QT-Net: Rethinking Evaluation of AI Models in Atomic Chemical Space

Pablo Martínez Crespo
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
pabloma@chalmers.se

Stefano Ribes
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Richard Beckmann
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Marisa Gliege
Chief Technology Office
EMD Electronics

Santiago Miret
Lila Sciences

Robert S. Jordan
Technology Research
Intel Corporation

Vijay Kris Narasimhan
M Ventures

Martin Rahm
Department of Chemistry and Chemical Engineering
Chalmers University of Technology

Rocío Mercado
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
and Science for Life Laboratory (SciLifeLab), Gothenburg
rocio.mercado@chalmers.se

Abstract

Atomic properties such as partial charges or multipoles encode chemically meaningful information that can inform downstream molecular property prediction, but their evaluation as machine learning targets has been complicated by the absence of a principled out-of-distribution evaluation protocol at the atomic level. In this work, we propose a held-out evaluation protocol that clusters atomic environments by SOAP descriptors and computes metrics accounting only for cluster labels unseen during training. Following this procedure, we use 5×5 cross-validation and Tukey's HSD to run a statistically rigorous comparison of E(3)-equivariant against non-equivariant, rotationally augmented models for predicting electron populations and multipoles of H, C, N, and O atoms. Building on our results, we introduce the Quantum Topological Neural Network (QT-Net), a rotationally augmented, non-equivariant graph neural network. We show that QT-Net can be used to infer properties of atoms in molecules from QM9 outside our training set, and that the atomic environments seen during training strongly affect the performance of molecular property prediction tasks. To further validate the framework, molecular dipole moments computed from QT-Net's per-atom outputs recover the ground-truth values reported in QM9. We release all code and data, including a JAX implementation of QT-Net.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Molecular property prediction [21, 31], accelerated simulations [20, 22], and *de novo* molecular design [40, 15] are among the main drivers of new artificial intelligence (AI) for science methods. However, out-of-distribution (OOD) generalization and broad sampling are among the most pressing challenges in the field. [11, 35, 2] Molecular representations built from first principles atomic properties are a promising avenue for tackling this research gap. [18, 36, 14, 23, 1]

Because atoms in molecules and materials interact as open quantum systems that exchange energy and charge with their neighbors, not all atoms of the same element can be treated as equal. Quantum chemists have developed several atomic descriptors that can also be effective for downstream property prediction, including conceptual density functional theory (DFT) reactivity indices such as Fukui functions [30], electrostatic surface potentials and local ionization energies [18], natural population analysis [36], interacting quantum atoms (IQA) energy decompositions [25, 7, 16], and the various definitions of atomic partial charge [27, 39, 5, 9]. Works by Vargas et al. [37] and Gee et al. [13] showed that descriptors derived from the quantum theory of atoms in molecules (QTAIM) can enhance regression tasks such as predicting activation energies in reactions, or formation and orbital energies of transition metal complexes with less than 1 000 training points. Recently, the AIMEl dataset [28], a subset of QM9 [33] with quantum topological atomic (QTA) properties for H, C, N and O such as electron populations, multipole moments, and localization indices, was released. Some open questions from these works are whether QTA properties inferred by a model would preserve their predictive power, and which properties are the most informative outside the region of chemical space that a model was trained on.

Several studies have focused on learning atomic partial charges and multipoles derived from first-principles. [27, 19, 39, 38, 5] Nevertheless, truly OOD evaluation is precluded by the usage of random splits, which cannot prevent atomic environments from leaking into test sets. Atoms, as the building blocks of chemistry and molecules, are open quantum systems with complex interactions of their own dictated by their electronic structure. In this work, we show that metrics that account for all atoms in a test set of molecules, rather than only atoms in environments unseen during training, lead to overconfident results that don't reflect true OOD performance.

In view of this, we introduce the Quantum Topological Neural Network (QT-Net) for learning QTA properties, a densely connected, non-equivariant graph neural network (GNN). To address the aforementioned limitations, our key contributions are:

- **Statistically significant benchmarking.** We apply repeated-measures ANOVA (RM-ANOVA) and Tukey's honestly significant difference (HSD) frameworks for comparing E(3)-equivariant against non-equivariant multitask models to predict scalar and tensor properties for topological atoms.
- **Faithful OOD evaluation.** Our evaluation set, built upon per-element clustering of smooth overlap of atomic positions (SOAP), contains atomic environments unseen during training, and our metrics only account for atoms in said unseen environments.
- **Inferential quality assessment.** We infer QTA properties for all the QM9 atoms outside of AIMEl, and show that the molecular dipole moment is recovered from its atomic contributions.

2 Related work

A productive line of work has applied machine learning (ML) to QTA properties drawn from QTAIM, the IQA formalism, and related branches of quantum chemical topology (QCT), establishing that these chemical descriptors are within reach of modern regressors. Early contributions used Gaussian process regression to predict the full set of IQA energy components for small biomolecular fragments [25], and were subsequently extended to clusters with the formamide-dimer transfer-learning study of Brown et al. [7], which expanded the property catalog to atomic multipole moments. The works by Gallegos et al. [9, 10] introduced neural networks for predicting QTAIM atomic charges of H, C, N, and O atoms, and one-body and two-body QTAIM and IQA descriptors. Most recently, Hoffmann et al. [16] have shown that equivariant graph networks can learn IQA-decomposed atomic and pairwise energies across diverse benchmark sets. Together, these works establish QTA properties as a realistic ML target and have steadily expanded the space of accessible descriptors.

A methodological pitfall nevertheless recurs across this literature, bearing directly on the OOD-evaluation contribution we make in this work. Mainly, these works rely on splits that do not isolate unseen atomic environments. Maxwell et al. [25] and Gallegos et al. [9] draw their splits at random from a common pool of geometries, which is bound to leak atomic environments seen during training into the test set. Their follow-up [10] probes extrapolation along reaction coordinates and through a supramolecular trajectory — useful as configurational-space stress tests, but uninformative about generalization across chemical space. Brown et al. [7] sample training and evaluation from a single molecular-dynamics trajectory of the same monomer. They show that their monomer-trained multipoles fail to transfer to the dimer regime, providing perhaps the clearest signal that environment-level OOD is a standing challenge.

3 Methods

Figure 1 summarizes the methodology followed to train and evaluate the proposed QT-Net models.

3.1 Target properties of topological atoms

QTAIM defines an atom as a region in real space $\Omega \subset \mathbb{R}^3$ such that $\nabla\rho(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) = 0 \; \forall \; \mathbf{r} \in \partial\Omega$, where ρ is the electron density and $\hat{\mathbf{n}}(\mathbf{r})$ is the normal vector to the surface defined by $\partial\Omega$, meaning there is no flux of electrons through the boundary of the basin. Atomic properties depend on the shape of this basin, as they are quantities integrated over it. In this work, we focus on electron population (N), atomic contribution to molecular dipole moment ($\boldsymbol{\mu}$), traceless quadrupole moment ($\mathbf{Q}$), and localization index (λ),

$$N = \int_\Omega \rho(\mathbf{r}) d\mathbf{r}, \quad \boldsymbol{\mu} = -\int_\Omega e\,\mathbf{r}^{(\Omega)} \rho(\mathbf{r}) d\mathbf{r} + \sum_{\Lambda}^{N_b(\Omega)} \left[\mathbf{R}_\Omega - \mathbf{R}_b(\Omega|\Lambda)\right] q(\Omega|\Lambda),$$

$$\text{and } Q_{\alpha\beta} = -\int_\Omega e\,\rho(\mathbf{r}) \left(3 r_\alpha^{(\Omega)} r_\beta^{(\Omega)} - \delta_{\alpha\beta} \|\mathbf{r}^{(\Omega)}\|^2\right) d\mathbf{r}.$$

In the expressions above, $\rho(\mathbf{r})$ is the electron density, e is the charge of the electron, indices $\alpha, \beta \in \{x, y, z\}$ denote Cartesian components s.t. $r_\alpha^{(\Omega)}$ is the α component of the displacement from the nuclear position $\mathbf{R}_\Omega$, $N_b(\Omega)$ is the number of bond critical points (BCPs) connected to atom Ω, $\mathbf{R}_b(\Omega|\Lambda)$ is the position of the BCP between atoms Ω and Λ and $q(\Omega|\Lambda)$ is the charge transfer between these atoms. The localization index λ is defined from the variance of the electron population within a basin as $\lambda = N - \sigma^2(N)$, and quantifies the extent to which electrons are localized within the basin Ω vs shared with neighboring basins. We refer the interested readers to the book by Matta and Boyd [24]. Here N is a scalar, μ is a 3-vector, Q is a rank-2 traceless symmetric tensor with 5 independent components, and λ is a scalar.

3.2 Dataset construction and atomic environments

We used the QTAIM atomic multipoles (N, $\boldsymbol{\mu}$, $\mathbf{Q}$) and localization indices (λ) reported in AIMEl [28]. The AIMEl Zenodo repository [29] releases data in two formats: a summary `.csv` file covering approximately 40 000 molecules but containing only scalar properties, and per-molecule `.sumviz` files (raw AIMAll output) that include the full tensor components of $\boldsymbol{\mu}$ and $\mathbf{Q}$. To enable the comparison of equivariant and non-equivariant models on tensorial targets, we parsed the `.sumviz` files directly. These were available for 31 223 of the molecules, of which we removed 411 entries with missing molecular property values. The resulting AIMEl-derived dataset contains 30 812 molecules.

To enable evaluation on atomic environments not seen during training, we group atoms into element-specific clusters of similar local environment, which are later used to construct held-out test sets of atoms whose cluster labels never appear in the training data. For each atom we computed the SOAP descriptors [3] (cutoff 4.5 Bohr, 8 radial basis functions, spherical harmonics up to $L = 6$, Gaussian smearing $\sigma = 0.7$ Bohr; 3696 features per atom). These were reduced using principal component analysis (PCA) to element-specific numbers of features preserving at least 99% of the variance, and clustered separately by atomic species using HDBSCAN [26]. Each atom in the dataset was then annotated with its element-centered environment label. Cluster statistics are reported in Table 4 and Figures 5 and 6 in the Appendix.

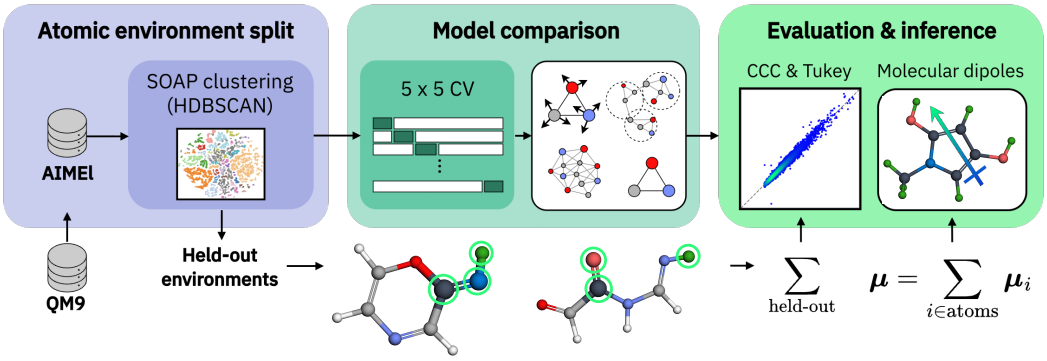

Figure 1: QT-Net training and evaluation pipeline on AIMEl for atomic (QTAIM) property prediction. The molecules at the bottom are sample molecules from the test set (left: `N=C1OC=CN=C1`; right: `N=CNC(=O)C=O`), where the circled atoms are in environments held-out from training and are the only ones contributing to the scores. The molecule with the dipole moment (`CN1C=C(O)C=C1O`) is part of the physics-based test outside AIMEl. Color for atoms in held-out environments: H_10, C_11, N_13, O_10. Color for the other atoms: H , C, N, O.

We now describe the co-occurrence closure construction of the held-out set for the evaluation of QTA models. For reproducibility purposes, we use A_i in the following to refer to the label of an atomic environment centered around element A, as these are the labels reported with our dataset. We selected N_13 first, as nitrogen was the limiting factor for holdout size; over 90% of molecules containing N_13 also contained C_11 and H_10, so both were added to prevent co-occurrence leakage. Including these three labels automatically absorbed all O_10-containing molecules, which we therefore retained as a fourth evaluation label at no additional cost in training-set size. The final test set contains 5 840 molecules, with at least one unseen cluster label per atom type (see the bottom of Fig. 1), forming a coherent OOD region of chemical space at both the molecular and atomic levels.

3.2.1 Data curation for molecular experiments

Our experiments on molecular property prediction run on QM9 [33], of which AIMEl is a subset. We exclude the 3 054 molecules reported to fail the geometry-consistency check (molecules whose DFT-optimized geometry does not match their input SMILES—flagged by the QM9 authors as a known issue). Molecules containing fluorine are also excluded to enable the study of how QTA-properties affect downstream molecular tasks, as there is no data for F in AIMEl. In total, the working dataset for these experiments contains 128 908 molecules, with 98 096 for which QTAIM properties were inferred using QT-Net. We retain four target properties: isotropic polarizability (α), HOMO–LUMO energy gap ($\Delta\varepsilon$), internal energy at 0 K (U_0), and heat capacity at 298.15 K (C_v). Together, these span both electrostatic-response and energetic properties, for which the estimation of atomic multipoles is plausibly informative.

These 128 908 molecules provide a total of 2 333 341 atoms which are clustered by environment following the protocol of Section 3.2, except for a SOAP cutoff of 6 Bohr to match the receptive radius of the molecular models. This alternative set of atomic labels allows us to analyze the extrapolation regime of the inferred QTAIM properties respect to the training distribution in atomic chemical space, and whether atomic environments play a role in molecular property prediction. At the same time, we enable OOD evaluation in molecular chemical space by assigning each molecule to a ‘chemotype’ cluster. For this, we compute binary Morgan fingerprints (radius 2, 2 048 bits), normalize them to unit L2 norm, project them on 50 principal components, and cluster them with k-means (k = 50, 10 initializations, fixed seed).

3.3 Cross-validation at the atomic and molecular levels

All models are evaluated using a 5-repeat, 5-fold cross-validation (5×5 CV) protocol, yielding 25 fold-level scores per model. A standard 5×5 CV protocol computes performance metrics on fold-specific holdouts drawn from the data. Here, predicting atomic properties in the data regime of the AIMEl dataset forces the usage of a common held-out set of atomic environments for all 25 folds. We justify

this design in the Limitations (Sec. 6) and in the Appendix A.6, and provide metrics of the statistical validity of the comparison in the Appendix A.5. The remaining AIMEl molecules are partitioned into fold-specific training and validation sets. Within each repeat, molecules are split into five folds using `GroupKFold` with the Bemis–Murcko scaffold as the grouping variable, so all molecules sharing a scaffold are assigned to the same fold; each fold's validation set therefore contains only scaffolds entirely absent from the training set. Scaffold-to-fold assignments are re-shuffled with an independent seed at each repeat, making the five repeats statistically distinct.

For the molecular experiments, we follow a standard 5×5 CV protocol drawing the folds from AIMEl and leaving the remainder of QM9 purely for inference. Each fold withholds a group of molecular chemotypes and a set of atomic environments at the same time. The 48 chemotypes present in AIMEl are dealt into five groups of whole chemotypes with balanced molecule counts. Environments are withheld through five environment folds, which automate the co-occurrence closure of Sec. 3.2 on the 6-Bohr labels. Fold (r, c) of repeat r withholds chemotype group c and environment fold $(r + c) \bmod 5$, making each chemotype group meet each environment fold once over the five repeats. Every molecule with an atom in a withheld environment leaves training, too. Because a molecule keeps its chemotype group while the withheld environment fold rotates between repeats, it is tested both by models that saw its environments and by models that did not. See Appendix A.4 for further details and t-SNE visualizations.

3.4 Statistical framework of the comparison of models for QTA properties

We compare model performance using a repeated-measures ANOVA (RM-ANOVA) with Tukey's honestly significant difference (HSD) as the post-hoc pairwise test. Data limitations force a common hold-out set across all folds (Sec. 6), on which we compute per-fold concordance correlation coefficients (CCC) for each (holdout label, property). Because the five folds within a repeat share this hold-out and overlap in training data, their scores are correlated. We therefore average them into a single repeat-level score, yielding five approximately independent observations per model in a balanced design with repeat as subject and model as treatment. Means and 95% confidence intervals are reported as $\bar{x} \pm t_{0.975,4} \cdot \mathrm{SEM}$, where SEM is the standard error of the five repeat-level means. Tukey HSD then defines a minimum significant difference (MSD) such that two models differ significantly only if the gap between their repeat-averaged means exceeds it. The same procedure applies to any subset of (atom × property) cells, and we report the diagnostics of normality, homoscedasticity and fold independence in Appendix A.5.

3.5 The QT-Net architecture

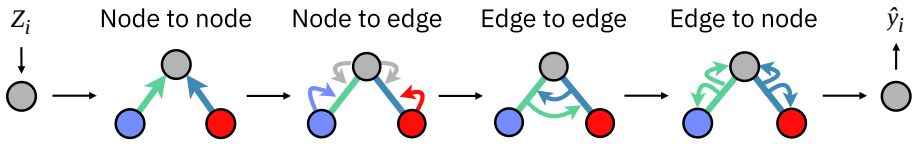

Figure 2: Schematic figure of a layer of message passing of QT-Net.

The set of transformations that compose a layer of QT-Net is represented in Figure 2. We focus on the transformations for the scalar versions of QT-Net, as that was the best-performing design. See Section A.10 for details on E(3)-equivariant models. Our architecture is inspired by the attention and gating mechanisms of the work by Irwin et al. [17], the separation of feature and geometric channels by Schütt et al. [34], and the chemical reminders, raw-component input, and per-layer readout by Pozdnyakov and Ceriotti [32].

We represent each atom by a node with scalar features h_i. For edge features, we follow the same convention, but with double subindices (e.g., h_{ij}). We denote the element-wise (Hadamard) product with $\odot$. In the equations below that can apply to either type of cell, we absorb the subindices of nodes and edges into the subscripts $r,\ s,\ t$ to refer to receivers, senders and intermediaries (respectively). In messages between nodes, the intermediary would be the edge connecting them (and the common node between two edges for messages between edges). The usage of Hadamard product implies that the output dimension of MLPs and geometric filters match the number of features of the receiver. We use SiLU as activation function between hidden layers.

In all models, node features are initialized with a message passing layer as follows:

$$h_i^{(0)} = \text{LayerNorm}\left(\sum_j \text{MLP}\left(\text{EMBED}(Z_i), \text{EMBED}(Z_j)\right) \odot \Gamma\left(\hat{\mathbf{r}}_{ij}, \text{RBF}(r_{ij})\right)\right), \quad (1)$$

where Z_i, Z_j, are the atomic species. Γ is a geometric gate with $\tanh$ activation output that takes as input the components of the inter-atom displacement vector, concatenated to a projection onto radial basis functions (RBFs) as implemented in DimeNet [12]:

$$\text{RBF}_n(r) = \sqrt{\frac{2}{c}} \frac{\sin(n\pi r/c)}{r} \frac{1+\cos(\pi r/c)}{2}.$$

Edge features are initialized from these node features as follows

$$h_{ij}^{(0)} = \text{LayerNorm}\left(\text{MLP}\left(\text{LayerNorm}\left(\sum_{k\in\{i,j\}} \text{MLP}\left(h_k^{(0)}\right)\right)\right) \odot \Gamma\left(\hat{\mathbf{G}}_{ij}, \text{RBF}(r_{ij})\right)\right), \quad (2)$$

where $\hat{\mathbf{G}}_{ij} \propto \hat{\mathbf{r}}_{ij} \otimes \hat{\mathbf{r}}_{ij}^T$ is the edge's normalized gyration tensor expressed in the traceless symmetric representation. See Section A.2.

After these, all layers follow a residual update structure, $h_r' = h_r + g\left(h_r\right) \odot M_r(\tilde{h}_r; \{\tilde{h}_s, \tilde{h}_t\})$, where g is a gating MLP with a sigmoid output that depends on the unnormed features, and $M_r(\tilde{h}_r; \{\tilde{h}_s, \tilde{h}_t\})$ is a message aggregation with $\tilde{h}_r, \tilde{h}_s, \tilde{h}_t$ being LayerNormed receiver, sender, and intermediary features, respectively. In general,

$$M_r\left(\tilde{h}_r; \{\tilde{h}_s, \tilde{h}_t\}\right) = \sum_{s,t\in\mathcal{N}_t(r)} \text{MLP}\left(\tilde{h}_r, \tilde{h}_s, \tilde{h}_t\right) \odot \Gamma(\mathbf{R}_r, \mathbf{R}_s),$$

where $(\mathbf{R}_r, \mathbf{R}_s)$ is some real-space geometric relation between receiver and sender. In node-to-node and node-to-edge messages, we use the same convention as in Eqs. 1 and 2, respectively. In edge-to-edge messages, $(\mathbf{R}_r, \mathbf{R}_s) \equiv \left(\hat{\mathbf{G}}_{rs}, P(\cos\left[\mathbf{G}_r, \mathbf{G}_s\right])\right)$, where $\hat{\mathbf{G}}_{rs} = (\mathbf{G}_r - \mathbf{G}_s)/\|\mathbf{G}_r - \mathbf{G}_s\|$ is the unit-norm relative gyration tensor and we project $\cos\left(\mathbf{G}_r, \mathbf{G}_s\right) = \text{Tr}\left(\hat{\mathbf{G}}_r^T \hat{\mathbf{G}}_s\right)$ onto a basis of Legendre polynomials, $\{P_n(x)\}$, since the gyration tensors contain angular information.

At the end of a messaging cycle, we apply a feed-forward (FF) residual update following the expression

$$h_r' = h_r + g(h_r) \odot \text{MLP}\left(\tilde{h}_r, W \cdot \text{EMBED}(Z_r)\right),$$

where we include a linear transformation of the initial embedding of the atomic species to serve as a chemical reminder (but that term is omitted for edges' FF).

Finally, the prediction of the model for the properties of a given atom, $\hat{y}_i = (\hat{N}_i, \hat{\lambda}_i, \hat{\boldsymbol{\mu}}_i, \hat{\mathbf{Q}}_i)$ is a readout that depends on an atomic head at the end of each layer:

$$\hat{y}_i^{(l)} = W \cdot \text{MLP}\left(\text{LayerNorm}\left(h_i^{(l)}\right)\right), \text{ and } \hat{y}_i = \sum_l \alpha^{(l)} \hat{y}_i^{(l)},$$

where the layer weights α come from a softmax, $\alpha^{(l)} = \exp(\omega^{(l)}) / \sum_{l'} \exp(\omega^{(l')})$.

4 Experiments

4.1 Atomic property prediction

Our objective was to find out which architectural features offered the best tradeoff between computational cost and quality of inferred QTA properties with respect to the ground truth distribution. In order to find the best candidate for this inference task, we trained a series of E(3)-equivariant and non-equivariant models for QTA prediction, both with fully connected and cutoff based models,

ensuring all atoms had at least 2 neighbors within cutoff (see Fig. 17). We notice that the models' cutoff radii are longer than the one used for clustering the atomic environments and show that this does not incur serious data leakage in Appendix A.6. Hyperparameter optimization was run for a first batch of experiments, after which we further tuned width and depth of the models to keep sensible, matched compute cost across models (see Appendix A.1). We trained for a number of epochs between 1500 and 3000, keeping training wall-time below 2 days per fold. Nevertheless, all models plateaued after epoch 1000. The parameters with the lowest validation loss were used for each fold's predictions.

Table 1: Architectural details of the main models in the discussion. E-models are equivariant. Data augmented models saw randomly rotated molecules every epoch.

	Architectural details			
Models	Cutoff (Bohr)	Max. n.n.	Depth (layers)	Data augmented
EGFC	-	-	4	-
SFC2	-	-	7	Yes
SGFC	-	-	7	No
SG-8-12	8	12	7	Yes
SG-8-5	8	5	7	Yes
EGNN	5.25	5	4	-
SGNN	5.25	5	7	No

4.2 Molecular properties from inferred QTA properties

We ran experiments on molecular property prediction on the AIMEl subset with a molecular version of QT-Net that used a 3.5 Bohr cutoff with 4 n.n. and 4 layers. The resulting receptive field reaches 98.3% of the atoms within a 6-Bohr environment while deliberately limiting model capacity. Each model was trained both with ('informed') and without ('blind') ground truth QTA properties as input. Every model is fed geometry through QT-Net's geometric filters, so an informed advantage cannot come from geometry implicit in the QTA properties. Both variants were trained with identical hyperparameters for 2 000 epochs on molecules randomly rotated at every epoch, under the 5×5 CV of Sec. 3.3 at four training fractions (200 models in total), and we kept the parameters with the lowest validation loss.

An ensemble of 5 models with the atomic QT-Net architecture was re-trained on the whole ground truth in AIMEl, having different validation sets for each member of the ensemble for early stopping. We use this ensemble to infer QTA properties for each of the atoms in the remainder of QM9. The quality of the multipoles and localization indices inferred by this ensemble was tested by two different experiments. In the first, we directly sum the inferred atomic contributions to the molecular dipole moments, and compare the resulting molecular dipole moments against the ground truth in QM9. In the second, every molecular model trained on AIMEl was evaluated, without retraining, on the 98 096 molecules of the remainder. The informed variant read the QTA properties inferred by the ensemble, normalized with the statistics of its own training set.

5 Discussion

5.1 Comparison of QTA property prediction on OOD atomic environments

For each model, element, and QTA property, we report Lin's concordance correlation coefficient (CCC), defined as $\rho_c = 2\,\mathrm{Cov}(y,\hat{y})/(\sigma_y^2+\sigma_{\hat{y}}^2+(\mu_y-\mu_{\hat{y}})^2)$. This metric jointly penalizes deviations from the identity line and mean bias, making it more sensitive to always-positive quantities such as $|\boldsymbol{\mu}|$ and $|\mathbf{Q}|$. Our discussion revolves around the interplay among cutoff, number of nearest neighbors (n.n.) and architectural constraints. The statistical validity of the comparison gets slightly compromised when digging into the peculiarities of each species, and therefore we defer this part of the discussion to Appendix A.8. More architectures were tested (including topological message passing [6, 8] and tensor product layers), which offered no advantage at increased computational cost. Interested readers are referred to the repository for more information.

Table 2: Tukey p-value matrix for CCC values averaged over atom types and properties, comparing data augmented, non-equivariant models with equivariant models. Values above 0.05 (marked in **bold**) indicate a statistically insignificant difference in performance. The best performing models are determined based on CCC mean value and statistical significance of the difference: 1st 2nd 3rd.

	Tukey p-values					
Models	EGFC	SFC2	SG-8-12	SG-8-5	EGNN	CCC ± 95% CI
EGFC	1.0000	0.0010	0.0010	**0.5729**	0.0010	0.798±0.006
SFC2	0.0010	1.0000	**0.8968**	0.0010	0.0010	0.895±0.006
SG-8-12	0.0010	**0.8968**	1.0000	0.0010	0.0010	0.888±0.006
SG-8-5	**0.5729**	0.0010	0.0010	1.0000	0.0010	0.809±0.013
EGNN	0.0010	0.0010	0.0010	0.0010	1.0000	0.754±0.028

The results in Table 2 show that non-equivariant, rotationally augmented models outperform the equivariant counterparts significantly. Looking at Table 1, one would be tempted to attribute this advantage to the greater depth of the scalar models. We argue that this depth is not being invested in passing chemical information, but rather in polishing the purely geometrical information that equivariant models understand by design. Interested readers are referred to Appendix Section A.8, where we discuss this possibility based on results from non-augmented models.

With the effect of the layer depth aside, the results suggest that the connectivity beyond 5 neighbors is mostly providing the non-equivariant models with geometric information, rather than chemical. On the one hand, notice that SG-8-5 performs on-par with EGFC. On the other, SG-8-12 outperforms EGFC and matches SFC2. Therefore, the increased amount of neighbors of SG-8-12 against SG-8-5 is helping the model achieve better geometric awareness, which doesn't improve beyond 12 neighbors. Given these results, we chose the SG-8-12 architecture for QT-Net.

5.2 Atomic environments determine the accuracy of molecular predictions

To test the selected architecture on unseen molecules, we ran a physics-based test where we computed the molecular dipole moments for the 98 096 molecules of the remainder from their inferred atomic contributions. Following Matta and Boyd [24], the molecular dipole moment can be obtained from the output of QT-Net for each atom Ω as $\boldsymbol{\mu} = \sum_{\Omega} \boldsymbol{\mu}(\Omega)$.

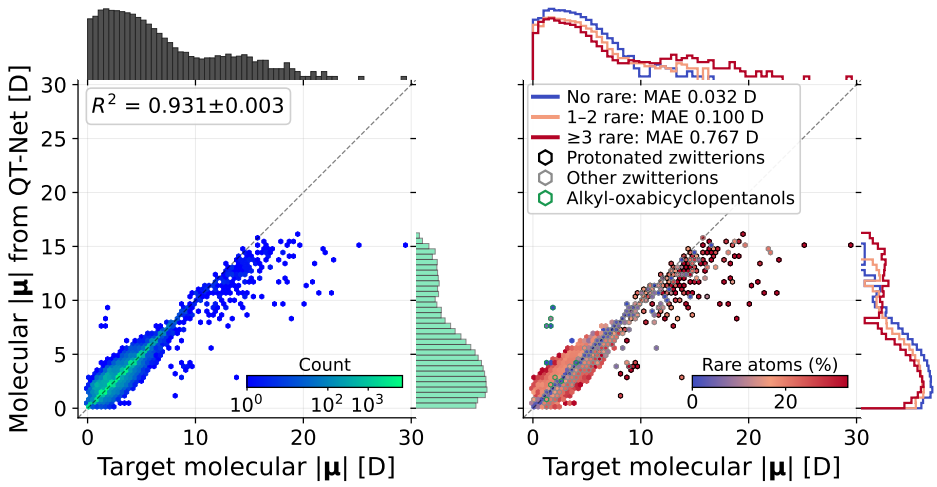

Figure 3: **(left)** Parity plot of molecular dipole moments of the 98 096 QM9 molecules outside AIMEl, computed from the atomic contributions inferred by QT-Net. **(right)** The same plot, colored by each molecule's percentage of rare atoms. Marginals split the molecules by their number of such atoms.

Although QT-Net *was never trained on molecular dipoles*, the sums recover the magnitudes reported in QM9 with $R^2 = 0.931$ (Fig. 3, left). This recovery is a strong physical consistency check: because molecular dipole moments arise from the vector sum of per-atom contributions, accurate reconstruction requires that QT-Net has correctly learned not only the magnitudes but also the orientations of the inferred atomic dipoles relative to one another. This indicates that the model has captured the underlying QTAIM partitioning of the electron density and not simply memorized artifacts of the AIMEl distribution.

Regarding molecular property prediction, our experiments show that QTA descriptors do not provide significant advantages in general. We dig deeper into this analysis in Appendix A.9. Our analysis revealed that the atomic environments seen during training matter more than the descriptors. As each molecule is predicted by models that saw its atomic environments or chemotype and by models that did not (Sec. 3.3), the cost of each holdout can be measured on the same molecules.

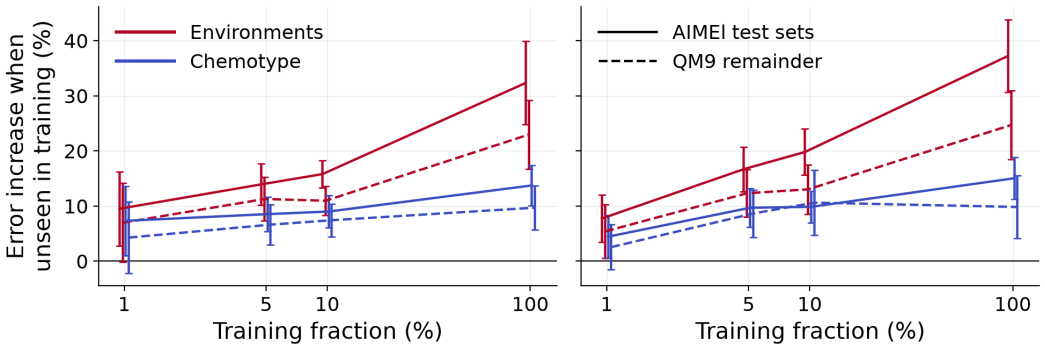

Figure 4: Error increase on the same molecules when their atomic environments or their chemotype were unseen in training, relative to the models that saw them, averaged over the four targets (Appendix A.9). **(left)** Blind and **(right)** informed models. Fold means with 95% t-intervals.

In both studies, we found the distance to the training distribution in atomic – rather than molecular – chemical space to be the main source of error. We measure the distance of an atom to a training set as the distance to its 15th-nearest training atom of the same element in the 6-Bohr descriptor space (see Appendix A.4), and say that an atom is 'far from AIMEl' or 'rare' when it sits beyond the 99th percentile of the same distance for in-distribution atoms. For the reconstructed dipoles, the errors concentrate where the remainder leaves the atomic chemical space of AIMEl. Molecules with 0, 1–2 and $\geq$3 atoms far from AIMEl have MAEs of 0.032, 0.100 and 0.767 D (Fig. 3, right). For 855 of the molecules with $\geq$3 rare atoms, the reconstruction is nearly exact. The rest of the molecules in this group are alkynes (14 molecules in AIMEl against 17 468 in the remainder 77% of the error) and zwitterions. QT-Net thus transfers to rare environments of familiar chemistry, but not to chemistry its training set lacks (see Appendix A.9). For the molecular models, molecules with $\geq$3 atoms far from their training set have 1.3 to 1.9 times the error of molecules with none on the AIMEl test sets, and up to 4.2 times on the remainder (see Fig. 19).

6 Limitations

Here, we justify the usage of a common held-out set for evaluating the 25 folds of the models in Section 5.1. In standard molecular benchmarks, scaffolds are randomly assigned to folds, so the expected test-set difficulty is approximately equal across folds. Between-fold variance therefore reflects sampling noise, and the ANOVA exchangeability assumption holds approximately. At the atomic level, atomic environments play the role of "scaffolds", but they co-occur within molecules: specific cluster labels are mutually enriched, so the presence of one label in a molecule raises the probability of finding another.

During training, showing the model an atomic label that is highly enriched with a held-out label would expose the model to atomic environments that co-reside in the same molecules as the held-out environments, allowing it to learn their structural context indirectly. Mutually enriched label sets must therefore be withheld together as a unit. In the AIMEl data regime, this constraint severely

limits the number of admissible holdout combinations and precludes randomizing label assignments across repeats.

As a result, the Tukey HSD comparisons presented in this work are conditional on the chosen holdout and cannot be interpreted as claims of general OOD superiority of either design, nor of general superiority of non-equivariant over E(3)-equivariant models. We use them as guidance for deciding on the final QT-Net design for the inference and downstream tasks.

7 Conclusion

Our work shows a statistically thorough analysis of the performance difference between E(3)-equivariant and purely scalar, rotationally augmented models for the task of learning scalar and tensorial QTA properties. Our results suggest that scalar models benefit from densely connected graphs for capturing geometrical information, especially when data augmentation is used. We therefore justify a non-equivariant design for the introduced architecture, QT-Net, which we demonstrate can reconstruct molecular dipole moments outside the set of training molecules from inferred atomic contributions, with errors concentrated in environments that AIMEl barely samples, most of all alkynes and zwitterions.

We want to emphasize that OOD evaluation of AI models for QTA properties requires careful labeling and tracking of the atomic environments. The properties of an atom depend on its environment, and the same element can appear in chemically distinct environments within a molecule. Our approach shows that SOAP-based clustering can separate these chemically different environments, and enables us to construct OOD splits by withholding molecules that contain environments absent from the training set. Furthermore, using metrics that specifically measure model performance on these unseen environments enables us to design an architecture based on faithful metrics of OOD performance. Additional results in Section A.7 (specifically in Fig. 16) show that **metrics that account for all atoms in a set of test molecules lead to overconfident results that do not evaluate a model's true OOD performance.** This pitfall has been repeatedly observed in the AI-for-QCT literature (see Section 2).

At the molecular level, the coverage of atomic environments matters more than atomic descriptors. Holding out atomic environments and chemotypes at the same time, from 803 training molecules on we found that unseen environments raise the error of the same molecule by 11–37%, 1.2–2.5 times the cost of an unseen chemotype. QTA properties, whether ground truth or inferred, do not make up for them. The atomic environments seen in training are therefore a more important hinge of molecular property prediction than atomic descriptors, and evaluations of molecular models should track them as carefully as evaluations of atomic ones.

Our results suggest that further progress on QTA property prediction is now primarily a data problem rather than a representation or architectural challenge: across the four properties studied here, the same architecture achieves CCC values ranging from $\sim$0.4 to $\sim$0.99 depending on element and property, indicating that the bottleneck has shifted from how to model these targets to which targets to model and on what chemistry. The dipole analysis makes this concrete: alkynes, with 14 samples in AIMEl against 17 468 in the remainder, account for 77% of the error of the reconstructed dipoles. Extending the evaluation protocols introduced here to other quantum-mechanically derived atomic descriptors—such as conceptual DFT reactivity indices, natural bond orbital populations, IQA decompositions—and to broader regions of chemical space is, in our view, the natural next step.

8 Code and data availability

Code and data for this project are available at `https://github.com/pablomcrespo/QT-Net`. It includes all seeds and information to reproduce the experiments of this work, besides relevant code to reproduce the statistical analysis. The source data is referenced in the text, but we include compressed versions of our curated files (mainly for reproducibility with the same cluster labels). The QTAIM properties inferred by QT-Net are released in the repository, too.

Acknowledgments and disclosure of funding

PMC and RB acknowledge funding from the Intel-Merck joint university research center for AI-Aware Pathways to Sustainable Semiconductor Process and Manufacturing Technologies (AWASES). RM acknowledges funding provided by the Wallenberg AI, Autonomous Systems, and Software Program (WASP), supported by the Knut and Alice Wallenberg Foundation. PMC acknowledges Lynne Ressel for the inspiration in naming QT-Net. The computations and data storage were enabled by resources provided by Chalmers e-Commons and by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. The authors declare no competing interests.

Author contributions

PMC performed the main research activities, collected and curated the dataset, analyzed the results, prepared the visualizations, and wrote the original draft of the manuscript. SR and RB contributed to the planning of the cross validation experiments, statistical analysis and visualizations. MG, SM, RSJ and VKN contributed with methodological discussion and supervision. RM managed resources and acquired funding for the project with support from SM and VKN. MR and RM contributed through conceptual input, critical discussion, supervision, and manuscript review and editing.

References

[1] Javier E Alfonso-Ramos, Rebecca M Neeser, and Thijs Stuyver. Repurposing quantum chemical descriptor datasets for on-the-fly generation of informative reaction representations: Application to hydrogen atom transfer reactions. *Digital Discovery*, 3(5):919–931, 2024.

[2] Evan Antoniuk, Shehtab Zaman, Tal Ben-Nun, Peggy Li, James Diffenderfer, Busra Sahin, Obadiah Smolenski, Everett Grethel, Tim Hsu, Anna Hiszpanski, et al. BOOM: benchmarking out-of-distribution molecular property predictions of machine learning models. *Advances in Neural Information Processing Systems*, 38, 2026.

[3] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

[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, 2023. URL `https://arxiv.org/abs/2206.07697`.

[5] Tristan Bereau, Denis Andrienko, and O Anatole Von Lilienfeld. Transferable atomic multipole machine learning models for small organic molecules. *Journal of Chemical Theory and Computation*, 11(7): 3225–3233, 2015.

[6] Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yuguang Wang, Pietro Lio, Guido F Montufar, and Michael Bronstein. Weisfeiler and Lehman go cellular: CW networks. *Advances in Neural Information Processing Systems*, 34:2625–2640, 2021.

[7] Matthew L Brown, Bienfait K Isamura, Jonathan M Skelton, and Paul LA Popelier. Incorporating noncovalent interactions in transfer learning gaussian process regression models for molecular simulations. *Journal of Chemical Theory and Computation*, 20(14):5994–6008, 2024.

[8] Pablo Martínez Crespo, Robert S. Jordan, Marisa Gliege, Santiago Miret, Vijay Kris Narasimhan, and Rocío Mercado. TopoMole: Topological message passing meets hyperedge messages. In *AI for Accelerated Materials Design - NeurIPS 2025*, 2025. URL `https://openreview.net/forum?id=k39fE6SD4u`.

[9] Miguel Gallegos, Jose Manuel Guevara-Vela, and Ángel Martín Pendás. NNAIMQ: A neural network model for predicting QTAIM charges. *The Journal of Chemical Physics*, 156(1), 2022.

[10] Miguel Gallegos, Valentin Vassilev-Galindo, Igor Poltavsky, Ángel Martín Pendás, and Alexandre Tkatchenko. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nature Communications*, 15(1):4345, 2024.

[11] Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: A benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems*, 35:21342–21357, 2022.

[12] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=B1eWbxStPH`.

[13] Winston Gee, Abigail Doyle, Santiago Vargas, and Anastassia N Alexandrova. Multi-level QTAIM-enriched graph neural networks for resolving properties of transition metal complexes. *Digital Discovery*, 4(11):3378–3388, 2025.

[14] Yanfei Guan, Connor W Coley, Haoyang Wu, Duminda Ranasinghe, Esther Heid, Thomas J Struble, Lagnajit Pattanaik, William H Green, and Klavs F Jensen. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chemical Science*, 12(6): 2198–2208, 2021.

[15] Jiazhen He, Helen Lai, Lakshidaa Saigiridharan, Gian Marco Ghiandoni, Kinga Jenei, Umur Gokalp, Ajsa Nukovic, Ola Engkvist, Jon Paul Janet, and Samuel Genheden. Democratising real-world drug discovery through agentic AI. *Drug Discovery Today*, page 104605, 2026.

[16] M Hoffmann, A Kazimir, T Oestereich, L Kaermer, F Engelberger, J Meiler, and C Lamers. A quantum lens on molecular design: A machine-learned energy function from interacting quantum atoms. *bioRxiv*, pages 2026–03, 2026.

[17] Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow – efficient 3D molecular generation with latent attention and equivariant flow matching. 258:3772–3780, 03–05 May 2025. URL `https://proceedings.mlr.press/v258/irwin25a.html`.

[18] Kjell Jorner, Tore Brinck, Per-Ola Norrby, and David Buttar. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chemical Science*, 12(3):1163–1175, 2021.

[19] Koichiro Kato, Tomohide Masuda, Chiduru Watanabe, Naoki Miyagawa, Hideo Mizouchi, Shumpei Nagase, Kikuko Kamisaka, Kanji Oshima, Satoshi Ono, Hiroshi Ueda, et al. High-precision atomic charge prediction for protein systems using fragment molecular orbital calculation and machine learning. *Journal of Chemical Information and Modeling*, 60(7):3361–3368, 2020.

[20] Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Yixuan Pu, Venkat Kapil, William C Witt, Ioan-Bogdan Magdau, Daniel J Cole, and Gábor Csányi. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025.

[21] Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, et al. The Open Molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

[22] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025.

[23] Shih-Cheng Li, Haoyang Wu, Angiras Menon, Kevin A Spiekermann, Yi-Pei Li, and William H Green. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? *Journal of the American Chemical Society*, 146(33):23103–23120, 2024.

[24] Chérif F Matta and Russell J Boyd. An introduction to the quantum theory of atoms in molecules. *The quantum theory of atoms in molecules*, 1, 2007.

[25] Peter Maxwell, Nicodemo di Pasquale, Salvatore Cardamone, and Paul LA Popelier. The prediction of topologically partitioned intra-atomic and inter-atomic energies by the machine learning method kriging. *Theoretical Chemistry Accounts*, 135(8):195, 2016.

[26] Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[27] Derek P Metcalf, Andy Jiang, Steven A Spronk, Daniel L Cheney, and C David Sherrill. Electron-passing neural networks for atomic charge prediction in systems with arbitrary molecular charge. *Journal of Chemical Information and Modeling*, 61(1):115–122, 2020.

[28] Brandon Meza-González, David I Ramírez-Palma, Pablo Carpio-Martínez, David Vázquez-Cuevas, Karina Martínez-Mayorga, and Fernando Cortés-Guzmán. Quantum topological atomic properties of 44k molecules. *Scientific Data*, 11(1):945, 2024.

[29] Brandon Meza-González, David I. Ramírez-Palma, Pablo Carpio-Martínez, David Vázquez-Cuevas, Karina Martinez-Mayorga, and Fernando Cortés-Guzmán. Aimel-db: Atomic properties for 44k small organic molecules, February 2024. URL `https://doi.org/10.5281/zenodo.11406726`.

[30] Robert G Parr and Weitao Yang. Density functional approach to the frontier-electron theory of chemical reactivity. *Journal of the American Chemical Society*, 106(14):4049–4050, 1984.

[31] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025.

[32] Sergey Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. *Advances in Neural Information Processing Systems*, 36:79469–79501, 2023.

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

[34] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[35] Peter Škrinjar, Jérôme Eberhardt, Gabriel Studer, Gerardo Tauriello, Torsten Schwede, and Janani Durairaj. Evaluating generalization in protein–ligand cofolding methods. *Nature Structural & Molecular Biology*, pages 1–13, 2026.

[36] Thijs Stuyver and Connor W Coley. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *The Journal of Chemical Physics*, 156(8), 2022.

[37] Santiago Vargas, Winston Gee, and Anastassia Alexandrova. High-throughput quantum theory of atoms in molecules (QTAIM) for geometric deep learning of molecular and reaction properties. *Digital Discovery*, 3(5):987–998, 2024.

[38] Max Veit, David M Wilkins, Yang Yang, Robert A DiStasio, and Michele Ceriotti. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *The Journal of chemical physics*, 153(2), 2020.

[39] Jike Wang, Dongsheng Cao, Cunchen Tang, Lei Xu, Qiaojun He, Bo Yang, Xi Chen, Huiyong Sun, and Tingjun Hou. DeepAtomicCharge: A new graph convolutional network-based architecture for accurate prediction of atomic charges. *Briefings in Bioinformatics*, 22(3):bbaa183, 2021.

[40] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

A Technical appendices and supplementary material

A.1 Training and computational details

We report training and compute statistics for all models compared in this work in Table 3. We characterize per-batch compute using GFLOPs (giga floating-point operations, 10^9 FLOPs), measured as the cost of a single forward pass through the model on one training batch. EGFC model trained on A100fat GPU for about 2 days per fold. All other models trained on either A40 or A100 (based on GPU availability) and took from 12 to 36 hours per fold.

Table 3: Training and computational details for the models compared in Section 4. Parameter counts were chosen to keep model sizes approximately matched across architectures, while batch size and epoch count were tuned per model based on convergence behavior. Abbreviations: learning rate (LR); weight decay for AdamW (WD).

Models	# Params.	Batch Size	GFLOPs/Batch	Epochs	LR	WD
EGFC	486 786	128	17.88	1500	1.6e-3	1.2e-4
SFC2	686 081	256	20.30	2000	2.7e-3	1.6e-4
SGFC	487 598	128	8.67	3000	2.7e-3	1.6e-4
SG-8-12	686 049	256	9.25	2000	3.8e-3	1.1e-4
SG-8-5	686 049	2048	17.99	2000	3.8e-3	1.1e-4
EGNN	615 090	512	8.89	3000	3.8e-3	1.1e-4
SGNN	677 089	1024	8.96	3000	5.5e-3	9.1e-4

EGFC, EGNN, SGFC, and SGNN ran in a first round of experiments that used for loss function

$$\mathcal{L} = \frac{1}{N_{at}} \sum_{i=1}^{N_{at}} \frac{1}{4} \left[(\hat{N}_i - N_i)^2 + (\hat{\lambda}_i - \lambda_i)^2 + \frac{\sum_{\alpha=1}^{3} (\hat{\mu}_i^{(\alpha)} - \mu_i^{(\alpha)})^2}{3} + \frac{\sum_{\alpha=1}^{5} (\hat{Q}_i^{(\alpha)} - Q_i^{(\alpha)})^2}{5} \right],$$

where N_{at} is the total number of atoms in the training fold. For the rotationally augmented models, the training was slightly refined, adding per-element weights w_{Z_i} and modifying the quadrupole term to account for the Frobenius weights:

$$\mathcal{L} = \frac{1}{\sum_{i=1}^{N_{at}} w_{Z_i}} \sum_{i=1}^{N_{at}} \frac{w_{Z_i}}{4} \left[(\hat{N}_i - N_i)^2 + (\hat{\lambda}_i - \lambda_i)^2 + |\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|^2 + \|\hat{\mathbf{Q}}_i - \mathbf{Q}_i\|_F^2 \right].$$

The weight of each atom was normalized so that the sum of all atoms of a same element in the fold summed to 1, following $w_{Z_i} = \sqrt{N_{at}/N_{Z_i}}$.

For all models, the target quantities were z-score normalized, using mean and standard deviation for scalar properties and dividing by root mean square for tensor properties.

All points in the comparison stand, given the direct comparison between non-rotationally augmented and equivariant models in Section A.8.

A.2 Conventions

Our convention for representing a traceless tensor in Cartesian coordinates is

$$[T_{xx}, T_{xy}, T_{xz}, T_{yy}, T_{yz}, T_{zz}] \rightarrow [T_{xy}, T_{xz}, T_{yz}, T_{an}, T_{zz}],$$

where the anisotropic component is obtained from $T_{an} = \frac{T_{xx} - T_{yy}}{2}$. The equivariance-preserving Frobenius norm is therefore computed as

$$\|T\| = \sqrt{2(T_{xy}^2 + T_{xz}^2 + T_{yz}^2 + T_{an}^2) + \frac{3}{2} T_{zz}^2}. \tag{3}$$

A.3 Clustering of AIMEl atomic environments

Element-specific HDBSCAN clustering parameters and resulting cluster counts are reported in Table 4, and the distributions of cluster labels across molecules in the dataset is shown in Figure 6.

Table 4: Element-specific details of HDBSCAN clustering. In all cases, we used the excess of mass method and no cluster merging.

	H	**C**	**N**	**O**
Number of atoms	278 943	199 660	30 805	40 894
PCA components	20	35	30	20
Variance retained (%)	99.11	99.01	99.18	99.04
Min. cluster size	850	850	100	100
Min. samples	200	200	30	30
Number of clusters	20	30	50	50
Noise points (%)	8.2	5.95	16.69	16.5

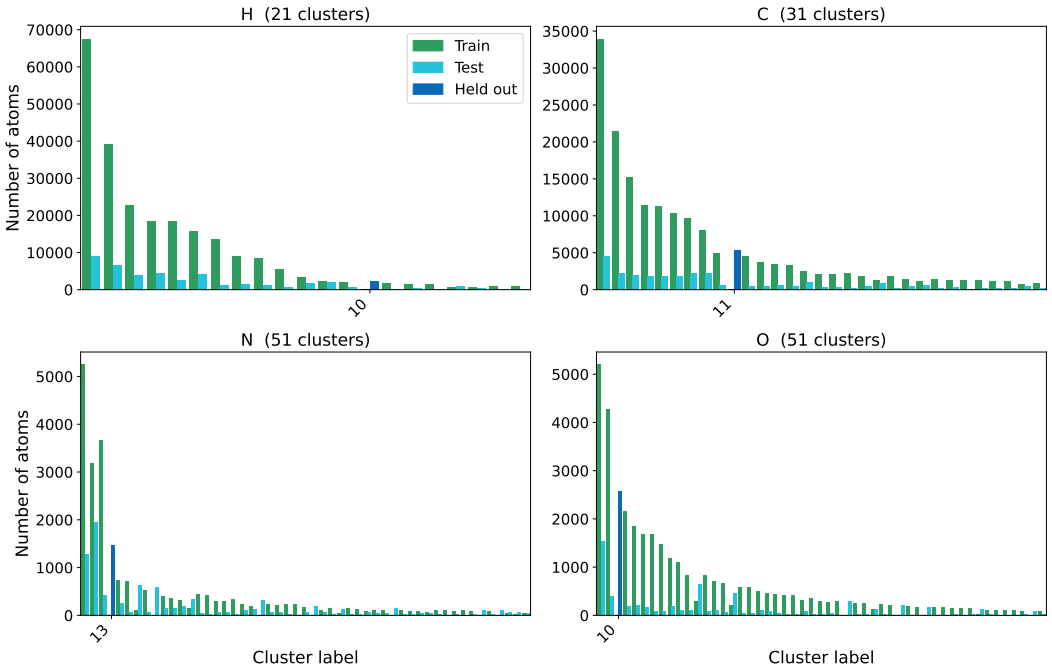

Figure 5: Number of atoms in each atomic environment. Only held out environment labels are included for clarity.

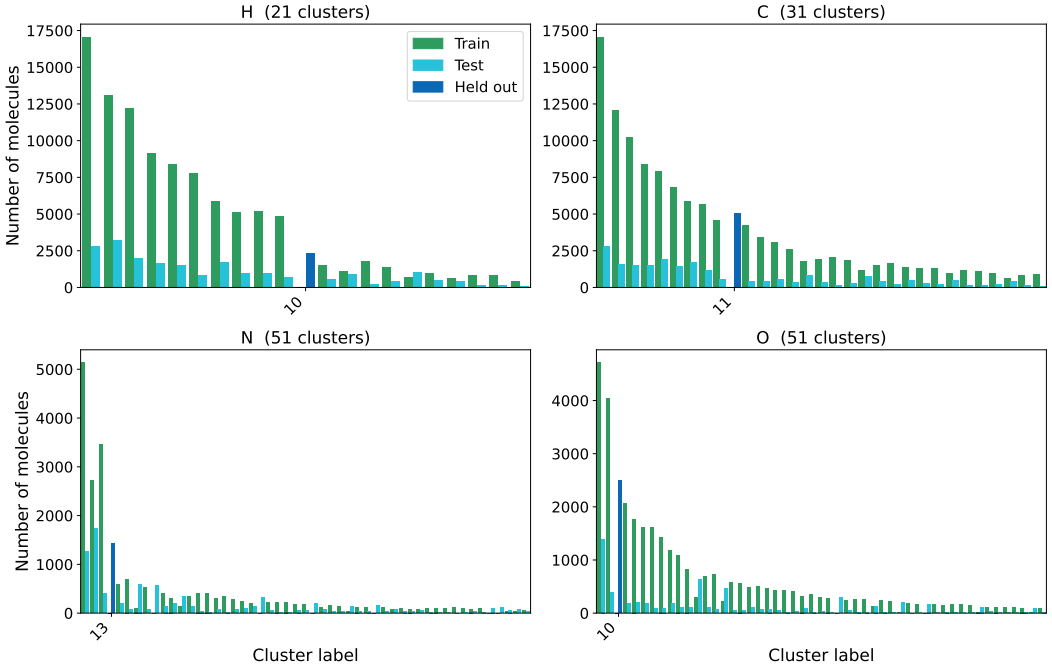

Figure 6: Number of molecules where an atomic environment occurs in the whole AIMEl subset. Only held out environment labels are included for clarity.

A.4 Clustering of QM9 atomic environments and chemotypes

Atomic environments. Element-specific details for the clustering over all atoms in the 128 908 molecules (30 812 in AIMEl and 98 096 in the remainder; 2 333 341 atoms) of the working dataset for molecular experiments are reported in Table 5. Because AIMEl and the remainder were clustered together, they share one set of labels. Seven environments (three of C, one of N and three of O) have no atom in AIMEl. These labels are numbered independently of the 4.5-Bohr labels of the atomic holdout; a shared name such as O_10 is a coincidence. Figures 7–10 map the environments of each element.

Distance to AIMEl. We measure how far an atomic environment lies from a training set in the PCA projection of the 6-Bohr SOAP vectors. The distance of an atom to a training set is its Euclidean distance to the 15th-nearest training atom of the same element. We calibrate it on in-distribution atoms of molecules outside the reference, and say that an atom is 'far from AIMEl' or 'rare' when its distance exceeds their 99th percentile. For QT-Net, the training set is AIMEl, and every AIMEl atom is scored without the atoms of its own molecule, so that 1% of the AIMEl atoms are rare by construction. For the molecular models, the training set is each model's training molecules and the calibration its validation molecules. Measured this way, 3.0, 7.3, 1.7 and 2.7% of the remainder's H, C, N and O atoms are far from AIMEl, and 35.8% of the remainder molecules contain at least one rare atom (14.9% three or more), against 14.7% (0.5%) of the AIMEl molecules by chance. Every remainder molecule in the two chemotypes absent from AIMEl, and in the three that AIMEl holds once or twice, has a rare atom.

Table 5: Element-specific details of the HDBSCAN clustering of the 6-Bohr environments of all molecules of the working dataset. In all cases, we used the excess of mass method and no cluster merging.

	H	**C**	**N**	**O**
Atoms in AIMEl	278 943	199 660	30 805	40 894
Atoms in the remainder	920 627	623 154	98 489	140 769
PCA components	48	76	67	52
Variance retained (%)	99.01	99.01	99.01	99.02
Min. cluster size	2 500	850	300	250
Min. samples	100	50	5	10
Number of clusters	38	85	91	103
Clusters without AIMEl atoms	0	3	1	3
Noise points (%)	22.9	14.1	20.8	19.7

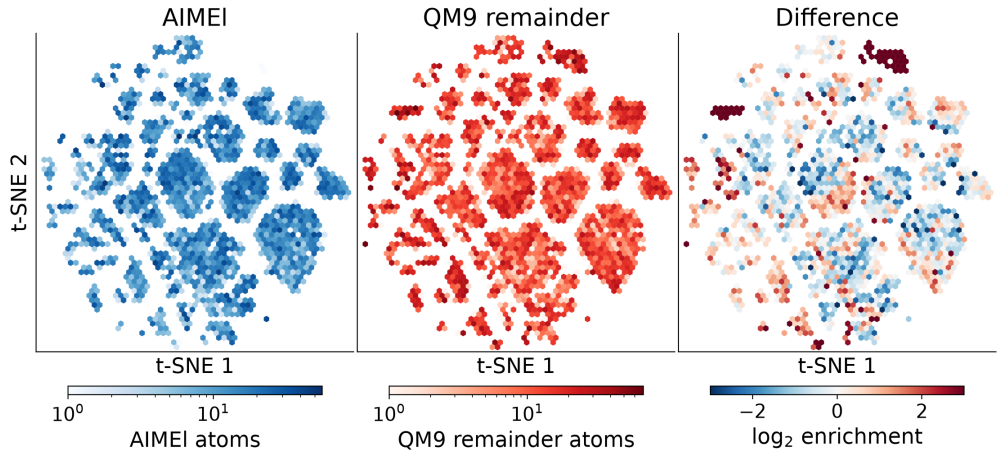

Figure 7: Atomic environments of C: t-SNE (perplexity 50, PCA initialization) of 20 000 AIMEl and 20 000 remainder atoms drawn at random, in the 6-Bohr SOAP space. From left to right: hexagonal-bin counts of the AIMEl atoms, the QM9 remainder atoms, and the $\log_2$ enrichment of the remainder over AIMEl in each bin, i.e. the $\log_2$ ratio of the remainder's share of its sample to AIMEl's share (red: denser in the remainder; blue: denser in AIMEl; bins with fewer than three atoms left blank). t-SNE preserves neighborhoods, not distances or densities across the map.

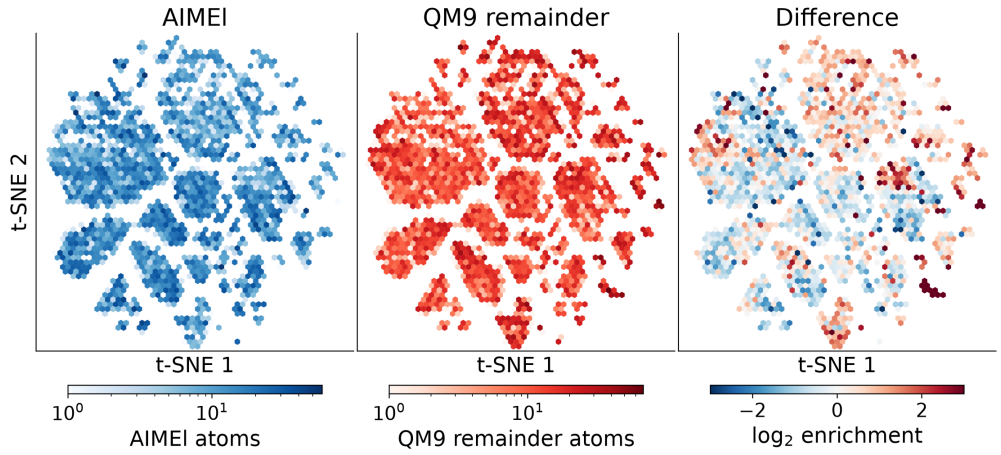

Figure 8: Atomic environments of H, as in Fig. 7.

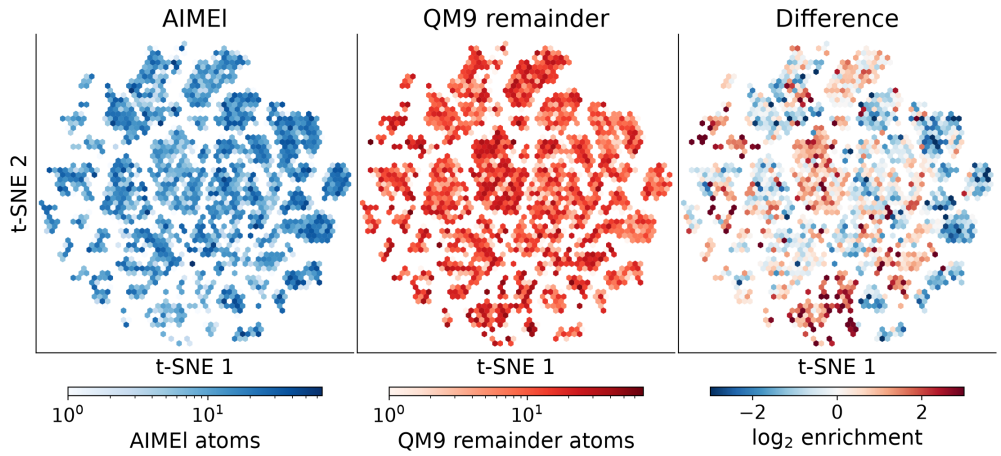

Figure 9: Atomic environments of O, as in Fig. 7.

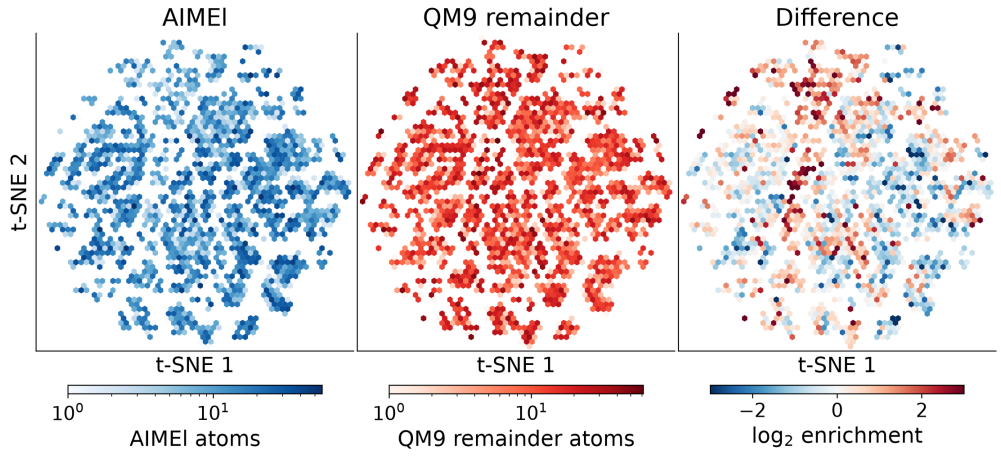

Figure 10: Atomic environments of N, as in Fig. 7.

Environment folds. Each environment fold is made of four labels, one per element, that co-occur in the same molecules. Following the logic of the atomic holdout of Sec. 3.2, withholding a fold removes from training every molecule that carries one of its labels so that environments that co-occur with the withheld ones cannot leak into training. Each fold was grown around a seed environment of N or O, the scarce elements, then adding C and H environments whose atoms co-occur with the seed in at least 95% of the seed atoms' molecules, and then the environment of the other scarce element with the largest share of its molecules inside the resulting set. Seeds were taken rarest first, alternating between N and O, from the environments holding 1–15% of their element's AIMEl atoms. A candidate was rejected when no C or H environment reached 95% co-occurrence, when it would withhold more than 25% of AIMEl, when it reused an environment of an accepted fold, when one of its environments had fewer than 50 atoms, or when more than half of its molecules already belonged to accepted folds. Noise atoms carry no label and were ignored in this design. Table 6 lists the five folds. They were designed on AIMEl, so none of the seven environments found only in the remainder belongs to a fold.

Table 6: The five environment folds. Molecules: AIMEl and remainder molecules that carry at least one of the fold's environments. Atoms: AIMEl atoms in the fold's environment of each element.

		Molecules		AIMEl atoms			
Fold	Environments (seed in bold)	AIMEl	Remainder	H	C	N	O
0	H_29, C_52, **N_65**, O_10	2 653 (8.6%)	8 620	3 211	3 472	252	78
1	H_21, C_72, N_26, **O_51**	3 497 (11.3%)	12 156	3 784	485	67	359
2	H_13, C_34, **N_27**, O_36	2 664 (8.6%)	9 845	1 793	2 122	503	65
3	H_28, C_84, **N_81**, O_16	6 700 (21.7%)	15 263	9 497	9 546	497	85
4	H_12, C_79, **N_24**, O_19	5 065 (16.4%)	15 683	792	5 176	584	202

Cross-validation. The chemotypes of Sec. 3.2 were fitted once on AIMEl and the remainder together and hold 1 251–5 164 molecules. Two of them contain 3 768 terminal alkynes, which have no molecule in AIMEl. Figure 12 maps the chemotype space. The 48 chemotypes present in AIMEl were dealt into five groups of whole chemotypes with balanced molecule counts (largest chemotype first, into the currently smallest group; 6 101–6 288 molecules per group). Fold (r, c) withholds chemotype group c and environment fold $(r + c) \bmod 5$, so that over the five repeats each chemotype group meets each environment fold once (Table 7, Figs. 11 and 13). The test set of each fold has four panels: (A) molecules of the withheld chemotypes without withheld environments; (B) molecules of the withheld chemotypes with withheld environments; (C) molecules of other chemotypes with withheld environments; and (D) 924 molecules (3% of AIMEl) drawn at random from the rest as an in-distribution reference. The remaining molecules form the training pool, cut to 17 855 molecules in every fold (the size of the smallest pool) with a chemotype-stratified subsample. A chemotype-stratified 10% of the pool served for validation and early stopping. The training fractions are chemotype-stratified subsamples of the rest, so that every chemotype labeled as seen is in training at every fraction. Test and validation sets are the same at every fraction and for both variants. On the QM9 remainder, each molecule falls into a panel depending on the training set of the model evaluating it.

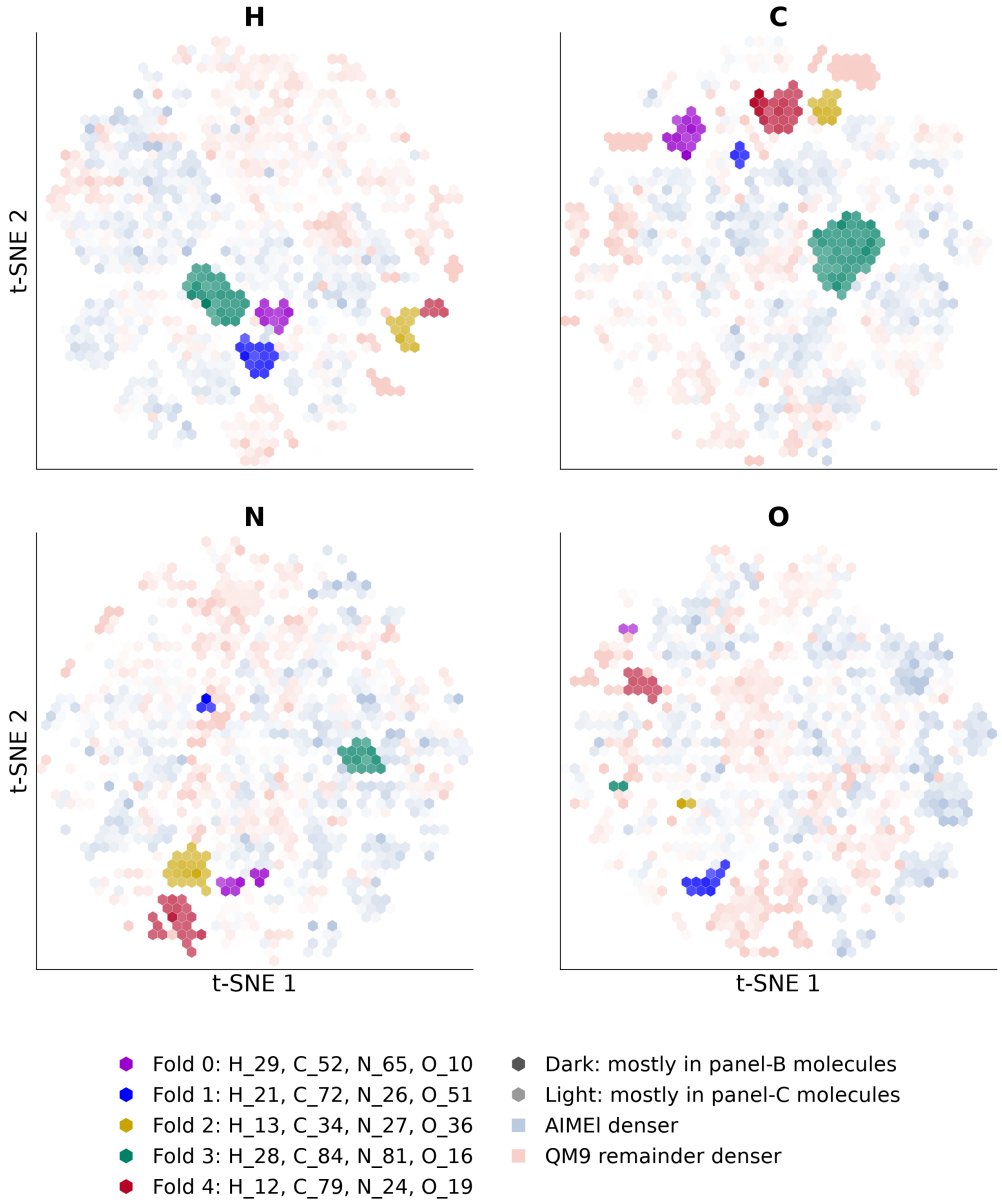

Figure 11: Environments withheld by the five folds of the first repeat on the t-SNE of each element (Figs. 7–10). Each fold withholds one environment per element. Hexagonal bins of the withheld atoms are colored by fold: dark where the atoms are mostly in panel-B molecules (chemotype also withheld), light where mostly in panel-C molecules (chemotype seen).

Table 7: The 25 folds of the molecular cross-validation: the withheld chemotype group and environment fold (identified by its seed environment, Table 6), and the size of each test panel. Every fold has a training pool of 17 855 molecules and a validation set of 1 785.

Fold	Repeat	Chemotype group	Environment fold (seed)	Withheld chemotypes	A	B	C	D	Test
0	0	0	0 (N_65)	9	5 256	1 032	1 621	924	8 833
1	0	1	1 (O_51)	8	4 870	1 287	2 210	924	9 291
2	0	2	2 (N_27)	9	5 590	516	2 148	924	9 178
3	0	3	3 (N_81)	12	4 306	1 795	4 905	924	11 930
4	0	4	4 (N_24)	10	4 443	1 717	3 348	924	10 432
5	1	0	1 (O_51)	9	5 957	331	3 166	924	10 378
6	1	1	2 (N_27)	8	5 927	230	2 434	924	9 515
7	1	2	3 (N_81)	9	4 876	1 230	5 470	924	12 500
8	1	3	4 (N_24)	12	5 616	485	4 580	924	11 605
9	1	4	0 (N_65)	10	5 790	370	2 283	924	9 367
10	2	0	2 (N_27)	9	5 136	1 152	1 512	924	8 724
11	2	1	3 (N_81)	8	5 192	965	5 735	924	12 816
12	2	2	4 (N_24)	9	5 312	794	4 271	924	11 301
13	2	3	0 (N_65)	12	6 014	87	2 566	924	9 591
14	2	4	1 (O_51)	10	5 581	579	2 918	924	10 002
15	3	0	3 (N_81)	9	5 333	955	5 745	924	12 957
16	3	1	4 (N_24)	8	5 743	414	4 651	924	11 732
17	3	2	0 (N_65)	9	5 332	774	1 879	924	8 909
18	3	3	1 (O_51)	12	6 081	20	3 477	924	10 502
19	3	4	2 (N_27)	10	5 521	639	2 025	924	9 109
20	4	0	4 (N_24)	9	4 633	1 655	3 410	924	10 622
21	4	1	0 (N_65)	8	5 741	416	2 237	924	9 318
22	4	2	1 (O_51)	9	5 145	961	2 536	924	9 566
23	4	3	2 (N_27)	12	5 974	127	2 537	924	9 562
24	4	4	3 (N_81)	10	4 900	1 260	5 440	924	12 524

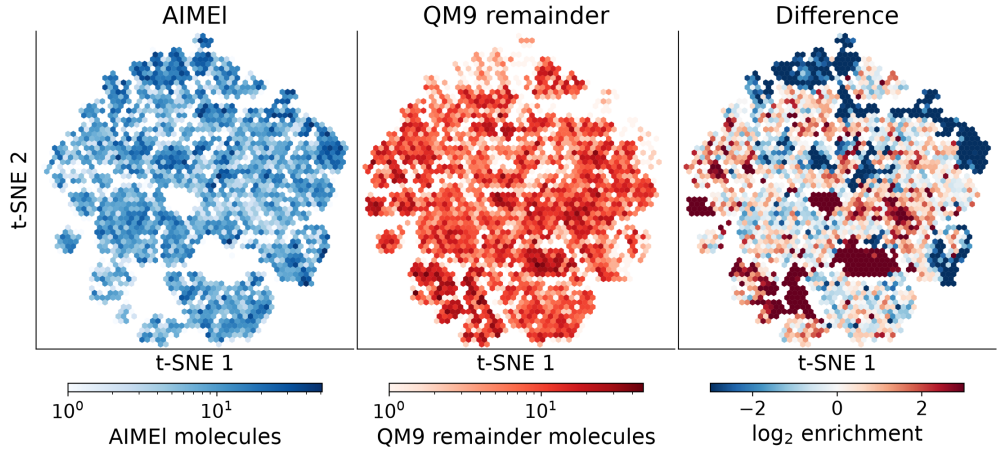

Figure 12: Chemotype space. t-SNE of 20 000 AIMEl and 20 000 remainder molecules drawn at random, in the space of the chemotype clustering (Morgan fingerprints, L2-normalized, 50 principal components). Panels as in Fig. 7, for molecules.

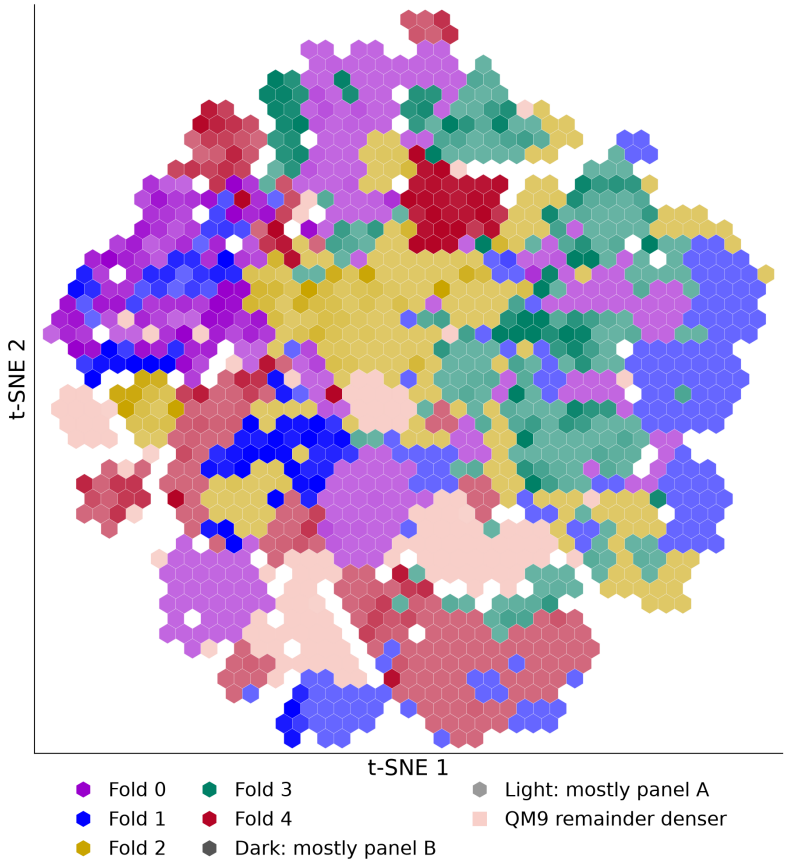

Figure 13: Chemotype groups withheld by the five folds of the first repeat, on the chemotype t-SNE (Fig. 12). Each hexagonal bin of withheld molecules is colored by the fold with most molecules in it: dark where they mostly carry the fold's withheld environments (panel B), light where they do not (panel A).

A.5 Statistical validity of the comparison of atomic models

The statistical validity of the model comparison based on a fixed holdout design was assessed through three diagnostics computed exclusively over atoms carrying the held-out cluster labels H_10, C_11, N_13, and O_10 — enforcing the condition that no metric is diluted by environments seen during training (see text and Tabs. 8 and 9). Because the 5 repeat means reflect only training-sample variability (not holdout-selection variability), Tukey's honest significant difference (HSD) error term is tighter than it would be under a random-holdout design. The resulting model rankings are statistically reliable for the defined OOD environments {H_10, C_11, N_13, O_10} and stable across training-set perturbations, as confirmed by the near-zero ICC. However, the confidence intervals do not capture uncertainty about the choice of holdout, so rankings should be interpreted as conditional on this specific evaluation set and not generalized to arbitrary OOD regimes.

Fold independence (ICC). The intraclass correlation $\mathrm{ICC}(1,1)$ across the five inner folds within each repeat measures whether training-set composition variation causes correlated test performance within a repeat. All models returned near-zero or mildly negative ICC values when averaging over atoms and properties, indicating that folds contribute essentially independent information despite the partial overlap in training molecules. Negative values — observed in several strata of the SI table (Tab. 10) — arise naturally when training-set variability is dominated by which environments are *excluded* rather than included, and do not indicate a violation. The effective sample size $n_{\mathrm{eff}} \in [16.5, 25.0]$ across models reflects this near-independence and confirms that the standard repeated-CV estimator remains close to unbiased. Sphericity corrections (Mauchly / Greenhouse–Geisser) are therefore not needed.

Normality and homoscedasticity (SW and Levene). Shapiro–Wilk normality tests on the five repeat-level means passed for 81.2–100% when averaging over atoms and properties, only descending below 75% when looking at per-atom type scores. Levene homoscedasticity tests passed for 93.8% when scores were averaged over property and atom label, only descending to 75% for certain atom types. Together, these diagnostics confirm that the RM-ANOVA and Tukey HSD assumptions are met in practice for CCC.

Table 8: ANOVA validity diagnostics for averaged CCC metrics of rotationally augmented and equivariant models. All quantities computed exclusively over atoms carrying held-out cluster label. **ICC**$(1,1)$: intraclass correlation across the five inner folds within each repeat; ideal range $[-0.10,\ 0.10]$ (near-independent folds, $n_{\mathrm{eff}} \approx 25$); orange if $|\mathrm{ICC}| > 0.20$, red if > 0.40. $n_{\mathbf{eff}}$: effective sample size $= 25\,/\,(1 + (k-1)\cdot\max(0, \mathrm{ICC}))$; orange if $<80\%$ of nominal (<20), red if $<60\%$ (<15). **SW**: Shapiro–Wilk pass rate (%) on five repeat-means; orange if $<80\%$, red if $<60\%$. **Lev.**: Levene homoscedasticity pass rate (%) across models; same thresholds.

Diagnostic	**EGNN**	**SG-8-5**	**SG-8-12**	**SFC2**	**EGFC**
$\mathrm{ICC}(1,1)$	0.008 ± 0.097	0.043 ± 0.137	-0.006 ± 0.126	0.008 ± 0.140	-0.050 ± 0.138
n_{eff}	22.3 ± 3.7	20.2 ± 4.7	22.0 ± 4.4	21.5 ± 4.5	22.7 ± 4.0
SW	100.0	87.5	81.2	100.0	100.0
Lev.	93.8	93.8	93.8	93.8	93.8

Table 9: ANOVA validity diagnostics for averaged CCC metrics of non-augmented models. All quantities computed exclusively over atoms carrying held-out cluster label. **ICC**$(1,1)$: intraclass correlation across the five inner folds within each repeat; ideal range $[-0.10,\ 0.10]$ (near-independent folds, $n_{\mathrm{eff}} \approx 25$); orange if $|\mathrm{ICC}| > 0.20$, red if > 0.40. $n_{\mathbf{eff}}$: effective sample size $= 25\,/\,(1 + (k-1)\cdot\max(0, \mathrm{ICC}))$; orange if $<80\%$ of nominal (<20), red if $<60\%$ (<15). **SW**: Shapiro–Wilk pass rate (%) on five repeat-means; orange if $<80\%$, red if $<60\%$. **Lev.**: Levene homoscedasticity pass rate (%) across models; same thresholds.

Diagnostic	**SGNN**	**SGFC**
$\mathrm{ICC}(1,1)$	-0.035 ± 0.129	-0.116 ± 0.064
n_{eff}	18.0 ± 3.6	24.8 ± 0.7
SW	93.8	93.8
Lev.	100.0	100.0

Table 10: Full ANOVA validity diagnostics of the models in the main discussion per held-out label and property, CCC metrics. All quantities computed exclusively over atoms carrying held-out cluster label. **ICC**$(1,1)$: intraclass correlation across the five inner folds within each repeat; ideal range $[-0.10,\ 0.10]$ (near-independent folds, $n_{\text{eff}} \approx 25$); orange if $|\text{ICC}| > 0.20$, red if > 0.40. $n_{\textbf{eff}}$: effective sample size $= 25\ /\ (1 + (k-1) \cdot \max(0, \text{ICC}))$; orange if $<80\%$ of nominal (<20), red if $<60\%$ (<15). **SW**: Shapiro–Wilk pass rate (%) on five repeat-means; orange if $<80\%$, red if $<60\%$. **Lev.**: Levene homoscedasticity pass rate (%) across models; same thresholds.

		H_10				C_11				N_13				O_10			
Model	Property	ICC$(1,1)$	n_{eff}	SW	Lev.	ICC$(1,1)$	n_{eff}	SW	Lev.	ICC$(1,1)$	n_{eff}	SW	Lev.	ICC$(1,1)$	n_{eff}	SW	Lev.
EGNN	N	-0.037	25.0	100.0	100.0	-0.025	25.0	100.0	100.0	-0.100	25.0	100.0	100.0	0.153	15.5	100.0	100.0
	λ	0.076	19.2	100.0	100.0	0.020	23.1	100.0	100.0	-0.205	25.0	100.0	100.0	0.209	13.6	100.0	0.0
	$\lvert\boldsymbol{\mu}\rvert$	-0.019	25.0	100.0	100.0	-0.016	25.0	100.0	100.0	-0.036	25.0	100.0	100.0	0.111	17.3	100.0	100.0
	$\lvert\mathbf{Q}\rvert$	0.001	24.9	100.0	100.0	0.042	21.4	100.0	100.0	-0.081	25.0	100.0	100.0	0.035	21.9	100.0	100.0
SG-8-5	N	-0.071	25.0	100.0	100.0	0.070	19.5	100.0	100.0	0.082	18.8	100.0	100.0	-0.017	25.0	100.0	100.0
	λ	-0.149	25.0	100.0	100.0	0.092	18.3	100.0	100.0	-0.132	25.0	100.0	100.0	0.129	16.5	0.0	0.0
	$\lvert\boldsymbol{\mu}\rvert$	-0.065	25.0	100.0	100.0	0.102	17.7	100.0	100.0	0.019	23.2	100.0	100.0	0.343	10.5	100.0	100.0
	$\lvert\mathbf{Q}\rvert$	-0.177	25.0	100.0	100.0	0.061	20.1	0.0	100.0	0.201	13.8	100.0	100.0	0.199	13.9	100.0	100.0
SG-8-12	N	0.049	20.9	100.0	100.0	-0.011	25.0	100.0	100.0	-0.143	25.0	100.0	100.0	-0.165	25.0	100.0	100.0
	λ	0.213	13.5	100.0	100.0	-0.100	25.0	100.0	100.0	-0.066	25.0	100.0	100.0	-0.196	25.0	0.0	0.0
	$\lvert\boldsymbol{\mu}\rvert$	0.146	15.8	100.0	100.0	0.045	21.2	100.0	100.0	0.079	19.0	100.0	100.0	0.255	12.4	100.0	100.0
	$\lvert\mathbf{Q}\rvert$	-0.073	25.0	100.0	100.0	0.002	24.9	0.0	100.0	-0.055	25.0	100.0	100.0	-0.068	25.0	0.0	100.0
SFC2	N	0.011	24.0	100.0	100.0	-0.098	25.0	100.0	100.0	0.143	15.9	100.0	100.0	-0.042	25.0	100.0	100.0
	λ	0.072	19.4	100.0	100.0	0.001	24.9	100.0	100.0	0.348	10.5	100.0	100.0	0.025	22.7	100.0	0.0
	$\lvert\boldsymbol{\mu}\rvert$	0.157	15.3	100.0	100.0	-0.160	25.0	100.0	100.0	0.069	19.6	100.0	100.0	0.124	16.7	100.0	100.0
	$\lvert\mathbf{Q}\rvert$	-0.193	25.0	100.0	100.0	-0.109	25.0	100.0	100.0	-0.043	25.0	100.0	100.0	-0.181	25.0	100.0	100.0
EGFC	N	-0.104	25.0	100.0	100.0	0.007	24.3	100.0	100.0	-0.127	25.0	100.0	100.0	0.050	20.8	100.0	100.0
	λ	-0.097	25.0	100.0	100.0	0.149	15.7	100.0	100.0	-0.208	25.0	100.0	100.0	-0.090	25.0	100.0	0.0
	$\lvert\boldsymbol{\mu}\rvert$	-0.178	25.0	100.0	100.0	-0.178	25.0	100.0	100.0	0.093	18.2	100.0	100.0	0.017	23.4	100.0	100.0
	$\lvert\mathbf{Q}\rvert$	-0.195	25.0	100.0	100.0	0.299	11.4	100.0	100.0	-0.154	25.0	100.0	100.0	-0.086	25.0	100.0	100.0

A.6 Analysis of environment leakage due to model cutoff

The out-of-distribution holdout — the environment clusters {H_10, C_11, N_13, O_10} — was defined by clustering SOAP descriptors computed at a cutoff of 4.5 Bohr. The atomic models, however, build their message-passing graph at a larger cutoff (5.25 Bohr for EGNN, 8 Bohr for QT-Net), and it is the cutoff that sets the graph topology, i.e. which atoms a central atom perceives. A held-out environment that is genuinely distinct from training at 4.5 Bohr may cease to be distinct once the neighborhood is widened, because the larger sphere pulls in context that makes it resemble a training environment. If that happens, the held-out set is less out-of-distribution at the model's radius than its 4.5-Bohr construction implies — it leaks. This appendix quantifies that effect with a single, clustering-free, calibrated measure of novelty, evaluated as a function of the cutoff r_c.

For each cutoff r_c and each element we recompute every atom's SOAP descriptor and project it onto the principal components of the full dataset, keeping enough components to retain the same per-element variance the original 4.5-Bohr projection retained (O 99.04%, H 99.11%, C 99.01%, N 99.18%); at 4.5 Bohr we reuse the exact per-element component counts of the original clustering (H 20, C 35, N 30, O 20). Holding the information budget fixed in this way ensures the only quantity varying along the sweep is the radius, not how much of the descriptor is retained.

We then rescale each principal component by the standard deviation of the *training* atoms along it, so that every direction contributes to distance in proportion to how much the training set actually varies along it (equivalently, Euclidean distance in the rescaled space is approximately the Mahalanobis distance with respect to the training covariance). For each held-out (query) atom we compute the distance d_{15} to its 15th nearest training neighbor. Since d_{15} is in arbitrary units that differ by element and radius, we split the training atoms into two disjoint groups: a reference set R ($\leq$ 150 000 atoms) and a calibration set C ($\leq$100 000 atoms). The distances from the calibration atoms to R give a reference distribution for atoms we know are training-typical, against which each held-out atom's distance to the same R is ranked. A held-out atom's novelty percentile is then the percentage of calibration atoms with smaller d_{15}: a value of 50 means the atom is as isolated as the median training atom (it sits inside training's typical density and is effectively seen), whereas a value near 100 means it lies in a tail that training barely covers (genuinely OOD).

In Figure 14 we show the median novelty percentile across the three radii. The carbon and hydrogen holdouts are essentially unaffected: their median novelty stays above 95 at every radius, they remain genuinely out-of-distribution at the models' perception radius. The nitrogen and oxygen holdouts behave very differently. As the cutoff widens from 4.5 to 8 Bohr, the median novelty of N_13 falls from 86 to 35 and that of O_10 from 95 to 58. To test whether this leakage is inflating the scores, we re-stratify the held-out N_13 and O_10 atoms at the 8-Bohr cutoff into a leaked subset (novelty percentile < 50) and a novel subset, and compute the mean CCC (see Tab. 11). Crucially, the leaked atoms are predicted less accurately than those that remain novel. The reported OOD metrics are therefore not inflated by the cutoff mismatch, and the model-comparison conclusions — already framed as conditional on the chosen holdout — are unaffected.

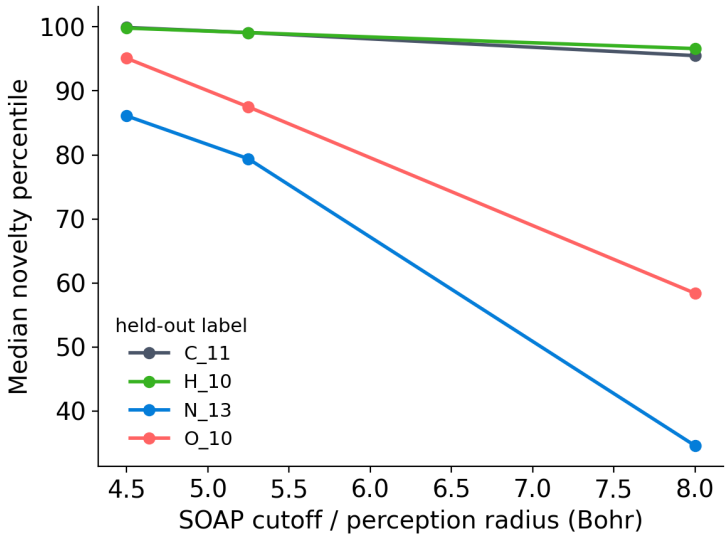

Figure 14: Median novelty percentile of the held-out environments at 4.5, 5.25 and 8 Bohr.

Table 11: Mean CCC of each model for each property of the leaked and novel atoms.

		N_13		O_10	
Model	Property	Leaked	Novel	Leaked	Novel
EGNN	N	0.590	0.806	0.852	0.906
	λ	0.595	0.810	0.830	0.878
	$\lvert\boldsymbol{\mu}\rvert$	0.526	0.569	0.921	0.927
	$\lvert\mathbf{Q}\rvert$	0.380	0.603	0.856	0.854
SG-8-5	N	0.839	0.919	0.941	0.958
	λ	0.832	0.904	0.949	0.954
	$\lvert\boldsymbol{\mu}\rvert$	0.683	0.776	0.928	0.946
	$\lvert\mathbf{Q}\rvert$	0.654	0.852	0.869	0.893
SG-8-12	N	0.884	0.961	0.949	0.967
	λ	0.897	0.955	0.961	0.971
	$\lvert\boldsymbol{\mu}\rvert$	0.821	0.888	0.954	0.971
	$\lvert\mathbf{Q}\rvert$	0.675	0.889	0.902	0.924
SFC2	N	0.852	0.955	0.958	0.974
	λ	0.879	0.946	0.964	0.975
	$\lvert\boldsymbol{\mu}\rvert$	0.818	0.898	0.951	0.968
	$\lvert\mathbf{Q}\rvert$	0.767	0.921	0.889	0.907
EGFC	N	0.654	0.853	0.917	0.941
	λ	0.715	0.859	0.925	0.938
	$\lvert\boldsymbol{\mu}\rvert$	0.634	0.724	0.846	0.880
	$\lvert\mathbf{Q}\rvert$	0.457	0.650	0.683	0.727

A.7 Additional plots and metrics for the discussion on predicting QTA properties

We include here numerical values of CCC (Tab. 12) and Tukey forest plots for the equivariant and rotationally augmented models (Fig. 15). Per-element and per-property differences must be interpreted keeping in mind that the statistical validity gets compromised at this stratum (see Tab. 10). It is worth noticing that, when accounting for all atoms in the test set, all models achieve CCC > 0.9 in all properties and atoms (Fig. 16).

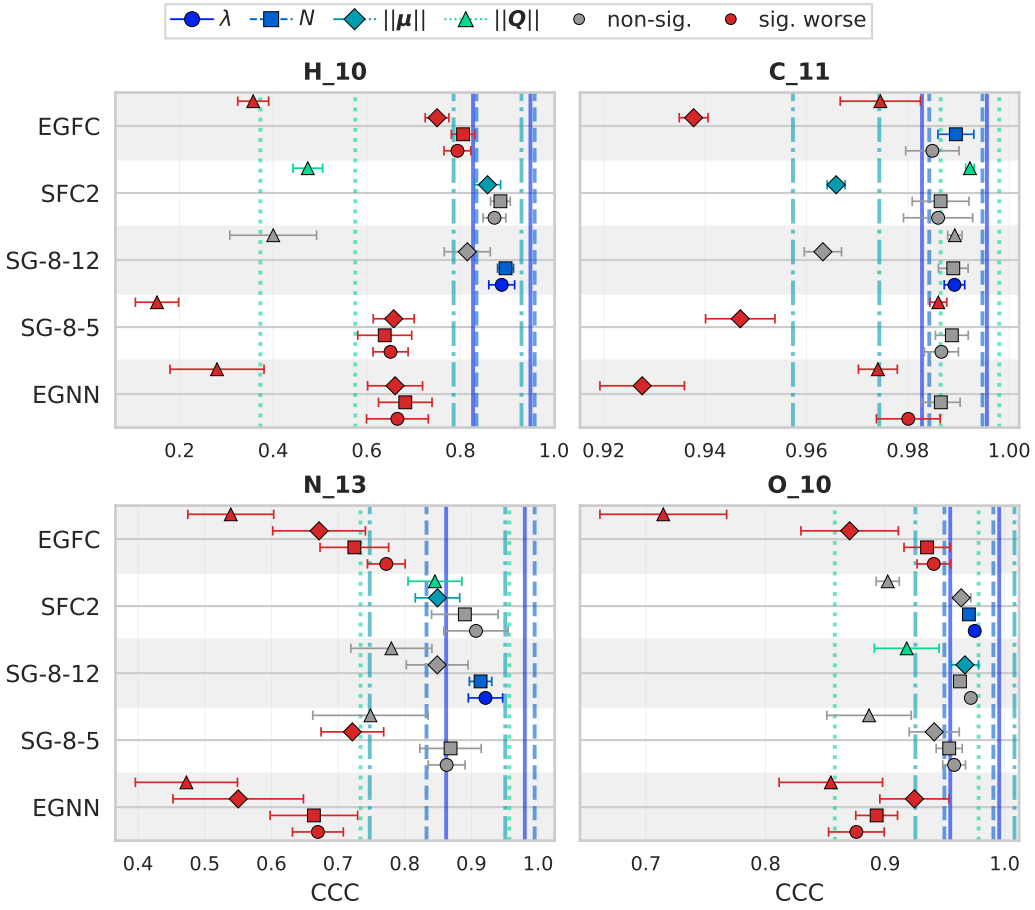

Figure 15: Tukey forest plots for CCC scores of each model, for each atom and property in the holdout set. The best performing model for each (atom, property) shows a colored symbol surrounded by the minimum significant difference.

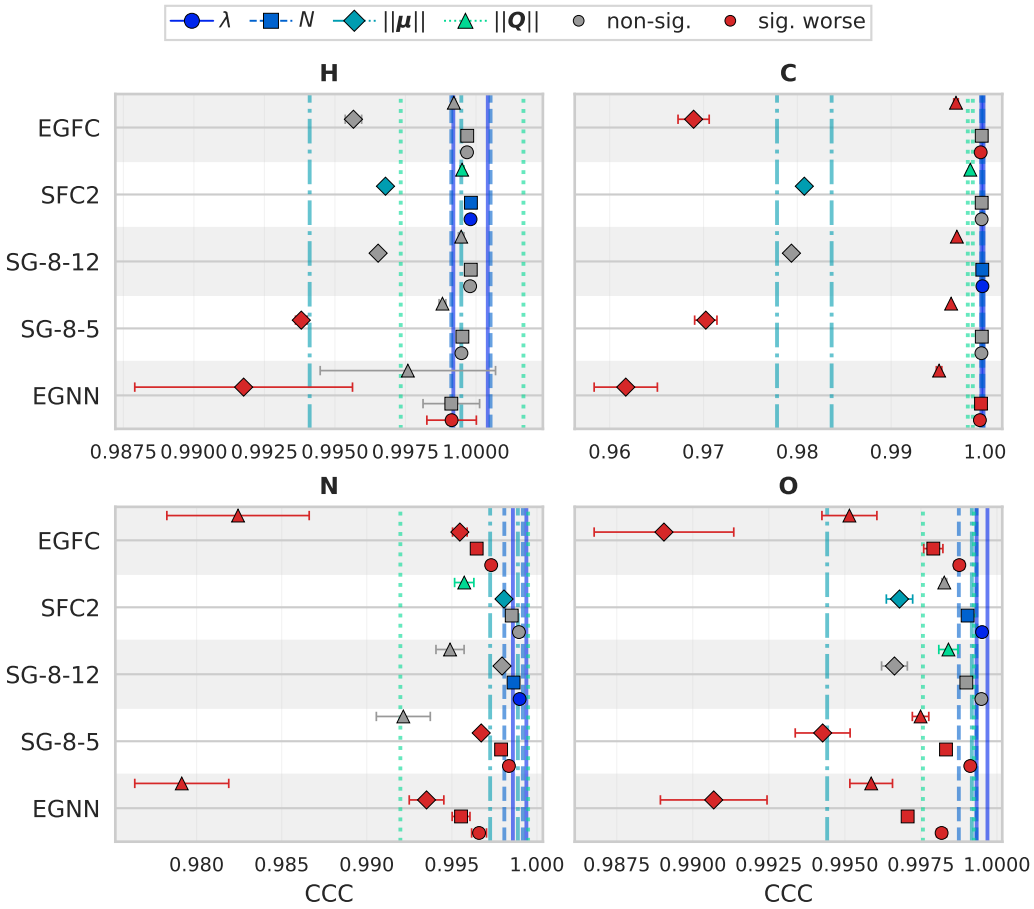

Figure 16: Tukey forest plots for CCC scores of each model, for each atom and property for all of the test set. The best performing model for each (atom, property) shows a colored symbol surrounded by the minimum significant difference.

Table 12: Model comparison on CCC (mean $\pm$ 95 % CI over $n = 5$ CV repeats). Cell color: 1st 2nd 3rd per row.

Cluster	Property	**EGNN**	**SG-8-5**	**SG-8-12**	**SFC2**	**EGFC**
H_10	N	0.682 ± 0.057	0.638 ± 0.058	0.897 ± 0.017	0.886 ± 0.021	0.806 ± 0.025
	λ	0.665 ± 0.066	0.651 ± 0.037	0.888 ± 0.028	0.873 ± 0.024	0.794 ± 0.029
	$\lvert\boldsymbol{\mu}\rvert$	0.661 ± 0.059	0.658 ± 0.044	0.815 ± 0.049	0.858 ± 0.028	0.750 ± 0.025
	$\lvert\mathbf{Q}\rvert$	0.280 ± 0.100	0.152 ± 0.046	0.400 ± 0.093	0.474 ± 0.032	0.357 ± 0.033
C_11	N	0.987 ± 0.004	0.989 ± 0.003	0.989 ± 0.003	0.986 ± 0.006	0.989 ± 0.004
	λ	0.980 ± 0.006	0.987 ± 0.003	0.989 ± 0.002	0.986 ± 0.007	0.985 ± 0.005
	$\lvert\boldsymbol{\mu}\rvert$	0.928 ± 0.008	0.947 ± 0.007	0.963 ± 0.004	0.966 ± 0.002	0.938 ± 0.003
	$\lvert\mathbf{Q}\rvert$	0.974 ± 0.004	0.986 ± 0.002	0.989 ± 0.001	0.992 ± 0.001	0.975 ± 0.008
N_13	N	0.664 ± 0.066	0.869 ± 0.046	0.914 ± 0.017	0.890 ± 0.050	0.724 ± 0.051
	λ	0.670 ± 0.038	0.863 ± 0.028	0.921 ± 0.026	0.907 ± 0.048	0.772 ± 0.028
	$\lvert\boldsymbol{\mu}\rvert$	0.550 ± 0.098	0.721 ± 0.047	0.849 ± 0.046	0.849 ± 0.033	0.671 ± 0.069
	$\lvert\mathbf{Q}\rvert$	0.473 ± 0.077	0.749 ± 0.086	0.780 ± 0.061	0.845 ± 0.040	0.539 ± 0.064
O_10	N	0.893 ± 0.017	0.954 ± 0.011	0.963 ± 0.003	0.970 ± 0.005	0.935 ± 0.019
	λ	0.876 ± 0.023	0.958 ± 0.010	0.972 ± 0.002	0.975 ± 0.004	0.941 ± 0.014
	$\lvert\boldsymbol{\mu}\rvert$	0.925 ± 0.029	0.941 ± 0.021	0.967 ± 0.011	0.964 ± 0.008	0.870 ± 0.041
	$\lvert\mathbf{Q}\rvert$	0.855 ± 0.043	0.887 ± 0.035	0.918 ± 0.027	0.902 ± 0.010	0.715 ± 0.053

A.8 Additional discussion on the effect of layer depth

Here, we provide arguments in favor that the greater depth of the scalar models is not being invested in passing chemical information, but rather in polishing the purely geometrical information that equivariant models understand by design. To confirm this, we trained two non-augmented, scalar GNNs matching the connectivity of EGNN and EGFC. We refer to these two non-augmented models as SGNN and SGFC. In Figure 17, we include radial distributions for reference.

The results in Tab. 13 show that the difference between EGNN and SGNN is not significant. Therefore, at equal connectivity and without data augmentation, the greater depth is being invested in geometric learning. The significant performance difference between SGFC and EGFC suggests that the geometric understanding of non-equivariant models can highly benefit from a denser connectivity.

Table 13: Tukey p-value matrix of CCC scores averaged over atom types and properties, comparing non-data augmented, non-equivariant models with equivariant models. A value below 0.05 indicates a statistically significant difference in performance. The best performing models are determined based on CCC mean value and statistical significance of the difference: 1st 2nd 3rd.

	Tukey p-values				
Models	EGFC	SGFC	EGNN	SGNN	CCC ± 95% CI
EGFC	1.0000	0.0010	0.0010	0.0010	0.798±0.006
SGFC	0.0010	1.0000	0.0010	0.0010	0.844±0.011
EGNN	0.0010	0.0010	1.0000	**0.6128**	0.754±0.028
SGNN	0.0010	0.0010	**0.6128**	1.0000	0.744±0.018

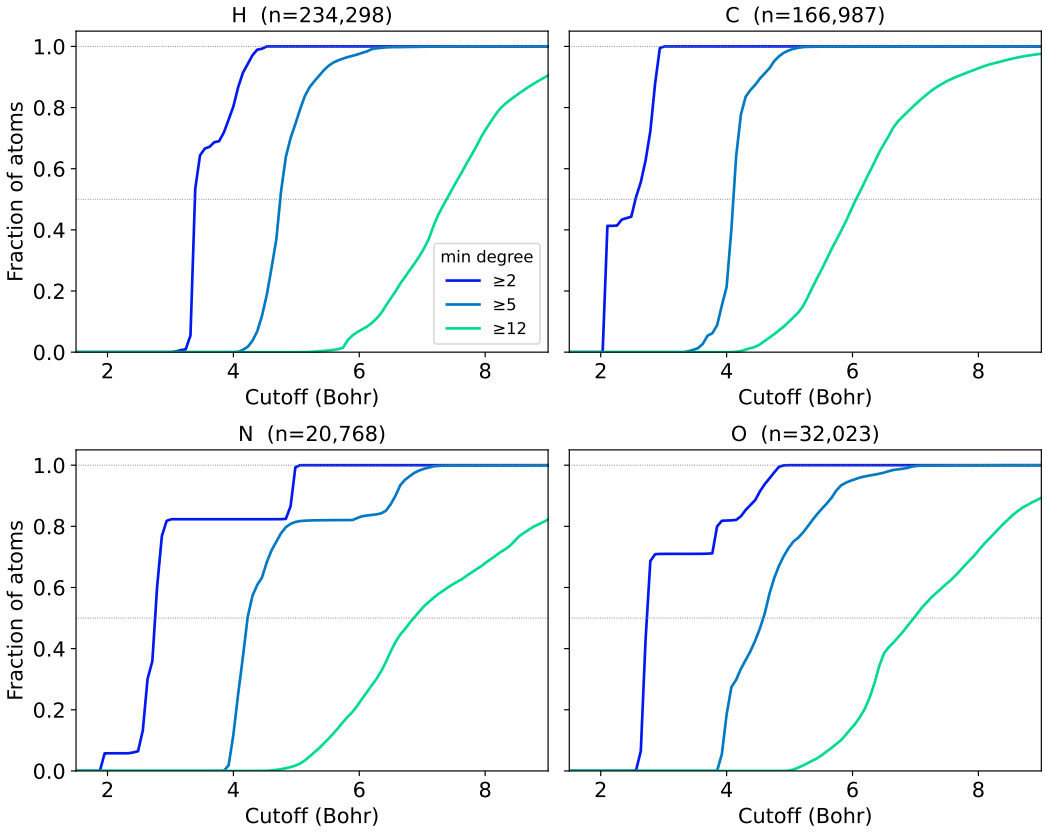

Figure 17: Distributions of fractions of atoms in training by minimum number of neighbors at each cutoff.

A.9 Additional plots and metrics for the discussion on molecular predictions

This section complements Sec. 5.2 with the sources of error of the reconstructed dipoles, the comparison of informed and blind models, the method behind the same-molecule comparison of Fig. 4, and the error of the molecular models against the distance to their training sets.

Reconstructed dipoles. Of the remainder's molecules, 15% have three or more atoms far from AIMEl. They carry 73% of the absolute error of the reconstructed dipoles (Fig. 3), with 99.3% of the molecules that are off by more than 1 D containing at least one rare atom. Two chemical groups explain most of this error (Table 14, Fig. 18). Most of the contribution comes from alkynes: AIMEl contains only 14 molecules with a C≡C bond, whereas the remainder contains 17 468. These carry 77% of the absolute and squared error and represent 95% of the molecules whose predictions are off by more than 1 D. Zwitterions also contribute significantly, with 83% of them containing a rare atom. They represent 0.54% of the remainder but carry 86 of the 100 largest errors and 20% of the squared error. The only large errors without rare atoms are four 4-alkyl-2-oxabicyclo[1.1.1]pentan-4-ols, off by 6.0–7.5 D: AIMEl contains their strained cage in 527 molecules, but their substitution pattern just once. QT-Net's receptive field spans a whole QM9 molecule, so what is new to it is the molecule rather than a 6-Bohr environment. The groups also differ in the sign of their errors: among the molecules off by more than 1 D, QT-Net underestimates the dipole of every zwitterion, and overestimates that of 77% of the alkynes and of all the oxabicyclopentanols.

Table 14: Chemical groups behind the largest errors of the reconstructed dipoles on the remainder. The groups do not overlap: zwitterions with a C≡C bond count as zwitterions. Share: of the total absolute and squared error. Off: molecules with an absolute error above 1 D. R^2 without: of the remainder without the group (0.931 with all molecules).

	Molecules			Share (%)			
Group	AIMEl	Remainder	MAE (D)	Abs.	Sq.	Off	R^2 without
Alkynes	14	17 468	0.674	77.1	76.9	4 888	0.981
Protonated zwitterions	13	195	2.468	3.2	16.1	125	0.934
Other zwitterions	37	332	0.904	2.0	4.0	103	0.929
Alkyl-oxabicyclopentanols	1	10	3.053	0.2	1.1	5	0.931

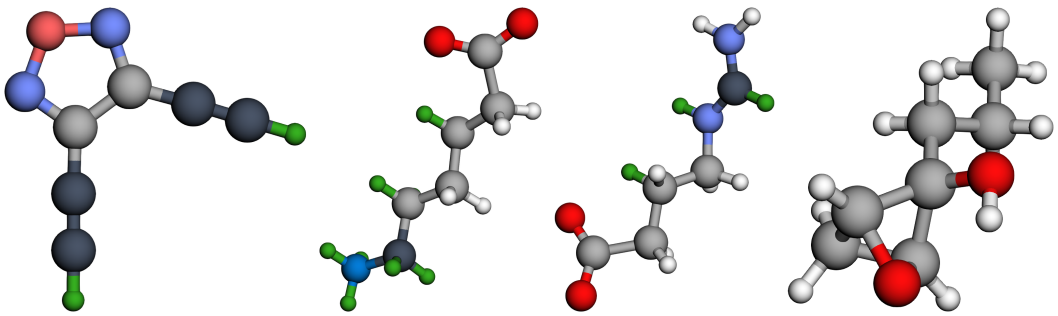

Figure 18: The molecule with the largest error of the reconstructed dipole in each group of Table 14. From left to right: an alkyne (`C#CC1=NON=C1C#C`; QM9 dipole 3.8 D vs reconstructed 0.6 D, whereas most alkynes off by more than 1 D are overestimated), a protonated zwitterion (`[NH3+]CCCCCC([O-])=O`; 29.6 vs 15.1 D), another zwitterion (`NC=[NH+]CCCC([O-])=O`; 21.9 vs 14.5 D) and an alkyl-oxabicyclopentanol (`CCCC1(O)C2CC1O2`; 1.7 vs 9.3 D). Atoms within AIMEl's 99th percentile of the distance of Appendix A.4 are drawn in the colors of Fig. 1 for atoms in seen environments (H , C , N , O), atoms beyond it in its colors for held-out environments (H , C , N , O).

Same molecules and distance to the training set. The panels of a fold hold different molecules, and the molecules that carry withheld environments are harder to predict for any model. In consequence, comparing panels mixes the cost of a holdout with the difficulty of the molecules it selects. Figure 4 instead compares each molecule with itself. For two panels X and Y that differ only in whether an

atomic environment or a chemotype was seen in training (Appendix A.4), the error increase in fold k is

$$\Delta_k^{X|Y} = \frac{100}{4} \sum_p \left(\frac{\sum_{i \in X_k} e_{ik}^p}{\sum_{i \in X_k} \bar{e}_{i,Y}^p} - 1 \right), \qquad e_{ik}^p = \frac{|\hat{y}_{ik}^p - y_i^p|}{\sigma_p},$$

where $\hat{y}_{ik}^p$ is the prediction of the model of fold k for target p of molecule i, σ_p is the standard deviation of target p over AIMEl, so that the four targets share one scale, X_k holds the molecules of panel X in fold k that sit in panel Y in other folds, and $\bar{e}_{i,Y}^p$ is the mean error of molecule i over those folds. The ratio compares the total error of these molecules with the total error that the same molecules receive from the models that saw the environment or the chemotype, so the difficulty of each molecule cancels. Each factor has two such contrasts: unseen environments are B|A (with the chemotype withheld) and C|D (with the chemotype seen). An unseen chemotype is A|D (environments seen) and B|C (environments withheld). All four are positive at every training fraction on both test sets, and for each target separately on the AIMEl test sets. Figure 4 shows the mean of the two contrasts of each factor over the 25 folds. On the QM9 remainder, every molecule is predicted by the 25 models of a fraction, which play the part of the folds, and its panel depends on each model's training set. The molecules of the two chemotypes absent from AIMEl are left out, since their chemotype is never seen.

The distance of Appendix A.4 grades how far each atom lies from a model's own training set. In every model configuration, the error increases with this distance. Molecules with ≥3 rare atoms have a larger error than molecules with none (Fig. 19). The penalty grows with the training set, as in Fig. 4: on the AIMEl test sets, from 1.3–1.4 times the error of molecules without rare atoms with 160 training molecules to 1.8–1.9 times with the full training set, and up to 4.2 times on the remainder. The more a model learns from the environments it has seen, the more the environments it has not seen cost.

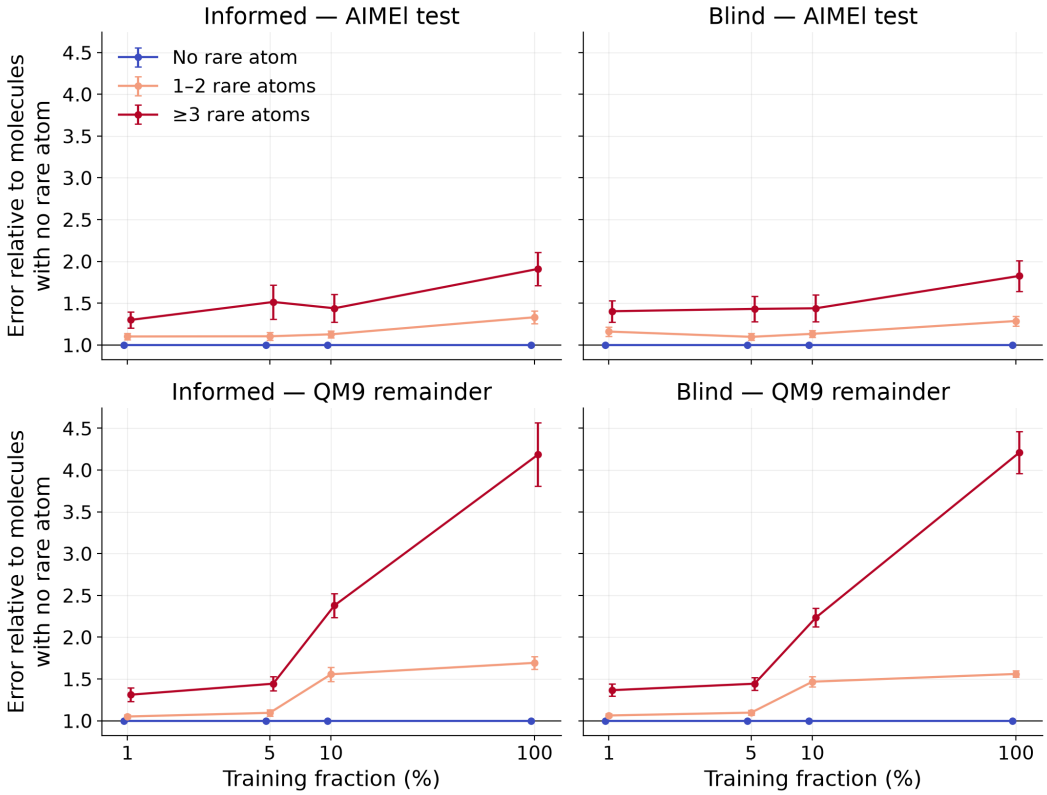

Figure 19: Error of test molecules with 1–2 (salmon) and ≥3 (red) atoms rare with respect to the model's training set, relative to the molecules with none (blue, at 1) for the same model, against the training fraction. Rows: AIMEl test sets (top) and 20 000 remainder molecules (bottom); columns: informed (left) and blind (right) models. Fold means with 95% t-intervals.

A.10 Equivariant models

In the following, we use L to denote the label of the irreducible representation (irrep) of $SO(3)$ with dimension $2L+1$. We represent each atom by a node with scalar features h_i, vector features $\mathbf{x}_i$ and traceless, symmetric features $\mathbf{T}_i$. For edge features, we follow the same convention, but with double subindices (e.g., h_{ij}). We denote the channel-wise (Hadamard) product with $\odot$ and the outer product with $\otimes$. We refer to the whole array of features by $f \equiv (h, \mathbf{x}, \mathbf{T})$, and to the $L > 0$ (non-scalar) features by $\mathbf{f} \equiv (\mathbf{x}, \mathbf{T})$. We refer to the concatenation of invariants by $\|f\| \equiv (h, \|\mathbf{x}\|, \|\mathbf{T}\|)$. Throughout the models, we apply normalization separately to each irreducible representation: $\mathrm{LayerNorm}$ for scalar features and the equivariant $\mathrm{E3Norm}$ of Batatia et al. [4] for vector and tensor features. We denote normalized features uniformly by $\tilde{f} = \mathrm{E3Norm}(f)$, with the understanding that $\mathrm{E3Norm}$ reduces to $\mathrm{LayerNorm}$ for the scalar case. For non-scalar features,

$$\tilde{\mathbf{f}}_{i,c} = w_c \cdot \frac{\mathbf{f}_{i,c}}{\sqrt{\langle \|\mathbf{f}_{i,c}\|^2 \rangle_c + \varepsilon}},$$

where i is the cell (atom) index, c indexes the parallel feature channels within a given irrep, w_c is a learnable per-channel scaling weight, and ε is a small constant for numerical stability. Because the normalization acts only on the magnitude of $\mathbf{f}_{i,c}$ and not on its components individually, the orientation of each vector or tensor feature is preserved — which is what allows $\mathrm{E3Norm}$ to be used inside an equivariant architecture without breaking the equivariance.

In all models, node features are initialized with a message passing layer as follows:

$$h_i^{(0)} = \mathrm{LayerNorm}\left(\sum_j m_{ij}^s \odot \Gamma_0^s\left(r_{ij}\right)\right) \tag{4}$$

$$\mathbf{x}_i{}^{(0)} = \mathrm{E3Norm}\left(\sum_j m_{ij}^v \odot \Gamma_0^v\left(r_{ij}\right) \otimes \hat{\mathbf{r}}_{ij}\right) \tag{5}$$

$$\mathbf{T}_i{}^{(0)} = \mathrm{E3Norm}\left(\sum_j m_{ij}^t \odot \Gamma_0^t\left(r_{ij}\right) \otimes \hat{\mathbf{G}}_{ij}\right) \tag{6}$$

where the messages m_{ij} are generated by a multi-layer perceptron (MLP) that takes as input the embeddings of the atomic species, $m_{ij} = \mathrm{MLP}\left(\mathrm{EMBED}(Z_i), \mathrm{EMBED}(Z_j)\right)$, and Γ^s, Γ^v, Γ^t are geometric filters that map a projection of the inter-atomic displacement onto a radial basis functions, and output through channel-wise sigmoid gates. Finally, $\hat{\mathbf{G}}_{ij} \propto \hat{\mathbf{r}}_{ij} \otimes \hat{\mathbf{r}}_{ij}$ is the unit-norm outer product of the relative position vector, expressed in the traceless symmetric representation as

$$\hat{\mathbf{G}}(\hat{\mathbf{r}}) = \sqrt{3/2}\,[\,r_x r_y,\ r_x r_z,\ r_y r_z,\ (r_x^2 - r_y^2)/2,\ r_z^2 - 1/3\,] \equiv [G_{xy}, G_{xz}, G_{yz}, G_{an}, G_{zz}].$$

The unit norm follows from the Frobenius norm in Eq. 3.

Edge features are initialized from these node features as follows

$$h_{ij}^{(0)} = \mathrm{LayerNorm}\left(\Psi^s\left(\mathrm{LayerNorm}\left(\sum_{k\in\{i,j\}} \psi^s\left(\|f_k^{(0)}\|\right)\right)\right) \odot \Gamma_1^s\left(r_{ij}\right)\right) \tag{7}$$

$$\mathbf{x}_{ij}^{(0)} = \mathrm{E3Norm}\left(\left(\sum_{k\in\{i,j\}} \psi^v\left(\|f_k^{(0)}\|\right) \odot \left(W^v \mathbf{x}_{\mathbf{k}}^{(\mathbf{0})}\right)\right) \odot \Gamma_1^v\left(r_{ij}\right)\right) \tag{8}$$

$$\mathbf{T}_{ij}^{(0)} = \mathrm{E3Norm}\left(\left(\sum_{k\in\{i,j\}} \psi^t\left(\|f_k^{(0)}\|\right) \odot \left(W^t \mathbf{T}_{\mathbf{k}}^{(\mathbf{0})}\right)\right) \odot \Gamma_1^t\left(r_{ij}\right)\right) \tag{9}$$

where the superscripts s, v, and t denote separate parameter sets for the scalar, vector, and tensor channels respectively; Ψ and ψ are MLPs acting on per-node feature norms; and W^v and W^t are

learnable linear maps between node and edge channel dimensions (taken as the identity when these dimensions match)

After these, all layers follow a residual update structure, $f'_r = f_r + g\left(\|f_r\|\right) \odot M_r(\tilde{f}_r; \{\tilde{f}_s, \tilde{f}_t\})$, where g is a gating MLP with a sigmoid output that depends on the unnormed features, and $M_r(\tilde{f}_r; \{\tilde{f}_s, \tilde{f}_t\})$ is a message aggregation with $\tilde{f}_r, \tilde{f}_s, \tilde{f}_t$ being $\mathrm{E3Normed}$ receiver, sender, and intermediary features, respectively. In general,

$$M_r(\tilde{f}_r; \{\tilde{f}_s, \tilde{f}_t\}) = \sum_{s,t \in \mathcal{N}_t(r)} m\left(\tilde{f}_r, \tilde{f}_s, \tilde{f}_t\right) \odot \Gamma(\mathbf{R}_r, \mathbf{R}_s) \odot W \tilde{f}_s, \tag{10}$$

where W is a linear map to match dimensions (if necessary), and the message generating function is an MLP

$$m\left(\tilde{f}_r, \tilde{f}_s, \tilde{f}_t\right) = \mathrm{MLP}\left(\tilde{h}_r, \tilde{h}_s, \|U\tilde{\mathbf{f}}_r\|, \|V\tilde{\mathbf{f}}_s\|, \cos(U\tilde{\mathbf{f}}_r, V\tilde{\mathbf{f}}_s), \|\tilde{f}_t\|\right), \tag{11}$$

with U and V being linear maps applied to the $L > 0$ channels, and $\cos(A, B)$ being the cosine similarity of the $L > 0$ channels. This way, the messages are generated based on both magnitude and directional information.

The geometric filter Γ takes as input an invariant geometric relation between the receiver and sender real-space geometric information, $\mathbf{R}_r, \mathbf{R}_s$, and is absent in layers where there is no obvious geometric relation or to avoid redundant information from previous transformations (e.g., edge-to-node messages in late layers). In node-to-node and node-to-edge messages, we use radial basis functions evaluated at the length of the edge (i.e. distance between node and receiver nodes). In edge-to-edge messages, we project $\cos\left(\mathbf{G}_r, \mathbf{G}_s\right) = \mathrm{Tr}\left(\hat{\mathbf{G}}_r^T \hat{\mathbf{G}}_s\right)$ onto a basis of Legendre polynomials, $\{P_n(x)\}$, since the gyration tensors contain angular information.

At certain steps, geometry is re-injected to avoid degradation of geometric signal. In this case, the residual connection gets two simultaneous updates, one for features and one for geometry,

$$f'_r = f_r + g_f(\|f_r\|) \odot M_r + g_{\mathbf{R}}(\|f_r\|) \odot \Delta\mathbf{f}_r, \tag{12}$$

where

$$\Delta\mathbf{f}_r = \sum_{s,t \in \mathcal{N}_t(r)} m\left(\tilde{f}_r, \tilde{f}_s, \tilde{f}_t\right) \odot \Gamma(\mathbf{R}_r, \mathbf{R}_s) \otimes \Delta\hat{\mathbf{R}}_{rs}. \tag{13}$$

Here, $\Delta\hat{\mathbf{R}}_{rs} = (\hat{\mathbf{r}}_{rs}, \hat{\mathbf{G}}_{rs})$ is a unit-norm tensor.

At the end of a messaging cycle, we apply a feed-forward residual update following the expression

$$f'_r = f_r + g(\|f_r\|) \odot \mathrm{MLP}(\|\tilde{f}_r\|, W \cdot \mathrm{EMBED}(Z_r)) \odot (\mathbb{I}_0 | \tilde{\mathbf{f}}_r) \tag{14}$$

where $\mathbb{I}_0$ is the identity over the scalar channels (i.e. scalar features are not gated). We include a linear transformation of the initial embedding of the atomic species whenever this layer is applied to nodes, but that term is omitted when applied to edges.

Finally, the prediction of the model for the properties of a given atom, $\hat{y}_i = (\hat{N}_i, \hat{\lambda}_i, \hat{\boldsymbol{\mu}}_i, \hat{\mathbf{Q}}_i)$ is a readout that depends on an atomic head at the end of each layer:

$$(\hat{N}_i^{(l)}, \hat{\lambda}_i^{(l)}) = \mathrm{MLP}(h_i^{(l)}), \quad \hat{\boldsymbol{\mu}}_i^{(l)} = \sum_c w_c^{v,(l)} \mathbf{x}_{i,c}^{(l)}, \quad \hat{\mathbf{Q}}_i^{(l)} = \sum_c w_c^{t,(l)} \mathbf{T}_{i,c}^{(l)},$$

$$\text{and } \hat{y}_i = \sum_l \alpha^{(l)} \hat{y}_i^{(l)},$$

where the layer weights α come from a softmax, $\alpha^{(l)} = \exp(\omega^{(l)}) / \sum_{l'} \exp(\omega^{(l')})$.

A.11 Parity plots for the OOD predictions of QT-Net on AIMEl

We show parity plots for QT-Net's predictions on the held-out atomic environments {H_10, C_11, N_13, O_10} for each of the four QTA targets: electron populations N (Fig. 20), localization indices λ (Fig. 21), atomic contributions to the molecular dipole moment $\boldsymbol{\mu}$ (Fig. 22), and quadrupole moments $\mathbf{Q}$ (Fig. 23).

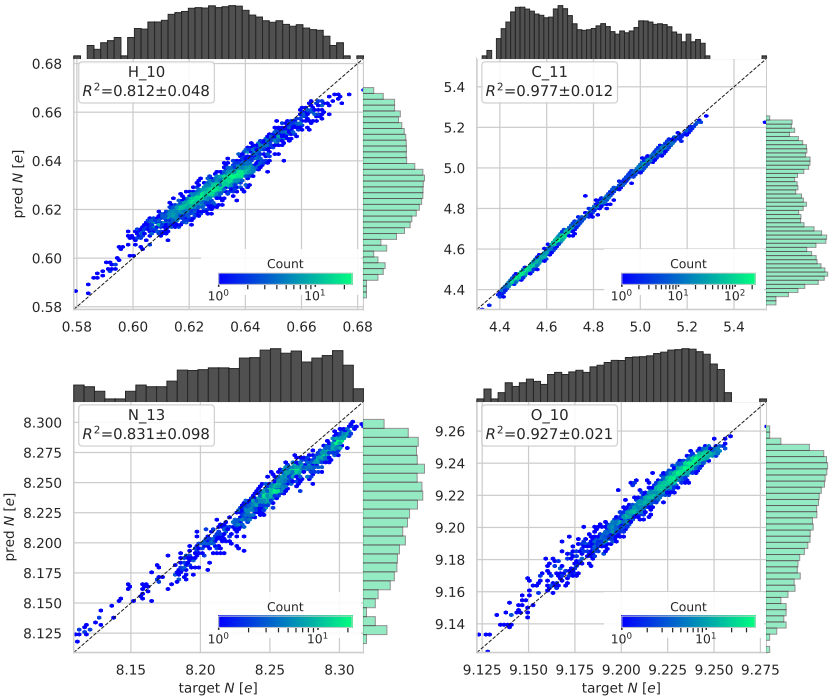

Figure 20: Parity plot for the OOD predictions of QT-Net for electron populations.

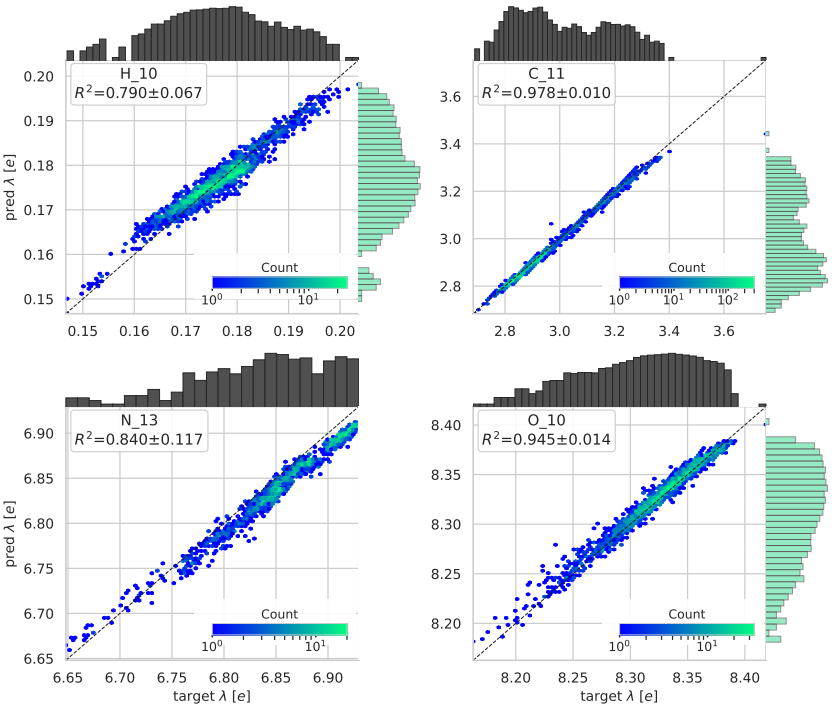

Figure 21: Parity plot for the OOD predictions of QT-Net for localization indices.

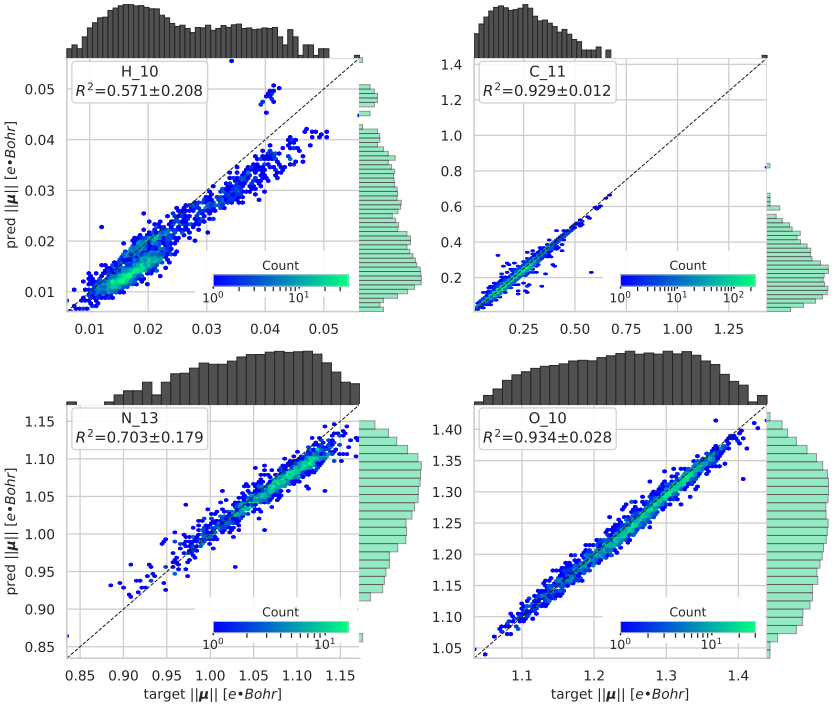

Figure 22: Parity plot for the OOD predictions of QT-Net for atomic contributions to the molecular dipole moment.

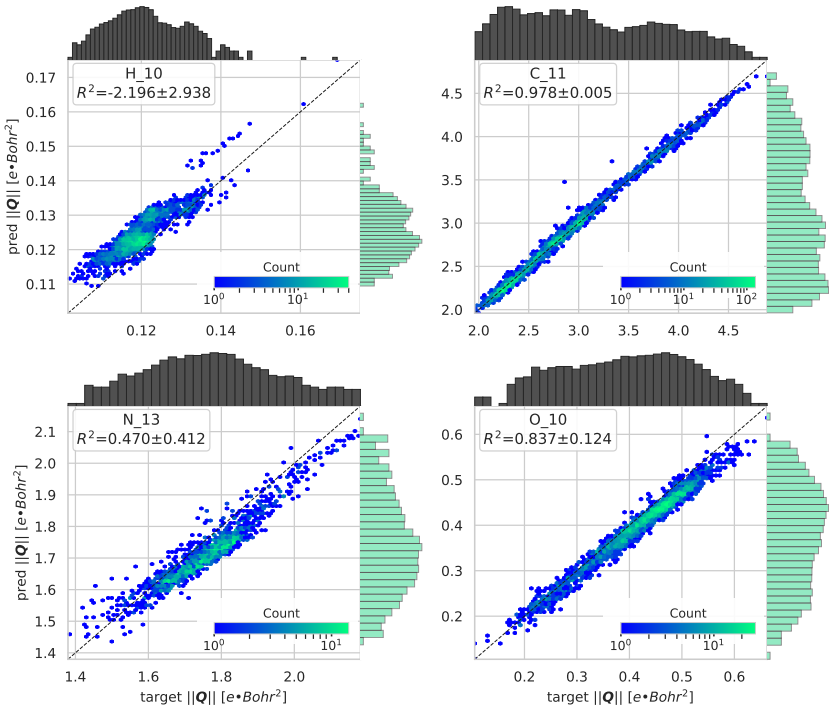

Figure 23: Parity plot for the OOD predictions of QT-Net for atomic quadrupole moments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The results in Fig. 16 show that metrics that account for all atoms in the test set have systematically higher values. The results in Section 5.2 match the claims about the inference quality of QT-Net.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 5.2, it is acknowledged that we cannot prove statistical significance of the improvement in molecular property prediction in the AIMEl subset. In Section 6, it is acknowledged that the model comparison only applies to this specific held-out set.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: No theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Cutoffs, number of nearest neighbors, number of epochs, learning rates and seeds are found across the paper and in the accompanying repository, referenced in Section 8.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All code and data is provided open access in the accompanying repository, referenced in Section 8.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 3.2 and repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section A.5 shows supporting metrics for the statistical significance of the comparison in Section 5.1. Section 5.2 mentions explicitly the usage of bootstrapping for ensemble benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details about training and computational cost are found in Section A.1. Seeds are found in the repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Conformity with the NeurIPS Code of Ethics has been checked.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The provided models and methodology cannot be used (at least in principle) with malicious intent.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The references to datasets are included in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All code is included in the accompanying repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.