Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Machine Learning Using Taguchi-QbD Data for Digital Design Space Mapping of Pregabalin Extended-Release Tablets

Deepender
School of Computer Applications, Lovely Professional University, Punjab, India https://orcid.org/0000-0002-0529-4007

Tarandeep Singh Walia
School of Computer Applications, Lovely Professional University, Punjab, India https://orcid.org/0000-0001-8127-3112

Sreenath Sriram
Vaagdevi College of Pharmacy Autonomous, Ramnagar, Hanamkonda, Telangana, India, 506001 https://orcid.org/0009-0007-3705-7536

Dr. Nabeel Ahmad
nabeel.biotech@gmail.com

Institute of Bio Science & Technology, Shri Ramswaroop Memorial University Lucknow , Uttar Pradesh-225003

Dr. Sonal Solanki
drsonalsolanki25@gmail.com

Hon, Shri Babanrao Pachpute Vichardhara Trust's Group of Institutions, Faculty of Pharmacy, Kashti, Ta-Shrigonda, Ahilyanagar, Maharastra, India, 414071 https://orcid.org/0009-0007-5535-3881

Dr. Nikunj Solanki
drnikunjsolanki@gmail.com

Hon, Shri Babanrao Pachpute Vichardhara Trust's Group of Institutions, Faculty of Pharmacy, Kashti, Ta-Shrigonda, Ahilyanagar, Maharastra, India, 414071 https://orcid.org/0009-0003-2271-9382

Dr Jyoti Maithani Kalra
School of Pharmaceutical Sciences, Shri Guru Ram Rai University, Dehradun, Uttarakhand, India https://orcid.org/0000-0002-4960-6551

Dr. Yamsani Shravan
Vaagdevi College of Pharmacy Autonomous, Ramnagar, Hanamkonda, Telangana, India, 506001 https://orcid.org/0000-0001-9729-6175

Dr. Sreekanth Thota

Vaagdevi College of Pharmacy Autonomous, Ramnagar, Hanamkonda, Telangana, India, 506001 https://orcid.org/0000-0002-7501-3987

Dr Kapil Kalra

Alpine College of Management and Technology Dehradun, Uttarakhand, India https://orcid.org/0000-0001-7739-7614

Kiran Dangwal

Alpine College of Management and Technology Dehradun, Uttarakhand, India https://orcid.org/0009-0007-8717-7691

Dr Shibha Shiphalika Morris

Alpine College of Management and Technology Dehradun, Uttarakhand, India https://orcid.org/0000-0002-1094-4203

Neetu Rajpoot

Alpine College of Management and Technology Dehradun, Uttarakhand, India https://orcid.org/0000-0003-3519-4610

Dr Saumya Mishra

Alpine College of Management and Technology Dehradun, Uttarakhand, India https://orcid.org/0009-0003-3467-4951

Research Article

Keywords: Explainable machine learning; Taguchi-QbD; Pregabalin extended-release tablets; Digital design space mapping; Permutation feature importance; Extended-release matrix tablets

Posted Date: May 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9683753/v2

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

Additional Declarations: The authors declare no competing interests.

Explainable Machine Learning Using Taguchi-QbD Data for Digital Design Space Mapping of Pregabalin Extended-Release Tablets

Deepender[1], Tarandeep Singh Walia[1], Sreenath Sriram[2], Dr. Nabeel Ahmad[3*], Dr. Sonal Solanki[4*], Dr. Nikunj Solanki[4*], Dr Jyoti Maithani Kalra[5], Dr. Yamsani Shravan[2], Dr. Sreekanth Thota[2], Dr Kapil Kalra[6], Kiran Dangwal[6], Dr Shibha Shiphalika Morris [6], Neetu Rajpoot[6], Dr Saumya Mishra[6]

1. Faculty of Computing, Guru Kashi University, Bathinda, Punjab, India

2. Vaagdevi College of Pharmacy Autonomous, Ramnagar, Hanamkonda, Telangana, India, 506001

3. Institute of Bio Science & Technology, Shri Ramswaroop Memorial University Lucknow , Uttar Pradesh-225003

4. Department of Pharmaceutics, Hon, Shri Babanrao Pachpute Vichardhara Trust's Group of Institutions, Faculty of Pharmacy, Kashti, Ta- Shrigonda, Ahilyanagar, Maharastra, India, 414071

5. School of Pharmaceutical Sciences, Shri Guru Ram Rai University, Dehradun, Uttarakhand, India

6. Alpine College of Management and Technology Dehradun, Uttarakhand, India

Corresponding Authors

1. Dr. Nikunj Solanki – drnikunjsolanki@gmail.com

2. Dr. Sonal Solanki – drsonalsolanki25@gmail.com

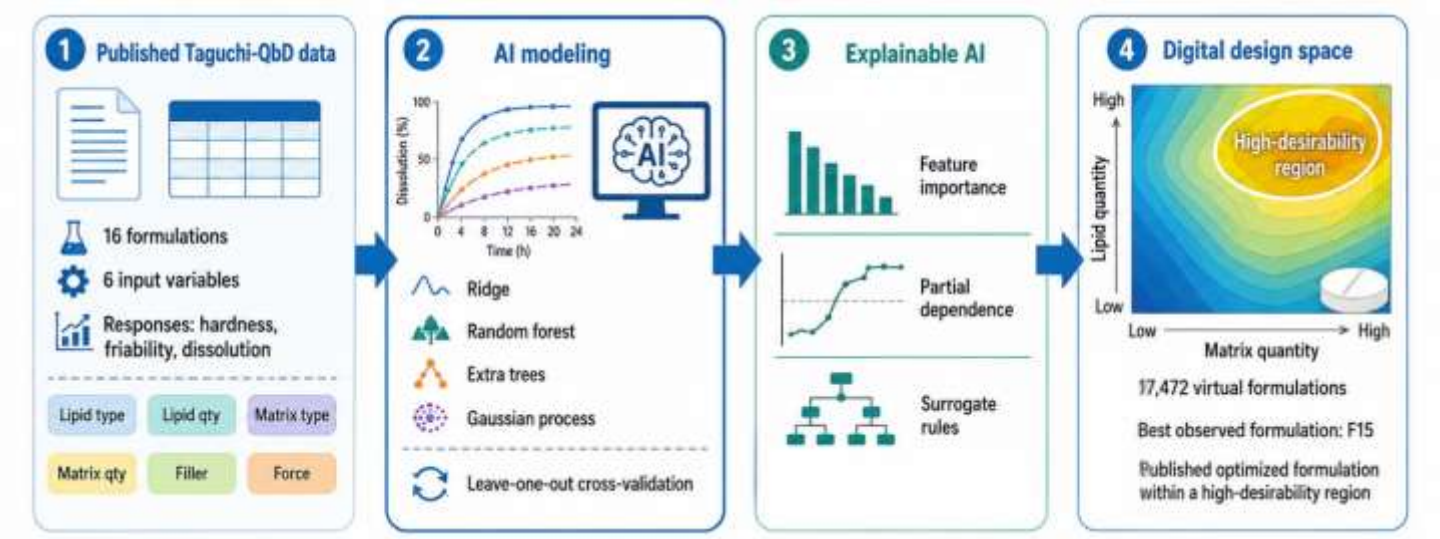
3. Dr. Nabeel Ahmad - nabeel.biotech@gmail.com

Deepender - ORCID: 0000-0002-0529-4007	Email - deependerduhan6@gmail.com
Tarandeep Singh Walia- ORCID : NA	Email- taran_walia2k@yahoo.com
Sreenath Sriram- ORCID: 0009-0007-3705-7536-	Email- ssriram@evaricpharma.com
Dr. Nabeel Ahmad – ORCID:- NA	Email - nabeel.biotech@gmail.com
Dr. Sonal Solanki – ORCID : 0009-0007-5535-3881	Email: drsonalsolanki25@gmail.com
Dr. Nikunj Solanki – ORCID: 0009-0003-2271-9382	Email: drnikunjsolanki@gmail.com
Dr Jyoti Maithani Kalra– ORCID: 0000-0002-4960-6551	Email -jyotimk123@gmail.com
Dr. Yamsani Shravan. ORCID: 0000- 0001-9729-6175	Email - shravanyamsani@gmail.com
Dr. Sreekanth Thota, ORCID: 0000-0002-7501-3987	Email - stsreekaththota@gmail.com
Dr Kapil Kalra - ORCID- 0000-0001-7739-7614	Email -kapil752003@yahoo.com
Kiran Dangwal, ORCID: - 0009-0007-8717-7691	Email- kiran97dangwal@gmail.com
Dr Shibha Shiphalika Morris , ORCID: 0000-0002-1094-4203	Email- shibamorris14@gmail.com
Neetu Rajpoot, ORCID: 0000-0003-3519-4610	Email:- neeturajpoot30@gmail.com
Dr Saumya Mishra ORCID: 0009-0003-3467-4951	Email: - Saumyamishra1988@gmail.com

Abstract

This study applied explainable machine learning to published Taguchi-QbD data for pregabalin extended-release tablets to support formulation understanding and digital design space mapping. No new experimental batches were prepared; the analysis used formulation-response data from a previously developed pregabalin extended-release tablet system. A secondary data set containing 16 Taguchi formulations was prepared using six formulation and process variables: lipid material type, lipid quantity, hydrophilic matrix type, hydrophilic matrix quantity, filler type, and compression force. Hardness, friability, and dissolution at 1, 2, 4, and 8 h were selected as modeling responses. Ridge regression, random forest regression, extra trees regression, and Gaussian process regression were evaluated using leave-one-out cross-validation. Model-independent explainability was performed using permutation feature importance, partial dependence analysis, and surrogate decision-tree rules. A desirability-based virtual formulation screening approach was then used to identify high-desirability regions within the original experimental design limits. Dissolution at 2 h and 4 h showed the most useful cross-validated predictive performance, with mean absolute errors of 10.25% and 8.10%, respectively. Explainability analysis identified hydrophilic matrix quantity as the dominant variable controlling early and intermediate drug release, whereas lipid type and lipid quantity were more influential for tablet hardness and late-stage release behavior. Among the original Taguchi formulations, F15 showed the highest observed desirability, although it did not fully satisfy all dissolution criteria. Virtual screening identified high-desirability in silico candidates characterized mainly by high hydrophilic matrix loading and adequate lipid content. The published optimized formulation was positioned within a high-desirability AI-predicted region, supporting consistency between the explainable ML workflow and the original QbD optimization. These findings indicate that small-data explainable machine learning can complement QbD-based formulation analysis by extracting formulation rules and prioritizing design-space regions; however, AI-predicted candidates require experimental confirmation.

Graphical Abstract



1. Introduction

Pregabalin is a centrally acting neuromodulating agent used in the management of neuropathic pain, fibromyalgia, and seizure-associated conditions [1,2]. Because of its relatively short biological half-life, immediate-release pregabalin formulations often require repeated administration, commonly two to three times daily, to maintain therapeutic exposure. Extended-release tablet systems can reduce dosing frequency, improve patient convenience, and support adherence. However, the development of an extended-release pregabalin tablet is technically challenging because pregabalin is highly water soluble [3]. Rapid penetration of aqueous medium into the dosage form can produce premature drug release unless the matrix system provides adequate diffusional resistance and erosion control [4].

The performance of extended-release matrix tablets is governed by the combined influence of formulation and process variables. Drug release may be affected by the type and quantity of hydrophilic polymer, the presence of lipidic or hydrophobic excipients, filler properties, tablet porosity, and compression force. Hydrophilic matrices form a hydrated gel layer that can regulate water ingress and drug diffusion, whereas lipidic components may reduce wettability and provide additional resistance to drug migration [3,4]. The tablet must also maintain acceptable mechanical integrity because a formulation with a suitable dissolution profile but inadequate hardness or high friability would be unsuitable for manufacturing, packaging, handling, or patient use [3,5]. Therefore, the development of pregabalin extended-release tablets requires simultaneous control of dissolution behavior and mechanical robustness.

Quality by Design (QbD) provides a systematic framework for pharmaceutical development by emphasizing predefined product objectives, risk assessment, product and process understanding, and control of critical quality attributes [4,6]. In extended-release tablets, dissolution is a major critical quality attribute because it reflects the intended release behavior of the dosage form [3,4]. Mechanical properties such as hardness and friability are also important because they determine the physical stability of the tablet [5]. A QbD-based strategy allows formulation scientists to define target performance criteria, identify critical material attributes and critical process parameters, and establish a design space in which product quality is expected to be maintained [6,7].

Design of experiments is commonly used within the QbD framework to evaluate the influence of formulation and process variables in a structured manner. Taguchi designs are useful when multiple factors must be screened using a limited number of experimental runs [3]. In pregabalin extended-release tablet development, a Taguchi-QbD data set can capture the effects of lipid type, lipid quantity, hydrophilic matrix type, hydrophilic matrix quantity, filler selection, and compression force on tablet performance [3]. Although such data sets are compact, they are information-rich and may support further computational interpretation when analyzed with appropriate methodological caution.

Ebeed et al. developed pregabalin extended-release tablets using a Taguchi-QbD approach and hot-melt granulation, and reported the effects of lipid material, hydrophilic matrix, filler type, and compression force on tablet quality and dissolution performance [3]. In the present study, the same published Taguchi-QbD formulation-response data were restructured into an artificial intelligence-ready data set and reanalyzed using explainable machine learning. The aim was not to redevelop the formulation experimentally, but to generate an interpretable computational layer for predicting selected critical quality attributes, identifying formulation rules, and mapping high-desirability digital design-space regions.

Conventional statistical analysis of design-of-experiments data can identify important factors and support formulation optimization. However, pharmaceutical formulation systems may show nonlinear behavior and factor interactions that are not fully described by simple main-effect interpretation. Machine learning can provide an additional analytical layer for extracting patterns from formulation data, predicting critical quality attributes, and exploring multidimensional design spaces [8,9]. In pharmaceutical development, machine learning should be viewed as complementary to QbD rather than as a replacement for experimental design or confirmatory testing.

Its value lies in response prediction, formulation ranking, and generation of interpretable decision rules that may support formulation understanding [9].

The application of machine learning to small pharmaceutical data sets requires careful methodological control. Data sets with limited experimental runs are not suitable for high-capacity black-box models or deep learning approaches. In such cases, the objective should be restricted to model-assisted interpretation and hypothesis generation. Leave-one-out cross-validation is useful for small data sets because each experimental run serves once as a validation point while the maximum amount of data remains available for model training [5,11]. Regularized regression, tree-based ensemble methods, and Gaussian process regression can provide practical options for analyzing tabular formulation data, provided that the results are interpreted cautiously [10,11].

Explainability is essential when machine learning is applied to formulation science [12,13]. A model that predicts hardness or dissolution without identifying the contributing formulation factors has limited practical value. Explainable machine learning can help determine which variables govern early, intermediate, and late drug release, as well as mechanical strength. Model-independent tools such as permutation feature importance and partial dependence analysis can translate model behavior into formulation-relevant information. These outputs can help determine whether performance is mainly governed by polymer quantity, lipid quantity, matrix grade, filler type, or compression force [5,12].

Digital design space mapping represents an extension of explainable machine learning within the QbD framework [9,12]. Once a model is evaluated within the experimental domain, it can be used to generate virtual formulations within the studied factor ranges and evaluate them against predefined quality criteria using desirability functions. This approach can identify candidate regions where acceptable hardness, low friability, and controlled dissolution profiles are predicted to coincide. These virtual candidates should be interpreted as in silico suggestions for future confirmation, not as experimentally validated optimized formulations [9,12].

Although Taguchi-QbD studies generate structured formulation data, such data are often used only for conventional statistical optimization. Their reuse for explainable machine-learning-based rule extraction and digital design space mapping remains underexplored, particularly for small data sets. The present study was therefore designed as a secondary explainable machine-learning analysis of published Taguchi-QbD data for pregabalin extended-release tablets. The objectives were to develop a cautious small-data modeling workflow, evaluate the predictability of selected critical quality attributes, identify the formulation variables governing dissolution and mechanical performance, and map high-desirability regions within the original experimental design space.

2. Materials and Methods

2.1. Data Source

This study was conducted as a secondary computational analysis of published Taguchi-QbD data for pregabalin extended-release tablets. No new laboratory experiments, animal studies, or clinical trials were performed. The source study reported the development of pregabalin extended-release tablets by hot-melt granulation using a Taguchi design of experiments and evaluated the influence of formulation and process variables on granule properties, tablet quality attributes, dissolution profiles, and optimized formulation performance.

The experimental formulation development was originally performed by Ebeed et al., who prepared pregabalin extended-release tablets using a Taguchi-QbD approach [3]. In the present study, the 16 Taguchi experimental formulations reported in that article were used as the primary modeling data set [3]. The post-optimization formulation reported in the same study was not included in the training data because it was generated after the initial Taguchi design. It was used only as a reference formulation for comparison with the high-desirability regions predicted by machine learning. The overall computational workflow used in the present study is illustrated in Figure 1.

2.2. Dataset Preparation

The formulation and response data were extracted from the published tables of the source study and organized into a structured data set. Each row represented one Taguchi formulation, from F1 to F16, while columns represented either an independent input variable or an experimentally reported response [3]. The summarized modeling data set is presented in Table 1.

Numerical values were recorded as the reported mean values. Standard deviations were not modeled separately because the objective of the present analysis was to predict central response trends rather than replicate-level variability. The data set was checked for transcription accuracy, unit consistency, and missing values before import into Python. The checked spreadsheet served as the master file for all subsequent computational analyses.

2.3. Input Variables and Responses

Six formulation and process variables were selected as model inputs based on the original experimental design. These variables included lipid material type, lipid quantity, hydrophilic matrix type, hydrophilic matrix quantity, filler type, and compression force. Lipid material type consisted of beeswax, cetyl alcohol, glyceryl monostearate, or stearic acid. Hydrophilic matrix type consisted of Methocel E50LV, K100LV, K15M, or K100M. Filler type consisted of lactose or Avicel. Lipid quantity and hydrophilic matrix quantity were recorded as mg/tablet, and compression force was recorded as kN.

The primary response variables selected for machine-learning modeling were tablet hardness, friability, and dissolution at 1, 2, 4, and 8 h. These responses were selected because they represent the principal mechanical and release-related critical quality attributes of the extended-release system. Hardness and friability describe tablet robustness, while the four dissolution time points characterize early, intermediate, and late drug release behavior. Additional variables, including bulk density, tapped density, Carr's index, angle of repose, thickness, and assay, were recorded in Table 1 for completeness but were not used as primary modeling targets in the final workflow [3].

2.4. Data Preprocessing

The data set was separated into independent input variables and dependent response variables before model development. Categorical input variables, including lipid material type, hydrophilic matrix type, and filler type, were transformed using one-hot encoding to enable their use in regression models [3]. Numerical input variables, including lipid quantity, hydrophilic matrix quantity, and compression force, were standardized within the modeling pipeline to reduce scale-related bias during model fitting [10].

All preprocessing steps were integrated into the cross-validation pipeline [11]. This approach prevented information leakage because encoding and scaling parameters were refitted within each validation cycle rather than estimated once from the entire data set [10,11]. The same preprocessing structure was used for final model training, explainability analysis, and virtual formulation prediction. Because the original Taguchi design contained only 16 formulations, no data augmentation, imputation, or random train-test split was performed. This ensured that the analysis remained restricted to the actual published experimental observations [3,11].

2.5. Machine Learning Models

Four regression models were evaluated to predict the selected critical quality attributes of the formulations: ridge regression, random forest regression, extra trees regression, and Gaussian process regression. Ridge regression was used as a regularized linear baseline model to assess mainly additive relationships between formulation variables and responses. Random forest regression and extra trees regression were included because they can capture nonlinear relationships and interaction effects between formulation and process variables without requiring a predefined functional form [8,9]. Gaussian process regression was included because it is suitable for small tabular data sets and can model smooth response behavior when interpreted with appropriate caution [9,11].

Separate response-specific models were developed for hardness, friability, and dissolution at 1, 2, 4, and 8 h. This approach was used because each response may be governed by different material attributes or process variables. Deep learning models were not used because the data set contained only 16 experimental formulations, which is

insufficient for stable neural network training. Conservative model settings were used to reduce the risk of overfitting in this small-data setting [10,11].

2.6. Model Validation

Model validation was performed using leave-one-out cross-validation. This strategy was selected because the data set contained only 16 formulations, and a conventional random train-test split would likely produce unstable and sample-dependent performance estimates. In each validation cycle, one formulation was held out as the validation point, while the model was trained using the remaining 15 formulations. This procedure was repeated until every experimental run in the Taguchi design had served once as the validation sample [5,11].

For each response, the cross-validated predictions were compared with the observed experimental values reported in the source study. The algorithm with the lowest mean absolute error was selected as the final response-specific model [10]. After model selection, the selected model for each response was retrained using the complete 16-formulation data set to support explainability analysis and virtual design space screening [9,12]. This cross-validation and modeling workflow is summarized in Figure 1.

2.7. Model Performance Evaluation

The predictive performance of the selected models was evaluated using mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). MAE was used as the primary criterion for model selection because it directly represents the average prediction error in the original unit of the response variable. RMSE was used to identify larger deviations between observed and predicted values. R^2 was reported as an additional measure of agreement between observed and predicted responses [9,11].

Because the data set was small ($n = 16$), R^2 was interpreted cautiously. Low or negative R^2 values can occur when the response range is narrow or when individual observations strongly influence the cross-validation result. Therefore, model selection and interpretation were based mainly on MAE, RMSE, and observed-versus-predicted plots [10,11]. The final performance metrics for the response-specific models are summarized in Table 2, and the corresponding observed-versus-predicted plots are presented in Figure 2.

2.8. Explainability Analysis

Model-independent explainability analysis was performed to identify the formulation and process variables contributing to each predicted response. Permutation feature importance was used as the primary global explainability method [9,12]. In this analysis, the values of one input variable were randomly permuted while the remaining variables were kept unchanged, and the resulting change in model error was calculated. A larger increase in error indicated a greater contribution of that variable to the model prediction. Permutation feature importance was applied to all six input variables: lipid material type, lipid quantity, hydrophilic matrix type, hydrophilic matrix quantity, filler type, and compression force. The feature importance results are summarized in Table 3 and Figure 3.

Partial dependence analysis was used to examine the average influence of numerical variables on dissolution responses. Lipid quantity, hydrophilic matrix quantity, and compression force were varied across their studied ranges, and the corresponding mean predicted dissolution values were calculated. This analysis was used to understand how each numerical variable affected early, intermediate, and late drug release behavior. The partial dependence plots are presented in Figure 4.

A surrogate decision tree was also generated for the dissolution responses to convert selected machine-learning model behavior into simplified formulation rules. The surrogate tree was used only as an interpretability tool for rule extraction and not as the primary prediction model.

2.9. Desirability Analysis

A desirability-based scoring method was used to rank both the observed experimental formulations and the AI-predicted virtual candidates [7,12]. Individual desirability values were calculated for hardness, friability, and

dissolution at 1, 2, 4, and 8 h. Hardness was treated as a higher-is-better response with a minimum acceptable threshold of 30 N. Friability was treated as a lower-is-better response with an upper acceptable limit of 1%.

Dissolution desirability was defined according to the target extended-release profile for pregabalin. Release at 1 h was targeted below 20%, release at 2 h between 20 and 40%, release at 4 h between 40 and 60%, and release at 8 h above 80%. Individual desirability values were combined using the geometric mean to generate an overall desirability score [6,7]. This approach ensured that a formulation with poor performance in any one critical response received a lower overall score even if other responses were acceptable. The ranking of the observed formulations is summarized in Table 4.

2.10. Digital Design Space Mapping

A virtual formulation space was generated within the experimental factor ranges of the original Taguchi design. Lipid quantity was varied from 20 to 80 mg, and hydrophilic matrix quantity was varied from 0 to 100 mg. Compression force was evaluated at the studied levels. Categorical variables were explored using the same lipid types, hydrophilic matrix grades, and filler types present in the published data set.

The selected response-specific models were used to predict hardness, friability, and multi-time-point dissolution responses for each virtual candidate [9,12]. The predicted responses were then converted into desirability scores using the predefined quality criteria. Virtual formulations were ranked according to their overall desirability scores to identify high-desirability regions within the formulation space. The digital design space was restricted to the original experimental factor ranges to avoid extrapolation outside the studied domain [10,11]. The highest-ranking in silico candidates are presented in Table 5.

2.11. Comparison with Published Optimized Formulation

The optimized formulation identified in the original QbD study was used as a reference point to evaluate the consistency of the machine-learning workflow. This reference composition consisted of stearic acid, Methocel K15M, and lactose, compressed at 10 kN, and was developed by the original authors after the initial Taguchi screening [3]. Because this formulation was generated as a post-optimization formulation rather than as part of the original 16-run Taguchi design, it was not included in model training, cross-validation, or response-specific model selection.

After virtual design space mapping, the published optimized composition was located within the AI-generated formulation space, and its predicted responses and overall desirability score were calculated. This comparison was used to determine whether the experimentally optimized formulation from the original QbD study was positioned within a high-desirability region predicted by the explainable machine-learning workflow [9,12]. The comparison was interpreted as a consistency check, not as independent experimental validation. The detailed comparison for this reference formulation is summarized in Table 6.

2.12. Software and Reproducibility

All computational analyses were performed using Python in a Jupyter notebook environment. Data handling and numerical operations were performed using pandas and NumPy. Preprocessing, regression modeling, leave-one-out cross-validation, performance evaluation, permutation feature importance, and surrogate decision-tree analysis were implemented using scikit-learn [8,12,15]. Figures were generated using matplotlib, and result tables were exported using openpyxl.

The computational workflow was organized to allow reproducible execution from data import to final result export. The notebook imported the checked data set, performed preprocessing and model validation, generated explainability outputs, constructed the virtual formulation space, ranked formulation candidates, and exported the resulting tables and figures. The checked data sheet, Python code, and generated outputs are intended to be provided as supplementary material to support peer verification and reuse [14–16].

3. Results

3.1. Data Set

The final data set consisted of 16 formulations of pregabalin extended-release tablets derived from the published Taguchi-QbD study [3]. Each experimental run was defined by six formulation and process input variables: lipid material type, lipid quantity, hydrophilic matrix type, hydrophilic matrix quantity, filler type, and compression force. The responses selected for machine-learning modeling were tablet hardness, friability, and dissolution at 1, 2, 4, and 8 h. The complete formulation-response matrix used for this computational analysis is presented in Table 1.

The extracted data showed substantial response variation across the 16 formulations. Tablet hardness ranged from 19 to 95 N, while friability ranged from 0.16% to 4.02%. Marked differences were also observed in dissolution behavior. Drug release at 1 h ranged from 25.98% to 93.06%, release at 2 h from 39.49% to 97.34%, release at 4 h from 61.98% to 99.65%, and release at 8 h from 89.09% to 100.80% (Table 1). These response ranges indicate that the original Taguchi design generated meaningful variation in both mechanical properties and release behavior.

The release pattern showed that most formulations achieved high final release by 8 h, whereas larger differences were observed at 1, 2, and 4 h. Therefore, the main formulation challenge in the data set was control of early and intermediate release rather than achieving complete final drug release. This observation supported the use of separate dissolution responses at 1, 2, 4, and 8 h as individual modeling targets.

3.2. Machine Learning Model Performance

Ridge regression, random forest regression, extra trees regression, and Gaussian process regression were evaluated using leave-one-out cross-validation. For each response, the model with the lowest mean absolute error was selected as the response-specific final model. The complete model performance summary is provided in Table 2, and the observed-versus-predicted plots for the selected models are shown in Figure 2.

For hardness, ridge regression provided the best performance, with a mean absolute error of 11.28 N, root mean square error of 14.04 N, and coefficient of determination of 0.430 (Table 2). This indicated moderate agreement between observed and predicted hardness values within the limited data set. For friability, Gaussian process regression gave the lowest mean absolute error of 0.551% and root mean square error of 0.958%. However, the coefficient of determination was negative, indicating that friability prediction should be interpreted cautiously because most formulations had low friability values and only a few showed markedly higher values.

For dissolution responses, model performance varied by time point. Gaussian process regression was selected for dissolution at 1 h, with a mean absolute error of 14.83%, root mean square error of 18.26%, and coefficient of determination of 0.408. Ridge regression was selected for dissolution at 2 h and 4 h, with mean absolute errors of 10.25% and 8.10%, respectively. The 2 h dissolution model showed the strongest cross-validated performance, with a coefficient of determination of 0.626. For dissolution at 8 h, Gaussian process regression produced a low mean absolute error of 3.30%, but the coefficient of determination was negative because the 8 h dissolution values were clustered within a narrow range (Table 2).

Overall, the models were most informative for describing formulation-dependent trends in the 2 h and 4 h dissolution responses. The observed-versus-predicted plots supported this interpretation, showing better agreement for intermediate dissolution responses than for friability and 8 h dissolution (Figure 2). These findings indicate that the workflow is suitable for small-data exploratory modeling and response ranking, but not for high-precision prediction without additional experimental data.

3.3. Formulation Ranking by Desirability

The 16 observed Taguchi formulations were ranked using an overall desirability score calculated from hardness, friability, and dissolution criteria. The full ranking is shown in Table 4. This analysis was useful because none of the original Taguchi formulations fulfilled all dissolution targets simultaneously.

Formulation F15 showed the highest observed desirability score of 0.979. This formulation contained stearic acid 40 mg, Methocel K15M 100 mg, lactose, and compression force of 10 kN. It showed hardness of 63 N and friability of 0.42%, indicating acceptable tablet robustness. The dissolution values of F15 were 25.98%, 39.49%,

61.98%, and 89.09% at 1, 2, 4, and 8 h, respectively (Table 4). Although F15 did not fully satisfy the 1 h and 4 h dissolution criteria, it was the closest formulation to the desired overall response profile among the original 16 formulations.

Formulation F9 ranked second, with an overall desirability score of 0.944. It contained glyceryl monostearate 80 mg, Methocel K100M 100 mg, lactose, and compression force of 5 kN. Formulation F6 ranked third, with a desirability score of 0.902, and contained cetyl alcohol 60 mg, Methocel K100LV 100 mg, Avicel, and compression force of 10 kN (Table 4).

The top-ranked formulations generally contained high hydrophilic matrix quantities. In contrast, formulations with no hydrophilic matrix or insufficient matrix quantity commonly showed rapid early release and received lower desirability scores. These results indicate that adequate matrix loading was the main requirement for approaching the desired extended-release profile in the observed Taguchi data set.

3.4. Feature Importance

Permutation feature importance was used to identify the relative contribution of each formulation and process variable to the selected response-specific models. The complete feature-importance results are presented in Table 3, and the corresponding plots are shown in Figure 3.

For hardness, lipid material type was the most influential variable, followed by lipid quantity and filler type (Table 3; Figure 3). The importance values for these variables were higher than those for hydrophilic matrix type, hydrophilic matrix quantity, and compression force. This indicates that tablet hardness was mainly influenced by the lipid binder system and filler-related compaction behavior within the studied formulation space.

For friability, the importance pattern was more distributed. Lipid material type, hydrophilic matrix type, lipid quantity, and hydrophilic matrix quantity showed comparable contributions (Table 3; Figure 3). This suggests that friability was affected by the combined influence of lipid binding, matrix grade, and matrix loading rather than by a single formulation variable.

For dissolution at 1 h, hydrophilic matrix quantity was the dominant variable, followed by lipid quantity, hydrophilic matrix type, lipid material type, filler type, and compression force (Table 3; Figure 3). A similar pattern was observed for dissolution at 2 h and 4 h, where hydrophilic matrix quantity remained the highest-ranked variable and lipid quantity was the second most important contributor. These results show that early and intermediate drug release were mainly controlled by hydrophilic matrix loading and, to a lesser extent, by lipid quantity.

For dissolution at 8 h, the importance values were more evenly distributed. Lipid material type, hydrophilic matrix type, lipid quantity, and hydrophilic matrix quantity all contributed to the predicted response (Table 3; Figure 3). This indicates that late-stage release was governed by the combined effect of lipid characteristics, matrix grade, and matrix loading. Overall, the feature-importance analysis confirmed that hydrophilic matrix quantity was the main factor for early and intermediate release control, whereas lipid-related variables were more relevant to hardness and late-stage release behavior.

3.5. Effect of Formulation Variables on Dissolution

Partial dependence analysis was performed to examine the average effect of lipid quantity, hydrophilic matrix quantity, and compression force on dissolution responses. The plots are shown in Figure 4. These results were interpreted only within the factor ranges present in the original Taguchi design.

For dissolution at 1 h, increasing lipid quantity reduced the mean predicted release from 61.31% at 20 mg to 45.60% at 80 mg. The effect of hydrophilic matrix quantity was stronger, with the mean predicted 1 h release decreasing from 76.55% at 0 mg matrix to 27.30% at 100 mg matrix. Compression force had only a minor effect, with the predicted 1 h release changing from 53.43% at 5 kN to 52.54% at 10 kN (Figure 4).

A similar trend was observed for dissolution at 2 h. Increasing lipid quantity reduced the mean predicted release from 77.01% to 60.12%, whereas increasing hydrophilic matrix quantity reduced the predicted release from

87.41% to 44.34%. Compression force showed a smaller effect, reducing the predicted 2 h release from 70.58% at 5 kN to 66.55% at 10 kN (Figure 4).

At 4 h, hydrophilic matrix quantity continued to show the strongest influence. The mean predicted 4 h release decreased from 95.48% at 0 mg matrix to 69.76% at 100 mg matrix. Lipid quantity reduced the predicted 4 h release from 89.32% to 79.13%. Compression force produced a comparatively smaller reduction, from 86.43% at 5 kN to 82.02% at 10 kN (Figure 4).

At 8 h, the predicted dissolution remained high across the studied ranges. Lipid quantity, hydrophilic matrix quantity, and compression force showed only limited effects on final release (Figure 4). These findings indicate that matrix composition was more important than compression force for controlling early and intermediate pregabalin release, while most formulations were predicted to achieve adequate final release by 8 h.

3.6. Virtual Formulation Screening

The selected final models were used to screen 17,472 virtual formulations generated within the original experimental factor ranges. The top-ranked virtual candidates are presented in Table 5. This analysis was used to identify high-desirability in silico regions rather than to define an experimentally confirmed optimized formulation.

Fourteen virtual formulations achieved an AI-predicted overall desirability score of 1.000. Seventy-nine formulations showed desirability scores of at least 0.990, and 915 formulations showed desirability scores of at least 0.950. These results indicate that the predicted high-desirability region was not limited to a single formulation point but was concentrated within a defined region of the studied design space.

The highest-ranked virtual candidate contained beeswax 80 mg, Methocel K15M 100 mg, Avicel, and compression force of 10 kN. The predicted responses for this candidate were hardness of 55.42 N, friability of 0.672%, and dissolution values of 15.42%, 27.19%, 58.22%, and 97.24% at 1, 2, 4, and 8 h, respectively (Table 5). These values satisfied the predefined desirability criteria for both tablet robustness and dissolution behavior.

The top virtual candidates generally shared two characteristics: high hydrophilic matrix loading and adequate lipid content. Most high-desirability candidates contained matrix quantities of 95–100 mg and lipid quantities of 70–80 mg. Compression at 10 kN was common among the highest-ranked candidates (Table 5). These results are consistent with the feature-importance and partial dependence findings, which identified hydrophilic matrix quantity and lipid quantity as key determinants of controlled release within the studied design space.

3.7. Comparison with the Published Optimized Formulation

The optimized formulation reported in the source study was evaluated as a reference point within the AI-generated virtual design space. This formulation contained stearic acid 80 mg, Methocel K15M 100 mg, lactose, and compression force of 10 kN [3]. It was not included in model training or leave-one-out cross-validation because it was generated after the original 16-run Taguchi design.

When this composition was evaluated using the selected response-specific models, the predicted hardness was 71.33 N and the predicted friability was 0.672%. The predicted dissolution values were 21.91%, 30.73%, 61.81%, and 96.99% at 1, 2, 4, and 8 h, respectively. The corresponding AI-predicted overall desirability was 0.988 (Table 6).

These results positioned the published optimized formulation within a high-desirability region of the AI-generated design space, although it was not the highest-ranked virtual candidate. The composition of this formulation was consistent with the main formulation pattern identified by the explainable machine-learning workflow, namely the combination of high hydrophilic matrix quantity with sufficient lipid content.

This comparison supports the consistency between the machine-learning analysis and the original QbD optimization. However, the AI prediction should be interpreted as a computational consistency check rather than as independent experimental validation. The AI-predicted virtual candidates and high-desirability regions require confirmatory experimental evaluation before being considered optimized formulations.

4. Discussion

The present study applied an explainable machine-learning workflow to published Taguchi-QbD data for pregabalin extended-release tablets [3]. The objective was not to redevelop the formulation experimentally, but to determine whether a small, structured formulation data set could be used to extract interpretable formulation rules and support digital design space mapping [6,22]. This distinction is important because the data set contained only 16 Taguchi runs. Therefore, the results should be interpreted as model-assisted formulation understanding and in silico prioritization rather than as definitive experimental optimization [9,12].

The value of the present work lies in adding an explainable computational layer to an already published QbD formulation study. The original study established the experimental feasibility of pregabalin extended-release tablets and selected an optimized formulation through conventional QbD analysis and confirmatory testing [3]. The present analysis does not replace that work. Instead, it demonstrates how the same Taguchi-QbD data can be reused to generate response-specific predictive models, rank formulation variables, visualize formulation trends, and prioritize virtual design-space regions for future confirmatory experiments [6,9,12,22].

The extracted data set provided a useful basis for exploratory machine learning because it contained clear variation in both mechanical and release-related responses. Hardness varied from 19 to 95 N, and friability ranged from acceptable values to clear failure values [3]. Dissolution also showed wide variation at 1, 2, and 4 h, while most formulations achieved high release by 8 h [3]. This pattern indicates that the main formulation challenge was control of early and intermediate release rather than final drug release [3,4]. Such behavior is expected for pregabalin because it is highly water soluble and can diffuse rapidly once the dissolution medium penetrates the matrix. Therefore, a successful extended-release formulation requires a matrix structure capable of delaying water ingress, gel erosion, and drug diffusion during the initial and intermediate stages [4,9].

The leave-one-out cross-validation results showed that the models were most useful for dissolution at 2 h and 4 h, whereas friability and 8 h dissolution required cautious interpretation [17–19]. The 2 h and 4 h dissolution responses showed the most meaningful predictive performance because these time points captured the formulation-sensitive phase of drug release. In contrast, 8 h dissolution values were clustered within a narrow range; therefore, the model could show low absolute prediction error but weak coefficient of determination [3,23]. This highlights a key issue in small-data formulation modeling: performance should not be judged by R^2 alone [18,19]. Mean absolute error and root mean square error are more informative because they express the prediction error in the original response units. The modest performance values also reinforce that the workflow should be considered exploratory rather than fully predictive [10,18].

The desirability ranking of the observed formulations provided a practical interpretation of the Taguchi design [3,7]. F15 showed the highest overall desirability among the original 16 formulations, with acceptable hardness and friability and the closest overall dissolution behavior to the desired extended-release profile. However, F15 did not completely satisfy all dissolution criteria, especially at 1 h and 4 h. This result is important because it shows that none of the original Taguchi runs represented a perfect formulation. Instead, the observed data provided response trends that could be used for further digital exploration. The high ranking of F15 also emphasized the importance of high hydrophilic matrix loading, particularly with Methocel K15M, in suppressing rapid drug release [3,12].

The explainability analysis produced formulation-relevant findings that were consistent with matrix tablet science. Hydrophilic matrix quantity was the dominant factor controlling dissolution at 1, 2, and 4 h [3,9]. This is mechanistically reasonable because a sufficient quantity of hydrophilic polymer is required to form a continuous gel layer after hydration. This gel layer acts as a barrier to water penetration and drug diffusion and can also regulate matrix erosion. The partial dependence plots further supported this interpretation, showing a marked reduction in predicted early and intermediate release as hydrophilic matrix quantity increased. The effect was strongest at 1 and 2 h, confirming that polymer loading is critical for reducing burst-like release from the matrix [3,4].

Lipid quantity also contributed to dissolution control, although its effect was weaker than that of hydrophilic matrix quantity. Increasing lipid quantity reduced predicted release at 1, 2, and 4 h, suggesting that lipid material provided additional diffusional resistance [3,9]. However, the explainability results indicate that lipid material alone was not the primary determinant of extended-release behavior. This agrees with the expected behavior of highly water-soluble drugs, where a hydrophobic phase by itself may not be sufficient to maintain controlled release unless supported by an effective gel-forming matrix. Therefore, the release mechanism appears to depend on a hydrophilic-lipophilic balance: the hydrophilic polymer forms the main hydrated barrier, while the lipid component modifies water access and drug migration through the matrix [3,4].

For hardness, lipid type was the most important variable, followed by lipid quantity and filler type. This suggests that the mechanical strength of the tablets was strongly influenced by the lipid binder system and the compactibility of the formulation blend [3,5]. Different lipid materials can alter granule density, deformation during compression, and interparticulate bonding. The contribution of filler type is also reasonable because lactose and Avicel differ in their compaction behavior. Compression force had a smaller effect within the studied range, indicating that formulation composition was more influential than the difference between 5 and 10 kN in this data set [3].

The virtual formulation screening extended the observed Taguchi findings by identifying predicted high-desirability regions within the original design limits [6,12,22]. Several virtual candidates achieved high desirability scores, and the top-ranked candidates generally contained high hydrophilic matrix quantities with sufficient lipid content. This finding is useful from a QbD perspective because it suggests that the acceptable design space may not be limited to a single formulation point [6,22]. Instead, a broader region defined by high matrix loading and adequate lipid content may support the desired extended-release behavior. However, these candidates remain in silico predictions and require experimental confirmation before they can be considered optimized formulations [10,22].

The comparison with the published optimized formulation further supported the consistency of the machine-learning workflow. The published optimized composition, containing stearic acid 80 mg, Methocel K15M 100 mg, lactose, and compression force of 10 kN, was predicted to fall within a high-desirability region, with an overall AI-predicted desirability of 0.988. The model did not rank this composition as the single highest virtual candidate, but this is not a weakness. The important result is that the experimentally optimized formulation was located within a favorable AI-predicted region [3,6]. This indicates that the explainable ML workflow recovered the same general formulation logic as the original QbD optimization [9,12].

The major limitation of this study is the small sample size [18]. Sixteen formulations are adequate for a cautious secondary analysis but not for building a broadly generalizable predictive model [17,18]. The analysis was also based on mean response values extracted from a published article, so replicate-level variability and experimental uncertainty were not modeled. In addition, the virtual design space was restricted to the original experimental ranges and should not be extrapolated beyond them. Future work should include experimental preparation of selected AI-predicted candidates, replicate-level modeling, and uncertainty estimation [10]. Integration of dissolution modeling with pharmacokinetic or biopredictive approaches may also strengthen the translational relevance of the workflow [9].

Overall, the study shows that explainable machine learning can complement QbD-based formulation analysis when applied with appropriate caution [6,12]. The models recovered key formulation principles, identified hydrophilic matrix quantity as the main driver of early and intermediate release control, highlighted lipid-related variables as important for hardness and late-stage release behavior, and positioned the published optimized formulation within a high-desirability digital design region [3,9]. These findings support the value of small-data explainable ML as a decision-support tool for formulation interpretation and design space prioritization [12,22].

5. Conclusion

This study demonstrated the utility of explainable machine learning as a complementary computational tool for reanalyzing published Taguchi-QbD data of pregabalin extended-release tablets. Using the same formulation-

response data generated in the original experimental study, the present work developed a small-data modeling workflow to predict selected critical quality attributes, interpret formulation-variable effects, and map high-desirability regions within the studied design space. The analysis was intentionally framed as a secondary computational investigation and not as a replacement for experimental formulation development.

Among the modeled responses, dissolution at 2 h and 4 h showed the most useful cross-validated predictive performance, whereas friability and 8 h dissolution required cautious interpretation because of their response distribution. The explainability analysis identified hydrophilic matrix quantity as the main factor controlling early and intermediate drug release, while lipid type and lipid quantity were more influential for tablet hardness and late-stage release behavior. These findings were consistent with the expected behavior of hydrophilic-lipophilic matrix tablets and supported the formulation logic reported in the original QbD study.

Desirability-based ranking showed that F15 was the best-performing formulation among the original 16 Taguchi runs, although it did not fully satisfy all dissolution criteria. Virtual formulation screening identified high-desirability in silico regions characterized mainly by high hydrophilic matrix loading and adequate lipid content. The published optimized formulation was also positioned within a high-desirability AI-predicted region, indicating consistency between the explainable machine-learning workflow and the original experimental optimization.

Overall, the results support the use of small-data explainable machine learning for formulation-rule extraction, digital design space exploration, and prioritization of candidate formulations for future testing. However, the AI-predicted virtual candidates should be considered hypothesis-generating outputs and require experimental confirmation before being regarded as optimized formulations. Future work should include confirmatory batch preparation, replicate-level modeling, uncertainty estimation, and integration of dissolution behavior with biopredictive or pharmacokinetic approaches.

References

1. Derry S, Bell RF, Straube S, Wiffen PJ, Aldington D, Moore RA. Pregabalin for neuropathic pain in adults. Cochrane Database Syst Rev. 2019;2019(1):CD007076. doi:10.1002/14651858.CD007076.pub3.
2. Toth C. Pregabalin: latest safety evidence and clinical implications for the management of neuropathic pain. Ther Adv Drug Saf. 2014;5(1):38–56. doi:10.1177/2042098613505614.
3. Ebeed M, Balata GF, Elghamry HA, Eissa NG. Quality by Design for cost effective production of pregabalin extended-release tablets: in vitro and in vivo study. J Pharm Innov. 2025;20(5). doi:10.1007/s12247-025-10058-9.
4. Sousa AS, Serra JM, Estevens C, Costa R, Ribeiro AJ. Unveiling swelling and erosion dynamics: early development screening of mirabegron extended release tablets. AAPS PharmSciTech. 2024;25(8). doi:10.1208/s12249-024-02994-5.
5. Katayama H, Maeda Y, Sato T, Mogi A, Itakura S, Todo H, et al. Design and optimization of scored tablets with concave surface and application of Bayesian estimation for solving scale-up problem. Chem Pharm Bull (Tokyo). 2021;69(11):1088–1096. doi:10.1248/cpb.c21-00538.
6. Yang S, Hu X, Zhu J, Zheng B, Bi W, Wang X, et al. Aspects and implementation of pharmaceutical Quality by Design from conceptual frameworks to industrial applications. Pharmaceutics. 2025;17(5):623. doi:10.3390/pharmaceutics17050623.
7. Ilyés K. Applying the principles of Quality by Design (QbD) coupled with multivariate data analysis (MVDA) in establishing the impact of raw material variability for extended release tablets. Farmacia. 2021;69(3):481–497. doi:10.31925/farmacia.2021.3.11.
8. Jiang J, Ma X, Ouyang D, Williams RO 3rd. Emerging artificial intelligence technologies used in the development of solid dosage forms. Pharmaceutics. 2022;14(11):2257. doi:10.3390/pharmaceutics14112257.
9. Ros H, Chan N, Cook MT, Shorthouse D. Artificial intelligence and machine learning guided optimization in drug delivery. Adv Drug Deliv Rev. 2026;115781. doi:10.1016/j.addr.2026.115781.

10. Sekar M, Rajendran M, Sekar AP. Selection of injectable drug product composition using machine learning models [Preprint]. 2026. doi:10.2196/preprints.91552.
11. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. J Chem Inf Comput Sci. 2003;43(2):579–586. doi:10.1021/ci025626i.
12. Jadhav SSP, Rathod VK, Nagane AR, Mujawar SS, Ishwara R. Artificial intelligence and machine learning in pharmaceutical research: transforming drug discovery, formulation development, and quality assurance [Internet]. Zenodo; 2026 [cited 2026 Mar]. doi:10.5281/zenodo.18838484.
13. Kottai Muthu A, Alagu Manivasagam. Artificial intelligence and machine learning in drug formulation development [Internet]. Zenodo; 2025 [cited 2025 Nov]. doi:10.5281/zenodo.17589193.
14. Schaduangrat N, Lampa S, Simeon S, Gleeson MP, Spjuth O, Nantasenamat C. Towards reproducible computational drug discovery. J Cheminform. 2020;12(1). doi:10.1186/s13321-020-0408-x.
15. Seal S, Yang H, Trapotsi M, Singh S, Carreras-Puigvert J, Spjuth O, et al. Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. J Cheminform. 2023;15(1). doi:10.1186/s13321-023-00723-x.
16. Hajal AA, Bustanji L, Bryce RA, Ghattas MA. An introduction to GitHub and its significance for AI-driven drug discovery. Expert Opin Drug Discov. 2026:1–6. doi:10.1080/17460441.2026.2641511.
17. Austin GI, Pe'er I, Korem T. Distributional bias compromises leave-one-out cross-validation [Preprint]. arXiv; 2024. Available from: http://arxiv.org/abs/2406.01652.
18. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning [Preprint]. arXiv; 2018. Available from: http://arxiv.org/abs/1811.12808.
19. Geroldinger A, Lusa L, Nold M, Heinze G. Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures: a simulation study. Diagn Progn Res. 2023;7(1). doi:10.1186/s41512-023-00146-0.
20. Seeger M. Gaussian processes for machine learning. Int J Neural Syst. 2004;14(2):69–106. doi:10.1142/S0129065704001899.
21. Chanana A, Singh M, Agarwal M, Singh RP, Chawra HS, Mishra A. PharmaQAI: a machine learning framework and interactive platform for process intelligence and quality prediction in pharmaceutical tablet compression [Preprint]. Preprints.org; 2026. doi:10.20944/preprints202603.0225.v1.
22. Zhu Z. Intelligent information management enables quality-by-design in pharmaceutical production. Sci Rep. 2025;15(1):44201. doi:10.1038/s41598-025-27879-w.
23. Cichosz P. Assessing the quality of classification models: performance measures and evaluation procedures. Open Eng. 2011;1(2):132–158. doi:10.2478/s13531-011-0022-9.

Table 1. Extracted Taguchi-QbD formulation-response data set used for machine-learning analysis

Formulation	Lipid type	Lipid (mg)	Matrix type	Matrix (mg)	Filler	Force (kN)	BD	TD	CI (%)	AR (°)	Hardness (N)	Friability (%)	Thickness (mm)	Assay (%)	D1h (%)	D2h (%)	D4h (%)	D8h (%)
F1	Beeswax	80	K100LV	0	Lactose	10	0.53	0.64	18.18	19.98	44	0.16	3.31	102.32	85.33	90.12	93.34	97.45
F2	Beeswax	60	K15M	25	Lactose	5	0.57	0.76	25.86	21.58	33	0.19	3.42	101.48	51.91	69.67	92.09	96.65
F3	Beeswax	40	K100M	50	Avicel	10	0.41	0.62	34.21	22.93	40	0.23	3.56	94.59	32.87	54.73	71.82	90.09
F4	Beeswax	20	E50LV	100	Avicel	5	0.55	0.74	26.67	29.06	19	1.47	3.67	103.12	28.04	51.42	74.90	98.21
F5	Cetyl alcohol	80	K15M	50	Avicel	5	0.37	0.60	38.04	29.17	63	0.27	3.78	105.27	32.12	50.75	74.58	100.26
F6	Cetyl alcohol	60	K100LV	100	Avicel	10	0.36	0.59	38.89	32.28	68	0.17	3.89	105.84	31.39	47.95	71.01	100.80
F7	Cetyl alcohol	40	E50LV	0	Lactose	5	0.63	0.93	32.73	24.20	41	4.02	3.98	98.15	91.27	91.35	92.30	93.48
F8	Cetyl alcohol	20	K100M	25	Lactose	10	0.56	0.78	28.57	26.82	34	0.55	3.87	99.69	76.52	89.22	94.67	94.94
F9	GMS	80	K100M	100	Lactose	5	0.47	0.68	30.77	23.09	47	0.44	3.76	99.12	28.78	44.29	65.78	94.40
F10	GMS	60	E50LV	50	Lactose	10	0.51	0.74	30.65	23.96	43	0.54	3.65	105.32	39.01	58.16	86.02	100.30
F11	GMS	40	K100LV	25	Avicel	5	0.40	0.63	35.56	25.46	45	0.48	3.54	104.62	64.23	84.60	95.88	99.64
F12	GMS	20	K15M	0	Avicel	10	0.36	0.54	32.22	11.57	51	0.39	3.43	105.84	93.06	97.34	98.50	100.80
F13	Stearic acid	80	E50LV	25	Avicel	10	0.37	0.49	25.00	23.77	81	0.35	3.32	104.09	35.82	54.41	77.73	99.14
F14	Stearic acid	60	K100M	0	Avicel	5	0.37	0.52	29.41	26.57	95	0.38	3.21	105.44	74.25	92.76	97.38	100.42
F15	Stearic acid	40	K15M	100	Lactose	10	0.40	0.61	34.21	27.51	63	0.42	3.12	93.54	25.98	39.49	61.98	89.09
F16	Stearic acid	20	K100LV	50	Lactose	5	0.47	0.68	31.25	16.70	41	0.69	3.23	104.92	57.81	80.81	99.65	99.92

BD, bulk density; TD, tapped density; CI, Carr's index; AR, angle of repose; D1h, D2h, D4h, and D8h, percent drug release at 1, 2, 4, and 8 h, respectively; GMS, glyceryl monostearate. Suggested layout: landscape.

Table 2. Leave-one-out cross-validation performance of candidate machine-learning models

Response	Model	MAE	RMSE	R^2
Hardness	Ridge	11.276	14.043	0.430
Hardness	Extra trees	12.666	15.473	0.308
Hardness	Random forest	14.812	17.997	0.063
Hardness	Gaussian process	15.291	18.045	0.058
Friability	Gaussian process	0.551	0.958	-0.096
Friability	Extra trees	0.576	1.095	-0.432
Friability	Random forest	0.691	1.054	-0.326
Friability	Ridge	0.816	1.099	-0.443
Dissolution 1 h	Gaussian process	14.829	18.256	0.408
Dissolution 1 h	Extra trees	15.191	17.909	0.430
Dissolution 1 h	Random forest	15.247	17.311	0.468
Dissolution 1 h	Ridge	18.817	20.899	0.224
Dissolution 2 h	Ridge	10.249	12.048	0.626
Dissolution 2 h	Gaussian process	10.589	12.781	0.579
Dissolution 2 h	Random forest	14.512	16.028	0.337
Dissolution 2 h	Extra trees	14.623	17.056	0.250
Dissolution 4 h	Ridge	8.104	9.243	0.444
Dissolution 4 h	Gaussian process	9.042	10.121	0.333
Dissolution 4 h	Random forest	9.482	11.394	0.155
Dissolution 4 h	Extra trees	10.069	11.901	0.078
Dissolution 8 h	Gaussian process	3.299	3.933	-0.141
Dissolution 8 h	Random forest	3.857	4.690	-0.623
Dissolution 8 h	Ridge	5.136	5.977	-1.636
Dissolution 8 h	Extra trees	5.413	6.277	-1.907

MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination. The model with the lowest MAE for each response was selected as the response-specific final model.

Table 3. Permutation feature importance for response-specific final models

Response	Feature	Mean importance	SD
Hardness	Lipid type	9.396	1.913
Hardness	Lipid quantity	6.902	1.796
Hardness	Filler type	4.569	1.129
Hardness	Compression force	0.883	0.574
Hardness	Matrix type	0.739	0.516
Hardness	Matrix quantity	0.577	0.517
Friability	Lipid type	0.188	0.044
Friability	Matrix type	0.186	0.047
Friability	Lipid quantity	0.184	0.044
Friability	Matrix quantity	0.175	0.040
Friability	Filler type	0.125	0.052
Friability	Compression force	0.120	0.060
Dissolution 1 h	Matrix quantity	20.426	3.645
Dissolution 1 h	Lipid quantity	6.788	1.585
Dissolution 1 h	Matrix type	4.475	1.036
Dissolution 1 h	Lipid type	4.315	0.829
Dissolution 1 h	Filler type	3.532	0.844
Dissolution 1 h	Compression force	2.506	0.799
Dissolution 2 h	Matrix quantity	15.338	3.006
Dissolution 2 h	Lipid quantity	4.496	1.200
Dissolution 2 h	Matrix type	2.547	0.925
Dissolution 2 h	Compression force	0.483	0.412
Dissolution 2 h	Lipid type	0.382	0.398
Dissolution 2 h	Filler type	0.238	0.346
Dissolution 4 h	Matrix quantity	8.065	1.864
Dissolution 4 h	Lipid quantity	2.274	0.874
Dissolution 4 h	Matrix type	1.433	0.562
Dissolution 4 h	Compression force	0.782	0.454
Dissolution 4 h	Filler type	0.279	0.280
Dissolution 4 h	Lipid type	0.082	0.259
Dissolution 8 h	Lipid type	2.046	0.358
Dissolution 8 h	Matrix type	2.010	0.397
Dissolution 8 h	Lipid quantity	1.834	0.323
Dissolution 8 h	Matrix quantity	1.824	0.364
Dissolution 8 h	Compression force	1.518	0.514
Dissolution 8 h	Filler type	1.396	0.391

Mean importance represents the mean change in model error after permutation of the corresponding input variable. SD represents the standard deviation across repeated permutations.

Table 4. Desirability ranking of the observed Taguchi formulations

Formulation	Lipid type	Lipid (mg)	Matrix type	Matrix (mg)	Filler	Force (kN)	Hardness (N)	Friability (%)	D1h (%)	D2h (%)	D4h (%)	D8h (%)	Overall desirability
F15	Stearic acid	40	K15M	100	Lactose	10	63	0.42	25.98	39.49	61.98	89.09	0.979
F9	GMS	80	K100M	100	Lactose	5	47	0.44	28.78	44.29	65.78	94.40	0.944
F6	Cetyl alcohol	60	K100LV	100	Avicel	10	68	0.17	31.39	47.95	71.01	100.80	0.902
F3	Beeswax	40	K100M	50	Avicel	10	40	0.23	32.87	54.73	71.82	90.09	0.874
F5	Cetyl alcohol	80	K15M	50	Avicel	5	63	0.27	32.12	50.75	74.58	100.26	0.873
F13	Stearic acid	80	E50LV	25	Avicel	10	81	0.35	35.82	54.41	77.73	99.14	0.835
F4	Beeswax	20	E50LV	100	Avicel	5	19	1.47	28.04	51.42	74.90	98.21	0.797
F10	GMS	60	E50LV	50	Lactose	10	43	0.54	39.01	58.16	86.02	100.30	0.755
F2	Beeswax	60	K15M	25	Lactose	5	33	0.19	51.91	69.67	92.09	96.65	0.626
F11	GMS	40	K100LV	25	Avicel	5	45	0.48	64.23	84.60	95.88	99.64	0.477
F8	Cetyl alcohol	20	K100M	25	Lactose	10	34	0.55	76.52	89.22	94.67	94.94	0.438
F1	Beeswax	80	K100LV	0	Lactose	10	44	0.16	85.33	90.12	93.34	97.45	0.414
F14	Stearic acid	60	K100M	0	Avicel	5	95	0.38	74.25	92.76	97.38	100.42	0.369
F16	Stearic acid	20	K100LV	50	Lactose	5	41	0.69	57.81	80.81	99.65	99.92	0.337
F7	Cetyl alcohol	40	E50LV	0	Lactose	5	41	4.02	91.27	91.35	92.30	93.48	0.301
F12	GMS	20	K15M	0	Avicel	10	51	0.39	93.06	97.34	98.50	100.80	0.229

D1h, D2h, D4h, and D8h represent percent drug release at 1, 2, 4, and 8 h, respectively. Overall desirability was calculated using the geometric mean of individual desirability scores for hardness, friability, and dissolution responses.

Table 5. Top AI-predicted virtual formulation candidates from digital design space screening

Rank	Lipid type	Lipid (mg)	Matrix type	Matrix (mg)	Filler	Force (kN)	Predicted hardness (N)	Predicted friability (%)	Predicted D1h (%)	Predicted D2h (%)	Predicted D4h (%)	Predicted D8h (%)	AI desirability
1	Beeswax	80	K15M	100	Avicel	10	55.42	0.672	15.42	27.19	58.22	97.24	1.000
2	Beeswax	75	E50LV	100	Avicel	10	48.34	0.672	11.34	28.21	59.83	97.25	1.000
3	Beeswax	70	K15M	100	Avicel	10	51.66	0.672	14.05	30.00	59.92	97.23	1.000
4	Stearic acid	80	K15M	100	Avicel	10	84.22	0.672	15.91	27.50	59.14	97.22	1.000
5	Beeswax	80	K100M	100	Avicel	10	56.62	0.672	15.04	31.94	58.72	97.21	1.000
6	Beeswax	75	K15M	100	Avicel	10	53.54	0.672	14.57	28.60	59.07	97.24	1.000
7	Beeswax	75	K100M	100	Avicel	10	54.74	0.672	13.68	33.35	59.57	97.19	1.000
8	Beeswax	80	K15M	95	Avicel	10	55.74	0.672	15.74	29.34	59.50	97.24	1.000
9	Stearic acid	75	K15M	100	Avicel	10	82.34	0.672	15.54	28.90	59.99	97.19	1.000
10	Cetyl alcohol	80	E50LV	100	Avicel	10	64.22	0.672	18.68	29.47	59.06	97.39	1.000
11	Stearic acid	80	E50LV	100	Avicel	10	79.02	0.672	11.82	27.11	59.90	97.26	1.000
12	Beeswax	80	E50LV	100	Avicel	10	50.22	0.672	12.36	26.81	58.98	97.24	1.000
13	Cetyl alcohol	75	E50LV	100	Avicel	10	62.34	0.672	18.47	30.88	59.91	97.47	1.000
14	Stearic acid	80	K100M	100	Avicel	10	85.42	0.672	16.85	32.25	59.64	97.24	1.000

D1h, D2h, D4h, and D8h represent predicted percent drug release at 1, 2, 4, and 8 h, respectively. These are in silico candidates generated within the original experimental factor ranges and require experimental confirmation before formulation selection.

Table 6. Comparison of the published optimized formulation with AI-predicted responses

Published optimized composition: stearic acid 80 mg, Methocel K15M 100 mg, lactose, and compression force of 10 kN.

Parameter	Published optimized formulation [3]	AI-predicted value
Hardness (N)	71.50	71.33
Friability (%)	0.66	0.672
Dissolution 1 h (%)	28.88	21.91
Dissolution 2 h (%)	43.58	30.73
Dissolution 4 h (%)	57.54	61.81
Dissolution 8 h (%)	88.13	96.99
Overall desirability	—	0.988

The published optimized formulation was not used for model training or leave-one-out cross-validation. It was evaluated only as a reference composition within the AI-generated virtual design space. The comparison should be interpreted as a computational consistency check, not as independent experimental validation.

Figure 1. Computational workflow for explainable machine-learning re-analysis of published Taguchi-QbD data for pregabalin extended-release tablets. The workflow included data extraction, preprocessing, response-specific model comparison, leave-one-out cross-validation, explainability analysis, desirability scoring, virtual design-space screening, and comparison with the published optimized formulation.

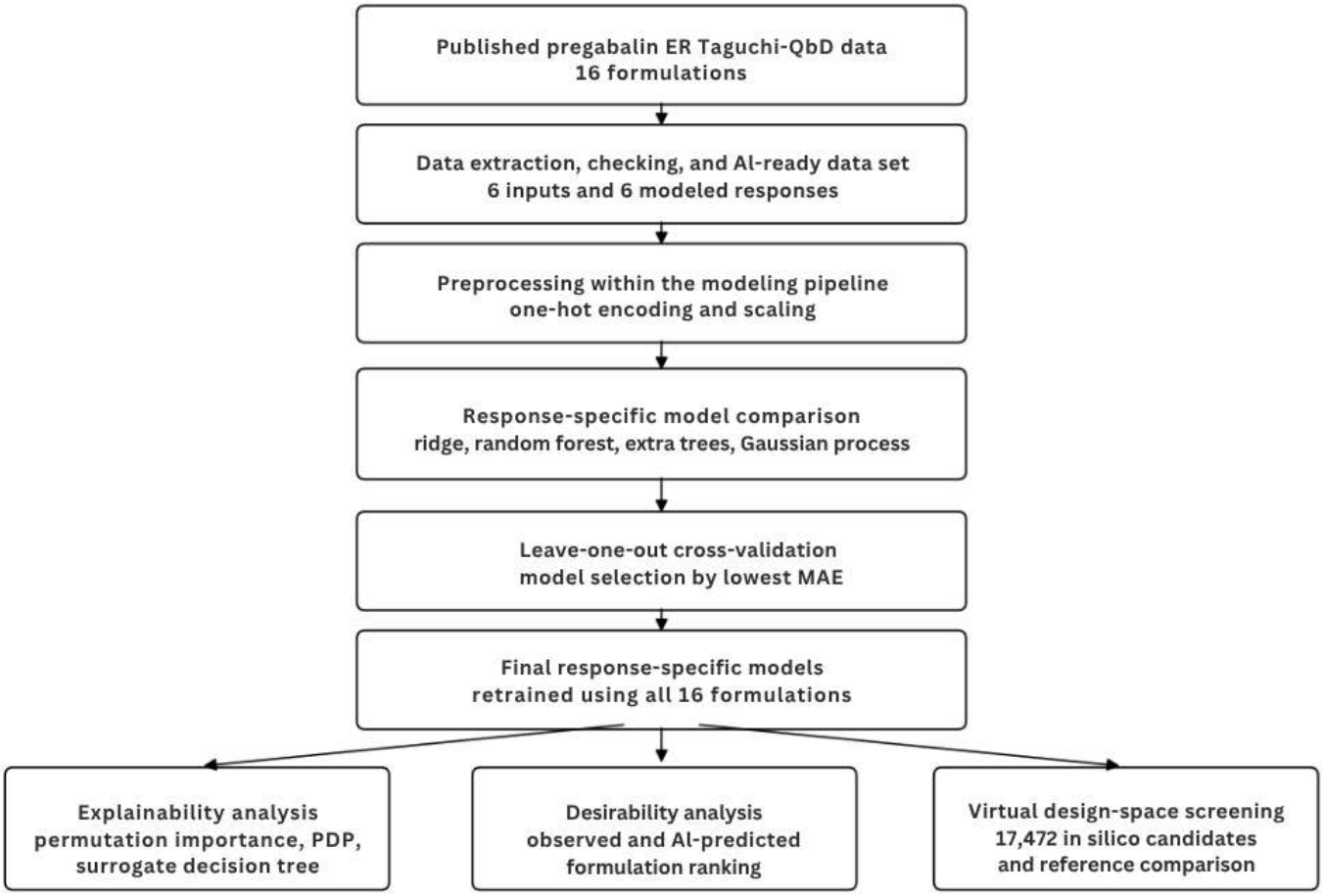


Figure 2. Observed-versus-predicted plots for the selected response-specific machine-learning models. Leave-one-out cross-validation was used to compare observed experimental values with predicted values for hardness, friability, and dissolution at 1, 2, 4, and 8 h.

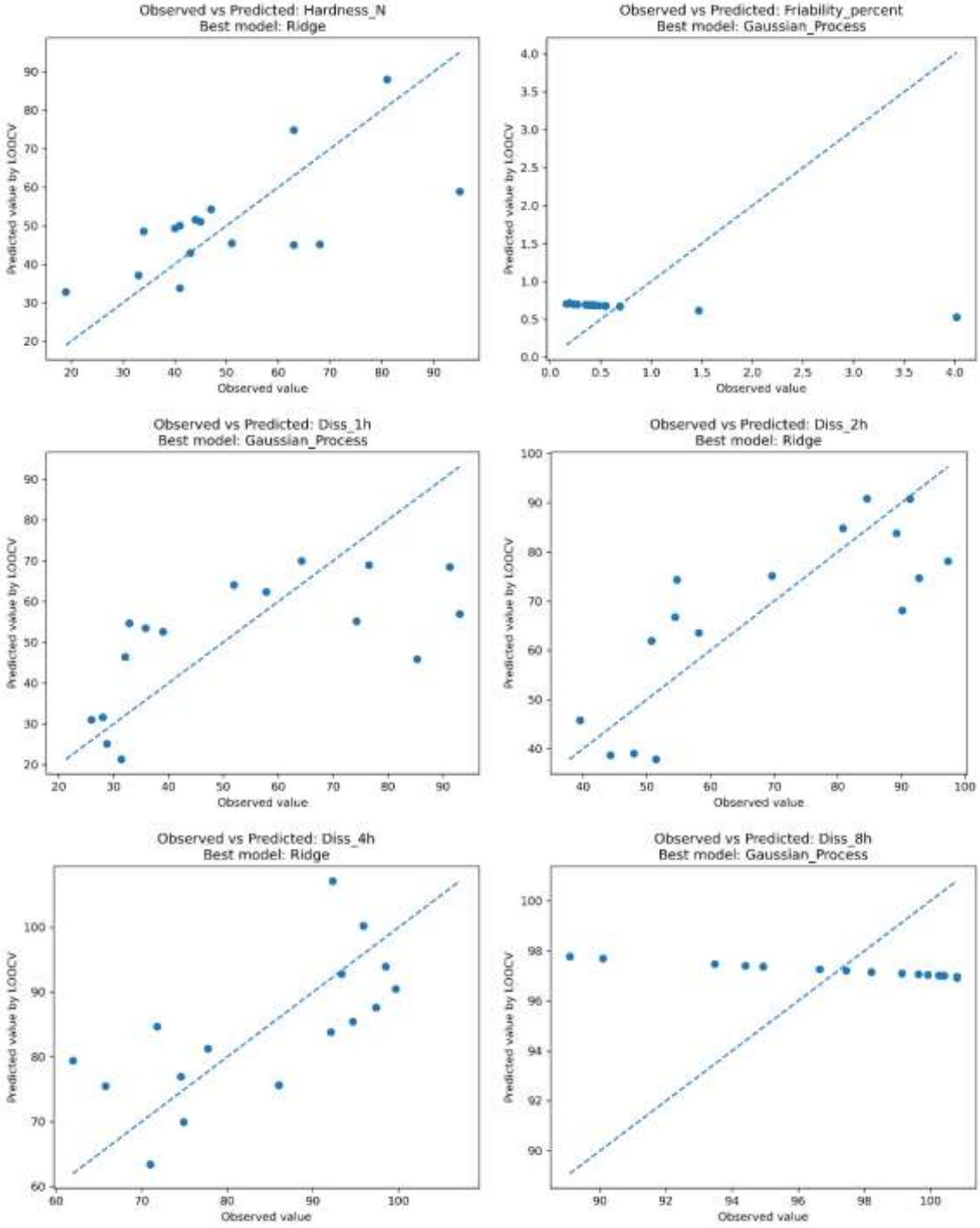


Figure 3. Permutation feature importance of formulation and process variables for response-specific final models.

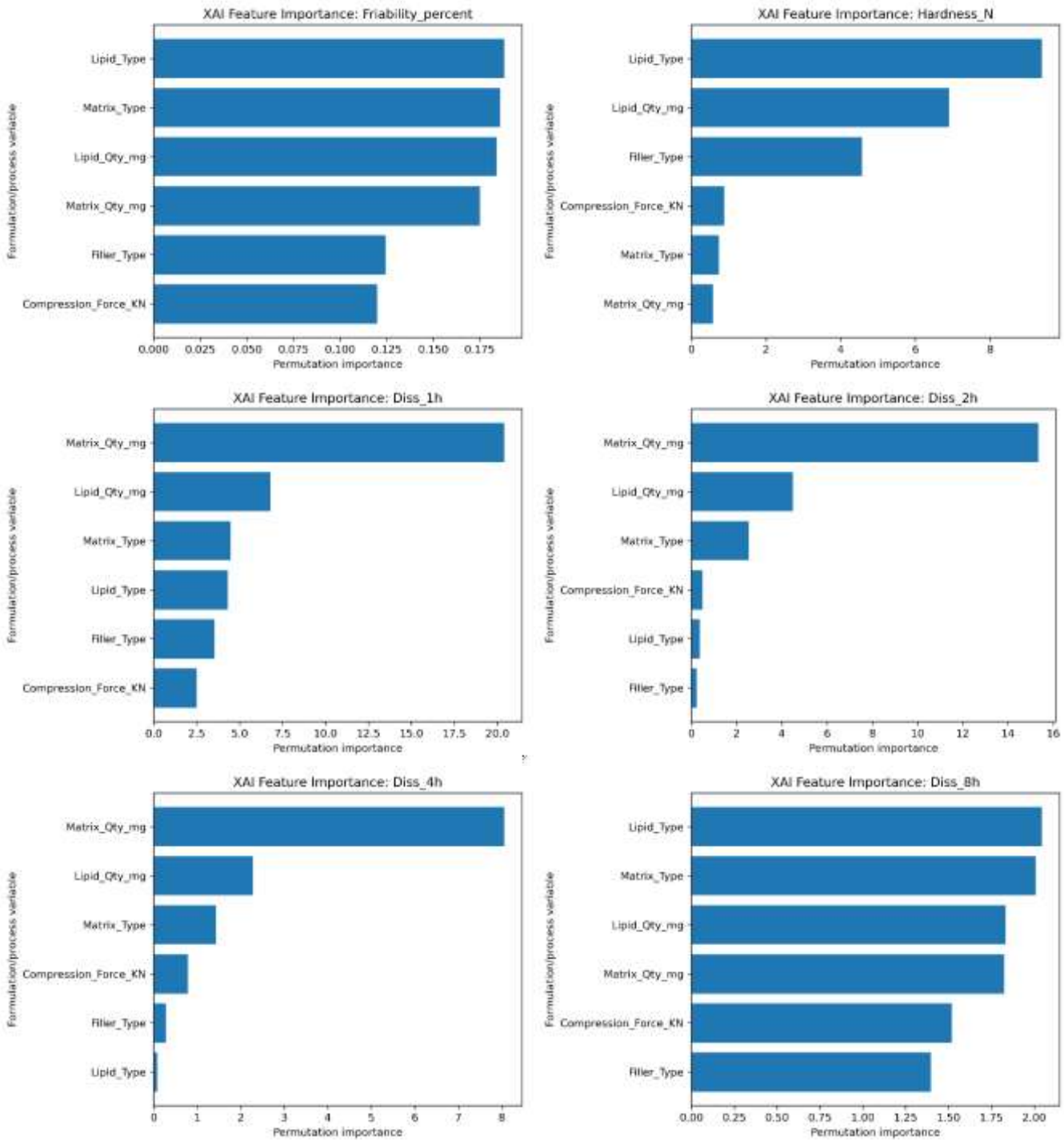


Figure 4. Partial dependence plots showing the effect of numerical formulation and process variables on predicted dissolution responses

