International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Data-Driven Machine Learning Surrogate Models for Predicting Dissolution and Gel Viscosity in Hot-Melt Extruded Polymeric Formulations: A Quality-by-Design Framework

Running Title: ML Surrogates for HME Dissolution and Viscosity

Mohammad Ali Riaz

Independent Researcher, 8472 Noreen Falls Drive, Colorado Springs, CO 80924, USA
Corresponding Author Email: *riazmohammadali[at]gmail.com*
ORCID: https://orcid.org/0009-0005-4366-1372

Abstract: ***Hot-melt extrusion (HME) is a solvent-free, continuous manufacturing technology that is increasingly central to the production of modified-release and abuse-deterrent oral dosage forms. The relationships among formulation composition, process conditions and critical quality attributes such as drug dissolution and gel viscosity are complex, nonlinear and costly to resolve by experimentation alone. This study develops and benchmarks tree-based machine-learning surrogate models, namely XGBoost, Random Forest and Gradient Boosting, for predicting the 30-min aqueous dissolution of hot-melt extruded polyethylene oxide/hydroxypropyl methylcellulose/carbomer matrices containing pseudoephedrine, and characterizes gel viscosity through exploratory regression. On a small curated dataset (42 records; five features), Random Forest achieved the best held-out performance ($R^2 = 0.485$), with pseudoephedrine drug loading identified as the dominant predictor. Five-fold hyperparameter optimization did not materially improve held-out accuracy, and negative cross-validated R^2 values revealed the variance inherent to small samples. Integrating a second data source (96 usable records; 17 features) yielded a markedly stronger optimized model ($R^2 = 0.984$; RMSE = 3.56%), although this gain partly reflects target leakage through the 60-min dissolution endpoint. SHAP analysis attributed predictive power to drug content uniformity and the 60-min readout, and residual diagnostics of the single-source model showed no systematic bias. These results establish a reusable, interpretable, quality-by-design workflow for surrogate modeling in pharmaceutical process engineering and highlight the safeguards needed before such models are used for formulation decisions.***

Keywords: hot-melt extrusion; machine learning; quality by design; dissolution prediction; SHAP

1. Introduction

Hot-melt extrusion (HME) has become one of the most versatile unit operations in pharmaceutical manufacturing. As a solvent-free, continuous process, HME combines thermal and mechanical energy to melt, mix and pump a formulation through a die, enabling the production of amorphous solid dispersions, taste-masked granules, sustained-release matrices, implants and films [1–7]. Its compatibility with continuous manufacturing and process analytical technology (PAT) has made HME a natural fit for modern quality-by-design (QbD) paradigms, in which product quality is designed into the process through a mechanistic or data-driven understanding of critical material attributes, critical process parameters and critical quality attributes (CQAs) [8–13]. Regulatory filings for HME products increasingly rely on this level of process understanding [14, 15].

Despite this promise, formulation development for HME remains largely empirical. The target product profile for modified-release and abuse-deterrent formulations depends on a delicate balance among polymer content, drug loading and processing history [16, 17]. Polyethylene oxide (PEO) and hydroxypropyl methylcellulose (HPMC) are hydrophilic, swellable, gel-forming polymers that govern both drug-release kinetics and the mechanical and viscous properties of the hydrated matrix [18–20]. High-molecular-weight PEO is particularly attractive for abuse-deterrent formulations because it forms a robust, high-viscosity gel that resists crushing, pulverization and solvent extraction [2, 21–23]. For pseudoephedrine-containing products, sustained release is a long-standing objective [24], and abuse deterrence is an increasingly prominent regulatory consideration for products with misuse potential [25]. The mapping from formulation variables to dissolution and gel viscosity is nonetheless nonlinear and coupled, and it is usually evaluated through costly dissolution and rheological experiments.

Machine learning (ML) offers a powerful complement to experimentation by learning these input–output relationships directly from data, producing surrogate models that can accelerate design-space exploration, screening and process control [26–28]. Tree-based ensemble methods, namely Random Forest [29], Gradient Boosting [30, 31] and Extreme Gradient Boosting (XGBoost) [32], are well suited to the small, tabular, heterogeneous datasets typical of pharmaceutical formulation studies, where interpretability and robustness to noise are paramount [33]. Recent work has demonstrated the feasibility of ML for predicting HME-induced amorphization and chemical stability [34], the physical stability of solid dispersions [35], formulation behavior and dissolution [36–38], the dissolution profiles of extended-release tablets [39] and the printability of drug-loaded filaments [40–42]. Interpretability methods such as Shapley additive explanations (SHAP) [43] allow these otherwise opaque ensembles to be interrogated for mechanistic insight, aligning them with the process-understanding expectations of QbD and with calls for

transparent models in high-stakes decisions [44]. Interpretable ML surrogates have also been combined with kinetic and computational fluid dynamics models to define risk-based design spaces within a computational QbD framework for complex drug- product manufacturing [45]. In continuous-manufacturing and digital-twin settings, such models can serve as soft sensors that map formulation and process inputs to CQAs in near real time [4, 8, 46–48].

The present study addresses three gaps. First, it provides a direct, side-by-side benchmark of XGBoost, Random Forest and Gradient Boosting for HME dissolution prediction within a single, consistent pipeline, complemented by exploratory regression of gel viscosity against formulation variables. Second, it couples hyperparameter optimization with interpretability (SHAP) and error analysis, moving beyond aggregate accuracy metrics to model diagnostics. Third, it critically evaluates the consequences of dataset construction, revealing both the gains available from data integration and the specific risk of target leakage when correlated endpoints are used as features.

The overall goal is a transferable, QbD-aligned and interpretable ML workflow that pharmaceutical process engineers can use to augment experimental screening with data-driven surrogate models.

2. Materials and Methods

2.1 Datasets and Experimental System

Two primary data sources were used. The first source (HME dataset; Table S1 in Online Resource 1) contains 42 records describing hot-melt extruded PEO/HPMC/carbomer matrices loaded with pseudoephedrine. Each record comprises five formulation/process features, namely PEO content (%), HPMC content (%), pseudoephedrine loading (mg), carbomer content (%) and tablet hardness (kp; 1 kp = 9.81 N), together with the response variables 30-min aqueous dissolution (%), ethanol and acetone extraction yields (%), gel viscosity (Pa·s), a binary gel- formation indicator and drug content uniformity (% relative standard deviation, RSD). Dissolution and gel viscosity served as the primary regression targets. All data analyzed in this study were secondary; no new laboratory experiments were performed by the author. The records were compiled from publicly available open-source datasets hosted on GitHub and DOI-registered data repositories, harmonized to common variable names and units and curated into a consolidated dataset that is openly available in figshare under a CC BY 4.0 license [49]. Materials, processing conditions and analytical methods (dissolution, gel viscosity, solvent extraction and hardness testing) are those reported by the original data contributors, and the reported values were used without modification. Where such details were not reported in the source records, they could not be independently verified, and this is acknowledged as a limitation (Section 4.5).

The second source (processing dataset) contained additional process and quality records and was likewise obtained from a public repository. Its 30-min dissolution column was harmonized with the HME dataset and the two tables were concatenated to form a combined table of 126 entries across partially overlapping features. After cleaning, the usable combined cohort comprised 96 samples and 17 features (Table 1).

2.2 Data Preprocessing

Data were loaded and cleaned using a purpose-built DataPipeline class implemented in Python. Cleaning replaced "NA" string values with missing values and, where a target column was specified, dropped rows with missing target values. For the single-source analysis, rows with any missing feature or target value were removed. For the combined analysis, missing feature values were imputed with the column mean to retain the maximum number of informative samples. Data were partitioned into training and held-out test sets using an 80/20 split with a fixed random seed (random_state = 42), yielding 33 training and 9 test samples for the single- source analysis. Features were standardized to zero mean and unit variance using StandardScaler fitted on the training partition only, to prevent information leakage from the test set.

2.3 Machine Learning Models

Three tree-ensemble regression algorithms were evaluated. Random Forest (RF) is a bagging ensemble in which each tree is trained on a bootstrap sample and each split considers a random subset of features [29]. Gradient Boosting (GB) constructs an additive ensemble of weak learners, with each new tree fitted to the residuals of the current ensemble [30]. XGBoost is a regularized, scalable implementation of gradient boosting with L1/L2 regularization [32].

Baseline hyperparameters were: XGBoost (n_estimators = 100, max_depth = 6, learning_rate = 0.1), RF (n_estimators = 100, max_depth = 15, min_samples_split = 5) and GB (n_estimators = 100, max_depth = 5, learning_rate = 0.1). Models were implemented with scikit-learn [50] and the xgboost library.

2.4 Model Training and Evaluation

Models were trained on the standardized training partition and evaluated on the held-out test partition. Four metrics were computed: the coefficient of determination, $R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2/\Sigma(y_i - \bar{y})^2$; the root-mean-square error, $RMSE = \sqrt{[(1/n)\Sigma(y_i - \hat{y}_i)^2]}$; the mean absolute error, $MAE = (1/n)\Sigma|y_i - \hat{y}_i|$; and the mean absolute percentage error, $MAPE = (100/n)\Sigma|(y_i - \hat{y}_i)/y_i|$, where y_i is the observed value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean observed value and n the number of test samples. R^2 was the primary objective metric; values below 0 indicate performance worse than predicting the mean.

2.5 Hyperparameter Optimization

Hyperparameters were optimized with GridSearchCV using five-fold cross-validation (CV) on the training data, maximizing R^2. Candidate grids covered learning_rate, max_depth and n_estimators for XGBoost and GB, and max_depth, min_samples_split and n_estimators for RF.

The configuration with the highest mean cross-validated R^2 was selected and re-evaluated on the held-out test set. Because model selection and performance estimation on the same folds

can bias error estimates [51], the held-out test set was not used during tuning.

2.6 Model Interpretability with SHAP

SHAP values were used to interpret the optimized RF models [43]. SHAP decomposes each prediction into additive feature contributions grounded in Shapley values from cooperative game theory. The tree-specific TreeExplainer was applied for computational efficiency [47].

2.7 Error Analysis

The residuals of the optimized RF model, defined as $r_i = y_i - \hat{y}_i$, were analyzed to assess model adequacy. A scatter plot of residuals versus predicted values was inspected for heteroscedasticity and systematic bias, and a histogram with a kernel-density estimate (KDE) was used to assess the residual distribution for approximate normality and centering near zero.

2.8 Software and Reproducibility

All analyses were performed in Python 3 using pandas (≥ 2.3), NumPy (≥ 2.4), scikit-learn (≥ 1.8) [50], xgboost (≥ 3.4), shap (≥ 0.52) and matplotlib (≥ 3.10). Data partitioning used random_state = 42 for reproducibility. The pipeline is modularized into DataPipeline (loading, cleaning, feature preparation) and ModelTrainer (splitting, training, evaluation, prediction) classes; the complete code, the consolidated dataset and the software environment are provided in Online Resource 1.

3. Results

3.1 Exploratory Analysis of Formulation–Property Relationships

Regression analysis of the HME dataset revealed a positive, approximately linear association between gel viscosity and both PEO content and HPMC content (Pearson r = 0.64 for each; Fig. 1). The influence of drug loading on gel viscosity was non-monotonic. Stratifying average gel viscosity by pseudoephedrine loading showed that intermediate loadings (3 mg and 10 mg) produced the highest average viscosities, whereas the lowest loading (0.075 mg) produced markedly lower viscosity. Dissolution spanned a wide dynamic range (approximately 8 –35% at 30 min). The fastest-dissolving records also showed high ethanol and acetone extraction yields (approximately 78–82%) and low gel viscosity (approximately 380–520 Pa·s), whereas slow-dissolving, high-viscosity records showed minimal extractable drug.

3.2 Baseline Model Performance

Three baseline models were trained on the 42-record HME dataset (33 training/9 test samples) using the five formulation/process features to predict 30-min dissolution. Held-out R^2 values were 0.404 (XGBoost), 0.485 (RF) and 0.419 (GB) (Table 2). Feature-importance analysis (Fig. 2) identified pseudoephedrine loading as the dominant predictor in both models; for XGBoost its importance was approximately 0.85, followed distantly by PEO content (approximately 0.09), whereas for RF tablet hardness ranked second.

To illustrate the deployment mode of the surrogates, three unseen candidate formulations (PEO/HPMC contents of 8/12%, 10/15% and 12/18% at a fixed 30-mg drug load) were passed through the trained baseline models. The models produced closely concordant predictions of about 20–25% dissolution, with wider inter-model disagreement for the higher-polymer candidate, reflecting greater extrapolation uncertainty in that region of the design space.

3.3 Hyperparameter Optimization

Five-fold grid-search optimization selected different configurations for each algorithm (Table 3), but the best cross-validated R^2 values were negative for all three models (−0.222, −0.351 and −0.187 for XGBoost, RF and GB, respectively; Table 2). On the held-out test set, the optimized models yielded R^2 values of 0.284, 0.474 and 0.397, respectively: XGBoost declined (0.404 to 0.284), whereas RF and GB changed marginally (Fig. 3).

3.4 Combined-Dataset Random Forest

An RF model trained on the combined dataset (96 usable samples; 17 features; mean imputation) achieved a test R^2 of 0.983 (RMSE = 3.76%, MAE = 2.38%, MAPE = 7.90%). After optimization, performance improved to R^2 = 0.984 (RMSE = 3.56%, MAE = 2.17%, MAPE = 7.54%), with a cross-validated R^2 of 0.978 (Table 4). The combined feature set included the 60- min dissolution endpoint.

3.5 Feature Importance and SHAP Interpretation

SHAP analysis of the optimized combined-data model identified drug content uniformity (% RSD) as by far the most influential feature, followed by the 60-min dissolution readout (data not shown). Dependence plots revealed nonlinear relationships, with SHAP contributions changing in magnitude and sign across the feature range. For the single-source model (Fig. 4), SHAP confirmed that drug loading dominates predictions, with polymer content and hardness contributing smaller effects, in agreement with the built-in feature importance (Fig. 2).

3.6 Error Analysis

Residual diagnostics for the optimized single-source RF model (n = 9 test samples) are shown in Fig. 5. Residuals were scattered about zero without an obvious trend across the prediction range, although a few points showed residuals of approximately ±7%. The residual histogram/KDE was approximately symmetric and centered near zero.

4. Discussion

4.1 Mechanistic Interpretation of Formulation Effects

The positive association between gel viscosity and polymer content is consistent with the role of PEO and HPMC as viscosity-enhancing, gel-forming excipients: a higher polymer mass fraction increases chain entanglement and

swelling, yielding a more viscous hydrated gel network [2, 18, 20]. The non-monotonic effect of drug loading suggests that the drug participates in the gel network indirectly through the local polymer-to-drug ratio. The inverse coupling between solvent extraction and gel robustness is meaningful for abuse-deterrent performance: a strong, high-viscosity gel resists solvent-based extraction, whereas a weakly gelling matrix is more readily extracted [2, 21–23]. The dominance of drug loading as a predictor of 30-min dissolution is likewise consistent with the established dependence of release rate on the drug-to- polymer ratio and the associated changes in matrix swelling, erosion and diffusion path length [20, 52].

4.2 Model Performance in the Small-Data Regime

RF achieved the strongest single-source performance, consistent with its robustness on small, high-variance datasets: bagging and feature subsampling mitigate the overfitting to which boosted models are prone when training data are scarce [29, 33]. These values must nevertheless be interpreted in the context of sample size, because a held-out estimate based on nine observations carries substantial variance [53, 54]. A negative cross-validated R^2 indicates that a model performed worse than a constant mean predictor, a signature of high variance and insufficient training data rather than algorithmic deficiency. Hyperparameter optimization driven by cross-validated R^2 can therefore select configurations that overfit the training folds without generalizing to unseen data, as illustrated by the decline of XGBoost after tuning. The consistently strong rank of RF supports its use as the default model class in this regime, while the in silico screening of candidate formulations highlights both the value of rapid screening and the need for calibrated uncertainty intervals outside densely sampled regions.

4.3 Data Integration and Target Leakage

The improvement from $R^2 \approx 0.47$ to 0.98 on the combined dataset must be interpreted with care. The near-parity between cross-validated and held-out performance suggests that the larger cohort is no longer variance-dominated, but two factors contribute: the combined dataset is larger and richer, and it also includes the 60-min dissolution endpoint as a feature. Because the 30- and 60-min dissolution values are correlated outputs of the same release process, using the later time point to predict the earlier one constitutes target leakage: the feature contains information unavailable at prediction time in a genuine prospective setting [55]. The reported R^2 of 0.984 therefore represents an upper bound on forward-predictive utility. Only features known before the prediction time point should be used, or a full profile should be predicted jointly rather than one time point from a later one. The dominance of content uniformity in the SHAP analysis is mechanistically plausible, because spatial heterogeneity in drug distribution modulates local dissolution rates and the reproducibility of 30-min release.

4.4 Engineering Implications and Comparison with Prior Work

This study demonstrates a complete, reusable data-driven QbD workflow: dataset consolidation, benchmark modeling, hyperparameter optimization, interpretability and diagnostic error analysis. The results align with and extend recent demonstrations of ML for HME [34, 35] and dissolution prediction [37–39], and they complement computational QbD frameworks that pair interpretable surrogates with mechanistic models [45], targeting continuous quality attributes that map directly onto CQAs and can serve as soft-sensor models in a digital-twin or real-time-release setting [4, 8, 12, 46]. Two practical recommendations follow. First, in small- data regimes RF should be the default baseline, with boosting methods reserved for larger, cleaner datasets. Second, data integration is the most effective lever for improving predictive power but must be accompanied by rigorous leakage auditing. The modular DataPipeline/ ModelTrainer architecture is amenable to industrial adoption as a soft sensor, but regulatory-facing surrogates should be (i) trained on a design space that brackets the intended operating range, (ii) validated on independent, prospectively manufactured batches and (iii) accompanied by quantified uncertainty so that extrapolation is flagged rather than silently accepted [9, 11, 44].

4.5 Limitations and Future Work

Several limitations bound these findings. The data are secondary and were compiled from public repositories, so the experimental conditions under which they were generated could not be independently verified or standardized across sources. The single-source dataset is small (42 records), and the residual analysis rests on nine test samples. The combined dataset contains missing values handled by mean imputation, which can attenuate variance and bias associations [56]; because the imputation means were not explicitly restricted to the training partition, minor information sharing between partitions cannot be excluded. The combined model's use of the 60- min endpoint introduces target leakage, and a single 80/20 split was used, whereas repeated or nested CV would provide more stable estimates [51, 54]. Future work should expand the dataset through active learning [28], replace mean imputation with principled multiple imputation [56, 57], adopt nested CV with conformal prediction for calibrated intervals, couple the tree ensembles with mechanistic release models such as Korsmeyer–Peppas or Higuchi kinetics [52, 58] to improve extrapolation, link inputs to in-line PAT signals [13] and validate predictions prospectively on independently manufactured batches.

5. Conclusion

This study developed and systematically benchmarked an ML workflow for predicting a CQA, the 30-min aqueous dissolution of hot-melt extruded PEO/HPMC/carbomer formulations, and characterized gel viscosity through exploratory regression. On a small single-source dataset, RF delivered the best held-out performance ($R^2 = 0.485$), with pseudoephedrine drug loading identified as the dominant predictive feature. Hyperparameter optimization did not overcome the fundamental limitation of small sample size, as shown by negative cross-validated R^2 across all models; algorithmic tuning cannot substitute for data. Integration of a second data source improved the optimized RF to $R^2 = 0.984$, with SHAP analysis attributing predictive power primarily to drug content uniformity and the 60-min dissolution readout;

the latter highlights a target-leakage caveat that must be managed in dissolution-profile modeling. Residual diagnostics of the single-source model showed no systematic bias. Overall, the work provides a transferable, interpretable, QbD-aligned template for surrogate modeling in pharmaceutical HME and concrete guidance for moving ML from exploratory analysis to process-engineering decision support.

Acknowledgements
Not applicable.

Funding
This research received no external funding.

Conflict of Interest
The author declares no conflicts of interest.

Author Contributions
Mohammad Ali Riaz: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – Original Draft, Writing – Review and Editing, Visualization, Project administration. Supervision and Funding acquisition: not applicable (single author; no external funding).

Ethics Statement
This study used only publicly available formulation and process data; no human participants or animals were involved, and ethics approval was not required.

Data Availability
The consolidated HME dataset, complete analysis code and software environment are openly available in figshare under a CC BY 4.0 license at https://doi.org/10.6084/m9.figshare.33946321 and are also provided as Online Resource 1 (analysis code, consolidated dataset and software environment). The underlying source data, including the processing dataset, are publicly available from the open-source repositories described in Section 2.1.

AI Declaration
An AI tool was used for AI-assisted copy editing (readability, grammar and formatting) of author-written text. No AI tool was used to generate data, analyses, figures or scientific content.

The author reviewed all edits and takes full responsibility for the final text.

References

[1] Crowley MM, Zhang F, Repka MA, Thumma S, Upadhye SB, Battu SK, et al. Pharmaceutical applications of hot-melt extrusion: part I. Drug Dev Ind Pharm. 2007;33:909–26. https://doi.org/10.1080/03639040701498759.

[2] Repka MA, Majumdar S, Battu SK, Srirangam R, Upadhye SB. Applications of hot-melt extrusion for drug delivery. Expert Opin Drug Deliv. 2008; 5: 1357–76. https://doi.org/10.1517/17425240802583421.

[3] Maniruzzaman M, Boateng JS, Snowden MJ, Douroumis D. A review of hot-melt extrusion: process technology to pharmaceutical products. ISRN Pharm. 2012; 2012: 436763. https://doi.org/10.5402/2012/436763.

[4] Maniruzzaman M. Pharmaceutical Applications of Hot-Melt Extrusion: Continuous Manufacturing, Twin-Screw Granulations, and 3D Printing. Pharmaceutics. 2019; 11: 218. https://doi.org/10.3390/pharmaceutics11050218.

[5] Repka MA, Battu SK, Upadhye SB, Thumma S, Crowley MM, Zhang F, et al. Pharmaceutical applications of hot-melt extrusion: part II. Drug Dev Ind Pharm. 2007; 33: 1043–57. https://doi.org/10.1080/03639040701525627.

[6] Patil H, Tiwari RV, Repka MA. Hot-Melt Extrusion: from Theory to Application in Pharmaceutical Formulation. AAPS PharmSciTech. 2016; 17: 20–42. https://doi.org/10.1208/s12249-015-0360-7.

[7] Tambe S, Jain D, Agarwal Y, Amin P. Hot-melt extrusion: Highlighting recent advances in pharmaceutical applications. J Drug Deliv Sci Technol. 2021; 63: 102452. https://doi.org/10.1016/j.jddst.2021.102452.

[8] Yu LX. Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. Pharm Res. 2008; 25: 781–91. https://doi.org/10.1007/s11095-007-9511-1.

[9] International Council for Harmonisation. ICH Harmonised Tripartite Guideline Q8(R2): pharmaceutical development. Geneva: ICH; 2009. https://database.ich.org/sites/default/files/Q8%28R2%29%20Guideline.pdf.

[10] Repka MA, Shah S, Lu J, Maddineni S, Morott J, Patwardhan K, et al. Melt extrusion: process to product. Expert Opin Drug Deliv. 2012; 9: 105–25. https://doi.org/10.1517/17425247.2012.642365.

[11] Yu LX, Amidon G, Khan MA, Hoag SW, Polli J, Raju GK, et al. Understanding Pharmaceutical Quality by Design. AAPS J. 2014; 16: 771–83. https://doi.org/10.1208/s12248-014-9598-3.

[12] Lee SL, O'Connor TF, Yang X, Cruz CN, Chatterjee S, Madurawe RD, et al. Modernizing Pharmaceutical Manufacturing: from Batch to Continuous Production. J Pharm Innov. 2015;10:191–9. https://doi.org/10.1007/s12247-015-9215-8.

[13] Saerens L, Dierickx L, Lenain B, Vervaet C, Remon JP, De Beer T. Raman spectroscopy for the in-line polymer–drug quantification and solid state characterization during a -pharmaceutical hot-melt extrusion process. Eur J Pharm Biopharm. 2011; 77: 158–63. - https://doi.org/10.1016/j.ejpb.2010.09.015.

[14] Simões MF, Pinto RMA, Simões S. Hot-melt extrusion in the pharmaceutical industry: toward filing a new drug application. Drug Discov Today. 2019;24:1749–68. https://doi.org/10.1016/j.drudis.2019.05.013.

[15] Simões MF, Pinto RMA, Simões S. Hot-Melt Extrusion: a Roadmap for Product Development. AAPS PharmSciTech. 2021; 22: 184. https://doi.org/10.1208/s12249-021-02017-7.

[16] Feng X, Zhang F. Twin-screw extrusion of sustained-release oral dosage forms and medical implants. Drug Deliv Transl Res. 2018; 8: 1694–713. https://doi.org/10.1007/s13346-017-0461-9.

[17] Maincent J, Williams RO 3rd. Sustained-release

amorphous solid dispersions. Drug Deliv Transl Res. 2018; 8: 1714–25. https://doi.org/10.1007/s13346-018-0494-8.

[18] Colombo P, Bettini R, Santi P, De Ascentiis A, Peppas NA. Analysis of the swelling and release mechanisms from drug delivery systems with emphasis on drug solubility and water transport. J Control Release. 1996;39:231–7. https://doi.org/10.1016/0168-3659(95)00158-1.

[19] Huang S, O'Donnell KP, Keen JM, Rickard MA, McGinity JW, Williams RO 3rd. A new extrudable form of hypromellose: AFFINISOL™ HPMC HME. AAPS PharmSciTech. 2016;17:106–19. https://doi.org/10.1208/s12249-015-0395-9.

[20] Siepmann J, Peppas NA. Modeling of drug release from delivery systems based on hydroxypropyl methylcellulose (HPMC). Adv Drug Deliv Rev. 2012;64:163–74. https://doi.org/10.1016/j.addr.2012.09.028.

[21] Maddineni S, Battu SK, Morott J, Majumdar S, Repka MA. Formulation optimization of hot- melt extruded abuse deterrent pellet dosage form utilizing design of experiments. J Pharm Pharmacol. 2014; 66: 309–22. https://doi.org/10.1111/jphp.12129.

[22] Butreddy A, Sarabu S, Dumpa N, Bandari S, Repka MA. Extended release pellets prepared by hot melt extrusion technique for abuse deterrent potential: Category-1 in-vitro evaluation. Int J Pharm. 2020;587:119624. https://doi.org/10.1016/j.ijpharm.2020.119624.

[23] Baronsky-Probst J, Möltgen CV, Kessler W, Kessler RW. Process design and control of a twin screw hot melt extrusion for continuous pharmaceutical tamper-resistant tablet production. Eur J Pharm Sci. 2016;87:14–21. https://doi.org/10.1016/j.ejps.2015.09.010.

[24] Ishida M, Abe K, Hashizume M, Kawamura M. A novel approach to sustained pseudoephedrine release: Differentially coated mini-tablets in HPMC capsules. Int J Pharm. 2008; 359: 46–52. https://doi.org/10.1016/j.ijpharm.2008.03.034.

[25] U.S. Food and Drug Administration. Abuse-deterrent opioids—evaluation and labeling: guidance for industry. Silver Spring (MD): FDA; 2015. https://www.fda.gov/media/84819/download.

[26] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018; 559: 547–55. https://doi.org/10.1038/s41586-018-0337-2.

[27] Lou H, Lian B, Hageman MJ. Applications of Machine Learning in Solid Oral Dosage Form Development. J Pharm Sci. 2021; 110: 3150–65. https://doi.org/10.1016/j.xphs.2021.04.013.

[28] Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Adv Drug Deliv Rev. 2021; 175: 113806. https://doi.org/10.1016/j.addr.2021.05.016.

[29] Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32. https://doi.org/10.1023/a:1010933404324.

[30] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001; 29: 1189–232. https://doi.org/10.1214/aos/1013203451.

[31] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013; 7: 21. https://doi.org/10.3389/fnbot.2013.00021.

[32] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[33] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003; 43: 1947–58. https://doi.org/10.1021/ci034160g.

[34] Jiang J, Lu A, Ma X, Ouyang D, Williams RO 3rd. The applications of machine learning to predict the forming of chemically stable amorphous solid dispersions prepared by hot-melt extrusion. Int J Pharm X. 2023;5:100164. https://doi.org/10.1016/j.ijpx.2023.100164.

[35] Han R, Xiong H, Ye Z, Yang Y, Huang T, Jing Q, et al. Predicting physical stability of solid dispersions by machine learning techniques. J Control Release. 2019;311-312:16-25 https://doi.org/10.1016/j.jconrel.2019.08.030.

[36] Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharm Sin B. 2019; 9: 177–85. https://doi.org/10.1016/j.apsb.2018.09.010.

[37] Wang S, Di J, Wang D, Dai X, Hua Y, Gao X, et al. State-of-the-Art Review of Artificial Neural Networks to Predict, Characterize and Optimize Pharmaceutical Formulation. Pharmaceutics. 2022; 14: 183. https://doi.org/10.3390/pharmaceutics14010183.

[38] Protopapa C, Siamidi A, Eneli AA, Elbadawi M, Vlachou M. Machine Learning Predicts Drug Release Profiles and Kinetic Parameters Based on Tablets' Formulations. AAPS J. 2025; 27: 124. https://doi.org/10.1208/s12248-025-01101-1.

[39] Lourenço AS, Schuster T, Lopes JA, Kirsch A. A non-linear modelling approach to predict the dissolution profile of extended-release tablets. Eur J Pharm Sci. 2025;204:106976. https://doi.org/10.1016/j.ejps.2024.106976.

[40] Muñiz Castro B, Elbadawi M, Ong JJ, Pollard T, Song Z, Gaisford S, et al. Machine learning predicts 3D printing performance of over 900 drug delivery systems. J Control Release. 2021;337:530–45. https://doi.org/10.1016/j.jconrel.2021.07.046.

[41] Elbadawi M, McCoubrey LE, Gavins FKH, Ong JJ, Goyanes A, Gaisford S, et al. Disrupting 3D printing of medicines with machine learning. Trends Pharmacol Sci. 2021;42:745–57. https://doi.org/10.1016/j.tips.2021.06.002.

[42] Elbadawi M, McCoubrey LE, Gavins FKH, Ong JJ, Goyanes A, Gaisford S, et al. Harnessing artificial intelligence for the next generation of 3D printed medicines. Adv Drug Deliv Rev. 2021; 175: 113805. https://doi.org/10.1016/j.addr.2021.05.015.

[43] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Red Hook (NY): Curran Associates; 2017. p. 4765–74. https://proceedings.neurips.cc/paper/2017/hash/8a20a86

21978632d76c43dfd28b67767- Abstract.html.
[44] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.
[45] Riaz MA. Quality by design (QbD) framework for manufacturing peptide-drug conjugates. Int J Sci Res. 2026;15:974–82. https://doi.org/10.21275/SR26723215912. Supplementary material: https://doi.org/10.6084/m9.figshare.33052985.
[46] Chen Y, Yang O, Sampat C, Bhalode P, Ramachandran R, Ierapetritou M. Digital Twins in Pharmaceutical and Biopharmaceutical Manufacturing: A Literature Review. Processes. 2020; 8:1088. https://doi.org/10.3390/pr8091088.
[47] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
[48] Rantanen J, Khinast J. The Future of Pharmaceutical Manufacturing Sciences. J Pharm Sci. 2015;104:3612–38. https://doi.org/10.1002/jps.24594.
[49] Riaz MA. Predicting dissolution and gel viscosity in hot-melt extruded polymeric formulations: a quality-by-design framework [dataset and supplementary information]. figshare; 2026. https://doi.org/10.6084/m9.figshare.33946321.
[50] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30. https://jmlr.org/papers/v12/pedregosa11a.html.
[51] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91.
[52] Korsmeyer RW, Gurny R, Doelker E, Buri P, Peppas NA. Mechanisms of solute release from porous hydrophilic polymers. Int J Pharm. 1983;15:25–35. https://doi.org/10.1016/0378- 5173(83)90064-9.
[53] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019;14:e0224365. https://doi.org/10.1371/journal.pone.0224365.
[54] Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage. 2018;180:68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061.
[55] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining. ACM Trans Knowl Discov Data. 2012;6:1–21. https://doi.org/10.1145/2382577.2382579.
[56] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. https://doi.org/10.1136/bmj.b2393.
[57] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45:1–67. https://doi.org/10.18637/jss.v045.i03.
[58] Higuchi T. Rate of Release of Medicaments from Ointment Bases Containing Drugs in Suspension. J Pharm Sci. 1961; 50: 874–5. https://doi.org/10.1002/jps.2600501018.

Figure Legends

Figure 1 Exploratory regression of gel viscosity against (a) PEO content and (b) HPMC content in the HME dataset (n = 42). Circles, individual records; line, least-squares fit (Pearson r = 0.64 for each panel)

Figure 2 Baseline feature importance for 30-min dissolution prediction by (a) XGBoost and (b) Random Forest. Tablet hardness is shown in kp (1 kp = 9.81 N)

Figure 3 Held-out R^2 of XGBoost, Random Forest and Gradient Boosting before (baseline) and after five-fold GridSearchCV optimization (HME dataset; n = 42, 33 training/9 test samples)

Figure 4 SHAP beeswarm summary plot for the optimized single-source Random Forest. Each point is one test sample; features are ordered by global importance and colour indicates feature value from low (blue) to high (red)

Figure 5 Residual diagnostics for the optimized single-source Random Forest (n = 9 test samples): (a) residuals versus predicted 30-min dissolution; (b) residual histogram with a kernel- density estimate

Table 1: Summary of the datasets used in this study

Dataset	Records	Features (modeling)	Primary target(s)
HME (Online Resource 1)	42	5 (PEO %, HPMC %, drug loading, carbomer %, hardness)	30-min dissolution; gel viscosity
Processing (public repository)	—	Process and quality variables	30-min dissolution (harmonized)
Combined	126 (96 usable)	17 (mean-imputed)	30-min dissolution

HME, hot-melt extrusion; HPMC, hydroxypropyl methylcellulose; PEO, polyethylene oxide

Table 2: Held-out predictive performance for 30-min dissolution (HME dataset, 42 records)

Model	Baseline R^2	Optimized R^2	Best CV R^2
XGBoost	0.404	0.284	−0.222
Random Forest	0.485	0.474	−0.351
Gradient Boosting	0.419	0.397	−0.187

CV, cross-validation (five-fold, training partition); held-out test set n = 9

Table 3: Best hyperparameters selected by five-fold GridSearchCV

Model	Selected hyperparameters
XGBoost	learning_rate = 0.01, max_depth = 5, n_estimators = 50
Random Forest	max_depth = 10, min_samples_split = 10, n_estimators = 200
Gradient Boosting	learning_rate = 0.01, max_depth = 3, n_estimators = 100

Table 4: Random Forest performance on the combined dataset (96 usable samples, 17 features)

Metric	Baseline	Optimized
Test R^2	0.983	0.984
RMSE (%)	3.76	3.56
MAE (%)	2.38	2.17
MAPE (%)	7.90	7.54
Best CV R^2	—	0.978

MAE, mean absolute error; MAPE, mean absolute percentage error; RMSE, root-mean-square error. The feature set includes the 60-min dissolution endpoint (see Discussion on target leakage)

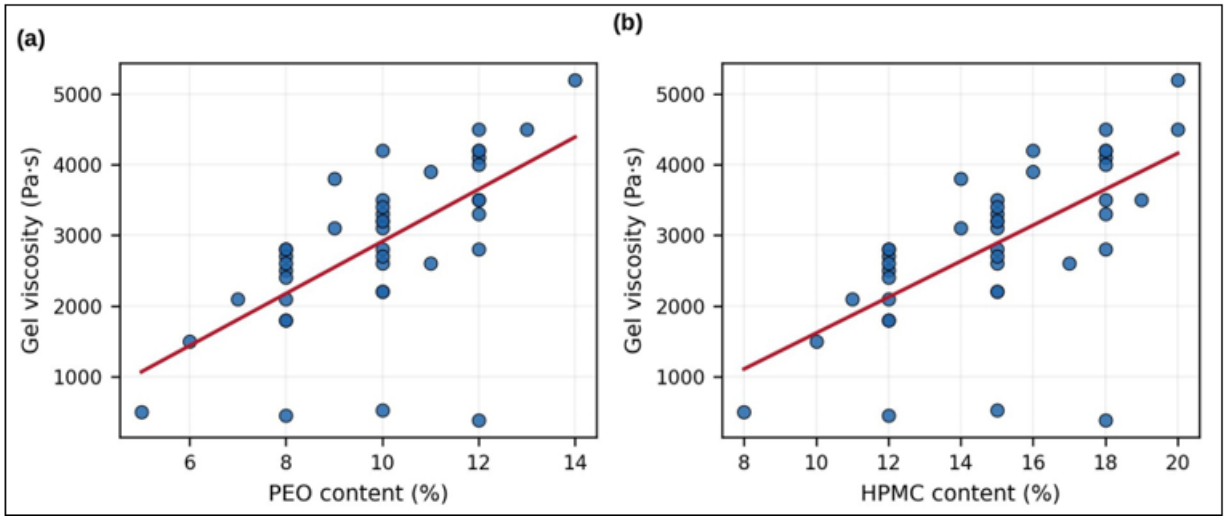


Figure 1 Exploratory regression of gel viscosity against (a) PEO content and (b) HPMC content in the HME dataset (n = 42). Circles, individual records; line, least-squares fit (Pearson r = 0.64 for each panel)

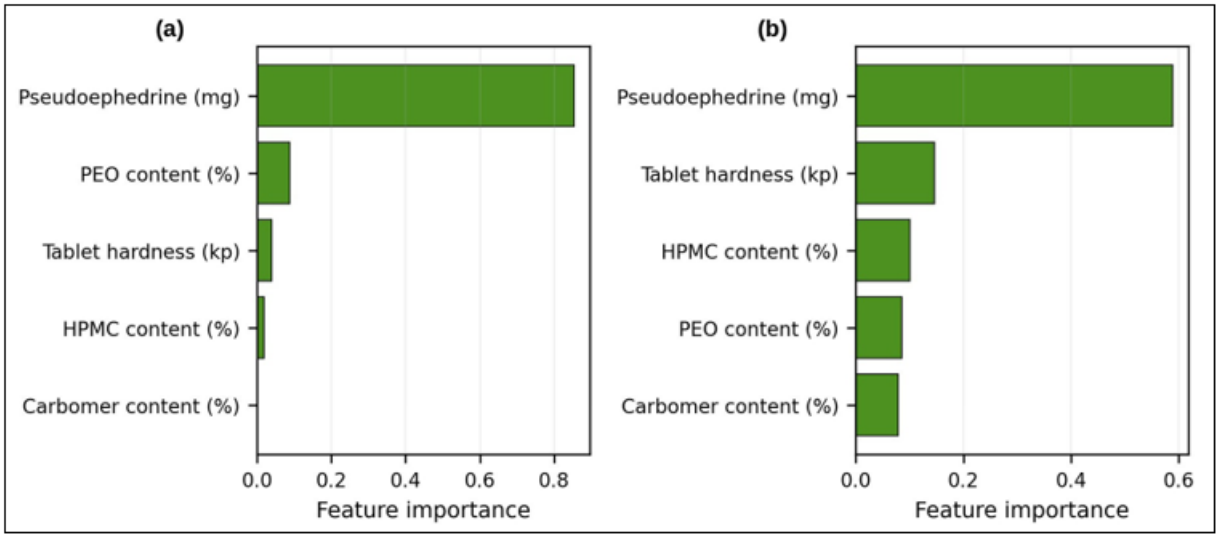


Figure 2: Baseline feature importance for 30-min dissolution prediction by (a) XGBoost and (b) Random Forest. Tablet hardness is shown in kp (1 kp = 9.81 N)

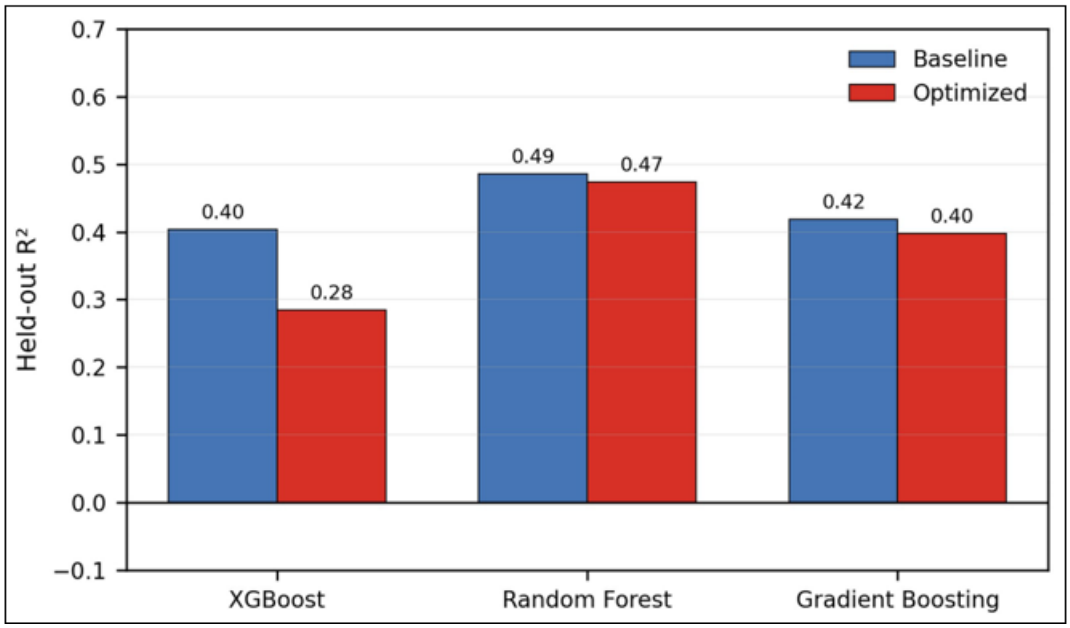


Figure 3: Held-out R^2 of XGBoost, Random Forest and Gradient Boosting before (baseline) and after five-fold GridSearchCV optimization (HME dataset; n = 42, 33 training/9 test samples)

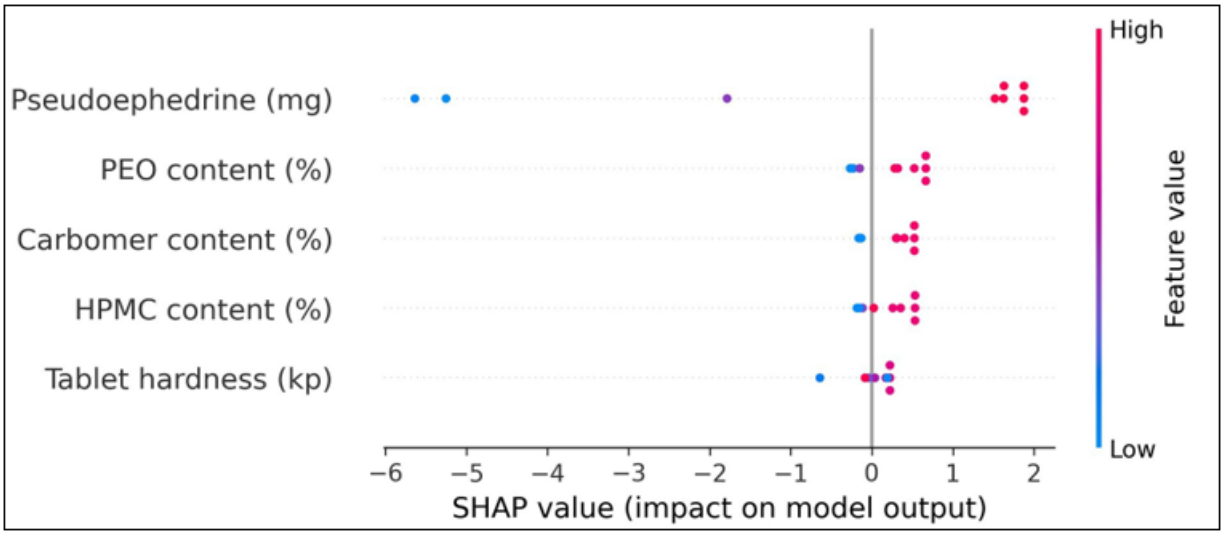


Figure 4 SHAP beeswarm summary plot for the optimized single-source Random Forest. Each point is one test sample; features are ordered by global importance and colour indicates feature value from low (blue) to high (red)

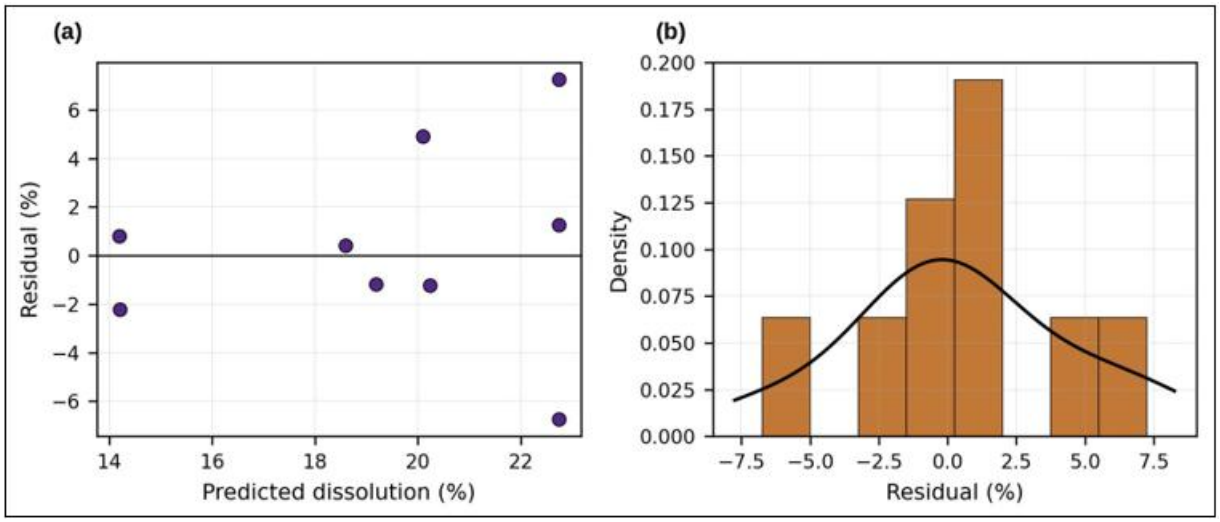


Figure 5: Residual diagnostics for the optimized single-source Random Forest (n = 9 test samples): (a) residuals versus predicted 30-min dissolution; (b) residual histogram with a kernel-density estimate