Rethinking evaluation of AI models in atomic chemical space

Pablo Martínez Crespo [1, †], **Stefano Ribes**[1], **Martin Rahm**[2], **Richard Beckmann**[1], **Robert S. Jordan**[3], **Marisa Gliege**[4], **Santiago Miret**[5], **Vijay Kris Narasimhan**[6], **Rocío Mercado**[1]

[1]Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg
[2]Department of Chemistry and Chemical Engineering, Chalmers University of Technology
[3]Intel Corporation, [4]EMD Electronics, [5]Lila Sciences, [6]M Ventures.
†First author's email: pabloma@chalmers.se

Atomic properties such as partial charges or multipoles encode chemically meaningful information that can inform downstream molecular AI tasks, but their evaluation as machine learning targets has been complicated by the absence of a principled out-of-distribution (OOD) evaluation method 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. We show that this procedure provides realistic benchmarks, as opposed to evaluations based on random splits, which systematically overestimate the performance of atomic AI models.

Following our evaluation protocol, 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 of 30k molecules from the AIMEl subset [1] of QM9. 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 98k QM9 molecules outside our training set, and that these inferred properties can yield improvement when used as input features for downstream molecular property prediction. To further validate the framework, molecular dipole moments computed from QT-Net's per-atom outputs recover the ground-truth values reported in QM9 with $R^2 = 0.931 \pm 0.003$, a physical consistency check the model was never trained on.

This contribution presents, for the first time, an environment-level OOD protocol for atomic targets and the resulting statistically controlled architecture comparison. With faithful evaluation in place, the accuracy we observe across elements and properties suggests that the bottleneck for quantum topological atomic property prediction has shifted from architecture to data coverage. We release all code and data, including a JAX implementation of QT-Net, to support the broader use of learned atomic properties as inductive biases for atomic-scale molecular machine learning.

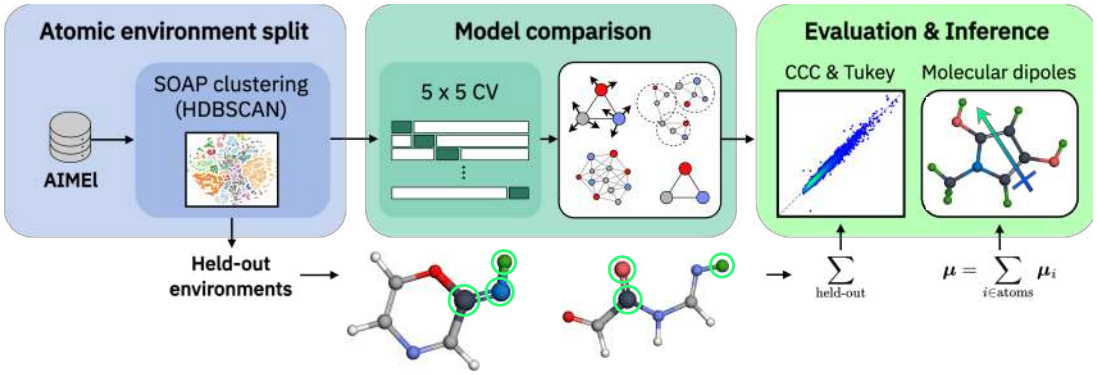

Figure 1: QT-Net training and evaluation pipeline on AIMEl for atomic (QTAIM) property prediction. The molecules at the bottom are sample molecules from the test set (left: `N=C1OC=CN=C1`; right: `N=CNC(=O)C=O`), where the circled atoms are in environments held-out from training and are the only ones 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.

Acknowledgments

PMC acknowledges funding from the Intel-Merck joint university research center for AI-Aware Pathways to Sustainable Semiconductor Process and Manufacturing Technologies (AWASES). Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), funded by the Swedish Research Council, and the Berzelius system funded by the Knut and Alice Wallenberg foundation and operated by NAISS. The authors declare no competing interests.

References

[1] B. Meza-González at al., Quantum topological atomic properties of 44k molecules, *Scientific Data* **11**, 1, 945 (2024).