LEARNING TO SELECT SOURCE DOMAINS: PROXY-REWARDED POLICY OPTIMIZATION FOR MOLECULAR OOD GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust molecular property prediction under extreme out-of-distribution (OOD) shifts faces two challenges. Conventional scaffold splits can retain local structural similarity across training and test molecules, while adaptation from heterogeneous source domains can lead to negative transfer. We introduce Policy Optimization for Multi-source Adaptation (POMA), a retrieve–compose–adapt framework for adaptive source domain selection when target property labels are unavailable. POMA uses unlabeled target structures to retrieve structurally similar labeled source scaffolds as proxy targets. A reinforcement learning policy learns to explore combinations of candidate source domains, using post-adaptation reductions in proxy-task mean absolute error (MAE) as rewards. The predictor is then adapted using the selected sources, with covariance alignment at the whole-molecule and BRICS-derived substructure levels. We also introduce SCOPE-Bench, a scaffold-cluster benchmark constructed in an explicit physicochemical descriptor space. Experiments on QM9 evaluate HOMO, LUMO, and HOMO–LUMO gap prediction across three 3D molecular backbones and 15 target-scaffold tasks. Prediction errors on SCOPE-Bench increase by up to $8.0\times$, with a mean of $5.9\times$, relative to conventional scaffold splitting. Relative to the supervised baseline, POMA reduces MAE by up to 11.2%, averaging 6.2% over the nine evaluated backbone–property combinations. Code is available at `https://anonymous.4open.science/r/Molecular-OOD-Code-73F6`.

1 INTRODUCTION

Recent 3D geometric models, including ViSNet, ETNN, and GotenNet, encode molecular geometry in learned representations and achieve strong performance on standard molecular benchmarks (Wang et al., 2024; Battiloro et al., 2025; Aykent & Xia, 2025). Yet benchmark accuracy alone does not establish reliability under structural distribution shifts (Liu et al., 2021; Li et al., 2025). We study how to evaluate molecular out-of-distribution (OOD) generalization beyond conventional scaffold splits and how to select source subsets when target property labels are unavailable.

Scaffold splitting groups molecules by their Bemis–Murcko scaffolds and assigns scaffold groups to disjoint data splits (Bemis & Murcko, 1996; Wu et al., 2018). However, distinct scaffolds may still share local substructures relevant to property prediction, so scaffold disjointness alone does not guarantee separation of these patterns. We introduce SCOPE-BENCH, a scaffold-cluster OOD benchmark that clusters scaffolds in an explicit physicochemical descriptor space and partitions the resulting clusters into source, validation, and target domains (Yang et al., 2022). This descriptor-based cluster split complements conventional scaffold splitting by evaluating prediction under cluster-level structural shifts.

Beyond evaluation, adaptation requires deciding which source domains to use. Transfer from heterogeneous labeled sources can be beneficial (Pan & Yang, 2009; Peng et al., 2019), but poorly matched sources may induce negative transfer (Rosenstein et al., 2005). Our ViSNet–HOMO comparisons illustrate this sensitivity. With the dual-scale adaptation design fixed across selection strategies, some strategies improve over the supervised baseline whereas others perform worse (Table 2). We therefore treat source subset selection as an explicit adaptation decision. Structural

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.

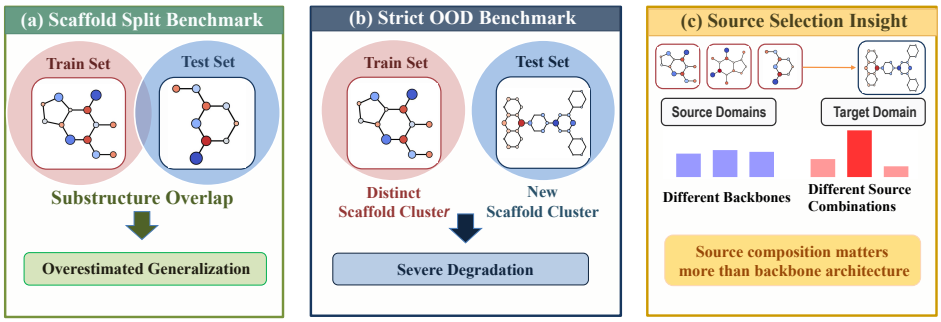


Figure 1: Conceptual motivation for molecular OOD evaluation and source selection. (a) Distinct scaffolds can share local substructures. (b) Descriptor-based scaffold clustering defines source and target partitions. (c) Source composition is an adaptation decision. All panels are schematic; panel (c) does not quantify the relative effects of source composition and backbone choice.

similarity can retrieve candidates but does not directly measure the downstream benefit of adapting with a source subset. Learning this benefit without target property labels therefore requires a surrogate feedback signal. Figure 1 illustrates these evaluation and source-selection motivations.

We propose Policy Optimization for Multi-source Adaptation (POMA), a retrieve–compose–adapt framework that learns source-subset selection from proxy-task adaptation feedback. Unlabeled target structures guide the retrieval of structurally similar scaffolds from labeled source domains as proxy targets. POMA evaluates sampled source subsets through adaptation to these proxy targets and uses post-adaptation reductions in proxy-task mean absolute error (MAE) as rewards. These rewards train a reinforcement learning policy over the combinatorial source-subset space using Group Relative Policy Optimization (Shao et al., 2024). Structural similarity thus guides retrieval, while proxy-task adaptation performance supplies the learning signal for source selection.

The selection policy is coupled with a dual-scale adaptation module. For each selected subset, POMA combines supervised regression on labeled source molecules with covariance alignment between source and target representations at the whole-molecule and BRICS-derived substructure levels (Degen et al., 2008; Sun & Saenko, 2016). The same adaptation procedure is used both to evaluate source subsets on proxy tasks and to adapt the predictor to the real target domain using labeled molecules from the selected sources and unlabeled target structures. Policy feedback is therefore tied to the downstream adaptation procedure, as summarized in Figure 2.

Our contributions are threefold:

(i) Scaffold-cluster OOD evaluation. We introduce SCOPE-BENCH to evaluate molecular prediction under descriptor-based scaffold-cluster shifts. Across the nine evaluated backbone–property combinations on QM9, the MAE under SCOPE-BENCH is as much as $8.0\times$ the corresponding MAE under conventional scaffold splitting; the mean degradation factor is $5.9\times$.

(ii) Proxy-reward-driven source selection. We develop POMA, which learns source-subset selection from proxy-task adaptation rewards and couples the learned policy with covariance alignment at the whole-molecule and BRICS-derived substructure levels.

(iii) Empirical evaluation. Experiments on QM9 cover HOMO, LUMO, and HOMO–LUMO gap prediction across three 3D molecular backbones and 15 target-scaffold tasks. Relative to the supervised baseline, POMA reduces MAE by up to 11.2%, with an average relative MAE reduction of 6.2% over the nine evaluated backbone–property combinations. Source-selection comparisons and alignment ablations examine the framework components within this setting.

2 Related Work

Molecular representation learning. Molecular representations include fixed fingerprints (Rogers & Hahn, 2010), graph convolutional networks (Kipf & Welling, 2016), and message-passing networks (Gilmer et al., 2017). SchNet uses continuous-filter convolutions for rotationally invariant energy predictions (Schütt et al., 2017), while PaiNN and EGNN use equivariant geometric message passing (Schütt et al., 2021; Satorras et al., 2021). ViSNet, ETNN, and GotenNet are recent 3D

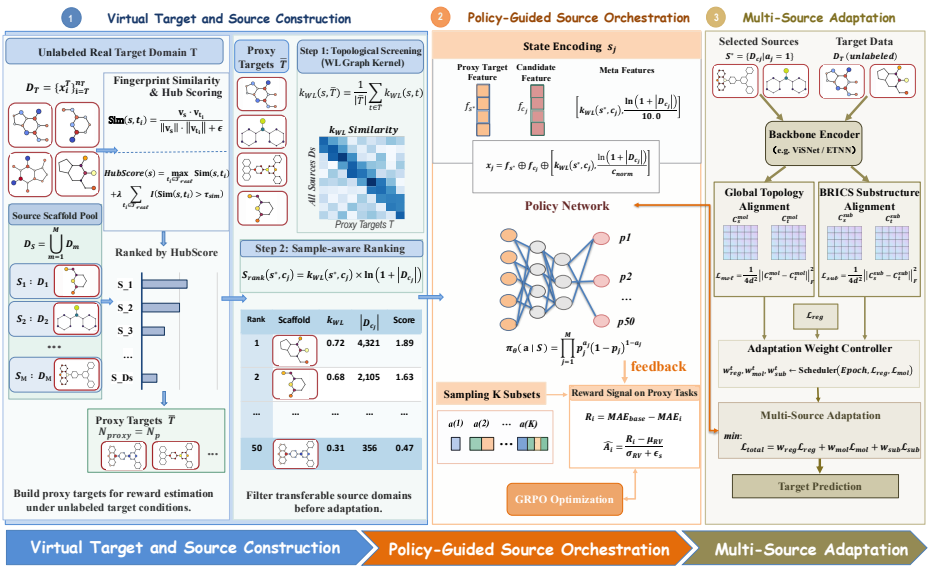


Figure 2: Overview of POMA. Unlabeled target structures guide the retrieval of scaffolds from labeled source domains as proxy targets. Sampled source subsets are evaluated through dual-scale adaptation on proxy tasks, and the resulting MAE reductions provide policy rewards. The trained policy selects a source subset for the real target domain; the predictor is then adapted using labeled molecules from the selected source domains and unlabeled real-target structures, with covariance alignment at the whole-molecule and BRICS-derived substructure levels.

geometric models (Wang et al., 2024; Battiloro et al., 2025; Aykent & Xia, 2025). Our focus is source-domain subset selection and adaptation under structural distribution shifts, rather than a new molecular backbone.

Molecular OOD evaluation. MoleculeNet uses scaffold-based splits (Bemis & Murcko, 1996; Wu et al., 2018); GOOD evaluates several graph OOD settings (Gui et al., 2022); and MoleOOD studies substructure-invariant molecular representations (Yang et al., 2022). DrugOOD addresses noisy drug–target binding-affinity prediction, while MOOD studies deployment-oriented molecular distribution shifts and uncertainty calibration (Ji et al., 2023; Tossou et al., 2024). SCOPE-BENCH complements these benchmarks by clustering scaffolds in a physicochemical descriptor space and assigning clusters to source, validation, and target domains.

Multi-source adaptation and source selection. Multi-source adaptation includes adversarial alignment and moment matching (Zhao et al., 2018; Peng et al., 2019), while Deep CORAL aligns second-order feature statistics (Sun & Saenko, 2016). Because source domains can differ in transfer utility, structural similarity alone may not identify useful source subsets (Rosenstein et al., 2005; Wang et al., 2019; Togninalli et al., 2019). Reinforcement learning provides a tool for combinatorial source selection (Bello et al., 2016), and Group Relative Policy Optimization estimates group-relative advantages without a separate value model (Shao et al., 2024). POMA separates candidate retrieval from proxy-task utility learning and uses dual-scale covariance alignment for adaptation.

3 METHOD

3.1 PROBLEM DEFINITION

The prediction task is defined over 3D molecular graphs $G_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{R}_i, \mathbf{Z}_i)$ with label $y_i \in \mathbb{R}$. The topological scaffold of each molecule is identified using Bemis–Murcko scaffold extraction, $s_i = \phi(G_i)$.

Let $\mathcal{S}$ and $\mathcal{T}$ denote the source and target scaffold sets, respectively, associated with labeled source molecules and unlabeled target molecules. We consider an extreme OOD setting in which $\mathcal{S} \cap \mathcal{T} = \emptyset$

and the two domains are separated through clustering and partitioning in an explicit physicochemical descriptor space.

Two challenges arise in this setting. First, distinct scaffolds can share local substructures, so scaffold disjointness alone does not ensure separation of local structural patterns relevant to property prediction. Second, heterogeneous source domains can differ in transfer utility, and adaptation with poorly matched sources may degrade target prediction. We therefore learn a policy to select a source subset $\mathcal{S}^* \subset \mathcal{S}$ for adaptation.

3.2 Construction of SCOPE-Bench

To construct the scaffold-cluster split based on physicochemical descriptors, SCOPE-Bench forms source, validation, and target domains through a three-step pipeline.

Step 1: Scaffold extraction and feature construction. Bemis–Murcko scaffolds are extracted from QM9 (Ramakrishnan et al., 2014) via RDKit (Landrum et al., 2024), yielding 1,247 unique scaffolds with at least 10 samples each. Each scaffold s_k is represented by a four-dimensional physicochemical descriptor:

$$f_k = V_{macro} \oplus V_{element} \oplus V_{conn} \oplus V_{flex} \tag{1}$$

where $\oplus$ denotes concatenation; V_{macro} encodes global topology (atom/ring count), $V_{element}$ captures elemental polarizability, V_{conn} reflects electron delocalization and rigidity, and V_{flex} measures conformational entropy through rotatable-bond ratios.

Step 2: Hierarchical clustering. Scaffolds are first grouped into five levels by ring count and maximum ring size. Within each level, the descriptor vectors are z-score normalized and clustered with k-means++ (Arthur et al., 2007). Cluster quotas are allocated proportionally to level size, with a sparsity compensation coefficient α_{comp} for underrepresented polycyclic levels, yielding a total of $K_{total} = 12$ disjoint scaffold clusters across the five level-wise normalized physicochemical descriptor spaces.

Step 3: Asymmetric partitioning. Six clusters form the source domain, one is used for validation, and the remaining five form the target domain. Target scaffolds with sufficient sample size are then used as target-scaffold evaluation tasks. These target scaffolds are excluded from labeled source training, while their unlabeled structures remain available for the later retrieval and adaptation steps; their property labels are not used by the source-selection policy.

The partitions are fixed by cluster assignment. The distance thresholds reported in Appendix C are used only for post hoc diagnostics and do not affect benchmark construction.

3.3 Policy Optimization for Multi-source Adaptation

Under extreme structural shifts, adaptation from heterogeneous source domains may lead to negative transfer, motivating adaptive source selection. We therefore propose Policy Optimization for Multi-source Adaptation (POMA), which organizes molecular knowledge transfer as an integrated retrieve–compose–adapt process. Rather than relying only on fixed similarity scores, POMA learns a source-selection policy over combinations of candidate source domains. The policy uses structural information from the unlabeled target domain to guide source-subset selection for subsequent adaptation.

3.3.1 Task-Specific Environment Construction

In the unlabeled target adaptation setting, direct policy optimization must cope with sparse rewards and a combinatorial source-subset space. To construct a task-specific surrogate feedback signal, structurally similar scaffolds from labeled source domains are selected as proxy targets using the unlabeled real-target structures.

Given N_T unlabeled real target scaffolds $\mathcal{T}_{real}$ and a sufficiently large source domain pool $\mathcal{S}_{pool}$, the Morgan fingerprint (Morgan, 1965) cosine similarity between a candidate scaffold $s \in \mathcal{S}_{pool}$ and a real target $t_i \in \mathcal{T}_{real}$ is computed as:

$$\mathrm{Sim}(s, t_i) = \frac{\mathbf{v}_s \cdot \mathbf{v}_{t_i}}{\|\mathbf{v}_s\| \cdot \|\mathbf{v}_{t_i}\| + \epsilon} \tag{2}$$

where $\mathbf{v}_s$ and $\mathbf{v}_{t_i}$ denote the fingerprint feature vectors of source scaffold s and target scaffold t_i, respectively. Here $\|\cdot\|$ is the L_2 norm, and ϵ is a small positive constant for numerical stability. A joint HubScore is defined by

$$\text{HubScore}(s) = \max_{t_i \in \mathcal{T}_{real}} \text{Sim}(s, t_i) + \lambda \sum_{t_i \in \mathcal{T}_{real}} \mathbb{I}(\text{Sim}(s, t_i) > \tau_{sim}) \tag{3}$$

where $\lambda > 0$ is a coverage-balancing coefficient, $\mathbb{I}(\cdot)$ is the indicator function, and τ_{sim} is the similarity threshold. The top-$N_p = 20$ scaffolds by HubScore form $\mathcal{T}_{proxy}$. This screening uses structural fingerprints and unlabeled target structures, not target property labels.

For each proxy target $s^* \in \mathcal{T}_{proxy}$, candidate source domains are ranked independently. Each candidate c_j is a source scaffold, and $\mathcal{D}_{c_j}$ contains the source molecules assigned to that scaffold. The source clusters define the benchmark partition, while selection operates on scaffolds within that partition. The Weisfeiler–Lehman (WL) graph kernel (Shervashidze et al., 2011), together with source-domain sample size, defines

$$S_{rank}(s^*, c_j) = k_{WL}(s^*, c_j) \times \ln(1 + |\mathcal{D}_{c_j}|) \tag{4}$$

where $k_{WL}(s^*, c_j)$ is the WL similarity between proxy target s^* and candidate domain c_j. The proxy-specific candidate pool is

$$\mathcal{C}(s^*) = \text{TopM}_{c_j}\ S_{rank}(s^*, c_j), \qquad |\mathcal{C}(s^*)| = M = 50.$$

Each s^* defines a policy-training environment with its own candidate pool and reward. Candidate pools are not merged, and rewards are not averaged across proxy tasks. One shared policy is updated sequentially across these environments.

3.3.2 Dual-Scale Decoupled Domain Adaptation

To model distribution differences at complementary molecular scales, the encoder constructs two representation paths: a whole-molecule path yielding features $h_{mol} \in \mathbb{R}^d$ and a substructure path yielding BRICS-derived features $h_{sub} \in \mathbb{R}^d$ through BRICS retrosynthetic cleavage (Degen et al., 2008). For each path, the discrepancy between a selected source domain and the target domain is measured by the squared Frobenius norm of the difference between their empirical covariance matrices, following the Deep CORAL criterion (Sun & Saenko, 2016).

For the environment associated with proxy target s^*, let an executed action $\mathbf{a}_i$ select $K_i = \|\mathbf{a}_i\|_0$ source domains $c_1, \ldots, c_{K_i}$, and let $S_k = S_{rank}(s^*, c_k)$. For a nonempty action, the ranking scores are normalized only over its selected sources: $\gamma_k = S_k / \sum_{\ell=1}^{K_i} S_\ell$. The dual-scale alignment objective for that action is

$$\mathcal{L}_{DA} = w_{mol} \sum_{k=1}^{K_i} \gamma_k \cdot \frac{\|\mathbf{C}_s^{mol,k} - \mathbf{C}_t^{mol}\|_F^2}{4d^2} + w_{sub} \sum_{k=1}^{K_i} \gamma_k \cdot \frac{\|\mathbf{C}_s^{sub,k} - \mathbf{C}_t^{sub}\|_F^2}{4d^2} \tag{5}$$

where $\mathbf{C}_s^{mol,k}$ and $\mathbf{C}_t^{mol}$ are the $d \times d$ mini-batch covariance matrices of whole-molecule features, and $\mathbf{C}_s^{sub,k}$ and $\mathbf{C}_t^{sub}$ are their substructure counterparts. We use the sample-covariance denominator $n - 1$, and $\|\cdot\|_F$ denotes the Frobenius norm. For the empty control ($K_i = 0$), alignment and γ normalization are skipped. The total objective is

$$\mathcal{L}_{total} = w_{reg}\, \mathcal{L}_{reg} + \mathcal{L}_{DA}. \tag{6}$$

Training proceeds in two phases. During the first $E_{warm} = 10$ epochs, only $\mathcal{L}_{reg}$ is used for supervised warm-up. From epoch 11 onward, the alignment objective is activated and a backbone-specific dynamic controller updates w_{reg}, w_{mol}, and w_{sub} once per epoch. The controller rules, including any applicable clipping bounds, are backbone-specific and are therefore not represented by a single shared formula.

3.3.3 Source-domain Combinatorial Decision via GRPO

For each proxy-specific environment, a state representation vector $\mathbf{x}_j \in \mathbb{R}^{d_{state}}$ with $d_{state} = 258$ is constructed for candidate scaffold c_j:

$$\mathbf{x}_j = f_{s^*} \oplus f_{c_j} \oplus \left[k_{WL}(s^*, c_j), \frac{\ln(1 + |\mathcal{D}_{c_j}|)}{c_{norm}} \right] \tag{7}$$

where f_k in Eq. (1) is the four-dimensional physicochemical descriptor used for scaffold clustering, whereas f_{s^*} and f_{c_j} here are 128-dimensional Morgan fingerprints used by the policy state. Together with the WL similarity and normalized log sample count, these components give $d_{state} = 128+128+2 = 258$. The shared policy network outputs a selection logit and a Bernoulli selection probability p_j for each candidate.

For each policy-training group, one of the $G = 33$ slots is reserved for the designated empty control $\mathbf{a}_0 = \mathbf{0}$. Each remaining slot first samples a proposal $\mathbf{b}_i \in \{0,1\}^M$ from independent Bernoulli variables with probabilities $p_1, \ldots, p_M$. Proposals selecting at most $K_{\max} = 5$ candidates are retained; proposals selecting more than five retain the entries with the largest policy logits. The resulting executed subset $\mathbf{a}_i$ has cardinality $K_i = \|\mathbf{a}_i\|_0 \in [0, 5]$.

For compactness, let $S = \mathbf{X}(s^*)$ denote the current proxy-specific policy state matrix.

The policy update uses the following factorized action score evaluated on the executed subset:

$$\ln \pi_\theta(\mathbf{a}_i \mid S) = \sum_{j=1}^{M} \left[a_{ij} \ln p_j + (1 - a_{ij}) \ln(1 - p_j) \right]. \tag{8}$$

Here $\pi_\theta(\mathbf{a}_i \mid S)$ denotes the factorized action score used by the policy update after the proposal has been transformed into the executed subset used by the proxy-task evaluator.

For each proxy target, all action slots in a group use the same supervised initialization and proxy-task evaluation split. A slot is considered valid only if its evaluator successfully returns a metric. The designated empty control follows the supervised path without selected-source alignment, and its returned error defines MAE_{base}. For each valid nonempty executed subset, MAE_i is measured after whole-molecule and BRICS-derived substructure alignment. The raw proxy reward is

$$R_i = \mathrm{MAE}_{base} - \mathrm{MAE}_i.$$

Each proxy target defines a separate policy-training task, and the same policy parameters are updated sequentially across these tasks. These rewards are surrogate feedback for source-subset selection and are not asserted to identify the optimal subset for the real target domain.

Let $V \subseteq \{0, \ldots, G-1\}$ denote the indices of action slots whose proxy evaluators successfully return a metric.

For each GRPO iteration, the G slots are evaluated in parallel and the intra-group normalized advantage is

$$\hat{A}_i = \frac{R_i - \mu_{R_V}}{\sigma_{R_V} + \epsilon_s} \tag{9}$$

where μ_{R_V} and σ_{R_V} are the mean and standard deviation of the rewards indexed by V.

This GRPO implementation does not train a separate critic network. To constrain policy update step sizes, a reference network π_{ref} and a KL-based reference penalty are introduced. Letting $\Delta_i = \ln \pi_\theta(\mathbf{a}_i \mid S) - \ln \pi_{ref}(\mathbf{a}_i \mid S)$, the penalty is estimated as:

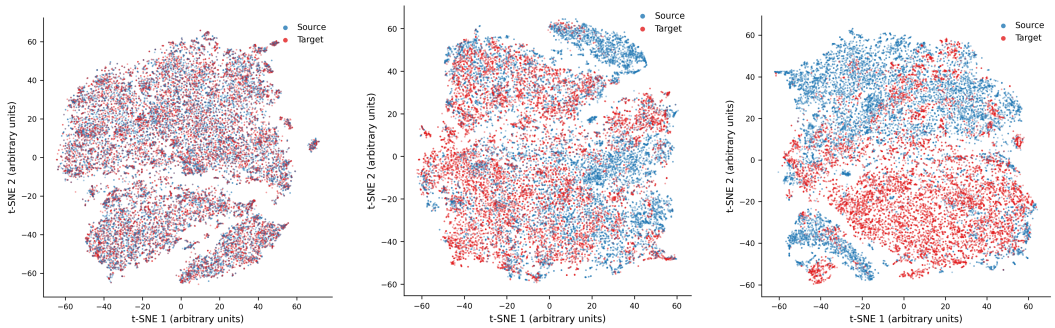
$$D_{KL}(\pi_\theta \,\|\, \pi_{ref}) \approx \frac{1}{|V|} \sum_{i \in V} \left(\exp(\Delta_i) - \Delta_i - 1 \right) \tag{10}$$

The GRPO optimization objective is:

$$\mathcal{L}_{GRPO}(\theta) = -\frac{1}{|V|} \sum_{i \in V} \min\Big(r_i(\theta)\, \hat{A}_i, \mathrm{clip}(r_i(\theta), 1 - \epsilon_{clip},\, 1 + \epsilon_{clip})\, \hat{A}_i \Big) + \beta\, D_{KL}(\pi_\theta \,\|\, \pi_{ref}) \tag{11}$$

where $r_i = \pi_\theta(\mathbf{a}_i \mid S)/\pi_{ref}(\mathbf{a}_i \mid S)$ is the score ratio used in the clipped objective, ϵ_{clip} is the clipping threshold, and $\beta > 0$ is the KL penalty coefficient. At deployment, for the current real-target scaffold t, candidate source domains are reranked using $S_{rank}(t, c_j)$ to construct the target-specific pool $\mathcal{C}(t)$ and state matrix $\mathbf{X}(t)$. The shared policy then deterministically selects the $K_{\mathrm{deploy}} = 5$ candidates with the largest selection probabilities, and the selected source domains are used for target-specific adaptation with $\mathcal{L}_{total}$. The full staged procedure is summarized in Algorithm 1 in the appendix.

The split diagnostics are reported in Section 4.2, with the feature-space coverage definition and additional results in Appendix C.



(a) Random split (b) Standard scaffold split (c) SCOPE-BENCH

Figure 3: t-SNE visualizations of L2-normalized molecular features. Each protocol includes all 12,000 sampled source molecules (blue) and 12,000 sampled target molecules (red). The same feature extractor and t-SNE settings are used, but embeddings are fitted independently. Projected distances are not used to quantify separation. Appendix C reports diagnostics in the original feature space.

Table 1: Overall mean absolute error (MAE, eV) on QM9 under conventional scaffold splitting and the strict OOD split used by SCOPE-BENCH. The degradation-factor row reports the strict-OOD supervised MAE divided by the corresponding scaffold-split MAE, and the final row reports the relative MAE reduction of POMA from the strict-OOD supervised baseline.

	ViSNet (Wang et al., 2024)			ETNN (Battiloro et al., 2025)			GotenNet (Aykent & Xia, 2025)		
	HOMO	LUMO	Gap	HOMO	LUMO	Gap	HOMO	LUMO	Gap
Scaffold split	0.0280	0.0230	0.0583	0.0262	0.0238	0.0387	0.0246	0.0257	0.0324
Strict OOD split	0.1621	0.1829	0.2303	0.1349	0.1312	0.1814	0.1580	0.1902	0.2059
Degradation factor	5.8×	8.0×	4.0×	5.1×	5.5×	4.7×	6.4×	7.4×	6.4×
+POMA	**0.1540**	**0.1673**	**0.2270**	**0.1233**	**0.1165**	**0.1681**	**0.1535**	**0.1728**	**0.2021**
Improve.	5.0%	8.5%	1.4%	8.6%	11.2%	7.3%	2.8%	9.1%	1.8%

3.4 COMPLEXITY ANALYSIS

Let D_{src} denote the number of source domains considered during retrieval, M the retained candidate-pool size, and $N_{selected}$ the number of molecules in the final selected source set. Candidate construction requires scoring source domains and retaining the top M candidates; its cost depends on the number and size of the source-domain graphs and the WL-kernel implementation. During policy training, each GRPO group evaluates G action slots. The dominant cost is the proxy adaptation associated with each slot, which depends on the selected source data, the number of proxy-adaptation epochs and minibatches, backbone forward/backward computation, and the whole-molecule and BRICS-derived substructure alignment terms. Thus, policy training cost scales with the number of evaluated groups and action slots, not merely with candidate-pool size. After the candidate pool is fixed, the policy network computation depends on M; proxy and final adaptation costs continue to depend on the data processed for the selected sources. At deployment, deterministic top-K_{deploy} selection operates over the M candidates, followed by target-specific adaptation on the selected source data. Per-query prediction then incurs the cost of the chosen molecular backbone; retrieval, selection, and adaptation are additional per-target setup costs.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

SCOPE-BENCH is built on QM9 (Ramakrishnan et al., 2014) following the protocol of Section 3.2. Three quantum chemical properties are evaluated: the highest occupied molecular orbital energy (HOMO), the lowest unoccupied molecular orbital energy (LUMO), and HOMO–LUMO gap (Gap), all in eV. All results are obtained with a fixed random seed of 42 for reproducibility. Two protocols are compared: a supervised baseline fine-tuned on the source domain for 200 epochs without adaptation, and POMA with offline policy optimization followed by target-specific adaptation.

Table 2: ViSNet–HOMO ablations on SCOPE-BENCH. The upper block compares source selection policies with full dual-scale adaptation. The lower block removes one alignment module while keeping the GRPO-learned policy fixed. Positive and negative Δ values indicate improvement and deterioration relative to the supervised baseline, respectively. Symbols ✓ and $\times$ denote enabled and disabled modules.

Policy	Mol-CORAL (M)	Sub-CORAL (S)	MAE	Δ
Only ViSNet	-	-	0.1621	-
Random selection	✓	✓	0.1708	-5.3%
Shallow feature matching	✓	✓	0.1663	-2.5%
Deep feature matching	✓	✓	0.1669	-2.9%
Physical descriptor	✓	✓	0.1641	-1.2%
Graph kernel similarity	✓	✓	0.1616	$+0.3\%$
Mixed strategy	✓	✓	0.1576	$+2.8\%$
POMA w/o M	$\times$	✓	0.1689	-4.2%
POMA w/o S	✓	$\times$	0.1672	-3.1%
POMA (Full)	✓	✓	0.1540	$+\mathbf{5.0}\%$

4.2 VERIFICATION OF SCOPE-BENCH

SCOPE-BENCH partitions 133,885 QM9 molecules into 12 disjoint scaffold clusters: 6 form the source domain (94,562 samples), 1 the validation set (18,326 samples), and 5 the target domain (19,894 samples), with 15 target-scaffold evaluation tasks based on scaffolds with $N \geq 200$. The t-SNE panels provide qualitative visualizations with equal sample counts across the three protocols (Figure 3); pairwise Tanimoto similarity (Bajusz et al., 2015) heatmaps provide a complementary qualitative diagnostic of cross-domain similarity patterns (Figure 6 in Appendix C). A post hoc nearest-source diagnostic on the diagnostic feature caches shows lower target coverage for SCOPE-BENCH than for the random and conventional scaffold protocols at each reported source-calibrated quantile (Appendix C). Because the cache-to-split mapping and molecule-level identity provenance are not established, we treat these measurements as descriptive diagnostics for the specified samples rather than formal validation of the full benchmark. These visualizations and coverage measurements do not establish complete chemical or substructure isolation; the quantitative effect on prediction error is reported separately in Table 1. As shown in the top section of Table 1, the MAE under SCOPE-BENCH exceeds the corresponding conventional scaffold-split MAE across all nine evaluated backbone–property combinations, with a mean degradation factor of $5.9\times$.

4.3 OVERALL RESULTS

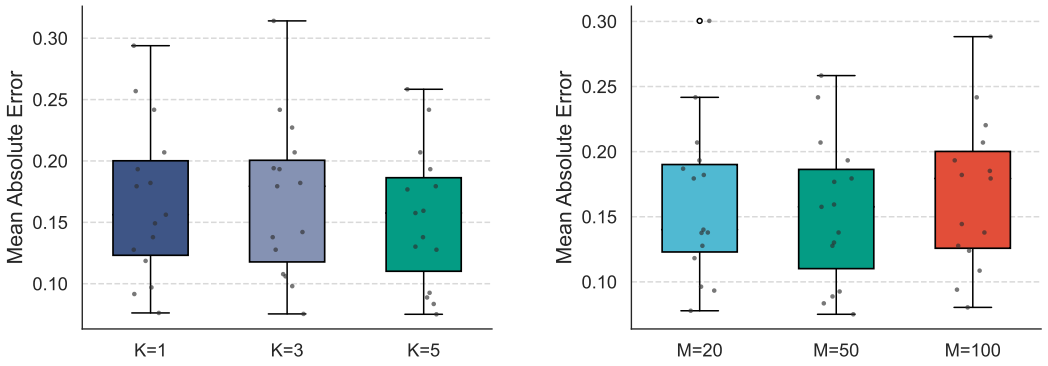
We evaluate POMA on the 15 target-scaffold tasks of SCOPE-BENCH on QM9 using three backbone architectures, with the supervised model trained on all available source data without adaptation as the baseline. Table 1 reports the main results, and Appendix B provides per-scaffold results for each backbone–property combination.

As shown in the bottom section of Table 1, POMA achieves lower reported MAE for all nine evaluated backbone–property combinations. For HOMO, the main-table entries are 5.0%, 8.6%, and 2.8% for ViSNet, ETNN, and GotenNet, respectively. For LUMO, the corresponding entries are 8.5%, 11.2%, and 9.1%. For the HOMO–LUMO gap, the ViSNet and GotenNet entries are 1.4% and 1.8%, respectively; the ETNN–Gap entry in the main table reports a 7.3% relative MAE reduction. Each relative reduction compares POMA with the corresponding strict-OOD supervised entry, rather than with the conventional scaffold-split model. The degradation factors and adaptation gains therefore have different reference points and answer different evaluation questions. The reported results use inherited backbone hyperparameters without backbone-specific tuning.

4.4 ABLATION STUDY

Learned source selection and dual-scale adaptation are the two core components of POMA.

Q1: How does learned source selection compare with the evaluated heuristics? We compare six non-RL heuristics on ViSNet with the full adaptation module fixed. Random selection provides a



(a) MAE versus subset size K at $M = 50$.

(b) MAE versus pool size M at $K = 5$.

Figure 4: Hyperparameter sensitivity on the HOMO task for ViSNet using SCOPE-BENCH. Box plots summarize MAE across the 15 target scaffolds.

baseline without learned source selection. Shallow-feat. uses cosine similarity of graph embeddings after 40 epochs, while Deep-feat. uses embeddings after 200 epochs. Physical ranks by atomic composition and orbital energy penalties. Graph-kernel uses WL similarity, and Mixed-strategy linearly combines WL and physical scores. In Table 2, Deep-feat. has an MAE of 0.1669, compared with 0.1621 for the supervised baseline. POMA has a lower reported mean MAE than the strongest evaluated heuristic (0.1540 vs. 0.1576).

Q2: What is the effect of removing each alignment scale? As shown in the lower block of Table 2, removing Mol-CORAL or Sub-CORAL increases aggregate MAE from 0.1540 for the full model to 0.1689 and 0.1672, respectively; both ablated values are above the supervised baseline of 0.1621. These results indicate that both alignment components contribute in the evaluated ViSNet–HOMO setting. The per-scaffold comparisons show that retaining both alignment scales does not yield the lowest error for every target. For Scaffold 4, removing Sub-CORAL gives an MAE of 0.1473, compared with 0.1576 for the full model (Appendix D).

4.5 HYPERPARAMETER ANALYSIS

The sensitivity of POMA to candidate pool size M and subset size K is evaluated on the HOMO task using the ViSNet architecture. The corresponding box plots are shown in Figures 4a and 4b. The default $K = 5$ setting achieves the lowest MAE among the compared values in this analysis. With $K = 5$ fixed, the default $M = 50$ setting provides the lowest MAE among the evaluated candidate-pool sizes and is used in the main experiments.

The box plots summarize variation across target scaffolds; Appendix E reports the individual values. The default 40 GRPO steps provide the reported performance, while additional steps yield limited improvement at higher cost.

5 CONCLUSION

In this paper, we introduced SCOPE-BENCH for evaluating molecular prediction under descriptor-based scaffold-cluster shifts and developed POMA, a policy-driven pipeline combining source selection with dual-scale covariance alignment. The policy uses proxy-task adaptation feedback to select source subsets for unlabeled target adaptation. On SCOPE-BENCH, the mean degradation factor relative to conventional scaffold splits is $5.9\times$; relative to the supervised fine-tuning baseline, POMA reduces MAE by an average of 6.2% over the nine evaluated backbone–property combinations. A current limitation is the computational cost of the GRPO training loop, which requires executing multiple proxy domain adaptation procedures per policy iteration. In future work, surrogate reward models will be explored to amortize this cost, and the framework will be extended to zero-proxy target scenarios.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation for language editing, formatting, consistency checking, proofreading, and structured review of methodological and experimental descriptions. All research decisions, implementations, experiments, results, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted outputs and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The benchmark construction and evaluation protocol are described in the main paper. Algorithmic and implementation details, hyperparameter settings, ablation results, and complete per-scaffold results are provided in the appendix. Anonymous code is available at the repository linked in the abstract.

REFERENCES

David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pp. 1027–1035, 2007.

Sarp Aykent and Tian Xia. Gotennet: Rethinking efficient 3d equivariant graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015.

Claudio Battiloro, Ege Karaismailoglu, Mauricio Tec, George Dasoulas, Michelle Audirac, and Francesca Dominici. E(n) equivariant topological neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 20267–20309, 2025.

Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Jörg Degen, Christof 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. doi: 10.1002/cmdc.200800178.

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. Pmlr, 2017.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, Houtim Lai, Wei Liu, Junzhou Huang, Shuigeng Zhou, Ping Luo, Peilin Zhao, and Yatao Bian. DrugOOD: Out-of-distribution dataset curator and benchmark for AI-aided drug discovery—a focus on affinity prediction problems with noise annotations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8023–8031, 2023. doi: 10.1609/aaai.v37i7.25970.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Greg Landrum et al. RDKit: Open-source chemoinformatics, 2024. URL `https://doi.org/10.5281/zenodo.12782092`.

Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10490–10512, 2025. doi: 10.1109/TPAMI.2025.3593897.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. To transfer or not to transfer. In *NIPS Workshop on Transfer Learning*, 2005.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yang Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler–lehman graph kernels. *Journal of Machine Learning Research*, 12(77):2539–2561, 2011.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein weisfeiler–lehman graph kernels. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6436–6446, 2019.

Prudencio Tossou, Cas Wognum, Michael Craig, Hadrien Mary, and Emmanuel Noutahi. Real-world molecular Out-Of-Distribution: Specification and investigation. *Journal of Chemical Information and Modeling*, 64(3): 697–711, 2024. doi: 10.1021/acs.jcim.3c01774.

Yusong Wang, Tong Wang, Shaoning Li, Xinheng He, Mingyu Li, Zun Wang, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1):313, 2024.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11293–11302, 2019.

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.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35: 12964–12978, 2022.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

A Training Algorithm and Implementation Details

All experiments are conducted on NVIDIA RTX A6000 (48 GB) GPUs. The computational overhead of POMA can be decoupled into two phases: (1) a one-time offline GRPO policy optimization, requiring approximately 72 GPU hours to converge; and (2) an online target adaptation phase, which leverages the pre-trained policy for source selection and dual-scale alignment. Under the default configuration ($M = 50$, $K = 5$, 40 GRPO steps), the online phase requires approximately 4.5 GPU hours per target scaffold.

Algorithm 1 POMA: Shared-policy Proxy Training and Target-specific Adaptation

Require: Source pool $\mathcal{S}_{pool}$, unlabeled target scaffolds $\mathcal{T}_{real}$, backbone f_θ, policy π_θ
Ensure: Model adapted without target property labels

Compute $\mathrm{HubScore}(s)$ for each $s \in \mathcal{S}_{pool}$ using Equation (3)
Select the top-$N_p = 20$ scaffolds as $\mathcal{T}_{proxy}$
Use one shared policy π_θ across all proxy tasks
for each proxy target $s^* \in \mathcal{T}_{proxy}$ **do**
    Rank candidate domains using $S_{rank}(s^*, c_j)$
    Construct the proxy-specific pool $\mathcal{C}(s^*) = \mathrm{TopM}_{c_j} S_{rank}(s^*, c_j)$, with $M = 50$
    Construct state matrix $\mathbf{X}(s^*)$
    for each GRPO iteration for this proxy task **do**
        Reserve one of the $G = 33$ slots as $\mathbf{a}_0 = \mathbf{0}$, the designated empty control
        Sample $G - 1$ independent Bernoulli proposals $\{\mathbf{b}_1, \ldots, \mathbf{b}_{G-1}\}$
        Apply the $K_{\max} = 5$ cardinality transform to obtain executed subsets $\mathbf{a}_1, \ldots, \mathbf{a}_{G-1}$
        Set $K_i = \|\mathbf{a}_i\|_0$ for each executed proposal
        Evaluate all G slots on the common proxy initialization and evaluation split; a slot is valid only if its evaluator returns a metric
        If the designated control is valid, include it in the runtime-valid set used for group statistics and the implementation-level update; it is a fixed control slot rather than a Bernoulli proposal
        Define MAE_{base} from the designated empty control
        For each valid nonempty slot, compute $R_i = \mathrm{MAE}_{base} - \mathrm{MAE}_i$ after dual-scale adaptation
        Compute group-relative advantages and update the shared π_θ using Equations (10)–(11)
    end for
end for
At deployment, for the current real-target scaffold t, rank candidate domains using $S_{rank}(t, c_j)$
Construct $\mathcal{C}(t) = \mathrm{TopM}_{c_j} S_{rank}(t, c_j)$ and the target-specific state matrix $\mathbf{X}(t)$
Apply the shared policy to $\mathbf{X}(t)$ and select the deterministic top-$K_{\mathrm{deploy}} = 5$ candidates by selection probability
Adapt f_θ on the selected source domains with $\mathcal{L}_{total}$
return the adapted model and its target predictions

B Detailed Performance Analysis across Individual Scaffolds

Table 3 reports per-scaffold MAE for the supervised baseline and POMA across three backbones and three properties. Relative performance varies across scaffolds and backbone–property combinations: POMA has lower MAE in many entries, while the baseline MAE is lower or equal in others. The table therefore provides a granular view of performance heterogeneity across the 15 target scaffolds.

Table 3: Detailed mean absolute error (MAE) on each of the 15 individual target scaffolds of SCOPE-BENCH. For each scaffold and property, the better performance between the supervised baseline and POMA is highlighted in bold.

Target	Method	ViSNet (Wang et al., 2024)			ETNN (Battiloro et al., 2025)			GotenNet (Aykent & Xia, 2025)		
		HOMO↓	LUMO↓	Gap↓	HOMO↓	LUMO↓	Gap↓	HOMO↓	LUMO↓	Gap↓
Scaffold 1	supervised baseline	0.1469	0.1165	0.1941	0.1138	0.0944	0.1555	0.1139	0.0897	**0.1446**
	POMA	**0.1302**	**0.1081**	**0.1625**	**0.1077**	**0.0912**	**0.1455**	**0.1044**	**0.0833**	0.1640
Scaffold 2	supervised baseline	0.1019	**0.0987**	**0.1206**	**0.0663**	**0.0895**	**0.1311**	**0.0748**	**0.0909**	**0.1154**
	POMA	**0.0835**	0.1008	0.1627	0.0706	0.1243	0.1417	0.0774	0.0928	0.1242
Scaffold 3	supervised baseline	0.1277	0.1177	0.1926	**0.1481**	0.1483	**0.1601**	0.1366	0.1819	0.1524
	POMA	0.1277	0.1177	0.1926	0.1515	**0.1301**	0.1886	0.1366	**0.1503**	0.1524
Scaffold 4	supervised baseline	0.2032	0.1890	**0.1894**	0.1035	**0.1626**	0.1875	**0.1321**	**0.1775**	0.2245
	POMA	**0.1576**	**0.1819**	0.1945	**0.0959**	0.1635	**0.1529**	0.1406	0.2160	**0.1792**
Scaffold 5	supervised baseline	0.1794	0.1923	0.1868	0.2205	0.1639	**0.1765**	0.2019	0.3214	0.1966
	POMA	0.1794	0.1923	0.1868	**0.1390**	**0.0922**	0.2103	0.2019	**0.2948**	0.1966
Scaffold 6	supervised baseline	0.1121	0.1629	0.1804	**0.1000**	0.0859	0.1185	0.1077	0.0849	**0.1057**
	POMA	**0.0926**	**0.1333**	**0.1314**	0.1030	**0.0792**	**0.1134**	**0.1011**	**0.0847**	0.1274
Scaffold 7	supervised baseline	0.2697	0.4440	0.7133	0.1704	0.2429	0.4447	0.2333	0.3014	0.5344
	POMA	**0.2584**	**0.3014**	**0.7077**	**0.1436**	**0.2242**	**0.1710**	**0.2173**	0.3014	**0.4900**
Scaffold 8	supervised baseline	0.0901	**0.0871**	**0.1026**	**0.0775**	0.0641	**0.0861**	**0.0825**	0.0669	0.0907
	POMA	**0.0888**	0.0923	0.1050	0.0817	**0.0502**	0.0891	0.0841	**0.0650**	**0.0899**
Scaffold 9	supervised baseline	0.2417	0.1897	0.2893	0.2486	0.1332	**0.2047**	0.3014	0.2630	0.2441
	POMA	0.2417	0.1897	0.2893	**0.2442**	**0.1312**	0.2288	0.3014	**0.1920**	0.2441
Scaffold 10	supervised baseline	0.1379	0.1978	0.2100	**0.1207**	0.1377	0.1674	0.1400	0.1927	0.1331
	POMA	0.1379	0.1978	0.2100	0.1254	**0.1038**	**0.1159**	0.1400	**0.1436**	0.1331
Scaffold 11	supervised baseline	0.1821	0.2451	0.2137	0.1607	0.2050	**0.1496**	0.1960	0.2615	0.2930
	POMA	**0.1768**	0.2451	0.2137	**0.1490**	**0.1253**	0.2421	0.1960	**0.2113**	0.2930
Scaffold 12	supervised baseline	0.1933	0.1821	0.2650	**0.1328**	**0.0981**	**0.1907**	0.2211	0.1474	0.1989
	POMA	0.1933	0.1821	0.2650	0.1415	0.0986	0.1992	0.2211	**0.1276**	0.1989
Scaffold 13	supervised baseline	0.2070	0.1576	0.2254	0.1367	0.1287	**0.1739**	0.2184	0.1414	0.1869
	POMA	0.2070	0.1576	0.2254	**0.1247**	**0.1175**	0.1886	0.2184	**0.1179**	0.1869
Scaffold 14	supervised baseline	**0.0574**	0.0969	**0.0921**	0.0722	0.0862	**0.1059**	0.0661	**0.0854**	0.0891
	POMA	0.0751	**0.0909**	0.1008	**0.0644**	**0.0845**	0.1163	**0.0642**	0.0981	**0.0826**
Scaffold 15	supervised baseline	0.1818	0.2653	0.2799	0.1515	**0.1272**	0.2235	0.1445	0.4472	0.3784
	POMA	**0.1594**	**0.2188**	**0.2581**	**0.1076**	0.1317	**0.2176**	**0.0985**	**0.4138**	**0.3689**

C FEATURE-SPACE DIAGNOSTICS OF THE SPLIT PROTOCOLS

This analysis is a post hoc diagnostic of fixed splits, not a criterion used to construct them. For each protocol-specific diagnostic cache, we sample 12,000 source and 12,000 target molecules without replacement. The source sample is divided into disjoint reference and calibration sets of 6,000 molecules each. We use the same fixed ViSNet checkpoint and its 128-dimensional molecular feature output for all three caches, and normalize each feature vector to unit L_2 norm. The diagnostic caches use row-level identifiers, and their mappings to the formal split counts and molecule identities have not been independently established. The diagnostic therefore describes the specified cache samples and does not by itself verify target-label non-exposure or encoder-training provenance.

Let p index a protocol, and let $\mathcal{R}_p$, $\mathcal{A}_p$, and $\mathcal{T}_p^{\mathrm{diag}}$ denote its source reference, source calibration, and target diagnostic samples. With $\widetilde{h}(x) = h(x)/\|h(x)\|_2$, define

$$d_p(x) = \min_{s \in \mathcal{R}_p} \|\widetilde{h}(x) - \widetilde{h}(s)\|_2, \qquad \rho_{p,q} = \mathrm{Quantile}_q\{d_p(s) : s \in \mathcal{A}_p\}.$$

The corresponding target coverage is

$$\widehat{C}_p(q) = \frac{1}{|\mathcal{T}_p^{\mathrm{diag}}|} \sum_{t \in \mathcal{T}_p^{\mathrm{diag}}} \mathbb{I}[d_p(t) \le \rho_{p,q}] \,.$$

We report $q \in \{0.90, 0.95, 0.99\}$ using linear interpolation to compute quantiles. These are calibration quantiles, not confidence levels. The resulting radii are used only for post hoc diagnostics and do not determine the split. Figure 5 shows the empirical cumulative distributions of $d_p(t)$ on a common absolute distance axis.

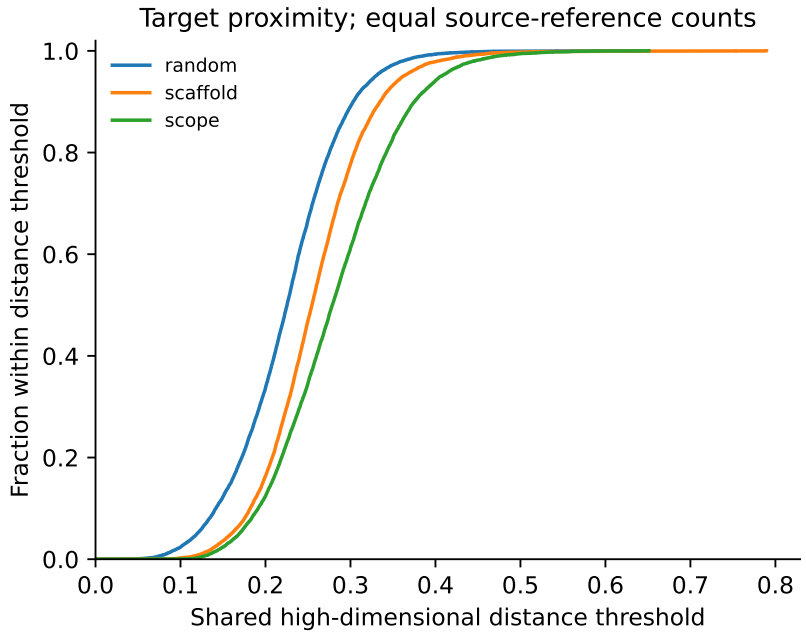


Figure 5: Empirical cumulative distributions of distances from target molecules to their nearest source reference molecules in the shared L2-normalized ViSNet feature space for the diagnostic caches. Each cache uses 12,000 target queries and 6,000 source reference molecules. At any horizontal-axis threshold, the vertical value gives the fraction of target queries with a nearest-source distance at or below that threshold. The horizontal axis is a common distance scale, unlike the cache-specific calibrated radii in Table 4. These descriptive curves do not provide confidence intervals or quantify chemical substructure overlap.

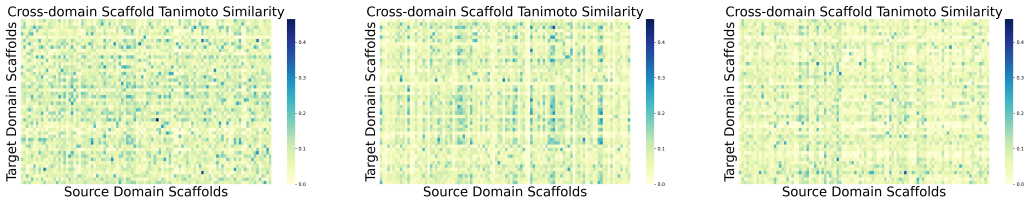
Table 4: Post hoc target coverage in the L2-normalized 128-dimensional ViSNet feature space for the diagnostic caches. Each cache uses 6,000 source reference molecules, 6,000 source calibration molecules, and 12,000 target queries. The radius $\rho_{p,q}$ is calibrated separately for each cache at source-distance quantile q; q is not a confidence level.

Protocol	q	$\rho_{p,q}$	Covered targets	Coverage
Random	0.90	0.30052	10,706 / 12,000	89.22%
Scaffold	0.90	0.29849	9,236 / 12,000	76.97%
SCOPE-BENCH	0.90	0.29129	6,816 / 12,000	56.80%
Random	0.95	0.32693	11,356 / 12,000	94.63%
Scaffold	0.95	0.32076	10,320 / 12,000	86.00%
SCOPE-BENCH	0.95	0.31643	8,323 / 12,000	69.36%
Random	0.99	0.38111	11,867 / 12,000	98.89%
Scaffold	0.99	0.37505	11,550 / 12,000	96.25%
SCOPE-BENCH	0.99	0.38116	10,946 / 12,000	91.22%

At $q = 0.95$, the protocol-specific diagnostic caches give radii of 0.32693, 0.32076, and 0.31643 for random, conventional scaffold, and SCOPE-BENCH, respectively. The corresponding coverage rates are 94.63%, 86.00%, and 69.36%; the median target-to-reference distances are 0.22535, 0.25454, and 0.27861. The same coverage ordering holds at $q = 0.90$ and $q = 0.99$ (Table 4). These are descriptive measurements for the specified samples, representation, and distance metric. They do not establish complete chemical or substructure isolation, invariance to alternative descriptor or clustering choices, or statistical significance of prediction results.

The t-SNE plots in Figure 3 are fitted independently to the concatenated source and target samples for each cache. We use perplexity 55, PCA initialization, learning rate 200, early exaggeration 12, Barnes–Hut angle 0.5, a maximum of 1,000 iterations, and random state 42. All sampled points are plotted with equal marker size and opacity; no local-purity or density mask is applied. Because the embeddings are fitted independently, their coordinates are not interpreted as a common distance scale.

Scaffold similarity. Figure 6 provides the complementary qualitative comparison of inter-domain scaffold Tanimoto similarity under the three splitting protocols.



(a) Random split (b) Standard scaffold split (c) SCOPE-BENCH

Figure 6: Inter-domain scaffold Tanimoto similarity heatmaps. Darker blue indicates higher similarity, whereas lighter yellow indicates lower similarity. The panels show qualitative differences in visible cross-domain similarity patterns across the three splitting protocols.

D DETAILED SOURCE-SELECTION AND ALIGNMENT ABLATIONS

Table 5 compares source-selection heuristics and alignment configurations across 15 target scaffolds. The relative performance of the heuristic strategies varies by target; no heuristic is uniformly best across all scaffolds.

In the aggregate ViSNet–HOMO results, removing Mol-CORAL or Sub-CORAL increases MAE from 0.1540 to 0.1689 and 0.1672, respectively. The full model does not outperform both single-scale variants on every scaffold. For Scaffold 4, removing Sub-CORAL gives an MAE of 0.1473, compared with 0.1576 for the full model.

The full configuration has the lowest aggregate MAE in Table 2, while individual heuristic strategies obtain lower MAE on some scaffolds, including Scaffolds 7 and 15. These results show that the relative performance of selection strategies is target dependent.

Table 5: Detailed ablation study results measured in mean absolute error across fifteen individual target scaffolds for the HOMO property using the ViSNet architecture. The best performance for each target is highlighted in bold. The full framework is denoted as POMA.

Target	Baseline	Heuristic Source Retrieval Strategies						Module Ablation & POMA		
		Random	Shallow	Deep	Physical	Graph-K.	Mixed	w/o Mol	w/o Sub	POMA
Scaffold 1	0.1469	0.1458	0.1470	0.1347	0.1416	0.1364	0.1364	0.1328	0.1359	**0.1302**
Scaffold 2	0.1019	0.0870	0.0939	0.1003	0.0986	0.0897	0.0897	0.0982	0.0872	**0.0835**
Scaffold 3	**0.1277**	0.1618	0.1672	0.1562	0.1573	0.1422	0.1422	**0.1277**	**0.1277**	**0.1277**
Scaffold 4	0.2032	0.1857	0.1683	0.1784	0.1802	0.1628	0.1628	0.1697	**0.1473**	0.1576
Scaffold 5	**0.1794**	0.2001	0.2085	0.2029	0.1937	0.1878	0.1937	**0.1794**	**0.1794**	**0.1794**
Scaffold 6	0.1121	0.1095	0.1079	0.1060	0.1078	0.1005	0.1005	0.1026	0.1028	**0.0926**
Scaffold 7	0.2697	0.2950	0.2593	0.2755	0.2456	**0.2386**	**0.2386**	0.3171	0.3274	0.2584
Scaffold 8	0.0901	0.0886	0.0936	0.0960	0.0916	**0.0871**	**0.0871**	0.0941	0.0995	0.0888
Scaffold 9	0.2417	0.2326	0.2614	**0.2299**	0.2434	0.2336	0.2434	0.2417	0.2417	0.2417
Scaffold 10	**0.1379**	0.1744	0.1896	0.2101	0.1868	0.1972	0.1695	**0.1379**	**0.1379**	**0.1379**
Scaffold 11	0.1821	0.1920	0.1830	0.1948	0.1816	0.1853	0.1816	0.1821	0.1821	**0.1768**
Scaffold 12	0.1933	0.2042	**0.1850**	0.1941	0.1938	0.2131	0.1938	0.1933	0.1933	0.1933
Scaffold 13	0.2070	0.2435	**0.1909**	0.2161	0.1966	0.2216	0.1966	0.2070	0.2070	0.2070
Scaffold 14	**0.0574**	0.0636	0.0597	0.0734	0.0661	0.0635	0.0635	0.0800	0.0775	0.0751
Scaffold 15	0.1818	0.1774	0.1785	**0.1349**	0.1772	0.1652	0.1652	0.2700	0.2611	0.1594

E GRANULAR SENSITIVITY ANALYSIS OF HYPERPARAMETERS

Table 6 reports MAE for each of the 15 target scaffolds under the evaluated hyperparameter configurations, supplementing the aggregate results in the main text.

The per-scaffold results vary across the evaluated candidate-pool sizes. The default $M = 50$ setting has MAE lower than or comparable to that of $M = 20$ and $M = 100$ on most reported scaffolds. On Scaffold 4, $M = 20$ gives lower MAE.

Performance also varies with subset size. A subset of three sources is competitive on some target scaffolds, whereas the default configuration has the lowest prediction error on many evaluated scaffolds. The default setting is retained for the main experiments.

Table 6: Detailed hyperparameter sensitivity analysis measured in mean absolute error across fifteen individual target scaffolds for the HOMO property using the ViSNet architecture. The best performance for each target is highlighted in bold. The default configuration used in the main experiments is shown in the final column.

Target	Candidate Pool Size M		Subset Size K		Default POMA
	$M = 20$	$M = 100$	$K = 1$	$K = 3$	$M = 50$ and $K = 5$
Scaffold 1	0.1376	0.1444	0.1492	0.1421	**0.1302**
Scaffold 2	0.0963	0.1086	0.0970	0.1078	**0.0835**
Scaffold 3	**0.1277**	**0.1277**	**0.1277**	**0.1277**	**0.1277**
Scaffold 4	**0.1401**	0.1852	0.1562	0.1941	0.1576
Scaffold 5	**0.1794**	**0.1794**	**0.1794**	**0.1794**	**0.1794**
Scaffold 6	0.1182	0.1239	0.1186	0.1060	**0.0926**
Scaffold 7	0.3004	0.2883	0.2939	**0.2272**	0.2584
Scaffold 8	0.0933	0.0940	0.0916	0.0980	**0.0888**
Scaffold 9	**0.2417**	**0.2417**	**0.2417**	**0.2417**	**0.2417**
Scaffold 10	**0.1379**	**0.1379**	**0.1379**	**0.1379**	**0.1379**
Scaffold 11	0.1821	0.1821	0.1821	0.1821	**0.1768**
Scaffold 12	**0.1933**	**0.1933**	**0.1933**	**0.1933**	**0.1933**
Scaffold 13	**0.2070**	**0.2070**	**0.2070**	**0.2070**	**0.2070**
Scaffold 14	0.0778	0.0804	0.0762	0.0754	**0.0751**
Scaffold 15	0.1869	0.2203	0.2569	0.3141	**0.1594**

F HYPERPARAMETER SETTINGS

Table 7 summarizes all key hyperparameters used across experiments.

Table 7: Hyperparameter settings for POMA.

Hyperparameter	Symbol	Value
Candidate pool size	M	50
Proxy targets per policy-training stage	N_p	20
Maximum training subset size	$K_{\max}$	5
Deployment subset size	K_{deploy}	5
GRPO group size (including empty control)	G	33
GRPO training steps		40
KL penalty coefficient	β	0.01
Clip threshold	ϵ_{clip}	0.2
HubScore threshold	τ_{sim}	0.45
HubScore balance coefficient	λ	0.3
Warm-up epochs	E_{warm}	10
Regression threshold	τ_{reg}	0.05
Momentum coefficient	β_m	0.9
Numerical stability	ϵ_s	10^{-8}
Backbone learning rate		3×10^{-4}
Batch size		64