PATHWISE RELATIONAL CONFORMER TENSORS FOR MOLECULAR ENSEMBLE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular conformer ensembles are often modeled as sets of three-dimensional structures, yet they possess substantially more structure than generic sets: all conformers share the same molecular graph, inducing exact correspondence among atoms, bonds, angles, and torsions. We argue that these aligned local structures provide natural sites for relational computation before ensemble aggregation. We introduce PaSTNet, a molecular ensemble architecture built around the Pathwise State Tensor (PaST) operator. For each molecular path, PaST lifts conformer path states into channel-valued symmetric and antisymmetric pair tensors, contextualizes these relations using permutation-equivariant tensor operations, and contracts them back into conformer updates. We characterize PaST's linear core and show that its structural maps form complete families of homogeneous permutation-equivariant linear operators within the symmetric and antisymmetric pair branches at fixed conformer cardinality. To center each block's ensemble response relative to a singleton reference, PaST further employs a singleton-centered residual that subtracts the response of the same relational operator re-evaluated on the corresponding singleton input, making each PaST block exactly the identity when only one conformer is available. PaSTNet alternates these relational updates with geometry-conditioned propagation across molecular paths, allowing cross-conformer relational signals induced by conformational differences to propagate through the atom–bond–angle–torsion hierarchy before conformer pooling. Together with reversal-invariant path encoding and invariant geometric descriptors, the resulting model is invariant to a common permutation of conformers, consistent atom relabeling, and independent proper rigid motions of each conformer. Across six molecular property tasks, PaSTNet achieves the best evaluated mean score on five. Identity, attention, pair-local, and no-subtraction controls favor full PaST in mean performance; attribution remains limited by coupled changes in some controls and uncertainty over three splits. PaSTNet therefore provides a principled framework for reasoning explicitly over aligned local relations in molecular conformer ensembles rather than treating conformers only as representations to be pooled.

1 INTRODUCTION

A molecular conformer ensemble is more structured than a generic set of three-dimensional geometries. Its members share the same molecular graph and therefore induce exact correspondence among atoms, bonds, angles, and torsions. While molecular property predictors commonly operate on a chemical graph or a single three-dimensional conformation (Gilmer et al., 2017; Schütt et al., 2017; Gasteiger et al., 2020), an ensemble exposes how the *same local molecular structure* varies across multiple geometries. A particular covalent path can thus be followed across conformers, providing a natural site at which to compare conformational variation before forming a molecular representation. Existing conformer-ensemble methods exploit multiple geometries through pooling, distribution summaries, attention, or interacting ensemble encoders (Zhu et al., 2024; Nguyen et al., 2024; Zhu et al., 2025; Nguyen et al., 2026). These approaches differ substantially in architecture, but do not make an explicit channel-valued relation at each aligned local structure the central intermediate representation. Our focus is the representation of relations at each aligned covalent path, rather than the absence of relational processing in prior models. This raises a natural representa-

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.

tion question: *can conformer ensembles be modeled by reasoning directly over relations between corresponding local structures, rather than only over the conformers?*

We address this question with PaSTNet, whose central object is a *path-indexed relational conformer tensor*. For each molecular path p, the PaST operator lifts conformer states $\{\mathbf{h}_c\}_{c=1}^{K}$ into $\mathbf{R}_p = (\mathbf{S}_p, \mathbf{A}_p)$, an ordered pair of symmetric and antisymmetric channel-valued tensors. It contextualizes each relation using other conformer-pair relations at the same path and contracts the resulting tensor back into conformer updates. PaST therefore implements *relations* $\rightarrow$ *contextualized relations* $\rightarrow$ *conformer updates*, making pairwise conformer structure an explicit computational state rather than an intermediate score that is immediately reduced. PaSTNet alternates PaST with geometry-conditioned propagation across molecular paths, allowing cross-conformer relational information to propagate through the atom–bond–angle–torsion hierarchy before conformer pooling. A shared spatial backbone provides complementary atom-level context, including nonbonded neighborhood information.

Our contributions are threefold. *First*, we introduce path-aligned conformer-pair states that are updated before contraction into conformer representations. Their branchwise homogeneous linear mixers are complete at fixed cardinality, specializing established permutation-equivariant tensor theory (Maron et al., 2019). *Second*, a shared singleton evaluation defines a block-local singleton-centered ensemble response with an exact identity at $K = 1$. *Third*, integrating these updates with reversal-invariant path encoding and geometry-conditioned propagation preserves conformer permutation, consistent atom relabeling, and independent proper rigid motions. We evaluate six property datasets, compare identity, attention, pair-local, and no-subtraction controls, and study separately trained conformer caps and model-pass resource costs.

2 Related Work

Geometric and higher-order molecular representations. Message-passing neural networks learn molecular representations by propagating information over chemical graphs and aggregating local states into molecular embeddings (Gilmer et al., 2017). Geometric extensions incorporate three-dimensional structure at increasing levels of detail: SchNet uses continuous distance-conditioned filters (Schütt et al., 2017), DimeNet introduces angular information (Gasteiger et al., 2020), GemNet develops directional message passing (Gasteiger et al., 2021), and PaiNN combines scalar and rotationally equivariant vector features (Schütt et al., 2021). Beyond atom-level graphs, higher-order architectures represent structured objects such as cells or molecular paths (Bodnar et al., 2021; Li et al., 2024). In particular, the Path Complex Neural Network (PCNN) (Li et al., 2024) organizes atoms, bonds, angles, and torsions into a hierarchy of covalent paths. In this work, PaSTNet builds on this higher-order view, but uses aligned paths not only to propagate geometric information within a conformer, but also as sites for relational computation across conformers.

Learning from molecular conformer ensembles. Conformer ensembles provide multiple geometries of the same molecule and have been incorporated through pooling, attention, representation fusion, and explicit cross-conformer interaction. Early work showed the benefit of learning from multiple conformations and introduced learned conformer attention (Axelrod & Gómez-Bombarelli, 2023). ConAN combines two- and three-dimensional molecular representations, with ConAN-FGW further using a fused Gromov–Wasserstein barycenter to couple structural information across conformers (Nguyen et al., 2024). FACET learns fragment-aware conformer representations with graph transformers and FGW-based supervision before aggregating them in a shared latent space (Nguyen et al., 2026), while SPiCE alternates geometric encoding with ensemble-level interaction under geometric and conformer-permutation symmetries (Zhu et al., 2025). These methods demonstrate several ways to couple multiple molecular geometries. PaSTNet takes a different representation view: because the same covalent path is aligned across conformers, it assigns that path an explicit channel-valued state for every conformer pair and contextualizes these relations before contracting them back into conformer representations.

Conformer pretraining, datasets, and ensemble summaries. Large-scale molecular pretraining provides another route to exploiting 3D structure. Uni-Mol pretrains a general-purpose 3D molecular encoder over a large collection of conformations (Zhou et al., 2023), whereas 3DGCL treats different conformers as related views for contrastive representation learning (Moon et al., 2023).

Complementary dataset efforts have made conformational information increasingly accessible: GEOM provides large energy-annotated conformer collections (Axelrod & Gómez-Bombarelli, 2022), and MARCEL establishes benchmarks specifically for learning from molecular conformer ensembles (Zhu et al., 2024). Rather than maintaining explicit pair states, ensemble structure can also be summarized statistically. Distribution Kernel Operators, for example, represent conformer features through distributional quantities such as their mean and covariance (Cheng et al., 2026). Such approaches highlight a complementary design choice: compress the ensemble into a global summary, or retain pairwise relational structure long enough for relations themselves to interact.

Set functions, equivariant relational operators. Generally, conformer ensembles are instances of learning over unordered collections. Deep Sets obtains permutation invariance through symmetric aggregation (Zaheer et al., 2017), while Set Transformer introduces interactions among set elements through attention (Lee et al., 2019). Pair-state retention also appears in Edge Transformers, which update pair vectors through triangular attention (Bergen et al., 2021), and Relational Attention, which updates edge vectors within transformer layers (Diao & Loynd, 2023). PaST instantiates relations between conformers at each aligned molecular path, with exchange-parity branches and a shared singleton-centered update. Its linear tensor operations are based on the permutation-equivariant constructions of Maron et al. (2019): diagonal, row, trace, and global contractions provide relational context while commuting with conformer permutations. We specialize this framework to symmetric and antisymmetric conformer-pair states and characterize the resulting branchwise linear operator families (App. F). The novelty is therefore not a new general theory of permutation-equivariant tensors, but their use as explicit path-aligned relational states, together with geometry-aware interaction and the singleton-centered PaST update. This distinction is architectural rather than an expressivity claim over arbitrary attention models, which may also retain pairwise states when designed to do so.

3 PaSTNet: Pathwise Relational Conformer Reasoning

Figure 1 summarizes the architecture. The state at each path is a conformer-indexed feature matrix; each PaST block temporarily lifts it into the pair tensors $(\mathbf{S}_p, \mathbf{A}_p)$ before returning an update of the same shape.

3.1 Motivation for explicit pair states

Pooling produces an ensemble summary; standard self-attention contracts pair scores into conformer states. PaST retains channel-valued pair states through two contextualization layers. For example, at a torsion with conformer angles $20°$, $25°$, and $170°$, a row summary lets the first–third relation depend on the first–second relation. This is an illustration of information flow, not evidence that a molecular target requires it or that pair-aware attention cannot represent it. Pair states are reconstructed at each stage from the updated conformer states.

3.2 Spatial backbone and path initialization

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an attributed molecular graph and $\mathcal{X} = \{\mathbf{X}_c\}_{c=1}^{K}$, with $\mathbf{X}_c \in \mathbb{R}^{|\mathcal{V}|\times 3}$, its conformers under a known atom correspondence. For $r \in \{0, 1, 2, 3\}$, the set $\mathcal{P}_r$ contains simple covalent paths corresponding to atoms, bonds, angles, and torsions. Each path is stored once modulo reversal using $\mathrm{canon}(p) = \min_{\mathrm{lex}}\{p, \mathrm{rev}(p)\}$. A coface $p \in \mathcal{P}_{r+1}$ connects its canonical prefix u and suffix v through both $u \to v \mid p$ and $v \to u \mid p$, so reversing the stored representative preserves its incidence set.

Each conformer is first processed by a shared SchNet backbone (Schütt et al., 2017), trained jointly from random initialization. Three interaction blocks produce atom states $\mathbf{u}_{c,i} \in \mathbb{R}^{d_s}$ with $d_s = 128$, using all nonself neighbors within 10 Å. This provides spatial context, including nonbonded interactions, before path-level reasoning. Building on PCNN's higher-order geometric representation (HRGE) (Li et al., 2024), we combine elemental and bond descriptors with internal distances, angles, signed torsion harmonics, and area/volume descriptors (Appendix A). Let $\mathbf{J}_r$ denote the feature action of path reversal, so $\mathbf{F}_{r,c,\mathrm{rev}(p)} = \mathbf{J}_r \mathbf{F}_{r,c,p}$ and $\mathbf{J}_r^2 = \mathbf{I}$, and let $\mathbf{s}_{r,c,p}$ be a reversal-invariant summary of the path's atom states (Appendix A.4). We enforce reversal invariance by averaging

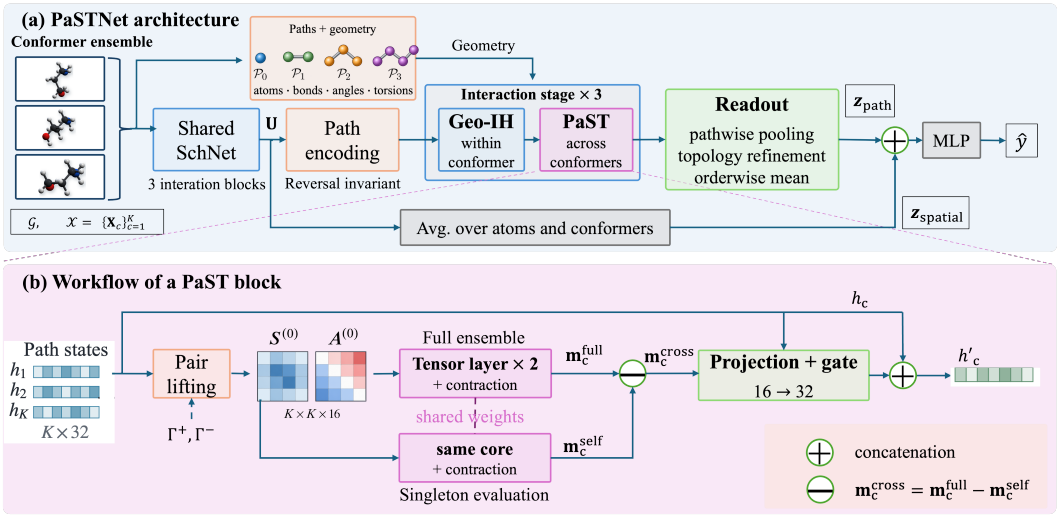


Figure 1: **PaSTNet architecture.** (a) Shared spatial features initialize reversal-invariant path states. Three stages alternate geometry-conditioned propagation within conformers (Geo-IH) with relational interaction across conformers (PaST), followed by pooling and readout. (b) At each path, PaST forms symmetric and antisymmetric pair tensors, contextualizes them, and subtracts the shared singleton response. For $K = 1$, the PaST block is exactly the identity. Readout concatenates the path summaries with the spatial bypass.

after a shared learned encoder:

$$\mathbf{H}^{(0)}_{r,c,p} = \tfrac{1}{2}\left[e_r(\mathbf{s}_{r,c,p}, \mathbf{F}_{r,c,p}) + e_r(\mathbf{s}_{r,c,p}, \mathbf{J}_r\mathbf{F}_{r,c,p})\right], \qquad \mathbf{H}^{(0)}_r \in \mathbb{R}^{K\times|\mathcal{P}_r|\times d_p}. \tag{1}$$

Equation (1) removes dependence on the chosen path orientation while retaining signed torsion information, since full path reversal preserves the signed torsion under our convention. Bond-direction metadata are transported with their original bonds under atom relabeling, not regenerated from a new drawing. Appendix A.4 specifies the feature actions and descriptor transport. We use path width $d_p = 32$ and relational width $d_r = 16$.

3.3 Geo-IH: Propagation across molecular paths

PaST compares conformers at a fixed molecular path, whereas geometry-conditioned interaction heads (Geo-IH) propagate information between paths within each conformer. Geo-IH adapts the IH message-passing primitive of PCNN (Li et al., 2024) to the shared canonical path hierarchy. Each stage applies

$$\widetilde{\mathbf{H}}^{(s)} = \mathsf{GeoIH}_s(\mathbf{H}^{(s-1)}), \qquad \mathbf{H}^{(s)} = \mathsf{PaST}_s(\widetilde{\mathbf{H}}^{(s)}), \qquad s = 1, \dots, L. \tag{2}$$

Consequently, ensemble context introduced by PaST can be propagated across molecular paths by a subsequent Geo-IH stage. For example, a torsion coface (i, j, k, l) links the angle paths (i, j, k) and (j, k, l). Geo-IH retains IH's source–destination–coface structure and augments it with explicit geometric conditioning of both routing and message values. Suppressing stage, conformer, order, and head indices, let $\mathbf{x}_u, \mathbf{x}_v, \mathbf{e}_p$ denote normalized path states, and define $\mathbf{c}_{uvp} = \mathbf{W}_i\mathbf{x}_u + \mathbf{W}_j\mathbf{x}_v + \mathbf{W}_e\mathbf{e}_p + \mathbf{b}_e$ and $\mathbf{t}_u = \mathbf{W}_N[\mathbf{x}_u \,\|\, \bar{\mathbf{e}}_u] + \mathbf{b}_N$, where $\bar{\mathbf{e}}_u$ is the mean state of cofaces incident to the source. An embedding $\mathbf{g}_p$ of the coface geometry—bond length, angle, or torsion—produces feature-wise affine modulation (FiLM) (Perez et al., 2018):

$$\mathbf{z}_{uvp} = \mathrm{SiLU}\big[(1 + \boldsymbol{\gamma}_p) \odot \mathbf{c}_{uvp} + \boldsymbol{\beta}_p\big], \tag{3}$$

$$\mathbf{m}_{uvp} = (1 + \boldsymbol{\lambda}_p) \odot \mathbf{t}_u + \boldsymbol{\mu}_p. \tag{4}$$

The modulation vectors are learned from $\mathbf{g}_p$. Softmax weights derived from $\boldsymbol{\eta}^\top\mathbf{z}_{uvp}$ aggregate $\mathbf{m}_{uvp}$ at each destination. Both directed incidences use the same reversal-invariant geometry, and averaging their directed relation states updates the canonical coface independently of its stored orientation.

Geo-IH therefore uses absolute geometry within each conformer, whereas PaST uses relative geometry across conformers. Conditioning both routing and values gives geometry a direct influence on message content. In particular, when a destination has only one incoming relation, its softmax weight is necessarily one, yet the modulated value can still depend on geometry. The sequential $\mathcal{P}_3 \to \mathcal{P}_2 \to \mathcal{P}_1 \to \mathcal{P}_0$ sweep immediately reuses updated states, allowing torsion-level information to reach atom states within a single Geo-IH stage. We use $L = 3$ stages with distinct Geo-IH and PaST parameters, while sharing the embedding bank for fixed absolute geometry across stages. Head dimensions, normalization, and residual scaling are detailed in Appendix B.

3.4 PaST: relational conformer reasoning

PaST operates independently at each fixed path $p \in \mathcal{P}_r$ for all four path orders. Suppressing stage, order, and path indices, let K be the actual conformer count, $\mathbf{h}_c \in \mathbb{R}^{d_p}$ its path state, and $\bar{\mathbf{h}}_c = N^C_{s,r}(\mathbf{h}_c)$ its RMS-normalized representation. Relative descriptors $\mathbf{\Gamma}^+_{r,cc'}$ and $\mathbf{\Gamma}^-_{r,cc'}$ are respectively even and odd under exchanging c and c'. They encode differences in bond length and angle/torsion harmonics with joint validity masks; for atoms, PaST uses content only. A key property of the relational core is that its branchwise linear mixers are complete within their corresponding permutation-equivariant pair spaces, as formalized in Theorem 1.

Pair lifting. PaST first lifts conformer states into symmetric and antisymmetric channel-valued pair tensors $\mathbf{S}^{(0)}, \mathbf{A}^{(0)} \in \mathbb{R}^{K\times K\times d_r}$:

$$\mathbf{u}^+_{cc'} = [\bar{\mathbf{h}}_c + \bar{\mathbf{h}}_{c'} \parallel |\bar{\mathbf{h}}_c - \bar{\mathbf{h}}_{c'}| \parallel \bar{\mathbf{h}}_c \odot \bar{\mathbf{h}}_{c'} \parallel \mathbf{\Gamma}^+_{r,cc'}], \qquad \mathbf{S}^{(0)}_{cc'} = N_+(f_+(\mathbf{u}^+_{cc'})), \tag{5}$$

$$\mathbf{u}^-_{cc'} = [\bar{\mathbf{h}}_c - \bar{\mathbf{h}}_{c'} \parallel \mathbf{\Gamma}^-_{r,cc'}], \qquad \mathbf{A}^{(0)}_{cc'} = N_-(f_-(\mathbf{u}^-_{cc'})). \tag{6}$$

The symmetric branch uses a shared MLP. In the antisymmetric branch, $f_-(\cdot)$ uses bias-free linear maps and an even gate, while $N_-(\cdot)$ is a normalization without additive bias; both preserve oddness. Consequently, $\mathbf{S}_{cc'} = \mathbf{S}_{c'c}$ and $\mathbf{A}_{cc'} = -\mathbf{A}_{c'c}$. The two branches capture exchange-symmetric comparisons and directed conformer differences, respectively; this parity concerns conformer exchange rather than spatial reflection.

Equivariant relational reasoning. Two tensor layers contextualize the pair states before contraction. For a current pair state, define diagonal states $\mathbf{d}_c = \mathbf{S}_{cc}$, row means $\boldsymbol{\rho}_c = K^{-1}\sum_j \mathbf{S}_{cj}$, mean trace $\boldsymbol{\tau} = K^{-1}\sum_j \mathbf{S}_{jj}$, global mean $\mathbf{g} = K^{-2}\sum_{ij} \mathbf{S}_{ij}$, and odd row means $\mathbf{q}_c = K^{-1}\sum_j \mathbf{A}_{cj}$. The symmetric branch uses the nine structural maps

$$\{\mathbf{S}_{cc'},\ \delta_{cc'}\mathbf{d}_c,\ \mathbf{d}_c + \mathbf{d}_{c'},\ \delta_{cc'}\boldsymbol{\rho}_c,\ \boldsymbol{\rho}_c + \boldsymbol{\rho}_{c'},\ \delta_{cc'}\boldsymbol{\tau},\ \boldsymbol{\tau},\ \delta_{cc'}\mathbf{g},\ \mathbf{g}\}\,. \tag{7}$$

Writing these maps as $\{\mathcal{B}_b\}_{b=1}^9$, the symmetric mixer is $\mathcal{L}^+_{22}(\mathbf{S}) = \sum_{b=1}^9 \mathcal{B}_b(\mathbf{S})\mathbf{W}_b$. The corresponding antisymmetric mixer is

$$\mathcal{L}^-_{22}(\mathbf{A})_{cc'} = \mathbf{A}_{cc'}\mathbf{V}_1 + (\mathbf{q}_c - \mathbf{q}_{c'})\mathbf{V}_2. \tag{8}$$

Theorem 1 (Complete branchwise linear pair mixers)**.** *For every fixed $K \geq 4$, the nine structural maps in Equation (7), with arbitrary channel mixing, form a basis of all homogeneous linear permutation-equivariant maps between symmetric pair tensors. For $K = 1, 2, 3$, they remain spanning, with scalar operator dimensions $1, 5, 8$. The two structural maps in Equation (8) span all homogeneous linear permutation-equivariant maps between antisymmetric pair tensors and are independent for $K \geq 3$. Their scalar operator dimensions at $K = 1, 2, 3$ are $0, 1, 2$.*

The proof classifies simultaneous-permutation orbits of symmetric input/output index pairs: nine equality patterns exist for $K \geq 4$. Antisymmetric sign constraints leave two independent maps for $K \geq 3$ (Appendices F.1–F.2), specializing Maron et al. (2019). Completeness is homogeneous, branchwise, and fixed-K: it excludes linear cross-branch maps such as $S \mapsto \rho_i - \rho_j$ and $A \mapsto q_i + q_j$, and implies neither nonlinear universality nor predictive benefit.

The two branches are coupled through parity-preserving products:

$$\mathbf{Z}^+ = \mathcal{L}^+_{22}(\mathbf{S}) + (\mathbf{A}\odot\mathbf{A})\mathbf{W}_{AA}, \qquad \mathbf{Z}^- = \mathcal{L}^-_{22}(\mathbf{A}) + (\mathbf{S}\odot\mathbf{A})\mathbf{W}_{SA}. \tag{9}$$

Since $\mathbf{A}\odot\mathbf{A}$ is symmetric and $\mathbf{S}\odot\mathbf{A}$ is antisymmetric under conformer exchange, these couplings preserve the two parity branches. Entrywise feed-forward residual updates preserve the same structure. Retaining the pair states through two tensor layers allows each relation (c, c') to incorporate context from other conformer-pair relations at the same molecular path before contraction.

Block-local singleton-centered residual. After relational reasoning, a $2 \to 1$ contraction produces

$$\mathbf{m}_c^{\text{full}} = \mathbf{d}_c \mathbf{U}_1 + \boldsymbol{\rho}_c \mathbf{U}_2 + \boldsymbol{\tau} \mathbf{U}_3 + \mathbf{g} \mathbf{U}_4 + \mathbf{q}_c \mathbf{U}_5. \tag{10}$$

We then re-evaluate the same tensor core and collapse, with shared parameters, on the singleton input $([\mathbf{S}_{cc}^{(0)}], [\mathbf{0}])$, using $K = 1$ in every contraction, to obtain $\mathbf{m}_c^{\text{self}}$. The incoming state $\mathbf{h}_c$ is held fixed and may already contain ensemble context from earlier stages, so this counterfactual is local to the current PaST block. We center the block response at this singleton reference:

$$\mathbf{m}_c^{\text{cross}} = \mathbf{m}_c^{\text{full}} - \mathbf{m}_c^{\text{self}}, \qquad \mathbf{h}'_c = \mathbf{h}_c + u_{s,r}(\mathbf{h}_c, \mathbf{m}_c^{\text{cross}}), \quad u_{s,r}(\mathbf{h}, \mathbf{0}) = \mathbf{0}. \tag{11}$$

Here $u_{s,r}$ is a gated residual whose message transform vanishes at zero (Appendix C). Hence $K = 1$ implies $\mathsf{PaST}(\mathbf{h}) = \mathbf{h}$ for all learned parameters. Masking the final diagonal is not equivalent to this singleton evaluation because the pair tensors may already contain row, global, and diagonal-derived context. The subtraction defines a block-local singleton-centered ensemble response. Cardinality-normalized contractions can change even for duplicated identical conformers, so it is neither a causal nor a diversity-only decomposition.

3.5 Readout, invariances, and complexity

After the final interaction stage, we pool conformers independently at each path:

$$\beta_{r,c,p} = \frac{\exp\left(\mathbf{w}_r^\top \mathbf{H}_{r,c,p}^{(L)}\right)}{\sum_{c'=1}^{K} \exp\left(\mathbf{w}_r^\top \mathbf{H}_{r,c',p}^{(L)}\right)}, \qquad \bar{\mathbf{H}}_{r,p} = \sum_{c=1}^{K} \beta_{r,c,p} \mathbf{H}_{r,c,p}^{(L)}. \tag{12}$$

Because the pooling weights are path-specific, the pooled states need not correspond to any single representative conformer. We therefore apply a final topology-only descending sweep using path incidence, without injecting additional raw geometry. Its inputs still encode geometry and need not describe one realizable pose. This yields refined states $\widehat{\mathbf{H}}_{r,p}$. The molecular representation combines orderwise path summaries with a spatial-backbone bypass:

$$\mathbf{z}_r = \begin{cases} |\mathcal{P}_r|^{-1} \sum_{p \in \mathcal{P}_r} \widehat{\mathbf{H}}_{r,p}, & |\mathcal{P}_r| > 0, \\ \mathbf{0}, & |\mathcal{P}_r| = 0, \end{cases} \qquad \mathbf{z}_{\text{spatial}} = \frac{1}{K|\mathcal{V}|} \sum_{c=1}^{K} \sum_{i \in \mathcal{V}} \mathbf{u}_{c,i},$$
$$\hat{y} = \mathsf{MLP}([\mathbf{z}_0 \parallel \mathbf{z}_1 \parallel \mathbf{z}_2 \parallel \mathbf{z}_3 \parallel \mathbf{z}_{\text{spatial}}]). \tag{13}$$

The resulting 256-dimensional representation is processed by a $256 \to 128 \to 64 \to 1$ SiLU head, whose scalar output is used directly for regression or as a binary-classification logit.

Three complementary invariances. The predictor is invariant to (i) a common permutation of conformers, (ii) independent proper rigid motions of each conformer, and (iii) any consistent relabeling of the attributed molecular graph together with its coordinates and descriptors. Shared conformer-wise operations and equivariant PaST updates preserve (i) before conformer pooling removes the conformer axis. Internal geometric quantities and the distance-based spatial backbone give (ii). Reversal-invariant path initialization, shared path operations, both directed incidences, and orderwise pooling give (iii). Appendix D provides the formal statements and proofs, including the required descriptor-transport assumption. Stochastic depth is disabled. Because signed torsions may change under reflection, these guarantees apply to proper rigid motions and do not imply reflection invariance.

Computational cost. Let $P = \sum_{r=0}^{3} |\mathcal{P}_r|$ and let $T = 2$ denote the number of tensor layers in each PaST block. Across L interaction stages, pair lifting, relational mixing, and contraction require $\mathcal{O}\big(LPK^2(d_p d_r + T d_r^2)\big)$, while the singleton evaluations add $\mathcal{O}\big(LPK(d_p d_r + T d_r^2)\big)$. Pair-state storage is $\mathcal{O}(PK^2 d_r)$ per stage. Mixed-size ensembles are padded to a configured conformer capacity and masked so that all reductions use each molecule's actual conformer count; conformers are never duplicated. Appendix E gives the Geo-IH and spatial-backbone costs, padded execution details, and scaling limitations.

4 EXPERIMENTAL EVALUATION

4.1 DATASETS AND EVALUATION PROTOCOL

We evaluate molecular property prediction on ESOL, FreeSolv, Lipophilicity (Lipo), BACE, BBBP, and ClinTox (Delaney, 2004; Wu et al., 2018). We use RMSE for regression and ROC-AUC in percent for classification; ClinTox uses the single CT_TOX endpoint. All predictive and resource measurements are from our own local runs, including the SchNet and Uni-Mol v1 baseline scores. Tables 1–3 share 81/9/10 train/validation/test random-scaffold splits (3 seeds), subject to whole-scaffold allocation; initialization uses seed 42. Scores are means and sample standard deviations across the three splits. PaSTNet uses training-only target standardization for regression and validation RMSE for checkpoint selection; classification uses binary cross-entropy on logits and validation ROC-AUC. Test evaluation follows checkpoint selection. All 314 ESOL molecules with empty scaffolds lie in training; its test sets therefore assess held-out ring-containing scaffolds, not acyclic generalization. Appendix G details model-specific recipes and the scope of the linked code release.

4.2 COMPARISON WITH MOLECULAR PREDICTION BASELINES

We compare PaSTNet with PCNN (Li et al., 2024), SchNet (Schütt et al., 2017), Uni-Mol v1 (Zhou et al., 2023), SPiCE (Zhu et al., 2025), ConAN and ConAN-FGW (Nguyen et al., 2024), FACET (Nguyen et al., 2026), 3DGCL (Moon et al., 2023), PaiNN (Schütt et al., 2021), and DKO (Cheng et al., 2026). Together, these baselines cover independent geometric encoding, pretrained 3D representations, interacting conformer ensembles, and distribution-based summaries. Table 1 reports all locally reproduced scores. The comparison is not architecture-matched: PCNN has fewer parameters and omits distance-contact edges, while DKO uses the local diagonal-variance, truncated-feature adaptation described in Appendix G. We therefore use this table to position overall predictive performance rather than to isolate the effect of PaST. Uni-Mol's target scaler was fitted on training and validation targets before splitting; test targets were excluded. Its reported scores retain this validation-inclusive preprocessing, whereas PaSTNet uses training-only standardization. No train-only Uni-Mol rerun is reported, so preprocessing is not fully harmonized. PaSTNet has the best evaluated mean score on five of the six tasks: ESOL, FreeSolv, Lipo, BACE, and BBBP. Relative to the strongest baseline on each task, the margins are 0.02, 0.03, and 0.01 RMSE on ESOL, FreeSolv, and Lipo, respectively, 1.97 ROC-AUC points on BACE, and 0.83 points on BBBP. On ClinTox, ConAN-FGW has the highest mean at $84.30 \pm 5.50\%$, while PaSTNet obtains $81.04 \pm 7.63\%$. Because the compared methods differ in features, pretraining, capacity, and tuning, these results should be interpreted as mean-score comparisons rather than evidence that the interaction operator alone explains the observed gaps. With only three split seeds and a fixed initialization seed, we likewise make no statistical-significance claim.

4.3 CONTROLLED ANALYSIS OF PAST INTERACTIONS

Table 2 compares four controls at $K = 5$ with the same conformer inputs and surrounding architecture (Appendix G). Against *No PaST*, which replaces every PaST block by identity, full PaST lowers mean ESOL and FreeSolv RMSE by 0.05 and 0.12, and raises BACE ROC-AUC by 2.04 points. This supports adding the complete block, without isolating its mechanisms or added capacity.

The *same-path attention* control uses two four-head attention/FFN layers with relative geometry as attention-logit biases. Its $667{,}710$ parameters closely match PaSTNet's $667{,}548$ (within 0.03%). PaST improves mean RMSE by 0.03 and 0.07, and BACE ROC-AUC by 0.95 points. Geometry placement, pair-state retention, and centering change together; this control does not retain channel-valued pair states.

The *pair-local* control retains two tensor layers, local/diagonal maps, parity coupling, singleton subtraction, and the full collapse, but removes cross-pair contextual maps. Its $618{,}396$ parameters are $49{,}152$ fewer than full PaST. The lower means therefore reflect a combined change in context and capacity, without isolating context or individual contraction families.

Removing *singleton subtraction* at unchanged parameter count increases mean ESOL and FreeSolv RMSE by 0.06 each and reduces BACE ROC-AUC by 1.51 points. These results favor singleton

Table 1: Molecular property prediction: local mean ± sample standard deviation over 81/9/10 random-scaffold splits (3 seeds). Bold and underlining mark the best and second-best means, without a significance claim. ClinTox reports the single CT_TOX endpoint, not a multitask average. Model-specific protocols are in Appendix G.

	Regression: RMSE ↓			Classification: ROC-AUC (%) ↑		
Model	ESOL	FreeSolv	Lipo	BACE	BBBP	ClinTox (CT_TOX)
PCNN (no contact edges)	0.95 ± 0.10	2.24 ± 0.08	0.69 ± 0.05	81.89 ± 0.68	87.56 ± 1.80	74.28 ± 3.60
SchNet	0.87 ± 0.05	2.73 ± 0.28	0.66 ± 0.04	83.44 ± 4.20	$\underline{89.86 \pm 4.80}$	68.19 ± 5.60
Uni-Mol v1	$\underline{0.78 \pm 0.08}$	$\underline{1.73 \pm 0.37}$	$\underline{0.57 \pm 0.04}$	$\underline{83.57 \pm 3.70}$	87.86 ± 3.70	80.73 ± 8.80
SPiCE	0.92 ± 0.08	2.15 ± 0.65	0.74 ± 0.12	77.34 ± 4.48	87.18 ± 4.50	66.14 ± 7.00
ConAN	0.93 ± 0.08	2.40 ± 0.19	0.69 ± 0.05	83.44 ± 2.37	87.09 ± 2.40	$\underline{83.33 \pm 4.00}$
ConAN-FGW	0.90 ± 0.12	2.37 ± 0.19	0.70 ± 0.06	81.74 ± 3.43	88.95 ± 3.40	$\mathbf{84.30 \pm 5.50}$
FACET	0.90 ± 0.11	2.02 ± 0.39	0.70 ± 0.09	82.98 ± 5.53	89.77 ± 5.00	80.08 ± 8.00
3DGCL	0.99 ± 0.15	2.32 ± 0.66	0.68 ± 0.13	63.76 ± 17.35	80.35 ± 9.00	60.76 ± 12.00
PaiNN	0.87 ± 0.08	2.17 ± 0.15	0.61 ± 0.05	81.60 ± 2.56	84.33 ± 2.60	78.92 ± 4.50
DKO (diag., truncated)	1.42 ± 0.03	5.82 ± 1.20	1.06 ± 0.10	70.39 ± 4.87	75.09 ± 4.90	54.09 ± 8.00
PaSTNet (ours)	$\mathbf{0.76 \pm 0.06}$	$\mathbf{1.70 \pm 0.42}$	$\mathbf{0.56 \pm 0.06}$	$\mathbf{85.54 \pm 3.41}$	$\mathbf{90.69 \pm 2.71}$	81.04 ± 7.63

Table 2: Interaction controls at $K = 5$: mean ± sample standard deviation over three random-scaffold splits. Bold/underlining mark the best/second-best means. The attention control is approximately parameter matched; removing singleton subtraction preserves the parameter count.

Interaction variant	ESOL RMSE ↓	FreeSolv RMSE ↓	BACE ROC-AUC (%) ↑
No PaST (identity)	0.81 ± 0.10	1.82 ± 0.30	83.50 ± 3.50
Same-path attention	$\underline{0.79 \pm 0.08}$	1.77 ± 0.42	$\underline{84.59 \pm 6.02}$
PaST w/o pair-to-pair context	0.85 ± 0.15	1.90 ± 0.26	82.02 ± 4.37
PaST w/o singleton subtraction	0.82 ± 0.14	$\underline{1.76 \pm 0.40}$	84.03 ± 3.72
Full PaST	$\mathbf{0.76 \pm 0.06}$	$\mathbf{1.70 \pm 0.42}$	$\mathbf{85.54 \pm 3.41}$

centering within the evaluated architecture, without establishing statistical significance or general necessity.

4.4 Conformer count and computational scaling

Table 3 evaluates caps $K \in \{1, 3, 5, 10\}$ with separate training and training-only bond calibration for every cap and split (Appendix G). This is not test-time sensitivity of a fixed checkpoint. The pattern is nonmonotonic: $K = 5$ yields the lowest ESOL and FreeSolv mean RMSE, while $K = 10$ yields the highest BACE mean ROC-AUC. At $K = 1$, each PaST block is the identity, but the backbone, Geo-IH, refinement, and readout remain active. Cap effects combine conformer availability, interaction, and pooling.

Table 4 compares model-pass costs on 100 selected ESOL training molecules with at least ten conformers. FACET has lower measured costs at every cap; at $K = 10$, PaSTNet uses $4.11\times$ the memory, $3.38\times$ the forward-plus-backward time, and $3.18\times$ the inference time. These measurements exclude optimizer steps and offline preparation, including FACET's teacher/FGW stage; they are not end-to-end training costs.

5 Discussion and Limitations

Evidence and attribution. Theorem 1 characterizes homogeneous branchwise linear maps at fixed K; it implies neither nonlinear universality nor superiority over pair-aware attention. The

Table 3: Separately trained conformer-cap study over three random-scaffold splits. Values are mean $\pm$ sample standard deviation; molecules with fewer available conformers use their actual count, and $K = 5$ is the default. Bold and underlining denote the best and second-best means, respectively.

Conformer cap K	ESOL RMSE ↓	FreeSolv RMSE ↓	BACE ROC-AUC (%) ↑
1	0.83 ± 0.12	1.78 ± 0.31	84.99 ± 2.76
3	0.78 ± 0.12	$\underline{1.71 \pm 0.20}$	84.55 ± 4.82
5	$\mathbf{0.76 \pm 0.06}$	$\mathbf{1.70 \pm 0.42}$	$\underline{85.54 \pm 3.41}$
10	$\underline{0.77 \pm 0.07}$	1.72 ± 0.24	$\mathbf{86.80 \pm 3.42}$

Table 4: Local model-pass resource measurements on 100 molecules from ESOL training split 0 using one RTX 5070 Ti, FP32, and batch size one. Memory is peak CUDA allocated memory after optimizer-state initialization; latency is the mean of five repeat means. Optimizer steps and offline preparation are excluded; repeat-level SDs are unavailable.

Model	K	Peak train memory (GiB) ↓	Fwd.+bwd. (ms/mol) ↓	Inference (ms/mol) ↓
FACET	1	0.04	17.40	9.71
FACET	3	0.07	18.09	10.22
FACET	5	0.10	17.77	9.65
FACET	10	0.19	17.77	9.74
PaSTNet	1	0.06	22.50	11.80
PaSTNet	3	0.15	27.50	14.20
PaSTNet	5	0.28	35.50	18.50
PaSTNet	10	0.78	60.00	31.00

subtraction ablation favors singleton centering in mean performance. Geometry placement, parity coupling, individual contraction families, the spatial bypass, and final refinement remain unisolated; no capacity-balanced pair-local variant or alternative pair-state comparator is reported. Baseline differences in pretraining, capacity, tuning, and target preprocessing limit operator-level attribution. Three scaffold splits with fixed initialization do not establish statistical significance; more split and initialization replicates are needed for stronger uncertainty analysis.

Physical and computational scope. Geometry-selected conformers and pathwise pooling weights are not Boltzmann populations, so predictive accuracy does not validate physical ensemble weights; evaluation on energy-annotated ensembles such as GEOM (Axelrod & Gómez-Bombarelli, 2022) remains open. The ESOL scaffold construction does not test held-out acyclic generalization, atom-relabeling invariance assumes consistent descriptor transport, and signed torsions permit reflection sensitivity. Explicit pair states require $\mathcal{O}(PK^2d_r)$ storage, with explicit hydrogens further increasing P. Profiling covers one ESOL subset through $K = 10$, where PaSTNet has higher model-pass costs than FACET; these measurements are not end-to-end runtime comparisons. Micro-batching does not remove the pair-storage requirement of an individual molecule. The bond scale is positive across the nine diagnostic splits for ESOL, FreeSolv, and BACE; a positive-floor alternative remains to be evaluated.

6 Conclusions

PaSTNet introduces path-aligned conformer-pair reasoning with complete homogeneous branch-wise linear mixers at fixed cardinality and a singleton-centered residual. Across six molecular property datasets, it achieves the best evaluated mean score on five tasks, while comparisons with identity, same-path attention, pair-local, and no-subtraction variants favor full PaST. These results support explicit relational computation at aligned molecular paths for conformer-ensemble learning. Finer component attribution, initialization uncertainty, physical ensemble validation, and efficient scaling beyond ten conformers remain open.

REPRODUCIBILITY STATEMENT

Section 3 and the method appendices specify the architecture, tensor operators, and symmetry assumptions. Section 4 and Appendix G document training, selection, and comparability limits. The result tables, completeness proofs, finite-size algebra checks, and synthetic-check procedures and outcomes are reported within this PDF. An anonymized repository provides the default E0 model and training pipeline at https://anonymous.4open.science/r/PaSTNet-6DD7/. The linked code supports up to five conformers; the separate $K = 10$, baseline, and interaction-control runners are not included. The local algebra-verification and additional synthetic-check scripts are also outside that release. A complete per-run checkpoint/configuration manifest and paired split scores for all table entries are not provided. The reported checks support their stated mathematical and numerical scope, not independent reproduction of every benchmark comparison.

ETHICS STATEMENT

The study uses public molecular property datasets and computationally generated conformers. It assesses predictions on the specified benchmark partitions. No human-participant study or prospective chemical intervention is reported.

AI USE STATEMENT

Generative AI assisted manuscript wording, LaTeX organization, literature lookup, and comparison of the method description with the implementation and available experimental records. It also assisted earlier encoder revisions, tests, and interpretation of experimental results. The authors are responsible for the final text, results, claims, and artifacts.

REFERENCES

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4. URL https://www.nature.com/articles/s41597-022-01288-4.

Simon Axelrod and Rafael Gómez-Bombarelli. Molecular machine learning with conformer ensembles. *Machine Learning: Science and Technology*, 4(3):035025, 2023. doi: 10.1088/2632-2153/acefa7. URL https://arxiv.org/abs/2012.08452.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928.

Leon Bergen, Timothy J. O'Donnell, and Dzmitry Bahdanau. Systematic generalization with edge transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2112.00578.

Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yu Guang Wang, Pietro Liò, Guido Montúfar, and Michael Bronstein. Weisfeiler and Lehman go cellular: CW networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/157792e4abb490f99dbd738483e0d2d4-Abstract.html.

Bryan Cheng, Austin Jin, and Jasper Zhang. When three-dimensional conformer ensembles improve molecular property prediction beyond two-dimensional fingerprints: A systematic study. *arXiv preprint arXiv:2606.08825*, 2026. URL https://arxiv.org/abs/2606.08825.

John S. Delaney. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, 2004. doi: 10.1021/ci034243x. URL https://doi.org/10.1021/ci034243x.

Cameron Diao and Ricky Loynd. Relational attention: Generalizing transformers for graph-structured tasks. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.05062.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2003.03123.

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. URL https://arxiv.org/abs/2106.08903.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1263–1272. PMLR, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Thomas A. Halgren. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry*, 20(7):720–729, 1999. doi: 10.1002/(SICI)1096-987X(199905)20:7<720::AID-JCC7>3.0.CO;2-X.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Longlong Li, Xiang Liu, Guanghui Wang, Yu Guang Wang, and Kelin Xia. Path complex neural network for molecular property prediction. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2024. URL https://openreview.net/forum?id=FlnGcMp6FL.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.09902.

Kisung Moon, Hyeon-Jin Im, and Sunyoung Kwon. 3D graph contrastive learning for molecular property prediction. *Bioinformatics*, 39(6):btad371, 2023. doi: 10.1093/bioinformatics/btad371. URL https://doi.org/10.1093/bioinformatics/btad371.

Duy M. H. Nguyen, Trung Q. Nguyen, Ha T. H. Le, Mai Thanh Nhat Truong, TrungTin Nguyen, Nhat Ho, Khoa D. Doan, Duy Duong-Tran, Li Shen, Daniel Sonntag, James Zou, Mathias Niepert, Hyojin Kim, and Jonathan E. Allen. FACET: A fragment-aware conformer ensemble transformer. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cpwbXHvd2h.

Duy Minh Ho Nguyen, Nina Lukashina, Tai Nguyen, An Thai Le, Trungtin Nguyen, Nhat Ho, Jan Peters, Daniel Sonntag, Viktor Zaverkin, and Mathias Niepert. Structure-aware E(3)-invariant molecular conformer aggregation networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 37736–37760. PMLR, 2024. URL https://proceedings.mlr.press/v235/nguyen24g.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, III, and W. M. Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 1992. doi: 10.1021/ja00051a040. URL https://doi.org/10.1021/ja00051a040.

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(12): 2562–2574, 2015. doi: 10.1021/acs.jcim.5b00654. URL https://doi.org/10.1021/acs.jcim.5b00654.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9377–9388. PMLR, 2021. URL https://proceedings.mlr.press/v139/schutt21a.html.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, 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. 992–1002, 2017. URL https://arxiv.org/abs/1706.08566.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Touvron_Going_Deeper_With_Image_Transformers_ICCV_2021_paper.html.

Shuzhe Wang, Jagna Witek, Gregory A. Landrum, and Sereina Riniker. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *Journal of Chemical Information and Modeling*, 60(4):2044–2058, 2020. doi: 10.1021/acs.jcim.0c00025. URL https://doi.org/10.1021/acs.jcim.0c00025.

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*, 9:513–530, 2018. doi: 10.1039/C7SC02664A. URL https://pubs.rsc.org/cs/content/articlelanding/2018/sc/c7sc02664a.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120:145301, 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://arxiv.org/abs/1710.10324.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1910.07467.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Yanqiao Zhu, Jeehyun Hwang, Keir Adams, Zhen Liu, Bozhao Nan, Brock Stenfors, Yuanqi Du, Jatin Chauhan, Olaf Wiest, Olexandr Isayev, Connor W. Coley, Yizhou Sun, and Wei Wang. Learning over molecular conformer ensembles: Datasets and benchmarks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.00115.

Yanqiao Zhu, Yidan Shi, Yuanzhou Chen, Fang Sun, Yizhou Sun, and Wei Wang. Symmetry-preserving conformer ensemble networks for molecular representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/62891522c00cf7323cbacb500e6cfc8d-Abstract-Conference.html.

A COMPLETE GEOMETRIC REPRESENTATION

This appendix specifies the backbone-equipped PaSTNet. Coordinates are in angstroms and angles in radians. All conformers use one atom correspondence and one enumeration of simple covalent paths of lengths zero through three, unique modulo reversal. The enumeration retains explicit hydrogens present in the supplied molecular graph. These atoms contribute to the spatial graph and to every applicable path order; their computational cost is discussed in Appendix E. For reference, the principal shapes and normalization conventions are:

Symbol	Meaning and shape (single molecule)
K, P_r	Actual conformer count and number of paths of order r.
H_r, h_c	Path states $K \times P_r \times 32$; one fixed-path state in $\mathbb{R}^{32}$.
$R_p = (S_p, A_p)$	Symmetric/antisymmetric pair states, each $K \times K \times 16$.
$\bar{h}_c = N^C_{s,r}(h_c)$	Stage/order-specific content RMSNorm, width 32.
$N^I_{s,r}, N^{\mathrm{geo}}_r$	Geo-IH content RMSNorm (32) and shared geometry RMSNorm (12).
N_+, N_-	Pair-lift RMSNorms (16), with stage/order indices suppressed.
$N^{\pm}_{s,r,t}$	Tensor-layer RMSNorms (16), no additive bias.
$m^{\mathrm{full}}, m^{\mathrm{self}}, m^{\mathrm{cross}}$	Collapsed messages, each $K \times 16$.

Stage s, order r, path p, and tensor-layer t indices are suppressed only where stated. Contractions use actual K, not padded capacity. No conformer is duplicated to reach a configured cap. The default cap is five; the conformer-count and resource studies also evaluate a cap of ten.

A.1 ATOMS AND BONDS

The atom feature $F_{0,c,i} \in \mathbb{R}^{92}$ is the fixed CGCNN-style elemental descriptor returned by the JARVIS feature adapter for the atom's chemical element (Xie & Grossman, 2018). This specifies an input descriptor, not a CGCNN prediction baseline. It is constant across conformers and is not an atom-specific charge, local environment, or chirality descriptor. The learned SchNet atom states already provide spatial context when this descriptor enters the path adapter. For a bond $p = (i, j)$, write $\ell = \|x_j - x_i\|_2$ and let $b_{ij} \in \mathbb{R}^{21}$ denote the bond descriptor. Its 21 slots contain seven bond-direction indicators, four bond-type indicators, two nominal-length metadata values, two ring-membership indicators, and six zero entries. The exact adapter indexes the type by the integer part of the numerical bond order minus one, so single and aromatic bonds occupy the same type slot. Its nominal-length lookup receives an RDKit bond-type object although its keys are strings; the implemented lookup therefore returns its default 1.0, making both nominal-length slots constant. These metadata values are distinct from the measured conformer-dependent length ℓ, which is encoded as follows:

$$F_{1,c,p} = [b_{ij} \parallel R_0(\ell) \parallel \cdots \parallel R_{15}(\ell)], \quad R_m(\ell) = \exp\left[-\left(\frac{15}{4}\right)^2 \left(\ell - \frac{4m}{15}\right)^2\right]. \tag{14}$$

We preserve this evaluated adapter for all reported PaSTNet results; no corrected bond-type or nominal-length lookup is substituted after training. The constant slots supply no bond-specific information. Adapter tests cover single, aromatic, double, and triple bonds, including the distinction between string lookup keys and the actual RDKit inputs. The 16 centers are fixed between 0 and 4 angstroms, and the output width is 37. The basis is evaluated continuously even for a distance outside that interval. The direction category is attached to the supplied RDKit bond object; it is not computed from the lexicographic orientation of the path. Atom relabeling transports this bond metadata with its bond. This convention is different from reassigning wedge directions or re-parsing an independently serialized molecular drawing.

A.2 ANGLES

For $p = (i, j, k)$, set $a = x_j - x_i$, $b = x_k - x_j$, and $o = (x_i + x_j + x_k)/3$. The conventional interior bond angle is

$$\theta = \operatorname{atan2}\big(\|(-a) \times b\|, (-a)^\top b\big), \qquad v_\theta = \mathbf{1}\{\|a\| > \epsilon_{\mathrm{geo}},\ \|b\| > \epsilon_{\mathrm{geo}}\}, \quad \epsilon_{\mathrm{geo}} = 10^{-8}. \tag{15}$$

When $v_\theta = 0$, a finite placeholder $\theta = 0$ is used and all angular encoding channels are masked to zero. Nonzero collinear bonds define a valid angle of zero or π. Define

$$E_\theta(\theta) = v_\theta[\theta/\pi, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, 1], \tag{16}$$

$$d_i = \|x_i - o\|, \quad d_j = \|x_j - o\|, \quad d_k = \|x_k - o\|, \qquad A_\triangle = \tfrac{1}{2}\|a \times (a+b)\|. \tag{17}$$

The complete 13-dimensional feature is

$$F_{2,c,p} = [E_\theta(\theta) \parallel d_i, d_j, d_k \parallel \|a\|, \|b\|, \|a+b\| \parallel A_\triangle]. \tag{18}$$

Only the angular block is validity-masked; finite distance and area channels remain.

A.3 Signed torsions

For $p = (i, j, k, l)$, set $a = x_j - x_i$, $b = x_k - x_j$, $c = x_l - x_k$. A torsion is valid when all three edge lengths exceed ϵ_{geo} and

$$\|a \times b\| > \epsilon_{\text{geo}}\|a\|\|b\|, \qquad \|b \times c\| > \epsilon_{\text{geo}}\|b\|\|c\|. \tag{19}$$

Let v_ϕ be the indicator of these conditions. On valid paths,

$$n_1 = \frac{a \times b}{\|a \times b\|}, \qquad n_2 = \frac{b \times c}{\|b \times c\|}, \qquad \widehat{b} = \frac{b}{\|b\|}, \qquad \phi = \operatorname{atan2}\left((n_1 \times n_2)^\top \widehat{b}, n_1^\top n_2\right). \tag{20}$$

On invalid paths $\phi = 0$ is a finite placeholder and all seven torsional encoding channels are zero:

$$E_\phi(\phi) = v_\phi[\cos\phi, \sin\phi, \cos 2\phi, \sin 2\phi, \cos 3\phi, \sin 3\phi, 1]. \tag{21}$$

The remaining descriptors are

$$V = \frac{|(a \times (a+b))^\top (a+b+c)|}{6}, \qquad d_{il} = \|a+b+c\|, \qquad A_4 = \frac{1}{2}\|(x_j - x_l) \times (x_k - x_i)\|. \tag{22}$$

We refer to A_4 as a cross-diagonal area descriptor. For nonplanar quadrilaterals it is not the area of a uniquely specified surface. The full width is 12:

$$F_{3,c,p} = [V \parallel E_\phi(\phi) \parallel d_{il} \parallel A_4 \parallel (\|a\| + \|b\|) \parallel (\|b\| + \|c\|)]. \tag{23}$$

Invalid angular geometry does not remove a conformer or suppress its content interactions. Nonfinite coordinates are rejected, rather than silently replaced.

A.4 Learned reversal averaging

An atom permutation can change which of p and $\mathrm{rev}(p)$ is lexicographically first. The encoder therefore treats these two representatives as one reversal orbit. Let J_r act on the HRGE feature vector of order r. For the descriptors above, $J_0 = I_{92}$ and $J_1 = I_{37}$, since element and transported bond metadata and measured bond length do not depend on the representative's direction. Using zero-based channel indices,

$$J_2 F = F[(0, 1, 2, 3, 4, 5, 8, 7, 6, 10, 9, 11, 12)], \tag{24}$$

$$J_3 F = F[(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 10)]. \tag{25}$$

For P_2, the exchanges are $d_i \leftrightarrow d_k$ and $\|a\| \leftrightarrow \|b\|$ simultaneously. For P_3, only $\|a\| + \|b\| \leftrightarrow \|b\| + \|c\|$ is exchanged. Reversing the torsion gives $(a', b', c') = (-c, -b, -a)$, $(n_1', n_2') = (-n_2, -n_1)$, and $\widehat{b}' = -\widehat{b}$. Hence both $(n_1' \times n_2')^\top \widehat{b}'$ and $(n_1')^\top n_2'$ equal their original values. Validity conditions also agree. The signed torsion and its encoding are therefore unchanged, including for the masked invalid case. Distances, absolute volume, and A_4 are unchanged as well. Bond, angle, and torsion inputs to Geo-IH and their relative PaST encodings inherit this reversal invariance.

Let $u_i \in \mathbb{R}^{128}$ be the final SchNet atom state in the current conformer. The path adapter first constructs the following reversal-invariant spatial summaries:

$$s_{0,i} = u_i, \tag{26}$$

$$s_{1,(i,j)} = [u_i + u_j \parallel |u_i - u_j|], \tag{27}$$

$$s_{2,(i,j,k)} = [u_j \parallel u_i + u_k \parallel |u_i - u_k|], \tag{28}$$

$$s_{3,(i,j,k,l)} = [u_i + u_l \parallel |u_i - u_l| \parallel u_j + u_k \parallel |u_j - u_k|]. \tag{29}$$

Their widths are $(128, 256, 384, 512)$. With one order-specific shared encoder,

$$e_r(s, F) = \mathrm{LN}_r(\mathrm{SiLU}(W_r[s \,\|\, F] + b_r)), \quad e_r : \mathbb{R}^{D_r^{\mathrm{in}}} \to \mathbb{R}^{32}, \tag{30}$$

the total input widths are $(D_0^{\mathrm{in}}, D_1^{\mathrm{in}}, D_2^{\mathrm{in}}, D_3^{\mathrm{in}}) = (220, 293, 397, 524)$. Equation (1) averages the two *encoded* orientations. For every F,

$$\tfrac{1}{2}[e_r(s, J_r F) + e_r(s, J_r^2 F)] = \tfrac{1}{2}[e_r(s, J_r F) + e_r(s, F)]. \tag{31}$$

This holds for arbitrary encoder parameters and preserves joint dependence among the features inside each orientation before averaging. In particular, it does not require averaging individual raw endpoint channels separately. The two evaluations share all weights, including LayerNorm, so the reversal average adds no trainable parameters. For $r = 0, 1$, the identical encoder evaluations are reused.

A.5 Relative geometry for PaST

Relative geometry is computed only between conformers of the same path. The ordered-pair convention is $\Delta t_{cc'} = t_c - t_{c'}$, including diagonal pairs. For atoms both geometry blocks are empty.

For bonds, the scale is the training-set statistic

$$s_\ell = Q_{.95}\big(\{|\ell_{m,c,p} - \ell_{m,c',p}| : m \in \mathcal{D}_{\mathrm{train}},\ p \in P_1(m),\ 1 \le c, c' \le K_m\}\big). \tag{32}$$

The implementation uses float64 and a linearly interpolated empirical quantile. Every ordered pair, including $c = c'$, is one sample; consequently singleton molecules contribute zeros and molecules with more bonds or conformers contribute more samples. There is no subtraction across different bonds or molecules. The fitted nonnegative scalar is frozen for validation, testing, and inference. An empty collection of training bond samples raises an error. The reported implementation does not impose a positive minimum scale; its only denominator stabilizer is $\epsilon_\ell = 10^{-8}$. Define

$$u_{cc'} = \frac{\ell_c - \ell_{c'}}{s_\ell + \epsilon_\ell}, \qquad R_j^{\Delta}(|u|) = \exp\left[-\left(\frac{7}{2}\right)^2 \left(|u| - \frac{2j}{7}\right)^2\right], \quad j = 0, \ldots, 7. \tag{33}$$

Then $\Gamma^+_{1,cc'} = [R_j^{\Delta}(|u_{cc'}|)]_{j=0}^{7}$ and $\Gamma^-_{1,cc'} = u_{cc'}\Gamma^+_{1,cc'}$, each of width 8.

Calibration limitation. Diagonal zeros can drive Equation (32) to zero when singleton or geometrically identical ensembles dominate the training sample. The resulting large normalized differences can suppress the Gaussian features; the small additive epsilon does not solve this statistical degeneracy. A robustness extension would estimate $Q_{.95}$ over $c \neq c'$ only and impose a positive, training-specified scale floor, using that floor also when no off-diagonal samples exist. Such a change requires a new calibration and evaluation and is not part of the reported PaSTNet results. Singletons must still retain their true $K = 1$ when applying PaST.

Diagnostics on the evaluated training splits. The diagonal-inclusive s_ℓ is strictly positive on all nine evaluated training splits of ESOL, FreeSolv, and BACE for the default five-conformer cache (Table 5), ranging from 3.01×10^{-3} to 6.49×10^{-3} Å. The off-diagonal-only $Q_{.95}$ is approximately 14–18% higher than the current statistic. Although roughly one quarter of the training molecules in ESOL and FreeSolv are singletons, diagonal zeros do not cause the zero-scale failure mode in these default-cap experiments. An off-diagonal quantile with a positive floor remains a robustness extension for future singleton-dominated datasets, not a change to the reported model or its predictive results.

For angle or torsion order $r \in \{2, 3\}$, let $t = \theta$ or ϕ, respectively, and $v_{cc'} = v_c v_{c'}$. The unified definition is

$$\Gamma^+_{r,cc'} = [v_{cc'}\cos(\Delta t_{cc'}), \ldots, v_{cc'}\cos(r\Delta t_{cc'}), v_{cc'}], \tag{34}$$

$$\Gamma^-_{r,cc'} = [v_{cc'}\sin(\Delta t_{cc'}), \ldots, v_{cc'}\sin(r\Delta t_{cc'})]. \tag{35}$$

Validity is applied before trigonometric evaluation when placeholders may be nonfinite. The implemented cached features themselves use finite placeholders. These definitions give $\Gamma^+_{cc'} = \Gamma^+_{c'c}$ and $\Gamma^-_{cc'} = -\Gamma^-_{c'c}$ exactly in real arithmetic. For torsions the harmonics avoid a discontinuity when equivalent angle representatives differ by 2π.

Table 5: Bond-scale diagnostics on the nine evaluated training splits with the default five-conformer cache. Both scale statistics are in Å; current s_ℓ uses Equation (32), whereas the off-diagonal quantile excludes $c = c'$. Singleton counts are relative to the number of training molecules. These diagnostics do not change the reported calibration or predictive results.

Dataset	Split	Current s_ℓ	Off-diagonal $Q_{.95}$	Singleton molecules
ESOL	0	0.004793011	0.005643504	226/904
ESOL	1	0.005637486	0.006640276	213/904
ESOL	2	0.005186476	0.006056724	213/904
FreeSolv	0	0.003120728	0.003565268	134/515
FreeSolv	1	0.003007472	0.003417192	149/515
FreeSolv	2	0.003051150	0.003473988	141/515
BACE	0	0.006321593	0.007257510	2/1179
BACE	1	0.006494845	0.007459786	2/1179
BACE	2	0.006493046	0.007445341	3/1179

B Complete Geo-IH and refinement operators

For compactness, fix stage s, conformer c, and relation order r. Let $\mathcal{N}^-(u) = \{(w, q) : w \to u \mid q\}$ be a multiset of incoming relations; distinct cofaces remain distinct even when they connect the same two paths. At each substep, normalize the current states using the stage/order-specific maps

$$x_u = N^I_{s,r}(H_{r,u}), \qquad e_p = N^I_{s,r+1}(H_{r+1,p}), \tag{36}$$

$$\bar{e}_u = \begin{cases} |\mathcal{N}^-(u)|^{-1} \sum_{(w,q)\in\mathcal{N}^-(u)} e_q, & \mathcal{N}^-(u) \neq \varnothing, \\ 0, & \mathcal{N}^-(u) = \varnothing. \end{cases} \tag{37}$$

RMSNorm throughout the deep backbone has epsilon 10^{-6}, a learned channelwise multiplicative weight, and no additive bias (Zhang & Sennrich, 2019). The absolute geometry bank uses

$$\Gamma^I_{0,p} = R(\ell_p), \quad \Gamma^I_{1,p} = E_\theta(\theta_p), \quad \Gamma^I_{2,p} = E_\phi(\phi_p), \tag{38}$$

$$g_p = N^{\mathrm{geo}}_r\big(W^{\mathrm{geo}}_{r,2}\,\mathrm{SiLU}(W^{\mathrm{geo}}_{r,1}\Gamma^I_{r,p} + b^{\mathrm{geo}}_{r,1}) + b^{\mathrm{geo}}_{r,2}\big). \tag{39}$$

Its hidden width is 32 and output width 12. The entire bank, including its three normalization maps and three learned geometry strengths, is shared across all three stages; its differentiable embeddings are computed once and reused.

Each of four heads, $a = 1, \ldots, 4$, has width 8 and first forms a content relation

$$c^a_{u\to v|p} = W^a_i x_u + W^a_j x_v + W^a_e e_p + b^a_e. \tag{40}$$

The source, destination, and coface transforms each map 32 channels to 32 and reshape the result into four 8-channel heads; W^a_i, W^a_j, W^a_e denote their head slices. Head-specific FiLM parameters are $(\gamma^a_p, \beta^a_p, \lambda^a_p, \mu^a_p) = a_{\mathrm{geo},r}\tanh(W^a_F g_p + b^a_F)$, where the learned order scalar $a_{\mathrm{geo},r}$ starts at 0.1. One $12 \to 128$ affine map produces four modulation vectors of width 8 for each of four heads. Geometry conditions routing and message values:

$$z^a_{u\to v|p} = \mathrm{SiLU}((1+\gamma^a_p)\odot c^a_{u\to v|p} + \beta^a_p), \qquad \omega^a_{u\to v|p} = \mathrm{softmax}_{(u,p):u\to v|p}((\eta^a)^\top z^a_{u\to v|p}), \tag{41}$$

$$t^a_u = W^a_N[x_u \,\|\, \bar{e}_u] + b^a_N, \qquad m^a_{u\to v|p} = (1+\lambda^a_p)\odot t^a_u + \mu^a_p. \tag{42}$$

Writing $\mathrm{concat}^4_{a=1}$ for concatenation of the head outputs, the lower update uses a bias-free $32 \to 32$ output projection W_O; the coface update concatenates the averaged relation states directly:

$$\Delta H_{r,v} = W_O\,\mathrm{concat}^4_{a=1}\left(\sum_{(u,p):u\to v|p} \omega^a_{u\to v|p} m^a_{u\to v|p}\right), \tag{43}$$

$$\Delta H_{r+1,p} = \mathrm{concat}^4_{a=1}\left[\tfrac{1}{2}(z^a_{u\to v|p} + z^a_{v\to u|p})\right]. \tag{44}$$

A single coface's geometry is used for both directed orientations. Incoming softmax normalization is over relations with a common destination, separately for each conformer and head. The context in the value transform belongs to the source. Both lower and upper increments are computed before either is added:

$$H_r \leftarrow H_r + a^I_{s,r} \odot \Delta H_r, \qquad H_{r+1} \leftarrow H_{r+1} + a^I_{s,r+1} \odot \Delta H_{r+1}. \tag{45}$$

The channelwise vectors $a^I_{s,r}$ start at 0.1. An order uses its same stage-specific norm and scale when it acts first as a lower path and later as a coface. Orders P_1 and P_2 can therefore receive two residual additions during one sweep. An empty coface set yields zero increments and its deep substep can be skipped.

Stochastic depth is disabled in all three stages during training and evaluation. There is no normalization after a deep residual addition. Binary conformer masks remove padded states from the routing and residual outputs; they are unrelated to angular validity masks, which suppress only undefined geometry channels.

B.1 Topology-only final sweep

After Equation (12), one new, unshared interaction module per relation order uses the pooled content directly. It follows the same incidence, source-context, routing, and bidirectional coface averaging pattern, with four 8-channel heads. There is no geometry encoder or FiLM modulation, and the residual uses post-normalization. LeakyReLU(0.01) is applied to both the content relation and the aggregated values before the output projection. This retains the activation convention of the existing topology refinement implementation. It is an implementation choice, not required by the symmetry proofs; no isolated activation ablation is reported. Explicitly, with pooled or current refined states x_u, e_p,

$$z^a_{u\to v|p} = \mathrm{LeakyReLU}_{.01}(W^a_i x_u + W^a_j x_v + W^a_e e_p + b^a_e), \tag{46}$$

$$\omega^a_{u\to v|p} = \mathrm{softmax}_{(u,p):u\to v|p}((\eta^a)^\top z^a_{u\to v|p}), \qquad t^a_u = W^a_N[x_u \,\|\, \bar{e}_u] + b^a_N, \tag{47}$$

$$\Delta H_{r,v} = W_O \,\mathrm{concat}^4_{a=1}\, \mathrm{LeakyReLU}_{.01}\left(\sum_{(u,p):u\to v|p} \omega^a_{u\to v|p} t^a_u\right), \tag{48}$$

$$\Delta H_{r+1,p} = \mathrm{concat}^4_{a=1}\left[\tfrac{1}{2}(z^a_{u\to v|p} + z^a_{v\to u|p})\right], \tag{49}$$

$$H_q \leftarrow \mathrm{LN}_q(H_q + \Delta H_q), \qquad q = r, r+1. \tag{50}$$

The sweep order is again $r = 2, 1, 0$. Its P_3 update is retained for readout. Unlike a skipped empty deep substep, this refinement still applies its stated LayerNorm when the increment is zero, including for isolated lower paths. No coordinates, geometry encodings, or choice of representative conformer enter this final module.

C Complete PaST tensor core

All operations below apply independently to one path. Structural mixers use row-vector channel notation; pointwise MLP formulas use column vectors. We omit stage and order indices on their separate weights. Path states have width $d_p = 32$; pair states and collapsed messages have width $d_r = 16$. The even pair MLP has input widths $(96, 104, 99, 100)$ and shape $\bullet \to 32 \to 16$; the odd pair MLP has widths $(32, 40, 34, 35)$ and shape $\bullet \to 16 \to 16$, with an additional $16 \to 16$ even gate. These widths follow from Equations (5)–(6). The input content uses a stage/order-specific RMSNorm $\bar{h}_c = N^C_{s,r}(h_c)$. The relative even/odd widths for orders 0–3 are $(0,0), (8,8), (3,2), (4,3)$. The even lift uses SiLU between its two affine maps. The odd lift is

$$f_-(u) = W_2[z \odot \sigma(W_g|z| + b_g)], \qquad z = W_1 u, \tag{51}$$

where W_1, W_2 have no biases, making $f_-(-u) = -f_-(u)$. The output normalizations N_+, N_- are RMSNorm without additive bias. After lifting, numerical projections $S \leftarrow (S + S^\top)/2$ and $A \leftarrow (A - A^\top)/2$ enforce the respective parities.

C.1 Structural maps and channel mixing

For a symmetric tensor S and antisymmetric tensor A, define

$$d_c = S_{cc}, \qquad \rho_c = \frac{1}{K}\sum_j S_{cj}, \qquad \tau = \frac{1}{K}\sum_j S_{jj}, \qquad g = \frac{1}{K^2}\sum_{ij} S_{ij}, \qquad q_c = \frac{1}{K}\sum_j A_{cj}. \tag{52}$$

All sums include diagonal entries and all denominators use the actual K. The nine symmetric maps, listed explicitly, are

$$B_1(S)_{cc'} = S_{cc'}, \qquad B_2(S)_{cc'} = \delta_{cc'} d_c, \qquad B_3(S)_{cc'} = d_c + d_{c'}, \tag{53}$$

$$B_4(S)_{cc'} = \delta_{cc'}\rho_c, \qquad B_5(S)_{cc'} = \rho_c + \rho_{c'}, \qquad B_6(S)_{cc'} = \delta_{cc'}\tau, \tag{54}$$

$$B_7(S)_{cc'} = \tau, \qquad B_8(S)_{cc'} = \delta_{cc'} g, \qquad B_9(S)_{cc'} = g. \tag{55}$$

The mixer is $L^+_{22}(S)_{cc'} = \sum_{b=1}^{9} B_b(S)_{cc'} W_b$, with independent $W_b \in \mathbb{R}^{16\times 16}$ and no biases. The odd mixer is $L^-_{22}(A)_{cc'} = A_{cc'} V_1 + (q_c - q_{c'})V_2$, also bias-free. At each fixed K, the nine maps span all homogeneous linear permutation-equivariant maps on symmetric pair tensors. They are independent for $K \geq 4$; the dimensions are 1, 5, and 8 at $K = 1, 2, 3$, respectively. The two odd maps likewise span the homogeneous linear equivariant operators on antisymmetric pair tensors, and are independent for $K \geq 3$. Appendix F proves these claims. Sharing the same channel matrices across different K imposes a consistent parameterization; fixed-K completeness does not mean arbitrary independently chosen operators at all K can be represented by one set of shared matrices. No nonlinear-network universality or expressivity ordering relative to attention follows from these linear results.

C.2 Two coupled tensor layers

Each tensor layer $t \in \{1, 2\}$ has its own mixer, coupling, and feed-forward weights. From the same incoming S, A, compute

$$Z^+ = L^+_{22}(S) + (A \odot A)W_{AA}, \qquad Z^- = L^-_{22}(A) + (S \odot A)W_{SA}. \tag{56}$$

Let

$$F^+_t(x) = W^+_{t,2}\,\mathrm{SiLU}(W^+_{t,1}x + b^+_{t,1}) + b^+_{t,2}, \tag{57}$$

$$F^-_t(x) = W^-_{t,2}\big[z \odot \sigma(W_{t,g}|z| + b_{t,g})\big], \qquad z = W^-_{t,1}x. \tag{58}$$

Both feed-forward hidden widths are 32; the first and last odd transformations have no biases. The updates are

$$S' = S + \eta^+_{s,r,t} \odot F^+_t(N^+_{s,r,t}(Z^+)), \qquad A' = A + \eta^-_{s,r,t} \odot F^-_t(N^-_{s,r,t}(Z^-)), \tag{59}$$

followed by $S' \leftarrow (S' + S'^\top)/2$ and $A' \leftarrow (A' - A'^\top)/2$. Both $N^\pm$ are RMSNorm, and the channelwise $\eta^\pm$ are initialized to 0.1. The two layers have distinct weights. All three stages and all four path orders have separate pair encoders, tensor layers, collapse maps, norms, and residual scales. Within a stage/order block, the same weights act on every path and conformer pair. There are no stochastic operations inside either tensor core or singleton evaluation.

C.3 Collapse and exact singleton evaluation

Recompute Equation (52) from the final pair tensors. Five bias-free matrices produce m^{full}_c in Equation (10). For each conformer c, set

$$S^{(0,c)}_{\mathrm{single}} \in \mathbb{R}^{1\times 1\times 16}, \quad (S^{(0,c)}_{\mathrm{single}})_{11} = S^{(0)}_{cc}, \qquad A^{(0,c)}_{\mathrm{single}} = 0. \tag{60}$$

The same two tensor layers, stage-specific states, and collapse matrices operate on this singleton. All contractions now use $K = 1$. Crucially, the input is the *initial* lifted diagonal. Removing diagonal entries from the final full tensor would not perform this counterfactual singleton computation, because diagonal entries have already received global and row context. Both m^{full}_c and m^{self}_c retain their gradient graphs before subtraction. The gated update in Equation (11) is

$$h'_c = h_c + a^C_{s,r} \odot \sigma\big(W_g[\bar{h}_c \,\|\, m^{\mathrm{cross}}_c] + b_g\big) \odot W_{\mathrm{up}} m^{\mathrm{cross}}_c. \tag{61}$$

Here the gate is a $48 \to 32$ affine map followed by a sigmoid, and W_{up} is a bias-free $16 \to 32$ linear map. Both consume the unnormalized cross message; only h_c is RMS-normalized before concatenation. Since $W_{\mathrm{up}}0 = 0$, the update vanishes at zero cross message. The outer channelwise scale starts at 0.1. This defines the zero-preserving function $u_{s,r}$ in the main text. For an actual singleton in a padded batch, the implementation reuses the computed singleton response as the full response before subtraction, eliminating a possible roundoff discrepancy between equivalent kernels without changing the algebraic operator. The singleton evaluation is a deterministic comparison between two inputs to the same operator, not an intervention on conformer generation, a physical ensemble, or the target.

D SYMMETRY STATEMENTS AND PROOFS

The following are algebraic properties of the specified operators in real arithmetic. Floating-point implementations are subject to rounding. They are not empirical claims about prediction accuracy, the usefulness of conformer diversity, or causal interpretation.

Proposition 1 (exchange parity). *At pair lifting and after every tensor layer, $S_{cc'} = S_{c'c}$ and $A_{cc'} = -A_{c'c}$; in particular $A_{cc} = 0$.*

Proof. The even lifted input is symmetric because sum, absolute difference, componentwise product, and Γ^+ are symmetric under exchanging indices. The odd lifted input changes sign. For $f_-(u) = W_2[z \odot \sigma(W_g|z| + b_g)]$, $z = W_1 u$, bias-free W_1 sends $-u$ to $-z$, the gate stays unchanged, and bias-free W_2 yields $f_-(-u) = -f_-(u)$. RMSNorm without additive bias is odd since its denominator depends only on squared values. Thus lifting has the stated parities.

Every $B_b(S)$ is symmetric, including its diagonal-restricted terms, and $A_{cc'}V_1 + (q_c - q_{c'})V_2$ is antisymmetric. Entrywise products satisfy $(A \odot A)_{c'c} = (A \odot A)_{cc'}$ and $(S \odot A)_{c'c} = -(S \odot A)_{cc'}$. The even feed-forward transform preserves equality of symmetric entries; the odd transform and odd RMSNorm preserve their sign relation. Residual sums and channelwise scales preserve both properties. The explicit transpose projections also enforce these subspaces. Induction over the two tensor layers completes the claim. □

Proposition 2 (conformer equivariance and invariant prediction). *Let π simultaneously permute the conformer axis of all content, coordinates, and validity inputs of a molecule. Every pre-pooling stage is equivariant to π, and the final prediction is invariant. Stochastic depth is disabled.*

Proof. The distance-based spatial backbone, feature construction, and initial projections act separately on each conformer with shared functions. Geo-IH uses the same molecular incidence and weights in every conformer, so permuting conformers permutes its outputs. Consider PaST at one fixed path. Pair lifting maps a permutation matrix Π to the pair action $S \mapsto \Pi S \Pi^\top$, $A \mapsto \Pi A \Pi^\top$, independently in each channel. The diagonal and row summaries permute by Π, while τ and g are unchanged. The Kronecker identity $\delta_{cc'}$ is unchanged by simultaneous index permutation. Every structural mixer therefore commutes with this action, as do pointwise channel maps, products, normalizations, and transpose projections. The collapse is equivariant because d, ρ, q permute and τ, g are broadcast constants. The collection of initial lifted diagonal entries permutes, so the identical singleton evaluations and their subtraction also permute. The gate and update act independently on each conformer. Permuting the binary validity mask together with padded content preserves these reductions and the actual-K denominators. Induction over stages establishes pre-pooling equivariance.

The scores and softmax weights in Equation (12) permute together, leaving their weighted sum unchanged. The subsequent topology sweep and readout receive identical pooled tensors. The spatial bypass averages the same atom states over conformers and is also invariant, so the complete scalar prediction is invariant. □

The full backbone requires a *common* conformer permutation across all paths to preserve which paths belong to the same conformation during Geo-IH. A standalone PaST operator or the final

pathwise pool can also be independently permuted at each path, but that stronger statement does not apply to the alternating backbone.

Proposition 3 (independent proper rigid motions). *For fixed atom correspondence and path representatives, replacing each conformer's coordinates by $x'_{c,i} = R_c x_{c,i} + t_c$, where $R_c \in SO(3)$ and $t_c \in \mathbb{R}^3$, leaves the prediction unchanged.*

Proof. Translations cancel in coordinate differences and in centroid-relative coordinates. Orthogonal matrices preserve norms and inner products. For $\det R_c = 1$, $(R_c a) \times (R_c b) = R_c(a \times b)$, so cross-product norms and the signed scalar triple product in Equation (20) are also unchanged. Consequently lengths, angles, torsions, validity tests, areas, and absolute volumes are unchanged, including the geometry encodings. Each conformer may have its own proper rigid motion because all these quantities are computed internally before any comparison across conformers. Their differences and relative encodings are therefore unchanged as well. The graph, chemical descriptors, and all HRGE inputs coincide. The backbone's cutoff graph, Gaussian distance features, and continuous-filter atom states are also unchanged, as is their mean bypass. The remaining network inputs therefore coincide, proving the statement. □

Proposition 4 (singleton identity). *For $K = 1$, every PaST block is the identity on its input, for every setting of the learned parameters.*

Proof. Antisymmetry implies $A_{11}^{(0)} = 0$. The full lifted tensor is therefore identical to Equation (60). Both evaluations use the same deterministic tensor layers, normalization, scales, and collapse with $K = 1$, hence $m_1^{\text{full}} = m_1^{\text{self}}$. Their difference is zero. The up-projection is bias-free, so $W_{\text{up}} 0 = 0$. The gate and residual scale multiply zero, leaving $h'_1 = h_1$. □

The entire network remains active on a singleton through its spatial backbone, geometry propagation, refinement, and prediction head. Singleton identity is not a guarantee of invariance to duplicating conformers: diagonal-restricted tensor maps and mean normalizers depend on collection cardinality, and duplicating a conformer changes its relative weight. Nor does the subtraction define an additive decomposition into independent pairwise effects.

Proposition 5 (arbitrary consistent atom relabeling). *Let $\sigma : V \to V'$ be any bijection. Transport the attributed graph, every conformer's coordinate rows, atom/bond metadata, and atom correspondence by σ; then rebuild the simple paths, incidences, and HRGE features in the new numerical order. With Equation (1), the final prediction is unchanged during training and evaluation.*

Proof. The induced map on unoriented paths is the bijection $\overline{\sigma}_r(p) = \text{canon}(\sigma(p))$. For a stored representative p, its new representative is either $\sigma(p)$ or $\text{rev}(\sigma(p))$. Element and bond attributes travel with their vertices and edges. Coordinate differences on corresponding oriented paths therefore reproduce the same raw HRGE features or their exact reversal $J_r F$, including masks. The SchNet cutoff graph is carried to the corresponding graph by the same bijection; atomic-number embeddings, distance filters, and neighbor sums preserve this correspondence. The spatial summaries in Appendix A.4 are unchanged by reversing their path representative. Equation (31) then gives $H'^{(0)}_{r,c,\overline{\sigma}_r(p)} = H^{(0)}_{r,c,p}$.

The canonical prefix/suffix pair of a coface maps to the corresponding pair, possibly exchanged. Both directed relations are present, so their multiset is carried bijectively by $\overline{\sigma}$; orientation does not choose different head parameters. The bond, angle, and signed torsion geometry used by Geo-IH is unchanged by full path reversal, as established in Appendix A.4. Its normalized states, source context means, destination-segmented softmax, and sums therefore commute with the induced path permutations. Averaging both directed relation states does the same for coface updates. Sequential residual substeps preserve this correspondence.

At each path, the conformer vector and relative geometry are now the same as at its image path. Shared PaST operators, including their singleton computations, consequently produce the same conformer states there. Induction over stages proves path equivariance through the backbone. Pathwise conformer pooling and the topology-only sweep preserve the correspondence, and each orderwise

Table 6: Configuration of the backbone-equipped PaSTNet. Widths distinguish the spatial atom encoder, path states, and relational pair states.

Component	Configuration
Path orders / actual conformers	P_0, P_1, P_2, P_3 / cap 5 by default; 10 in the scaling study
Spatial encoder	SchNet, 3 interaction blocks, $d_s = 128$, trained from scratch
Spatial graph / distance basis	All nonself neighbors within 10 angstroms; 50 Gaussians and cosine cutoff
HRGE widths / projected width	$92, 37, 13, 12$ / $d_p = 32$
Initial adapter	Inputs $220, 293, 397, 524$; affine–SiLU–LayerNorm with reversal averaging
Deep stages / weight sharing	3 / separate Geo-IH and PaST weights in each stage
Geo-IH heads / head width	4 / 8; concatenation and a $32 \to 32$ output projection
Shared geometry encoders	$(16, 6, 7) \to 32 \to 12$, one per relation order
Even pair lift / odd pair lift	$\bullet \to 32 \to 16$ / $\bullet \to 16 \to 16$ with an even gate
Pair tensors / tensor layers	Two $K \times K \times 16$ tensors / 2 distinct layers
Symmetric / antisymmetric mixers	9 / 2 channelwise structural transforms
Tensor feed-forward hidden widths	32 for both even and odd branches
Collapse	Four symmetric terms plus the antisymmetric row term; width 16
PaST gate / message projection	$48 \to 32$ with sigmoid / bias-free $16 \to 32$
RMSNorm / LayerScale initialization	$\epsilon = 10^{-6}$ / every channel initialized at 0.1
Geometry-strength initialization	One learned scalar per relation order, initialized at 0.1
Stochastic depth	Disabled in all three stages
Final refinement	One topology sweep; four 8D heads; LeakyReLU and residual LayerNorm
Readout / scalar head	Four 32D path means plus a 128D spatial mean / $256 \to 128 \to 64 \to 1$ with SiLU

mean removes the remaining path permutation. The atomwise and conformerwise means in the spatial bypass are also unchanged. The concatenated embedding and final prediction are identical. □

Corollary (three simultaneous invariances). The predictor is invariant under compositions of any consistent atom bijection, any common conformer permutation, and an arbitrary independent proper rigid motion for each conformer. Propositions 2, 3, and 5 each apply to every valid input, so successive applications preserve the prediction. The atom statement concerns attributed-graph relabeling, including transported bond-direction metadata. It does not claim invariance to changing that metadata, substituting a different stereochemical representation, or regenerating a different conformer ensemble.

D.1 Limits of the symmetry guarantees

Reflection. For an orthogonal transform Q with $\det Q = -1$, cross products transform as $(Qa) \times (Qb) = \det(Q)Q(a \times b)$. The torsion cosine is unchanged, while its signed sine changes sign; equivalently, $\phi \mapsto -\phi$ modulo 2π. Thus the absolute sine harmonics and, under a common reflection, the relative torsion sine harmonics can change. The even/odd PaST construction constrains exchange of *conformer indices*; it does not constrain these spatial-reflection transformations. Consequently the current predictor has no general reflection-invariance guarantee. This also does not establish that it can distinguish all enantiomers or that reflection sensitivity improves a given target.

E Implementation dimensions and computational accounting

E.1 Sharing and execution

The three stages own separate Geo-IH transformations, pair encoders, tensor layers, collapse maps, gates, and up-projections. These parameters are order-specific and shared over molecules, paths of that order, and conformer indices. Each block's full and singleton branches use the same parameters and normalization states. All content and PaST RMSNorm parameters and all deep LayerScale

parameters are stage-specific. Only the absolute-geometry encoder bank, including its normalization and geometry strengths, is shared across stages. Deep residual scales start at 0.1 (Touvron et al., 2021); stochastic depth is disabled.

Mixed-count mini-batches use a disjoint union of molecular path complexes with offsets on path and atom IDs, padded to the configured conformer capacity κ (five by default and up to ten in the scaling study). Path tensors have shape $[\kappa, P_r^{\text{batch}}, 32]$ and pair tensors $[P_r^{\text{batch}}, \kappa, \kappa, 16]$. A binary mask $M_{p,c}$ marks actual conformers; $M_{p,c}M_{p,c'}$ masks pair states before contractions and after tensor updates. Every mean uses $K_p = \sum_c M_{p,c}$ (or K_p^2), and conformer pooling excludes padded logits from the softmax. Singleton computations are vectorized as $[P_r^{\text{batch}}\kappa, 1, 1, 16]$ with their corresponding validity masks. These rules reproduce the stated unpadded operator for each molecule, including the exact singleton subtraction. No conformer is duplicated. The research cap study includes capacity ten; the default E0 code in the linked anonymous repository enforces capacity at most five. The separate scaling-study runner is not included in that release (Appendix G).

The spatial backbone contains only actual conformer graphs, with $\sum_m K_m|V_m|$ atom rows. All conformer comparisons stay within the same original path, all propagation relations stay within the same molecule, and both readouts separate molecules before averaging. This implementation avoids splitting a mini-batch into separate groups by K. Padding still incurs work and storage at capacity κ, rather than actual K_m; no measured throughput advantage over grouping is claimed. Static geometry and incidence tensors are cached. Mixed precision uses BF16 channel transformations with sensitive reductions accumulated in float32.

The training runner supports molecular micro-batches: sample-weighted losses accumulate over one logical batch, followed by one gradient clip and one optimizer step. This reduces the number of simultaneously retained molecular activation graphs without discarding conformers or changing the logical batch objective. Gradient checkpointing is not used in the reported implementation; Table 4 reports memory and timing measurements through conformer cap ten.

E.2 Complexity and edge cases

Let $P = \sum_r |P_r|$ and $Q = \sum_{r=1}^{3} |P_r|$, so there are $2Q$ directed incidences. For an unpadded molecule, a Geo-IH sweep with four fixed heads costs $O(K(P+Q)d_p^2)$ up to the fixed geometry width. With $T = 2$ tensor layers, pair lifting, mixing, and collapse cost $O(PK^2(d_p d_r + Td_r^2))$ per stage; reductions cost $O(PK^2 d_r)$ and singleton evaluation adds $O(PK(d_p d_r + Td_r^2))$. There are $L = 3$ stages. The reversal-averaged adapters add $O(K\sum_r |P_r| D_r^{\text{in}} d_p)$ initialization work. Pair-state storage per stage is $O(PK^2 d_r)$; ordinary backpropagation retains layer activations across the depth, giving an $O(LTPK^2 d_r)$ pair-activation term for fixed feed-forward expansion. The quadratic channel factor belongs to dense-map arithmetic, not to the size of each stored pair tensor. In the padded implementation, replace K by κ in path and pair allocation costs; masking corrects the operator, not this allocation.

For $N = |V|$, let E_c^{sp} count directed cutoff edges in conformer c. The three-block SchNet backbone additionally requires $O(3\sum_c (N + E_c^{\text{sp}}) d_s^2)$ work, up to the fixed 50-channel radial basis. The cached graph builder evaluates all atom-pair distances, with $O(KN^2)$ construction work and temporary distance storage, and has no neighbor cap. These costs are separate from the covalent path count. Total training memory also includes backbone activations, geometry caches, optimizer state, and temporary kernel workspaces.

The number of paths depends on graph degree: enumerating simple paths of length at most three is $O(|V|\Delta^3)$ under a maximum-degree bound Δ, so "linear in paths" is not an unconditional linear-in-atoms claim for arbitrary graphs. Explicit hydrogens increase N, degrees, and the realized numbers of bond, angle, and torsion paths; hydrogen-terminated torsions can therefore be a substantial multiplier even at fixed K. Retaining them exposes their local geometry to the model, but no hydrogen-removal ablation establishes their necessity or a particular steric benefit. Covalent Geo-IH incidences are complemented by nonbonded contacts in the spatial backbone. Neither representation guarantees distinguishability of all molecular geometries. Empty path orders preserve empty tensors through propagation, contribute the zero vector to the path readout, and leave the 256-dimensional combined embedding well-defined.

The evaluated caps reach ten conformers. Narrow relational width, masked batching, and optional molecular micro-batching limit the workload without removing PaST's quadratic dependence on ensemble size. Molecular micro-batching cannot resolve a single molecule's pair-tensor memory requirement. Further scaling may benefit from checkpointed layers or path-wise chunking with unchanged same-path contractions; these strategies require separate validation. Conformer subsampling or a sparse/low-rank pair approximation would change the modeled ensemble or operator and must be evaluated as such.

F Completeness of the PaST linear tensor mixers

This section characterizes the homogeneous linear parts of the two PaST branches. The argument specializes the permutation-orbit construction of Maron et al. (2019) to symmetric and antisymmetric matrices. The result identifies the implemented structural maps as complete families on their respective spaces; it does not introduce a new general theory of equivariant tensors. The proof first enumerates symmetric coefficient orbits, then handles the antisymmetric sign constraints and contractions. The final examples illustrate operator behavior; they are not empirical evidence of predictive advantage.

F.1 Symmetric branch: nine maps in the stable regime

Let $\mathrm{Sym}_K = \{S \in \mathbb{R}^{K\times K} : S^\top = S\}$, with the action $S \mapsto \Pi S \Pi^\top$ of conformer permutations $\Pi \in \mathfrak{S}_K$. For this section, $B_1, \dots, B_9$ are exactly the maps in Equation (55), including their $1/K$ and $1/K^2$ normalizers.

Theorem (complete symmetric linear mixer). *For every fixed $K \geq 1$, every linear permutation-equivariant map $L : \mathrm{Sym}_K \otimes \mathbb{R}^{d_{\mathrm{in}}} \to \mathrm{Sym}_K \otimes \mathbb{R}^{d_{\mathrm{out}}}$ has the form*

$$L(S)_{ij} = \sum_{b=1}^{9} B_b(S)_{ij} W_b, \qquad W_b \in \mathbb{R}^{d_{\mathrm{in}}\times d_{\mathrm{out}}}. \tag{62}$$

The nine scalar structural maps are linearly independent for $K \geq 4$. For $K = 1, 2, 3$, the scalar equivariant-operator dimensions are $1, 5, 8$, respectively, and the nine maps remain spanning but redundant.

Proof. First take one input and one output channel. Index independent symmetric-matrix coordinates by unordered pairs with repetition, $\{i, j\}$ with $i \leq j$. The coefficient of an input coordinate in an output coordinate must be constant on simultaneous-permutation orbits of these two unordered pairs. Conversely, arbitrary coefficients on these orbits specify an equivariant linear map. The corresponding orbit indicators have disjoint support and are linearly independent whenever their orbits exist.

Write $D_i = S_{ii}$, $R_i = \sum_j S_{ij} = K\rho_i$, $T = \sum_i S_{ii} = K\tau$, and $U = \sum_{ij} S_{ij} = K^2 g$. The nine possible orbit sums are listed below; $\delta = \delta_{ij}$ and every off-diagonal input coordinate is counted once.

Map	Output / input relation	Value at (i, j)	Min. K
O_1	diagonal / same diagonal	δD_i	1
O_2	diagonal / other diagonal	$\delta(T - D_i)$	2
O_3	diagonal / incident edge	$\delta(R_i - D_i)$	2
O_4	diagonal / nonincident edge	$\delta[(U - T)/2 - R_i + D_i]$	3
O_5	edge / endpoint diagonal	$(1 - \delta)(D_i + D_j)$	2
O_6	edge / other diagonal	$(1 - \delta)(T - D_i - D_j)$	3
O_7	edge / same edge	$(1 - \delta)S_{ij}$	2
O_8	edge / one shared endpoint	$(1 - \delta)(R_i + R_j - D_i - D_j - 2S_{ij})$	3
O_9	edge / disjoint edge	$(1 - \delta)[(U - T)/2 - R_i - R_j + D_i + D_j + S_{ij}]$	4

Here "edge" denotes a pair of distinct conformer indices, not a molecular bond. There are two diagonal–diagonal orbits, two diagonal–edge orbits, two edge–diagonal orbits, and three edge–edge

orbits. This exhausts the possible equality patterns of input and output indices. Each orbit map is in the span of the implemented maps:

$$\begin{aligned}
&O_1 = B_2, \qquad O_2 = KB_6 - B_2,\\
&O_3 = KB_4 - B_2,\\
&O_4 = \tfrac{1}{2}(K^2B_8 - KB_6) - KB_4 + B_2,\\
&O_5 = B_3 - 2B_2, \qquad O_6 = K(B_7 - B_6) - O_5,\\
&O_7 = B_1 - B_2, \qquad O_8 = K(B_5 - 2B_4) - O_5 - 2O_7,\\
&O_9 = \tfrac{1}{2}[K^2(B_9 - B_8) - K(B_7 - B_6)] - O_8 - O_7.
\end{aligned} \tag{63}$$

All identities hold for every $K \geq 1$; an absent orbit gives the zero map. Thus $B_1, \ldots, B_9$ span every equivariant linear operator. At $K \geq 4$ all nine independent orbit maps exist, so the nine spanning B_b form a basis. The minimum-K column gives dimensions $1, 5, 8$ at $K = 1, 2, 3$. Finally, permutation acts trivially on feature channels. Applying the scalar argument to each input–output channel pair gives independent matrices W_b and dimension $d_{\text{in}}d_{\text{out}}$ times the scalar dimension. □

F.2 Antisymmetric branch: two complete maps

Let $\text{Skew}_K = \{A \in \mathbb{R}^{K\times K} : A^\top = -A\}$ and $q_i = K^{-1}\sum_j A_{ij}$.

Proposition (complete antisymmetric linear mixer). *Every linear permutation-equivariant map between channel-valued antisymmetric matrices has the form*

$$L(A)_{ij} = A_{ij}V_1 + (q_i - q_j)V_2. \tag{64}$$

For scalar channels, the equivariant-operator dimension is 0 *for* $K = 1$*,* 1 *for* $K = 2$*, and* 2 *for every* $K \geq 3$*.*

Proof. For $K \geq 3$, consider output $(1, 2)$. Permutations fixing indices 1 and 2 force a common coefficient on all A_{1k} for $k \geq 3$, and another common coefficient on all A_{2k}. Any input pair entirely outside $\{1, 2\}$ has zero coefficient: swapping its two indices changes that input's sign while fixing the output. Swapping 1 and 2 negates the output, so the two incident coefficients must be opposites. Consequently the scalar output is

$$L(A)_{12} = aA_{12} + b\sum_{k\geq 3}(A_{1k} - A_{2k}) = (a - 2b)A_{12} + Kb(q_1 - q_2). \tag{65}$$

Relabeling gives the same expression at every off-diagonal output; the diagonal is zero by antisymmetry. Both displayed maps are equivariant. They are independent for $K \geq 3$: an input with only $A_{23} = 1$ and $A_{32} = -1$ has $A_{12} = 0$ but $q_1 - q_2 = -1/K$. For $K = 2$, $q_1 - q_2 = A_{12}$, so there is one independent map. For $K = 1$ the input and output spaces are zero. The channelwise extension follows as above. □

Scope of linear completeness. These statements concern all homogeneous linear maps within each parity branch at a *fixed* K. Mean normalization rescales contractions by nonzero constants and therefore preserves the span at that K. One set of weights shared across different K does not parameterize arbitrary independently chosen operators at all cardinalities simultaneously. The theorems do not characterize every coupled map between symmetric and antisymmetric spaces, the nonlinear parity couplings, or the expressive power of the whole network. In particular, linear completeness implies neither universal function approximation nor an expressivity ordering relative to attention. It establishes that the branchwise linear mixer omits no permutation-equivariant linear operation of its specified input/output type. For example, $S \mapsto (\rho_i - \rho_j)_{ij}$ is a linear equivariant map from symmetric to antisymmetric tensors, while $A \mapsto (q_i + q_j)_{ij}$ maps antisymmetric to symmetric tensors. These cross-branch maps are excluded from the stated operator spaces. PaST instead couples the branches through the parity-preserving nonlinear products in Equation (9).

F.3 Relational context and singleton counterfactuals

The structural maps make pair-to-pair context explicit. For example, the linear operation

$$S'_{13} = S_{13} + \lambda(\rho_1 + \rho_3) \tag{66}$$

can change the relation $(1, 3)$ when S_{12} changes, even with S_{13} held fixed. Successive tensor layers retain these updated pair states before collapsing to conformer states. This is an architectural distinction from a conventional attention block whose retained state is indexed by individual tokens; attention also computes pairwise scores and contextual token updates. Equation (66) explains the relational representation and does not assert that attention cannot learn the corresponding function.

The singleton subtraction evaluates the same deterministic operator on the initial pair state that one conformer would supply alone in the current block. The incoming conformer state is held fixed and can already contain ensemble information from preceding stages. This block-local comparison therefore centers the ensemble computation at a specified singleton reference; it does not rerun the entire network with other conformers removed. An elementary example shows why masking the final diagonal does not perform this centering. Take

$$S = \begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}, \qquad T_K(S)_{ij} = \rho_i + \rho_j, \tag{67}$$

and collapse by reading the diagonal of $T_K(S)$. The full response is $(4, 5)$; reevaluating the same map on singleton inputs $[1]$ and $[2]$ gives $(2, 4)$. The ensemble-minus-singleton response is $(2, 1)$. Masking the final diagonal instead gives a zero collapsed response, discarding the ensemble context already written to the diagonal. This example uses one of the complete symmetric structural maps and therefore does not depend on a special nonlinear construction.

For $K = 1$, the full and singleton inputs and operators coincide, making the subtraction exactly zero; the zero-preserving residual map then leaves the conformer state unchanged. For $K > 1$, this subtraction is a model-defined *counterfactual ensemble residual*, not a causal effect, a leave-one-out attribution, or a guarantee that every self-dependent term disappears. It does not imply invariance to duplicated conformers. For example, let $S = s\mathbf{1}\mathbf{1}^\top$ and retain only $T_K(S)_{ij} = \delta_{ij}S_{ii}$, followed by row-mean collapse. The full message is s/K, whereas its singleton counterpart is s; hence the centered response is $(1/K - 1)s$ even for identical inputs. This illustrates cardinality sensitivity of the operator family, not a measured effect in a trained checkpoint.

Finite-size algebra checks. We checked the displayed mean-normalized maps over exact rational numbers for $K = 1, \ldots, 6$. To construct each scalar operator matrix, apply $B_1, \ldots, B_9$ to the symmetric coordinate basis E_{ii} and $E_{ij}+E_{ji}$ $(i < j)$, or the two odd maps to the antisymmetric basis $E_{ij} - E_{ji}$. Vectorizing these operator matrices and applying rational Gaussian elimination gives ranks $(1, 5, 8, 9, 9, 9)$ for the symmetric branch and $(0, 1, 2, 2, 2, 2)$ for the antisymmetric branch, in increasing order of K. Independent coefficient-orbit counts, including the antisymmetric sign constraints, give the same dimensions.

For each adjacent transposition Π and each coordinate-basis input X, the checks verify $L(\Pi X \Pi^\top) = \Pi L(X)\Pi^\top$ exactly. They also verify every orbit-change-of-basis identity in Equation (63) on those inputs and the numerical singleton-versus-diagonal-masking example above. These finite checks implement the formulas independently of the learned PyTorch core; they are algebra checks, not a production-code audit or predictive experiment. The proofs above establish the statements for all K.

G Experimental protocol and interpretation limits

Reporting scope and reproducibility. All predictive results and resource measurements are obtained from our own local runs, including the SchNet and Uni-Mol v1 baseline scores. All predictive comparisons in Tables 1–3 use the same random-scaffold protocol with nominal 81/9/10 training/validation/test proportions and 3 seeds; initialization uses seed 42. Predictive scores are unweighted means and sample standard deviations across the three splits ($n - 1$ denominator). Shared partitions and reporting conventions do not imply identical pretraining, model capacity, optimization schedules, or hyperparameter-search budgets. The reported dispersion measures split variation with

a fixed initialization; it does not estimate initialization variability. Paired per-split scores are not reported, so no paired confidence intervals or significance tests are inferred from these means and standard deviations.

Reporting and code availability. The tables in this PDF report aggregate measurements and the following sections specify their protocols. Raw predictions, timing traces, and a complete run-level checkpoint/configuration manifest are not provided. The linked anonymous E0 code supports conformer counts up to five; the separate $K = 10$ cap-study runners, model-specific baseline modifications, and interaction-control runners are outside that release. The code therefore does not independently reproduce every table entry.

Partitions and topological generalization. All models read the same dataset-specific split CSVs rather than generate a new partition for each architecture. The splitter groups chirality-aware Bemis–Murcko scaffolds, shuffles the groups using the split seed, and allocates whole groups to the test and validation partitions before training. It requests 10% of all molecules for testing and 10% of the remaining 90% for validation, giving nominal 81/9/10 proportions; whole-scaffold allocation and the fixed conformer-failure exclusions can slightly change the realized proportions. The ESOL files contain 1,115 molecules, with 904/100/111 training/validation/test molecules per split. Their hashes still match the saved canonical-molecule and chirality-aware Bemis–Murcko scaffold audit (Bemis & Murcko, 1996), which finds no cross-partition overlap. All 314 empty-scaffold molecules are in training in every ESOL split. Therefore, these test sets measure transfer to held-out ring-containing scaffold families and do not evaluate held-out acyclic molecules. Redistributing acyclic molecules would define an additional split protocol: assigning the same empty scaffold to several partitions cannot simultaneously preserve strict scaffold disjointness. Such an audit should report ring-scaffold disjointness separately from molecule-disjoint acyclic evaluation. No such additional result is claimed here.

The prepared Lipo and ClinTox partitions exclude five and thirteen molecules, respectively, with failed conformer preparation. These fixed exclusions are applied after scaffold splitting without reshuffling the retained molecules. The preparation manifests record the excluded molecules and split hashes; selection does not use targets. Table 1 reports results on all six datasets: ESOL, FreeSolv, Lipo, BACE, BBBP, and ClinTox. ClinTox uses the single `CT_TOX` endpoint. All ClinTox table entries are reported for this endpoint; the `FDA_APPROVED` endpoint and two-task averages are outside the comparison. The released single-target reader rejects missing or nonfinite target values rather than imputing them or silently filtering rows. The thirteen recorded ClinTox exclusions are conformer-preparation failures. A separate endpoint/label-handling audit of every baseline wrapper is not reported; this endpoint convention does not establish independent verification of those implementations.

Conformer preparation and physical scope. The PaSTNet caches for all six datasets share the same preparation settings. Baseline coordinate generation and conformer pooling are model-specific. ETKDGv3 (Riniker & Landrum, 2015; Wang et al., 2020) embeds up to 20 candidates using a deterministic molecule-specific seed derived from global seed 42 and the canonical molecule identifier, with a deterministic random-coordinate retry if necessary. Force-field optimization uses MMFF94s (Halgren, 1999), falling back to UFF (Rappé et al., 1992) where supported, for at most 1,000 iterations. Only converged candidates with finite energies and coordinates are usable. Selection starts with the lowest-energy usable conformer, then maximizes minimum aligned heavy-atom RMSD to the selected set, stopping below 0.5 Å or at the configured cap (five by default). Alignment uses fixed atom correspondence without reflections or atom remapping. Actual conformer counts are used without duplicating small ensembles. Explicit hydrogens and a shared atom table are retained. Training reads the saved coordinates rather than regenerate conformers.

Conformer-cap study. Table 3 trains a separate PaSTNet model for each cap $K \in \{1, 3, 5, 10\}$ and each split. The $K = 10$ configuration retains the same ETKDGv3, MMFF94s/UFF, and aligned heavy-atom RMSD selection pipeline, increasing only the maximum retained conformer count. Ensembles smaller than the cap are not duplicated. All configurations use the same 3 seeds and model initialization seed 42; scores are mean $\pm$ sample standard deviation across these three splits. The bond scale in Equation (32) is fitted separately for each cap on each corresponding training partition, then frozen for validation and testing; the $K = 5$ scale is not reused for other caps. At $K = 1$,

this statistic is zero because all pairs are diagonal, but all bond differences are also zero and PaST's singleton residual vanishes exactly. The positive-scale diagnostic in Table 5 concerns the default $K = 5$ caches, not the singleton-only configuration.

This diversity filter is a geometric sampling rule, not a draw from a Boltzmann distribution. Force-field energies select candidates but do not enter PaSTNet as predictive features or thermodynamic pooling weights. Consequently, even the FreeSolv results do not validate thermodynamic populations or physical correctness of learned conformer weights. No energy-annotated GEOM evaluation (Axelrod & Gómez-Bombarelli, 2022) is reported. A separate physical-ensemble study would need reference ensembles or energy-aware sampling under an explicit temperature and energy convention, with no unaccounted diversity filtering of their populations.

Model and attention control. PaSTNet uses hidden width 32, pair width 16, three Geo-IH/PaST stages, and two tensor layers per PaST block. Its SchNet backbone is trained from scratch jointly with the path network. Stochastic depth is disabled; residual and geometry-strength scales initialize to 0.1. The same-path attention control retains the same spatial encoder, path adapters, topology propagation, pooling, spatial bypass, and prediction head. It replaces each PaST block with two standard self-attention and FFN layers, with four heads at width 32, pre-RMSNorm, SiLU, and no dropout. FFN widths by path order are 96, 98, 96, and 97, giving 667,710 parameters compared with PaSTNet's 667,548. Relative even and odd geometry becomes a learned per-head attention-logit bias. Self-attention includes the query conformer; there is no singleton-counterfactual subtraction. The unchanged modules retain their reference initialization when the attention modules are constructed. This is a controlled comparison of two complete interaction blocks, not a separate ablation of every PaST design choice.

Multi-conformer identity control. The No PaST (identity) control in Table 2 keeps the default conformer cap $K = 5$, the same input conformers, SchNet backbone, path encoder, Geo-IH stages, pathwise conformer pooling, final topology sweep, spatial bypass, and prediction head. Every PaST block is replaced by the identity, so Equation (2) becomes $H^{(s)} = \widetilde{H}^{(s)} = \mathsf{GeoIH}_s(H^{(s-1)})$. The model retains all available conformer views up to the cap, with no PaST pair lifting, tensor processing, or singleton subtraction. Thus it tests the addition of the complete PaST block at fixed conformer inputs and surrounding architecture. Pooling still aggregates conformers; this is not the separately trained $K = 1$ model. The identity itself has no trainable parameters, and this comparison does not match the interaction capacity of full PaST. It follows the shared 81/9/10 random-scaffold protocol with 3 seeds; the aggregate record does not specify its total trainable parameter count.

Pair-local context control. The "without pair-to-pair context" variant replaces each $2 \to 2$ tensor layer in every stage/order weight group by a pair-local layer. Only the first two symmetric maps and the direct odd map survive:

$$Z^{+}_{ij} = S_{ij}W_1 + \delta_{ij}S_{ij}W_2 + (A_{ij} \odot A_{ij})W_{AA}, \tag{68}$$

$$Z^{-}_{ij} = A_{ij}V_1 + (S_{ij} \odot A_{ij})W_{SA}. \tag{69}$$

Thus $B_3, \ldots, B_9$ and the odd row-difference map V_2 are removed. Both tensor layers, RMSNorm, LayerScale, the two feed-forward updates, parity couplings, and numerical parity projections are retained. The full $2 \to 1$ collapse in Equation (10) is unchanged: it still aggregates diagonal, row, trace, global, and odd-row information. "No context" therefore refers specifically to contextualization of pair states inside the $2 \to 2$ layers, not to the absence of ensemble aggregation. The singleton branch uses the initial diagonal pair, zero odd input, and the same modified core and collapse as the full branch; subtraction and the exact $K = 1$ identity remain. The configuration is `pair_to_pair_context=False`, with two symmetric and one odd structural maps per layer. Removing eight 16×16 channel maps across three stages, four orders, and two layers removes 49,152 parameters, leaving 618,396 trainable parameters. This is not a capacity-matched ablation.

Singleton-subtraction ablation. The "without singleton subtraction" control omits the singleton reference from the residual in Equation (11): it uses $\mathbf{h}'_c = \mathbf{h}_c + u_{s,r}(\mathbf{h}_c, \mathbf{m}^{\text{full}}_c)$ instead of subtracting $\mathbf{m}^{\text{self}}_c$. The full-ensemble pair lifting, tensor layers, collapse, gated residual, and all surrounding modules are retained. Because the singleton branch shares its weights, removing its subtraction leaves the trainable parameter count unchanged at 667,548. The control follows the same $K = 5$

inputs, three scaffold splits, initialization seed 42, optimization, and checkpoint-selection protocol as full PaST. This intervention removes the algebraic guarantee that the block is the identity at $K = 1$; the guarantee still applies to full PaST. The control obtains 0.82 ± 0.14 ESOL RMSE, 1.76 ± 0.40 FreeSolv RMSE, and $84.03 \pm 3.72\%$ BACE ROC-AUC. These aggregate scores favor centering in this setting, but do not establish its general necessity or statistical significance.

Optimization and checkpoint selection. Table 7 summarizes the recorded local recipes. PaSTNet regression uses MSE on training-only z-scores with population standard deviation; validation RMSE and final test RMSE use original target units. BACE, BBBP, and ClinTox are binary-classification tasks. PaSTNet uses unweighted binary cross-entropy on raw logits, with no label standardization, and selects checkpoints by validation ROC-AUC. Classification results are reported as ROC-AUC percentages. The local trainers reload the selected checkpoint before test evaluation. Auxiliary objectives in SPiCE, FACET, and 3DGCL retain their model-specific roles. FACET stage 2 selects the largest validation Pearson correlation between teacher and learned pair distances; 3DGCL pretraining selects the smallest training contrastive loss. These auxiliary selection rules precede supervised property-checkpoint selection.

Table 7: Local training recipes. Batch sizes count molecules; epoch and patience values are maximum training epochs and early-stopping patience. Rows with multiple stages list them in execution order. SchNet and Uni-Mol v1 entries describe their source-defined training recipes; the author-specified ensemble and initialization settings are detailed below. DKO's learning rate is selected from the search described below.

Model / phase	Optimizer	Initial LR	Schedule	Batch	Epochs	Patience
PaSTNet, same-path attention	Adam	10^{-3}	Fixed	16	300	40
PCNN	Adam	10^{-3}	Fixed	16	300	40
SchNet	Adam	10^{-3}	Plateau	32	300	100
Uni-Mol v1	Adam	10^{-4}	Warmup + linear	32	100	10
SPiCE	AdamW	$2 \cdot 10^{-4}$	Warmup + cosine	16	300	40
ConAN	Adam	10^{-3}	Plateau	96	150	50
ConAN-FGW	Adam	$5 \cdot 10^{-4}$	Plateau	24	80	50
FACET, stage 1	Adam	$2 \cdot 10^{-3}$	Plateau	96	150	50
FACET, stage 2	AdamW	10^{-4}	Fixed	1	700	–
FACET, stage 3	Adam	$5 \cdot 10^{-4}$	Plateau	24	400	100
3DGCL, supervised	Adam	$2 \cdot 10^{-3}$	Exponential	32	100	100
PaiNN	Adam	10^{-3}	Plateau	32	150	50
DKO	AdamW	Search	Cosine	32	300	30

For SchNet, plateau scheduling multiplies the learning rate by 0.5 after 25 epochs without validation improvement. Its weight decay is 10^{-5} and gradient clipping is 1. Uni-Mol v1 uses Adam with $\epsilon = 10^{-6}$, zero weight decay, 3% linear warmup followed by linear decay, and gradient clipping at 5. For the other plateau-scheduled baselines, the rate is multiplied by 0.8 after 10 epochs without monitored improvement. SPiCE uses five warmup epochs, a minimum learning-rate factor of 0.01, and weight decay 10^{-3}. FACET stage 2 uses weight decay 0.01. Supervised 3DGCL training uses weight decay 10^{-3} and exponential factor 0.95. DKO uses the weight decay selected by its validation search; the remaining recipes not specified above use zero weight decay. Gradient clipping is 5 for PaSTNet, its attention control, PCNN, and SPiCE; 1 for ConAN, ConAN-FGW, supervised FACET, PaiNN, and DKO; and disabled for 3DGCL. PaSTNet uses BF16 autocast in training on all six datasets, with BF16 evaluation on ESOL and FreeSolv and FP32 evaluation on Lipo, BACE, BBBP, and ClinTox. For Lipo and the three classification datasets, it uses sample-weighted gradient accumulation over microbatches of 4 molecules to preserve each logical batch. Its attention control uses BF16 training and FP32 evaluation on BACE. The recorded BACE recipes use microbatches of 4 molecules for PaSTNet, its attention control, ConAN-FGW, and FACET; 16 for PCNN, SPiCE, and 3DGCL; and 8 for ConAN and PaiNN. DKO uses physical batches of 32 in FP32.

SchNet and Uni-Mol v1 implementation details. The reported runs use initialization seed 42, a cap of five conformers, and mean pooling across conformers. The SchNet implementation uses

width 128, six interaction blocks, 128 filters, 50 radial basis functions and a 10 Å cutoff. Its atom-wise output head is followed by an atom sum and then a mean of conformer predictions; it is trained from scratch. Regression output is affinely calibrated with training-target mean and sample standard deviation, and MSE is applied after returning predictions to original units. Classification uses raw logits and binary cross-entropy. These details differ from PaSTNet's three-block backbone and standardized regression loss.

Uni-Mol v1 starts from `mol_pre_all_h_220816.pt`, retains hydrogens, and fine-tunes the complete encoder and prediction head. Its reported five-conformer adaptation mean-pools embeddings before the head; the inspected base entry point accepts a SMILES-derived single-conformer input, so this adaptation is a separate run-specific modification. The source-defined recipe is in Table 7. Resolved run configurations, the multi-conformer wrapper, and dataset-registry additions for BBBP/ClinTox are not included in the linked code release.

The reported Uni-Mol setting retains the upstream Uni-Mol Tools automatic target preprocessing (`target-normalize=auto`). The scaler is fitted before the split is applied; in our fixed-split wrapper, its input contains the training and validation targets. Test data remain separate. This is inherited upstream behavior and differs from PaSTNet's training-only target standardization. Evaluating a common training-only preprocessing protocol would require a separate run. The pretrained checkpoint's corpus membership and overlap with the downstream molecules have not been independently audited; absence of pretraining overlap is not established. No training-only target-preprocessing rerun is included in Table 1. Validation targets therefore influence the fitted preprocessing state in the reported Uni-Mol protocol, but test targets do not. This difference limits preprocessing comparability and is not evidence of test-label leakage.

Baseline adaptations and fairness. PCNN averages its encoder's 122-dimensional conformer embeddings before its prediction head and has 206,827 parameters. Its terminal sigmoid is replaced by an identity for standardized regression or raw classification logits. Covalent-path features still encode geometry, but omitting distance-contact edges restricts nonbonded spatial information. Therefore, the PCNN comparison alone cannot establish an advantage over a contact-enabled, capacity-controlled PCNN. The local SPiCE configuration trains PaiNN/GIN encoders and eight sparse experts from scratch, activating two experts per input without pretrained expert upcycling. ConAN uses a SchNet/GAT model trained from scratch, while ConAN-FGW starts from the corresponding split's validation-selected ConAN checkpoint and adds FGW aggregation. FACET uses its three-stage SchNet/fragment-transformer procedure, with the learned transformer frozen in the final supervised stage; the main table uses this final-stage result. No external pretraining is used in these local ensemble recipes.

3DGCL performs contrastive pretraining on the training partition only, then supervised fine-tuning, and predicts from the first cached, lowest-energy conformer. Its pretraining budget is 100 epochs, learning rate 10^{-3}, temperature 0.2, and exponential factor 0.95 per batch; the batch size is 400 for ESOL/FreeSolv and 128 for BACE. PaiNN averages conformer predictions from a width-128, three-interaction PaiNN with a 5 Å cutoff and 20 radial basis functions. These choices differ in capacity, pretraining, conformer readout, and training cost. Retaining distinct optimization schedules avoids imposing PaSTNet's fixed schedule on every model, but it does not establish equally effective tuning or equal search budgets. Only PaSTNet and its attention control are approximately parameter matched; no capacity-matching claim is made for the complete baseline set. A stronger family-level comparison would require prespecified validation-only search budgets and a contact-enabled PCNN control.

Distribution kernel operator baseline. DKO uses the `DKOGatedFusion` implementation (Cheng et al., 2026), trained from scratch on the same cached conformers. Its geometric feature extractor uses a 4 Å distance cutoff; variable-length feature vectors are right-padded with zeros or truncated to 1,024 dimensions. Feature z-scores are fitted on the training partition with equal molecule weights. Each molecule is represented by its uniformly weighted conformer mean and the ten largest diagonal variances, with variance regularization 0.01 and centered feature values clipped to $[-10, 10]$. This uses the implementation's diagonal branch for feature dimensions above 256. It differs from a full-covariance DKO; feature truncation can also discard trailing geometric or atom features. Thus, the reported comparison concerns this specific DKO representation and adaptation.

Eight prespecified hyperparameter configurations are evaluated using validation performance independently within each split. Hidden widths 128 and 256 are each combined with four tuples of (learning rate, weight decay, dropout, feature-z-score clipping threshold): $(10^{-4}, 10^{-5}, 0.1, \infty)$, $(3{\cdot}10^{-4}, 10^{-4}, 0.3, 5)$, $(10^{-3}, 10^{-4}, 0.3, 5)$, and $(10^{-4}, 10^{-3}, 0.5, 5)$, where ∞ denotes no clipping. Every fit uses AdamW with cosine decay to 10^{-6}, at most 300 epochs, and early-stopping patience 30. The best validation checkpoint across these configurations is evaluated on the test partition once. A final singleton training batch is merged into its preceding batch to support batch normalization without dropping or duplicating molecules. This search gives DKO an additional validation-tuning budget; equal search cost across models is not claimed.

Matched-hardware resource protocol. Table 4 profiles FACET and PaSTNet locally on one NVIDIA RTX 5070 Ti with 16 GB of device memory. Both models use FP32, batch size one molecule, and microbatch size one. The fixed profiling subset contains 100 molecules from ESOL training split 0, each with at least ten available conformers, so each tested cap uses exactly K conformers. This subset is used only for resource profiling; the predictive study uses the full corresponding dataset partitions and allows smaller ensembles.

Each timing run uses 20 warm-up iterations followed by 100 measured iterations; the entire procedure is repeated five times. Latencies in the table are means of the five repeat means, in milliseconds per molecule. Repeat-level standard deviations are not available in the reported record and are not estimated here. Forward-plus-backward timing includes the forward pass, loss computation, and backward pass, but excludes the optimizer step. Peak training memory is CUDA peak *allocated* memory, measured after optimizer states have been materialized. Timings exclude data loading, conformer generation, preprocessing/cache construction, and FACET's offline teacher/FGW preparation. Thus, the measurements compare the profiled model passes and do not quantify total training cost. The FP32, single-molecule profiling setup is distinct from the predictive-training batch and precision settings above.

Resource interpretation and remaining evidence. The resource results apply only to the stated hardware, precision, batch size, and molecular subset through $K = 10$. Masked batching uses actual conformer counts in reductions (Appendix E); microbatching preserves one optimizer update per logical batch but does not remove per-molecule quadratic pair storage. Explicit hydrogens increase path counts. Gradient checkpointing and path-wise chunking remain prospective; conformer subsampling changes the input ensemble. No end-to-end timing is reported.

Table 2 includes a singleton-subtraction ablation alongside the identity, attention, and pair-local controls. No capacity-balanced pair-local variant, alternative channel-valued pair-state comparator, or isolated parity-coupling study is reported. Multiple initialization seeds and paired split-level inference are also absent.

Synthetic implementation checks. We performed local CPU FP32 checks using synthetic inputs and a randomly initialized model, rather than trained checkpoints. One mixed batch contains explicit-hydrogen molecules with SMILES `CCO`, `C/C=C/C`, and `C[C@H](O)Cl`, with $K = 1, 3, 5$, respectively. Coordinates are standard-normal samples generated in FP64 with seed 2718; model initialization uses seed 42, bond scale 0.01 Å, and evaluation mode. We compare predictions after conformer permutations, independent proper rotations and translations, and atom relabeling with transported RDKit attributes, as well as packed versus individual execution. Maximum absolute/relative prediction errors are respectively $7.45 \cdot 10^{-9}/9.28 \cdot 10^{-8}$ for conformer permutation, $2.24{\cdot}10^{-8}/4.29{\cdot}10^{-7}$ for independent proper rotations and translations, and $2.98{\cdot}10^{-8}/5.72{\cdot}10^{-7}$ for atom relabeling with transported RDKit attributes. Packed-versus-individual prediction error is $2.24 \cdot 10^{-8}/4.29 \cdot 10^{-7}$. Relative errors divide by $\max(|\hat{y}|, 10^{-8})$; assertions use absolute tolerance $2 \cdot 10^{-6}$ and relative tolerance $2 \cdot 10^{-5}$. Reflection is tested separately and changes predictions by up to $2.21 \cdot 10^{-4}$, consistent with the absence of a reflection-invariance guarantee; this is not a benchmark chirality result.

Further tests cover path reversal of descriptors and learned initialization, coincident points, collinear and near-threshold paths, rejection of nonfinite coordinates, masked invalid angle/torsion channels, and finite predictions and model-parameter gradients. Coordinates are fixed inputs; coordinate derivatives at degenerate geometry are not claimed. Bond-adapter tests lock the evaluated single/aromatic slot sharing and constant nominal-length channels. Zero and small bond scales (0,

10^{-12}, 10^{-8}) produce finite parity-consistent relative descriptors on the tested inputs without changing the calibration rule. All 15 checks pass with Python 3.10, PyTorch 2.5.1, RDKit 2026.03.3, and SciPy 1.14.1. These local CPU checks use a different environment from benchmark training. The test source and raw report are not included in the linked anonymous code release. A trained-checkpoint duplicate/diversity diagnostic remains unreported. Table 2 includes removal of the complete PaST block and ablation of singleton subtraction at $K = 5$; these predictive comparisons are distinct from the synthetic checks. The recorded checks compare model outputs; they do not separately assert the identity of every PaST block at $K = 1$ across all four path orders, including padded singleton inputs. The singleton identity remains the algebraic statement in Appendix D, with its implementation specified in Appendix C. No additional block-level test result is claimed.