MoDANet: A Motif-Aware Distance-Masked Attention Network for Molecular Property Prediction

Anonymous authors
Paper under double-blind review

Abstract

Reliable molecular property prediction is essential for drug discovery, yet standard graph Transformers use unrestricted global attention without explicitly organizing atom interactions by topological range. We introduce Motif-Aware Distance-Masked Attention Network (MoDANet), a dual-branch multi-scale graph Transformer whose global branch constructs fixed, nested, and permutation-equivariant neighborhoods from shortest-path distances and learns a layer-wise convex mixture over local, intermediate, and global attention scopes. Its motif branch directly encodes functional-group fragments obtained through BRICS decomposition. MoDANet provides an inspectable, distance-structured attention mechanism, is evaluated under scaffold splits on seven MoleculeNet benchmarks, and achieves state-of-the-art transfer performance after pretraining on a 250K-molecule ZINC subset.

1 Introduction

Molecular property prediction is central to drug discovery and materials design. As molecules are naturally represented as two-dimensional graphs with atoms as nodes and bonds as edges, graph neural networks and graph transformers have become the dominant tools (Gilmer et al., 2017; Rong et al., 2020; Rampášek et al., 2022), learning representations that encode topological and chemical semantics for predicting toxicity, solubility, bioactivity, and other properties from structure alone.

Accurate ground-truth labels, however, are expensive to obtain through wet-lab experiments, which has motivated adapting the pretrain-then-finetune paradigm from natural language processing and computer vision (Devlin et al., 2019; He et al., 2022) to molecular learning: pretrain on large unlabeled molecular collections, then finetune on scarce labeled data. Success on benchmarks such as MoleculeNet (Wu et al., 2018) hinges on whether the pretrained representation captures chemically meaningful inductive biases.

Most molecular self-supervised methods are purely data-driven, reconstructing masked atoms or bonds (Hu et al., 2020), predicting topological properties (Rong et al., 2020), or contrasting augmented views (You et al., 2020; Sun et al., 2022a), without explicitly modeling the structural hierarchy that governs molecular behavior; moreover, since adding or removing a single bond can change a molecule's identity, graph augmentation can catastrophically disrupt chemical topology and corrupt the learned signals (Trivedi et al., 2022; Sun et al., 2022b).

Beyond the pretraining objective, the architecture itself faces two challenges (Figure 1). First, graph transformers apply global self-attention over all atom pairs (Rong et al., 2020; Ying et al., 2021; Rampášek et al., 2022), treating distant atoms as equally relevant as bonded ones and diluting local chemical signals, while existing distance-biased or sparse attention variants (Ying et al., 2021; Shirzad et al., 2023; Zaheer et al., 2020) rely on fixed rules and cannot learn which topological scopes are most useful at each layer. Second, atom-level models do not explicitly represent the substructure hierarchy of molecules: motif-based methods introduce it either through closed fragment vocabularies or rule-based decomposition (Jin et al., 2018; Zhang et al., 2021; Degen et al., 2008; Lewell et al., 1998), but the former cannot represent fragments outside their lookup sets and the latter impose fixed cleavage priors.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

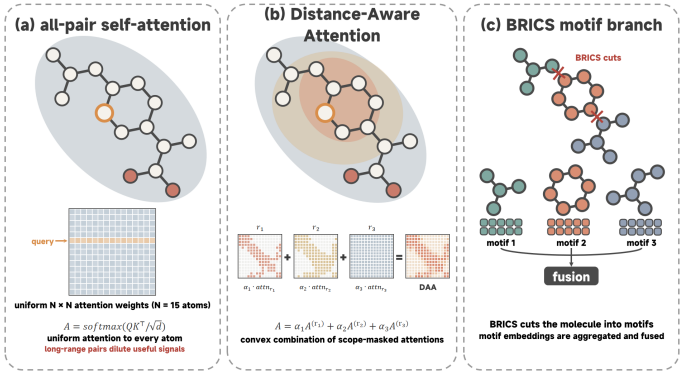


Figure 1: Two limitations of existing molecular graph transformers addressed by MoDANet. (a) All-pair self-attention treats distant atom pairs as equally relevant as bonded ones, diluting local chemical signals. (b) Our distance-masked attention decomposes each query's attention into nested, shortest-path-ranked neighborhoods defined by per-node distance quantiles, with coverage ratios selected on the validation set, and learns a per-layer convex mixture α over them. (c) BRICS decomposition breaks a molecule into retrosynthetically motivated functional-group fragments (colored subgraphs; red crosses mark the broken bonds), which are encoded directly from their atom-level graphs and fused with the global representation through hierarchical motif fusion.

To address these limitations, we propose MoDANet (Motif-Aware Distance-Masked Attention Network), a dual-branch graph transformer that learns controlled attention scopes and directly encoded fragment representations. Its global branch constructs coverage-based, permutation-equivariant masks from nested shortest-path neighborhoods, with the coverage ratios selected on the validation set, and learns a layer-wise convex mixture over local, intermediate, and global attention scopes; its motif branch applies the fixed BRICS cleavage rules and encodes each resulting functional-group fragment directly from its atom-level graph, without a closed fragment vocabulary. The two views are fused hierarchically before the prediction head, where a global-context-guided attention dynamically decides how much motif-level information to incorporate for each molecule. The model is pretrained on a 250K-molecule ZINC subset by regressing an automatically computed molecular property target, then finetuned on downstream tasks. Our contributions are:

- a distance-masked attention mechanism that constructs coverage-based, permutation-equivariant local, intermediate, and global scopes and learns a layer-wise convex mixture over them;
- a BRICS-based motif branch that encodes functional-group fragments directly from atom-level graphs and fuses them with the global representation, without a closed fragment vocabulary;
- state-of-the-art transfer performance on seven MoleculeNet benchmarks under scaffold splits after pretraining on a 250K-molecule ZINC subset.

2 Related Work

Molecular Representation Learning. Early pipelines relied on fixed fingerprints or expert-crafted descriptors fed into shallow models (Rogers & Hahn, 2010). Graph neural networks then became dominant on molecular graphs (Gilmer et al., 2017), with richer message and aggregation functions improving expressiveness (Dwivedi et al., 2023; Corso et al., 2020). Graph transformers (Dwivedi & Bresson, 2021; Rong et al., 2020; Kreuzer et al., 2021) combine local message passing with global self-attention, but their global attention is often indiscriminate and may obscure chemically relevant local patterns.

Self-Supervised Learning on Molecular Graphs. Molecular self-supervised pretraining is largely data-driven in nature: generative methods reconstruct masked atoms or bonds (Hu et al.,

2020; Rong et al., 2020),contrastive methods align augmented views (Hassani & Khasahmadi, 2020; You et al., 2020; Sun et al., 2022a), and predictive methods regress computed descriptors (Ramsundar et al., 2017). These methods rarely encode chemical domain knowledge and rely on augmentations that can alter molecular identity (Trivedi et al., 2022; Sun et al., 2022b); 3D denoising pretraining achieves strong transfer (Zaidi et al., 2023; Liu et al., 2022) but requires costly 3D conformers. We instead pretrain on 2D molecular graphs from a 250K-molecule ZINC subset (Sterling & Irwin, 2015) and inject chemical hierarchy through a motif branch.

Motif-Based Molecular Models. Motif-based methods represent recurring substructures either through predefined fragment vocabularies (Jin et al., 2018; Zhang et al., 2021) or through fixed cleavage rules such as BRICS and RECAP (Degen et al., 2008; Lewell et al., 1998). We adopt the fixed BRICS rules but encode every fragment directly from its atom-level graph, so previously unseen fragments can be represented without a closed vocabulary.

Attention Mechanisms for Graphs. Distance-biased attention injects graph distance into attention scores (Ying et al., 2021), sparse expander-based attention restricts each query to a few nodes (Shirzad et al., 2023), and block-sparse or hashing-based methods apply predefined patterns (Zaheer et al., 2020; Kitaev et al., 2020). These methods apply fixed rules rather than learning which neighborhoods to attend to at each layer; our distance-masked attention instead learns a layer-specific convex mixture over coverage-quantile scopes, yielding inspectable layer-level preferences.

3 METHODOLOGY

MoDANet is a dual-branch graph transformer for molecular property prediction (Figure 2): a global encoder with VDAA learns atom-level representations, a motif encoder encodes BRICS fragments from their atom-level graphs, and the two views are integrated by hierarchical motif fusion before the prediction head.

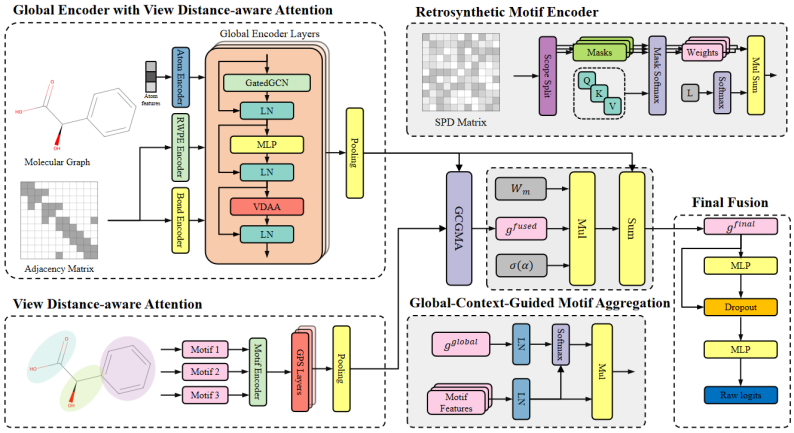


Figure 2: MoDANet architecture. The global branch applies View-Distance-Aware Attention over the full atom graph. The motif branch encodes BRICS fragments with a separate GPS encoder. The two representations are fused via hierarchical motif fusion before the prediction head.

3.1 PRELIMINARIES

We first introduce preliminary information on molecular graph representation learning and the building blocks underlying MoDANet, alongside the notation used in later sections.

Molecular graphs. A molecular graph is represented as an attributed graph $G = (V, E, X)$, where $V = \{v_1, \ldots, v_N\}$ is the set of $N = |V|$ atoms, $E \subseteq V \times V$ is the set of chemical bonds, and

$X \in \mathbb{R}^{N \times d_x}$ is the atom feature matrix whose i-th row $x_i \in \mathbb{R}^{d_x}$ encodes atomic properties of atom i (e.g., element type, degree, formal charge, hybridization). The topological structure is captured by the adjacency matrix $A_{\text{adj}} \in \{0,1\}^{N \times N}$ with $[A_{\text{adj}}]_{ij} = 1$ iff $(i,j) \in E$, and the immediate neighbors of atom i are denoted by $\mathcal{N}(i) = \{j \in V \mid (i,j) \in E\}$. Each bond $(i,j) \in E$ has an edge feature $e_{ij} \in \mathbb{R}^{d_e}$ encoding bond type and stereochemistry. Given the graph G, a graph encoder f typically returns a node representation matrix $Z = f(G) \in \mathbb{R}^{N \times d}$. We write $h_i^{(l)}$ for the representation of atom i after the l-th encoder layer, with $h_i^{(0)}$ produced by the input encoding. In molecular property prediction, a single graph-level representation is required, which is typically obtained by pooling all node representations (Hamilton, 2020). The learning target is a task-specific label y.

Graph Transformers. MoDANet adopts the GraphGPS layer template (Rampášek et al., 2022), where each layer integrates a local message-passing neural network (MPNN) (Gilmer et al., 2017) defined over the molecular bond graph E with global self-attention applied to all atom pairs. Given the atom representations $Z^{(i)}$ at the i-th layer, the local MPNN first updates the representations according to

$$\widetilde{Z}^{(i+1)} = \mathrm{LN}\left(Z^{(i)} + \mathrm{MPNN}\left(Z^{(i)}, E\right)\right). \tag{1}$$

The resulting intermediate representations are subsequently refined through global self-attention:

$$Z^{(i+1)} = \mathrm{LN}\left(\widetilde{Z}^{(i+1)} + \mathrm{Attn}\left(\widetilde{Z}^{(i+1)}\right)\right). \tag{2}$$

A position-wise feed-forward network is applied afterward but is omitted here for brevity. The local MPNN component preserves the inductive bias associated with molecular bond connectivity, whereas the global self-attention component captures long-range interactions between atoms that cannot be efficiently propagated through only a few message-passing layers.

BRICS decomposition. We use the RDKit implementation of BRICS (Degen et al., 2008) to identify retrosynthetically breakable bonds. Removing these bonds partitions the molecule into connected components $\{M_1, \ldots, M_K\}$, which we treat as functional-group fragments; if no bond is identified, the complete molecule is retained as a single fragment.

3.2 Input Encoding

Atom and bond attributes are represented by discrete categorical features. We encode each atom with an embedding table, or a sum of per-feature embedding tables when multiple atom features are used, as $h_i^{(0)} = \mathrm{AtomEncoder}(x_i) \in \mathbb{R}^{d-p}$, where d is the model hidden size and p is the positional-encoding dimension. Bond features are embedded into the model dimension as $e'_{ij} = \mathrm{BondEncoder}(e_{ij}) = \sum_{f=1}^{F_e} \mathrm{Emb}_f(e_{ij,f}) \in \mathbb{R}^d$. We use $F_e = 3$ categorical features (bond type, stereochemistry, and conjugation) for the downstream MoleculeNet datasets, following the OGB featurization scheme, and $F_e = 1$ (four bond types) for ZINC pretraining. The encoded bonds are supplied to the first message-passing layer and updated by GatedGCN (Dwivedi et al., 2023) at every subsequent layer. To inject structural inductive bias, we append random-walk positional encodings (RWPE), computed from the diagonal entries of powers of the row-normalized adjacency matrix A_{adj}, to the atom embeddings, followed by layer normalization: $h_i^{(0)} \leftarrow \mathrm{LN}\left(\left[h_i^{(0)} \,\|\, \mathrm{RWPE}_i\right]\right)$.

3.3 Global Encoder with VDAA

The global branch is a stack of L GPS layers equipped with View-Distance-Aware Attention (VDAA). Each layer first applies a local message-passing network (GatedGCN by default), then performs VDAA over all atom pairs, and finally applies a feed-forward network, with residual connections and layer normalization. While the local module respects the bond graph, standard global self-attention exposes every query to all nodes and does not explicitly distinguish whether a key atom is two bonds away or ten bonds away within its attention support. We therefore organize candidate keys into explicit, coverage-based topological attention scopes derived from shortest-path distances.

Distance-Aware Attention Scopes. For each graph, we compute the all-pairs shortest-path distance matrix $\Delta \in \mathbb{N}^{N\times N}$ over the covalent-bond graph once before training and cache it with the graph data. MoDANet uses three distance-thresholded attention scopes ($R = 3$), where the thresholds are defined *per node* as quantiles of the distance distribution rather than as fixed integer hop counts. Concretely, each scope r is characterized by a ratio $\rho_r \in (0, 1]$; for a query node i, the attention scope contains the nearest $\lceil \rho_r N \rceil$ nodes (ties at the cutoff distance are all included):

$$V_i^{(r)} = \{j \in V \mid \Delta_{ij} \le \tau_i(\rho_r)\}, \tag{3}$$

where $\tau_i(\rho_r)$ is the k-th smallest value of $\{\Delta_{ij}\}_{j=1}^{N}$ with $k = \lceil \rho_r N \rceil$. The corresponding binary mask is

$$M_{ij}^{(r)} = \mathbb{I}\Big[j \in V_i^{(r)}\Big] = \mathbb{I}[\Delta_{ij} \le \tau_i(\rho_r)], \tag{4}$$

and the last scope is global ($\rho_R = 1$). Because $\rho_1 < \rho_2$, the scopes are nested, and because each mask depends only on the pairwise distances Δ_{ij}, the construction is invariant to the ordering of nodes. For pretraining, the ratios are searched once on the ZINC validation split; for finetuning, they are re-searched per dataset. The ratios $(\rho_1, \rho_2, 1.0)$ are selected by a genetic-algorithm search over percentile pairs on the validation split of each downstream dataset and are then fixed; the same set of scopes is shared by all layers. The scopes remain fixed throughout; among the VDAA components, only the per-layer mixture weights $\alpha^{(l)}$ in Eq. equation 7, obtained by a softmax over learnable logits, are learned by gradient descent.

VDAA. Given queries Q, keys K, and values V, the attention logits at layer l are

$$S^{(l)} = \frac{QK^\top}{\sqrt{d_h}}, \tag{5}$$

where d_h is the dimension of each attention head. For clarity, we omit the head index h from Q, K, V, $S^{(l)}$, and $A^{(l,r)}$; the masked softmax is computed independently for each head, while $\alpha^{(l)}$ is shared across heads. For each scope $r \in \{1, \dots, R\}$, we apply the corresponding mask:

$$A_{ij}^{(l,r)} = \frac{M_{ij}^{(r)} \exp\Big(S_{ij}^{(l)}\Big)}{\sum_{j'} M_{ij'}^{(r)} \exp\Big(S_{ij'}^{(l)}\Big)}, \tag{6}$$

where masked positions ($M_{ij}^{(r)} = 0$) receive exactly zero attention weight; equivalently, this is implemented as an additive $-\infty$ mask applied to the logits before the softmax. Each layer learns a convex mixture of the masked attention matrices:

$$\alpha^{(l)} = \mathrm{softmax}\left(\ell^{(l)}\right), \qquad A^{(l)} = \sum_{r=1}^{R} \alpha_r^{(l)} A^{(l,r)}. \tag{7}$$

The mixture logits are initialized to zero, yielding a uniform mixture at initialization. The attention output is $\mathrm{Attn}(Q, K, V) = A^{(l)}V$. By learning layer-specific mixture weights, the model can emphasize different topological scopes at different depths. The weights $\alpha^{(l)}$ represent a layer-level preference over the predefined scopes. Because the attention scopes are nested and each masked attention matrix is normalized separately, $\alpha_r^{(l)}$ should not be interpreted as the fraction of attention mass assigned to a particular hop-distance range. In our implementation, one set of mixture weights is learned per layer and shared across attention heads and input molecules; the weights are updated during both pretraining and finetuning and are therefore layer-specific rather than input-adaptive. Because the global scope is retained, VDAA reweights rather than eliminates long-range interactions, allowing the model to preserve potentially useful nonlocal dependencies.

3.4 Retrosynthetic Motif Encoder

Functional-group fragments provide a complementary view to atom-level molecular representations. We therefore introduce a Retrosynthetic Motif Encoder (RME) based on the fixed BRICS decomposition rules. RME does not learn fragment boundaries; instead, it directly encodes every BRICS-derived fragment from its atom-level graph rather than retrieving an embedding from a closed fragment vocabulary.

For a molecule with fragments $\{M_1, \ldots, M_K\}$, each fragment M_k is the induced subgraph over its atoms. We encode all fragments of a batch jointly as one large disconnected graph with a second GPS encoder that has independent parameters, uses half the number of layers of the global encoder ($L/2$), and applies standard full self-attention without distance masking, since fragments are small and all pairwise interactions within a fragment are potentially relevant. The fragment encoder shares the atom/bond input encoding described above; during pretraining we omit positional encodings on fragment atoms, and motif RWPE can be optionally enabled for downstream tasks. Each fragment is pooled into a vector $g_k \in \mathbb{R}^d$, and the fragments of one molecule are packed into a matrix

$$G_{\text{motif}} = [g_1, \ldots, g_K]^\top \in \mathbb{R}^{K \times d}. \tag{8}$$

A padding mask records which positions are valid for each graph.

3.5 Hierarchical Motif Fusion

The global branch provides a graph-level vector $g_{\text{global}} \in \mathbb{R}^d$ by sum-pooling atom representations. We use this global context to guide motif aggregation through Global-Context-Guided Motif Aggregation (GCMA), which forms the core aggregation operation in Hierarchical Motif Fusion (HMF).

Global-Context-Guided Motif Aggregation. GCMA uses the global graph representation as the aggregation query and all motif representations as keys and values. For compactness, let $\widehat{g} = \mathrm{LN}(g_{\text{global}})$ and $\widehat{G} = \mathrm{LN}(G_{\text{motif}})$. Then

$$g_{\text{fused}} = \mathrm{Attn}\left(\widehat{g}, \widehat{G}, \widehat{G}\right). \tag{9}$$

For batched computation, padded motif positions are excluded from the attention softmax using a standard key-padding mask; this implementation detail is omitted from the single-molecule equation above. For molecules with multiple motifs, the attention weights measure the relative relevance of each motif conditioned on the global molecular representation. If no valid motif is available, g_{fused} falls back to g_{global}.

Final fusion. The global and fused motif representations are combined before the prediction head. During pretraining, when the fusion modules are still randomly initialized, we use a gated residual connection in which a learnable logit η initialized to -2 produces a gating factor $\sigma(\eta) \approx 0.12$ on a projected motif residual, so that the motif contribution is initially weak and becomes stronger only when the training signal supports it:

$$g_{\text{final}} = \mathrm{LN}\left(g_{\text{global}} + \sigma(\eta) \cdot W_m\, g_{\text{fused}}\right). \tag{10}$$

The gated form acts as a scaffold that protects the randomly initialized fusion modules from being dominated by an uncalibrated motif pathway early in pretraining. For finetuning, the motif pathway is already pretrained and calibrated, and we use the plain residual sum $g_{\text{final}} = \mathrm{LN}(g_{\text{global}} + g_{\text{fused}})$. Learned convex gates, concatenation, or dropping the motif branch are available as alternative fusion strategies.

3.6 Pretraining and Finetuning

Pretraining. The model is pretrained on a 250K-molecule ZINC subset with a predictive objective and, optionally, a motif-aware auxiliary objective.

Property regression (main objective). We minimize the mean-squared error between the model output and the ZINC molecular property target, the penalized logP score $y_{\text{raw}} = \mathrm{logP} - \mathrm{SA} - \mathrm{ring\ penalty}$, computed deterministically from the 2D structure and standardized using the training-split statistics, $y = (y_{\text{raw}} - \mu_{\text{train}})/\sigma_{\text{train}}$:

$$\mathcal{L}_{\text{ZINC}} = \frac{1}{B} \sum_{i=1}^{B} (\widehat{y}_i - y_i)^2. \tag{11}$$

We use this predictive objective because it provides an automatically computed, downstream-label-free supervision signal without requiring molecular graph augmentation.

BRICS bond prediction (auxiliary objective, ablated). To inject motif awareness directly into pretraining, we also explore an auxiliary objective that predicts whether a bond is BRICS-breakable. For each bond $(u, v) \in E$, a scorer over the global-branch atom representations $[h_u \,\|\, h_v \,\|\, h_u \odot h_v]$ predicts a breakability logit, trained with binary cross-entropy, with the positive class reweighted by the negative-to-positive ratio $n_{\text{neg}}/n_{\text{pos}}$:

$$\mathcal{L}_{\text{BRICS}} = -\frac{1}{|E|} \sum_{(u,v)\in E} \Big[b_{uv} \log \widehat{b}_{uv} + (1 - b_{uv}) \log(1 - \widehat{b}_{uv}) \Big], \tag{12}$$

where $b_{uv} \in \{0, 1\}$ indicates whether bond (u, v) is broken by BRICS. The total pretraining loss is $\mathcal{L} = \mathcal{L}_{\text{ZINC}} + \lambda \mathcal{L}_{\text{BRICS}}$. In our experiments, however, this auxiliary term consistently degrades pretraining quality, so the reported model sets $\lambda = 0$, i.e., plain property regression.

We use AdamW with cosine annealing, and all modules are trained jointly.

Finetuning. For each downstream MoleculeNet dataset, the task-specific prediction head (a two-layer MLP) is reinitialized while all other parameters are loaded from the pretrained checkpoint, and the full model is then fine-tuned end-to-end with AdamW under a single learning rate. An optional warmup schedule that freezes the encoders for the first epochs is supported but not used in the reported experiments. Classification tasks use binary cross-entropy (with optional per-task positive weights or asymmetric loss for imbalanced labels), while regression tasks use mean-squared error. Motif RWPE, which is disabled during pretraining, is enabled at this stage. Models are evaluated under a scaffold split (80/10/10) to test out-of-distribution generalization. Transfer results are reported in Table 1.

4 Experiments

We evaluate MoDANet on the MoleculeNet benchmark (Wu et al., 2018) through comprehensive finetuning experiments, ablation studies, and analysis.

4.1 Experimental Setup

Pretraining. We pretrain MoDANet on a 250K-molecule ZINC subset by regressing the penalized-logP-style molecular property target with mean-squared error. The model uses a hidden dimension $d = 64$, $L = 4$ global GPS layers with $H = 4$ attention heads, an $L/2 = 2$-layer motif encoder, RWPE of dimension $p = 20$, and coverage ratios selected by the genetic-algorithm search on the ZINC validation split. We use AdamW with learning rate 5×10^{-4}, weight decay 10^{-4}, batch size 128, and cosine decay for up to 200 epochs, with early stopping (patience 30) on the validation MAE.

Finetuning. We finetune on MoleculeNet classification benchmarks (BBBP, BACE, ClinTox, HIV, SIDER, Tox21, MUV) under the scaffold split (80/10/10), which places structurally distinct Bemis–Murcko scaffolds in different folds and thus measures out-of-distribution generalization. The prediction head is reinitialized per dataset while all other parameters are loaded from the pretrained checkpoint, and the full model is fine-tuned end-to-end. For multi-task datasets (ClinTox, SIDER, Tox21, MUV) we report the average ROC-AUC over the labeled tasks; for single-task datasets (BBBP, BACE, HIV) we report ROC-AUC.

Baselines. We compare against GNN baselines (GIN (Xu et al., 2019), GPS (Rampášek et al., 2022)), self-supervised molecular pretraining methods (GraphLoG (Xu et al., 2021), GraphCL (You et al., 2020), SimGRACE (Xia et al., 2022), D-SLA (Kim et al., 2022), MolCLR (Wang et al., 2022), GraphMVP (Liu et al., 2022), 3D Infomax (Stärk et al., 2022), MGSSL (Zhang et al., 2021), GROVER (Rong et al., 2020), KANO (Fang et al., 2022), KPGT (Li et al., 2022), GraphMAE (Hou et al., 2022), MoleculeSTM (Liu et al., 2023a)), and the recent SimSGT model (Liu et al., 2023b). We additionally evaluate regression performance on the MoleculeNet regression datasets; results are reported in the appendix (Table 6).

Table 1: Transfer learning ROC-AUC (%) scores on seven MoleculeNet datasets, averaged across three random seeds with one standard deviation. All tasks are evaluated by ROC-AUC (higher is better) under the scaffold split (80/10/10 train/validation/test). For multi-task datasets (ClinTox, SIDER, Tox21, MUV) we report the average ROC-AUC over the labeled tasks. All methods are evaluated under the same protocol. Best results in **bold**, second best underlined.

Method	BBBP	ClinTox	MUV	HIV	BACE	Tox21	SIDER	Avg.
GIN	65.8±4.5	58.0±4.4	71.8±2.5	75.3±1.9	70.1±5.4	74.0±0.8	57.3±1.6	67.5
GPS	64.8±3.0	87.2±0.9	69.8±3.8	73.1±3.3	78.0±3.0	74.5±0.6	60.8±0.6	72.6
GraphLoG	72.3±0.9	74.7±2.2	74.2±1.8	75.4±0.6	82.2±0.9	75.1±0.7	61.2±1.1	73.6
D-SLA	72.6±0.8	80.2±1.5	76.6±0.9	78.6±0.4	83.8±1.0	76.8±0.5	60.2±1.1	75.5
MolCLR	73.5±0.4	90.4±1.7	75.5±1.8	77.6±3.2	83.5±1.8	76.7±2.1	60.7±5.7	76.8
GraphMVP	72.4±1.6	79.1±2.8	77.7±0.6	77.0±1.2	81.2±0.9	75.9±0.5	63.9±1.2	75.3
GROVER	69.5±0.1	76.2±3.7	67.3±1.8	68.2±1.1	81.0±1.4	73.5±0.1	**65.4±0.1**	71.6
KANO	69.9±1.9	90.7±2.2	74.7±2.0	75.7±0.3	82.7±0.9	75.8±0.5	60.2±1.4	75.7
SimGRACE	72.3±0.6	65.4±1.5	71.4±2.9	74.1±0.7	76.0±1.0	73.1±0.3	64.2±0.2	70.9
MGSSL	73.9±0.9	76.2±1.7	74.6±2.7	74.8±1.6	82.6±0.9	73.2±0.4	59.5±0.4	73.5
SimSGT	71.6±0.6	84.9±1.6	**78.9±0.7**	77.5±1.0	83.5±0.1	76.7±0.2	62.8±0.6	76.6
GraphCL	72.6±0.5	74.7±0.2	71.4±1.3	75.9±1.3	80.3±1.1	73.0±0.2	61.4±0.2	72.8
KPGT	71.4±0.7	88.8±2.9	75.7±1.4	78.5±1.7	81.8±2.7	**78.5±0.5**	64.7±1.0	77.1
3DInfomax	69.3±0.7	60.7±2.9	74.8±1.6	77.9±0.2	81.2±0.6	73.6±0.5	55.3±7.7	70.4
GraphMAE	71.1±4.6	80.5±2.5	77.2±1.2	77.0±1.2	83.5±0.3	75.3±0.3	61.1±0.4	75.1
MoleculeSTM	71.0±1.0	91.6±0.1	75.2±0.4	78.1±0.9	80.4±4.2	74.4±0.3	60.1±0.8	75.8
MoDANet	**74.7±0.4**	**94.4±0.9**	77.0±0.9	**79.3±0.8**	**87.0±1.7**	77.6±0.1	64.8±0.2	**79.3**

4.2 MoleculeNet Finetuning Results

Table 1 reports the transfer learning results. MoDANet achieves an average ROC-AUC of 79.3% across the seven benchmarks, the highest among the compared methods, and obtains the best result on four datasets: BBBP (74.7%), ClinTox (94.4%), BACE (87.0%), and HIV (79.3%).

The strongest gains appear on BBBP, ClinTox, and BACE, where local functional groups and related molecular substructures may provide useful predictive signals. On ClinTox, MoDANet improves over the strongest baseline (MoleculeSTM, 91.6%) by 2.8 points,on BACE it improves over D-SLA (83.8%) by 3.2 points, and on BBBP it improves over MGSSL (73.9%) by 0.8 points. This observation is consistent with the use of BRICS-derived fragment representations.On HIV (79.3%), MoDANet improves over the strongest baseline (D-SLA, 78.6%) by 0.7 points. On SIDER (64.8%), MoDANet ranks second behind GROVER (65.4%). On Tox21 (77.6%), MoDANet ranks second behind KPGT (78.5%). On MUV (77.0%), MoDANet remains behind SimSGT (78.9%) and is comparable to GraphMVP (77.7%) and GraphMAE (77.2%) within one standard deviation; MUV spans 17 highly diverse bioactivity targets, where the relevant substructures may exceed what the 250K-molecule pretraining set covers.

MoDANet achieves state-of-the-art average transfer performance after pretraining on a 250K-molecule ZINC subset. Several baselines use larger or richer pretraining corpora; since they also differ in architecture and training objectives, controlled scaling experiments are needed to establish relative sample efficiency.

4.3 Ablation Studies

Table 2 examines the contribution of each major component on five datasets: BBBP, ClinTox, BACE, Tox21, and SIDER, which span single-task and multi-task benchmarks with different levels of substructure dependence. The variants remove, respectively, the distance-masked attention (w/o VDAA, replaced by standard full self-attention), the motif branch (w/o motif), GCMA (w/o GCMA, replaced by masked mean pooling of motif vectors), the shared input encoding between the two branches (w/o shared GPS), and motif RWPE (w/o motif RWPE). We also replace the GatedGCN local encoder

with GCN or GINE, use concatenation or the pretraining-time gated connection as the fusion strategy (w/ concat fusion, w/ gated fusion), and remove the pretrained weights entirely (w/o pretrain, trained from scratch).

Table 2: Ablation results (ROC-AUC %, mean ± std) of MoDANet variants on BBBP, ClinTox, BACE, Tox21, and SIDER.

Variant	BBBP	ClinTox	BACE	Tox21	SIDER
full	**74.7±0.4**	**94.4±0.9**	**87.0±1.7**	**77.6±0.1**	**64.8±0.2**
w/o VDAA	67.0±3.0	85.7±2.2	63.8±7.8	73.4±1.1	54.9±0.6
w/o motif	65.9±2.4	52.6±8.3	77.0±1.5	72.9±1.2	50.6±1.5
w/o GCMA	68.9±1.6	78.0±1.1	72.8±3.4	74.1±1.1	56.5±0.5
w/ concat fusion	68.0±2.1	83.7±0.6	75.0±2.8	74.6±0.4	56.8±1.0
w/ gated fusion	69.6±2.1	59.9±1.0	70.4±5.4	74.9±0.6	55.1±2.7
w/o shared GPS	69.0±1.2	84.9±1.7	73.6±3.9	73.9±0.9	56.8±1.9
w/o motif RWPE	69.3±1.6	86.6±0.8	75.7±1.9	73.2±0.9	53.8±3.6
GCN local enc.	67.9±1.8	88.0±0.3	73.4±5.5	74.2±0.5	56.8±1.3
GINE local enc.	62.8±4.7	83.5±0.3	75.3±5.5	73.4±0.5	57.8±1.1
w/o pretrain	69.9±1.8	85.0±2.7	80.2±2.3	75.7±0.3	57.5±4.3

Removing VDAA causes a clear drop on all five datasets (e.g., $87.0 \rightarrow 63.8$ on BACE), indicating that the learned mixture over fixed, coverage-based topological attention scopes is more effective than standard full self-attention in this setting. The motif branch is the largest single contributor: without it, ClinTox collapses from 94.4 to 52.6, BACE drops from 87.0 to 77.0, and BBBP drops from 74.7 to 65.9, which explains why MoDANet's strongest gains over the baselines appear on these substructure-dominated datasets. Replacing GCMA with masked mean pooling (w/o GCMA) also degrades results substantially (e.g., 78.0 on ClinTox), indicating that global molecular context improves the aggregation of motif representations. The two alternative fusion strategies (w/ concat fusion and w/ gated fusion), a separate input encoding for the motif branch (w/o shared GPS), disabling motif RWPE, and replacing GatedGCN with GCN or GINE as the local encoder all reduce performance, confirming the design choices of the full model. Training from scratch loses 1.9–9.4 points across datasets, confirming that pretraining transfers meaningful knowledge in the evaluated setting. We also note that several ablated variants exhibit larger variance (e.g., ± 8.3 for w/o motif on ClinTox), suggesting that the full model's components jointly stabilize training. Controlled comparisons across pretraining corpus sizes, as well as a full ablation of the BRICS auxiliary objective, are left to future work.

5 Conclusion

We proposed MoDANet, a dual-branch graph transformer for molecular property prediction. MoDANet addresses two limitations of existing graph transformers: indiscriminate global attention and the lack of explicit motif-level structure. The global branch constructs fixed, coverage-based, permutation-equivariant topological attention scopes and learns a layer-wise convex mixture over them, whose weights summarize each layer's preference over the predefined scopes. The motif branch directly encodes retrosynthetically motivated functional-group fragments produced by the fixed BRICS cleavage rules, without relying on a closed fragment lookup vocabulary, and Hierarchical Motif Fusion incorporates them through Global-Context-Guided Motif Aggregation.

Experiments on seven MoleculeNet benchmarks under scaffold splits show that MoDANet achieves state-of-the-art average transfer performance after pretraining on a 250K-molecule ZINC subset. The learned attention-scope mixtures and explicit BRICS-derived motif representations provide structural transparency into the model's organization. Controlled experiments across multiple pretraining-set sizes remain necessary to determine the model's sample efficiency relative to alternative architectures.

REFERENCES

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems*, 2020.

Jörg Degen, Christophe Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using "drug-like" chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. In *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023.

Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3968–3976, 2022.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272, 2017.

William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159, 2020.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pp. 4116–4126, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pp. 2323–2332, 2018.

Dongki Kim, Jinheon Baek, and Sung Ju Hwang. Graph self-supervised learning with accurate discrepancy learning. In *Advances in Neural Information Processing Systems*, 2022.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21618–21629, 2021.

Xiao Qing Lewell, Duncan B. Judd, Stephen P. Watson, and Michael M. Hann. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, 38(3):511–522, 1998.

Han Li, Dan Zhao, and Jianyang Zeng. KPGT: Knowledge-guided pre-training of graph transformer for molecular property prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 857–867, 2022.

Shengchao Liu, Hongyu Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023a.

Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Rethinking tokenizer and decoder in masked graph modeling for molecules. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25854–25875, 2023b.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, 2022.

Bharath Ramsundar, Bowen Liu, Zhenqin Wu, Andreas Verras, Matthew Tudor, Robert P. Sheridan, and Vijay Pande. Is multitask deep learning practical for pharma? *Journal of Chemical Information and Modeling*, 57(8):2068–2076, 2017.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12559–12571, 2020.

Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J. Sutherland, and Ali Kemal Sinop. Exphormer: Sparse transformers for graphs. In *International Conference on Machine Learning*, pp. 31613–31632, 2023.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D Infomax improves GNNs for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502, 2022.

Teague Sterling and John J. Irwin. ZINC 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015.

Mengying Sun, Jing Xing, Huijun Wang, Bing Chen, and Jiayu Zhou. MoCL: Contrastive learning on molecular graphs with multi-level domain knowledge. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022a.

Ruixiang Sun, Hanjun Dai, and Adams Wei Yu. Does GNN pretraining help molecular representation? In *Advances in Neural Information Processing Systems*, 2022b.

Puja Trivedi, Ekdeep Singh Lubana, Yujun Yan, Yaoqing Yang, and Danai Koutra. Augmentations in graph contrastive learning: Current methodological flaws & towards better practices. In *Proceedings of the ACM Web Conference*, 2022.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z. Li. SimGRACE: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM Web Conference*, pp. 1070–1079, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*, pp. 11548–11558, 2021.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888, 2021.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823, 2020.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, 2020.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2023.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15870–15882, 2021.

A AI USE STATEMENT

In accordance with the ICLR 2027 policy on AI use, we disclose the following. Generative AI tools were used (i) to aid and polish the writing of this manuscript, including grammar and readability edits; (ii) for retrieval and discovery of related work, including locating and verifying bibliographic sources; (iii) to support research ideation and execution, such as discussing methodological design choices and debugging analysis code; and (iv) to draft parts of the paper, which were subsequently reviewed and substantially edited by the authors. All scientific claims, model designs, experiments, and numerical results were produced and verified by the authors without generative AI involvement. The authors have reviewed all AI-assisted content and take full responsibility for the final manuscript, including any text or artifacts produced with the aid of generative AI.

B VIEW-FIELD SEARCH FOR VDAA

B.1 SEARCH SPACE AND ENCODING

VDAA partitions the attention field of each node into $R = 3$ scopes: two distance-thresholded scopes followed by a global scope. We parameterise the two thresholds by quantiles of the pairwise shortest-path distance distribution rather than by integer hop counts, so that the same setting adapts to graphs of different sizes and densities. Concretely, a genome is a pair of integers (r_1, r_2) with $r_1 < r_2$ and $r \in \{1, \ldots, 19\}$, where r denotes the $5r$-th percentile; the third scope is fixed to the 100% (i.e., global) field. This yields $\binom{19}{2} = 171$ legal genomes, e.g., genome $(1, 6)$ corresponds to view fields of 5% and 30%.

Figure 3 walks through the construction on a molecule from BACE, using the genome $(2, 10)$ selected for this dataset (Table 4). From the shortest-path distance matrix Δ, each scope admits the $\lceil \rho_r N \rceil$ nearest atoms of every query atom, with all ties at the cutoff distance included, so that a scope may contain slightly more than $\lceil \rho_r N \rceil$ atoms (cf. Eq. equation 3); the 10% scope restricts attention to a tight topological neighbourhood while the 50% scope approaches the full graph; VDAA then mixes the per-scope attention maps with the learned layer-wise weights of Eq. equation 7.

B.2 SEARCH PROCEDURE, ILLUSTRATED ON BBBP

For each downstream dataset, the ratios are selected by an independent genetic-algorithm search over percentile pairs on that dataset's validation split; we illustrate the procedure on BBBP here. The fitness of a genome is the best validation ROC-AUC of a single training run under the downstream protocol (molecular scaffold split, seed 42, 80 epochs, plateau scheduler, and the central hyperparameters of Table 5), which takes roughly 30 minutes per genome on one GPU. We use a standard real-encoded genetic algorithm: population 10 for 6 generations, elite retention of the top 2, uniform crossover, and a 30% mutation rate that resamples one coordinate uniformly at random. Evaluated genomes are cached and skipped. On BBBP, the search consumed 23 unique evaluations ($\approx$12 GPU-hours).

Table 3: Top view-field genomes on BBBP by validation ROC-AUC (seed 42). The default configuration $(6, 12)$ is underlined.

Rank	Genome (r_1, r_2)	View fields	Val
1	$(1, 6)$	$5\%, 30\%$	0.9651
2	$(2, 8)$	$10\%, 40\%$	0.9604
3	$(4, 6)$	$20\%, 30\%$	0.9600
4	$(3, 6)$	$15\%, 30\%$	0.9591
5	$(3, 10)$	$15\%, 50\%$	0.9591
...			
	$\underline{(6, 12)}$	$\underline{30\%, 60\%}$	$\underline{0.9474}$

The search converges rapidly: the best genome changes only once, from $(2, 8)$ at generation 0 to $(1, 6)$ at generation 1 ($0.9604 \rightarrow 0.9651$), and remains unchanged through generation 5, which is

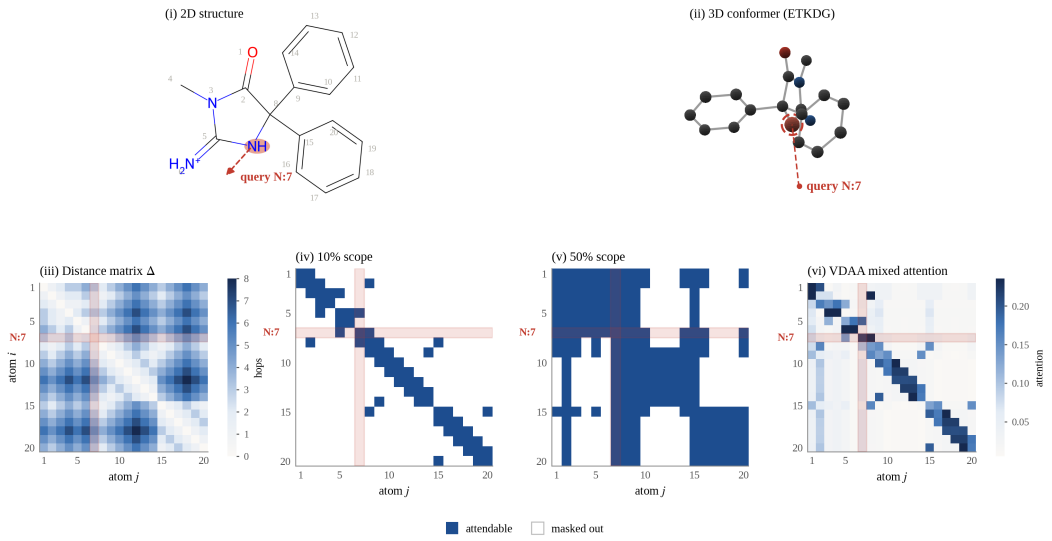


Figure 3: VDAA view-field masking on a molecule from the BACE benchmark (20 heavy atoms), using the scopes selected by the genetic-algorithm search for this dataset (genome $(2, 10)$, i.e., 10% and 50% view fields; Table 4). (i) 2D structure with atom indices; the query atom is N:7. (ii) 3D conformer obtained with ETKDG. (iii) Shortest-path distance matrix Δ; the highlighted row corresponds to the query atom. (iv–v) Binary view-field masks at the 10% and 50% scopes: cells in dark blue are attendable for the query atom, blank cells are masked out; the red row/column marks the query atom itself. (vi) Resulting VDAA mixed attention, obtained as the learned convex combination of the per-scope attention maps. Nested scopes restrict attention to progressively larger topological neighbourhoods of the query atom.

consistent with a locally smooth fitness landscape, though we do not claim coverage of the global optimum (Figure 4). The top-ranked genomes among the 23 evaluated are listed in Table 3.

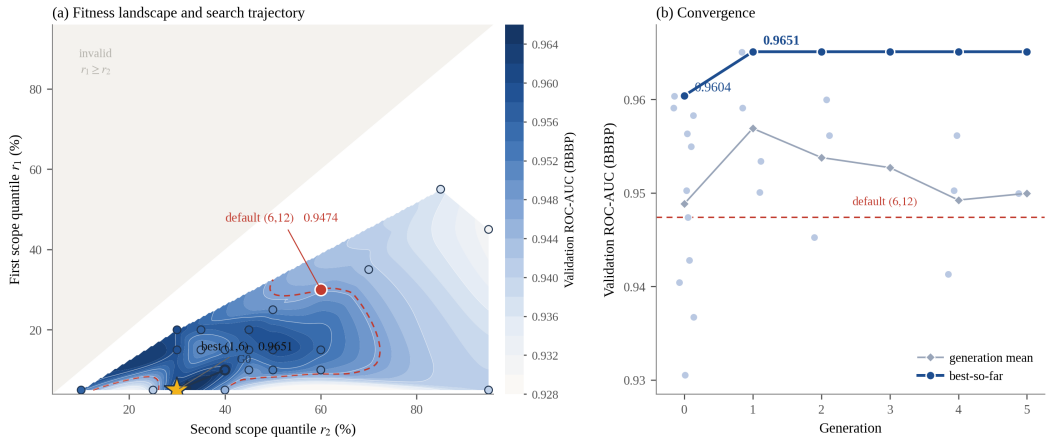


Figure 4: Genetic-algorithm search on BBBP. (a) Fitness landscape over the (r_1, r_2) genome space: each circle is an evaluated genome, colored by its validation ROC-AUC; the star marks the best genome $(1, 6)$ and the red dot marks the default configuration $(6, 12)$; the dashed curve indicates the search trajectory. The gray upper-left region contains invalid genomes with $r_1 \geq r_2$. (b) Convergence of the search: best-so-far validation ROC-AUC (blue) and generation mean (gray); the dashed line marks the default configuration.

B.3 PER-DATASET SELECTION

Because the usefulness of a view field is dataset-dependent, the genome selected on each dataset's validation split differs across the seven datasets (Table 4); the selected genomes consistently outperform the default view fields by an average of $+4.5$ test points. On BBBP, the search evaluated only 23 of the 171 admissible genomes ($\sim 13\%$ of the space) before converging, and the searches on the remaining datasets exhibit the same rapid convergence.

Table 4: Final per-dataset configurations (genomes selected on the validation split) and the best test results obtained with these configurations. n denotes the number of completed seeds.

Dataset	Genome (r_1, r_2)	View fields	Test (mean±std)	n
BBBP	$(1, 6)$	$5\%, 30\%$	0.747 ± 0.004	3
ClinTox	$(1, 19)$	$5\%, 95\%$	0.944 ± 0.009	3
MUV	$(3, 7)$	$15\%, 35\%$	0.770 ± 0.009	3
HIV	$(4, 6)$	$20\%, 30\%$	0.793 ± 0.008	3
BACE	$(2, 10)$	$10\%, 50\%$	0.870 ± 0.017	3
Tox21	$(3, 12)$	$15\%, 60\%$	0.776 ± 0.001	3
SIDER	$(5, 10)$	$25\%, 50\%$	0.648 ± 0.002	3

These configurations are exactly those used to produce the main results in Table 1.

B.4 HYPERPARAMETERS

Table 5: Central hyperparameters. Pretraining uses weight decay 10^{-4}; the view-field search protocol uses weight decay 10^{-3} with all other entries shared.

Item	Value
Hidden dimension d	64
Global layers L / heads H	4 / 4
Motif encoder layers	$L/2 = 2$
RWPE dimension p	20
Learning rate	5×10^{-4}
Dropout / attention dropout	0.3 / 0.3
Weight decay	10^{-4} (pretrain), 10^{-3} (search)
Batch size	128

C RESULTS ON REGRESSION DATASETS

We complement the classification benchmarks of Table 1 with six molecular property regression datasets from MoleculeNet (Wu et al., 2018): ESOL, FreeSolv, and Lipo (physicochemical properties), and QM7, QM8, and QM9 (quantum properties). Following the protocol of prior work, we report the test MAE (lower is better). We compare against GIN (Xu et al., 2019) and GPS (Rampášek et al., 2022), as well as representative self-supervised pre-training methods (GraphMVP (Liu et al., 2022), SimSGT (Liu et al., 2023b), MolCLR (Wang et al., 2022), 3DInfomax (Stärk et al., 2022), MGSSL (Zhang et al., 2021), GraphMAE (Hou et al., 2022), GraphCL (You et al., 2020), MoleculeSTM (Liu et al., 2023a), and SimGRACE (Xia et al., 2022)).

Table 6 reports the results. MoDANet attains the lowest MAE on ESOL (1.039) and QM9 (0.00125); on QM9 the margin is substantial, $1.8\times$ lower than the strongest baseline (GPS, 0.0022) and an order of magnitude lower than most self-supervised baselines, suggesting that motif-aware attention is particularly effective at capturing the structural regularities underlying quantum properties. MoDANet is second on FreeSolv and remains competitive on Lipo (0.735 vs. 0.702) and QM8 (0.0105 vs. 0.0099); we attribute the advantage of GraphMVP and MoleculeSTM on these two datasets to their geometric inductive bias, as the predicted properties depend strongly on 3D geometry. On QM7,

Table 6: Test MAE (↓) on six regression datasets. Best in bold, second best underlined.

Method	ESOL ↓	FreeSolv ↓	Lipo ↓	QM7 ↓	QM8 ↓	QM9 ↓
GIN	1.210±0.036	2.925±0.370	0.800±0.039	179.5±8.7	0.0114±0.0007	0.00930±0.00030
GPS	1.196±0.032	2.483±0.102	0.820±0.021	129.9±7.3	0.0116±0.0003	0.00220±0.00040
GraphMVP	1.047±0.010	**1.973**±0.176	**0.702**±0.012	158.5±3.4	<u>0.0100</u>±0.0000	0.00800±0.00021
SimSGT	1.074±0.007	2.690±0.152	0.758±0.010	129.1±3.2	0.0126±0.0003	0.00890±0.00000
MolCLR	1.347±0.048	2.865±0.304	<u>0.732</u>±0.009	140.3±4.6	0.0107±0.0003	0.00859±0.00020
3DInfomax	2.007±1.573	2.248±0.429	0.936±0.298	128.1±4.7	0.0136±0.0014	0.01780±0.00980
MGSSL	1.401±0.035	3.148±0.163	0.775±0.016	140.6±6.9	0.0107±0.0003	0.00930±0.00000
GraphMAE	1.276±0.020	2.913±0.121	0.773±0.006	148.7±1.8	0.0104±0.0001	0.00966±0.00022
GraphCL	1.332±0.039	3.182±0.198	0.790±0.019	<u>110.0</u>±4.9	0.0108±0.0003	0.00961±0.00032
MoleculeSTM	1.192±0.004	2.798±0.133	0.751±0.004	110.3±2.4	**0.0099**±0.0001	0.00949±0.00020
SimGRACE	1.420±0.014	3.143±0.126	0.812±0.010	**91.3**±3.7	0.0120±0.0002	0.01048±0.00038
MoDANet (Ours)	**1.039**±0.059	<u>1.995</u>±0.022	0.735±0.009	118.6±0.6	0.0105±0.0004	**0.00125**±0.00025

SimGRACE, GraphCL, and MoleculeSTM perform best, while MoDANet still outperforms the remaining methods. Averaged over the six datasets, MoDANet achieves the best mean rank (2.5), ahead of GraphMVP (3.3) and all other self-supervised baselines, demonstrating consistently competitive performance on both physicochemical and quantum property prediction.

D Visualization of Learned Representations

To better understand the learned representation space, we visualize MoDANet embeddings with t-SNE on four datasets, coloring each point by its property value (Figure 5). On the continuous-target datasets (Lipo, QM9, QM8), the embeddings form smoothly graded manifolds: points that are close in the embedding space exhibit similar property values, as reflected by the coherent color gradients. This indicates that the representation space is organized according to physicochemical and quantum properties rather than arbitrary cluster structure. On the binary-label BACE dataset, the two classes form partially separated regions with substantial overlap, consistent with the known difficulty of this benchmark.

E Case Study: Attention Aligns with Known BACE Pharmacophores

We inspect the VDAA mixed attention of individual BACE test molecules (Figure 6). Attention is extracted from a single fine-tuned checkpoint by replicating the forward pass of the distance-masked attention module; atoms are colored by the normalised attention they receive (column sums of the layer- and head-averaged mixed attention matrix, normalised within each molecule), and the three highest-attention atoms of each molecule are circled and annotated with their matched substructure. Cases are selected by prediction outcome (two true positives, two true negatives, one false positive, one false negative).

The correct predictions are chemically coherent. In both true positives, attention concentrates on motifs that match established BACE-1 pharmacophore elements: the nitrile nitrogen and the fused ring system in (i)—nitriles are a known warhead that engages the catalytic dyad (Asp32/Asp228)—and the hydroxyl and amide carbonyl in (ii), a canonical hinge hydrogen-bonding pattern, with additional coverage of the hydrophobic pocket. In the true negatives (iii–iv), attention is diffuse and no warhead-like fragment attracts a concentrated response, consistent with the model's low predicted probability of inhibition. The failure modes are similarly interpretable. In the false positive (v), attention anchors on a sulfonamide, a fragment that frequently occurs among real BACE inhibitors, but the molecule lacks the surrounding pharmacophore and the model is misled by the local feature; in the false negative (vi), attention spreads over an aniline-like region without consolidating into strong evidence, indicating that the model's recognition of this scaffold family remains unstable. Taken together, the cases suggest that both the successes and the failures of the model are attributable to identifiable substructural evidence, which is the intended behaviour of an inspectable attention mechanism.

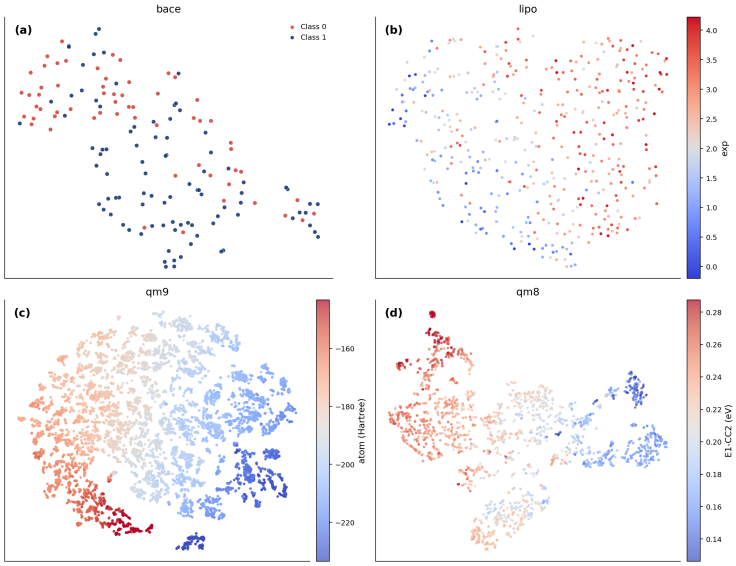


Figure 5: t-SNE visualization of MoDANet molecular representations. (a) BACE, colored by binary label; (b) Lipo, colored by lipophilicity; (c) QM9, colored by atomization energy (Hartree); (d) QM8, colored by the E1-CC2 excitation energy (eV). Continuous property values form smooth color gradients in the embedding space, indicating that the learned representations are aligned with the target properties.

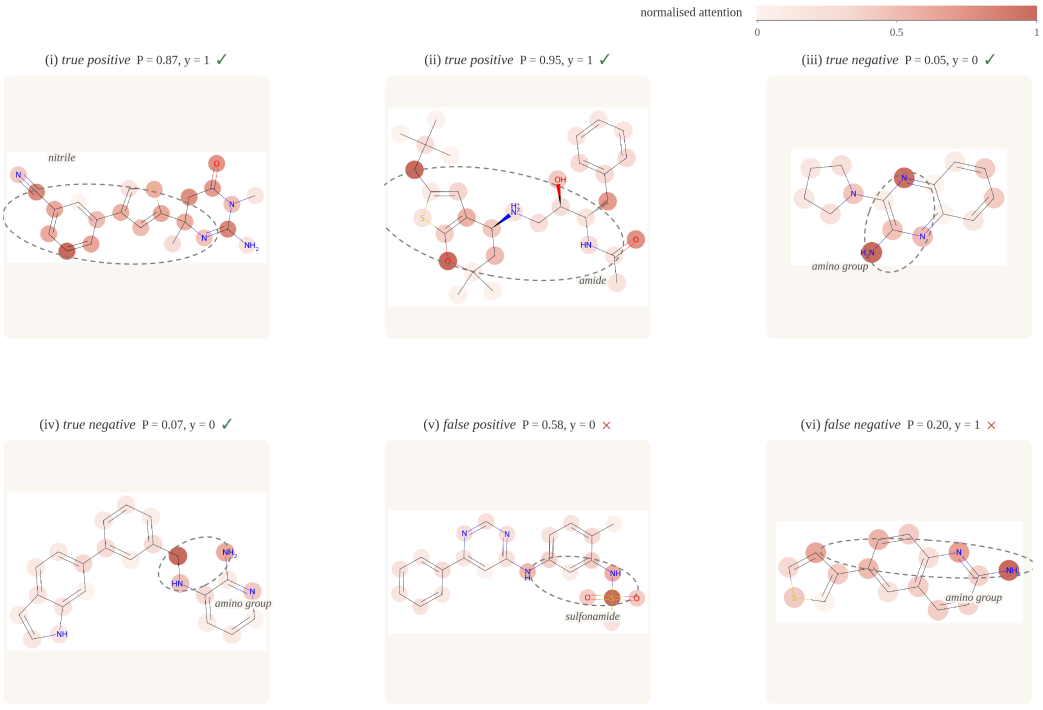


Figure 6: Case study of VDAA attention on six BACE test molecules, selected by prediction outcome: (i–ii) true positives, (iii–iv) true negatives, (v) false positive, (vi) false negative. Atoms are colored by received normalised attention (scale at top); dashed circles mark the three highest-attention atoms, annotated with the matched substructure. P is the predicted inhibition probability, y the ground-truth label, and the symbol marks correctness at the 0.5 threshold.