Evaluation of Machine-Learning Models in Polymer Chemistry with Prediction of Not Reported Measurements

SINGH Shivani[1], MONDAL Sourov[2], NIEVES Juan Carlos[1], TORRA Vicenç[1]

[1]*Department of Computing Science, Umeå University, Sweden*
E-mail: ssingh@cs.umu.se, jcnieves@cs.umu.se, vtorra@cs.umu.se
[2]*Department of Chemistry, Umeå University, Sweden*
E-mail: sourov.mondal@umu.se

We present a data-centric evaluation framework tailored for heterogeneous, unevenly distributed scientific datasets, using polymer nanoparticles as a case study. This work aims to extend the application of machine learning (ML) within the chemical sciences. In many chemistry-related problems, the primary challenge lies not in the selection of algorithms but in the evaluation process, particularly when datasets are limited in size, comprise mixed data types (categorical and numerical) and exhibit uneven data distribution. ML enables the modeling of these dependencies and guides experiments through data-driven predictions. We evaluate multiple ML models to identify the one that most effectively captures the underlying patterns in polymer data and yields accurate predictions of polymer characteristics, as measured by low mean absolute error (MAE). The selected model is then used to identify key features that influence the target properties and to impute missing ("Not Reported") values, thus completing the dataset for downstream analyses. Our findings regarding algorithm behavior are specific to the configurations evaluated, whereas the evaluation framework can be applied to other datasets with similar characteristics beyond polymer chemistry.

1 Introduction

Machine Learning offers a robust tool for capturing complex data relationships and directing experimental efforts through predictive insights. Polymeric nanoparticle performance results from the interaction of multiple factors, including polymer architecture, composition, and processing conditions. Traditional experimental testing is costly and time-consuming, especially when measurements are incomplete or vary widely in scale. By learning patterns across chemical datasets, ML accelerates discovery, reduces cost, and reveals mechanistic signals that are hard to see by experiment alone.

In their survey, Ge et al. [1] report that ML helps connect polymer structures to their properties, but progress is often slowed by small, messy, and inconsistently prepared datasets. To make real use of ML, chemical data need to be translated into clear, machine-readable descriptors and follow FAIR principles—findable, accessible, interoperable, and reusable [6]. The field grows faster when chemists and data scientists work together to refine models for specific polymer systems, improve data quality, and apply ML to study solid-state behavior, solution properties, composites, drug delivery, and polymer–biology interactions. Prior works [3, 5] have improved the accuracy of chemical property, but reproducibility in limited sample and mixed

Acknowledgement. This research has been conducted in the project "Understanding and designing Block Copolymers using AI" funded by the KEMPE foundation. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

https://www.doi.org/10.15452/978-80-7599-515-5.2026.20

Model	Fold										Avg Test
	1	2	3	4	5	6	7	8	9	10	MAE
XGB_{Confg}	3.90	28.42	8.91	29.66	11.69	17.20	96.57	58.78	13.28	114.93	38.34 ±36.99
MLP_{Confg}	64.18	43.03	13.29	71.23	34.93	53.15	124.91	117.38	50.49	78.70	65.13 ±33.12

Table 1: Fold-wise test MAE (10 folds and mean±SD) for XGB_{Confg} and MLP_{Confg} on EM Nanoparticle Size(nm)

type data still depends on careful evaluation. It requires checking how well prediction errors generalize across folds, identifying the key descriptors that influence polymer behavior, and accurately predicting missing features. Building on these principles, our study delivers a compact, reproducible pipeline and demonstrates its benefits on evaluation of polymeric characterized properties while maintaining a balanced chemical and computational perspective.

To address this setting, we adopt a data-centric ML workflow that balances chemical interpretability with statistical robustness. We curate the diblock copolymer dataset defining descriptors and based on those descriptors their polymer sizes. We benchmark five regressors spanning linear, tree/ensemble, gradient-boosted, and neural models under consistent settings with k-fold cross validation to ensure fair comparison [2]. We use model-based attribution to highlight key features, and finally predict the "Not Reported" measurements to advance later-stage analysis.

2 Methodology

Polymer nanoparticle dataset is curated though surveyed publications up to 2024, containing 118 diblock copolymer samples with 13 descriptors as features/attributes (chemical composition, core/corona DP, core cLogP; concentration(wt%), characterization pH, LCST/UCST, and related conditions) and two particle-size outputs, namely, EM nanoparticle size and Hydrodynamic diameter. We repeatedly cleaned and adjusted the dataset in several meetings to make the chemistry data suitable for ML, identify problematic samples, and improve the results until the predictions became reliable. To obtain unbiased performance estimates under heterogeneous and uneven data distribution constraints, we employed 10-fold cross-validation. We benchmarked five regression models with complementary inductive biases: Ridge and Lasso regression (both with default parameters), Random Forest (`n_estimators`=500, `random_state`=42), XGBoost (`n_estimators`=500, `learning_rate`=0.05, `max_depth`=4, `subsample`=0.8, `colsample_bytree`=0.8, `random_state`=42), and a Multilaye Perceptron (MLP) as an artificial-neural-network baseline(`hidden_layer_sizes`=(20,10), `max_iter`=600, `early_stopping`=True, `random_state`=42). XGB_{Confg} and MLP_{Confg} represent the tested XGB configuration and tested MLP configuration, respectively. We report test MAE for each fold and summarize uncertainty as the mean ± SD across folds to compare the performance of each regressor. Predicted-versus-observed plots for models (one per fold) were used to visualize fit quality. For each plot, we record the input-feature combinations of these outliers to assess whether specific chemical compositions or conditions systematically challenge the regressor. Also, using fold-wise error spread, we select the most stable model and examine which descriptors are most important for the nanoparticle sizes. Then we refit the model on all labeled data to estimate previously unreported measurements.

Model	Fold										Avg Test MAE
	1	2	3	4	5	6	7	8	9	10	
XGB_{Confg}	28.79	32.53	41.12	5.54	22.25	99.78	18.78	40.14	25.88	73.37	38.92 ±26.70
MLP_{Confg}	29.62	66.51	45.84	80.38	55.04	172.17	18.94	36.85	25.88	18.65	54.98 ±43.62

Table 2: Fold-wise test MAE (10 folds and mean±SD) for XGB_{Confg} and MLP_{Confg} on Hydrodynamic Diameter(nm)

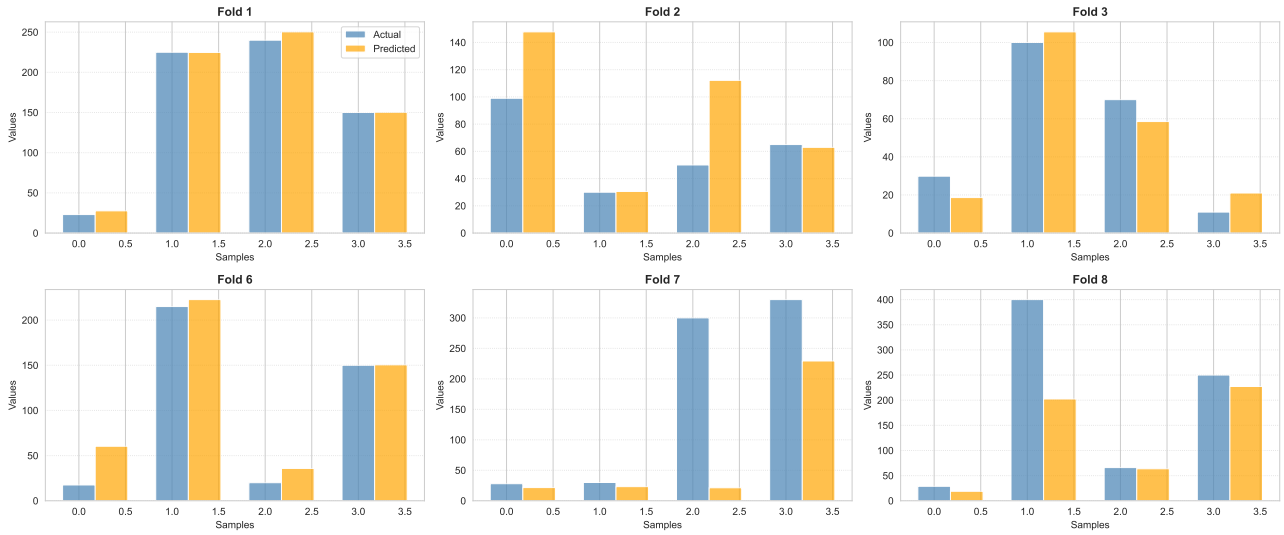

Figure 1: Actual vs Predicted Test Values (Folds 1-3 and 6-8) with XGB_{Confg} on EM Nanoparticle Size(nm)

3 Results and Discussion

We compute train and test set MAE's for each regressors. With the limiliting space, we report only a subset of results. Tables 1–2 report test MAE per target for XGB_{Confg} and MLP_{Confg}, showing that the tree-based boosting model XGB_{Confg} outperforms the neural-network model MLP_{Confg} on this polymeric dataset. Figure 1 shows predicted-vs-observed plots under 10-fold cross-validation for the top performer XGB_{Confg} model on electron microscopy (EM) nanoparticle size for folds 1-3 and 6-8. These fold-wise panels indicate that most folds fit well, while sparse regimes in certain folds degrade performance and inflate the average MAE. From these fold-wise plots, we flag individual samples whose predicted diameters deviate substantially from their true values.

In Figure 2, we present XGB_{Confg}'s gain-based feature attribution to identify the most influential descriptors. For each split, the algorithm computes the reduction in loss achieved by splitting on a feature and averages these gains across all trees. Features with higher average gain are interpreted as more influential on the predicted output. For the size of EM nanoparticle, 'temperature' is most impactful attribute, followed by 'characterization pH', 'corona DP', 'LCST', and 'concentration (wt%)'. For Hydrodynamic diameter, the most influential attributes are 'core cLogP', 'temperature', 'concentration (wt%)', 'end-group charge', and 'core DP'. Afterwards, to complete the record, we refit the configured XGB_{Confg} model on all labeled samples and generated predictions for entries with "Not Reported" polymeric sizes.

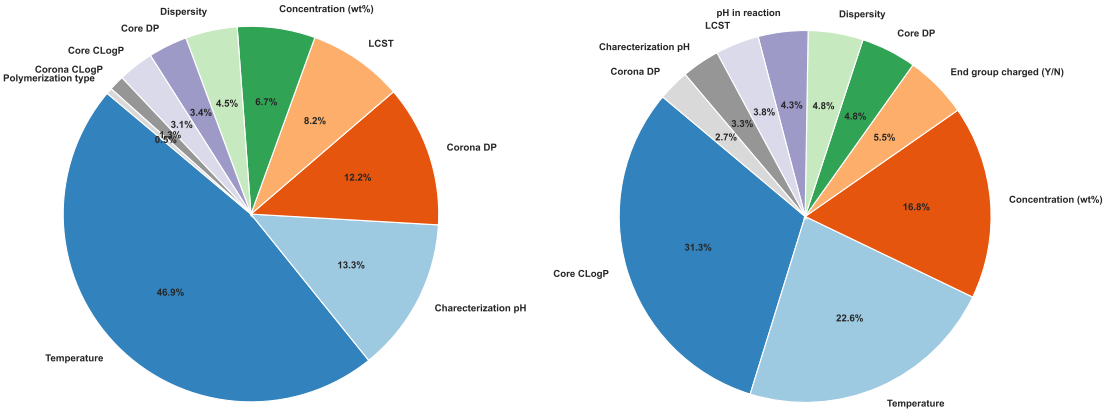

Figure 2: XGB_{Confg} Gain-based Feature Attribution for (left) EM Nanoparticle Size(nm), (right) Hydrodynamic Diameter(nm)

4 Conclusion

This work showcases the application of machine learning to polymer–nanoparticle data. Using 10-fold cross-validation, we evaluated the generalization error of multiple regression models under identical preprocessing pipelines and fixed hyperparameter settings. In our experiments, the tested configuration of XGB_{Confg} achieved the lowest test MAE and the most consistent performance across folds, whereas the tested neural baseline MLP_{Confg} showed higher fold-to-fold variance. We do not claim superiority of entire method families (e.g., all boosting or all ANN); results are specific to this dataset and configurations. Furthermore, model-based feature attribution identified the most influential predictors of polymer size. Finally, after selecting the most stable regressor, we retrained it on the full labeled dataset and successfully estimated previously unreported polymer particle-size values. Overall, we present a straightforward and reproducible ML pipeline for interpretable chemical modeling in settings where data are limited and unevenly distributed.

References

[1] Ge, W., De Silva, R., Fan, Y., Sisson, S. A., & Stenzel, M. H. (2025). Machine learning in polymer research. *Advanced Materials*, 37(11), 2413695.

[2] Hastie, T., Tibshirani, R., Friedman, J. (2009) The Elements of Statistical Learning, *Springer*.

[3] Kimmig, J., Schuett, T., Vollrath, A., Zechel, S., & Schubert, U. S. (2021). Prediction of nanoparticle sizes for arbitrary methacrylates using artificial neuronal networks. *Adv. Science*, 8, 2102429.

[4] Lu, Y., Yalcin, D., Pigram, P. J., Blackman, L. D., & Boley, M. (2023). Interpretable machine learning models for phase prediction in polymerization-induced self-assembly. *Journal of Chemical Information and Modeling*, 63(11), 3288-3306.

[5] Youshia, J., Ali, M. E., & Lamprecht, A. (2017). Artificial neural network based particle size prediction of polymeric nanoparticles. *Euro. J. Pharm. and Biopharmaceutics*, 119, 333-342.

[6] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et. al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.