SIMPLE MODELS CAN STILL COMPETE IN OUT-OF-DISTRIBUTION MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

This study began as a replication of BOOM, a recent benchmark for out-of-distribution (OOD) molecular property prediction (Antoniuk et al., 2025), and developed into a broader investigation of how model choice, evaluation design, and metrics affect conclusions about molecular OOD performance. We reproduce and extend BOOM using six models spanning linear, tree-based, message-passing, pretrained language, and 3D-equivariant approaches. A simple Elastic Net model based on 2D molecular descriptors performs competitively with Chemprop on the original property-based OOD tasks, showing that increased model complexity does not necessarily translate into better extrapolation. This conclusion remains qualitatively robust across repeated runs and after validation-based hyperparameter tuning of Chemprop. In contrast, Random Forest and XGBoost perform poorly on these tasks, illustrating that the ability to extrapolate beyond the training target range, rather than model complexity alone, strongly influences performance. We further show that BOOM's original OOD evaluation measures extrapolation to extreme property values rather than structural novelty. When we introduce structure-based OOD splits, all evaluated models perform substantially better and differences between them narrow. Finally, we identify a metric discrepancy in the original benchmark: values reported as R^2 correspond to squared Pearson correlation (ρ^2). Using the coefficient of determination instead reveals many strongly negative OOD scores and a more severe extrapolation problem. Together, these results show that simple models remain important baselines for molecular property prediction and that conclusions about OOD performance depend critically on the type of distribution shift, the models included, and the evaluation metric.

1 INTRODUCTION

Out-of-distribution (OOD) molecular property prediction is central to data-driven molecular discovery: models are often applied to compounds that differ from those used during training, and the relevant distribution shift can arise in molecular structure, target-property values, or both. BOOM (Antoniuk et al., 2025) provides a broad benchmark of this problem across ten molecular endpoints and a diverse collection of machine-learning models. Its principal conclusion—that reliable OOD molecular property prediction remains difficult—is important and is also supported by the experiments reported here.

This work began as an attempt to reproduce the BOOM evaluation and subsequently expanded in response to questions that arose during that process. Our purpose is therefore complementary to BOOM: we examine how three choices—the evaluation metric, the definition of OOD, and the set of baseline models—affect the conclusions drawn from the benchmark. We focus on the following observations.

First, the values reported as R^2 in the original BOOM OOD results are closely reproduced by the square of the Pearson correlation coefficient, ρ^2, rather than by the coefficient of determination. These quantities can differ substantially when predictions have systematic offset or scale error (Alexander et al., 2015). Recomputing the coefficient of determination produces many negative OOD values. Importantly, this correction primarily changes the magnitude and interpretation of performance rather than overturning BOOM's main conclusion; OOD prediction remains difficult under either metric.

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.

Second, we broaden the empirical comparison to six models: four models included in BOOM—Chemprop, Random Forest, GotenNet, and MolFormer—and two additional baselines, Elastic Net and XGBoost. On BOOM's property-tail split, Elastic Net is competitive with Chemprop, whereas Random Forest and XGBoost are poor extrapolators across all ten endpoints. This contrast suggests that the ability of a model class to extrapolate beyond the observed target range is an important part of what this split measures. We additionally repeat four models across three random seeds and perform validation-based Chemprop hyperparameter tuning on six endpoints to assess the robustness of the Elastic-Net–Chemprop comparison.

Third, we add a structure-based OOD experiment. The original BOOM split deliberately selects extreme target-property values; this is a useful form of extrapolation, but it is distinct from holding out structurally separated molecules. We therefore construct a structure-only split based on Morgan fingerprints, UMAP, and density-based clustering, inspired by recent work on structure-aware molecular data splitting (Guo et al., 2024; 2025). Under this particular split, all six evaluated models obtain high positive R^2 values and the differences between models narrow. We view the property-based and structure-based experiments as complementary rather than as competing definitions of a single "correct" OOD setting.

The overall message is deliberately modest: model comparisons in molecular OOD prediction can depend strongly on what is held out and how predictions are scored. Simple baselines remain informative, but different simple models behave very differently depending on whether the task requires target-space extrapolation.

Their practical value also extends beyond predictive accuracy. When a lightweight model reduces the cost per prediction, a fixed computational budget can support screening more compounds. Simple models can also be easier to interpret: descriptor-based linear models expose explicit feature contributions that can help inspect predictions and generate chemical hypotheses. These potential advantages motivate evaluating cost and interpretability alongside accuracy, rather than treating model complexity as an end in itself.

2 Background and evaluation setting

2.1 The BOOM benchmark

BOOM evaluates molecular property prediction on ten endpoints (Antoniuk et al., 2025): heat of formation (HoF) and crystal density from CSD-10k, and eight quantum-chemical properties from QM9—dipole moment μ, polarizability α, HOMO, LUMO, the HOMO–LUMO gap (GAP), zero-point vibrational energy (ZPVE), electronic spatial extent $\langle R^2 \rangle$, and heat capacity C_v. CSD-10k contains approximately 10,000 compounds and QM9 approximately 130,000 small molecules.

For each endpoint, BOOM fits a Gaussian kernel-density estimate to the scalar target values and assigns the lowest-density, most extreme observations to the OOD set. This gives approximately 1,000 OOD molecules for the CSD-10k endpoints and 10,000 for the QM9 endpoints. Five percent of the remaining molecules are used as an in-distribution (ID) test set and the rest for training. The procedure is repeated independently for every endpoint. BOOM reports ID performance and an OOD score computed separately for the lower and upper target tails and then averaged.

2.2 Property-space and structure-space OOD

We use *property OOD* (or Y-space OOD) for a shift defined by unusual target values and *structural OOD* (or X-space OOD) for a shift defined by molecular structure. Neither notion subsumes the other. Property extrapolation is directly relevant when the objective is to discover compounds with unusually high or low values of a property. Structural generalization is relevant when models are applied to new chemical regions or scaffold families. A benchmark can therefore be informative about one axis without directly measuring the other.

This distinction is particularly relevant to model comparison. Random Forest predictions are averages of training responses and therefore remain within the observed response range. Gradient-boosted trees are more flexible, but remain piecewise-constant functions of the input features and are not naturally suited to smooth trend extrapolation beyond regions defined by training splits. A

linear model does not share these restrictions. Consequently, a property-tail split can distinguish models partly by their target-space extrapolation behavior, even when their ability to generalize across molecular structures is similar.

2.3 R^2 AND ρ^2

For observations y_i and predictions $\hat{y}_i$, the coefficient of determination is

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{1}$$

It equals zero for a predictor that always returns the test-set mean and can be arbitrarily negative when prediction errors exceed that baseline. In contrast, the squared Pearson correlation

$$\rho^2 = \left(\frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \tag{2}$$

is bounded in $[0, 1]$ and is invariant to affine rescaling of the predictions. Thus, a model can have high ρ^2 despite substantial bias or scale error. This distinction has previously been emphasized for QSAR/QSPR evaluation (Alexander et al., 2015).

3 METHODS

3.1 DATASETS AND PROPERTY-BASED SPLITS

We use the same CSD-10k and QM9 endpoints and reproduce the BOOM property-based split procedure described above. All property-based OOD scores use BOOM's lower/upper-tail aggregation, but with the coefficient of determination computed by `scikit-learn`'s `r2_score`. We also report RMSE and compute ρ^2 where needed for comparison with the originally reported values.

3.2 STRUCTURE-BASED SPLIT

To probe structural rather than target-value shift, we construct a single structure-only OOD split for each dataset. We compute 2048-bit radius-2 Morgan fingerprints (Rogers & Hahn, 2010) and embed them in two dimensions with UMAP using $n_{\text{neighbors}} = 15$, $\text{min_dist} = 0.1$, Euclidean distance, spectral initialization, $\text{spread} = 1.0$, and random seed 42; remaining UMAP parameters use library defaults. We cluster the embedding using HDBSCAN with minimum cluster size 20. Clusters are ranked by their separation from the rest of the embedding: for each cluster, we compute the shortest distance from any cluster point to any non-cluster point using a KD-tree. We then hold out the most separated complete clusters until approximately 10% of molecules are assigned to the structural OOD set; all remaining clusters and HDBSCAN noise points form the training set. This heuristic follows the general motivation of UMAP-based molecular splitting (Guo et al., 2025), while using a density-based clustering criterion to avoid selecting a fixed Tanimoto threshold.

The split seed is independent of the model-training seeds, and the generated split is cached and reused across models. The structural split is intentionally presented as one complementary stress test rather than as a definitive benchmark: its absolute difficulty depends on the fingerprint, embedding, clustering, and holdout rule. Further implementation details are provided in Appendix A.2.

3.3 MODELS

We evaluate six models spanning different inductive biases and representations.

Elastic Net. RDKit 2D descriptors are cleaned by removing columns containing missing values and a set of highly correlated descriptors, then standardized. The pipeline selects the top 80 descriptors by univariate F-score and expands them to degree-2 interaction features (approximately 3,160 features). We cross-validate four `l1_ratio` values $(0.1, 0.5, 0.9, 1.0)$ and 30 log-spaced regularization strengths using three folds (Zou & Hastie, 2005).

Random Forest. We use 500 trees with `max_features=sqrt`, matching the original reproduction setup.

XGBoost. We add a second tree-based baseline (Chen & Guestrin, 2016). We use library defaults for the main boosting hyperparameters, with row and column subsampling both set to 0.8; no hyperparameter search is performed.

Chemprop. We use Chemprop v2 (Yang et al., 2019; Heid et al., 2024). The default configuration uses message-passing depth 3, hidden dimension 300, a two-layer feed-forward network with hidden dimension 300, batch normalization, batch size 1,024, maximum learning rate 10^{-3}, and up to 100 epochs with patience 10. A random 90/10 training/validation split is used for early stopping.

GotenNet. We re-implement the 3D-equivariant GotenNet model (Aykent & Xia, 2025). BOOM does not provide runnable GotenNet training code; we match its documented 50-epoch duration. The released reproduction implementation uses AdamW, a cosine learning-rate schedule, validation-based checkpoint selection, and RDKit ETKDGv3 conformers followed where possible by MMFF optimization. Because geometry provenance can materially affect geometry-derived QM9 targets, we separately compare our RMSE with BOOM's reported values in Appendix D.

MolFormer. We fine-tune the pretrained molecular language model MolFormer (Ross et al., 2022), which was pretrained on up to 1.1 billion SMILES. The released reproduction implementation fine-tunes the encoder and regression head end-to-end with LAMB and cosine scheduling. We use 30 fine-tuning epochs for HoF and Density and 20 epochs for the eight QM9 endpoints. In our code, the training script uses 30 epochs unless a different value is specified; the script that starts the QM9 experiments sets this value to 20. Appendix A.1 describes the implementation and identifies settings that could not be verified from the retained execution records. Appendix D compares our results with those reported by BOOM.

3.4 Seeds and Chemprop tuning

The six-model tables use seed 42, because GotenNet and MolFormer were run once. To assess stochastic variability for the central comparison, Elastic Net, Random Forest, XGBoost, and default Chemprop were also run with seeds 43 and 44. We summarize variability in Section 4.4 and give endpoint-level Chemprop results in Appendix B.

We additionally perform a validation-selected Chemprop search on six endpoints (HoF, Density, HOMO, α, μ, and C_v). For each endpoint, 50 random configurations vary message-passing depth (2–6), dropout (0–0.40), feed-forward depth (1–3 layers), and hidden size (300–2400). The reported configuration is selected using validation loss, without consulting OOD test performance. Four endpoints (LUMO, GAP, ZPVE, and $\langle R^2 \rangle$) did not complete this search within the available compute budget; we therefore report the partial tuning analysis explicitly rather than extrapolating from it.

4 Results

4.1 Metric reconstruction: R^2 and ρ^2 give different OOD magnitudes

Our independently computed ρ^2 values closely reproduce the pattern of the values reported as R^2 in BOOM, whereas the coefficient of determination frequently becomes negative on the property-tail OOD sets. The discrepancy is small on ID data: across our comparison, the maximum absolute difference between ID R^2 and ρ^2 is 0.05. It is much larger OOD. Among the 52 seed-42 model–endpoint cells for which both metrics were available in our direct comparison, 38 have negative OOD R^2, while ρ^2 is necessarily non-negative. Illustrative examples include Elastic Net on $\langle R^2 \rangle$ ($R^2 = -100.56$, $\rho^2 = 0.01$), Random Forest on LUMO (-8.44 vs. 0.02), and GotenNet on μ (-7.90 vs. 0.15).

The correction changes the apparent magnitude more than the ordering of models. Per-endpoint Spearman correlation between R^2 and ρ^2 rankings is at least 0.8 on eight of ten endpoints in our analysis, and Random Forest is the lowest-ranked model under both metrics on seven of ten. The largest ordering changes occur for μ and GAP. Thus, the metric reconstruction strengthens the evidence that absolute property-tail extrapolation is difficult, while leaving BOOM's broad qualitative conclusion intact.

Table 1: Property-based (Y-space) OOD binned R^2, seed 42. Higher is better. $\langle R^2 \rangle$ denotes the QM9 electronic spatial-extent target, not the R^2 metric.

Endpoint	Chemprop	Elastic Net	XGBoost	RF	GotenNet	MolFormer
HoF	0.27	0.68	-0.74	-2.42	0.62	-0.70
Density	0.48	0.54	-1.53	-6.67	0.95	-1.88
HOMO	-1.11	-2.30	-2.65	-5.12	-0.89	-2.54
LUMO	-2.32	-7.84	-6.42	-8.44	-2.21	-3.29
GAP	-1.98	-5.02	-3.84	-5.55	-1.99	-3.58
ZPVE	0.58	0.82	-0.82	-3.08	0.91	0.22
$\langle R^2 \rangle$	-2.08	-100.56	-0.82	-2.54	-0.43	-1.02
α	0.05	0.32	-0.66	-2.03	0.36	-0.74
μ	-7.03	-17.61	-13.46	-12.97	-7.90	-16.53
C_v	-0.23	0.83	-0.19	-2.46	0.80	-0.11
Positive R^2	4/10	5/10	0/10	0/10	5/10	1/10

4.2 Property-based OOD: extrapolation capacity matters

Table 1 reports seed-42 binned OOD R^2 for all six models. Elastic Net is higher than default Chemprop on five endpoints and lower on five. It has positive OOD R^2 on five endpoints, compared with four for Chemprop and five for GotenNet. Random Forest and XGBoost are negative on all ten endpoints. MolFormer is positive on one endpoint (ZPVE) despite competitive performance on the structural split described below.

These results should not be read as evidence that simple models uniformly outperform more complex ones. Rather, the comparison separates two issues. Elastic Net is simple but can extrapolate linearly beyond observed response values; Random Forest and XGBoost are also relatively simple baselines but are not naturally suited to smooth target-space extrapolation. GotenNet is the strongest model on several endpoints. Together, these results show that a well-chosen simple baseline can remain competitive, while highlighting the importance of matching a model's extrapolation capabilities to the distribution shift being evaluated.

The original three-model RMSE comparison leads to the same Chemprop–Elastic-Net head-to-head pattern as R^2: Elastic Net has lower OOD RMSE on HoF, Density, ZPVE, α, and C_v, while Chemprop has lower OOD RMSE on HOMO, LUMO, GAP, μ, and $\langle R^2 \rangle$. This agreement indicates that the comparison is not an artifact of using R^2 alone.

4.3 Structural OOD: high scores and narrower model differences

Table 2 reports R^2 on the structure-based OOD split. On nine of ten endpoints, all six models lie between 0.84 and 1.00; μ is the common exception, with values between 0.45 and 0.70. The tree-based models that perform poorly on the property-tail split recover to high R^2 under this split. GotenNet is strongest or tied for strongest on most endpoints, but the overall spread is much smaller than in Table 1.

This does not imply that structural OOD is generally easier than property OOD. The two experiments use different selection rules, and the difficulty of a structural split depends on how chemical separation is defined. The result instead demonstrates empirically that BOOM's property-tail evaluation and our structure-based evaluation probe different failure modes: strong performance on one does not guarantee strong performance on the other.

4.4 Repeated seeds and validation-based Chemprop tuning

Across seeds 42, 43, and 44, the median per-endpoint standard deviation of property-OOD binned R^2 is 0.00 for Elastic Net, 0.01 for Random Forest, 0.10 for XGBoost, and 0.09 for Chemprop. Chemprop is the most variable of these models on an individual endpoint, with a maximum standard deviation of 1.13 on LUMO. The qualitative pattern is stable, although α lies close to zero and changes sign across seeds (0.05, -0.04, and -0.01). Endpoint-level Chemprop values are given in Appendix B.

Table 2: Structural (X-space) OOD R^2, seed 42, using the UMAP/HDBSCAN split described in Section 3.2.

Endpoint	Chemprop	Elastic Net	XGBoost	RF	GotenNet	MolFormer
HoF	0.94	0.99	0.96	0.94	0.99	0.94
Density	0.98	0.97	0.96	0.96	1.00	0.95
HOMO	0.94	0.85	0.88	0.89	0.96	0.84
LUMO	0.96	0.85	0.92	0.92	0.98	0.89
GAP	0.94	0.84	0.89	0.90	0.97	0.85
ZPVE	0.99	0.99	0.99	0.99	1.00	0.98
$\langle R^2 \rangle$	0.94	0.91	0.94	0.91	0.98	0.92
α	0.95	0.97	0.98	0.97	0.99	0.95
μ	0.70	0.45	0.54	0.57	0.70	0.47
C_v	0.95	0.98	0.98	0.97	1.00	0.96

Validation-based tuning provides a second robustness check for the Elastic-Net–Chemprop comparison. On the six endpoints for which the 50-configuration Chemprop search completed, tuned Chemprop is higher on Density, HOMO, and μ, while Elastic Net is higher on HoF, α, and C_v: a 3–3 split. Relative to default Chemprop, tuning changes the winner on Density but not on the other five completed endpoints. If these six tuned values are substituted into the ten-endpoint comparison while retaining default Chemprop on the four unfinished searches, the aggregate head-to-head changes from 5–5 to 6–4 in favor of Chemprop. We therefore avoid claiming an Elastic-Net advantage; the more stable conclusion is that Elastic Net remains competitive on these property-tail tasks despite its substantially simpler form.

The tuning cost is also different. For HoF, Density, and HOMO, the completed 50-configuration Chemprop searches took 4h38m, 3h48m, and 29h37m, respectively, on the same system, versus 23s, 2m36s, and 5m10s for Elastic-Net tuning. These measurements are not a comprehensive efficiency benchmark, but they illustrate that comparable model-selection opportunities can have very different computational costs. Full tuning and timing results are in Appendix C.

5 Discussion and limitations

What the simple baselines show. The six-model comparison does not support treating "simple" as a homogeneous model class. On property-tail OOD, Elastic Net and the two tree ensembles behave very differently because they have different extrapolation properties; on the structural split, the tree ensembles are again competitive. A useful practical implication is that molecular OOD benchmarks benefit from including more than one classical baseline and, in particular, at least one baseline whose mathematical form does not preclude the type of extrapolation being tested.

This interpretation is consistent with broader evidence that strong classical representations and baselines remain difficult to displace uniformly. In a recent benchmark of 25 pretrained molecular embedding models, nearly all neural embeddings showed negligible or no improvement over an ECFP fingerprint baseline under the authors' statistical comparison framework (Praski et al., 2025). CheMeleon, a descriptor-pretrained D-MPNN foundation model, recently achieved the highest win rate on the Polaris benchmark set, but the margin over a strong Random Forest baseline was modest: 75% (21/28 tasks) versus 68% (19/28) (Burns et al., 2026). That Random Forest used 2048-dimensional radius-2 Morgan count fingerprints augmented with RDKit physicochemical descriptors. The CheMeleon study explicitly states that no benchmark-specific hyperparameter optimization was performed for any model; models were instead evaluated with fixed training configurations, with validation-based procedures such as early stopping where applicable. The gap was much larger on MoleculeACE (97% versus 50% win rate for Random Forest), so the Polaris result should not be generalized to all tasks. It nevertheless illustrates that substantial pretraining and model complexity do not necessarily yield a large practical improvement over a strong classical baseline on every benchmark.

A subsequent OpenADMET technical blog post explored extensive changes to CheMeleon's pretraining data, featurization, message-passing architecture, and automated optimization (West & Burns, 2026). The authors report no statistically significant across-the-board improvement over

the original model; even the final autoresearch-derived model was not significantly better on the full benchmark. We view this as suggestive, not conclusive, evidence that this particular model family may currently be near a local performance plateau: incremental changes can yield task-specific gains, while broad improvements appear difficult. Together with the relatively small CheMeleon–Random-Forest gap on Polaris, this motivates continued comparison against inexpensive classical baselines rather than assuming that additional model scale will produce a practically meaningful gain.

Screening cost and interpretability. For large-scale virtual screening, the relevant resource constraint is often the number of compounds that can be scored within a fixed budget. A lightweight descriptor-based predictor can make a substantially larger candidate library computationally accessible than a more expensive model, provided that its end-to-end cost per compound is lower. This comparison must include descriptor calculation or conformer generation, not just the prediction step, as well as hardware and batching. Our tuning-time measurements demonstrate a model-selection cost difference; they do not quantify inference throughput or the resulting increase in screening coverage.

Interpretability provides a second motivation, particularly for regularized linear models such as Elastic Net. The fitted coefficients make the dependence on individual descriptors and interaction terms explicit, making the model more amenable to inspection than a large neural model. This advantage is not automatic: correlated descriptors and thousands of interaction features complicate interpretation, and coefficient associations are not evidence of causal chemical mechanisms. Tree ensembles likewise are not necessarily transparent merely because they are classical models. Where predictive performance is comparable, lower screening cost and a more inspectable prediction rule can nevertheless favor a simpler model in practice.

Property OOD and structural OOD are complementary. Property extrapolation is scientifically meaningful when the objective is to find compounds with unusually high or low properties. Structural novelty is a separate axis that matters when entering new chemical regions. The strong contrast between Tables 1 and 2 argues for reporting both kinds of shift when possible, rather than assuming that either one is a proxy for the other.

Metric interpretation. The R^2/ρ^2 discrepancy is consequential for absolute interpretation because ρ^2 cannot reveal a worse-than-mean predictor and is insensitive to affine miscalibration. At the same time, our ranking analysis shows that the correction does not generally reverse which models are stronger. We therefore view it as a clarification of benchmark severity rather than a reason to discount BOOM. Indeed, the more negative coefficient-of-determination values reinforce BOOM's central observation that property-tail OOD extrapolation remains challenging.

Coverage and reproduction fidelity. Our study evaluates six models, four of which were included in BOOM (Chemprop, Random Forest, GotenNet, and MolFormer); this remains substantially narrower than BOOM's full fourteen-model comparison. GotenNet and MolFormer were run for one seed, and their original BOOM fine-tuning pipelines were not available in the original BOOM study repository. Our GotenNet runner uses RDKit-generated conformers, while the geometry source used for the BOOM GotenNet results could not be confirmed from the available materials. Most GotenNet ID RMSE values are of the same order as BOOM's reported averages, but the $\langle R^2 \rangle$ OOD result differs markedly (Appendix D). MolFormer likewise shows some reproduction differences, particularly on ID RMSE. These models therefore broaden the comparison, but their numerical values should not be interpreted as exact replications of the original implementations.

Structural split limitations. The UMAP/HDBSCAN split is a heuristic, single split. Its separation is defined in a 2D embedding of Morgan fingerprints, and different fingerprint, UMAP, clustering, or cluster-selection settings could change the held-out region and difficulty. In addition, the structural experiment uses approximately 10% held out as OOD with the remainder used for training, whereas BOOM's property-based procedure also reserves an ID test subset. The property-based results also use the lower/upper-tail binned R^2 aggregation, whereas the structural split uses one R^2 over the held-out structural set. Direct differences in absolute score between the two split types should therefore be interpreted qualitatively. A stronger future evaluation would compare several structure-aware splitting strategies and repeat the structural split itself.

Incomplete tuning. Validation-based Chemprop tuning completed for six of ten endpoints. The remaining four endpoints are reported only with the default Chemprop configuration, so the ten-endpoint tuned-versus-Elastic-Net comparison is necessarily partial. We expose this limitation rather than imputing missing tuned results. More generally, tuning budgets can be difficult to equalize across model classes because the computational cost of one configuration differs substantially.

Dataset scope. Eight of the ten endpoints are from QM9. GotenNet was developed and evaluated on QM9 among other 3D molecular datasets (Aykent & Xia, 2025), so its strong performance here may not extend to other molecular datasets or tasks. Likewise, MolFormer was pretrained at very large scale (Ross et al., 2022), yet does not dominate these OOD tasks. This single benchmark is therefore insufficient to infer a general relationship between pretraining scale, architectural complexity, and OOD performance.

6 Conclusion

Revisiting BOOM with a reconstructed evaluation pipeline, additional baselines, repeated seeds, partial validation-based tuning, and a complementary structural split leads to three conclusions. First, values reported as R^2 in the original BOOM OOD results are reproduced by ρ^2; using the coefficient of determination exposes substantially poorer absolute property-tail extrapolation while largely preserving model ordering. Second, Elastic Net remains a competitive baseline against Chemprop on these tasks, including under the robustness checks completed here, whereas tree ensembles perform poorly specifically when target-space extrapolation is required. Third, under our structure-based split, all six models obtain high positive R^2 values and differences narrow, showing that property-tail and structural OOD probe distinct behaviors. Together, these observations support careful use of simple baselines, explicit definitions of distribution shift, and metrics that reflect both association and calibration when evaluating molecular OOD generalization.

AI use statement

We used generative AI tools to assist with implementing and debugging analysis code, literature search and summarization, and drafting, restructuring, editing, and formatting the manuscript. AI-generated text and code were treated as suggestions rather than as authoritative outputs. We supervised the AI-assisted work, ran and inspected the analyses, checked numerical results and citations against the underlying outputs and sources, and manually reviewed the scientific claims and final wording. We take responsibility for the final content of this work, including text, code, analyses, and claims produced with the aid of generative AI.

Reproducibility statement

The code attached to the submission as supplementary material is a fork of the original BOOM study's repository, which served as the starting point for our reproduction and extensions. Code accompanying the submission as supplementary material contains the reproduction and extension pipeline used in this study, including data preprocessing and split generation, training and evaluation scripts for the evaluated model families, computation of R^2, ρ^2, and RMSE, repeated-seed experiments, validation-based Chemprop tuning, the UMAP/HDBSCAN structural split, and scripts for regenerating reported tables from stored result files. The datasets are publicly available through the original BOOM resources and their underlying sources. Sections 2–4 and Appendix A document the data splits, model configurations, metrics, and robustness analyses; Appendix D records the remaining reproduction-fidelity limitations for GotenNet and MolFormer.

References

David L. J. Alexander, Alexander Tropsha, and David A. Winkler. Beware of r^2: Simple, unambiguous assessment of the prediction accuracy of qsar and qspr models. *Journal of Chemical Information and Modeling*, 55(7):1316–1322, 2015. doi: 10.1021/acs.jcim.5b00206.

Evan R. Antoniuk, Shehtab Zaman, Tal Ben-Nun, Peggy Li, James Diffenderfer, Busra Sahin, Obadiah Hersh Smolenski, Everett Grethel, Tim Hsu, Anna M. Hiszpanski, Kenneth Chiu, Bhavya

Kailkhura, and Brian Van Essen. BOOM: Benchmarking Out-of-Distribution Molecular Property Predictions of Machine Learning Models. In *Advances in Neural Information Processing Systems*, 2025.

Sarp Aykent and Tian Xia. GotenNet: Rethinking Efficient 3D Equivariant Graph Neural Networks. In *International Conference on Learning Representations*, 2025.

Jackson W. Burns, Akshat Shirish Zalte, Charlles R. A. Abreu, Jochen Sieg, Christian Feldmann, Miriam Mathea, and William H. Green. Deep learning foundation models from classical molecular descriptors. *arXiv preprint arXiv:2506.15792v2*, 2026. URL https://arxiv.org/abs/2506.15792.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Qianrong Guo, Saiveth Hernandez-Hernandez, and Pedro J. Ballester. Scaffold splits overestimate virtual screening performance. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, pp. 58–72. Springer Nature Switzerland, 2024.

Qianrong Guo, Saiveth Hernandez-Hernandez, and Pedro J. Ballester. UMAP-based clustering split for rigorous evaluation of ai models for virtual screening on cancer cell lines. *Journal of Cheminformatics*, 17:94, 2025. doi: 10.1186/s13321-025-01039-8.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Hao Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1): 9–17, 2024.

Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning. *arXiv preprint arXiv:2508.06199*, 2025.

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

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Devany West and Jackson W. Burns. Throwing everything and the kitchen sink at chemeleon. OpenADMET technical blog post, August 2026. URL https://openadmet.ghost.io/throwing-everything-and-the-kitchen-sink-at-chemeleon/. Published 4 August 2026; DOI: 10.5281/zenodo.21828040.

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

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

A ADDITIONAL EXPERIMENTAL DETAILS

A.1 BASE MODEL CONFIGURATIONS

Table 3 summarizes the documented configurations used in this study. For GotenNet and MolFormer, the table distinguishes settings recorded in our released reproduction implementation from per-run provenance that cannot be reconstructed independently without complete execution logs. Parameters not listed here should be taken from the released code rather than inferred from defaults.

Table 3: Model configurations used in the experiments.

Model	Setting	Value
Chemprop	Message-passing depth	3
	Hidden dimension	300
	FFN layers / hidden dimension	2 / 300
	Batch normalization	yes
	Maximum learning rate	10^{-3}
	Batch size	1,024
	Maximum epochs / patience	100 / 10
	Train/validation split	90/10 random
Random Forest	Number of trees	500
	`max_features`	`sqrt`
Elastic Net	Feature selection	top 80 by univariate F-score
	Polynomial expansion	degree 2, interaction terms only
	`l1_ratio`	$\{0.1, 0.5, 0.9, 1.0\}$
	Regularization strengths	30 log-spaced values
	Cross-validation folds	3
XGBoost	Main boosting parameters	library defaults in reported runs
	`subsample`	0.8
	`colsample_bytree`	0.8
	Hyperparameter search	none
GotenNet	Epochs	50
	Batch size	64
	Optimizer / learning rate	AdamW / 5×10^{-4}
	Weight decay	10^{-5}
	Scheduler	cosine annealing over 50 epochs
	Validation / selection	$\sim$10% seeded holdout; lowest validation MSE
	3D geometry in released runner	RDKit ETKDGv3 + attempted MMFF optimization
MolFormer	Epochs	30 (HoF, Density); 20 (eight QM9 endpoints)
	Batch size	32
	Optimizer / learning rate	LAMB / 10^{-5}
	Weight decay / gradient clipping	0 / global norm 1.0
	Scheduler	cosine annealing over requested epochs
	Validation / selection	$\sim$10% seeded holdout; lowest validation MSE
	Fine-tuning	encoder and regression head end-to-end

The GotenNet and MolFormer entries in Table 3 describe our released reproduction implementation; they are not intended to imply that BOOM used the same settings. Where complete execution logs are unavailable, we report only settings supported by the released code and recorded experimental configuration and avoid inferring additional command-line overrides.

GotenNet implementation. The released runner trains a newly initialized model for 50 epochs with AdamW (learning rate 5×10^{-4}, weight decay 10^{-5}), batch size 64, MSE on standardized targets, cosine annealing, and selection of the lowest-validation-MSE state from an approximately 10% seeded validation holdout. The architecture uses atom basis 128, four interaction blocks, $l_{\max} = 2$,

32 radial basis functions, eight attention heads, and cutoff 5.0. Coordinates are generated with RDKit ETKDGv3 (seed 42), with a fallback embedding attempt and MMFF optimization when possible.

MolFormer implementation. The released runner loads `N-Step-Checkpoint_3_30000.ckpt` with `strict=False` and fine-tunes the encoder and regression head end-to-end using LAMB (learning rate 10^{-5}), batch size 32, cosine annealing, and gradient clipping at 1.0; the lowest-validation-MSE state from an approximately 10% seeded holdout is restored. The encoder has 12 layers, 12 heads, dimension 768, and 32 random features, followed by masked mean pooling and a $768 \to 768 \to 1$ regression head. We report 30 epochs for HoF and Density and 20 epochs for the eight QM9 endpoints. In our code, the training script uses 30 epochs unless a different value is specified; the script that starts the QM9 experiments sets this value to 20.

For both runners, property-OOD and structural-OOD experiments fit separate models, and target scaling is computed from the outer training set before the internal validation holdout. Complete command logs were not retained for every run, so we do not infer unrecorded command-line overrides, checkpoint hashes, or effective RNG state beyond what is documented in the released implementation.

A.2 Structural split construction

For each dataset, the structure-only split is generated as follows:

1. Compute a 2048-bit radius-2 Morgan fingerprint for every molecule using RDKit.
2. Embed fingerprints in two dimensions with UMAP using $n_{\text{neighbors}} = 15$, $\text{min_dist} = 0.1$, Euclidean distance, spectral initialization, $\text{spread} = 1.0$, and random seed 42; all remaining UMAP parameters use library defaults.
3. Cluster the embedding with HDBSCAN using minimum cluster size 20.
4. For every non-noise cluster, compute its detachment score as the shortest Euclidean distance in the 2D embedding between any point in that cluster and any point not in the cluster. The nearest-neighbor search is implemented with a KD-tree.
5. Rank clusters from most to least detached and hold out complete clusters until approximately 10% of molecules are assigned to the OOD set. Remaining clusters and HDBSCAN noise points form the training set.

This is a heuristic variant of the UMAP-based splitting idea of Guo et al. (2025). It was chosen to produce coherent held-out structural regions without selecting a fixed Tanimoto threshold. The UMAP/split seed is independent of model-training seeds. Once generated, each structural split is cached and reused across models, so comparisons in Table 2 use identical train/OOD membership. The experiments use a single split rather than cross-validation over split definitions.

B Repeated-seed results

Table 4 summarizes variability of property-OOD binned R^2 over seeds 42, 43, and 44 for the four models that were repeated.

Table 4: Across-endpoint variability of property-OOD binned R^2 over seeds 42/43/44.

Model	Median std.	Max std.	Comment
Elastic Net	0.00	0.16	Largest value occurs on $\langle R^2 \rangle$, whose score is about -100
Random Forest	0.01	0.03	Low variability under bagging
XGBoost	0.10	0.26	Stochastic row/column subsampling
Chemprop	0.09	1.13	Maximum at LUMO

Chemprop, the most variable model in the repeated analysis, has the endpoint-level results in Table 5.

Table 5: Chemprop property-OOD binned R^2 across three seeds.

Endpoint	Seed 42	Seed 43	Seed 44	Mean $\pm$ std.
HoF	0.27	0.31	0.25	0.28 ± 0.03
Density	0.48	0.42	0.60	0.50 ± 0.09
HOMO	-1.11	-1.22	-1.04	-1.13 ± 0.09
LUMO	-2.32	-2.34	-4.28	-2.98 ± 1.13
GAP	-1.98	-1.88	-1.91	-1.92 ± 0.05
ZPVE	0.58	0.77	0.64	0.66 ± 0.10
$\langle R^2 \rangle$	-2.08	-1.70	-2.05	-1.94 ± 0.21
α	0.05	-0.04	-0.01	0.00 ± 0.05
μ	-7.03	-6.61	-6.53	-6.72 ± 0.26
C_v	-0.23	-0.09	-0.13	-0.15 ± 0.07

C VALIDATION-BASED CHEMPROP TUNING

The Chemprop search samples 50 configurations per endpoint, varying message-passing depth (2–6), dropout (0–0.40), feed-forward depth (1–3 layers), and hidden size (300–2400). Selection is by lowest validation loss. Six of ten endpoint searches completed within the available compute budget.

Table 6: Default Chemprop, validation-selected Chemprop, and tuned Elastic Net on the six completed endpoints. Binned R^2 and RMSE use the property-based OOD split, seed 42.

Endpoint	Metric	Chemprop default	Chemprop tuned	Elastic Net
HoF	Binned R^2	0.27	0.37	0.68
	RMSE	78.9	85.8	29.2
Density	Binned R^2	0.48	0.71	0.54
	RMSE	0.0598	0.0434	0.0487
HOMO	Binned R^2	-1.11	-0.99	-2.30
	RMSE	0.0180	0.0175	0.0224
α	Binned R^2	0.05	0.14	0.32
	RMSE	5.31	5.02	3.01
μ	Binned R^2	-7.03	-6.04	-17.61
	RMSE	2.40	2.27	3.01
C_v	Binned R^2	-0.23	-0.77	0.83
	RMSE	2.40	2.63	0.73

On these six endpoints, validation-selected Chemprop and Elastic Net each obtain the higher binned R^2 on three. The result also illustrates that validation-based tuning does not move Chemprop uniformly in one direction: it improves several endpoints, leaves some qualitative rankings unchanged, and lowers C_v performance in this run.

Table 7: Measured tuning wall-clock time on the same machine (one NVIDIA A10G GPU and 32 CPU cores). Elastic Net did not use the GPU.

Endpoint	Chemprop search	Elastic-Net search	Wall-clock ratio
HoF	4h38m	23s	726×
Density	3h48m	2m36s	88×
HOMO	29h37m	5m10s	344×

The sampled Chemprop hidden dimension ranges from 300 to 2400, corresponding in these tuning experiments to models of roughly 0.3M to 29M parameters; the largest sampled networks took approximately 3.4–6.5 times longer per run than the smallest. We report these timings as descriptive measurements rather than a general efficiency benchmark.

D REPRODUCTION-FIDELITY CHECKS FOR GOTENNET AND MOLFORMER

BOOM did not provide runnable training code for the GotenNet or MolFormer rows used in its reported comparison. We therefore compare our single-seed RMSE values against the three-seed averages reported in BOOM's Table 7. This is an approximate fidelity check rather than an exact implementation match.

Table 8: GotenNet RMSE: BOOM three-seed average versus our seed-42 run. BOOM "OOD" is the property-value split.

Endpoint	BOOM ID	Ours ID	BOOM OOD	Ours OOD
HoF	5.44	9.56	15.54	58.57
Density	0.0070	0.0043	0.0360	0.0122
HOMO	0.0052	0.0036	0.0126	0.0171
LUMO	0.0039	0.0044	0.0118	0.0205
GAP	0.0088	0.0050	0.0229	0.0275
ZPVE	0.00043	0.00103	0.00053	0.00423
$\langle R^2 \rangle$	11.80	32.00	16.70	319.77
α	0.553	0.468	1.825	3.798
μ	0.319	0.554	2.173	2.438
C_v	0.197	0.207	0.302	0.894

Most ID values are of the same order of magnitude, but several OOD values differ. The largest discrepancy is the $\langle R^2 \rangle$ endpoint. Our runner uses RDKit-generated conformers, whereas the geometry source for BOOM's GotenNet evaluation is not documented in the material available to us. Because $\langle R^2 \rangle$ is strongly geometry-dependent, geometry provenance is one plausible contributor, but we cannot establish it as the cause from the available evidence.

Table 9: MolFormer RMSE: BOOM three-seed average versus our seed-42 run.

Endpoint	BOOM ID	Ours ID	BOOM OOD	Ours OOD
HoF	10.94	22.75	93.60	110.02
Density	0.0273	0.0177	0.0770	0.1289
HOMO	0.0050	0.0079	0.0236	0.0233
LUMO	0.0052	0.0098	0.0256	0.0318
GAP	0.0064	0.0117	0.0275	0.0339
ZPVE	0.0011	0.0042	0.0199	0.0115
$\langle R^2 \rangle$	40.10	58.69	314.10	311.51
α	1.05	1.66	7.36	6.67
μ	0.6020	0.7736	2.2320	2.8851
C_v	0.7850	0.6832	3.6670	2.0575

Our MolFormer ID RMSE is higher than BOOM's reported average on eight of ten endpoints and lower on Density and C_v. OOD RMSE is lower on five endpoints and higher on five. Because the available BOOM materials do not specify the fine-tuning schedule used for this table, and our runs use 30 epochs on HoF/Density and 20 elsewhere, we treat these results as a broader model-family comparison rather than an exact numerical reproduction.

E ORIGINAL THREE-MODEL ID RESULTS

For completeness, Table 10 reports the ID R^2 values from the original three-model analysis.

These ID results are uniformly strong relative to the property-tail OOD values in Table 1, which is consistent with the central BOOM observation that extrapolation to the held-out property tails is substantially more difficult than prediction within the sampled target range.

Table 10: ID R^2 from the original three-model analysis (seed 42).

Model	HoF	Density	HOMO	LUMO	GAP	ZPVE	$\langle R^2 \rangle$	α	μ	C_v
Chemprop	.95	.94	.94	.99	.98	1.00	.97	.99	.77	.99
Random Forest	.91	.89	.86	.97	.96	.99	.94	.98	.67	.99
Elastic Net	.98	.98	.78	.93	.91	.98	.68	.95	.54	.97

F Additional metric-comparison details

The direct R^2/ρ^2 comparison shows that the discrepancy is concentrated in the property-based OOD columns. ID predictions are sufficiently well calibrated that the two measures are close (maximum absolute difference 0.05 in the comparison). For OOD, selected cells are:

Table 11: Illustrative property-OOD cells under the coefficient of determination and squared Pearson correlation.

Model / endpoint	OOD R^2	OOD ρ^2
Elastic Net / $\langle R^2 \rangle$	-100.56	0.01
Elastic Net / μ	-17.61	0.21
XGBoost / μ	-13.46	0.14
Random Forest / μ	-12.97	0.15
Random Forest / LUMO	-8.44	0.02
GotenNet / μ	-7.90	0.15

Across the 52 seed-42 model–endpoint cells for which both metrics were available in that direct comparison, 38 have negative OOD R^2. Per-endpoint rank correlation between the two metrics is at least 0.8 for eight of ten endpoints; the main effect is therefore a change in absolute interpretation rather than a wholesale reordering of models.

G Complete ρ^2 and RMSE results from the original three-model analysis

For completeness, Tables 12 and 13 report the complete ρ^2 and RMSE results from the original three-model analysis. They provide a direct bridge between the initial reproduction and the expanded six-model comparison.

Table 12: Squared Pearson correlation ρ^2 for the original three-model analysis (seed 42). OOD values use the lower/upper-tail aggregation.

Split	Model	HoF	Density	HOMO	LUMO	GAP	ZPVE	$\langle R^2 \rangle$	α	μ	C_v
ID	Chemprop	.97	.98	.94	.99	.98	1.00	.97	.99	.78	.99
ID	Random Forest	.92	.92	.87	.97	.96	.99	.94	.98	.67	.99
ID	Elastic Net	.98	.98	.78	.93	.91	.98	.73	.95	.54	.97
OOD	Chemprop	.58	.92	.21	.18	.05	.73	.10	.30	.13	.48
OOD	Random Forest	.15	.02	.02	.02	.04	.11	.06	.09	.15	.00
OOD	Elastic Net	.76	.73	.18	.08	.09	.87	.01	.56	.21	.88

Table 13: RMSE for the original three-model analysis (seed 42). Values are shown in each endpoint's original units and are rounded to the precision used in the original analysis.

Split	Model	HoF	Density	HOMO	LUMO	GAP	ZPVE	$\langle R^2 \rangle$	α	μ	C_v
ID	Chemprop	16.95	.02	.00	.00	.01	.00	32.05	.67	.56	.27
ID	Random Forest	24.11	.02	.01	.01	.01	.00	49.15	.99	.68	.40
ID	Elastic Net	11.41	.01	.01	.01	.01	.00	110.65	1.42	.80	.56
OOD	Chemprop	78.88	.06	.02	.02	.03	.01	310.96	5.31	2.40	2.40
OOD	Random Forest	139.44	.18	.03	.04	.04	.02	406.29	8.64	2.88	3.33
OOD	Elastic Net	29.16	.05	.02	.03	.04	.01	1133.45	3.01	3.01	.73