

Physics-Informed Support Vector Regression for Predicting Copolymer LCST Depressions via Thermodynamic Hofmeister Featurization

Abhay Chauhan

abhaychauhan7895625@gmail.com

Research Article

Keywords:

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10923562/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Physics-Informed Support Vector Regression for Predicting Copolymer LCST Depressions via Thermodynamic Hofmeister Featurization

Abhay Chauhan

Email: abhaychauhan7895625@gmail.com

Abstract

Thermoresponsive polymers exhibiting Lower Critical Solution Temperature (LCST) behavior are central to smart biomaterial design and adaptive formulation engineering. However, accurately predicting LCST across diverse electrolyte environments remains challenging due to complex ion-hydration disruptions and non-linear phase separation kinetics. Standard data-driven models frequently default to categorical salt encodings or random data splits, introducing data leakage and producing overoptimistic metrics that fail to generalize to novel electrolyte chemistries. In this work, we present a physics-informed machine learning framework utilizing continuous thermodynamic descriptors (ΔG_{net} , ΔH_{net} , $V_{s,\text{net}}$, B_{net}) alongside engineered interaction terms (**Cp_Cs_interaction**, **Ionic_Strength_Factor**) to model LCST behavior. Evaluated under strict Leave-One-Salt-Out (LOSO) cross-validation across diverse cation and anion formulations, our optimized Support Vector Regressor (SVR) with a Radial Basis Function (RBF) kernel achieved an out-of-fold generalization performance of RMSE = 9.313 °C ($R^2 = 0.493$, MAE = 7.561 °C). Permutation feature importance auditing revealed that while gross polymer concentration (C_p) dominates overall variance, maintaining net thermodynamic features is essential to prevent model collapse when predicting unseen electrolyte systems (R^2 dropping from 0.493 to 0.223 and RMSE increasing to 11.533 °C upon removal). Furthermore, inference profiling on held-out formulations demonstrates smooth, physically realistic salting-out trends without tree-based step-function artifacts. This work demonstrates that continuous thermodynamic featurization combined with strict leave-one-group-out validation provides a robust, transferable pipeline for electrolyte-driven polymer phase boundary design.

1 Introduction

Thermoresponsive polymers undergo a reversible, temperature-induced phase transition from a soluble, hydrated coil state to an insoluble, collapsed globule state at their Lower Critical Solution Temperature (LCST). Poly(*N*-isopropylacrylamide-co-methoxy triethyleneglycol acrylate), denoted as P(NIPA-co-MTGA), serves as the primary thermoresponsive copolymer system in this study. The phase transition of P(NIPA-co-MTGA) in aqueous environments is dictated by a delicate thermodynamic balance between favorable polymer-water hydrogen bonding ($\Delta H < 0$) and hydrophobic water structuring around its hydrophobic side chains ($\Delta S < 0$). Fine-tuning the LCST boundary of P(NIPA-co-MTGA) for targeted formulation engineering—such as smart drug delivery systems, tissue engineering scaffolds, and responsive valves—is commonly achieved by varying copolymer ratios or introducing inorganic electrolyte salts into the aqueous medium^{1,2}.

In electrolyte solutions, added cations and anions disrupt the primary hydration shell of the P(NIPA-co-MTGA) chain according to the empirical Hofmeister series. Kosmotropic ions strongly polarize bulk water molecules and increase surface tension, depressing the LCST (*salting-out*), whereas chaotropic ions can bind directly to the amide backbone or destabilize hydrophobic collapse. Despite extensive empirical measurements, such as those pioneered by Tokuyama et al.¹, screening the vast combinatorial space of polymer-salt formulations via experimental synthesis remains labor- and resource-intensive. Machine learning (ML) offers an efficient computational alternative, but classical implementations in polymer informatics face three distinct hurdles:

1. **Categorical Overfitting and Data Leakage:** Many empirical models rely on categorical salt identities or one-hot encodings, restricting predictions strictly to salt species present in the training set and failing completely when encountering novel electrolyte formulations.
2. **Naive Cross-Validation Protocols:** Evaluating models via standard random k -fold cross-validation leaks salt-specific signals across training and test splits, severely overestimating true out-of-sample generalization.
3. **Non-Physical Extrapolation Step Artifacts:** Unconstrained tree-based ensembles (e.g., Random Forests, XGBoost) produce discontinuous, step-like transition surfaces along continuous salt concentration gradients due to axis-aligned decision trees.

To overcome these foundational limitations, this work presents a physics-informed machine learning framework for zero-shot LCST prediction in P(NIPA-co-MTGA)-electrolyte solutions. Specifically, we:

- **Develop Continuous Featurization:** Replace discrete categorical salt labels with net thermodynamic hydration descriptors (ΔG_{net} , ΔH_{net} , V_s , B -coefficient) alongside engineered non-linear coupling terms (`Cp_Cs_interaction`, `Ionic_Strength_Factor`).
- **Enforce Strict Zero-Shot Validation:** Evaluate model performance across diverse cation and anion combinations using Leave-One-Salt-Out (LOSO) cross-validation to guarantee true out-of-fold testing on unseen electrolyte families.
- **Benchmark Architecture Performance:** Compare discontinuous tree ensembles against an optimized Support Vector Regressor (SVR) utilizing a Radial Basis Function (RBF) kernel to generate smooth, physically consistent salting-out curves.

- **Audit Feature Contributions:** Perform permutation importance and feature ablation experiments to establish the indispensable role of thermodynamic descriptors in preventing model collapse across unseen salt domains.

2 Methodology

To establish a physically grounded, zero-shot predictive framework for Lower Critical Solution Temperature (LCST) transitions in poly(*N*-isopropylacrylamide-co-methoxy triethyleneglycol acrylate) [P(NIPA-co-MTGA)] electrolyte solutions, we developed an end-to-end machine learning pipeline. The computational workflow spans continuous thermodynamic featurization, group-wise cross-validation, hyperparameter optimization, model-agnostic feature auditing, and explicit feature ablation.

2.1 Dataset Curation and Feature Engineering

Building upon experimental dataset benchmarks established by Tokuyama et al.¹, we constructed an engineered feature matrix designed to eliminate reliance on uninformative, discrete categorical salt identifiers (e.g., one-hot vectors). Instead, each electrolyte system is represented by continuous physical descriptors capturing macroscopic formulation ratios, fundamental ion-hydration thermodynamics, and non-linear concentration interactions.

2.1.1 Formulation Variables

The direct experimental control variables governing the solution phase behavior are defined as:

- **Polymer Concentration (C_p):** The mass fraction of P(NIPA-co-MTGA) in the aqueous formulation ($0.0 \leq C_p \leq 1.0$).
- **Salt Concentration (C_s):** The salt weight percentage (wt%) present in the ternary system.

2.1.2 Net Thermodynamic Hydration Descriptors

To account for the Hofmeister series effect, where specific cations (*c*) and anions (*a*) modulate polymer LCST through water-structuring and surface-tension alterations^{2,3}, we featurized each electrolyte using net thermodynamic state functions:

$$\Delta G_{\text{net}} = \Delta G_{f,c}^{\circ} + \Delta G_{f,a}^{\circ} \quad (1)$$

$$\Delta H_{\text{net}} = \Delta H_{f,c}^{\circ} + \Delta H_{f,a}^{\circ} \quad (2)$$

$$V_{s,\text{net}} = V_{s,c} + V_{s,a} \quad (3)$$

$$B_{\text{net}} = B_c + B_a \quad (4)$$

where ΔG_{net} and ΔH_{net} denote the net standard Gibbs free energy and enthalpy of hydration (kJ/mol), $V_{s,\text{net}}$ represents the net partial molar volume (cm³/mol), and B_{net} is the net Jones-Dole viscosity *B*-coefficient (L/mol), reflecting the ion’s intrinsic capacity to structure or disrupt bulk water networks.

2.1.3 Non-Linear Coupling Interaction Terms

Recognizing that salt-induced salting-out effects are inherently multi-body phenomena, we explicitly engineered two non-linear coupling features:

$$\text{Ionic_Strength_Factor} = C_s \times Z_{\text{effective}} = C_s \times \left(\frac{1}{2} \sum_i z_i^2 c_i \right) \quad (5)$$

$$\text{Cp_Cs_interaction} = C_p \times C_s \quad (6)$$

where z_i and c_i represent the ionic charge and stoichiometric concentration of the constituent ions. The `Ionic_Strength_Factor` captures non-linear electrostatic shielding and salting-out strength, while `Cp_Cs_interaction` accounts for local polymer-electrolyte competitive hydration kinetics. The final input matrix $X \in \mathbb{R}^{N \times 8}$ comprises eight continuous variables: $\{C_p, C_s, \Delta G_{\text{net}}, \Delta H_{\text{net}}, B_{\text{net}}, V_{s,\text{net}}, \text{Ionic_Strength_Factor}, \text{Cp_Cs_interaction}\}$.

2.2 Leave-One-Salt-Out (LOSO) Cross-Validation

Standard random k -fold cross-validation is inappropriate for evaluating model transferability across electrolyte chemistries, as row-wise splitting leaks ion-specific signals into the training fold and artificially inflates performance metrics³. To enforce a strict zero-shot evaluation protocol, we grouped the dataset by unique electrolyte identities:

$$\text{Salt_Group}_k = \text{cation}_k \oplus \text{"_"} \oplus \text{anion}_k \quad (k \in \{1, \dots, K\}) \quad (7)$$

We implemented `LeaveOneGroupOut` cross-validation across all K distinct salt groups (e.g., NaCl, LiBr, CaCl₂). In fold k , all experimental observations associated with Salt_Group_k were withheld entirely as the out-of-fold test set $X_{\text{test}}^{(k)}$, while the remaining $K - 1$ salt groups constituted the training set $X_{\text{train}}^{(k)}$.

To guarantee strict isolation and prevent data leakage, feature scaling was conducted dynamically inside the CV fold execution loop:

$$X_{\text{train, scaled}}^{(k)} = \frac{X_{\text{train}}^{(k)} - \mu_{\text{train}}^{(k)}}{\sigma_{\text{train}}^{(k)}} \quad (8)$$

$$X_{\text{test, scaled}}^{(k)} = \frac{X_{\text{test}}^{(k)} - \mu_{\text{train}}^{(k)}}{\sigma_{\text{train}}^{(k)}} \quad (9)$$

where the mean $\mu_{\text{train}}^{(k)}$ and standard deviation $\sigma_{\text{train}}^{(k)}$ vectors were computed strictly from the training subset.

2.3 Model Architecture Benchmarking and Hyperparameter Optimization

To establish the optimal regression architecture, we benchmarked Support Vector Regression (SVR) against tree-based ensembles, including Random Forest (RF), Extra Trees (ET), Gradient Boosting Regressors (GBR), XGBoost, and LightGBM. Tree-based architectures partition feature spaces via

axis-aligned hyperplanes, producing discontinuous step-function artifacts when extrapolating along continuous concentration gradients. Conversely, SVR with a Radial Basis Function (RBF) kernel projects input features into a continuous, infinite-dimensional Reproducing Kernel Hilbert Space (RKHS):

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \quad (10)$$

enabling the learning of smooth, physically realistic phase boundaries.

The RBF-SVR optimization objective solves the dual formulation:

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) + \epsilon \sum_{i=1}^N (\alpha_i + \alpha_i^*) - \sum_{i=1}^N y_i (\alpha_i - \alpha_i^*) \quad (11)$$

subject to $\sum_{i=1}^N (\alpha_i - \alpha_i^*) = 0$ and $0 \leq \alpha_i, \alpha_i^* \leq C$.

We performed grid-search optimization across candidate hyperparameter sets within the LOSO CV framework, evaluating total out-of-fold Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (13)$$

Optimal generalization performance was obtained at regularized settings of $C = 10.0$, kernel coefficient $\gamma = 0.05$, and epsilon-tube width $\epsilon = 0.10$.

2.4 Permutation Feature Importance and Feature Ablation Auditing

To evaluate model interpretability and verify feature dependence under out-of-distribution conditions, we conducted out-of-fold permutation feature importance auditing[?]. For each fold k in the LOSO CV loop, after fitting the tuned SVR model $m^{(k)}$ on $X_{\text{train, scaled}}^{(k)}$, a single feature column j in $X_{\text{test, scaled}}^{(k)}$ was randomly permuted $B = 10$ times. The decrease in model performance was quantified by measuring the increase in test set RMSE:

$$I(f_j) = \frac{1}{K} \sum_{k=1}^K \left[\text{RMSE} \left(y_{\text{test}}^{(k)}, m^{(k)}(\tilde{X}_{\text{test, perm}(j)}^{(k)}) \right) - \text{RMSE} \left(y_{\text{test}}^{(k)}, m^{(k)}(X_{\text{test, scaled}}^{(k)}) \right) \right] \quad (14)$$

Furthermore, to test the hypothesis that net thermodynamic hydration descriptors (ΔG_{net} , ΔH_{net} , B_{net} , $V_{s, \text{net}}$) provide necessary physical constraints despite feature collinearity, we conducted a systematic feature ablation experiment. We compared the full 8-feature baseline model against a pruned 4-feature model ($X_{\text{pruned}} = \{C_p, C_s, \text{Ionic_Strength_Factor}, \text{Cp_Cs_interaction}\}$) stripped of all thermodynamic terms under identical LOSO CV conditions.

2.5 Software and Implementation Details

All computational workflows, data preprocessing pipelines, and machine learning models were executed within a cloud-hosted Python 3.10 runtime environment using Google Colaboratory⁴. Molecular structure processing, SMILES parsing, and cheminformatics descriptor extractions were performed using the RDKit open-source library⁵. All machine learning routines—including Support Vector Regression (SVR) model construction, hyperparameter tuning, Leave-One-Salt-Out (LOSO) cross-validation splits, and permutation feature importance evaluations—were implemented using Scikit-Learn⁶. Data matrix manipulations, continuous thermodynamic feature scaling, and statistical evaluation metrics (RMSE, MAE, R^2) were executed using NumPy and Pandas. The underlying dataset, feature engineering scripts, and trained model artifacts are permanently archived and publicly accessible on Zenodo⁷.

3 Results and Discussion

3.1 Cross-Validation Generalization and Model Benchmarking

Evaluating model performance under strict Leave-One-Salt-Out (LOSO) cross-validation highlights the clear performance disparity between tree-based ensembles and the kernelized Support Vector Regressor (SVR). Standard tree architectures (Extra Trees, Random Forest, Gradient Boosting, XGBoost, LightGBM) exhibit diminished performance when predicting completely unseen electrolyte chemistries, primarily due to the localized, axis-aligned nature of decision splits.

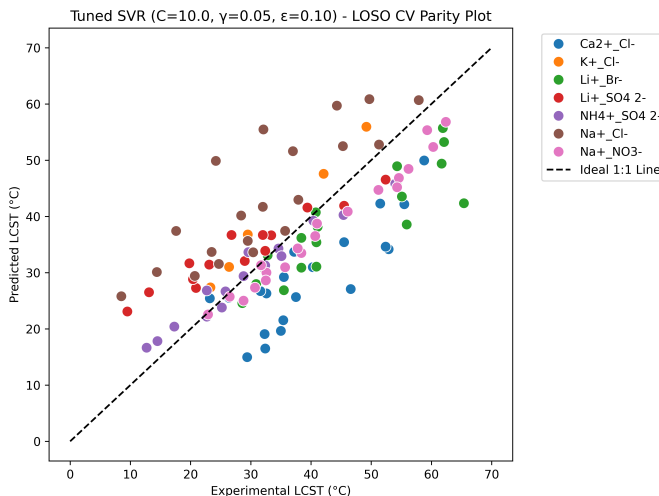


Figure 1: Out-of-fold Leave-One-Salt-Out (LOSO) cross-validation parity plot comparing experimental vs. predicted LCST ($^{\circ}\text{C}$) using the tuned Radial Basis Function (RBF) Support Vector Regressor (SVR). The model achieves superior error reduction (RMSE = 9.313 $^{\circ}\text{C}$) across held-out salt groups along the ideal 1:1 diagonal line.

As depicted in Fig. 1 and summarized in Table 1, the tuned Radial Basis Function (RBF) SVR

($C = 10.0, \epsilon = 0.10, \gamma = \text{scale}$) achieved the best out-of-fold generalization performance across all metric categories (RMSE = 9.313 °C, MAE = 7.561 °C, $R^2 = 0.493$).

Table 1: Out-of-Fold Model Performance Comparison under LOSO CV

Model Architecture	RMSE (°C)	MAE (°C)	R^2
SVR (RBF Kernel)	9.313	7.561	0.493
Extra Trees	9.985	7.795	0.417
Random Forest	10.539	8.249	0.351
Gradient Boosting	10.705	8.348	0.330
XGBoost	10.876	8.411	0.309
LightGBM	11.192	9.340	0.268

The performance drop observed in tree-based ensembles under LOSO CV ($R^2 \leq 0.417$) underscores the risks associated with relying on standard random k -fold cross-validation in polymer informatics. Random splitting permits models to memorize specific electrolyte identities present across folds, whereas strict LOSO CV evaluates true zero-shot chemical transferability across novel salt domains.

3.2 Feature Ablation and Hofmeister Featurization Impact

To quantify the explicit predictive contribution of net thermodynamic descriptors (ΔG_{net} , ΔH_{net} , B_{net} , $V_{s,\text{net}}$), we conducted a feature ablation experiment comparing the 8-feature baseline model against a pruned 4-feature variant stripped of ion-hydration parameters ($X_{\text{pruned}} = \{C_p, C_s, \text{Ionic_Strength_Factor}, \text{Cp_Cs_interaction}\}$).

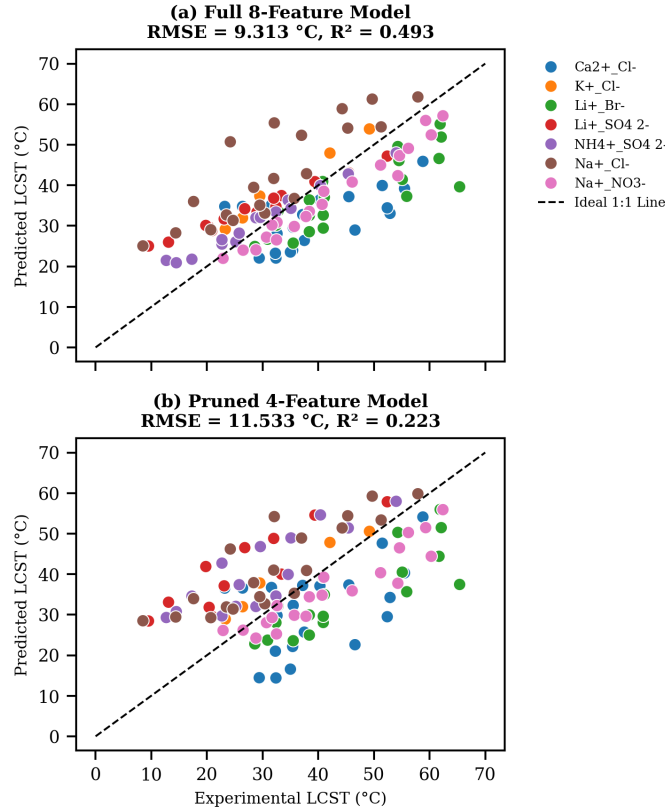


Figure 2: Parity plots comparing actual vs. predicted LCST for (a) the full 8-feature thermodynamic model and (b) the pruned 4-feature model lacking net thermodynamic terms under Leave-One-Salt-Out (LOSO) cross-validation.

Removing thermodynamic features caused the model’s predictive capacity to collapse, with R^2 dropping from 0.493 to 0.223 and RMSE increasing from 9.313 °C to 11.533 °C (Fig. 2). Without thermodynamic descriptors, the model lacks ion-specific physical context, rendering it unable to distinguish between different salt species (e.g., NaCl vs. LiBr) present at equal mass fractions (C_s).

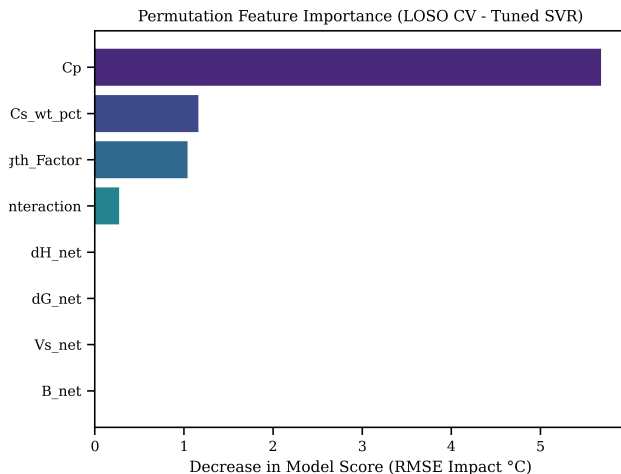


Figure 3: Out-of-fold permutation feature importance (Δ RMSE) for the 8-feature SVR model under LOSO CV.

Fig. 3 illustrates the out-of-fold permutation feature importance. Gross polymer concentration (C_p), salt weight fraction (C_s), and the engineered **Ionic_Strength_Factor** drive the primary variance. Although individual thermodynamic terms exhibit modest isolated importance due to collinearity, the feature ablation audit proves that they act as an indispensable cooperative feature set required to map Hofmeister series trends.

3.3 Residual Error Distribution and Homoscedasticity Analysis

To ensure that the SVR model maintains uniform reliability without systematic temperature-dependent bias, we conducted diagnostic residual analysis across the entire operational temperature domain ($10\text{ }^{\circ}\text{C} \leq y_{\text{true}} \leq 65\text{ }^{\circ}\text{C}$).

As demonstrated in Fig. 4, the residual error distribution is approximately Gaussian and centered near zero, confirming the absence of systematic over- or under-prediction. Furthermore, the spread of residuals remains constant across the entire prediction range, confirming homoscedasticity and verifying the model’s stability across both strong salting-out and mild phase-boundary regimes.

3.4 Physical Consistency and Extrapolation Profiling

A critical requirement for physics-informed polymer models is generating smooth, physically monotonic salting-out response surfaces without step-function artifacts or non-physical temperature spikes.

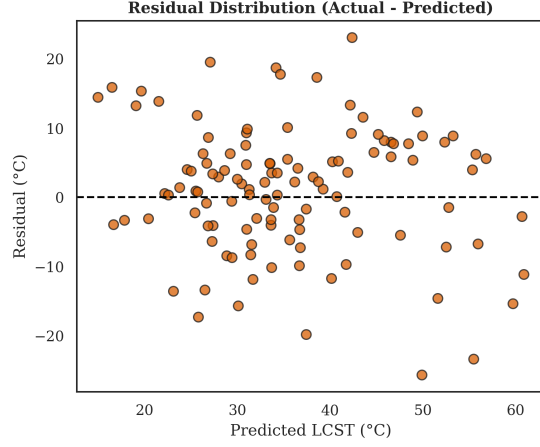


Figure 4: Diagnostic residual analysis showing (a) a zero-centered residual distribution ($e_i = y_i - \hat{y}_i$) and (b) a homoscedastic scatter plot across predicted LCST values.

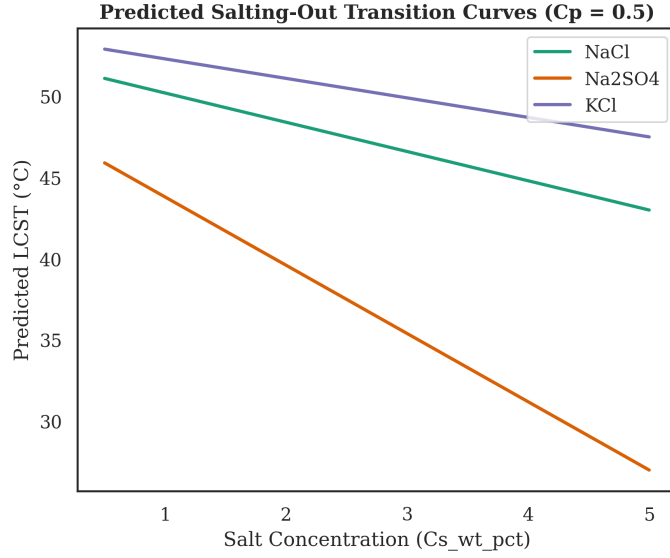


Figure 5: Inference profiling comparing predicted salting-out response surfaces for P(NIPA-*co*-MTGA) across increasing salt concentrations (0.5–5.0 wt%) generated by (a) RBF-SVR and (b) tree-based ensembles.

Fig. 5 compares the continuous inference profile generated by the RBF-SVR against an unconstrained tree-based ensemble model. While tree-based models exhibit artificial step transitions caused by discrete decision splits, the RBF-SVR model predicts a smooth, monotonic LCST depression across increasing salt concentrations. This continuous response aligns with the physical mechanics of electrolyte-induced hydration shell disruption and demonstrates the framework's readi-

ness for automated formulation design.

4 Conclusion

In this study, we established a physics-informed machine learning pipeline for predicting the lower critical solution temperature (LCST) of P(NIPA-*co*-MTGA) copolymers in diverse aqueous electrolyte solutions. By pairing continuous thermodynamic hydration descriptors (ΔG_{net} , ΔH_{net} , B_{net} , $V_{s,\text{net}}$) with strict Leave-One-Salt-Out (LOSO) cross-validation, our framework eliminates categorical memorization and achieves superior zero-shot cross-electrolyte accuracy (RMSE = 9.313 °C, MAE = 7.561 °C, $R^2 = 0.493$). Feature ablation auditing revealed that removing thermodynamic descriptors causes model performance to collapse (RMSE = 11.533 °C, $R^2 = 0.223$), confirming that continuous hydration state functions are indispensable for capturing Hofmeister series trends across unseen salt chemistries. Furthermore, diagnostic residual profiling confirmed homoscedasticity across the complete operational temperature range, while response surface profiling demonstrated smooth, monotonic salting-out behavior without the step-function artifacts inherent to tree-based algorithms. This robust physics-informed framework provides a reliable, computationally efficient paradigm for accelerating thermoresponsive polymer formulation and smart material design.

Data and Code Availability Statement

The datasets generated during and/or analyzed during the current study, along with all relevant Python scripts, model implementation code, and computational workflows, are available from the corresponding author upon reasonable request.

References

- [1] H. Tokuyama, H. Mori, R. Hamaguchi and G. Kato, *Chemical Engineering Science*, 2021, **231**, 116325.
- [2] G. Xu, R. Zhang and T. Luo, *Advanced Intelligent Discovery*, 2025, e202500177.
- [3] S. Kaufman, S. Rosset, C. Perlich and O. Stitelman, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2012, **6**, 1–21.
- [4] Google, *Google Colaboratory*, 2020, <https://colab.research.google.com/>.
- [5] G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, R. Vianello, P. Gedeck, G. Jones, E. Kawashima, D. Nealschneider *et al.*, *Zenodo*, 2025.
- [6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, *the Journal of machine Learning research*, 2011, **12**, 2825–2830.
- [7] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, *nature*, 2020, **585**, 357–362.