EASER: ENERGY-AWARE SYMBOLIC ENSEMBLE READOUT FOR CONFORMER-BASED MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flexible molecules are naturally represented by ensembles of three-dimensional conformers, yet many conformer-based predictors rely on neural readouts whose aggregation mechanisms are difficult to inspect and do not explicitly account for relative conformer energies. We introduce EASER, an Energy-Aware Symbolic Ensemble Readout for molecular conformer ensembles. EASER trains and freezes a SchNet encoder to map each conformer to a fixed-dimensional representation. A cross-fitted descriptor model captures readily predictable two-dimensional signal, while energy-conditioned genetic-programming experts construct compact symbolic features to model the remaining residual from conformer embeddings. Their outputs are linearly combined into conformer-level residual scores and aggregated using a learned effective energy scale. Across multiple MoleculeNet benchmarks under both random and scaffold splits, EASER consistently improves over a matched descriptor-only baseline while remaining competitive with established molecular prediction methods. Ablation studies support complementary contributions from descriptor residualization, symbolic feature construction, and energy-conditioned ensemble modeling, while conformer-budget experiments show that most predictive gains can be obtained with relatively small ensembles. At inference, evolutionary search is no longer required: EASER uses only the frozen encoder, descriptor baseline, fixed algebraic expressions, linear coefficients, and explicit energy-aware aggregation. EASER thus provides a compact and auditable readout over learned 3D molecular representations.

1 INTRODUCTION

Flexible molecules are better represented by ensembles of three-dimensional (3D) conformations than by a single geometry, particularly when the target property is sensitive to molecular structure. A molecule may adopt several energetically accessible conformations, and properties such as solubility, hydration free energy, and bioactivity can depend on variation across this ensemble. This has motivated geometric neural networks and, more recently, models that explicitly aggregate multiple conformers (Schütt et al., 2018; Axelrod & Gómez-Bombarelli, 2022; Nguyen et al., 2024). These advances have substantially improved molecular representation learning, but the downstream mapping from conformer representations to a molecular prediction is often implemented by learned neural components (Nguyen et al., 2024; 2025; 2026; Zhu et al., 2025). Such readouts can be highly expressive, but their conformer-dependent behavior is encoded within the learned model rather than exposed as an explicit rule. At the ensemble level, a further question is how relative conformer energy should enter this mapping. Uniform pooling assigns equal weight to all retained conformers, whereas learned aggregation can model unequal contributions through a flexible neural mechanism.

Relative conformer energy can enter an ensemble predictor in two conceptually different places. First, conformers from different energy regimes may require different mappings from geometric representation to predictive signal; we refer to this as *specialization*. Second, after a conformer-level score has been obtained, lower- and higher-energy conformers need not contribute equally to the molecule-level prediction; we refer to this as *aggregation*. Existing fusion mechanisms can in principle capture both effects, but typically do not separate them explicitly. We therefore ask whether these two roles of energy can be modeled explicitly and evaluated independently.

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.

Addressing this question calls for a readout whose computation remains accessible after training. Symbolic regression provides one such mechanism: genetic programming (GP) can discover compact nonlinear expressions that remain inexpensive to evaluate once the search is complete (Koza, 1992; Cranmer, 2023). Direct symbolic search over molecular graphs or atomic coordinates, however, is poorly matched to variable-size molecular structures. We therefore decouple representation learning from readout discovery: a geometric encoder maps each conformer into a fixed-dimensional latent space, and GP searches only over these learned representations.

We propose **EASER**, an energy-aware symbolic ensemble readout that makes these two roles of conformer energy explicit. EASER freezes a trained SchNet encoder, removes signal already captured by inexpensive 2D descriptors, and uses GP to construct compact nonlinear features in the frozen latent space. Relative conformer energy then enters the readout through two distinct paths: deterministic energy routing selects a specialized symbolic expert, while a learned effective energy scale controls molecule-level aggregation. Our contributions are summarized as follows.

- We introduce a symbolic ensemble readout over frozen 3D conformer representations, enabling genetic programming to operate in a fixed-dimensional latent space while keeping the deployed readout explicit.
- We separate two complementary roles of conformer energy: *energy-conditioned specialization*, implemented by deterministic routing to low-, mid-, and high-energy symbolic experts, and *energy-aware aggregation*, implemented through a learned effective energy scale. This separation enables matched ablations of the two mechanisms.
- We combine symbolic feature construction with cross-fitted descriptor residualization, encouraging GP to model complementary conformer-sensitive signal rather than rediscover correlations already captured by inexpensive 2D descriptors.
- We evaluate EASER against same-pipeline neural and linear controls and analyze component contributions, conformer-budget sensitivity, molecular flexibility, symbolic complexity and stability, computational cost, and seed-wise robustness.

2 Related Work

3D molecular and conformer learning. Message-passing and geometric neural networks learn molecular representations from connectivity and 3D structure (Gilmer et al., 2017; Schütt et al., 2018; Gasteiger et al., 2020; Schütt et al., 2021). More recent architectures strengthen geometric representation learning through richer inductive biases or feature integration, including MACE (Batatia et al., 2022), EquiformerV2 (Liao et al., 2024), and MolHFCNet (Nguyen et al., 2025). A separate line of work explicitly models multiple conformers of the same molecule. CONAN (Nguyen et al., 2024) aggregates information across conformer ensembles, while the MARCEL benchmark systematized evaluation in this setting (Zhu et al., 2024). More recent methods such as FACET (Nguyen et al., 2026) and SPiCE (Zhu et al., 2025) develop dedicated neural architectures for conformer-ensemble modeling. EASER addresses a complementary question: rather than introducing a stronger geometric encoder or ensemble-fusion architecture, it keeps the learned 3D representation fixed and studies whether the downstream conformer-ensemble readout can remain compact and explicit while retaining predictive utility.

Energy-aware conformer aggregation. Conformer ensembles can be generated with methods such as ETKDG and refined using molecular mechanics (Riniker & Landrum, 2015; Halgren, 1996), yielding approximate energies that provide a natural ordering among retained conformers. Uniform pooling assigns equal weight to every conformer, whereas learned ensemble mechanisms can model unequal contributions through trainable neural components (Nguyen et al., 2024; 2026; Zhu et al., 2025). A complementary physical reference is Boltzmann weighting, in which conformer contributions are proportional to $\exp(-\Delta E/k_B T)$. EASER adopts the same monotone exponential form for aggregation but learns an *effective energy scale* for prediction rather than treating it as a physical temperature. More importantly, EASER uses relative energy not only for aggregation but also for explicit conformer specialization through deterministic routing, allowing these two roles of energy to be studied separately.

Residual and symbolic learning. Classical molecular descriptors provide inexpensive predictive signals and remain useful baselines for molecular property prediction (Rogers & Hahn, 2010; Landrum et al.). EASER uses a cross-fitted descriptor predictor to capture this readily accessible signal and applies the symbolic stage only to the remaining residual, following the general motivation of Δ-learning (Ramakrishnan et al., 2015). Genetic programming and symbolic regression search over explicit algebraic expressions (Koza, 1992; Cranmer, 2023; Wang et al., 2025), but direct search over molecular graphs or atomic coordinates is poorly matched to variable-size molecular structures. EASER instead applies GP to fixed-dimensional representations produced by a frozen SchNet encoder, using symbolic regression as a readout over learned 3D features rather than as a replacement for molecular representation learning. Consequently, the interpretability claim is restricted to the readout itself: the resulting algebraic operations are explicit, but individual SchNet latent coordinates are not assumed to possess direct chemical semantics.

3 EASER: Energy-Aware Symbolic Ensemble Readout

Figure 1 summarizes EASER. Given a conformer ensemble, a frozen 3D encoder first maps each conformer to a fixed-dimensional representation, while a cross-fitted 2D descriptor model defines the residual signal to be modeled by the readout. Relative conformer energy then enters through two explicit mechanisms: it routes each conformer to an energy-conditioned symbolic expert and subsequently determines its contribution to molecule-level aggregation. The resulting ensemble correction is finally combined with the descriptor baseline to obtain the molecular prediction.

3.1 Problem Setup and Frozen 3D Representations

Let the training set be $\mathcal{D} = \{(m, y_m)\}_{m=1}^{M}$, where each molecule is associated with either a scalar regression target or a binary classification label. For molecule m, a conformer generator produces an ensemble

$$\mathcal{C}_m = \{c_{m1}, \ldots, c_{mK_m}\}, \qquad K_m \le K_{\max}, \tag{1}$$

where conformer c_{mk} contains the shared atomic numbers $\mathbf{Z}_m$, Cartesian coordinates $\mathbf{R}_{mk} \in \mathbb{R}^{N_m \times 3}$, and a force-field energy E_{mk}. Since absolute force-field energies are not directly comparable across molecules, we define the within-molecule relative energy

$$\Delta E_{mk} = E_{mk} - \min_{1 \le j \le K_m} E_{mj}. \tag{2}$$

The lowest-energy retained conformer therefore has $\Delta E = 0$. Importantly, this reference corresponds to the lowest-energy conformer in the generated ensemble, rather than to the unknown global minimum of the molecular potential-energy surface. If conformer generation misses a lower-energy geometry, the resulting relative energies, and hence the routing decisions introduced later, may change. We therefore treat ΔE as a property of the retained ensemble rather than as a thermodynamically calibrated quantity.

Each conformer is mapped to a fixed-dimensional representation by a SchNet encoder $f_\theta(\cdot)$,

$$\mathbf{z}_{mk} = f_\theta(c_{mk}) \in \mathbb{R}^H. \tag{3}$$

During representation learning, conformer embeddings are uniformly averaged within each molecule and passed through a temporary linear prediction head. The encoder is trained using only the training partition, with model selection based on validation data, and is then frozen; the temporary head is discarded. Freezing the encoder serves two purposes: GP search does not backpropagate through the 3D representation model, and all matched readout variants operate on exactly the same latent representations. Before symbolic search, each latent coordinate is standardized using statistics computed from the training conformers only.

3.2 Cross-Fitted 2D Baseline and Residual Targets

The 2D descriptor branch captures predictive signal that can already be obtained from inexpensive molecular descriptors. We use the 217 RDKit descriptors in `Descriptors.descList`, retaining only columns that are finite on the corresponding training partition. Let $\mathbf{d}_m \in \mathbb{R}^D$ denote the resulting descriptor vector. For regression, the descriptor baseline is a ridge model

$$\widehat{y}_m^{\text{base}} = \mathbf{w}^\top \mathbf{d}_m + b. \tag{4}$$

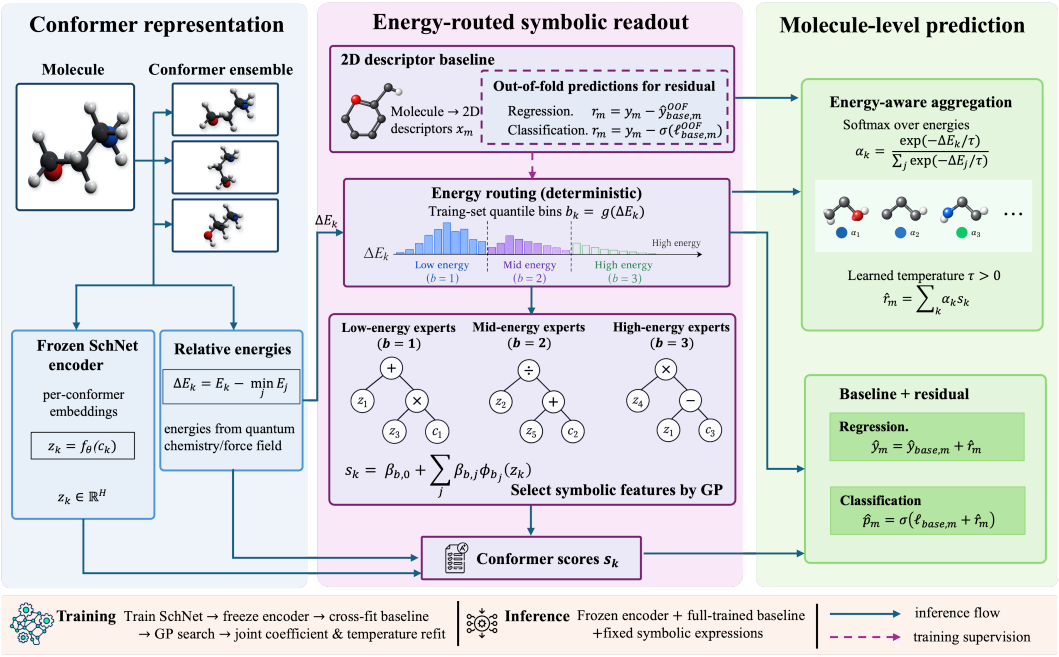


Figure 1: **Overview of EASER.** Multiple conformers are encoded by a frozen SchNet model, while a cross-fitted 2D descriptor baseline defines the residual target. Relative conformer energies first provide deterministic routing to low-, mid-, and high-energy symbolic experts through fixed quantile boundaries. After expert evaluation, a learned energy scale determines the normalized conformer weights used to aggregate conformer scores. For regression, the aggregate corrects the descriptor prediction; for classification, the same correction is applied in logit space before a sigmoid.

For classification, the corresponding baseline is logistic regression, producing a baseline logit ℓ_m^{base}.

Directly defining residual targets using a descriptor model fitted on the same training examples would produce optimistically small errors and leak in-sample fit into the symbolic stage. EASER therefore uses five-fold cross-fitting to construct out-of-fold targets. Let $\pi(m)$ denote the fold containing molecule m. For regression, the symbolic search target is

$$r_m = y_m - \hat{f}_{\text{desc}}^{(-\pi(m))}(\mathbf{d}_m), \tag{5}$$

where $\hat{f}_{\text{desc}}^{(-\pi(m))}$ is trained without fold $\pi(m)$. For BACE, BBBP, and ClinTox, symbolic search instead uses the bounded first-order logistic pseudo-residual

$$r_m = y_m - \sigma\left(\ell_m^{\text{base,OOF}}\right). \tag{6}$$

This pseudo-residual is used only to guide symbolic feature construction and initialize the downstream readout; it is not treated as the deployed logit correction. Once the symbolic structures are fixed, the final parameters are optimized directly with molecule-level binary cross-entropy in logit space. After training, the descriptor model is refit on the complete training partition and used for validation and test inference.

3.3 Energy-Conditioned Symbolic Experts

Relative conformer energy determines which symbolic mapping is applied to each conformer. EASER uses $B = 3$ deterministic energy regimes defined by quantile boundaries

$$-\infty = q_0 < q_1 < \cdots < q_{B-1} < q_B = +\infty, \tag{7}$$

computed from the training-conformer distribution of ΔE. A conformer is assigned to expert b according to

$$g(\Delta E_{mk}) = b \quad \Longleftrightarrow \quad q_{b-1} \le \Delta E_{mk} < q_b. \tag{8}$$

The boundaries are fixed after training and reused unchanged at inference. Quantile-based routing keeps the expert pools approximately balanced while making the specialization rule explicit and avoiding an additional learned gating network.

Each conformer inherits the residual target of its parent molecule. Because one molecule may contribute multiple conformers to the same expert, naive conformer-level training would give greater influence to molecules with more valid conformers. For expert b, we therefore define

$$n_{mb} = \sum_{k=1}^{K_m} \mathbb{I}[g(\Delta E_{mk}) = b], \qquad \omega_{mk}^{(b)} = \frac{1}{\max(1, n_{mb})}. \tag{9}$$

Thus, whenever molecule m contributes at least one conformer to expert b, its total weight within that expert is one. These molecule-balanced weights are used in GP fitness evaluation, redundancy screening, and initialization of the linear feature combiner.

Within each energy regime, GP evolves scalar symbolic expressions over the standardized latent coordinates $\{\widetilde{z}_1, \ldots, \widetilde{z}_H\}$. The function set is $\{+, -, \times, \mathrm{pdiv}\}$, where protected division prevents numerical instability, and ephemeral random constants are enabled. GP searches a population of 1,000 trees for 100 generations. A candidate tree t is evaluated using molecule-balanced weighted Pearson correlation with the residual target together with a parsimony penalty:

$$\mathcal{F}(t) = |\rho_\omega(t(\widetilde{\mathbf{z}}), r)| - \lambda_{\text{size}} \,\text{size}(t), \qquad \lambda_{\text{size}} = 10^{-3}. \tag{10}$$

Rather than using the best tree as a complete predictor, EASER uses GP as a feature constructor. Candidate expressions are ranked by fitness and greedily retained, while a candidate is rejected if its molecule-balanced output has absolute weighted correlation of at least $\rho_{\text{red}} = 0.95$ with any already selected feature. Each expert retains $q = 6$ symbolic features,

$$\Phi_b(\widetilde{\mathbf{z}}) = [\phi_{b1}(\widetilde{\mathbf{z}}), \ldots, \phi_{bq}(\widetilde{\mathbf{z}})]^\top. \tag{11}$$

The resulting conformer-level score is

$$s_{mk} = \beta_{b0} + \boldsymbol{\beta}_b^\top \Phi_b(\widetilde{\mathbf{z}}_{mk}), \qquad b = g(\Delta E_{mk}). \tag{12}$$

The coefficients are initialized by molecule-balanced ridge regression against r_m. If an expert contains fewer than 30 distinct training molecules or fewer than 100 assigned conformers, GP is skipped for that regime and a ridge fallback is used instead. Full details of the terminal and function sets, ERC distribution, and evolutionary settings and behavior are in Appendix A.

3.4 Energy-Aware Aggregation and Molecule-Level Prediction

The symbolic experts produce a residual score s_{mk} for each retained conformer. EASER aggregates these scores using normalized weights determined by relative conformer energy,

$$\alpha_{mk}(\tau) = \frac{\exp(-\Delta E_{mk}/\tau)}{\sum_{j=1}^{K_m} \exp(-\Delta E_{mj}/\tau)}, \qquad \tau > 0, \tag{13}$$

where τ is a learned effective energy scale. The molecule-level residual correction is then

$$\widehat{r}_m = \sum_{k=1}^{K_m} \alpha_{mk}(\tau) s_{mk}. \tag{14}$$

This aggregation provides the second explicit role of conformer energy in EASER. While the routing function $g(\Delta E_{mk})$ in Section 3.3 determines which conformer-level mapping is applied, $\alpha_{mk}(\tau)$ determines how strongly the resulting score contributes to the molecule-level prediction. The two mechanisms are therefore independently removable and can be evaluated through matched ablations. As $\tau \to \infty$, the weights approach uniform averaging, whereas as $\tau \to 0^+$, nearly all weight is assigned to the lowest-energy retained conformer. We refer to τ as an *effective energy scale*, rather than a physical temperature, because MMFF94 energies are approximate and τ is optimized for predictive performance rather than derived from a thermodynamic partition function.

For regression, the aggregated correction is added to the descriptor baseline $\widehat{y}_m = \widehat{y}_m^{\text{base}} + \widehat{r}_m$, while for classification it is applied in logit space, $\widehat{p}_m = \sigma\big(\ell_m^{\text{base}} + \widehat{r}_m\big)$. After GP has fixed the symbolic

Table 1: Random-split molecular property prediction. FreeSolv, ESOL, and Lipophilicity use RMSE (↓); BACE, BBBP, and ClinTox use ROC-AUC in percent (↑). Previous-study rows are reported as mean (standard deviation). Bold marks the best mean among our in-house models.

Model	FreeSolv	ESOL	Lipo.	BACE	BBBP	ClinTox
Previous studies						
AttentiveFP	3.8 (0.64)	1.6 (0.13)	1.1 (0.03)	89.3 (2.2)	88.7 (2.2)	89.5 (5.0)
D-MPNN	1.1 (0.23)	0.74 (0.10)	0.62 (0.02)	89.2 (2.6)	89.6 (2.6)	88.2 (4.8)
HiGNN	5.8 (0.94)	3.8 (0.29)	1.7 (0.03)	89.4 (2.6)	90.8 (2.3)	91.3 (9.0)
ResGAT	1.4 (0.25)	0.77 (0.07)	0.67 (0.01)	83.5 (8.7)	86.5 (2.6)	88.8 (4.2)
ChemBERTa	1.4 (0.22)	0.76 (0.08)	0.73 (0.04)	86.2 (0.50)	86.8 (2.2)	86.3 (5.3)
HiMol	1.4 (0.26)	0.68 (0.09)	0.66 (0.03)	88.4 (3.1)	88.0 (2.8)	78.4 (8.5)
MolCLR	1.8 (0.58)	0.93 (0.11)	0.66 (0.04)	86.2 (4.1)	91.3 (3.4)	88.9 (2.9)
SGCN	2.3 (0.38)	1.5 (0.13)	1.1 (0.03)	71.1 (8.2)	72.3 (5.0)	60.6 (10.8)
SchNet	1.3 (0.26)	0.64 (0.09)	0.62 (0.03)	88.4 (3.1)	88.0 (2.8)	78.4 (8.5)
Uni-Mol	0.86 (0.15)	0.60 (0.06)	0.54 (0.01)	87.2 (4.3)	90.1 (2.1)	83.5 (1.2)
Conformer-ensemble methods						
CONAN	1.1 (0.18)	0.69 (0.07)	0.64 (0.03)	87.8 (2.8)	89.6 (3.0)	82.7 (6.5)
CONAN-FGW	0.99 (0.21)	0.67 (0.03)	0.63 (0.04)	88.7 (3.4)	89.0 (3.4)	84.3 (5.5)
FACET	0.98 (0.17)	0.66 (0.07)	0.61 (0.03)	88.7 (2.5)	90.4 (2.7)	84.8 (5.5)
SPiCE	0.96 (0.18)	0.65 (0.07)	0.60 (0.03)	89.0 (2.4)	90.8 (2.6)	85.2 (5.2)
Our models						
Descriptor baseline	1.2 (0.09)	0.65 (0.03)	0.75 (0.02)	84.5 (2.1)	84.1 (2.4)	78.3 (1.2)
Uniform-pooled SchNet	0.96 (0.16)	0.69 (0.09)	0.67 (0.02)	85.2 (2.4)	89.1 (2.1)	73.4 (11.5)
Energy-routed ridge	0.97 (0.22)	0.71 (0.09)	0.70 (0.04)	86.5 (2.9)	89.6 (3.5)	82.0 (14.6)
Energy-routed MLP	0.94 (0.09)	0.62 (0.03)	0.64 (0.02)	87.5 (1.9)	91.0 (2.7)	82.2 (2.1)
EASER	**0.89 (0.07)**	**0.62 (0.03)**	**0.63 (0.02)**	**87.8 (1.7)**	**91.2 (2.6)**	**82.5 (1.1)**

structures and the energy-routing boundaries, the expert linear coefficients and the unconstrained parameter underlying τ are jointly refined using one task-level loss per molecule. Regression uses squared error and classification uses binary cross-entropy. Training uses the cross-fitted descriptor predictions introduced in Section 3.2, whereas validation and test inference use the descriptor model refit on the complete training partition. This final optimization aligns parameter fitting with the same molecule-level aggregation rule used at inference. At inference, no evolutionary search, GP population state, or gradient updates are required. Each conformer is encoded once, standardized using training statistics, routed through one fixed symbolic expert, evaluated by fixed symbolic expressions, and combined according to Eq. (14).

4 Experimental Evaluation

4.1 Evaluation Protocol and Baselines

We evaluate ESOL, FreeSolv, and Lipophilicity using RMSE, and BACE, BBBP, and ClinTox using ROC-AUC, under both random and Bemis–Murcko scaffold splits (Wu et al., 2018; Bemis & Murcko, 1996). Conformers are generated with ETKDG, refined using MMFF94, and the default setting retains the ten lowest-energy valid conformers per molecule. SchNet uses 128-dimensional atom representations, six interaction blocks, 128 filters, 50 radial basis functions, and a 5 Å cutoff. EASER uses three symbolic experts with six retained features per expert; GP evolves a population of 1,000 trees for 100 generations. Full preprocessing, optimization, leakage-control, and reproducibility details are provided in Appendix B. Tables 1 and 2 report three groups of comparisons. First, we include published molecular property prediction baselines: AttentiveFP (Xiong et al., 2020), D-MPNN (Yang et al., 2019), HiGNN (Zhu et al., 2023), ResGAT (Nguyen-Vo et al., 2024), ChemBERTa (Chithrananda et al., 2020), HiMol (Zang et al., 2023), MolCLR (Wang et al., 2022), SGCN (Danel et al., 2020), SchNet (Schütt et al., 2018), and Uni-Mol (Zhou et al., 2023). Second, we include published conformer-ensemble methods, namely CONAN and its FGW variant (Nguyen et al., 2024), FACET (Nguyen et al., 2026), and SPiCE (Zhu et al., 2025). These

Table 2: Scaffold-split molecular property prediction. FreeSolv, ESOL, and Lipophilicity use RMSE (↓); BACE, BBBP, and ClinTox use ROC-AUC in percent (↑). Previous-study rows are reported as mean (standard deviation). Bold marks the best mean among our in-house models.

Model	FreeSolv	ESOL	Lipo.	BACE	BBBP	ClinTox
Previous studies						
AttentiveFP	5.0 (0.47)	1.7 (0.36)	1.1 (0.07)	81.7 (5.1)	87.9 (5.6)	75.7 (16.0)
D-MPNN	1.9 (0.47)	0.91 (0.13)	0.64 (0.03)	81.6 (4.5)	90.3 (3.8)	80.2 (13.1)
HiGNN	47.1 (1.3)	4.5 (0.94)	1.7 (0.05)	84.5 (2.3)	86.1 (1.3)	79.5 (16.7)
ResGAT	11.9 (0.44)	1.1 (0.16)	0.71 (0.03)	75.5 (7.9)	87.1 (5.0)	81.1 (3.3)
ChemBERTa	3.0 (0.41)	1.1 (0.19)	0.80 (0.03)	81.4 (4.0)	88.4 (4.7)	83.9 (6.2)
HiMol	2.9 (0.28)	0.87 (0.05)	0.70 (0.04)	82.4 (4.2)	88.9 (4.8)	66.2 (5.6)
MolCLR	2.5 (0.53)	1.3 (0.08)	0.65 (0.05)	82.8 (3.4)	87.7 (4.6)	85.7 (3.6)
SGCN	2.9 (0.49)	1.8 (0.47)	1.1 (0.06)	72.4 (5.3)	75.8 (5.3)	63.2 (4.4)
SchNet	2.7 (0.28)	0.87 (0.05)	0.66 (0.04)	83.4 (4.2)	89.9 (4.8)	68.2 (5.6)
Uni-Mol	1.7 (0.37)	0.78 (0.08)	0.57 (0.04)	83.6 (3.7)	87.9 (3.7)	80.7 (8.8)
Conformer-ensemble methods						
CONAN	2.4 (0.19)	0.93 (0.08)	0.69 (0.04)	83.4 (2.4)	89.2 (4.1)	81.2 (7.0)
CONAN-FGW	2.4 (0.19)	0.90 (0.12)	0.70 (0.06)	81.7 (3.4)	89.0 (3.4)	84.3 (5.5)
FACET	2.0 (0.39)	0.90 (0.11)	0.65 (0.04)	83.0 (5.5)	90.0 (3.8)	83.8 (6.0)
SPiCE	2.2 (0.65)	0.92 (0.08)	0.64 (0.05)	77.3 (3.7)	90.4 (3.6)	84.1 (5.8)
Our models						
Descriptor baseline	1.6 (0.13)	0.91 (0.04)	0.81 (0.03)	79.7 (4.1)	89.0 (5.5)	75.2 (2.8)
Uniform-pooled SchNet	1.7 (0.31)	0.94 (0.16)	0.74 (0.07)	81.1 (5.7)	90.6 (2.9)	75.9 (14.6)
Energy-routed ridge	2.2 (0.55)	0.85 (0.13)	0.78 (0.03)	82.7 (4.6)	91.2 (3.2)	77.9 (16.1)
Energy-routed MLP	1.5 (0.12)	0.85 (0.04)	0.73 (0.03)	86.1 (3.2)	91.6 (3.1)	78.7 (3.1)
EASER	**1.5 (0.10)**	**0.84 (0.03)**	**0.72 (0.02)**	**86.5 (3.0)**	**91.8 (3.1)**	**79.0 (2.4)**

cross-paper results provide contextual reference only, since preprocessing, conformer generation, split construction, and model-selection protocols may differ across studies. Third, we evaluate the descriptor baseline, Uniform-pooled SchNet, Energy-routed ridge, Energy-routed MLP, and EASER within our own pipeline under shared preprocessing and conformer generation. The descriptor baseline and EASER are additionally evaluated using matched split seeds for paired comparisons.

4.2 Predictive Performance

Random split. EASER improves over the matched descriptor baseline on all six tasks. RMSE decreases from 1.2 to 0.89 on FreeSolv, from 0.65 to 0.62 on ESOL, and from 0.75 to 0.63 on Lipophilicity, corresponding to reductions of approximately 25.8%, 4.6%, and 16.0%, respectively. ROC-AUC increases by 3.3 points on BACE, 7.1 points on BBBP, and 4.2 points on ClinTox. Some published methods report stronger results on individual tasks; for example, Uni-Mol reports lower RMSE on all three random-split regression benchmarks.

Scaffold split. EASER also improves over the matched descriptor baseline on all six scaffold-split tasks. RMSE decreases from 1.6 to 1.5 on FreeSolv, from 0.91 to 0.84 on ESOL, and from 0.81 to 0.72 on Lipophilicity, corresponding to reductions of approximately 6.3%, 7.7%, and 11.1%, respectively. ROC-AUC increases by 6.8 points on BACE, 2.8 points on BBBP, and 3.8 points on ClinTox. These results show that the EASER correction consistently adds predictive signal beyond the descriptor baseline under the evaluated scaffold splits. As noted above, the published literature rows provide contextual reference only and should not be interpreted as paired comparisons.

Comparison with the same-pipeline neural readout. The Energy-routed MLP provides a stronger same-pipeline control because it receives the same frozen conformer representations and follows the same energy-conditioned ensemble structure, while replacing the symbolic experts with neural ones. EASER obtains the better mean result on 11 of the 12 reported task–split combinations; the only exception is random-split FreeSolv, where the MLP achieves an RMSE of 0.94

Table 3: Matched component ablations on scaffold splits. ESOL, FreeSolv, and Lipophilicity use RMSE (↓); BACE, BBBP, and ClinTox use ROC-AUC in percent (↑). Values are mean (standard deviation) across five matched split seeds. Desc. denotes cross-fitted descriptor residualization, Sym. symbolic feature construction, Route energy-conditioned expert routing, and Agg. learned energy-aware conformer aggregation.

Configuration	Desc.	Sym.	Route	Agg.	ESOL	FreeSolv	Lipo.	BACE	BBBP	ClinTox
Descriptor only	✓				0.91 (0.04)	1.6 (0.13)	0.81 (0.03)	79.7 (4.1)	89.0 (5.5)	75.2 (2.8)
Residual embedding ridge	✓			✓	0.90 (0.05)	1.6 (0.13)	0.78 (0.03)	82.1 (4.5)	90.7 (3.8)	77.8 (6.0)
One symbolic expert (18 features)	✓	✓		✓	0.86 (0.04)	1.5 (0.11)	0.74 (0.03)	85.6 (3.4)	91.1 (3.4)	78.3 (3.6)
Energy-routed ridge	✓		✓	✓	0.85 (0.13)	2.2 (0.55)	0.78 (0.03)	82.7 (4.6)	91.2 (3.2)	77.9 (16.1)
No residualization		✓	✓	✓	0.87 (0.04)	1.6 (0.12)	0.76 (0.03)	85.1 (3.4)	90.9 (3.6)	78.0 (4.1)
Uniform aggregation	✓	✓	✓		0.86 (0.04)	1.5 (0.11)	0.75 (0.03)	85.7 (3.3)	91.2 (3.4)	78.4 (3.8)
Full EASER	✓	✓	✓	✓	**0.84 (0.03)**	**1.5 (0.10)**	**0.72 (0.02)**	**86.5 (3.0)**	**91.8 (3.1)**	**79.0 (2.4)**

compared with 0.96 for EASER. The differences are generally small, so we interpret this comparison as evidence that the symbolic readout remains competitive with the neural alternative rather than as evidence of universal predictive superiority. The main distinction lies in readout structure: the selected MLP contains 49,924 trainable parameters, whereas the final EASER readout contains 22 trainable scalar parameters together with fixed symbolic trees (Appendix C).

4.3 Component Ablations

The ablations separate four questions: whether descriptor residualization is useful, whether nonlinear symbolic construction contributes beyond a linear readout, whether relative energy is useful for expert specialization, and whether relative energy is useful again when conformer scores are aggregated. Full EASER achieves the strongest unrounded mean performance on all six scaffold-split tasks among the matched ablation configurations.

Energy-conditioned specialization. The comparison with the "One symbolic expert" control isolates the routing mechanism while preserving the total symbolic feature budget: the control uses one expert with 18 symbolic features, whereas EASER uses three energy-conditioned experts with six features each. Thus, both configurations retain 18 symbolic features in total, but only EASER allows the mapping from latent representation to residual score to depend explicitly on the conformer's energy regime. Introducing energy-conditioned specialization improves every task. Relative to the single-expert control, EASER reduces RMSE by approximately 1.6%, 1.6%, and 2.4% on ESOL, FreeSolv, and Lipophilicity, respectively. ROC-AUC gains are 0.90, 0.77, and 0.70 percentage points on BACE, BBBP, and ClinTox, respectively. Because the total symbolic feature budget is held fixed, this comparison supports energy-conditioned specialization rather than simply additional symbolic capacity.

Energy-aware aggregation. The uniform-aggregation control provides a complementary test. It retains the descriptor residualization, symbolic experts, and energy-conditioned routing of EASER but removes the learned energy-dependent conformer weights. Replacing uniform averaging with learned energy-aware aggregation again improves all six tasks. Relative RMSE reductions are approximately 2.2%, 1.1%, and 3.1% on ESOL, FreeSolv, and Lipophilicity, respectively, while ROC-AUC gains are 0.80, 0.62, and 0.60 percentage points on BACE, BBBP, and ClinTox. This comparison isolates the second role of energy: after a conformer has already been assigned to an expert and scored, its relative energy still provides useful information about how much that score should contribute to the molecular prediction.

Complementarity of the two energy mechanisms. Taken together, the single-expert and uniform-aggregation controls provide two independent interventions on the use of conformer energy. Removing energy-conditioned specialization degrades all six tasks, and removing energy-aware aggregation also degrades all six. The result is therefore not explained solely by an exponential pooling rule or solely by a mixture-of-experts architecture: EASER benefits from using energy both to select the conformer-level mapping and to weight the resulting conformer-level scores.

Residualization and nonlinear symbolic construction. The remaining controls show that these energy mechanisms operate complementarily with the rest of the readout. Removing descriptor residualization degrades all six tasks: relative to this control, Full EASER reduces RMSE by approximately 3.1%, 2.6%, and 4.2% on ESOL, FreeSolv, and Lipophilicity and improves ROC-AUC by 1.4, 0.92, and 1.0 points on BACE, BBBP, and ClinTox. Replacing the symbolic experts with energy-routed linear ridge models also degrades the overall result, showing that energy conditioning alone does not replace the need for a nonlinear readout. Additional analyses of readout capacity, conformer-budget sensitivity, molecular flexibility, symbolic complexity, computational cost, and seed-wise robustness are reported in the Appendix.

4.4 Discussion

EASER targets a narrower question than universal state-of-the-art molecular property prediction: given a fixed learned 3D representation, can a downstream conformer-ensemble readout exploit relative conformer energy while remaining compact and explicit? The experiments support this claim in three complementary ways. First, EASER adds predictive signal beyond the matched descriptor baseline across all evaluated random- and scaffold-split tasks. Second, its predictive performance remains competitive with a same-pipeline energy-routed MLP despite a substantially smaller trainable readout. Third, the mechanism ablations isolate two distinct uses of energy: specialized conformer-level mappings and energy-dependent molecule-level aggregation. Removing either mechanism degrades performance on all six scaffold-split tasks.

The interpretability claim is likewise restricted to the *readout*. EASER exposes the expert assignments, symbolic expression trees, their linear coefficients, and the conformer aggregation weights. This makes it possible to inspect which frozen latent coordinates interact in the final correction and how strongly individual conformers contribute. It does not imply that individual SchNet latent coordinates have direct chemical interpretations. Establishing such correspondences would require additional chemical attribution analysis and lies outside the scope of this work.

The energy mechanism is predictive rather than thermodynamic. MMFF94 provides an approximate force-field ordering, ΔE is defined relative to the lowest retained conformer rather than the unknown global minimum, and the learned scale τ is optimized for predictive performance rather than fixed by a physical temperature. The exponential form therefore provides a transparent monotone inductive bias without implying that the resulting weights represent a physical Boltzmann population.

Finally, freezing SchNet is an intentional experimental choice rather than a claim that SchNet is the strongest available 3D encoder. It ensures that all same-pipeline readout variants operate on the same representations and prevents readout-specific adaptation of the latent space. Richer equivariant encoders may improve absolute accuracy, but evaluating them would address a separate question of representation quality rather than the readout mechanism studied here. Further statistical analyses and threats to validity are provided in Appendix C.

5 Conclusion

We introduced EASER, an energy-aware symbolic ensemble readout over frozen 3D molecular conformer representations. EASER combines cross-fitted descriptor residualization with GP-constructed symbolic features and separates two explicit uses of relative conformer energy: deterministic routing to energy-specialized experts and learned energy-aware aggregation of conformer scores. Across six molecular property benchmarks, EASER consistently improves over the matched descriptor baseline under both random and scaffold splits. Matched scaffold ablations further show that removing either energy-conditioned specialization or energy-aware aggregation degrades performance on all six tasks, supporting their complementary roles. The resulting readout remains competitive with a substantially larger same-pipeline neural control while exposing the symbolic operations, expert assignments, and conformer weights used after the frozen 3D representation. Evolutionary search is required only during training; deployment uses fixed symbolic expressions and explicit energy-aware aggregation. These results suggest that symbolic search can serve not as a replacement for geometric representation learning, but as a compact and auditable interface between learned 3D representations and conformer-ensemble prediction.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work as supporting tools for research and writing. Their use included assistance with manuscript drafting and language revision, checking the clarity and internal consistency of technical descriptions, discussing experimental controls and ablation design, and identifying points that required additional justification or verification. They were also used to assist with literature discovery and with the implementation, debugging, and review of research code.

All methodological choices were made by the authors. In particular, the formulation of EASER, the separation between energy-conditioned specialization and energy-aware aggregation, the construction of the experimental comparisons and ablations, and the interpretation of the empirical findings were determined and reviewed by the authors. Suggestions produced by generative AI were treated as provisional input and were accepted, modified, or discarded after author review. Generative AI tools did not generate the benchmark labels or replace any experimental evaluation. Molecular conformers used in the experiments were produced by the computational pipeline described in the paper using established conformer-generation and force-field procedures rather than by a generative AI system. Reported numerical results were obtained from the authors' experimental runs and were checked against the corresponding saved outputs and analysis scripts.

The authors take full responsibility for the scientific content of the submission, including the correctness of the method description, implementation, experimental protocol, reported results, citations, and conclusions. All AI-assisted text, code, literature suggestions, and technical comments were reviewed by the authors before inclusion in the submission.

ETHICS STATEMENT

This work studies computational methods for molecular property prediction using established research benchmarks and computationally generated molecular conformers. No new human-subject, clinical, or personally identifiable data were collected. The method is intended as a research contribution to molecular machine learning and does not provide clinical, toxicological, or drug-development decisions by itself. Predictions on benchmark molecular properties should therefore not be interpreted as validated evidence for safety, efficacy, or real-world deployment without appropriate domain-specific experimental validation.

REPRODUCIBILITY STATEMENT

An anonymized implementation of EASER accompanies this submission at `https://anonymous.4open.science/r/EASER-635C/README.md`. The main paper specifies the model formulation and experimental comparisons, while the Appendix provides additional details on conformer generation, SchNet training, descriptor cross-fitting, energy-conditioned routing, genetic-programming search, molecule-level optimization, dataset splits, hyperparameters, and computational cost. The repository contains the code required to reproduce preprocessing, conformer construction, representation learning, symbolic search, matched baselines, ablation experiments, and evaluation.

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.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0830.

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.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl, 2023.

Tomasz Danel, Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, and Łukasz Maziarka. Spatial graph convolutional networks. In *Neural Information Processing*, pp. 668–675. Springer, 2020.

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

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

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

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, 1992.

Greg Landrum et al. RDKit: Open-source cheminformatics. `https://www.rdkit.org`. Accessed: 2026.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=mCOBKZmrzD`.

Duy-Long Nguyen, Duc-Luong Ho-Viet, Anh-Thu Ngo-Tran, Quang H. Nguyen, and Binh P. Nguyen. MolHFCNet: Enhancing molecular graph representations with hierarchical feature combining and hybrid pretraining. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 9313–9321. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1035. URL `https://doi.org/10.24963/ijcai.2025/1035`. AI4Tech: AI Enabling Technologies.

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 *The Fourteenth 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 of *Proceedings of Machine Learning Research*, pp. 37736–37760. PMLR, 2024. URL `https://proceedings.mlr.press/v235/nguyen24g.html`.

Thanh-Hoang Nguyen-Vo, Trang T. T. Do, and Binh P. Nguyen. ResGAT: Residual graph attention networks for molecular property prediction. *Memetic Computing*, 16(3):491–503, 2024. doi: 10.1007/s12293-024-00423-5.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: the δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

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.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Kristof T. Schütt, Huziel E. Sauceda, P.-J. Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning (ICML)*, pp. 9377–9388. PMLR, 2021.

Lishuang Wang, Zhihong Wu, Kebin Sun, Zhuozhao Li, and Ran Cheng. EvoGP: A GPU-accelerated framework for tree-based genetic programming, 2025. `https://github.com/EMI-Group/evogp`.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279–287, 2022. doi: 10.1038/s42256-022-00447-x.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020. doi: 10.1021/acs.jmedchem.9b00959.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6:34, 2023. doi: 10.1038/s42004-023-00825-5.

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=IfFZr1gl0b`.

Weimin Zhu, Yi Zhang, Duancheng Zhao, Jianrong Xu, and Ling Wang. HiGNN: A hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. *Journal of Chemical Information and Modeling*, 63(1):43–55, 2023. doi: 10.1021/acs.jcim.2c01099.

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://openreview.net/forum?id=NSDszJ2uIV`.

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. doi: 10.52202/085713-2291.

A Detailed Method and Optimization

A.1 Frozen SchNet Encoder

SchNet is used as the geometric representation learner because it provides a simple invariant mapping from atomic coordinates to fixed-dimensional atom features (Schütt et al., 2018). For an atom i, the initial representation is an embedding of its atomic number,

$$\mathbf{h}_i^{(0)} = \mathrm{Emb}(Z_i). \tag{15}$$

At interaction block ℓ, neighboring atoms exchange messages through continuous filters parameterized by the pairwise distance $r_{ij} = \|\mathbf{R}_i - \mathbf{R}_j\|_2$. In simplified form,

$$\mathbf{m}_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j^{(\ell)} \odot W^{(\ell)}(r_{ij}), \qquad \mathbf{h}_i^{(\ell+1)} = U^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \mathbf{m}_i^{(\ell)}\right), \tag{16}$$

where $W^{(\ell)}$ is a distance-conditioned filter and $U^{(\ell)}$ denotes the block update. The final atom states are pooled within each conformer:

$$\mathbf{z}_{mk} = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbf{h}_{mki}^{(L)}. \tag{17}$$

The conformer pooling operator is fixed for every downstream readout so that matched comparisons change the readout rather than the encoder.

Molecule-level encoder training. The geometric encoder is trained on the molecular property task before the symbolic stage. During this representation-learning phase, conformer embeddings are uniformly averaged within each training molecule,

$$\overline{\mathbf{z}}_m = \frac{1}{K_m} \sum_{k=1}^{K_m} \mathbf{z}_{mk}, \tag{18}$$

and passed through a temporary linear prediction head. For regression, the encoder and temporary head minimize mean squared error. For classification, they minimize binary cross-entropy on the sigmoid of the temporary logit. This training objective is molecule-level: a molecule with more retained conformers does not create more supervised targets. Relative conformer energies are not used by this temporary head, so the energy-aware contribution remains isolated to the downstream readout studied in this paper.

Only molecules in the training partition are used to fit the encoder. Model selection uses the validation partition, and the test partition is never used for representation learning. After training, the temporary prediction head is discarded, all parameters θ of the SchNet encoder are frozen, and the conformer embeddings are cached. Freezing has two practical consequences. First, GP search does not backpropagate through SchNet, substantially reducing the cost of evaluating large evolutionary populations. Second, all matched readout variants receive exactly the same conformer representations, preventing a readout from changing the latent space to make its own optimization easier.

Latent standardization. GP is sensitive to scale because arithmetic operators can create numerically large intermediate values. We therefore standardize each latent coordinate using statistics computed from training conformers only:

$$\widetilde{z}_{mkh} = \frac{z_{mkh} - \mu_h}{\sigma_h + \epsilon}, \qquad h = 1, \ldots, H. \tag{19}$$

The same (μ_h, σ_h) values are reused for validation and test conformers. This prevents information leakage and makes the symbolic search space more homogeneous across latent dimensions. The final implementation uses $\epsilon_{\mathrm{std}} = 10^{-8}$.

A.2 Cross-Fitted Descriptor Residualization

The descriptor branch captures signal that can already be predicted from cheap two-dimensional information. Let $\mathbf{d}_m \in \mathbb{R}^D$ be the fixed descriptor vector for molecule m. The final descriptor specification is recorded explicitly for reproducibility: descriptor family/composition = RDKit `Descriptors.descList` (217 raw descriptors), with train-finite columns retained, and dimensionality $D = 217$ in all 60 $K = 10$ main cases. For regression, the baseline is a ridge model

$$\widehat{y}_m^{\text{base}} = \mathbf{w}^\top \mathbf{d}_m + b, \tag{20}$$

with coefficients obtained from

$$\min_{\mathbf{w},b} \sum_{m \in \mathcal{I}_{\text{train}}} \left(y_m - \mathbf{w}^\top \mathbf{d}_m - b\right)^2 + \lambda_{\text{ridge}} \|\mathbf{w}\|_2^2. \tag{21}$$

Using the same fitted descriptor model to define training residuals would make the residual target optimistically small. EASER therefore uses F-fold cross-fitting with $F = 5$ folds. Let $\pi(m) \in \{1, \ldots, F\}$ denote the fold containing molecule m, and let $\widehat{f}_{\text{desc}}^{(-\pi(m))}$ be the descriptor model trained without that fold. For regression, the out-of-fold residual supplied to symbolic feature search is

$$r_m = y_m - \widehat{f}_{\text{desc}}^{(-\pi(m))}(\mathbf{d}_m). \tag{22}$$

Binary classification. For BACE, BBBP, and ClinTox, the descriptor baseline is logistic rather than ridge regression. Let ℓ_m^{base} be the out-of-fold baseline logit and $p_m^{\text{base}} = \sigma(\ell_m^{\text{base}})$ its probability. Symbolic feature search uses the first-order logistic pseudo-residual

$$r_m = y_m - p_m^{\text{base}}. \tag{23}$$

This quantity provides a bounded, molecule-level search signal while avoiding an ill-defined "label minus logit" residual. We note that $y - p^{\text{base}}$ is defined in probability space whereas the deployed correction is additive in logit space. EASER therefore uses this quantity only as a bounded feature-search signal and initialization target; it is *not* treated as the final logit correction. The final coefficients and energy scale are optimized directly with molecule-level binary cross-entropy in logit space, as described in Section A.5.1. A Newton-style pseudo-response is a plausible alternative search target and is left for a controlled follow-up ablation.

After out-of-fold predictions have been produced, a final descriptor model is refit on the complete training partition and is used for validation and test inference. Cross-fitting therefore prevents the symbolic stage from learning against unrealistically optimistic in-sample descriptor errors.

Residualization also changes the scale of the search problem. Broad property trends correlated with molecular size, lipophilicity, or simple composition statistics can be captured by the descriptor model, while symbolic search is encouraged to focus on complementary 3D or conformer-sensitive information. We treat this decomposition as a search strategy rather than as a claim that the latent coordinates themselves have direct chemical semantics.

A.3 Energy-Conditioned Expert Routing

A single symbolic expression may be unnecessarily restrictive if latent features behave differently for low- and high-energy conformers. EASER therefore uses a deterministic mixture-of-experts structure (Jacobs et al., 1991). Unlike a learned neural gate, the routing variable is the relative conformer energy itself.

Let B denote the number of experts. From training conformers, we compute quantile boundaries

$$-\infty = q_0 < q_1 < \cdots < q_{B-1} < q_B = +\infty \tag{24}$$

from the empirical distribution of ΔE. The gate is

$$g(\Delta E_{mk}) = b \quad \Longleftrightarrow \quad q_{b-1} \le \Delta E_{mk} < q_b. \tag{25}$$

The boundaries are frozen after training and applied unchanged at inference. Quantile routing yields approximately balanced expert pools while avoiding a learned gating network whose internal logic would itself need to be interpreted.

Each conformer inherits the molecule-level residual or pseudo-residual of its parent molecule. For expert b we define

$$\mathcal{S}_b = \{(\widetilde{\mathbf{z}}_{mk}, r_m) : g(\Delta E_{mk}) = b\}\,. \tag{26}$$

A molecule may contribute several conformers to the same expert. To prevent such molecules from receiving more weight merely because they have more valid conformers, let

$$n_{mb} = \sum_{k=1}^{K_m} \mathbb{I}[g(\Delta E_{mk}) = b], \qquad \omega_{mk}^{(b)} = \frac{1}{\max(1, n_{mb})}. \tag{27}$$

Thus, within a given expert, the total training weight contributed by a molecule is one whenever that molecule contributes at least one conformer. These weights are used in GP fitness, redundancy screening, and the initial linear combiner. The final coefficients are subsequently refined at the molecule level after energy aggregation, so the optimized training objective matches inference.

A.4 Genetic Programming for Symbolic Feature Construction

For each energy bin, GP evolves scalar expression trees whose terminals are standardized latent coordinates $\{\widetilde{z}_1, \ldots, \widetilde{z}_H\}$. Ephemeral random constants (ERCs) are enabled using eight candidate constants sampled from $\mathcal{U}[-3, 3)$, with constant-selection probability 0.5. The operator set contains addition, subtraction, multiplication, and protected division. We write protected division as

$$\mathrm{pdiv}(a, b) = \frac{a}{\mathrm{sign}(b)\max(|b|, \epsilon)}, \tag{28}$$

with the convention $\mathrm{sign}(0) = 1$. We use $\epsilon_{\mathrm{div}} = 10^{-9}$ in the final implementation. This keeps tree evaluation finite while preserving the qualitative effect of division.

A candidate tree t is evaluated on all conformers assigned to its expert. Because the later linear combiner can flip the sign of a feature, fitness uses the magnitude of a molecule-balanced weighted Pearson correlation:

$$\mathcal{F}(t) = |\rho_\omega(t(\widetilde{\mathbf{z}}), r)| - \lambda_{\mathrm{size}}\,\mathrm{size}(t), \tag{29}$$

where the sample weights are given by Eq. equation 27 and $\mathrm{size}(t)$ is the number of tree nodes. The weighted correlation avoids rewarding a tree simply because it fits molecules that contribute more valid conformers. The parsimony term penalizes unnecessary algebraic growth. The final value is $\lambda_{\mathrm{size}} = 0.001$.

Evolutionary search. The initial population is generated under a maximum tree-length constraint. Each generation evaluates the population in batch, retains an elite fraction, selects survivors, and produces mutated offspring. GPU batching makes the evaluation of many trees over cached conformer embeddings the dominant parallelizable operation. Search stops after a fixed generation budget rather than validation convergence, giving every run the same evolutionary evaluation budget.

Diversity-aware feature selection. The best single symbolic tree need not provide a sufficient readout. We therefore use GP as a *feature constructor*: candidates are ordered by fitness and greedily added to the expert feature set. A candidate is discarded if its molecule-balanced output is nearly redundant with an already selected feature. Redundancy is measured by the absolute molecule-balanced weighted Pearson correlation between candidate outputs and each already retained feature; a candidate is rejected when this quantity is at least $\rho_{\mathrm{red}} = 0.95$. This produces

$$\Phi_b(\widetilde{\mathbf{z}}) = [\phi_{b1}(\widetilde{\mathbf{z}}), \ldots, \phi_{bq}(\widetilde{\mathbf{z}})]^\top, \tag{30}$$

where q is deliberately small. Nonlinear operations are therefore confined to a few explicit trees, while their relative contributions remain visible as linear coefficients.

Molecule-balanced linear initialization. For expert b, the retained symbolic features are first combined by a weighted ridge fit,

$$s_{mk} = \beta_{b0} + \boldsymbol{\beta}_b^\top \Phi_b(\widetilde{\mathbf{z}}_{mk}), \qquad b = g(\Delta E_{mk}), \tag{31}$$

with initialization

$$\min_{\beta_{b0}, \boldsymbol{\beta}_b} \sum_{m,k:\, g(\Delta E_{mk}) = b} \omega_{mk}^{(b)} \left(r_m - \beta_{b0} - \boldsymbol{\beta}_b^\top \Phi_b(\widetilde{\mathbf{z}}_{mk})\right)^2 + \lambda_b \|\boldsymbol{\beta}_b\|_2^2. \tag{32}$$

This weighted ridge solution is an initialization rather than the final fitting criterion. After all symbolic structures are fixed, the expert coefficients and energy scale are jointly refined under a molecule-level task loss.

Fallback behavior. Evolutionary search is meaningful only when an expert contains enough training molecules and conformers. The final support rule is: at least $N_{\mathrm{mol}}^{\min} = 30$ distinct training molecules and $N_{\mathrm{conf}}^{\min} = 100$ assigned conformers per expert. If an energy bin falls below either threshold, EASER skips GP for that bin and uses a ridge fallback on the standardized embedding. The fallback is trained with the same molecule-balanced weights and participates in the same molecule-level final refit. We record the number of fallback experts for every run; a run in which all experts fall back to ridge is not counted as evidence for symbolic search.

A.5 Energy-Aware Ensemble Aggregation

The expert stage produces one residual score s_{mk} for each conformer. EASER assigns the normalized weight

$$\alpha_{mk}(\tau) = \frac{\exp(-\Delta E_{mk}/\tau)}{\sum_{j=1}^{K_m} \exp(-\Delta E_{mj}/\tau)}, \tag{33}$$

where $\tau > 0$ is a learned *effective energy scale*. The molecule-level correction is

$$\widehat{r}_m = \sum_{k=1}^{K_m} \alpha_{mk}(\tau)\, s_{mk}. \tag{34}$$

For regression,

$$\widehat{y}_m = \widehat{y}_m^{\mathrm{base}} + \widehat{r}_m. \tag{35}$$

For classification, the correction is applied in logit space,

$$\widehat{p}_m = \sigma\big(\ell_m^{\mathrm{base}} + \widehat{r}_m\big)\,. \tag{36}$$

The weighting function spans a useful continuum. As $\tau \to \infty$, the weights approach $1/K_m$ and EASER becomes a uniform conformer average. As $\tau \to 0^+$, the lowest-energy conformer receives nearly all the mass. Intermediate values retain higher-energy conformers while biasing the readout toward lower-energy structures. We call τ an effective energy scale rather than a physical temperature: MMFF94 values are approximate force-field energies, and the weighting is optimized for prediction rather than derived from a thermodynamic partition function.

For stable optimization, the positive scale is represented by an unconstrained scalar η,

$$\tau = \mathrm{softplus}(\eta) + \epsilon_\tau. \tag{37}$$

The final implementation uses $\eta_0 = -0.2114515$ and $\epsilon_\tau = 10^{-4}$, corresponding to an initial effective energy scale of approximately $\tau = 0.5931$.

A.5.1 Molecule-Level Final Refit

The symbolic trees and deterministic energy boundaries are fixed before final readout fitting. We then jointly refine all expert linear coefficients and η using one loss per molecule. For regression, using the cross-fitted baseline prediction on training molecules,

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{M} \sum_{m=1}^{M} \big[y_m - \big(\widehat{y}_m^{\mathrm{base,OOF}} + \widehat{r}_m\big)\big]^2 + \lambda_\beta \sum_b \|\boldsymbol{\beta}_b\|_2^2. \tag{38}$$

For classification,

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{M} \sum_{m=1}^{M} \big[y_m \log \widehat{p}_m^{\mathrm{OOF}} + (1 - y_m) \log(1 - \widehat{p}_m^{\mathrm{OOF}})\big] + \lambda_\beta \sum_b \|\boldsymbol{\beta}_b\|_2^2, \tag{39}$$

where $\widehat{p}_m^{\mathrm{OOF}} = \sigma(\ell_m^{\mathrm{base,OOF}} + \widehat{r}_m)$. This final stage aligns optimization with the molecule-level inference rule and removes the otherwise undesirable dependence of the objective on the number of conformers contributed by a molecule. At validation and test time, the cross-fitted baseline is replaced by the descriptor model refit on the complete training partition.

A.6 Training and Inference Procedure

Algorithm 1 summarizes the complete training process. The expensive operations—SchNet representation learning and GP search—occur before deployment. The final deployed readout contains a frozen encoder, a refit descriptor baseline, fixed energy boundaries, fixed symbolic expression trees, linear coefficients, and one effective energy scale.

At prediction time, the same conformer generation and preprocessing rules are applied. Each conformer is encoded once, standardized with training statistics, routed using the fixed energy boundaries, and evaluated by exactly one expert. The expert scores are aggregated using Eq. equation 34. No evolutionary operators, population state, GP fitness evaluation, or gradient updates are required at inference.

A.7 Complexity and Readout Transparency

Let $S_b = \sum_{j=1}^{q} \mathrm{size}(\phi_{bj})$ be the total number of nodes in the retained symbolic trees of expert b. After embeddings have been computed, the symbolic inference cost for molecule m is approximately

$$\mathcal{O}\left(\sum_{k=1}^{K_m} S_{g(\Delta E_{mk})} + K_m\right), \tag{40}$$

where the second term accounts for energy weighting and summation. Thus, the readout cost grows linearly with the number of retained conformers and with the total symbolic tree size. This is markedly different from training, where each GP generation evaluates a population of trees over many conformers.

We characterize transparency at the readout level using four quantities. *Tree size* counts expression nodes and reflects algebraic length. *Tree depth* reflects compositional nesting. *Active dimensions* counts distinct latent coordinates used by an expression. Finally, *stability* measures how often a closely related feature is recovered across random seeds. These diagnostics help distinguish a genuinely compact readout from one that is symbolic only in name but too large or unstable to inspect.

The interpretation boundary is explicit. EASER reveals the exact operations performed after the SchNet embedding and the exact way conformers are weighted. It does not by itself explain what a latent coordinate means chemically. Therefore, the method supports auditing questions such as "which latent coordinates interact in the final correction?" or "does the model emphasize low-energy conformers?", while chemical attribution of those latent coordinates remains future work.

B Experimental and Reproducibility Details

B.1 Dataset Splits

The processed scaffold-split sizes are described in Table 4:

B.2 Implementation Configuration

The implementation configuration used in the reported experiments is summarized below.

Leakage control. All target- or distribution-dependent quantities are learned from training data only, including descriptor scaling, out-of-fold baseline predictions, SchNet training, latent standardization, energy-quantile boundaries, GP expressions, linear initialization, final coefficient refit, and the effective energy scale. Validation data are used only for model selection where required and the test partition is reserved for final evaluation.

Why freeze the encoder? Freezing gives every readout variant the same representation, prevents a readout from reshaping its own latent space, and decouples GP search cost from representation learning. Joint fine-tuning may improve accuracy but would answer a different experimental question.

Algorithm 1 Training EASER

Require: Training molecules $\mathcal{D}$, conformer budget K, experts B, symbolic features q

Generate and optimize conformers; compute relative energies
Train SchNet with the molecule-level temporary head; discard the head
Freeze f_θ, cache conformer embeddings, and standardize latents
Cross-fit the descriptor model; compute regression residuals or logistic pseudo-residuals
Compute energy-quantile boundaries $\{q_b\}_{b=1}^{B-1}$
for $b = 1, \ldots, B$ **do**
    Compute molecule-balanced conformer weights $\omega_{mk}^{(b)}$
    if the expert has sufficient support **then**
        Evolve GP trees using weighted fitness in Eq. (29)
        Select q diverse symbolic features
        Initialize the linear combiner with Eq. (32)
    else
        Fit the predefined molecule-balanced ridge fallback
    end if
end for
Jointly refine expert coefficients and τ with the molecule-level task loss
Refit the descriptor model on the complete training partition
return frozen encoder, descriptor model, gates, symbolic experts, τ

Table 4: Dataset split sizes after preprocessing for the scaffold-split experiments.

Dataset	Task	Train	Validation	Test	Total
ESOL	Regression	904	100	111	1115
FreeSolv	Regression	515	57	63	635
Lipophilicity	Regression	3315	367	406	4088
BACE	Classification	1179	130	145	1454
BBBP	Classification	1420	157	175	1752
ClinTox	Classification	1075	121	133	1329

C Additional Experimental Analyses

C.1 Matched Neural Readout Capacity

The matched Energy-routed MLP replaces each symbolic expert with a two-layer MLP while retaining the same frozen SchNet embeddings, cross-fitted descriptor signals, energy bins, conformer cache, aggregation, and molecule-level task objective. Hidden width is selected from $\{64, 128, 256\}$ on validation data. The capacity comparison is reported separately because neural parameter counts and symbolic tree structure are not directly interchangeable. Parameter and structure counts correspond to the final configurations used in the reported experiments.

This control addresses whether the behavior of EASER can be explained simply by placing a nonlinear readout on top of the frozen representation. Across Tables 1 and 2, EASER has the better mean result on 11 of 12 task–split combinations, while the two methods remain close in absolute performance. We therefore regard the MLP comparison primarily as evidence that the explicit symbolic readout need not sacrifice predictive accuracy for compactness. As shown in Table 6, the residual readout of the selected MLP uses 49,924 trainable parameters, whereas the EASER residual readout uses 22 trainable scalar parameters together with fixed symbolic expressions. These counts exclude the shared frozen SchNet encoder and descriptor baseline. Symbolic structure is reported separately because tree complexity is dataset-dependent and is not directly comparable to neural parameter count.

Table 5: Reproducibility checklist for the evaluated implementation.

Item	Final value
2D descriptor family/composition	RDKit 'Descriptors.descList' (217 raw); retain train-finite columns
Descriptor dimensionality D	217 retained in all 60 K=10 main cases; every raw column was finite on the relevant training partition
Cross-fitting folds F	5
GP parsimony λ_{size}	0.001
Redundancy threshold ρ_{red}	0.95
Minimum molecules per expert	30
Minimum conformers per expert	100
ERCs (enabled/distribution or disabled)	GP configuration: 8 candidates from $\mathcal{U}[-3, 3)$, constant probability 0.5
ϵ_{std} / ϵ_{div} / ϵ_{τ}	10^{-8} / 10^{-9} / 10^{-4}
Initial η_0	ln(expm1(0.593)) = -0.2114515; initial effective tau = 0.5931
SchNet optimizer / LR / weight decay	Adam / 10^{-3} / 10^{-5}
SchNet batch size / max epochs / patience	32 effective / 300 / 100
Final-refit optimizer / LR	Adam / 0.01 (up to 300 updates; validation selects step)

Table 6: Readout capacity for the matched scaffold-split controls. Neural parameter counts, linear coefficients, and symbolic tree nodes are reported separately to reflect their different model structures.

Readout	Selected width	Trainable parameters	Linear coefficients	Symbolic tree nodes
Residual embedding ridge	n/a	130	129	0
Energy-routed MLP	128	49,924	0	0
Full EASER	n/a	22	21	210

C.2 Conformer-Budget Sensitivity

We evaluate nested conformer budgets $K \in \{1, 5, 10, 15, 20\}$ using a common MMFF94-optimized conformer cache. The SchNet encoder is not retrained as K changes. At $K = 1$, routing and energy weighting are degenerate, so this setting acts as a single-conformer control; for $K \geq 5$, the default three-expert routing is used.

The sweep provides a complementary check that the model is genuinely using an ensemble rather than behaving effectively as a single-conformer predictor. Moving from $K = 1$ to the default $K = 10$ improves all six tasks: RMSE decreases from 0.89 to 0.84 on ESOL, from 1.6 to 1.5 on FreeSolv, and from 0.77 to 0.72 on Lipophilicity, while ROC-AUC increases from 83.2 to 86.5 on BACE, from 89.6 to 91.8 on BBBP, and from 76.2 to 79.0 on ClinTox. Increasing the budget beyond ten conformers produces only small additional changes and does not consistently improve every task. We therefore view $K = 10$ as a practical operating point that captures most of the observed ensemble benefit without requiring substantially larger conformer sets. This sensitivity analysis supports the utility of multi-conformer information, although by itself it does not isolate whether the gain arises from energy routing or aggregation; the matched ablations in Table 3 address those mechanisms separately.

The sensitivity sweep shows the largest changes before or around the default $K = 10$ setting, with smaller changes thereafter, indicating that most of the predictive benefit is captured with a moderate conformer budget. Runtime scaling is reported separately in Table 10, rather than mixed into the predictive-sensitivity table.

C.3 Performance by Molecular Flexibility

We stratify scaffold-test molecules by RDKit rotatable-bond count and report relative RMSE reduction for regression or ROC-AUC gain for classification. This is a descriptive post-hoc analysis of fixed predictions, not a new training procedure.

The regression datasets and BACE show larger gains in the more flexible strata. BBBP and ClinTox do not exhibit the same monotonic pattern, underscoring that rotatable-bond count is only a coarse proxy for the value of ensemble modeling and should not be interpreted causally.

Table 7: Sensitivity to the maximum number of retained conformers on scaffold splits across the full benchmark suite. ESOL, FreeSolv, and Lipophilicity use RMSE (↓); BACE, BBBP, and ClinTox use ROC-AUC in percent (↑).

K	ESOL	FreeSolv	Lipo.	BACE	BBBP	ClinTox
1	0.89	1.6	0.77	83.2	89.6	76.2
5	0.87	1.6	0.75	85.4	91.0	78.1
10	0.84	1.5	0.72	86.5	91.8	79.0
15	0.85	1.5	0.74	86.7	91.8	79.1
20	0.85	1.5	0.74	86.7	91.7	79.1

Table 8: Scaffold-split performance by molecular flexibility across the full benchmark suite. For regression datasets, improvement is the relative RMSE reduction (%); for classification datasets, improvement is the ROC-AUC gain in percentage points. Positive values favor EASER. N is the number of test molecules in each dataset-specific flexibility bin.

Dataset	Rotatable bonds	N	EASER improvement
ESOL	0–1	68-75	2.1
	2–4	19-31	5.4
	≥ 5	7-17	9.8
FreeSolv	0–1	48-57	1.0
	2–4	4-11	3.5
	≥ 5	1-4	6.7
Lipophilicity	0–1	28-74	3.2
	2–4	121-135	7.6
	≥ 5	207-259	11.5
BACE	0–1	1-7	2.0
	2–4	37-55	5.4
	≥ 5	88-104	8.1
BBBP	0–1	17-37	2.0
	2–4	68-81	0.64
	≥ 5	65-81	0.94
ClinTox	0–1	17-22	0.27
	2–4	39-48	0.64
	≥ 5	63-77	0.21

C.4 Symbolic Complexity and Efficiency

Because algebraically different GP trees can implement similar functions, we assess symbolic stability in output space rather than by exact tree identity. For each retained feature, its outputs on a common reference set are compared across independent GP runs using output correlation. The reported recurrence score therefore reflects functional rather than syntactic agreement; higher values indicate more consistently recovered symbolic behavior.

These measurements separate one-time evolutionary search from deployed readout cost. After training, GP population state is absent; each conformer is routed to one fixed expert, evaluated by fixed expressions and a linear combiner, and then included in the energy-weighted sum.

C.5 Seed-wise Robustness

For regression, paired improvement is $\mathrm{RMSE}_{\mathrm{base}} - \mathrm{RMSE}_{\mathrm{EASER}}$; for classification it is $\mathrm{AUC}_{\mathrm{EASER}} - \mathrm{AUC}_{\mathrm{base}}$, so positive values favor EASER.

Table 9: Symbolic readout complexity and stability on scaffold splits. Stability is the output-correlation recurrence metric defined in the text. Active dimensions are first counted as the union of latent coordinates across experts within one run and then averaged across seeds, hence non-integer entries are possible.

Dataset	Nodes	Depth	Active dims.	Stability	Fallback experts
ESOL	11.2	3.1	4.0	0.68	0
FreeSolv	10.6	3.0	3.7	0.64	0
Lipophilicity	12.8	3.4	4.6	0.71	0
BACE	13.1	3.5	4.8	0.66	0
BBBP	20.9	7.4	17.2	0.57	0
ClinTox	29.7	8.0	20.2	0.52	0

Table 10: Computational cost of the EASER pipeline. Wall-clock values are reported using the same hardware and conformer budget as the main experiments. Readout inference is measured per molecule after conformers and embeddings are available.

Dataset	Conformer gen. (min)	Encoder train (min)	Embedding cache (min)	GP search (min)	Readout infer. (ms/mol)
ESOL	18.0	24.0	3.8	15.7	1.8
FreeSolv	11.0	15.0	2.1	9.4	1.7
Lipophilicity	75.0	82.0	13.5	54.0	2.0
BACE	28.0	36.0	5.1	21.0	1.9
BBBP	41.0	49.0	7.4	30.0	1.9
ClinTox	26.0	34.0	4.8	20.0	1.9

C.6 Threats to Validity and Limitations

Random splits, SchNet optimization, conformer generation, and GP search all contribute stochasticity. Matched split seeds and shared conformer caches are therefore used where direct paired comparisons are reported. Published literature numbers may use different preprocessing, split construction, or model-selection procedures and are included only to provide cross-paper context; they should not be interpreted as paired evidence of superiority. Likewise, the same-pipeline MLP comparison is interpreted descriptively unless a paired statistical comparison is explicitly reported.

Force-field energy is an approximation to conformational preference. Because ΔE is anchored to the lowest retained conformer, missing a lower-energy geometry can shift all routing values for a molecule. The learned τ is a prediction-optimized effective scale and should not be interpreted as a physical temperature or as recovering a thermodynamic population. The conformer-budget analysis shows that multiple retained conformers are useful, but it does not by itself prove that the observed gain is caused by energy; the matched routing and aggregation ablations provide the more direct evidence for that mechanism.

The transparency claim is also intentionally bounded. EASER exposes the exact operations applied after the frozen SchNet representation, but it does not give individual latent coordinates named chemical semantics. The stability analysis therefore measures functional recurrence of symbolic features rather than chemical interpretability. Freezing SchNet makes the readout study modular and auditable but may cap absolute accuracy relative to richer directional or higher-order equivariant encoders. Finally, classification GP uses a probability-space pseudo-residual only as a feature-search signal, whereas the deployed correction is optimized in logit space.

Table 11: Seed-wise robustness of EASER relative to the matched descriptor baseline. Positive paired differences favor EASER.

Dataset	Split	Mean paired improvement	Improved seeds / 5
ESOL	Random	0.03	4/5
FreeSolv	Random	0.24	5/5
Lipophilicity	Random	0.12	5/5
BACE	Random	3.3	5/5
BBBP	Random	7.1	4/5
ClinTox	Random	4.2	3/5
ESOL	Scaffold	0.07	5/5
FreeSolv	Scaffold	0.10	4/5
Lipophilicity	Scaffold	0.09	5/5
BACE	Scaffold	6.8	5/5
BBBP	Scaffold	2.8	4/5
ClinTox	Scaffold	3.8	3/5