QSTRUCT: QUANTUM-CIRCUIT STRUCTURAL PRIORS FOR DATA-EFFICIENT MOLECULAR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable molecular prediction under scarce labels and chemical distribution shift benefits from structural biases beyond local geometry. We introduce QStruct, a hybrid framework in which a pooled geometric GNN embedding sets bounded rotation angles for a shallow eight-qubit ring circuit. Eight single-qubit Z expectations and seven adjacent ZZ correlations form a 15-dimensional feature vector fused with the graph representation. Exact classical state-vector evaluation supplies this inductive bias without claiming quantum advantage. Across QM9, MD17, PCQM4Mv2, and OC20, reported results support data efficiency and extrapolation while retaining competitive full-data accuracy. With 100 QM9 training molecules, MAE is 0.059, versus 0.070–0.074 for strong baselines; full-data OC20 OOD-Adsorbate/Catalyst errors decrease from $0.58/0.62$ to $0.54/0.59$. Controlled ablations favor learned circuit features over frozen or classical alternatives, while coverage curves reveal remaining calibration gaps. Separate descriptive records illustrate uncertainty ranking, computational trade-offs, and target-dependent residuals without adding independently verified benchmark evidence. Code: `https://anonymous.4open.science/r/QCMRL-7203`.

1 INTRODUCTION

Scarce labels and chemical distribution shift challenge molecular prediction. Geometric graph networks encode local interactions and symmetries (Gilmer et al., 2017; Schütt et al., 2021), but benchmark studies still identify sensitivity to chemical shift (Xia et al., 2023) and uncertainty miscalibration (Busk et al., 2021). These limitations motivate additional constraints on the representation, beyond simply increasing model capacity.

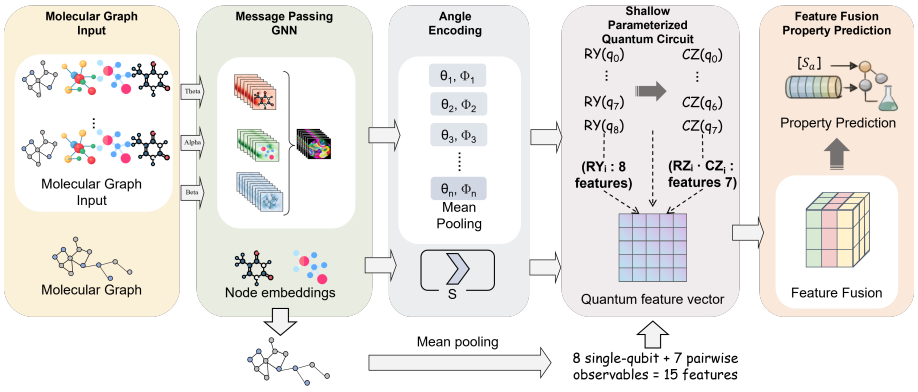


Figure 1: Property-prediction pipeline with quantum-induced priors.

We introduce *quantum-circuit-induced structural priors*: bounded molecular encodings are transformed by shallow parameterized circuits and summarized through fixed measurement operators.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Figure 1 illustrates this representation pathway. The circuit topology, depth, and observables restrict which encoded correlations can enter the predictor. This use of variational circuits (Cerezo et al., 2020) complements geometric biases (Bronstein et al., 2021; Geiger & Smidt, 2022) and physics-informed learning (Meng et al., 2022); it does not identify the learned state with an electronic wavefunction or establish quantum computational advantage.

QStruct concatenates local and pairwise circuit measurements with classical graph features. Mean pooling makes the circuit input independent of molecular size, while the classical branch preserves local geometry. The circuit is evaluated by exact classical simulation.

Our contributions are (i) a shallow-circuit structural prior, (ii) an explicit observable-graph fusion interface, and (iii) four-benchmark evaluation of data efficiency, calibration, chemical extrapolation, and parameter-matched controls. We assess whether this restricted feature family justifies its simulation cost; additional descriptive diagnostics characterize behavior without adding independent benchmark evidence.

2 Preliminaries and Problem Setup

We predict molecular properties from graphs $G = (V, E)$ with atom attributes and optional coordinates $\{\mathbf{r}_i \in \mathbb{R}^3\}_{i \in V}$, following the supervised setting of MoleculeNet (Wu et al., 2017). Given $\mathcal{D} = \{(G_n, y_n)\}_{n=1}^{N}$, a graph representation $\phi_\theta(G)$ feeds a task-specific predictor $f_\theta(G)$. We study few-shot training and extrapolation where $\mathcal{P}_{\text{test}} \neq \mathcal{P}_{\text{train}}$ along chemical dimensions such as composition or size. Figure 2 summarizes the restricted hypothesis space: circuit depth, entangling topology, and observable selection constrain the measurement-derived representation. These restrictions supplement backbone symmetries; symmetry preservation still depends on the classical encoding, as detailed in Appendix B.

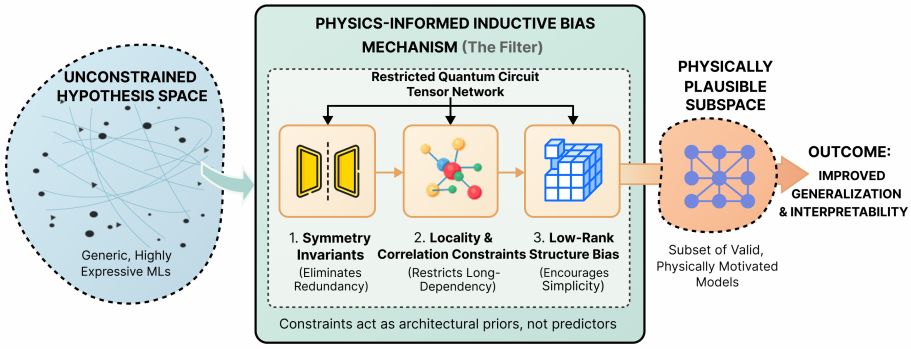


Figure 2: Structural priors from restricted quantum circuits.

3 Method

Encoding and Circuit. Let $\mathbf{s}$ be the mean of the penultimate-layer node states. An affine encoder produces $\boldsymbol{\alpha} = \mathbf{W}_e\mathbf{s} + \mathbf{b}_e$, with bounded angles $a_i = \pi(\tanh \alpha_i + 1)$. We use $n_q = 8$ qubits, one encoding block without data re-uploading, and $L_q = 3$ trainable layers:

$$U_{\text{enc}}(\mathbf{a}) = \bigotimes_{i=1}^{n_q} R_Y(a_i), \qquad |\psi\rangle = U_{L_q} \cdots U_1 U_{\text{enc}}(\mathbf{a}) \, |0\rangle^{\otimes n_q},$$
$$U_\ell = U_{\text{CZ}} \bigotimes_{i=1}^{n_q} [R_Z(\phi_{\ell i}) R_Y(\theta_{\ell i})], \qquad U_{\text{CZ}} = \prod_{i=1}^{n_q} \text{CZ}_{i,1+(i \bmod n_q)}. \tag{1}$$

The ring includes the $(n_q, 1)$ edge. Figure 3 shows the circuit and readout. Restricted depth and entangling topology constrain the encoded interactions accessible to each observable (Cerezo et al., 2020); Appendix A.3 gives the corresponding light-cone bounds. The encoded coordinates summarize the whole graph, so circuit adjacency describes latent coordinates rather than individual atomic bonds.

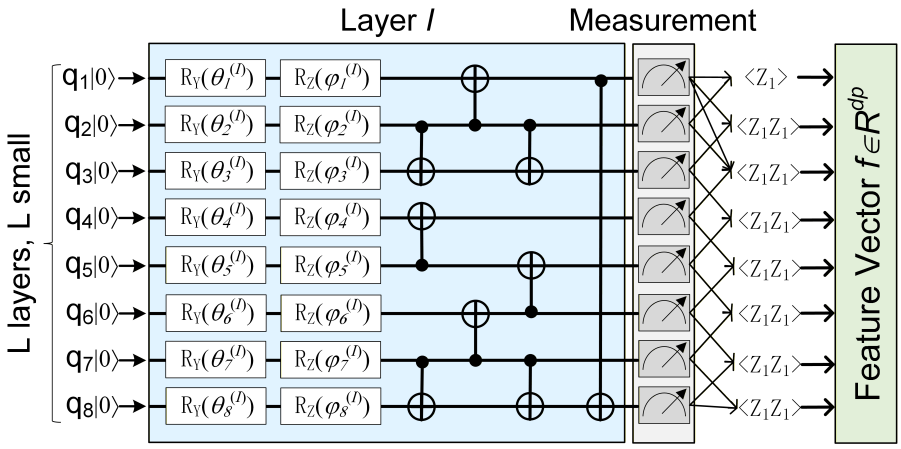


Figure 3: Local and pairwise quantum-circuit observables.

Measurement and Fusion. Exact noiseless state-vector evaluation supplies eight expectations $\langle Z_i \rangle$ and seven open-chain correlations $\langle Z_i Z_{i+1} \rangle$, $i = 1, \ldots, 7$, concatenated into $\mathbf{z}_q \in \mathbb{R}^{15}$. The readout omits the wrap-around pair although the entangler retains it. These fixed observables summarize marginal imbalance and adjacent agreement without accessing every amplitude. With the final pooled graph embedding $\mathbf{g}$, prediction uses

$$\mathbf{f} = \mathrm{ReLU}(\mathrm{LN}(\mathbf{W}_f[\mathbf{g}; \mathbf{z}_q] + \mathbf{b}_f)) \in \mathbb{R}^{256}, \qquad \widehat{y} = h_\omega(\mathbf{f}). \tag{2}$$

Layer normalization balances the two feature scales; concatenation retains the classical prediction path. The head has two 256-dimensional hidden layers and task-dependent outputs. Figure 4 traces the complete molecular-encoding, circuit, measurement, and fusion pathway.

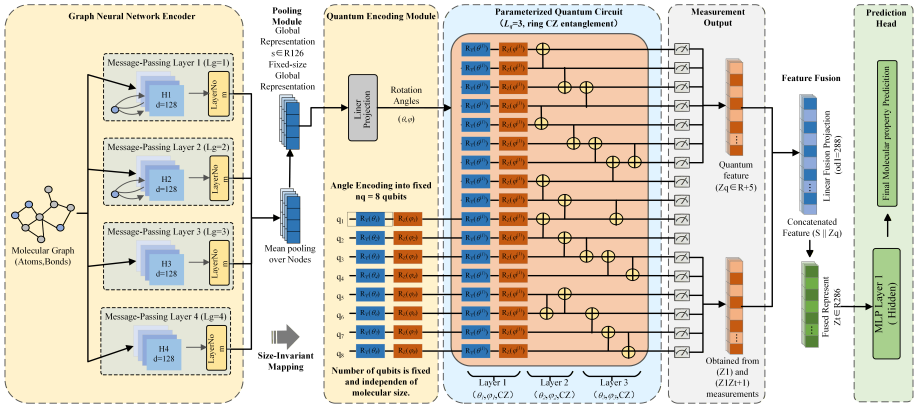


Figure 4: Hybrid fusion of quantum measurements and pooled graph features.

Learning. All graph, circuit, fusion, and head parameters are optimized jointly by reverse-mode differentiation through the exact simulator. Scalar uncertainty prediction adds a positive scale output and Gaussian negative log likelihood; energy-force tasks follow their stated supervision. Fixed-parameter inference has no sampling noise. Appendices A.2-A.6 specify encoder inputs, fusion, objectives, gradient propagation, and simulation complexity. Gate matrices, initialization, and measurement identities are in Appendix B.

4 Experimental Evaluation

4.1 Benchmark Setup

We evaluate QM9 energy, dipole moment, and gap; MD17 aspirin (ASP) and malondialdehyde (MDA) energies/forces; PCQM4Mv2 gap; and OC20 adsorption energy. Results average seeds $\{1, 2, 3, 4, 5\}$ on fixed splits. QM9 uses 104,000 training and 13,000 test molecules, with nested few-shot subsets of 50-5,000; its lowest-similarity 20% test subset defines the harder OOD ablation. Table 3 uses full training sets, whereas Figure 6 uses 100,000 OC20 training structures and 20,000 evaluation examples per split and seed; these protocols are not numerically interchangeable. We report MAE/RMSE, regression ECE, and interval coverage. Energy/gap units are eV, dipole units Debye, and MD17 energy/force units meV and meV/Å. Lower errors are better. A validation-fitted variance scale is frozen for ID/OOD testing. Parameters are matched within 1% only for controlled variants, not all published baselines. Benchmark circuit evaluations use exact classical state-vector simulation. Complete splits, metrics, baseline references, optimization, calibration, statistical, and timing protocols are specified in Appendix C; expanded numerical analysis is in Appendix E.

Table 1: QM9 predictive performance.

Model	Energy MAE ↓	Energy RMSE ↓	MAE@100 ↓	ECE ↓	Run Std ↓	μ MAE ↓	Gap MAE ↓
MPNN	0.026	0.046	0.083	0.142	0.004	0.037	0.068
SchNet	0.026	0.047	0.081	0.142	0.004	0.037	0.066
DimeNet++	0.022	0.041	0.074	0.118	0.003	0.033	0.062
PaiNN	0.021	0.040	0.071	0.115	0.003	0.032	0.061
Graph Transformer	0.025	0.045	0.078	0.127	0.005	0.034	0.065
PhysNet	0.021	0.039	0.070	0.111	0.003	0.031	0.061
SOAP + KRR	0.028	0.050	0.085	0.136	0.004	0.038	0.067
Ours	**0.021**	**0.038**	**0.059**	**0.089**	**0.002**	**0.030**	**0.060**

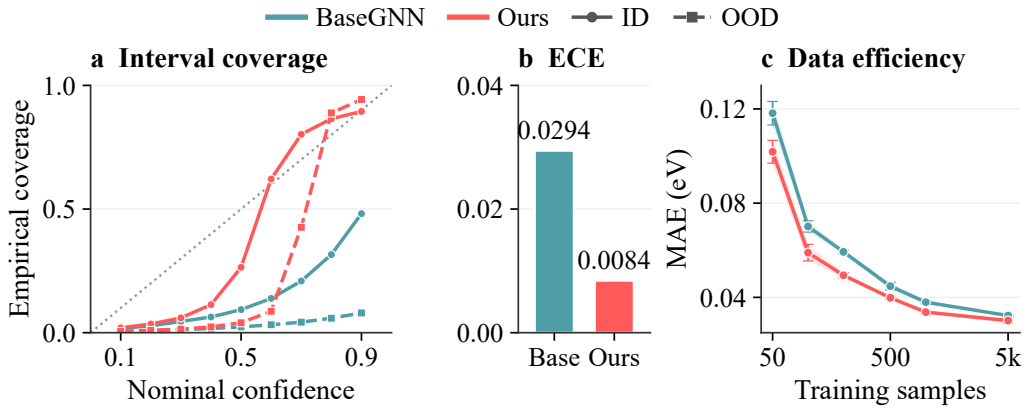


Figure 5: QM9 calibration and sample efficiency.

4.2 Prediction and Generalization

QM9. Table 1 combines full-data metrics with MAE@100. Ours ties the best full-data energy MAE at 0.021 eV, while reducing 100-shot MAE from PhysNet's 0.070 to 0.059 (15.7%), pre-calibration ECE from 0.111 to 0.089, and run standard deviation from 0.003 to 0.002. Figure 5(a-b) instead uses 100-shot checkpoints: post-calibration ECE is 0.0294 for BaseGNN ("Base") and 0.0084 for Ours. Panel (c) shows a larger absolute benefit with fewer labels: MAE changes from 0.1182 to 0.1018 at 50 training molecules, versus 0.0323 to 0.0301 at 5,000. At nominal confidence 0.9, ID/OOD coverage improves from 0.4814/0.0798 to 0.8948/0.9429. However, Ours covers only 0.2644/0.0401 at confidence 0.5; calibration remains uneven across levels. The largest relative

improvements concern scarce supervision and reliability, while full-data point-error differences are small. Full-data/pre-calibration and few-shot/post-calibration scores are kept distinct.

Table 2: MD17 prediction errors.

Model	ASP		MDA	
	E ↓	F ↓	E ↓	F ↓
SchNet	0.18	0.42	0.12	0.33
DimeNet++	0.15	0.36	0.10	0.28
PaiNN	0.14	0.34	0.09	0.26
PhysNet	0.14	0.33	0.09	0.25
Ours	**0.12**	**0.30**	**0.08**	**0.23**

Table 3: Large-scale prediction errors.

Model	PCQM4Mv2	OC20		
	MAE ↓	ID	OOD-A	OOD-C
MPNN	0.128	0.56	0.71	0.77
SchNet	0.123	0.52	0.66	0.71
SE(3)-GNN	0.114	0.49	0.63	0.68
PaiNN	0.103	0.45	0.58	0.63
PhysNet	0.102	0.46	0.59	0.64
SpookyNet	0.101	0.45	0.58	0.62
Ours	**0.098**	**0.43**	**0.54**	**0.59**

MD17. Table 2 reports energy/force MAEs of $0.12/0.30$ on aspirin and $0.08/0.23$ on malondialdehyde, compared with PhysNet's $0.14/0.33$ and $0.09/0.25$. These are reductions of $14.3\%/9.1\%$ and $11.1\%/8.0\%$, respectively. Mean error improves for both molecules and targets; this alone does not establish improved tail behavior.

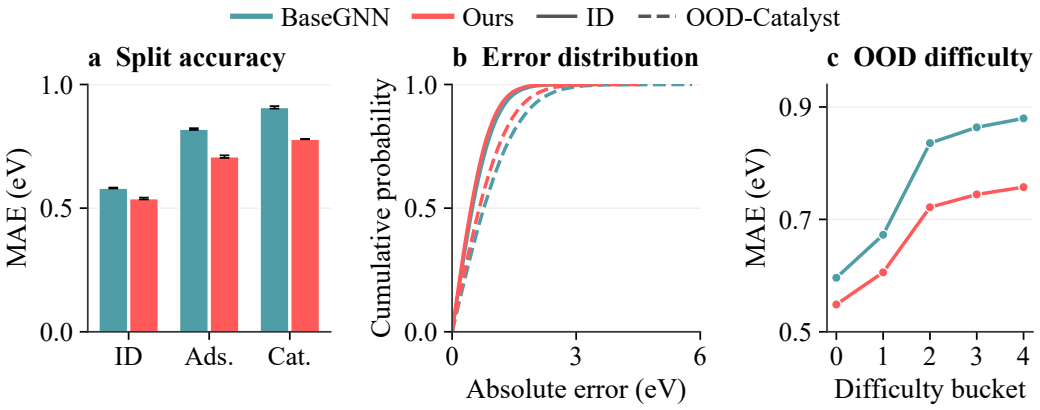


Figure 6: OC20 errors under distribution shift.

Large-Scale and OOD Results. In Table 3, PCQM4Mv2 MAE decreases from SpookyNet's 0.101 to 0.098 eV (3.0%).

Full-data OC20 ID/OOD-Adsorbate/OOD-Catalyst errors decrease from the best baselines' $0.45/0.58/0.62$ to $0.43/0.54/0.59$ ($4.4\%/6.9\%/4.8\%$). Figure 6 uses the separate reduced-data protocol; Ads./Cat. denote the two OOD splits. Panel (a) gives BaseGNN errors $0.581/0.820/0.908$ versus $0.539/0.708/0.779$ for Ours. In panel (b), OOD-Catalyst median/P90 errors decrease from $0.768/1.866$ to $0.659/1.607$; panel (c)'s hardest-bucket MAE decreases from 0.8797 to 0.7574 (13.9%). The mean, central quantile, and upper tail all favor Ours within this protocol. The full cross-benchmark heatmap and metric-by-metric comparisons appear in Appendix D.

4.3 Ablations and Optimization

Structural and Quantum Components. Figure 7(a) gives maximum feature-energy correlation 0.2148 for Z_0Z_1 and top-five mean $|r| = 0.1600$; these are diagnostic associations, not causal physical evidence. In panel (b), removing geometry/correlation/hierarchy raises complete-test MAE from 0.059 to $0.066/0.072/0.064$ eV. Correlation removal has the largest penalty. Panel (c)'s frozen circuit, classical bottleneck, and no-quantum controls give $0.067/0.065/0.071$, all above the full model. Table 4 evaluates the separate low-similarity OOD subset: the full circuit's 0.081 improves on the base GNN's 0.097 (16.5%), the MLP's 0.089, low-rank's 0.086, and frozen features' 0.090. Global compression helps, with a further gain from the learned circuit in these controls.

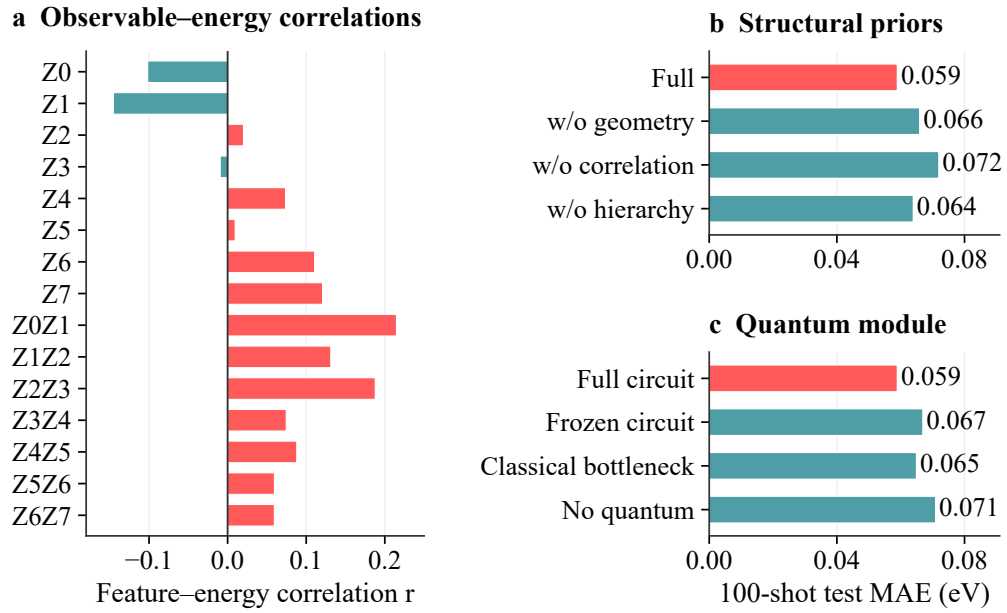


Figure 7: Observable correlations and component ablations.

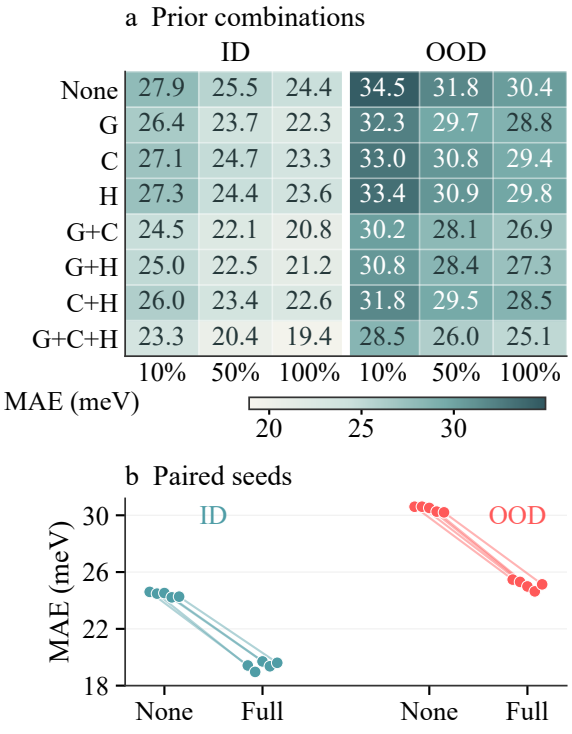


Figure 8: Structural-prior combinations.

Table 4: Quantum-module controls.

Variant	MAE ↓
No circuit (BaseGNN)	0.097
MLP projection	0.089
Low-rank bottleneck	0.086
Frozen circuit	0.090
Learned circuit (ours)	**0.081**

Factorial Priors. G/C/H denote geometry/correlation/hierarchy; Full combines all three. Figure 8 uses three training fractions and five matched entries. At 100%, Full reduces ID/OOD MAE from 24.42/30.44 to 19.41/25.11 meV (20.5%/17.5%) and leads all six settings. At 10%, the corresponding errors decrease from 27.88/34.47 to 23.28/28.49 meV. G alone gives the largest single-prior ID gain at 100% (8.7%). These descriptive comparisons do not establish significant interactions; details and provenance are in Appendix D.

The OOD value 0.081 and complete-test value 0.059 are different evaluation quantities; leave-one-out ablations alone do not establish significant interactions between priors.

Representation and Gradients. Figure 9 contains 500 standardized 15-dimensional feature vectors. Within-block mean absolute correlations are 0.766/0.750/0.728, versus 0.063 between blocks.

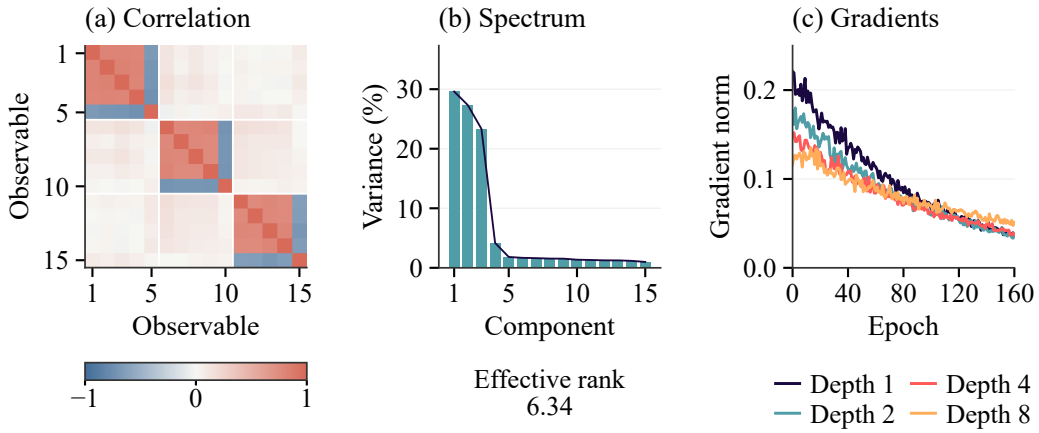


Figure 9: Feature structure and gradient trajectories.

The first three components explain 80.3% of variance, with effective rank 6.34. At depths 1/2/4/8, gradient norms fall from 0.2210/0.1766/0.1480/0.1161 to 0.0358/0.0373/0.0371/0.0510 over 160 epochs. These descriptive records show block structure and depth-dependent trajectories, without establishing quantum entanglement or favorable scaling at untested sizes. Appendix D retains the complete analysis.

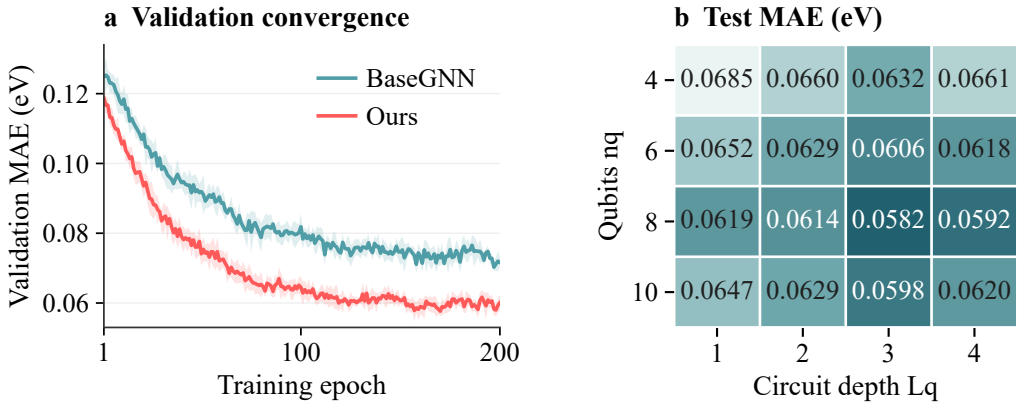


Figure 10: Training convergence and circuit sensitivity.

Convergence and Circuit Size. Figure 10(a) records all 200 epochs: final validation MAE is 0.0716 for BaseGNN and 0.0601 for Ours (16.2% lower). Reaching MAE < 0.08 takes 80 versus 38 epochs, about $2.1\times$ fewer epochs, not an equivalent wall-clock speedup. Panel (b) spans 0.0582-0.0685 test MAE over 16 configurations. Eight qubits with three/four layers give 0.0582/0.0592; extra depth does not monotonically improve accuracy. The 0.0601 validation value is not the 0.059 complete-test estimate.

4.4 Additional Diagnostic Records

The author-confirmed records originated as illustrative layout inputs; original run logs are unavailable. These are descriptive comparisons, not independently verified benchmark evidence, and are not pooled with the benchmark protocols. Provenance and full analyses appear in Appendices D-E.

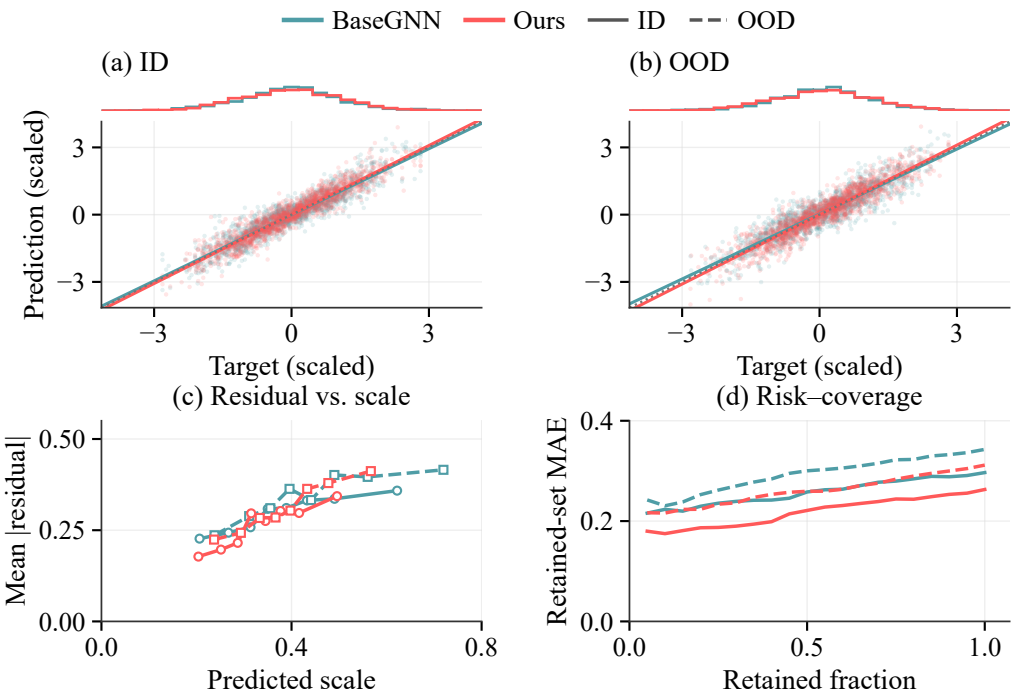


Figure 11: Prediction error and uncertainty.

Prediction and Uncertainty. Figure 11 summarizes 1,500 records per model/domain in scaled units. Full-coverage ID/OOD MAEs are $0.263/0.312$ for Ours versus $0.297/0.343$ for BaseGNN ($11.4\%/9.1\%$ lower). Retaining the least-uncertain 50% gives $0.221/0.259$ versus $0.258/0.300$; Ours improves by $15.9\%/16.8\%$ relative to full coverage. Its first-to-last uncertainty-bin residuals rise from 0.178 to 0.343 on ID and 0.224 to 0.412 on OOD. These records support error ranking and selective prediction, although ID bins are not strictly monotonic and ranking does not establish probabilistic calibration.

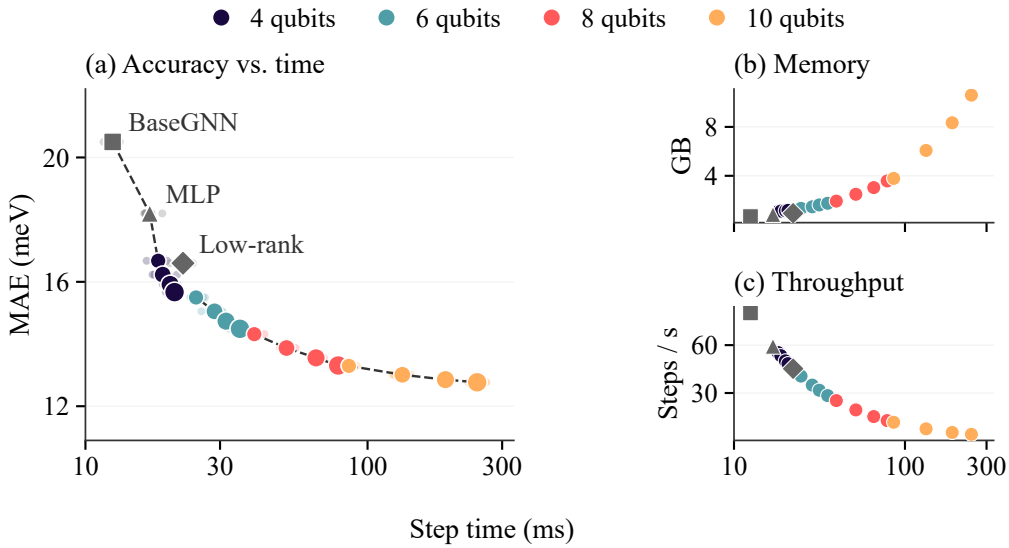


Figure 12: Accuracy-compute trade-offs.

Accuracy and Cost. Figure 12 compares 19 configurations with ten timing entries each. At eight qubits, depth one to four lowers MAE from 14.32 to 13.31 meV, while step time rises from 39.67 to 78.69 ms and memory from 1.93 to 3.58 GB. Ten qubits at depth four attain 12.77 meV with 244.65 ms, 10.60 GB, and 4.10 steps/s: only 4.1% further error reduction at 3.11× the eight-qubit step cost. The recorded trade-off favors moderate configurations under cost constraints. Dataset, device, and batch identifiers are absent, so it does not establish measured hardware efficiency.

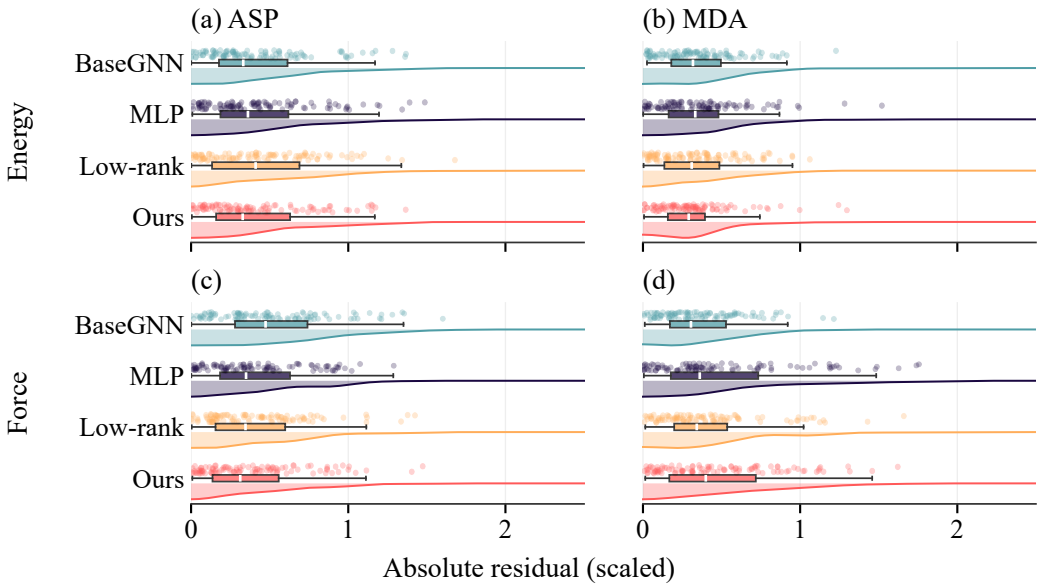


Figure 13: MD17 residual distributions.

MD17 Residuals. Figure 13 uses 100 records per molecule/model/target, in scaled units that cannot be compared directly with Table 2. Ours reduces ASP/MDA energy means from 0.426/0.354 to 0.417/0.322 and P90 from 0.959/0.702 to 0.884/0.562. ASP force has the lowest mean/median, 0.401/0.313, but low-rank's P90 is better (0.780 versus 0.867). MDA force reverses the advantage: Ours has mean 0.481 versus 0.365 and P90 1.022 versus 0.714 for BaseGNN. Thus, accuracy and tails depend on target and comparator; individual frames are not additional independent training seeds.

5 Related Work

Message passing and geometric models encode local molecular structure (Gilmer et al., 2017; Schütt et al., 2021). Variational circuits and quantum kernels define alternative feature families (Cerezo et al., 2020; Kübler et al., 2021). Appendix F compares architectural, physics-informed, uncertainty, and quantum-learning approaches with QStruct's observable-graph fusion.

6 Conclusion and Limitations

QStruct augments molecular graph representations with fifteen shallow-circuit observables. Reported benchmarks and classical controls support this feature family for data efficiency and chemical extrapolation, while calibration remains imperfect. The additional records illustrate cost and target-dependent behavior without adding independent benchmark evidence. Broader chemical coverage and scalable circuit evaluation remain important tests.

Limitations. Exact simulation scales exponentially; finite-shot estimation and hardware noise are untested. Four benchmarks and five seeds do not cover arbitrary chemical shifts or reactions. Encoder, observable, and calibration choices limit transfer. Illustrative-origin diagnostics lack run logs; larger circuits need renewed accuracy, cost, and trainability studies (Appendix B.4).

AI USE STATEMENT

Generative AI tools were used for language polishing and for adapting the manuscript's LaTeX formatting to the ICLR 2027 submission template. The authors are responsible for the final content of the paper.

ETHICS STATEMENT

Code of Conduct. We commit to the ICLR Code of Conduct throughout submission, review, and conference participation. We support respectful scientific discussion, fair treatment of all participants, and an environment free from harassment, discrimination, and personal attacks, including in online exchanges.

Code of Ethics. We commit to the ICLR Code of Ethics and transparent reporting of methods, results, and limitations. This study uses public molecular benchmarks without human participants or personal data. Molecular predictors may enable harmful chemical applications, and dataset bias and distribution shift can compromise reliability; downstream use therefore requires expert oversight and independent experimental validation, with attention to computational and environmental costs.

REFERENCES

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450.

Simon L. Batzner, Tess E. Smidt, Lixin Sun, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. Se(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *CoRR*, abs/2101.03164, 2021. URL https://arxiv.org/abs/2101.03164.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL https://arxiv.org/abs/2104.13478.

Lukas Burgholzer, Alexander Ploier, and Robert Wille. Simulation paths for quantum circuit simulation with decision diagrams what to learn from tensor networks, and what not. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(4):1113–1122, 2022.

Jonas Busk, Peter Bjørn Jørgensen, Arghya Bhowmik, Mikkel N. Schmidt, Ole Winther, and Tejs Vegge. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. *CoRR*, abs/2107.06068, 2021. URL https://arxiv.org/abs/2107.06068.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *CoRR*, abs/2012.09265, 2020. URL https://arxiv.org/abs/2012.09265.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon M. Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary W. Ulissi. The open catalyst 2020 (OC20) dataset and community challenges. *CoRR*, abs/2010.09990, 2020. URL https://arxiv.org/abs/2010.09990.

Stefan Chmiela, Huziel Enoc Sauceda Felix, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.*, 240:38–45, 2019. doi: 10.1016/J.CPC.2019.02.007. URL https://doi.org/10.1016/j.cpc.2019.02.007.

Fabian Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*

2020, December 6-12, 2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 6790–6802, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/35cf8659cfcb13224cbd47863a34fc58-Abstract.html.

Mario Geiger and Tess E. Smidt. e3nn: Euclidean neural networks. *CoRR*, abs/2207.09453, 2022. doi: 10.48550/ARXIV.2207.09453. URL https://doi.org/10.48550/arXiv.2207.09453.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL http://proceedings.mlr.press/v70/gilmer17a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/db8e1af0cb3aca1ae2d0018624204529-Abstract-round2.html.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *CoRR*, abs/2011.01938, 2020. URL https://arxiv.org/abs/2011.01938.

Shengli Jiang, Shiyi Qin, Reid C. Van Lehn, Prasanna Balaprakash, and Victor M. Zavala. Uncertainty quantification for molecular property predictions with graph neural architecture search. *CoRR*, abs/2307.10438, 2023. doi: 10.48550/ARXIV.2307.10438. URL https://doi.org/10.48550/arXiv.2307.10438.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL http://arxiv.org/abs/1609.02907.

Johannes Klicpera, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *CoRR*, abs/2011.14115, 2020. URL https://arxiv.org/abs/2011.14115.

Jonas M. Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 12661–12673, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/69adc1e107f7f7d035d7baf04342e1ca-Abstract.html.

Lukasz Maziarka, Tomasz Danel, Slawomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanislaw Jastrzebski. Molecule attention transformer. *CoRR*, abs/2002.08264, 2020. URL https://arxiv.org/abs/2002.08264.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *CoRR*, abs/1803.11173, 2018. URL http://arxiv.org/abs/1803.11173.

Chuizheng Meng, Sungyong Seo, Defu Cao, Sam Griesemer, and Yan Liu. When physics meets machine learning: A survey of physics-informed machine learning. *CoRR*, abs/2203.16797, 2022. doi: 10.48550/ARXIV.2203.16797. URL https://doi.org/10.48550/arXiv.2203.16797.

Albert Musaelian, Simon L. Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *CoRR*, abs/2204.05249, 2022. doi: 10.48550/ARXIV.2204.05249. URL https://doi.org/10.48550/arXiv.2204.05249.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Blai Bonet and Sven Koenig (eds.), *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 2901–2907. AAAI Press, 2015. doi: 10.1609/AAAI.V29I1.9602. URL https://doi.org/10.1609/aaai.v29i1.9602.

Jigyasa Nigam, Guillaume Fraux, and Michele Ceriotti. Unified theory of atom-centered representations and graph convolutional machine-learning schemes. *CoRR*, abs/2202.01566, 2022. URL https://arxiv.org/abs/2202.01566.

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *CoRR*, abs/2208.09481, 2022. doi: 10.48550/ARXIV.2208.09481. URL https://doi.org/10.48550/arXiv.2208.09481.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/94aef38441efa3380a3bed3faf1f9d5d-Abstract.html.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL http://proceedings.mlr.press/v139/satorras21a.html.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 991–1001, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Kristof Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9377–9388. PMLR, 2021. URL http://proceedings.mlr.press/v139/schutt21a.html.

Alain B. Tchagang and Julio J. Valdés. Prediction of the atomization energy of molecules using coulomb matrix and atomic composition in a bayesian regularized neural networks. In Igor V. Tetko, Vera Kurková, Pavel Karpov, and Fabian J. Theis (eds.), *Artificial Neural Networks and Machine Learning - ICANN 2019 - 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019, Proceedings - Workshop and Special Sessions*, volume 11731 of *Lecture Notes in Computer Science*, pp. 793–803. Springer, 2019. doi: 10.1007/978-3-030-30493-5_75. URL https://doi.org/10.1007/978-3-030-30493-5_75.

Oliver T. Unke, Stefan Chmiela, Michael Gastegger, Kristof T. Schütt, Huziel E. Sauceda, and Klaus-Robert Müller. Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *CoRR*, abs/2105.00304, 2021. URL https://arxiv.org/abs/2105.00304.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. Moleculenet: A benchmark for molecular machine learning. *CoRR*, abs/1703.00564, 2017. URL http://arxiv.org/abs/1703.00564.

Jun Xia, Lecheng Zhang, Xiao Zhu, Yue Liu, Zhangyang Gao, Bozhen Hu, Cheng Tan, Jiangbin Zheng, Siyuan Li, and Stan Z. Li. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/cc83e97320000f4e08cb9e293b12cf7e-Abstract-Conference.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 28877–28888, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f1c1592588411002af340cbaedd6fc33-Abstract.html.

Peng-Fei Zhou, Rui Hong, and Shi-Ju Ran. Automatically differentiable quantum circuit for many-qubit state preparation. *CoRR*, abs/2104.14949, 2021. URL https://arxiv.org/abs/2104.14949.

A Quantum Module Details

A.1 Structural-Prior Rationale

Our formulation is intentionally independent of specific quantum hardware and instead focuses on structural properties that can be represented and efficiently accessed through restricted circuits. Restricting representations to those realizable by shallow circuits filters the hypothesis space through circuit depth, entanglement structure, and observable selection. The resulting measurement functions exhibit structured correlation patterns rather than unrestricted dependencies.

Several prior classes arise from this view. Molecular Hamiltonians obey symmetries that include permutation invariance of identical atoms and rotational symmetry in three-dimensional space. Appropriate input encodings and measurements preserve these symmetries and reduce representational redundancy. Restricted entangling patterns also bias representations toward sparse correlation structure among atomic degrees of freedom. Finally, tensor-network simulation exploits restricted correlations to represent circuit evolution compactly (Burgholzer et al., 2022), while differentiable circuit-based state preparation provides a related learnable low-complexity representation (Zhou et al., 2021). These properties motivate circuit-induced representations as low-complexity global summaries. They do not imply that the circuit represents an electronic wavefunction.

A.2 Classical Encoder Inputs

Node attributes include an embedding of atomic number and, when available, formal charge and aromaticity. Edge attributes encode either bond types or distance-based neighborhoods constructed from 3D coordinates. With 3D coordinates, relative displacement vectors and radial distance features provide local spatial context. Learned edge-conditioned transformations and permutation-invariant aggregation ensure invariance to node ordering and stability across molecular sizes. Mean pooling isolates the quantum module as an explicitly global component and avoids coupling circuit input size to graph cardinality.

A.3 Restricted Trigonometric Feature Map

An expectation of any measured observable O is equivalently a finite sum of products of $\sin(a_i)$ and $\cos(a_i)$. Moreover, the coefficient of a frequency vector is zero whenever its support lies outside the backward light cone of O. With one nearest-neighbor ring-CZ entangler per circuit layer,

$$|\mathcal{C}(Z_i)| \le \min(n_q, 2L_q + 1), \qquad |\mathcal{C}(Z_i Z_{i+1})| \le \min(n_q, 2L_q + 2). \tag{3}$$

Consequently, a depth-L_q observable contains interactions among only finitely many encoded coordinates, and increasing depth expands the interaction order in a controlled manner. For $n_q = 8$ and $L_q = 3$, the light cones can summarize global pooled information but remain tied to the sparse ring topology. Single encoding excludes higher input harmonics, shallow depth limits the degree and spatial support of cross-coordinate products, the ring excludes arbitrary all-to-all couplings, and fixed one- and two-site observables prevent a learned readout from accessing every amplitude.

A.4 Feature Fusion

The quantum feature vector $\mathbf{z}_q$ is concatenated with the final pooled GNN representation $\mathbf{g}$:

$$\mathbf{f} = \mathrm{ReLU}(\mathrm{LN}\left(\mathbf{W}_f[\mathbf{g}; \mathbf{z}_q] + \mathbf{b}_f\right)), \qquad \mathbf{f} \in \mathbb{R}^{256}. \tag{4}$$

Layer normalization (Ba et al., 2016) prevents the bounded quantum coordinates from being dominated by the scale of the classical embedding. Concatenation retains the purely classical path, allowing the model to attenuate uninformative quantum features without destroying local geometric information. The task head has two 256-dimensional hidden layers with ReLU activations and is fixed across tasks up to its output dimension.

A.5 Optimization Details

For uncertainty-aware scalar regression, the prediction head returns a mean $\widehat{y}$ and raw scale r, with $\sigma = \mathrm{softplus}(r) + 10^{-6}$. The loss is

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2B} \sum_{j=1}^{B} \left[\frac{(y_j - \widehat{y}_j)^2}{\sigma_j^2} + 2 \log \sigma_j \right]. \tag{5}$$

Gradients pass through the prediction head and fusion projection to both branches. $\partial \mathcal{L} / \partial \mathbf{z}_q$ is differentiated through the exact state-vector operations, while $\partial \mathcal{L} / \partial \mathbf{a}$ continues through the bounded encoder, mean pooling, and message-passing layers. We use reverse-mode automatic differentiation rather than finite shots or finite differences. Unless otherwise specified, all parameters are optimized jointly with Adam (Kingma & Ba, 2015), learning rate 10^{-3}, batch size 64, and at most 300 epochs. Removing the quantum branch leaves the same graph encoder; in parameter-matched controls, only the head width is adjusted as specified in the experimental protocol.

A.6 Simulation Complexity

Let n_q be the number of qubits, L_q the circuit depth, and B the molecular batch size. The circuit has $G = n_q(1 + 3L_q)$ elementary gates: n_q encoding rotations and, per layer, $2n_q$ trainable rotations plus n_q CZ gates. Applying a one- or two-qubit gate to a dense state vector costs $\mathcal{O}(2^{n_q})$, and all $d_q = 2n_q - 1$ diagonal observables can be accumulated from the final basis probabilities. A forward circuit batch therefore costs

$$\mathcal{O}(B(G + d_q)2^{n_q}) = \mathcal{O}(BL_q n_q 2^{n_q}) \tag{6}$$

time and $\mathcal{O}(B2^{n_q})$ state memory. Reverse-mode differentiation has the same asymptotic time with a constant-factor backward pass; straightforward automatic differentiation stores $\mathcal{O}(BG2^{n_q})$ intermediate amplitudes, reducible to $\mathcal{O}(B2^{n_q})$ by recomputation or checkpointing. The exponential dependence is explicit, but at the fixed $n_q = 8$ used here the circuit state has only 256 complex amplitudes per molecule, making exact simulation a small, deterministic component of the full graph-learning pipeline.

B Additional Representation and Method Details

B.1 Representation Assumptions

A molecular representation defines a hypothesis space before the downstream predictor is fitted. Geometric symmetries restrict that space, but do not uniquely determine an appropriate representation for a particular chemical target. Under scarce supervision or chemical shift, learning global correlations entirely from data can be difficult. Our circuit construction supplies an additional restriction on the latent feature map, complementing geometric architectures (Bronstein et al., 2021; Geiger & Smidt, 2022). Its usefulness is an empirical question rather than a consequence of quantum terminology.

The encoding uses pooled graph features, not an electronic Hamiltonian or a molecular wavefunction. Permutation invariance follows from invariant graph aggregation. Rotational invariance requires the pooled input to be invariant; applying a quantum circuit to arbitrary rotation-dependent coordinates would not supply that property. The circuit topology controls interactions among encoded coordinates, which need not correspond one-to-one to atoms or bonds. Claims about physically meaningful global structure therefore refer to an architectural motivation and observed predictive behavior, rather than recovery of electronic correlations.

The broader motivation is to constrain a data-driven representation with a tractable family of measurement functions. At the selected circuit size these functions admit exact classical simulation. Limited depth, sparse entanglement, and fixed readouts constrain the feature family, but do not independently prove better sample complexity, a low effective rank, or robust extrapolation. Structural-prior rationale and light-cone bounds are provided in Appendices A.1 and A.3; predictive comparisons and matched classical controls assess whether the chosen restrictions help in the reported protocols.

B.2 Gate Matrices and Initialization

The trainable layer applies a Y rotation followed by a Z rotation on each qubit, then the ring-CZ entangler. With the usual Pauli matrices, the rotations are

$$R_Y(\beta) = e^{-\mathrm{i}\beta Y/2} = \begin{bmatrix} \cos(\beta/2) & -\sin(\beta/2) \\ \sin(\beta/2) & \cos(\beta/2) \end{bmatrix}, \\ R_Z(\gamma) = e^{-\mathrm{i}\gamma Z/2} = \mathrm{diag}(e^{-\mathrm{i}\gamma/2}, e^{\mathrm{i}\gamma/2}). \tag{7}$$

We initialize each trainable angle $\theta_{\ell i}, \phi_{\ell i}$ independently from $\mathcal{U}[-0.1, 0.1]$. The input encoder $a_i = \pi(\tanh \alpha_i + 1)$ maps finite inputs to $(0, 2\pi)$ and has derivative $\pi \operatorname{sech}^2 \alpha_i$, bounded by π. Bounding the angles excludes repeated full rotations of an unbounded affine encoding, although saturation can reduce the encoder gradient. No data re-uploading occurs between trainable layers. Initialization and bounded angles are implementation choices, not guarantees against barren plateaus or other optimization failures.

B.3 Measurement Interpretation

For computational-basis probabilities $p(\mathbf{b}) = |\langle \mathbf{b} \mid \psi \rangle|^2$,

$$\langle Z_i \rangle = 1 - 2\Pr(b_i = 1), \qquad \langle Z_i Z_{i+1} \rangle = 1 - 2\Pr(b_i \neq b_{i+1}). \tag{8}$$

Thus the first block measures marginal imbalance and the second measures agreement of adjacent encoded coordinates. Each coordinate lies in $[-1, 1]$. The eight local terms and seven open-chain pair terms give $2n_q - 1 = 15$ features. The ring entangler includes the $(8, 1)$ boundary, while the selected readout does not include $\langle Z_8 Z_1 \rangle$. These observables are correlation summaries, not entanglement witnesses. Exact noiseless evaluation uses `shots=None`, so fixed parameters and inputs yield deterministic expectations up to floating-point precision. The backpropagation path and the exponential simulation cost are specified in Appendices A.5 and A.6.

B.4 Evaluation Scope and Limitations

Technical scope. We evaluate shallow eight-qubit circuits with an exact differentiable state-vector simulator, without finite-shot sampling or hardware noise; the results therefore support circuit-induced features as a computational prior, not quantum advantage or deployment-ready hardware performance. Exact simulation scales exponentially, the selected Z/ZZ observables may miss dependencies needed by larger systems, and outcomes depend on the encoder, circuit ansatz, fusion rule, and optimization. Scaling may require selective encodings, approximate simulation, or hardware estimation, each with renewed trainability and cost studies. Hardware use would also require validation under sampling variance, limited connectivity, compilation, gate and readout errors, and device drift. QM9, MD17, PCQM4Mv2, and OC20 cover several few-shot and distribution-shift settings but not the full diversity of chemical space, reaction pathways, conformational change, or noisy experimental measurements. Five fixed seeds support controlled comparison but cannot eliminate sensitivity to alternative splits, preprocessing, or baseline implementations. Reproduction should preserve sample identifiers, splits, seeds, configurations, checkpoints, software versions, optimization traces, predictions, and uncertainty estimates.

C Detailed Experimental Protocols

This section specifies the benchmark protocols summarized in the main text. The additional diagnostic records have distinct provenance and are not pooled with these benchmark measurements.

Metrics and Units. For N predictions, we report

$$\mathrm{MAE} = \frac{1}{N}\sum_j |y_j - \widehat{y}_j|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_j (y_j - \widehat{y}_j)^2}. \tag{9}$$

For regression calibration, examples are sorted by predicted standard deviation and partitioned into $K = 15$ equal-frequency bins. With s_y denoting the training-set target standard deviation, we use

$$\mathrm{ECE} = \frac{1}{s_y} \sum_{k=1}^{K} \frac{|B_k|}{N} \left| \sqrt{\frac{1}{|B_k|} \sum_{j \in B_k} (y_j - \widehat{y}_j)^2} - \sqrt{\frac{1}{|B_k|} \sum_{j \in B_k} \sigma_j^2} \right|. \tag{10}$$

This regression ECE follows the usual reliability-gap principle (Naeini et al., 2015; Guo et al., 2017) but compares empirical residual scale with predicted scale, rather than class confidence with accuracy. For nominal central confidence $c \in \{0.1, 0.2, \ldots, 0.9\}$, empirical coverage is $\widehat{C}(c) = N^{-1} \sum_j \mathbb{1}[y_j \in I_j(c)]$. We summarize an entire coverage curve by $\mathrm{CovMAD} = 9^{-1} \sum_c |\widehat{C}(c) - c|$ and $\mathrm{MaxDev} = \max_c |\widehat{C}(c) - c|$; coverage at $c = 0.9$ is reported separately. P90 denotes the 90th percentile of per-example absolute error. For QM9 and PCQM4Mv2, energy-related properties and the HOMO-LUMO gap are reported in eV, while the dipole moment μ is reported in Debye. For MD17, energy MAE is reported in meV and force MAE in meV/Å. For OC20 (Chanussot et al., 2020), energy MAE is reported in eV following the official benchmark protocol. Lower values indicate better performance (↓).

Implementation Details. We use the Adam optimizer (Kingma & Ba, 2015) with learning rate 10^{-3} and no learning-rate tuning per method. Batch sizes are 64 for QM9, 32 for MD17, 256 for PCQM4Mv2, and 32 for OC20. QM9 and MD17 run for at most 300 epochs; PCQM4Mv2 and full OC20 run for at most 100 epochs. We stop when validation MAE has not improved by at least 10^{-5} for 30 epochs (10 epochs for the two large-scale datasets) and restore the lowest-validation-MAE checkpoint. Figure 10(a) is the sole exception: it records all 200 epochs without early termination so that convergence trajectories remain comparable. Inputs and targets are standardized using training statistics only. The benchmark experiments use NVIDIA A100 GPUs; the eight-qubit module is evaluated by an exact differentiable state-vector simulator with no measurement shots.

Datasets and Protocols. We use one fixed master split (seed 2027) and model/subsampling seeds $\{1, 2, 3, 4, 5\}$. For QM9 (Wu et al., 2017), preprocessing retains 130,000 valid molecules, partitioned into 104,000/13,000/13,000 train/validation/test examples. The primary energy target is internal energy at 0 K; the dipole moment and HOMO-LUMO gap are secondary targets. Full-data columns in Table 1 use all 104,000 training molecules. Few-shot subsets of size $\{50, 100, 200, 500, 1000, 5000\}$ are nested within that training pool for each model seed, while validation and test indices remain fixed. To obtain a reproducible structural-shift diagnostic, we compute each test molecule's maximum Tanimoto similarity to the 100-shot training set using radius-2, 2,048-bit Morgan fingerprints; the lowest-similarity 20% (2,600 molecules) is the OOD subset used in Figure 5(a) and Table 4. It is never used for fitting or calibration.

For MD17 (Chmiela et al., 2019), Table 2 uses the aspirin (ASP) and malondialdehyde (MDA) trajectories, with 1,000 training frames, 1,000 validation frames, and 10,000 test frames per molecule sampled without replacement after the master-seed shuffle. We predict potential energy and Cartesian forces jointly and average the standardized energy and force MAE losses. For PCQM4Mv2 (Hu et al., 2021), the target is the DFT HOMO-LUMO gap and we train on the official 3,378,606 training molecules; because public test labels are unavailable, Table 3 reports the official 73,545-molecule validation MAE. For OC20 (Chanussot et al., 2020), we use the IS2RE task, whose target is relaxed adsorption energy, and train on all 460,328 official training structures. Table 3 uses the complete ID, OOD-Adsorbate, and OOD-Catalyst validation splits. Figure 6 is a separate controlled diagnostic: every method is trained on the same 100,000-example subset and evaluated on 20,000 examples per displayed split and seed. This reduced-data protocol must not be compared numerically with the full-data OC20 rows in Table 3.

Uncertainty and Regression Calibration. Each uncertainty-enabled run uses the two-output heteroscedastic head described above and produces $(\widehat{y}_j, \sigma_j)$. After training, we fit one positive scalar τ on the validation set by minimizing Gaussian NLL with $\widehat{y}$ and all network weights frozen. The calibrated scale is $\sigma_j^{\mathrm{cal}} = \tau \sigma_j$, and a central interval is $I_j(c) = [\widehat{y}_j - q_c \sigma_j^{\mathrm{cal}}, \widehat{y}_j + q_c \sigma_j^{\mathrm{cal}}]$, where $q_c = \Phi^{-1}((1 + c)/2)$. The same τ is frozen for ID and OOD tests, so OOD labels never influence interval width. Table 1 reports pre-calibration full-data ECE, whereas Figure 5(b) reports post-calibration ECE for the 100-shot BaseGNN and our model; the two quantities therefore have different protocols and are not intended to coincide.

Repetition and Statistical Analysis. Every benchmark model is trained independently with seeds $\{1, 2, 3, 4, 5\}$. We compute a test metric separately for each seed and report its arithmetic mean and sample standard deviation (denominator $S - 1$). A two-sided 95% confidence interval for a mean is $\bar{x} \pm t_{0.975,S-1} s/\sqrt{S}$; for $S = 5$, the multiplier is 2.776. Model comparisons use seed-paired differences, an exact two-sided sign-permutation test over all 2^5 sign assignments, and Holm correction across the primary endpoints within each dataset. Per-example bootstrap intervals are used only for CDF quantiles and are formed from 10,000 stratified resamples within each seed before pooling the five seed estimates.

Parameter Matching and Timing. Controlled comparisons keep graph layers, hidden dimension, pooling, optimizer, and data order fixed. The BaseGNN head is widened, and the classical projection and low-rank controls are dimensioned, so that each differs from the full model's trainable parameter count by at most 1%; frozen quantum angles are not counted as trainable parameters. Benchmark baselines retain their standard published architecture and are therefore not claimed to be exactly parameter matched. Runtime is measured after 50 warm-up optimization steps, followed by 200 timed steps, and repeated 10 times with identical batches. We synchronize the GPU immediately before and after each timed block and report the repeat mean, sample standard deviation, and Student-t 95% interval; peak allocated GPU memory is reset and recorded for each repeat.

Raw Logs. The benchmark recording schema includes the dataset and target, split and sample identifiers, master and model seeds, epoch and global step, batch size, learning rate, train/validation loss, MAE, RMSE, raw and calibrated ECE, coverage at every nominal level, CovMAD, MaxDev, y, $\widehat{y}$, raw σ, calibrated σ, interval endpoints, wall-clock time, throughput, peak GPU memory, checkpoint path, trainable parameter count, n_q, L_q, observable set, simulator backend, software environment, source commit, and a hash of the complete configuration. These fields are sufficient to reconstruct the benchmark aggregates, error bars, coverage curves, and paired tests.

Baselines. Graph baselines are MPNN (Gilmer et al., 2017), GCN (Kipf & Welling, 2016), GIN (Xu et al., 2019), and SchNet (Schütt et al., 2017). Geometric baselines are SE(3)-GNN (Geiger & Smidt, 2022) and PaiNN (Schütt et al., 2021), together with DimeNet++ (Klicpera et al., 2020). The remaining neural baselines are Graph Transformer (Ying et al., 2021), 3D-aware Molecule Transformer (Maziarka et al., 2020), PhysNet (Reiser et al., 2022), SpookyNet (Unke et al., 2021), and SOAP + KRR (Nigam et al., 2022). We also evaluate Coulomb Matrix + KRR (Tchagang & Valdés, 2019). All neural baselines are trained with the protocol above; exact parameter matching is applied only to controlled BaseGNN variants, while established benchmark models retain their standard configurations.

D Supplementary Experiments

D.1 Cross-Benchmark Overview

Figure 14 summarizes the 15 metrics in Tables 1-3. Ours has the lowest displayed value on 14 metrics and ties the best QM9 energy MAE at 0.021 eV. Among the improved metrics, reductions relative to the best available baseline range from 1.6% for the QM9 gap to 33.3% for run-to-run standard deviation; they are 15.7% for 100-shot MAE, 19.8% for ECE, and 3.0% for PCQM4Mv2. The largest relative differences therefore occur in limited-data accuracy, calibration, and variation across runs. The heatmap scales colors separately within each metric; its columns retain their own units and training protocols, and missing entries are not interpreted as zero errors.

a QM9 performance

	Energy MAE	Energy RMSE	100-shot MAE	ECE	Run SD	Dipole MAE	Gap MAE
MPNN	0.026	0.046	0.083	0.142	0.004	0.037	0.068
SchNet	0.026	0.047	0.081	0.142	0.004	0.037	0.066
DimeNet++	0.022	0.041	0.074	0.118	0.003	0.033	0.062
PaiNN	0.021	0.040	0.071	0.115	0.003	0.032	0.061
Graph Transformer	0.025	0.045	0.078	0.127	0.005	0.034	0.065
PhysNet	0.021	0.039	0.070	0.111	0.003	0.031	0.061
SOAP + KRR	0.028	0.050	0.085	0.136	0.004	0.038	0.067
Ours	**0.021**	**0.038**	**0.059**	**0.089**	**0.002**	**0.030**	**0.060**

b Cross-benchmark performance

	ASP E	ASP F	MDA E	MDA F	PCQM gap	OC20 ID	OC20 OOD-A	OC20 OOD-C
MPNN	--	--	--	--	0.128	0.56	0.71	0.77
SchNet	0.18	0.42	0.12	0.33	0.123	0.52	0.66	0.71
DimeNet++	0.15	0.36	0.10	0.28	--	--	--	--
SE(3)-GNN	--	--	--	--	0.114	0.49	0.63	0.68
PaiNN	0.14	0.34	0.09	0.26	0.103	0.45	0.58	0.63
PhysNet	0.14	0.33	0.09	0.25	0.102	0.46	0.59	0.64
SpookyNet	--	--	--	--	0.101	0.45	0.58	0.62
Ours	**0.12**	**0.30**	**0.08**	**0.23**	**0.098**	**0.43**	**0.54**	**0.59**



Best in column — Worst in column

Colors scale within each column; labels retain original units.
Energy/gap: eV; MD17 E/F: meV and meV/Å; dipole: Debye.

Figure 14: Performance across molecular benchmarks.

D.2 Factorial Prior Comparison

Figure 8(a) compares all eight combinations of geometry (G), correlation (C), and hierarchy (H), using three training fractions and five seed-indexed records per setting. At the 100% setting, G, C, and H alone reduce ID MAE from 24.42 meV to 22.29, 23.34, and 23.62 meV; G has the largest individual decrease in this comparison (8.7%). G+C reaches 20.80 meV, and G+C+H reaches 19.41 meV, a 20.5% reduction from no priors. On OOD, the full combination reduces 30.44 to 25.11 meV (17.5%). At the 10% setting, the corresponding changes are 27.88 to 23.28 meV on ID and 34.47 to 28.49 meV on OOD. Panel (b) shows lower errors for the full combination in each of the five matched entries at 100%. The full combination has the lowest mean in all six tabulated settings.

D.3 Representation and Gradient Diagnostics

Representation Structure and Optimization. Figure 9(a-b) summarizes 500 vectors with 15 standardized features. Within the three displayed five-feature blocks, mean absolute correlation is 0.766, 0.750, and 0.728, versus 0.063 between blocks. The first three correlation-matrix components explain 29.6%, 27.3%, and 23.3% of variance (80.3% cumulatively), with entropy effective rank 6.34. In panel (c), gradient norms at depths $1/2/4/8$ change from $0.2210/0.1766/0.1480/0.1161$ at epoch zero to $0.0358/0.0373/0.0371/0.0510$ at epoch 160, reductions of $83.8\%/78.9\%/74.9\%/56.0\%$. The stored features therefore combine strong within-group correlations with several substantial variance directions, and the gradient trajectories depend on depth. These correlations do not establish quantum entanglement, and the finite trajectories do not establish favorable gradient scaling at untested sizes.

E Extended Numerical Analysis

The following analysis retains the detailed comparisons behind the compact main-text discussion. Figures and tables retained in the main text are referenced rather than repeated here.

QM9 Accuracy and Data Efficiency. Table 1 averages five seeds on the same 13,000-molecule test set, but its columns cover two training regimes. Energy MAE/RMSE, ECE, Run Std, dipole moment, and gap use 104,000 training molecules; MAE@100 uses 100. Ours ties the best full-data energy MAE at 0.021 eV. Compared with PhysNet, RMSE decreases from 0.039 to 0.038 eV (2.6%), dipole MAE from 0.031 to 0.030 Debye (3.2%), and gap MAE from 0.061 to 0.060 eV (1.6%). The 100-shot MAE decreases from 0.070 to 0.059 eV (15.7%). Full-data ECE decreases from 0.111 to 0.089 (19.8%), and Run Std from 0.003 to 0.002 (33.3%). Thus, the larger relative gains concern scarce supervision and reliability; the full-data point-error gains are comparatively small and are not claimed to be statistically significant. Figure 5 uses 100-shot checkpoints in panels (a-b) and nested training subsets of size 50-5,000 in panel (c); ID is the fixed random-split test set, OOD is its lowest-similarity 20% subset, and all curves and error bars aggregate five seeds. On the separate post-calibrated comparison in panel (b), BaseGNN ("Base" in panel (b)) and our model obtain ECE 0.0294 and 0.0084, respectively, a 71.4% reduction. Figure 5(c) further shows that the gain is largest in the most data-scarce regime: mean MAE changes from 0.1182 to 0.1018 at 50 samples and from 0.0701 to 0.0590 at 100 samples, while the corresponding values converge to 0.0323 and 0.0301 at 5,000 samples. This pattern indicates a larger absolute benefit in the low-data regime than at the largest displayed training size.

Using the validation-fitted variance scale frozen for both test splits, Figure 5(a) evaluates central prediction intervals from the same 100-shot checkpoints. On the 13,000-example ID test and its 2,600-example low-similarity OOD subset, our method reduces CovMAD from 0.3453 to 0.1335 and from 0.4691 to 0.2584, respectively, corresponding to 61.3% and 44.9% relative reductions. Similarly, MaxDev decreases from $0.4910/0.8203$ (ID/OOD) to $0.2868/0.5133$. At a representative nominal confidence of 0.9, the empirical coverage improves from 0.4814 to 0.8948 (ID) and from 0.0798 to 0.9429 (OOD). However, at nominal confidence 0.5, our coverage is only 0.2644 on ID and 0.0401 on OOD. The improvement is therefore uneven across confidence levels and does not imply that the entire coverage curve is well calibrated.

Molecular Dynamics Generalization. Table 2 reports five-seed MD17 results in meV for energy and meV/Å for forces. With 1,000 training, 1,000 validation, and 10,000 test frames per trajectory, the proposed method improves both energy and force prediction accuracy, achieving $0.12/0.30$

(ASP-E/F) and $0.08/0.23$ (MDA-E/F), compared to $0.14/0.33$ and $0.09/0.25$ for PhysNet (Table 2). These correspond to energy/force reductions of $14.3\%/9.1\%$ on aspirin and $11.1\%/8.0\%$ on malondialdehyde. The reported mean improvements cover both targets and trajectories; tail errors require a separate distributional analysis.

Large-Scale and OOD Results. Table 3 reports five-seed validation results after training on the complete official PCQM4Mv2 and OC20 training splits, containing 3,378,606 molecules and 460,328 structures, respectively. The proposed method reduces PCQM4Mv2 MAE from the best listed baseline, SpookyNet at 0.101, to 0.098 eV (3.0%). Its OC20 errors are 0.43 on ID, 0.54 on OOD-Adsorbate, and 0.59 on OOD-Catalyst, compared with the best baseline's 0.45, 0.58, and 0.62. The corresponding relative reductions are 4.4%, 6.9%, and 4.8%, with the largest improvement on OOD-Adsorbate.

Figure 6 addresses a different question under its parameter-controlled 100,000-example training subset. Labels Ads. and Cat. denote OOD-Adsorbate and OOD-Catalyst. In panel (a), BaseGNN obtains 0.581, 0.820, and 0.908 on ID, OOD-Adsorbate, and OOD-Catalyst; the corresponding values for our method are 0.539, 0.708, and 0.779. Beyond average MAE, the error distribution shifts left and degrades more gracefully (Figure 6(b-c)): median/P90 decreases from $0.492/1.197$ to $0.456/1.114$ on ID and from $0.768/1.866$ to $0.659/1.607$ on OOD-Catalyst, and the hardest-bucket mean MAE drops from 0.8797 to 0.7574 (13.9%). The mean, central quantile, and upper-tail comparisons all favor Ours within this reduced-data protocol.

E.1 Ablation Analysis

Quantum Feature Interpretability. Figure 7(a) reports correlations between the 15 measured features and energy on the QM9 100-shot test set. The largest absolute correlation is 0.2148 for Z_0Z_1, followed by 0.1878 for Z_2Z_3; the top five have mean absolute correlation 0.1600. These modest values show target-related variation across observables and support inspecting their joint contribution, without establishing a causal physical interpretation.

Structural and Quantum Components. In Figure 7(b), removing geometry, correlation, or hierarchy increases complete-test MAE from 0.059 to 0.066, 0.072, or 0.064 eV, respectively: degradations of 11.9%, 22.0%, and 8.5%. Removing correlation incurs the largest penalty in this leave-one-component-out comparison. Panel (c) gives 0.067 for frozen random circuit features, 0.065 for a classical bottleneck, and 0.071 without the quantum module, versus 0.059 for the full circuit. The full model therefore has the lowest reported error among these controls; these ablations alone do not identify statistically significant interactions between priors.

Quantum Circuit Module. Table 4 isolates the effect of the quantum circuit module on the low-similarity OOD subset of the QM9 100-shot test set. This deliberately harder 2,600-molecule subset explains why its 0.081 full-model MAE differs from the 0.059 complete-test MAE in Table 1 and Figure 7; the values are not alternative estimates of the same test metric. The full circuit reduces OOD MAE from the base GNN's 0.097 to 0.081 (16.5%). The classical MLP, low-rank bottleneck, and frozen-feature controls obtain 0.089, 0.086, and 0.090, so the full circuit improves on them by 9.0%, 5.8%, and 10.0%, respectively. Global compression thus improves over the base model, and the learnable circuit provides a further advantage over the listed alternatives within this protocol. Figure 10 reports 200-epoch validation trajectories in panel (a) and five-seed mean test MAE across qubit/depth settings in panel (b), jointly assessing convergence and circuit sensitivity. From Figure 10(a), the final validation MAE (mean over 5 seeds) is 0.0716 for the base GNN versus 0.0601 for our model, a relative reduction of 16.2%. The 0.0601 value is the final-epoch validation mean, not the complete-test MAE of 0.059 reported in Table 1 and Figure 7. Moreover, reaching a target error $\mathrm{MAE} < 0.08$ requires 80 epochs for the base GNN but only 38 epochs for our method (about $2.1\times$ faster in terms of epochs-to-threshold). Figure 10(b) spans test MAEs of 0.0582-0.0685 across the 16 configurations. At eight qubits, depths $1/2/3/4$ yield $0.0619/0.0614/0.0582/0.0592$ eV. Depth three is best in the displayed grid; adding a fourth layer does not improve this result. Faster convergence and the non-monotonic depth response favor a moderate configuration within the tested range.

E.2 Further Analysis of the Diagnostic Records

Prediction and Uncertainty Ranking. Figure 11 summarizes 1,500 prediction records per model and domain in scaled units. Panels (a-b) show predictions against targets and their marginal distributions. At full coverage, Ours has ID/OOD MAEs of $0.263/0.312$, compared with $0.297/0.343$ for BaseGNN, corresponding to reductions of $11.4\%/9.1\%$. Retaining the 50% of predictions with the lowest predicted uncertainty in panel (d) lowers these errors to $0.221/0.259$, versus $0.258/0.300$ for BaseGNN. For Ours, this selection reduces MAE by 15.9% on ID and 16.8% on OOD relative to full coverage. Across the eight uncertainty bins in panel (c), its first-to-last-bin mean absolute residual changes from 0.178 to 0.343 on ID and from 0.224 to 0.412 on OOD. The supplied records thus support useful error ranking and selective prediction, although the bin sequence on ID is not strictly monotonic and does not establish probabilistic calibration.

Accuracy and Computational Cost. Figure 12 compares 19 configurations with ten timing entries per configuration. The supplied BaseGNN, MLP, and low-rank records have MAEs of 20.50, 18.20, and 16.60 meV at mean step times of 12.49, 16.90, and 22.17 ms. Increasing the eight-qubit configuration from depth one to four reduces its tabulated MAE from 14.32 to 13.31 meV (7.0%), while mean step time increases from 39.67 to 78.69 ms and memory from 1.93 to 3.58 GB. Ten qubits at depth four give the lowest listed MAE, 12.77 meV, with 244.65 ms, 10.60 GB, and mean throughput of 4.10 steps/s. Relative to eight qubits at depth four, that is a further 4.1% error reduction at 3.11 times the step cost. These records illustrate diminishing accuracy returns with larger configurations. Dataset, device, and batch identifiers are absent from this table, so the comparison does not establish a measured hardware-efficiency advantage.

Molecule-Specific Residuals. Figure 13 compares 100 absolute-residual records per molecule, model, and target in scaled units; these units are not directly comparable to the physical-unit errors in Table 2. For aspirin energy, Ours reduces the mean residual from 0.426 to 0.417 and P90 from 0.959 to 0.884 relative to BaseGNN. For malondialdehyde energy, the corresponding values decrease from 0.354 to 0.322 and from 0.702 to 0.562. On aspirin forces, Ours has the lowest mean and median, 0.401 and 0.313, although low-rank has the lower P90, 0.780 versus 0.867. The trend reverses for malondialdehyde forces: Ours has a mean of 0.481 versus 0.365 for BaseGNN and a P90 of 1.022 versus 0.714. Thus, average error is lower on three targets in these records, while force accuracy and tail behavior depend on the molecule and comparator. Individual frame records are not additional independent training seeds.

F Extended Related Work

Geometric architectures include NequIP, Allegro, SE(3)-Transformers, and EGNNs (Batzner et al., 2021; Musaelian et al., 2022; Fuchs et al., 2020; Satorras et al., 2021). Attention-based alternatives include Graphormer, GROVER, and MAT (Ying et al., 2021; Rong et al., 2020; Maziarka et al., 2020). These models provide the broader architectural context for the compact main-text discussion.

Molecular Representation Learning. Molecular representation learning is central to molecular property prediction and materials discovery, as the choice of representation largely determines the balance between expressivity, inductive bias, and generalization. Earlier pipelines relied on handcrafted descriptors combined with linear models or kernel methods, whereas modern approaches increasingly operate directly on molecular graphs and 3D geometries. Message-passing neural networks (MPNNs) (Gilmer et al., 2017) and geometry-aware variants have become dominant due to their ability to model bonded structure, local chemical environments, and higher-order geometric interactions. In particular, DimeNet++ (Klicpera et al., 2020), PaiNN (Schütt et al., 2021), and GemNet (Gasteiger et al., 2021) encode directional, angular, or rotational information to model local orientation and higher-body interactions. For interatomic potentials and quantum-mechanical observables, neural potentials achieve strong accuracy-efficiency trade-offs by learning energy surfaces and forces with physically motivated parameterizations and training objectives. Representative examples include SchNet (Schütt et al., 2017), SpookyNet (Unke et al., 2021), and PhysNet (Reiser et al., 2022). At the same time, kernel methods and Gaussian process-style predictors remain competitive in data-scarce regimes, largely because carefully designed descriptors and kernels impose strong inductive biases and regularization. SOAP-like atom-centered representations (Nigam et al., 2022) are analyzed within a unified representation framework. MoleculeNet (Wu et al., 2017) standardizes multiple molecular tasks, while molecular generalization studies report substantial sensitiv-

ity to data regimes and distribution shifts (Xia et al., 2023). This has motivated work on uncertainty estimation, calibration, and reliability for scientific decision making, including calibrated message-passing ensembles (Busk et al., 2021) and uncertainty-aware graph neural architecture search (Jiang et al., 2023).

Inductive Biases and Generalization Inductive biases such as symmetry, equivariance, locality, and physics-informed constraints are widely recognized as key drivers of sample efficiency and robustness under distribution shift in scientific machine learning (Bronstein et al., 2021; Geiger & Smidt, 2022). In molecular settings, incorporating structure can reduce hypothesis complexity, stabilize optimization, and improve transfer across chemical space, particularly when supervision is limited or noisy. A broad family of physics-informed approaches incorporates physical laws, conservation principles, or differential-equation constraints via regularization, constrained parameterizations, or loss design (Meng et al., 2022). Related ideas include enforcing energy conservation, equivariant force fields, or consistency relations between predicted energies and forces in atomistic learning. In contrast, we introduce inductive bias primarily at the representation level: quantum-circuit-induced structure constrains the space of admissible molecular representations in a data-driven yet physically motivated manner. This representation-centric view is complementary to physics-informed losses: it can be applied even when explicit governing equations are unavailable, and it can be integrated into diverse downstream learning objectives without changing the target-specific supervision.

Quantum-Inspired Methods Quantum machine learning explores parameterized circuits, feature maps, and hybrid optimization procedures. Variational quantum algorithms (VQAs) (Cerezo et al., 2020) provide a common framework for these models, while analyses of quantum learning emphasize the role of data-dependent inductive bias (Huang et al., 2020). Quantum kernels (Kübler et al., 2021) offer a distinct feature-space construction. A recurring theme is the tension between expressivity and trainability, including barren plateaus (McClean et al., 2018) in parameterized circuits. Our approach does not seek quantum advantage, and it does not require quantum hardware. Instead, we use restricted quantum circuits as structured feature extractors whose measurement statistics define a controlled family of invariants and global correlation summaries. These features are computed efficiently via classical simulation at the circuit scales considered, serving as an architectural prior that can augment standard molecular learning models rather than replacing them with a standalone quantum learner.

Low-Rank Structure and Representations Tensor-network simulation represents quantum states and circuit evolution through structured contractions, providing a concrete connection between circuit structure and compact classical simulation (Burgholzer et al., 2022). Differentiable quantum-circuit state preparation (Zhou et al., 2021) further shows how restricted quantum representations can be trained end to end. Inspired by these developments, we translate low-complexity quantum structure into inductive biases for molecular representation learning, rather than designing tensor-network solvers or simulation algorithms. Concretely, by restricting circuit depth and measurement operators, we bias representations toward globally structured, low-complexity correlation summaries while retaining compatibility with classical geometric backbones. This perspective complements work on robustness and distribution shift in molecular prediction (Xia et al., 2023), motivating quantum-circuit-induced structural priors as a practical mechanism for stabilizing extrapolation in data-driven molecular models.