0.004)	(0.005)	(0.013)	(0.007)	(0.002)	(0.004)	(0.008)	(0.009)	(0.012)
ALL	0.869	0.761	0.931	0.812	0.713	0.671	1.31	0.537	0.512
	(0.001)	(0.005)	(0.002)	(0.003)	(0.001)	(0.008)	(0.005)	(0.001)	(0.001)

Supplementary Table 3 Raw data corresponding to **Fig. 3d of the manuscript**: average SHAP values of the most important modality features for the top 30 molecules with the highest prediction scores on the BBBP dataset.

Rank	Molecular index	Modality	Mean shaply value	Rank	Molecular index	Modality	Mean shaply value
1	597	1D_SMILES	0.0042	16	1100	3D_Geo	0.0017
2	154	2D_Graph	0.0031	17	1042	3D_Geo	0.0017
3	1421	3D_Geo	0.0026	18	1156	3D_Geo	0.0017
4	685	1D_SMILES	0.0026	19	1503	3D_Geo	0.0017
5	647	1D_SMILES	0.0023	20	1180	3D_Geo	0.0017
6	18	2D_Graph	0.0021	21	370	2D_Graph	0.0017
7	1329	3D_Geo	0.0017	22	49	2D_Graph	0.0017
8	354	2D_Graph	0.0017	23	886	1D_SMILES	0.0017
9	1294	3D_Geo	0.0017	24	231	2D_Graph	0.0017
10	1445	3D_Geo	0.0017	25	1411	3D_Geo	0.0017
11	162	2D_Graph	0.0017	26	85	2D_Graph	0.0017
12	1259	3D_Geo	0.0017	27	1237	3D_Geo	0.0017
13	1288	3D_Geo	0.0017	28	493	2D_Graph	0.0017
14	360	2D_Graph	0.0017	29	8	2D_Graph	0.0017
15	657	1D_SMILES	0.0017	30	743	1D_SMILES	0.0016

Supplementary Table 4 Raw data corresponding to **Figs. 3 and 5 of the manuscript**: SMARTS patterns of the functional groups involved.

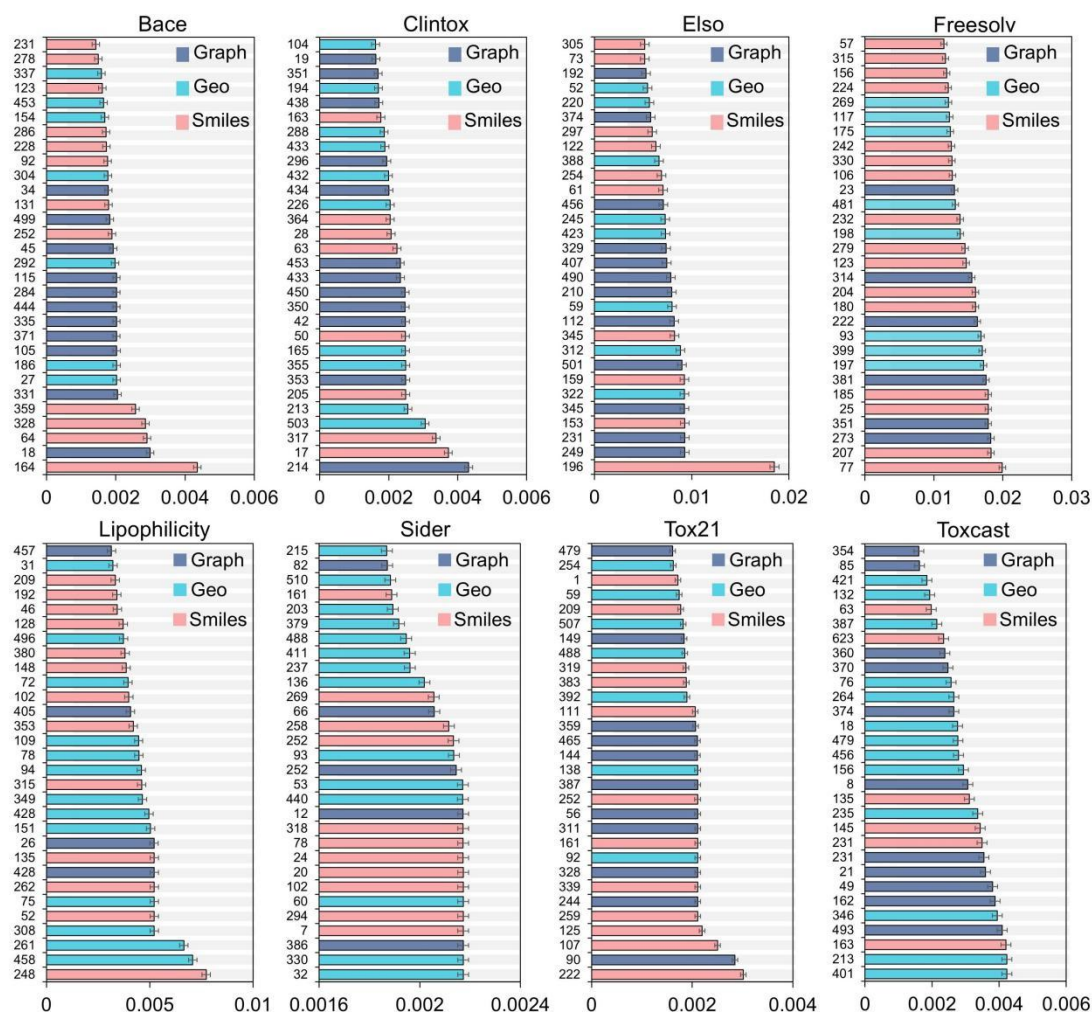
Functional Group	SMARTS
Carboxylic acid	<chem>C(=O)[O;H1]</chem>
Ester	<chem>C(=O)O[#6]</chem>
Amide	<chem>C(=O)N</chem>
Ketone	<chem>C(=O)C</chem>
Aldehyde	<chem>[CX3H1](=O)[#6]</chem>
Hydroxyl	<chem>[OX2H]</chem>
Ether	<chem>[OD2]([#6])[#6]</chem>
Amine	<chem>[NX3;H2,H1,H0;!\$(NC=O)]</chem>
Aromatic ring	<chem>a1aaaaa1</chem>
Phenyl	<chem>c1ccccc1[*]</chem>
Halogen	<chem>[F,Cl,Br,I]</chem>
Nitro	<chem>[N+](=O)[O-]</chem>
Carbonyl	<chem>[CX3]=[OX1]</chem>
Benzene	<chem>c1ccccc1</chem>

Supplementary Table 5 Detailed parameter configurations.

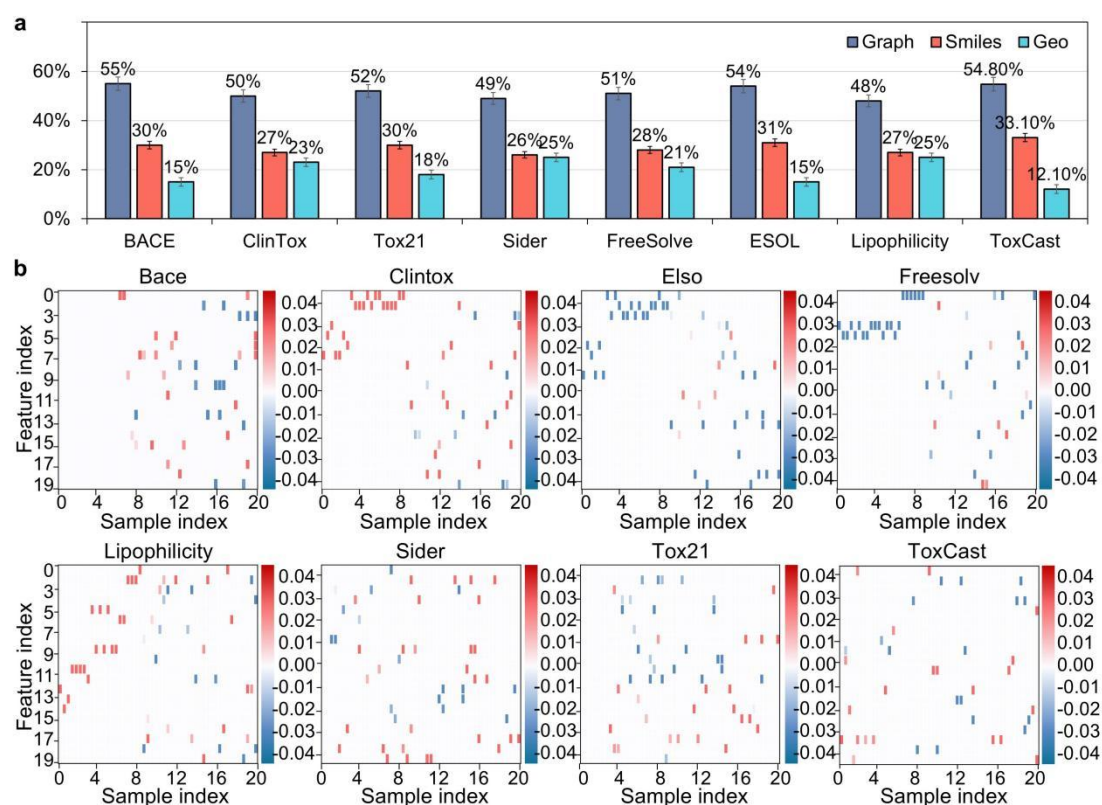
Parameters	Description	Range/options
pseudo_pair_hard_negative_k	The number of hard negative samples selected must be less than batch_size	choices: [1, 4, 8, 16, 32, 64, 128, 256, 384, 512, 768, 1024]; default: 10
task_settings.task_type	Task type: classification or regression	choices: ['classification', 'regression']; default: classification
data_settings.data_path	Data file path	CSV format
data_settings.target_column	Target column name	adjusted according to the dataset
data_settings.split_type	Data partitioning method	choices: ['scaffold', 'random']; default: scaffold
data_settings.train_val_test_ratio	The data set partition ratio	the sum must be 1.0
data_settings.normalize_targets	Whether to normalize the target value (regression tasks only)	choices: [True, False]; default: False
model_paths.bert_path	ChemBERTa Pre-trained model path	value: /root/autodl-tmp/SemMol/ChemBERTa-77M-MLM
model_paths.grover_path	GROVER Pre-trained model path	value: /root/autodl-tmp/SemMol/GROVER
training_core.epochs	Number of training rounds	default: 50; range: {'min': 20, 'max': 200}; suggested: [50, 100]
training_core.batch_size		choices: [16, 32, 64, 128]; default: 32
training_core.lr	Learning rate — the most important hyperparameter	default: 0.0005; range: {'min': 1e-05, 'max': 0.003, 'logscale': True}; suggested: [0.0001, 0.0003, 0.0005, 0.001]
training_core.weight_decay	L2 regularization weights	default: 0.0001; range: {'min': 0.0, 'max': 0.001}; suggested: [0.0, 1e-05, 0.0001, 0.0005]
training_core.scheduler		choices: ['none', 'cosine', 'step', 'plateau']; default: cosine; recommendations: {'classification': 'plateau', 'regression': 'cosine'}
training_core.seed	Random seed for reproducible results	default: 42; range: {'min': 1, 'max': 9999}; suggested: [42, 3407, 2025]
model_architecture.fusion_type	Feature fusion strategy	choices: ['mean', 'concat_mlp', 'attention']; default: concat_mlp; recommendations: {'classification': 'concat_mlp', 'regression': 'attention'}
model_architecture.dropout_rate	Dropout rate to prevent overfitting	default: 0.2; range: {'min': 0.0, 'max': 0.5}; suggested: [0.1, 0.2, 0.3]
model_architecture.output_activation	Output layer activation function	choices: ['none', 'sigmoid', 'relu']; default: none
loss_configuration.loss_function		choices: ['mse', 'mae', 'huber']; default: mse; recommendations: {'noisy_data': 'huber', 'clean_data': 'mse'}
loss_configuration.huber_delta	The delta parameter of Huber loss	default: 1.0; range: {'min': 0.1, 'max': 2.0}
pseudo_pair_settings.use_pseudo_pairs	Whether to enable pseudo pair generation (core innovation)	choices: [True, False]; default: True; warning: The current code uses store_true and defaults to True, which cannot be disabled in the command line. It is recommended to modify it to: parser.add_argument('--use-pseudo-pairs', dest='use_pseudo_pairs', action='store_true'); parser.add_argument('--no-use-pseudo-pairs', dest='use_pseudo_pairs', action='store_false'); parser.set_defaults(use_pseudo_pairs=True)
pseudo_pair_settings.pseudo_weight	Pseudo-contrastive loss weight	default: 0.1; range: {'min': 0.0, 'max': 1.0}; suggested: [0.05, 0.1, 0.2]

Parameters	Description	Range/options
pseudo_pair_settings.alignment_weight	Feature alignment loss weight	default: 0.01; range: {'min': 0.0, 'max': 0.1}; suggested: [0.005, 0.01, 0.02]
pseudo_pair_settings.temperature	Comparative learning temperature parameters	default: 0.07; range: {'min': 0.03, 'max': 0.2}; suggested: [0.05, 0.07, 0.1, 0.15]
pseudo_pair_settings.learnable_temperature	Is the temperature parameter learnable?	choices: [True, False]; default: True
pseudo_pair_settings.use_projection_head	Whether to use a projection head	choices: [True, False]; default: True
pseudo_pair_settings.use_feature_alignment	Whether to use feature alignment	choices: [True, False]; default: True
pseudo_pair_settings.similarity_metric	Similarity measurement method	choices: ['cosine', 'euclidean', 'dot']; default: cosine
pseudo_pair_settings.hard_negative_mining	Whether to enable hard negative sample mining	choices: [True, False]; default: True
pseudo_pair_settings.momentum	Momentum coefficient of the momentum encoder	default: 0.999; range: {'min': 0.0, 'max': 0.999}; suggested: [0.0, 0.9, 0.99, 0.999]
pseudo_pair_settings.use_memory_bank	Whether to use memory bank to store negative samples	choices: [True, False]; default: False
pseudo_pair_settings.queue_size	memory size	choices: [1024, 2048, 4096, 8192]; default: 4096
pseudo_pair_settings.warmup_epochs	Warm up the number of rounds and gradually increase the pseudo-pair loss weight	default: 5; range: {'min': 0, 'max': 10}
early_stopping.patience	Early stopping patience value	default: 10; range: {'min': 5, 'max': 30}
early_stopping.early_stop_metric		choices: ['mse', 'rmse', 'mae', 'r2']; default: rmse
system.output_dir	Output directory	value: ./outputs
system.num_workers	Number of data loading processes	default: 4; range: {'min': 0, 'max': 16}; suggested: [0, 2, 4, 8]

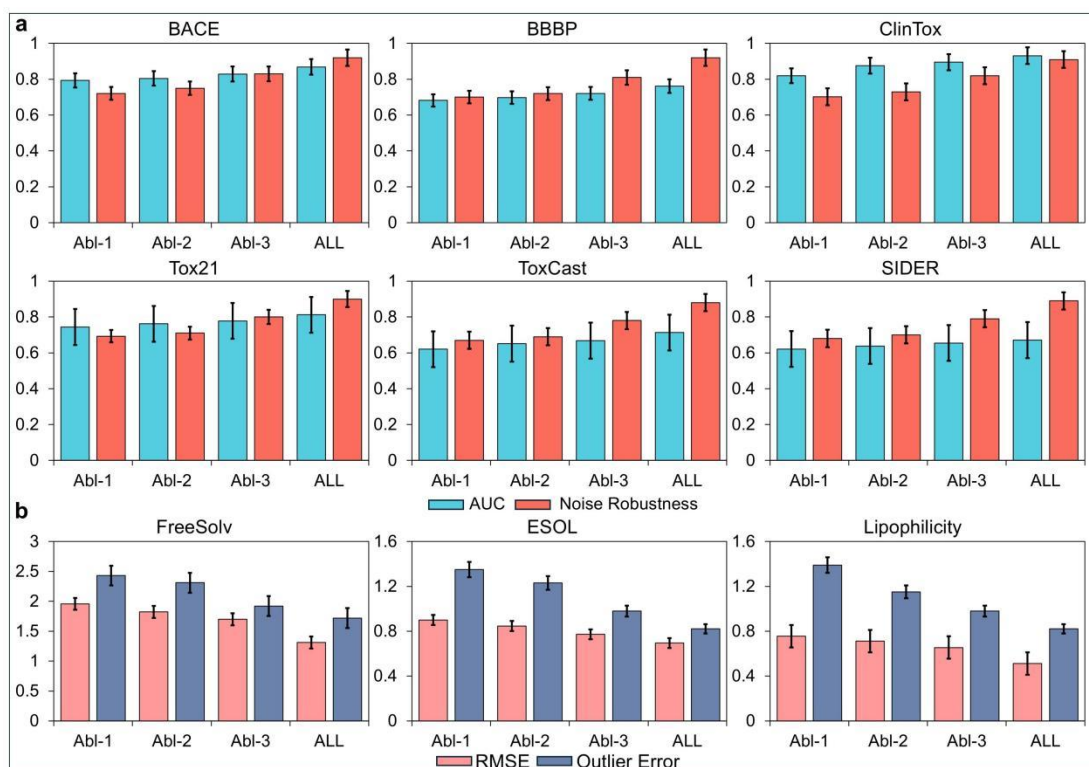
Supplementary Figures



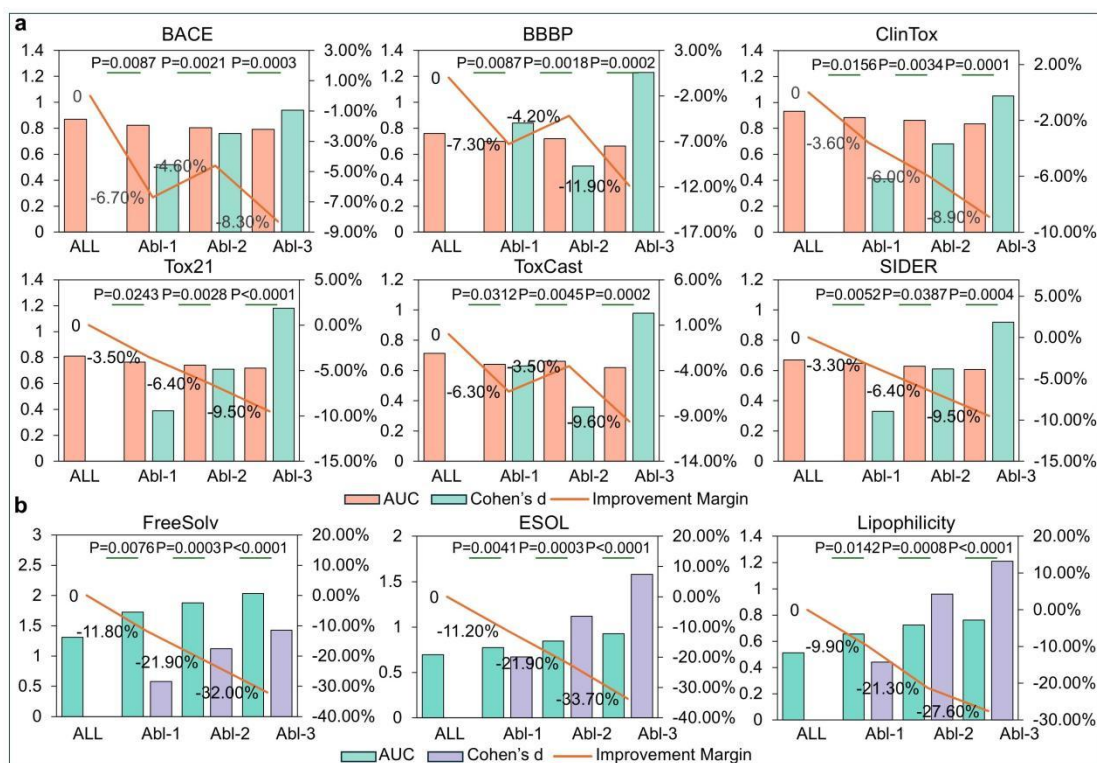
Supplementary Fig. 1 | Modality importance analysis on the BACE, ClinTox, ESOL, FreeSolv, Lipophilicity, SIDER, Tox21, and ToxCast datasets. The top 30 molecules were used as statistical units, and the average SHAP values of their features are reported. Feature prefixes are defined as follows: C denotes 1D SMILES, G denotes 2D graph structures, and 3D denotes three-dimensional geometry. The numbers on the left side of each subfigure represent the indices of the corresponding molecules in the dataset.



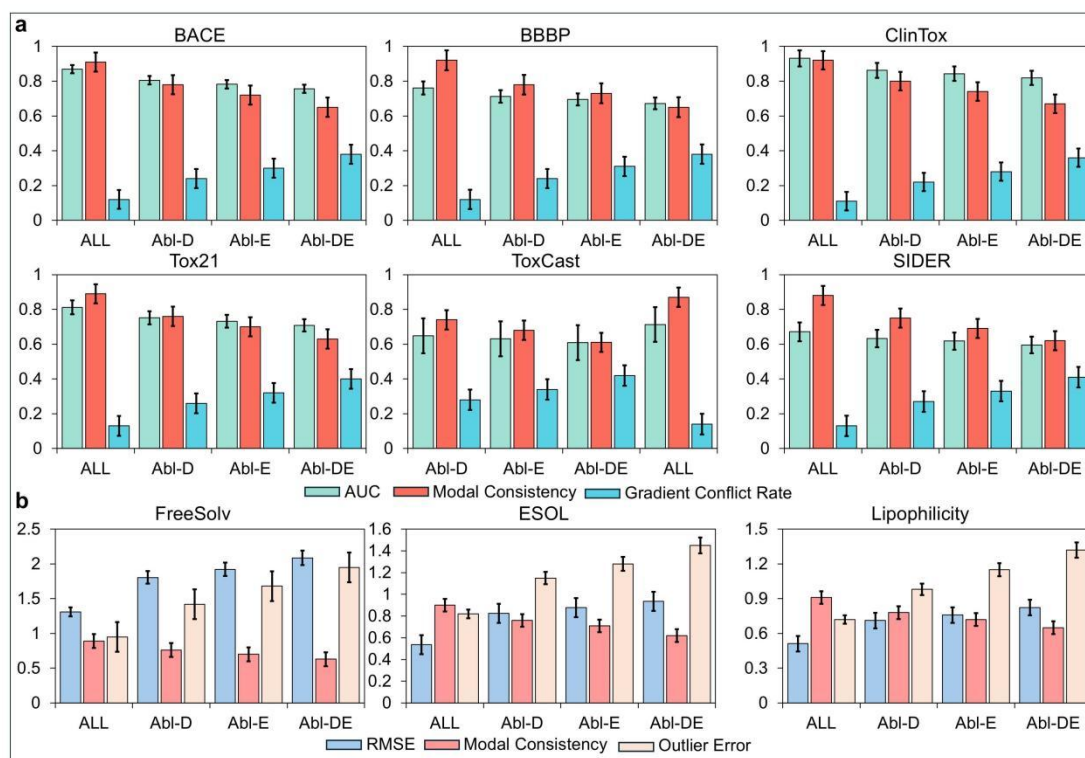
Supplementary Fig. 2 | **a**, Overall contribution proportions of molecular graphs, SMILES sequences, and 3D descriptors across the BACE, ClinTox, ESOL, FreeSolv, Lipophilicity, SIDER, Tox21, and ToxCast datasets. **b**, Interpretability of SemMol in molecular intervention experiments. For each dataset, the top 20 molecules and their top 50 features are visualized with SHAP heatmaps (red/blue = positive/negative contribution, darker/lighter = stronger/weaker intensity), covering the same eight datasets.



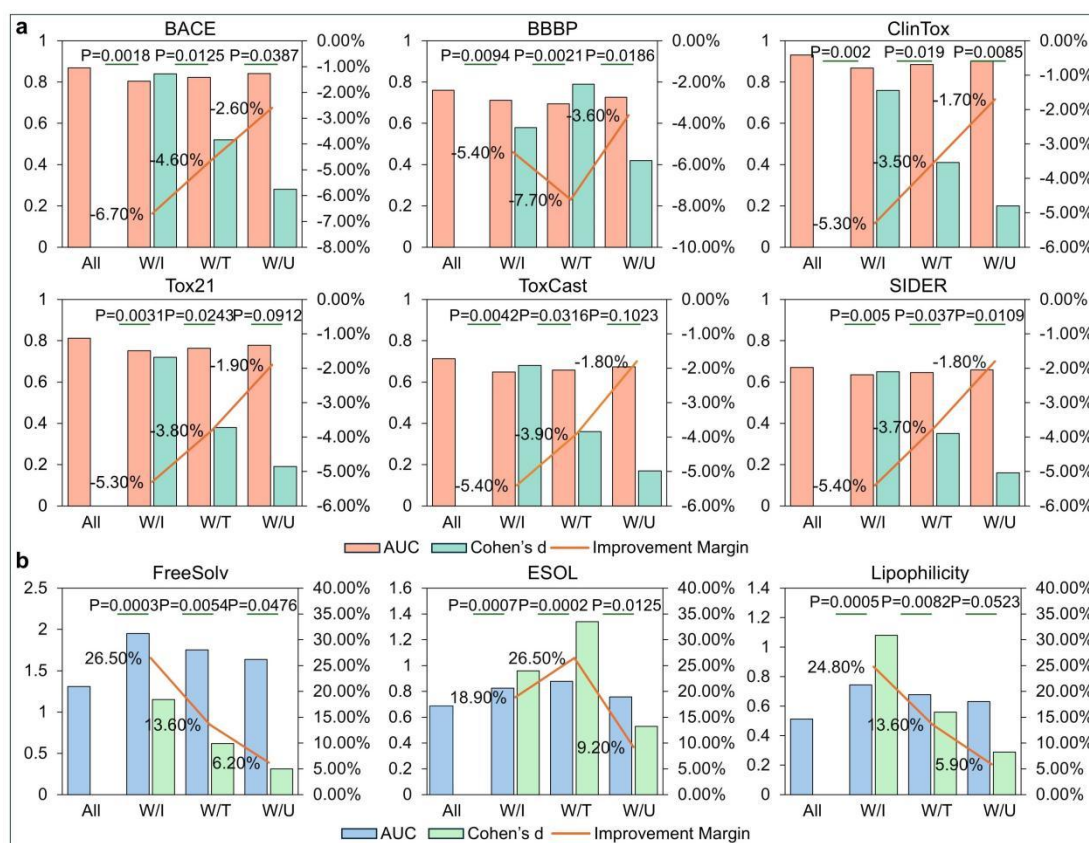
Supplementary Fig. 3 | Ablation study of the Dynamic Central Library (DCL) within the SemMol framework, with statistical significance tests validating its effectiveness. Performance comparison of the full model (ALL), Abl-1, Abl-2, and Abl-3 across **a** six MoleculeNet classification and **b** three regression datasets, respectively. “Abl-1” removes the entire DCL module and replaces it with random centers. “Abl-2” removes the periodic update mechanism of dynamic centers, fixing initial clustering centers. “Abl-3” weakens the update mechanism by setting the EMA momentum coefficient β in Eq. 4 of the manuscript to a value close to 1 (e.g., 0.999). The first six datasets are classification tasks, and the last three are regression tasks. Error bars indicate standard deviation across 10 independent runs.



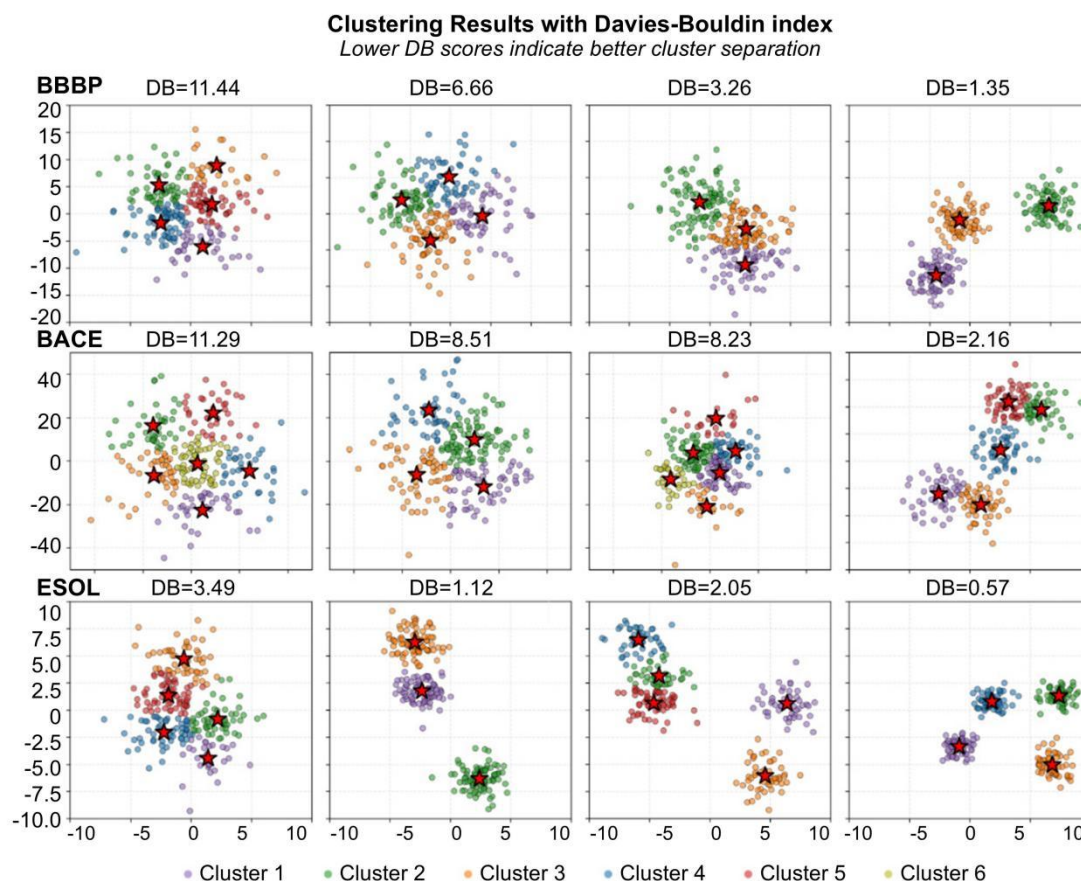
Supplementary Fig. 4 | Ablation study of the Dynamic Central Library (DCL) within the SemMol framework, with statistical significance tests validating its effectiveness. Statistical significance analysis comparing ALL, Abl-1, Abl-2, and Abl-3 across **a** six MoleculeNet classification and **b** three regression datasets, respectively. “Abl-1” removes the entire DCL module and replaces it with random centers. “Abl-2” removes the periodic update mechanism of dynamic centers, fixing initial clustering centers. “Abl-3” weakens the update mechanism by setting the EMA momentum coefficient β in Eq. 4 of the manuscript to a value close to 1 (e.g., 0.999). The first six datasets are classification tasks, and the last three are regression tasks. Error bars indicate standard deviation across 10 independent runs.



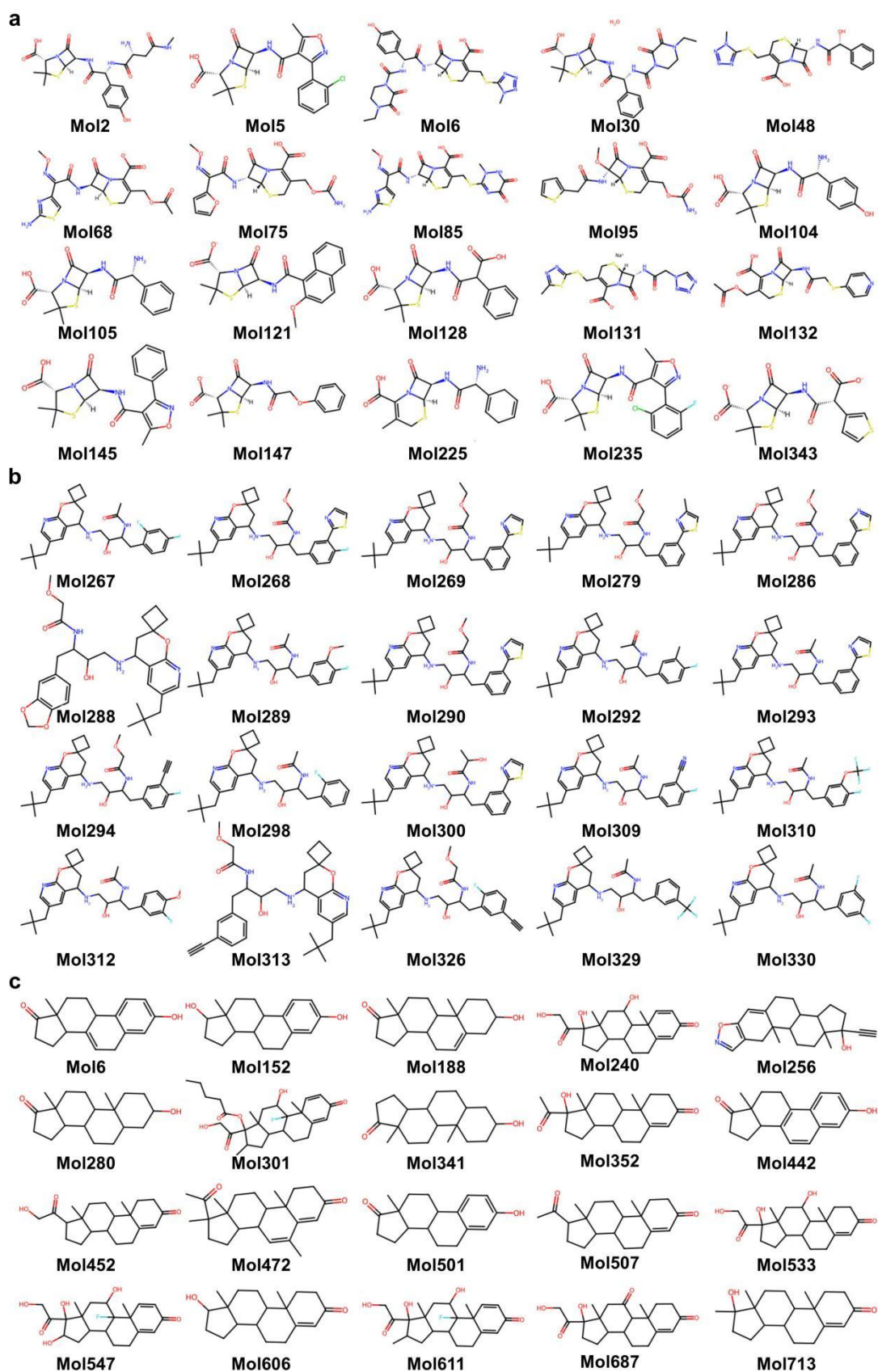
Supplementary Fig. 5 | Ablation study on the effectiveness of the ACSM mechanism in the SemMol framework. Synergistic effects of ACSM and unified spatial projection. Performance comparison of the full model (ALL), “Abl-D,” “Abl-E,” and “Abl-DE” across **a** six MoleculeNet classification and **b** three regression datasets, respectively. “Abl-D” removes the unified spatial projection within DCL component; “Abl-E” removes the ACSM component (**Fig. 2b of the manuscript**); “Abl-DE” removes both and falls back to instance-level contrastive learning (**Figs. 2b and 2c of the manuscript**). Bars show mean performance across 10 independent runs, with error bars indicating standard deviation.



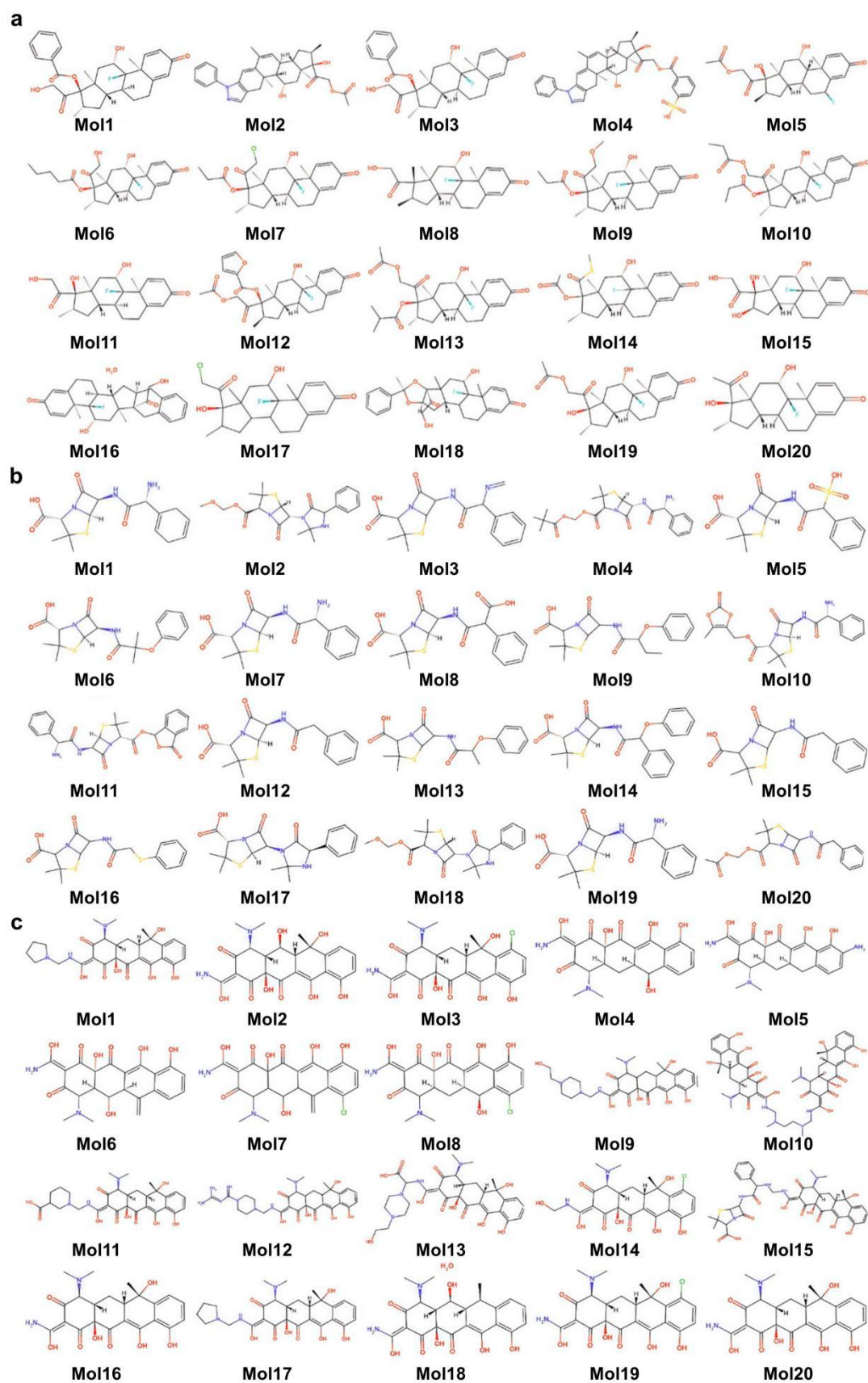
Supplementary Fig. 6 | Ablation study on the effectiveness of the soft matching strategies in ACSM mechanism. Effectiveness of positive sample generation strategies in the soft matching mechanism across **a** six MoleculeNet classification and **b** three regression datasets, respectively.. Under unified spatial projection, performance of “ALL,” “W/U,” “W/T,” and “W/I” is compared via significance testing. “ALL” (full model) uses similarity-weighted fusion of retrieved centers; “W/U” applies uniform weighting; “W/T” retrieves only the Top-1 center; “W/I” applies traditional instance-level contrastive learning. Bilateral paired t-tests assess performance differences, with Cohen’s d measuring effect size. The significance level is $\alpha=0.05$, where $d \approx 0.2$, 0.5, and 0.8 indicate weak, moderate, and strong effects, respectively.



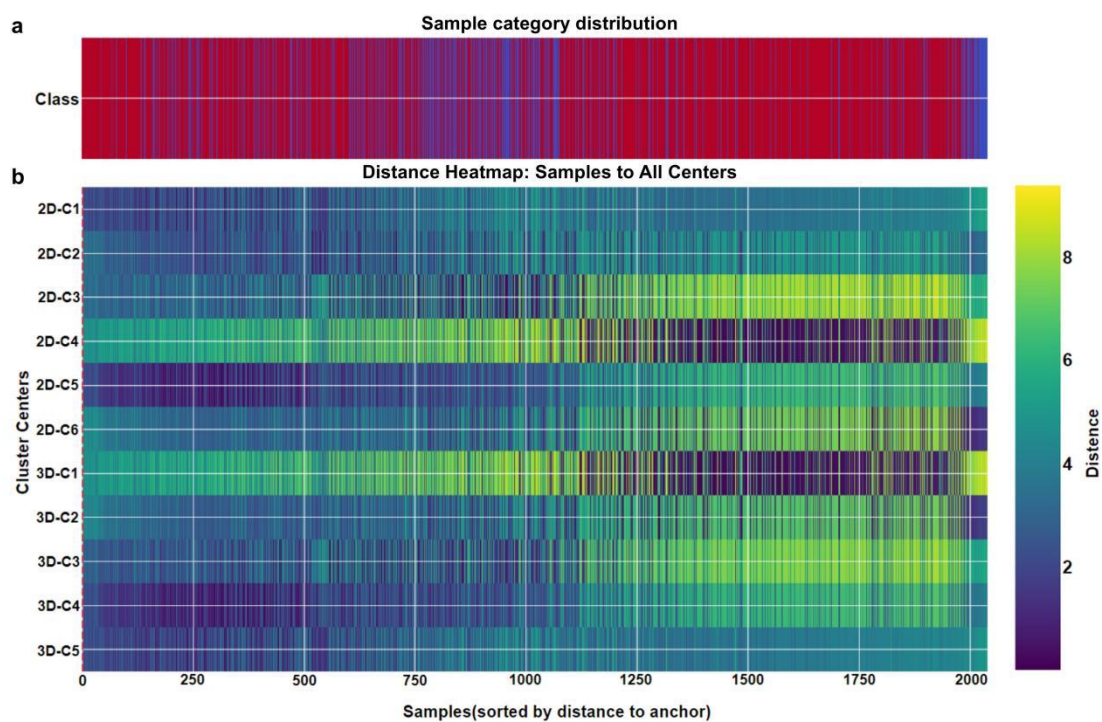
Supplementary Fig. 7 | Clustering analysis of molecular representations learned by the model on the BBBP, BACE, and ESOL datasets. Representations were projected into two dimensions using t-SNE for visualization, followed by clustering to evaluate distributional patterns across molecules.



Supplementary Fig. 8 | **a**, Visualization of molecules in Cluster 3 of the BBBP dataset; **b**, Visualization of molecules in Cluster 1 of the BACE dataset; **c**, Visualization of molecules in Cluster 3 of the ESOL dataset.



Supplementary Fig. 9 | Three collections **a**, **b** and **c** of representative molecular structures corresponding to clusters **b**, **c**, and **d** in **Fig. 5** of the manuscript on the BBBP dataset.



Supplementary Fig. 10 | For all molecules in the BBBP dataset shown in Figure 6 of the manuscript: **a**, distribution of classes; and **b**, heatmap of their distances to the cluster centers. Red indicates positive samples, blue indicates negative samples, and darker colors correspond to shorter distances.