THE SHIFT–UTILITY GAP: REPRESENTATION SHIFT DOES NOT DETERMINE PREDICTIVE UTILITY IN 3D MOLECULAR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Three-dimensional (3D) molecular models can exploit geometric information beyond coordinate-independent two-dimensional (2D) topology, but they also inherit a deployment vulnerability: predictions depend on the coordinate realization supplied at inference time. We study a problem distinct from representation-shift detection itself. A strongly shifted 3D representation can still outperform a 2D expert for a particular molecule, while an apparently in-distribution representation need not provide additional predictive value. We call this mismatch the *shift–utility gap*. On OpenADMET-ExpansionRx, end-to-end Uni-Mol falls from macro $R^2 = 0.5292$ on original coordinates to -0.4738 after coordinate randomization; a matched Frad analysis drops from 0.360646 to -3.422540. Across six external regression datasets, clean-versus-random shift is detected almost perfectly (AUROC 0.9957–1.0000), whereas clean sample-level utility discrimination remains near chance (AUROC 0.4430–0.5256) and randomized-condition utility varies substantially (0.4206–0.6539). We introduce **CORE (COordinate-shift-aware Routing of Experts)**, which combines shift evidence with 2D/3D predictions, disagreement, endpoint context, and a bounded 2D-anchored robust path. Under randomized coordinates, CORE reaches macro $R^2 = 0.5710$ versus 0.5465 for endpoint-fixed static mixing (BH-adjusted $q = 0.0133$); under original coordinates static mixing is slightly higher (0.6022 versus 0.5943), but not significantly so ($q = 0.3256$). Full-architecture transfer on Lipo and ESOL shows the same qualitative failure-containment pattern beyond OpenADMET. The main contribution is therefore the empirical distinction between representation shift and relative predictive utility; CORE is one selective-routing response to that distinction.

1 INTRODUCTION

Three-dimensional (3D) molecular representations can encode geometric information unavailable to coordinate-independent two-dimensional (2D) topology (Wu et al., 2018; Rogers & Hahn, 2010; Yang et al., 2019; Schütt et al., 2017; Klicpera et al., 2020; Gasteiger et al., 2021; Zhou et al., 2023). The same coordinate dependence that makes 3D models informative also creates a deployment risk: different coordinate realizations of the same molecular graph can induce different representations and predictions even when molecular identity, endpoint, and target are unchanged.

Inference-time coordinates are not uniquely defined. Practical pipelines depend on distance geometry, conformer sampling, force-field relaxation, random initialization, hydrogen treatment, and software-specific fallbacks (Riniker & Landrum, 2015; Axelrod & Gómez-Bombarelli, 2022; Hamakawa & Miyao, 2025; Qiao et al., 2025). A model can therefore encounter a coordinate distribution different from the one seen during training.

The obvious response is to detect representation shift and suppress the 3D pathway when shift is large. This conflates two tasks. *Representation shift* is observable from the current input or latent representation. *Relative predictive utility* asks whether 3D is more accurate than a coordinate-independent 2D expert for the current molecule and endpoint, which depends on the unknown target. A shifted representation can still yield the lower-error prediction, while an apparently in-distribution representation can add little or hurt.

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.

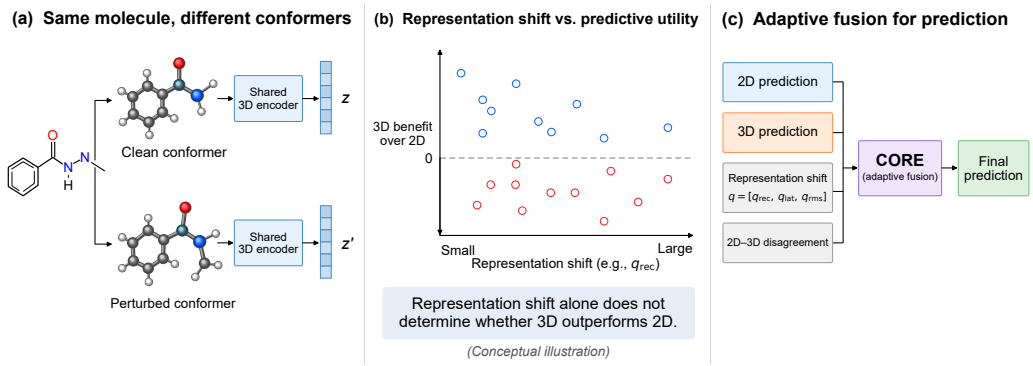


Figure 1: **Problem setting and the shift–utility gap.** The same molecular identity can yield different 3D representations when coordinates change. Representation shift is observable, whereas relative 3D-versus-2D predictive utility is label dependent. In panel (b), blue points denote samples for which the 3D expert has lower squared error than the 2D expert ($\Delta U > 0$), whereas red points denote samples for which the 2D expert is more accurate ($\Delta U < 0$). Their overlap across similar shift levels illustrates that shift magnitude alone does not determine sample-level relative utility.

We call this mismatch the **shift–utility gap**. Figure 1 illustrates the distinction. The claim is not statistical independence: shift can contain utility information in some datasets. Rather, the mapping from shift magnitude to relative 2D-versus-3D utility is not stable enough to serve as a universal sample-level decision rule.

As illustrated in Figure 1, similar levels of representation shift can correspond to opposite 2D-versus-3D utility outcomes. We establish the gap in three steps. First, on OpenADMET-ExpansionRx, end-to-end Uni-Mol falls from macro $R^2 = 0.5292$ on original coordinates to -0.4738 after randomization; matched Frad falls from 0.360646 to -3.422540. Second, on OpenADMET, three shift descriptors remain close to chance for sample-level 2D-versus-3D utility discrimination across five coordinate conditions. Third, across six external regression datasets, clean-versus-random shift detection is nearly perfect, yet clean utility discrimination remains near chance and randomized-condition utility varies substantially. A validation-selected hard threshold on q_{rec} consequently performs poorly under clean OpenADMET inputs.

Motivated by this evidence, we introduce **CORE (COordinate-shift-aware Routing of Experts)**. CORE combines a coordinate-independent 2D pathway, a coordinate-sensitive 3D pathway, a bounded 2D-anchored robust pathway, and a learned router using shift descriptors together with expert predictions, disagreement, and endpoint context. The router receives neither coordinate-condition labels nor perturbation magnitudes at inference time.

Our contributions are:

1. We identify the shift–utility gap: representation shift and sample-level relative predictive utility are distinct deployment quantities in conformation-sensitive molecular prediction.
2. We validate this distinction on OpenADMET and six external regression datasets and reproduce severe coordinate fragility with both Uni-Mol and Frad.
3. We show that a validation-selected shift-only fallback is not a reliable expert-selection policy.
4. We introduce CORE as a selective-routing response, identify its robustness trade-off against static fusion, and reproduce the same qualitative failure-containment pattern with full-architecture transfer on Lipo and ESOL.

2 Related Work

2D and 3D molecular representations. Morgan fingerprints and learned 2D graph models provide coordinate-independent representations (Rogers & Hahn, 2010; Yang et al., 2019), whereas geometric neural networks and pretrained 3D encoders exploit distances, directions, or coordinate-dependent

attention (Schütt et al., 2017; Klicpera et al., 2020; Gasteiger et al., 2021; Zhou et al., 2023; Feng et al., 2023). Conformation-aware pretraining and multimodal 2D–3D learning largely seek stronger representations or better modality integration (Liu et al., 2022; Qiao et al., 2025; Jing et al., 2025). Our focus is different: once a coordinate-sensitive representation is deployed, how should its contribution be regulated when the coordinate realization changes?

Distribution shift and uncertainty. Out-of-distribution detection asks whether a test representation departs from a reference distribution (Lee et al., 2018; Guan et al., 2023); predictive-uncertainty methods characterize uncertainty in the resulting prediction (Scalia et al., 2020; Hirschfeld et al., 2020). Our shift descriptors are narrower: they summarize how unusual the frozen 3D representation is relative to fold-specific clean training representations. We do not interpret them as calibrated uncertainty, conformer quality, or prediction correctness.

Mixture-of-experts routing. Mixture-of-experts methods combine specialized predictors through learned gates (Jacobs et al., 1991; Jordan & Jacobs, 1994; Shazeer et al., 2017). CORE applies this framework to a reliability problem in which a coordinate-sensitive expert has sample-dependent utility. The key design choice is to supply shift evidence jointly with prediction-level evidence rather than convert shift magnitude into a hard accept/reject rule.

3 Problem Formulation

Let i, k, and c index molecule, endpoint, and coordinate realization, and let y_{ik} be invariant to c. With coordinate-independent molecular input x_i, atom identities A_i, and coordinates $X_i^{(c)}$,

$$\hat{y}_{ik}^{2D} = f_{2D}(x_i, k), \qquad z_i^{(c)} = E_{3D}(A_i, X_i^{(c)}), \qquad \hat{y}_{ikc}^{3D} = f_{3D}(z_i^{(c)}, k). \tag{1}$$

Changing coordinates therefore leaves the 2D pathway unchanged but can alter both the 3D representation and its prediction.

For analysis, define the post-hoc relative 3D advantage

$$\Delta U_{ikc} = \left(y_{ik} - \hat{y}_{ik}^{2D}\right)^2 - \left(y_{ik} - \hat{y}_{ikc}^{3D}\right)^2. \tag{2}$$

Positive ΔU means the 3D expert has lower squared error; negative ΔU favors 2D. Because ΔU depends on the unknown target, it is unavailable at inference and is used only for analysis.

By contrast, a representation-shift vector $q_i^{(c)}$ can be computed from the current 3D representation relative to a fold-specific clean-training reference:

$$\underbrace{q_i^{(c)}}_{\text{observable shift evidence}} \quad \neq \quad \underbrace{\Delta U_{ikc}}_{\text{label-dependent relative utility}} \quad . \tag{3}$$

The inequality denotes different roles rather than independence. A deployable rule must act on observable quantities such as endpoint identity, shift descriptors, expert predictions, and disagreement.

4 CORE: A Selective-Routing Response

4.1 Experts and Shift Evidence

Figure 2 summarizes the full routing architecture. The 2D expert uses a radius-2, 2048-bit Morgan fingerprint (Rogers & Hahn, 2010); the 3D expert uses a frozen 512-dimensional Uni-Mol representation (Zhou et al., 2023). Both receive a learned 32-dimensional endpoint embedding e_k. Within each outer fold, 3D embeddings are standardized using original-coordinate inner-training statistics and clipped to $[-10, 10]$. The 3D prediction expert itself is trained only on original-coordinate training conformers.

Shift evidence is summarized by

$$q_i^{(c)} = [q_{\text{rec}}, q_{\text{lat}}, q_{\text{rms}}],$$

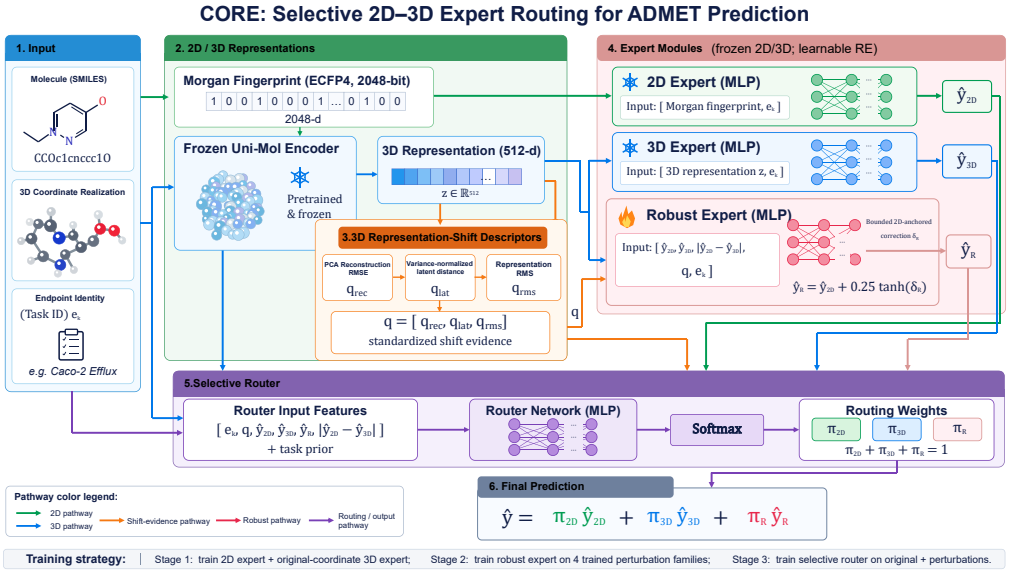


Figure 2: **CORE architecture.** A coordinate-independent 2D path and a coordinate-sensitive 3D path are combined with shift evidence, a bounded 2D-anchored robust path, and a sample-dependent router.

derived from PCA reconstruction RMSE, variance-normalized PCA latent distance, and RMS representation magnitude. The PCA basis and both scalers are fitted exclusively on original-coordinate inner-training data. Exact definitions and optimization settings are given in Appendix D and Table 5. None of the descriptors is assumed to estimate ΔU directly.

4.2 Bounded Robust Path and Router

A third robust pathway is anchored to the 2D prediction:

$$\hat{y}^R_{ikc} = \hat{y}^{2D}_{ik} + 0.25 \tanh(\delta_{R,ikc}), \qquad |\hat{y}^R_{ikc} - \hat{y}^{2D}_{ik}| \le 0.25. \tag{4}$$

Its input contains the shift vector, the 2D and 3D predictions, their absolute disagreement, and endpoint context. The bound limits how far this auxiliary path can depart from the coordinate-independent prediction on the standardized target scale; it is not a bound on final prediction error.

The router receives

$$s_{ikc} = [e_k, q^{(c)}_i, \hat{y}^{2D}_{ik}, \hat{y}^{3D}_{ikc}, \hat{y}^R_{ikc}, |\hat{y}^{2D}_{ik} - \hat{y}^{3D}_{ikc}|]$$

and produces

$$\boldsymbol{\pi}_{ikc} = \mathrm{softmax}(g_\theta(s_{ikc}) + a_k), \qquad \hat{y}_{ikc} = \pi^{2D}\hat{y}^{2D}_{ik} + \pi^{3D}\hat{y}^{3D}_{ikc} + \pi^R\hat{y}^R_{ikc}. \tag{5}$$

Unlike a hard threshold, the router can use q together with expert outputs, disagreement, and endpoint context.

The bounded robust path and the consistency term are *safety-oriented inductive constraints*, not components claimed to independently raise mean accuracy. The ablations show no significant aggregate loss when either is removed. We therefore interpret the bounded path as a smooth, explicitly limited fallback around the 2D prediction, while consistency discourages unnecessary prediction variation across matched coordinate realizations.

4.3 Three-Stage Training

Stage 1 trains the coordinate-independent 2D expert and the original-coordinate 3D expert. Stage 2 freezes both and trains the bounded robust expert on the four artificial perturbation families $\mathcal{P} = \{0.5\ \text{Å}, 1.0\ \text{Å}, 2.0\ \text{Å}, \text{random}\}$. Stage 3 freezes all experts and trains only the router.

The training loader is constructed from *observed molecule–endpoint records*: for endpoint k, a molecule i contributes a row only when y_{ik} is finite. Thus a minibatch $\mathcal{B}$ contains labeled pairs

(i, k) rather than molecules with a dense nine-endpoint target vector. Let y_{ik} denote the endpoint-standardized target for such a row. The router losses implemented in code are

$$\mathcal{L}_{\text{orig}} = \frac{1}{|\mathcal{B}|} \sum_{(i,k)\in\mathcal{B}} \left(\hat{y}_{ik}^{(0)} - y_{ik}\right)^2, \tag{6}$$

$$\mathcal{L}_{\text{pert}} = \frac{1}{|\mathcal{B}|\,|\mathcal{P}|} \sum_{(i,k)\in\mathcal{B}} \sum_{c\in\mathcal{P}} \left(\hat{y}_{ik}^{(c)} - y_{ik}\right)^2, \tag{7}$$

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{B}|\,|\mathcal{P}|} \sum_{(i,k)\in\mathcal{B}} \sum_{c\in\mathcal{P}} \left(\hat{y}_{ik}^{(c)} - \hat{y}_{ik}^{(0)}\right)^2. \tag{8}$$

The optimized objective is

$$\mathcal{L} = \mathcal{L}_{\text{orig}} + 0.25\,\mathcal{L}_{\text{pert}} + 0.05\,\mathcal{L}_{\text{cons}}. \tag{9}$$

Within a minibatch, every observed molecule–endpoint record has equal weight; there is no additional per-endpoint reweighting. Consequently, endpoints with more observed training labels contribute proportionally more records to the optimization objective, even though evaluation macro-R^2 later gives all nine endpoints equal weight. The perturbation conditions are equally weighted inside $\mathcal{L}_{\text{pert}}$ and $\mathcal{L}_{\text{cons}}$. Because the loader contains only finite-label rows, the consistency term is also evaluated only for molecule–endpoint pairs with an observed target, although its algebraic form does not directly use y_{ik}.

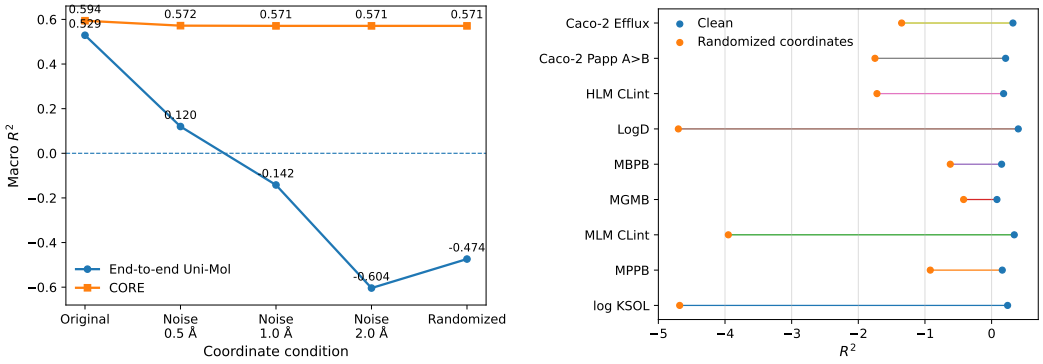
Stage 2 follows the same record-level convention: the robust expert minimizes MSE over observed molecule–endpoint rows for each non-clean condition, and the four condition losses are averaged equally. The consistency term in Stage 3 regularizes matched routed predictions, not routing weights. The endpoint-fixed static-mixture baseline uses the same frozen experts, labeled rows, original/perturbed curriculum, and objective, but replaces the sample-dependent router by learned endpoint-specific mixture logits; this isolates adaptive routing from differences in curriculum or expert training. Because the four artificial perturbation families participate in Stages 2–3, those results are trained-family stress tests on held-out molecules, not zero-shot corruption generalization. ETKDGv3 and MMFF-lowest conformers are held out from training and model selection.

5 Experimental Setup

OpenADMET benchmark. We use OpenADMET-ExpansionRx (OpenADMET Consortium, 2026). After canonical-SMILES processing and endpoint-specific preprocessing, the modeling table contains 7,608 unique molecules and nine regression endpoints: LogD, KSOL, HLM CLint, MLM CLint, Caco-2 Papp A>B, Caco-2 Efflux, MPPB, MBPB, and MGMB. Positive KSOL, clearance, and Caco-2 targets are modeled on the $\log_{10}$ scale; LogD and the three percent-unbound endpoints retain their provided scales. Missing labels are masked. Appendix Table 4 lists the endpoint-specific modeling scales.

Coordinate conditions. The fixed original-coordinate archive is constructed before perturbation or downstream representation extraction. Artificial stress tests alter only Cartesian coordinates while preserving molecule identity, atom identity/order, connectivity, labels, and scaffold membership. For each molecule, one Gaussian direction is reused and scaled by $\sigma \in \{0.5, 1.0, 2.0\}$ Å; randomized coordinates use an independent stream with components sampled from $\mathcal{U}(-5, 5)$ Å. Three ETKDGv3 alternatives (Riniker & Landrum, 2015) and one MMFF-lowest alternative (Halgren, 1996) are excluded from expert/router training, shift-transform fitting, checkpoint selection, and threshold tuning.

External validation. We additionally evaluate Lipo (N = 4,200), ESOL (1,117), FreeSolv (642), Caco2 (901), VDss (1,111), and LD50 (7,341). Each uses a coordinate-independent Morgan-MLP 2D expert and a clean-trained frozen-Uni-Mol MLP 3D expert under five fixed scaffold folds, each evaluated with model-initialization seeds 42, 123, and 2024. Clean coordinates use a fixed RDKit ETKDGv3 plus MMFF/UFF protocol; randomized coordinates preserve identity while drawing coordinates from $\mathcal{U}(-5, 5)$ Å. All six datasets test transfer of the shift–utility relationship. We additionally retrain full CORE on Lipo and ESOL using the archived external expert pipeline and the same 15 seed–fold blocks. Because only clean and randomized external embeddings were archived,



(a) OpenADMET coordinate-stress response.

(b) Matched Frad clean vs. randomized R^2.

Figure 3: **Coordinate fragility in two 3D encoders.** (a) End-to-end Uni-Mol degrades sharply under coordinate perturbation, whereas CORE remains stable across the trained stress families. (b) Frad shows the same qualitative fragility under randomized coordinates.

the external router objective uses $\mathcal{P}_{\text{ext}} = \{\text{random}\}$, i.e., $\mathcal{L}_{\text{clean}} + 0.25\mathcal{L}_{\text{random}} + 0.05\mathcal{L}_{\text{cons}}$. The robust path and router retain the corresponding Table 5 settings; the archived 2D/3D expert-head training settings are preserved for exact baseline alignment. Full implementation details and dataset preparation are given in Appendix J and Table 9.

Evaluation and statistical testing. We use five fixed Bemis–Murcko outer scaffold folds (Bemis & Murcko, 1996). Each outer fold is evaluated under three model-initialization seeds (42, 123, and 2024), yielding 15 seed–fold evaluation blocks; the outer-test membership for a given fold is shared across the three seeds. Each outer-training fold contains a scaffold-aware inner validation split. Scaling, PCA fitting, reliability standardization, target normalization, early stopping, model selection, and threshold selection use training/validation data only. For OpenADMET, R^2 is first computed separately for each endpoint on its available outer-test labels after inversion of fold-specific target standardization. The primary macro R^2 for a seed–fold–condition block is the unweighted arithmetic mean of the nine endpoint-level R^2 values, so endpoints contribute equally regardless of their sample counts. Reported OpenADMET means and standard deviations summarize these 15 seed–fold macro values. Paired OpenADMET comparisons use a Nadeau–Bengio-style corrected paired test (Nadeau & Bengio, 2003) with Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) within predefined comparison families. External Lipo/ESOL are single-task datasets, so no endpoint averaging is applied; their reported means and standard deviations are over the 15 seed–fold R^2 values.

6 Results

6.1 Coordinate Shift Can Cause Predictive Collapse

Figure 3 summarizes the coordinate-stress response of Uni-Mol and the matched Frad validation. End-to-end Uni-Mol falls from macro $R^2 = 0.5292$ on original coordinates to 0.1205, −0.1423, −0.6040, and −0.4738 under 0.5, 1.0, 2.0 Å perturbations and randomized coordinates, respectively.

A matched Frad experiment (Feng et al., 2023) shows that the phenomenon is not unique to Uni-Mol. Frad macro R^2 falls from 0.360646 to −3.422540 between matched original and randomized records. All 135 endpoint–seed–fold blocks degrade, all 15 seed–fold macro comparisons agree in direction, and the corrected two-sided test gives $p = 1.09 \times 10^{-7}$. We use this as second-encoder failure-mode validation, not as evidence of encoder-agnostic CORE efficacy.

Population collapse motivates reliability control, but routing requires the harder question of whether shift magnitude predicts *relative* 2D-versus-3D accuracy.

Table 1: **OpenADMET sample-level utility AUROC.** Positive class is "2D has lower squared error." Values are means over 15 seed–fold macro blocks.

Condition	q_{lat}	q_{rec}	q_{rms}
Original	0.5087	0.4994	0.5024
Noise 0.5 Å	0.5028	0.5061	0.4981
Noise 1.0 Å	0.4986	0.5066	0.4977
Noise 2.0 Å	0.4898	0.5005	0.4956
Randomized	0.5097	0.5148	0.5135

Table 2: **External validation of shift detection and sample-level relative utility.** Shift AUROC is the range across three descriptors. Random-condition utility uses raw descriptors to avoid standardized-score saturation.

Dataset	N	2D R^2	3D clean R^2	3D random R^2	Shift AUROC	Clean utility	Random utility
Lipo	4200	0.4493	0.3026	−2.3024	0.9996–1.0000	0.4966–0.5012	0.4529–0.4913
ESOL	1117	0.3943	0.7306	−1.8464	0.9975–0.9999	0.4671–0.4935	0.6212–0.6539
FreeSolv	642	0.5174	0.7373	−0.4688	0.9967–1.0000	0.4430–0.4524	0.4891–0.5039
Caco2	901	0.4808	0.3949	−2.7734	0.9961–0.9991	0.5065–0.5256	0.4206–0.5111
VDss	1111	0.3551	0.3218	−1.4411	0.9957–0.9980	0.4713–0.4805	0.5207–0.5556
LD50	7341	0.3655	0.3431	−2.5973	0.9982–0.9999	0.5125–0.5156	0.4572–0.4745

6.2 Shift Detection Transfers, but Utility Mapping Does Not

On OpenADMET, q_{lat}, q_{rec}, and q_{rms} yield utility AUROC 0.5087, 0.4994, and 0.5024 under original coordinates and 0.5097, 0.5148, and 0.5135 after randomization; intermediate perturbations show the same near-chance pattern (Table 1). Across all five conditions, the largest absolute mean Spearman correlation with ΔU is 0.0407. Raw-descriptor controls are nearly identical, so clipping does not explain the result.

Across the six external datasets, clean-versus-random shift is separable with AUROC 0.9957–1.0000, yet clean utility AUROC is only 0.4430–0.5256 and randomized-condition raw-descriptor utility AUROC spans 0.4206–0.6539 (Table 2). ESOL is an important counterexample to an overly strong interpretation: under randomization its utility AUROC reaches 0.6212–0.6539, so shift can contain nontrivial utility information. Other datasets are near chance or reverse orientation. The supported conclusion is therefore that representation shift does not *reliably determine* predictive utility across datasets and conditions, not that the quantities are independent.

Despite severe random-coordinate population degradation, 3D still has lower squared error than 2D for 16.13%–26.52% of molecules (mean 21.54%), so population failure does not imply a uniform sample-level decision.

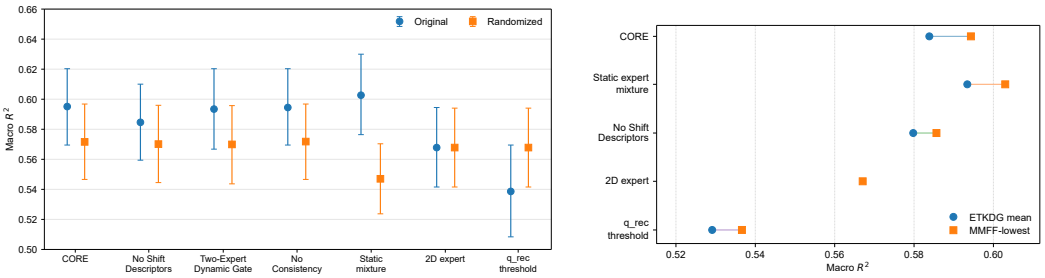
6.3 A Shift-Only Hard Fallback Is Not a Reliable Routing Policy

For every seed–fold–endpoint block, a threshold τ is selected on validation data; 3D is used when $q_{\mathrm{rec}} \leq \tau$, otherwise prediction falls back to 2D. Under original coordinates, this rule reaches macro $R^2 = 0.538223$, below full CORE (0.594331, BH-adjusted $q = 0.000340$) and the 2D expert (0.567139, $q = 0.025533$). Only 29.79% of clean samples are accepted into the 3D branch; among accepted samples, 56.24% have larger 3D squared error than 2D, while 42.39% of fallback samples would have been predicted better by 3D.

Under randomization, fallback becomes nearly universal and the threshold model is effectively No3D ($R^2 = 0.567139$). CORE reaches 0.570981, but the difference is not significant ($q = 0.155029$). The hard threshold therefore fails most clearly when 3D usefulness is heterogeneous and shift magnitude alone does not identify the better expert.

6.4 Dynamic Routing Provides a Robustness Trade-off

Figure 4 visualizes the clean/randomized trade-off and held-out conformer results, while Table 3 summarizes the key corrected comparisons.



(a) Original vs. randomized ablations.

(b) Held-out ETKDG/MMFF conformers.

Figure 4: **Where selective routing helps.** (a) Static mixing is slightly higher than CORE on original coordinates but significantly worse after randomization; removing shift descriptors reduces clean performance. (b) Static mixing remains numerically competitive on held-out alternative conformers, while CORE stays above the 2D expert. Detailed values and corrected tests are in the appendix.

Table 3: **Key corrected comparisons.** Positive ΔR^2 favors CORE unless noted.

Comparison	Condition	ΔR^2	BH q
CORE vs. 2D expert	Original	+0.0272	0.00093
CORE vs. Static mixture	Original	−0.0079	0.32559
CORE vs. Static mixture	Randomized	+0.0245	0.01330
CORE vs. q_{rec} fallback	Original	+0.0561	0.00034

Under original coordinates, CORE reaches macro $R^2 = 0.594331$ and exceeds the 2D expert by +0.027192 (BH-adjusted $q = 0.000930$). Static mixing is numerically higher at 0.602184, but its difference from CORE is not significant ($q = 0.325591$). Under randomized coordinates, CORE reaches 0.570981 versus 0.546518 for static mixing; the paired difference of +0.024463 is significant after correction ($q = 0.013301$). Thus, the clearest measured advantage of sample-dependent routing is under severe randomized-coordinate stress.

Removing shift descriptors lowers original-coordinate performance from 0.594331 to 0.583972 ($q = 0.001659$), supporting reliability information as router context; the randomized difference is not significant ($q = 0.716582$). Removing the bounded robust path or consistency term yields no significant independent loss, so neither is claimed as a separately established accuracy driver.

Full-CORE transfer beyond OpenADMET. Using the aligned external pipelines and the same 15 seed–fold blocks, Lipo gives clean/random $R^2 = 0.4839/0.4509$ for CORE, compared with 0.4831/0.4432 for static mixing and 0.4493/0.4493 for the 2D expert; the 3D expert collapses from 0.3026 to −2.3024. On ESOL, CORE gives 0.7202/0.3467, compared with 0.6541/0.1993 for static mixing and 0.3943/0.3943 for 2D, while 3D changes from 0.7306 to −1.8464. The bounded robust path alone reaches 0.4461 on randomized ESOL, higher than full CORE, reinforcing its interpretation as a conservative fallback rather than a universally accuracy-improving component. These cross-dataset runs therefore support transfer of the selective-fusion behavior without implying universal dominance of the complete architecture. Appendix Table 10 reports all external full-CORE values.

6.5 Transfer to Held-Out Alternative Conformers

ETKDGv3 and MMFF-lowest are excluded from expert/router training, shift-transform fitting, checkpoint selection, and threshold tuning. Across the three ETKDG alternatives, CORE reaches macro $R^2 = 0.5839$; on MMFF-lowest it reaches 0.5942. Relative to 2D, the ETKDG gain is +0.016771 with BH-adjusted $q = 0.01579$ and 14/15 paired directions favoring CORE. Static mixing remains numerically higher on ETKDG mean (0.5934 versus 0.5839) and MMFF-lowest (0.6027 versus 0.5942), but these differences are not significant after correction ($q \approx 0.1509$). Thus,

CORE retains gains over 2D on held-out conformers, while its clearest advantage over static fusion remains severe randomization.

7 Discussion and Limitations

The central result is a separation between two deployment tasks. Shift detection asks whether the 3D representation is atypical relative to a clean reference; relative utility asks which expert has lower error for the current sample. Across six external datasets, clean-versus-random shift is detected with AUROC 0.9957–1.0000, whereas clean utility is near chance and randomized-condition utility varies substantially. Shift evidence can therefore be useful context without serving as a stable stand-alone utility oracle.

This explains why CORE combines shift descriptors with expert outputs, disagreement, and endpoint context. The hard-threshold audit shows that many accepted 3D predictions are worse than 2D while many rejected 3D predictions would have been better. The measured benefit of dynamic routing is conditional: static mixing remains competitive on original and held-out alternative conformers, whereas CORE becomes significantly stronger under severe OpenADMET randomization. The Lipo and ESOL full-architecture transfers show that the same qualitative failure-containment behavior is not restricted to the original benchmark: on Lipo, CORE nearly returns to the 2D safety level under randomization, while on ESOL it substantially limits the collapse of both 3D and static fusion.

The ablations also clarify the role of the bounded robust path and consistency loss. Neither shows a significant independent OpenADMET gain, so we treat them as safety-oriented architectural constraints rather than separately established performance drivers. The ESOL random-coordinate result, where the robust path alone exceeds full CORE, is consistent with a conservative fallback role.

Several limitations remain. The four artificial OpenADMET perturbation families participate in router training; randomized coordinates are intentionally severe; external clean conformers are computationally generated; full-CORE transfer is presently limited to Lipo and ESOL and uses a clean/random external curriculum rather than the full OpenADMET perturbation curriculum; and Frad validates a second-encoder failure mode rather than encoder-agnostic CORE transfer. Future work should test experimental conformer ensembles, additional encoders, realistic geometry shifts, and routing objectives that estimate relative expert utility more directly from observable context.

8 Conclusion

We identify a shift–utility gap in conformation-sensitive molecular prediction: observable representation shift can diagnose severe 3D distributional degradation without providing a universally reliable answer to which expert is more accurate for an individual molecule. Across six external datasets, shift AUROC reaches 0.9957–1.0000, whereas sample-level utility is weak under clean inputs and dataset dependent under randomization. CORE responds by combining shift evidence with prediction-level information for learned 2D–3D routing. Its clearest OpenADMET advantage over static mixing occurs under randomized-coordinate stress (0.5710 versus 0.5465 macro R^2, $q = 0.0133$), while static fusion remains competitive under original and held-out alternative conformers. Full-architecture transfer on Lipo and ESOL reproduces the same qualitative containment pattern beyond OpenADMET, although the external stress curriculum is narrower. These results support selective routing as a failure-containment strategy when geometry-dependent reliability changes across inputs, not as a universal replacement for static fusion.

AI Use Statement

Generative AI tools were used as an aid to improve clarity and style in the writing of this paper and to assist with the design of the figures. All AI-assisted content was reviewed and verified by the authors.

Reproducibility Statement

Data preprocessing, split construction, coordinate generation, model training, shift descriptors, threshold selection, alternative conformers, external validation, and statistical inference are documented in the appendix and anonymous supplementary artifact.

REFERENCES

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9:185, 2022. doi: 10.1038/s41597-022-01288-4.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39:2887–2893, 1996. doi: 10.1021/jm9602928.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, and Wei-Ying Ma. Fractional denoising for 3D molecular pre-training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9938–9961. PMLR, 2023.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xiaoyuan Guan, Zhouwu Liu, Wei-Shi Zheng, Yuren Zhou, and Ruixuan Wang. Revisit PCA-based technique for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19431–19439, 2023. doi: 10.1109/ICCV51070.2023.01780.

Thomas A. Halgren. Merck molecular force field. I. basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5–6):490–519, 1996. doi: 10.1002/(SICI)1096-987X(199604)17:5/6⟨490::AID-JCC1⟩3.0.CO;2-P.

Y. Hamakawa and T. Miyao. Understanding conformation importance in data-driven property prediction models. *Journal of Chemical Information and Modeling*, 65:3388–3404, 2025. doi: 10.1021/acs.jcim.5c00018.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60:3770–3780, 2020. doi: 10.1021/acs.jcim.0c00502.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Z. Jing, Y. Sun, Y. Y. Li, S. Janarthanan, A. Deng, and P. Hu. Structure-aware fusion with progressive injection for multimodal molecular representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0986.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shengchao Liu, Hanchen 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.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52: 239–281, 2003. doi: 10.1023/A:1024068626366.

OpenADMET Consortium. OpenADMET-ExpansionRx Challenge FULL Dataset. Hugging Face and Zenodo, 2026. Hugging Face DOI: 10.57967/hf/9687; Zenodo DOI: 10.5281/zenodo.21504733.

J. Qiao, J. Jin, D. Wang, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. *Nature Communications*, 16:4382, 2025. doi: 10.1038/s41467-025-59634-0.

Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55: 2562–2574, 2015. doi: 10.1021/acs.jcim.5b00654.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 2010. doi: 10.1021/ci100050t.

Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60:2697–2717, 2020. doi: 10.1021/acs.jcim.9b00975.

Kristof T. 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. In *Advances in Neural Information Processing Systems*, volume 30, pp. 991–1001, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, et al. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9:513–530, 2018. doi: 10.1039/C7SC02664A.

Kevin Yang, Kyle Swanson, Wengong Jin, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59:3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, et al. Uni-Mol: A universal 3D molecular representation learning framework. In *International Conference on Learning Representations*, 2023.

A Dataset and Target Preprocessing

OpenADMET-ExpansionRx is an industry-derived multi-endpoint ADMET regression dataset released publicly after the challenge (OpenADMET Consortium, 2026). We canonicalize SMILES, aggregate duplicate canonical SMILES by endpoint-wise median, and then apply endpoint-specific transformations. Missing endpoint labels are masked throughout training and evaluation. Target normalization is fitted on inner-training labels only and inverted before evaluation. Table 4 summarizes the endpoint transformations.

Table 4: **Endpoint transformations used in the OpenADMET modeling table.**

Endpoint	Modeling scale
LogD	Original provided value (already logarithmic)
KSOL	$\log_{10}$ for positive values
HLM CLint	$\log_{10}$ for positive values
MLM CLint	$\log_{10}$ for positive values
Caco-2 Papp A>B	$\log_{10}$ for positive values
Caco-2 Efflux	$\log_{10}$ for positive values
MPPB	Original % Unbound
MBPB	Original % Unbound
MGMB	Original % Unbound

B Coordinate Construction and Alignment

The clean coordinate archive is constructed from an MMFF conformer SDF and fixed before artificial perturbation or downstream representation extraction. For each molecule, one standard-normal noise matrix $\epsilon \sim \mathcal{N}(0, I)$ is generated and reused across the three Gaussian levels:

$$X^{(\sigma)} = X^{\text{clean}} + \sigma\epsilon, \qquad \sigma \in \{0.5, 1.0, 2.0\}\ \text{Å}.$$

Thus, the three Gaussian conditions share perturbation direction and differ only in magnitude. Randomized-coordinate inputs use an independent random stream and sample each Cartesian component independently from $\mathcal{U}(-5, 5)$ Å. Atom identity, atom ordering, molecular connectivity, target labels, and scaffold membership remain unchanged across artificial conditions.

Every molecule–condition coordinate record is generated once and stored. All downstream systems consume the same aligned records. The end-to-end Uni-Mol baseline consumes the coordinates directly; CORE consumes frozen 512-dimensional Uni-Mol embeddings extracted from those same coordinates.

C Scaffold Splits and Leakage Control

We use five fixed Bemis–Murcko outer scaffold folds. Each fold is rerun with model-initialization seeds 42, 123, and 2024, giving 15 seed–fold evaluation blocks but only five distinct outer-test partitions. Within every outer-training partition, a scaffold-aware inner split assigns approximately 10% of molecules to validation. Outer-test molecules do not participate in feature scaling, PCA fitting, representation-shift standardization, target normalization, early stopping, model selection, router-prior initialization, or q-threshold selection. All fold-specific transforms are fitted on inner-training data and then frozen for validation, outer-test data, and all coordinate conditions.

For the formal external shift–utility experiments, acyclic molecules with empty Murcko scaffolds are distributed across folds using molecule-specific fallback keys rather than being collapsed into one artificial super-scaffold. The saved external fold files form complete, non-overlapping test partitions.

D Full Architecture and Shift-Descriptor Details

Let $\tilde{z}_i^{(c)}$ denote the fold-standardized 512-dimensional 3D representation. A 64-component PCA basis is fitted only on original-coordinate inner-training representations. We compute three raw

descriptors:

$$r_{\mathrm{rec},i}^{(c)} = \sqrt{\frac{1}{d}\sum_j \left(\tilde{z}_{ij}^{(c)} - \hat{\tilde{z}}_{ij}^{(c)}\right)^2}, \tag{10}$$

$$r_{\mathrm{lat},i}^{(c)} = \sqrt{\frac{1}{m}\sum_\ell \frac{\left(u_{i\ell}^{(c)}\right)^2}{\lambda_\ell}}, \tag{11}$$

$$r_{\mathrm{rms},i}^{(c)} = \sqrt{\frac{1}{d}\sum_j \left(\tilde{z}_{ij}^{(c)}\right)^2}. \tag{12}$$

Here $\hat{\tilde{z}}$ is the PCA reconstruction, u is the PCA latent representation, and λ_ℓ is the corresponding explained variance. The raw descriptor triple is standardized by a second scaler fitted only on original-coordinate inner-training descriptors and clipped to $[-10, 10]$, yielding

$$q_i^{(c)} = [q_{\mathrm{rec}}, q_{\mathrm{lat}}, q_{\mathrm{rms}}].$$

Raw descriptors are retained for analyses in which standardized scores saturate under extreme randomization. Table 5 lists the architecture and optimization settings.

Table 5: **CORE architecture and optimization settings on OpenADMET.** External full-CORE transfer preserves these robust-path/router settings while retaining the archived external expert-head optimizer settings described in Appendix J.

Component	Setting
2D input	radius-2, 2048-bit Morgan fingerprint
Endpoint embedding	32 dimensions
3D encoder	frozen Uni-Mol, 512-dimensional output
PCA components	64
2D / 3D expert hidden size	256
Robust expert / router hidden size	128
Embedding clipping	$[-10, 10]$ after fold-specific standardization
Expert learning rate	10^{-3}
Router learning rate	5×10^{-4}
Weight decay	10^{-4}
Batch size / dropout	512 / 0.15
Gradient clipping	5.0
Maximum epochs / patience	100 / 15

Aligned 3D representations are extracted with `unimol_tools` using `UniMolRepr(data_type='molecule', remove_hs=False, batch_size=32)`. PCA is used only to construct representation-shift descriptors, not as the predictive 3D bottleneck. Early stopping uses inner-validation performance only.

E OpenADMET Shift–Utility Analysis

Table 6 reports the condition-wise sample-level utility AUROC values used in the main shift–utility analysis.

Table 6: **Sample-level shift–utility association on OpenADMET.** Values are mean AUROC over 15 seed–fold macro blocks. Positive class is "2D has lower squared error."

Condition	q_{lat}	q_{rec}	q_{rms}
Original	0.5087	0.4994	0.5024
Noise 0.5 Å	0.5028	0.5061	0.4981
Noise 1.0 Å	0.4986	0.5066	0.4977
Noise 2.0 Å	0.4898	0.5005	0.4956
Randomized	0.5097	0.5148	0.5135

The largest absolute mean Spearman correlation among these standardized descriptors and conditions is 0.0407. Raw-descriptor analyses yield nearly identical OpenADMET conclusions, showing that the near-chance utility discrimination is not an artifact of clipping.

F Hard-Threshold Baseline and Decision Decomposition

The q_{rec}-threshold baseline uses only the standardized PCA reconstruction descriptor:

$$\hat{y}^{QThr} = \begin{cases} \hat{y}^{3D}, & q_{\text{rec}} \le \tau, \\ \hat{y}^{2D}, & q_{\text{rec}} > \tau. \end{cases}$$

A separate threshold is selected for every seed × fold × endpoint block using validation data only. The threshold model contains no robust expert, learned router, or prediction-disagreement feature.

On original-coordinate evaluation blocks, the coordinate-aligned audit gives an accepted fraction of 0.2979 and a fallback fraction of 0.7021. Among accepted samples, 0.5624 have larger 3D squared error than 2D; among fallback samples, 0.4239 would have been predicted better by 3D. Under noise 0.5 Å, 1.0 Å, 2.0 Å, and randomized coordinates, fallback is nearly universal, while approximately 0.2833, 0.2702, 0.2538, and 0.2526 of fallback samples, respectively, would still have been predicted better by 3D. Table 7 summarizes the resulting macro-R^2 comparison.

Table 7: **Macro-R^2 comparison for the hard fallback, full CORE, and the coordinate-independent 2D expert.**

Model	Original	Randomized
Full CORE	0.594331 ± 0.025585	0.570981 ± 0.024900
q_{rec} fallback	0.538223 ± 0.030354	0.567139 ± 0.026233
2D expert (No3D)	0.567139 ± 0.026233	0.567139 ± 0.026233
Two-expert dynamic gate	0.592983 ± 0.026965	0.569236 ± 0.026020

Corrected comparisons give BH-adjusted $q = 0.000340$ for full CORE versus the hard fallback under original coordinates, $q = 0.025533$ for hard fallback versus No3D under original coordinates, and $q = 0.155029$ for full CORE versus the hard fallback under randomized coordinates.

G Full Ablation Results

Table 8 reports the complete original/randomized ablation family and its BH-adjusted comparisons.

Table 8: **Complete ablation summary under original and randomized coordinates.** Values are mean ± SD over 15 seed–fold macro-R^2 blocks. BH-adjusted q values compare full CORE with each predefined comparator.

Model	Original R^2	Randomized R^2	q_{orig}	q_{rand}
Full CORE	0.594331 ± 0.025585	0.570981 ± 0.024900	–	–
No Shift Descriptors	0.583972 ± 0.025416	0.569610 ± 0.025693	0.001659	0.716582
Two-expert dynamic gate	0.592983 ± 0.026965	0.569236 ± 0.026020	0.749297	0.716582
No Consistency	0.594202 ± 0.025522	0.570989 ± 0.024895	0.853110	0.853110
Static expert mixture	0.602184 ± 0.026569	0.546518 ± 0.023443	0.325591	0.013301
2D expert (No3D)	0.567139 ± 0.026233	0.567139 ± 0.026233	0.000930	0.322977

The ablations do not support the claim that every CORE component independently improves predictive accuracy. Reliability context has a significant original-coordinate contribution, while the bounded robust path and consistency term do not show independent significant gains in the predefined family. We therefore interpret the latter two as safety-oriented inductive constraints: the bounded path supplies a smooth, explicitly limited 2D-anchored fallback, while consistency discourages unnecessary prediction variation across matched coordinate realizations.

H Alternative-Conformer Evaluation

We generate three ETKDGv3 conformers using the ETKDG distance-geometry framework (Riniker & Landrum, 2015) and one MMFF-lowest conformer (Halgren, 1996) from the same SMILES. The three ETKDG conformer-specific macro-R^2 values are first computed separately within each seed–fold block and then averaged within the block before the 15-block mean and SD are summarized. We do not average the three conformer predictions before computing R^2.

Hydrogens are added before embedding, chirality enforcement is enabled, and embedding uses deterministic molecule-specific random seeds. If a single-conformer ETKDG embedding fails, it is retried with `useRandomCoords=True`; this is an RDKit conformer-generation fallback and is distinct from the artificial randomized-coordinate corruption. For MMFF-lowest, candidate conformers are optimized with MMFF94s and ranked by force-field energy, with UFF fallback when MMFF properties are unavailable. These structures are computationally generated alternatives, not experimental ground-truth conformers.

I Second-Encoder Fragility Validation with Frad

To test whether severe coordinate fragility is specific to Uni-Mol, we evaluate Frad representations (Feng et al., 2023) for matched original and randomized coordinate records. This experiment is not a transfer of CORE to Frad; it is a second-encoder failure-mode validation.

Macro R^2 decreases from 0.360646 under original coordinates to −3.422540 under randomized coordinates. All 135 seed–fold–endpoint blocks degrade, and all 15 seed–fold macro comparisons have the same direction. The Nadeau–Bengio-style corrected paired two-sided test gives $p = 1.09 \times 10^{-7}$.

For sample-level utility analysis under randomized coordinates, standardized shift descriptors saturate and are therefore not used as the primary evidence. Raw descriptors yield mean Spearman correlations of approximately −0.0284, −0.0294, and −0.0358 for r_{lat}, r_{rec}, and r_{rms}, respectively, with utility AUROC 0.5059, 0.4984, and 0.5024. Thus, strong population degradation again does not imply a strong sample-level mapping from representation-shift magnitude to relative expert utility.

J External Validation Details

The external shift–utility validation uses the same conceptual comparison between a coordinate-independent Morgan-MLP and a clean-trained frozen-Uni-Mol MLP. For each fold, 3D standardization and PCA/reliability transforms are fitted only on clean inner-training data. Randomized coordinates preserve molecular identity, atom identities, and atom counts while replacing Cartesian components by independent samples from $\mathcal{U}(-5, 5)$ Å. Table 9 summarizes the external dataset preparation and target treatment.

Table 9: **External data preparation and target treatment.**

Dataset	Cleaned N	Evaluated N	Target used for modeling
Lipo	4200	4200	provided logD
ESOL	1117	1117	provided regression target
FreeSolv	642	642	provided regression target
Caco2	906	901	provided target
VDss	1111	1111	$\log_{10}$(raw VDss in L/kg)
LD50	7342	7341	provided target

For Caco2 and LD50, geometry screening excludes five and one molecules, respectively, from both expert paths. VDss is log-transformed because its cleaned raw values span 0.01–700 L/kg with median 0.95 and a single maximum of 700; this is an analysis choice for squared-error modeling and is not claimed to be the benchmark's default target definition.

Across the six datasets, clean-versus-random shift detection reaches AUROC 0.9957–1.0000. Clean sample-level utility discrimination spans AUROC 0.4430–0.5256 with a mean interval midpoint of 0.4888. Under randomized coordinates, raw-descriptor utility AUROC spans 0.4206–0.6539 with a

mean interval midpoint of 0.5127. The fraction of molecules for which randomized-coordinate 3D remains more accurate than 2D ranges from 16.13% to 26.52%, with mean 21.54%.

Full-CORE external transfer. We additionally retrain the complete routing architecture on Lipo and ESOL. The formal archived external expert pipelines are preserved exactly: Morgan fingerprints and clean 3D embeddings are standardized on inner-training data only; the saved five outer scaffold folds are evaluated with model seeds 42, 123, and 2024; and all early stopping remains inner-validation only. To reproduce the archived baselines, the external 2D/3D heads use a one-hidden-layer 256-unit MLP with Adam learning rate 10^{-3}, weight decay 10^{-5}, batch size 128, maximum 250 epochs, and patience 20. The full-CORE additions retain the main architecture settings for the bounded robust path and router: 32-dimensional task context, 128 hidden units, dropout 0.15, robust/router learning rates $10^{-3}/5 \times 10^{-4}$, weight decay 10^{-4}, batch size 512, maximum 100 epochs, patience 15, and gradient clipping at 5.0.

The external archive contains clean and randomized embeddings but not the three Gaussian perturbation families used by OpenADMET. Hence Stage 2 trains the bounded robust path on randomized representations only, while Stage 3 uses

$$\mathcal{L}_{\text{ext}} = \mathcal{L}_{\text{clean}} + 0.25\,\mathcal{L}_{\text{random}} + 0.05\,\mathcal{L}_{\text{cons}}, \qquad \mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(\hat{y}_i^{(\text{random})} - \hat{y}_i^{(\text{clean})}\right)^2.$$

The external static-mixture baseline uses the same three frozen experts and the same clean/random objective, but replaces the sample-dependent router with one learned three-logit mixture vector for the single task. Thus, differences between Static and CORE reflect adaptive routing rather than different experts, splits, or stress curricula. This experiment tests cross-dataset transfer of the *architecture and selective-fusion behavior*, not exact transfer of the OpenADMET perturbation curriculum. Table 10 reports the 15-block full-CORE transfer results.

Table 10: **Full-CORE transfer on Lipo and ESOL.** Values are mean ± SD over 15 seed–fold test blocks. These results are reported descriptively; no additional corrected significance claim is made for this post-hoc external transfer experiment.

Dataset	Method	Clean R^2	Randomized R^2
Lipo	2D	0.4493 ± 0.0376	0.4493 ± 0.0376
	3D	0.3026 ± 0.0276	−2.3024 ± 3.3852
	Robust	0.4490 ± 0.0381	0.4495 ± 0.0380
	Static	0.4831 ± 0.0367	0.4432 ± 0.0393
	CORE	**0.4839 ± 0.0314**	**0.4509 ± 0.0367**
ESOL	2D	0.3943 ± 0.2740	0.3943 ± 0.2740
	3D	**0.7306 ± 0.1224**	−1.8464 ± 1.3961
	Robust	0.4237 ± 0.2777	**0.4461 ± 0.2593**
	Static	0.6541 ± 0.0598	0.1993 ± 0.5116
	CORE	0.7202 ± 0.0905	0.3467 ± 0.4687

The two datasets cover complementary regimes. On Lipo, clean 3D is weaker than 2D; CORE matches static fusion on clean inputs and returns to essentially the 2D safety level after randomization. On ESOL, clean 3D is much stronger; CORE retains near-3D clean performance while substantially limiting the random-coordinate collapse relative to both 3D and static fusion. The robust path alone is strongest under randomized ESOL, consistent with its intended conservative fallback role.

As an implementation audit, the rerun 2D/3D experts reproduce the archived external predictions to numerical precision. The maximum absolute block-level R^2 discrepancy is below 1.1×10^{-6} on Lipo and below 1.4×10^{-5} on ESOL across 2D, clean-3D, and randomized-3D checks.

K Statistical Inference Details

For endpoint k, model seed s, outer fold f, and condition c, $R^2_{s,f,c,k}$ is computed from the outer-test molecules with an observed label for endpoint k, on the endpoint's stated modeling scale after undoing only the fold-specific target standardization. Missing labels therefore affect only the corresponding

endpoint calculation. We then define

$$R^2_{\text{macro},s,f,c} = \frac{1}{9}\sum_{k=1}^{9} R^2_{s,f,c,k}.$$

This is an unweighted endpoint macro-average: an endpoint with fewer labeled test molecules receives the same macro weight as an endpoint with more labeled test molecules, and predictions are not pooled across endpoints before computing R^2. The paper-level mean and SD for a condition/model are computed over the 15 seed–fold macro values (five distinct outer folds, each evaluated with three model seeds).

The statistical implementation applies a Nadeau–Bengio-style corrected paired t test to the 15 paired seed–fold macro differences d_j. With $n = 15$ and the five-fold test-to-train ratio $n_{\text{test}}/n_{\text{train}} \approx 1/4$, the code uses

$$SE_{\text{corr}} = s_d\sqrt{\frac{1}{15} + \frac{1}{4}}, \qquad t = \frac{\bar{d}}{SE_{\text{corr}}}, \qquad df = 14,$$

followed by a two-sided t-test and Benjamini–Hochberg adjustment within the predefined comparison family. This statement describes the implemented test exactly. Because the three model seeds reuse the same five outer-test partitions, the 15 paired values are repeated model fits over five distinct data splits rather than 15 independently resampled partitions; the correction does not introduce an additional hierarchical seed-within-fold term. The primary ablation family contains CORE versus five predefined comparators under original and randomized conditions, for 10 macro-level comparisons. Alternative-conformer comparison families are treated separately according to their predefined analysis groups. Nonsignificant differences are not interpreted as evidence of equivalence.

The added Lipo/ESOL full-CORE transfer is a post-hoc external robustness check motivated by cross-dataset generalization concerns. We report mean ± SD over the same 15 seed–fold blocks but do not attach a new corrected significance claim to this supplemental transfer analysis.

L Reproducibility and Provenance Notes

The main OpenADMET experiment covers 3 model seeds × 5 fixed outer scaffold folds × 9 endpoints × 5 artificial coordinate conditions. Thus, the 15 seed–fold values are repeated model fits over five distinct test partitions, not 15 independent scaffold partitions. Archived sample-level predictions can be used to recompute each endpoint-level outer-test R^2, the equal-weight nine-endpoint macro R^2 for every seed–fold block, and the reported 15-block summaries. The 2D expert is coordinate-independent and therefore produces identical predictions across artificial coordinate conditions.

The Lipo/ESOL full-CORE transfer uses 3 model seeds × 5 saved scaffold folds for each dataset. Each run archives the aligned 2D/3D expert predictions, clean/random shift descriptors, robust-path outputs, router outputs, routing weights, block-level metrics, and the provenance check against the previously archived external expert predictions. The transfer code does not overwrite the earlier external shift–utility experiment.

The anonymous submission artifact is intended to archive the processed modeling table; scaffold assignments; coordinate-generation scripts and fixed coordinate archives; aligned 512-dimensional Uni-Mol embeddings; all fold-specific predictions; CORE and ablation checkpoints; q_{rec}-threshold selections; alternative-conformer outputs; external-validation inputs; Lipo/ESOL full-CORE transfer outputs; statistical-analysis scripts; exact dependency versions; and the exact commands used to reproduce the final tables and figures.