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.

# Machine Learning Modeling of D-Optimal Design Data for Metformin Hydrochloride Orally Disintegrating Tablets

**Sanju**

1. SR University, Ananthasagar, Hasanparthy (PO), Warangal-506371, Telangana, India
https://orcid.org/0000-0002-9148-077X

**Suresh Yarlagadda**

1. SR University, Ananthasagar, Hasanparthy (PO), Warangal-506371, Telangana, India
https://orcid.org/0009-0007-7096-4488

**Vinay Kumar**

2. Department of Statistics, Central University of Haryana, Mahendergarh, Haryana, India
https://orcid.org/0000-0003-1787-5172

**Sreenath Sriram**

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

**Dr. Nabeel Ahmad**

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

**Dr. Sonal Solanki**

5. 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

5. 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**

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

**Dr. Yamsani Shravan**

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

**Dr. Sreekanth Thota**

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

**Dr Kapil Kalra**

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

**Kiran Dangwal**

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

**Dr Shibha Shiphalika Morris**

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

**Neetu Rajpoot**

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

**Dr Saumya Mishra**

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

---

**Research Article**

**Keywords:** Metformin hydrochloride, orally disintegrating tablets, D-optimal design, Quality by Design, machine learning, Gaussian Process Regression, explainable modeling, tablet performance

**Posted Date:** May 14th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9699705/v1

**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.

---

# Abstract

Metformin hydrochloride orally disintegrating tablets require an appropriate balance between mechanical strength and rapid disintegration, particularly because the drug is administered at a relatively high dose and has unfavorable taste characteristics. The present study applied machine learning to published D-optimal design data for metformin hydrochloride orally disintegrating tablets to evaluate whether formulation and process variables could predict powder-flow and tablet-performance responses. The dataset contained 30 experimental runs with four inputs: co-processed metformin hydrochloride ratio, co-processed excipient ratio, compression pressure, and excipient brand. Five responses were modeled separately: angle of repose, Hausner's ratio, Carr's index, hardness, and disintegration time. Polynomial Ridge Regression, Random Forest, Extra Trees, Gradient Boosting, Support Vector Regression, Gaussian Process Regression, and k-nearest neighbors were compared using leave-one-out cross-validation. Model performance was assessed using mean absolute error, root mean square error, and cross-validated $R^2$. Gaussian Process Regression was the best model for Hausner's ratio, Carr's index, hardness, and disintegration time, while Support Vector Regression performed best for angle of repose. Tablet hardness and disintegration time were predicted with the strongest performance, with $R^2$ values of 0.840 and 0.858, respectively. In contrast, angle of repose was weakly predicted, with an $R^2$ of 0.158. Permutation importance and partial dependence analysis indicated that compression pressure and excipient brand were the main contributors to tablet-level responses. The findings show that machine learning can support interpretation of D-optimal formulation data, especially for tablet-performance attributes, while predictions remain limited by dataset size and absence of new external validation.


# 1. Introduction

Orally disintegrating tablets are patient-centered solid dosage forms designed to disintegrate rapidly in the oral cavity without the need for water. They are particularly useful for patients who experience difficulty swallowing conventional tablets or capsules, including elderly patients, patients with dysphagia, and individuals receiving long-term therapy. In chronic diseases, adherence is influenced not only by pharmacological efficacy but also by convenience, acceptability, and ease of administration. Metformin hydrochloride is widely used in the management of type 2 diabetes mellitus; however, conventional metformin tablets can be inconvenient for some patients because of the relatively high dose and large tablet size. For such patients, an orally disintegrating tablet may provide a practical alternative while retaining the advantages of a solid oral dosage form [1,2].

The formulation of metformin hydrochloride as an orally disintegrating tablet remains technically demanding. Metformin hydrochloride is highly water soluble, intensely bitter, and generally requires high drug loading. These properties make it difficult to achieve an acceptable balance between palatability, mechanical strength, rapid disintegration, and tablet uniformity. A robust orally disintegrating tablet must withstand handling, packaging, and transport, but must also disintegrate quickly after contact with saliva. Increasing compression pressure may improve hardness and reduce friability, but excessive

compaction can reduce porosity and delay disintegration. Similarly, the type and proportion of co-processed excipients can influence wetting, swelling, powder flow, compactability, and tablet breakup. Therefore, formulation development requires a systematic understanding of how formulation composition and process conditions affect critical quality attributes [2–4].

Quality by Design provides a structured framework for developing such understanding. Within this framework, design of experiments is widely used to study the effects of multiple formulation and process variables with an efficient number of experimental runs. D-optimal designs are particularly useful when the experimental space contains mixture variables, continuous process variables, and categorical material variables. In orally disintegrating tablet development, this type of design can be used to evaluate the combined effects of drug-to-excipient ratio, compression pressure, and excipient type on powder-flow properties and tablet performance. The resulting models can support formulation optimization and help identify conditions that provide an acceptable balance between hardness, disintegration time, and flow-related properties [2].

Although D-optimal modeling is valuable, formulation systems may contain nonlinear relationships and interactions that are not always fully represented by conventional polynomial models. Tablet performance is governed by several interacting mechanisms, including particle rearrangement, deformation, fragmentation, interparticulate bonding, water penetration, disintegrant swelling, and matrix porosity. These mechanisms may be affected simultaneously by formulation ratio, compression pressure, and excipient-specific material properties. When commercial co-processed excipients are used, interpretation becomes more complex because each excipient system may differ in composition, particle engineering, porosity, compressibility, and disintegrant functionality. These features provide a rationale for applying complementary computational approaches to existing design-of-experiments datasets [5–7].

Machine learning can support formulation modeling by identifying nonlinear and interaction-driven relationships without requiring a fixed equation to be specified in advance. In pharmaceutical development, machine-learning models have been used to predict critical quality attributes, compare response predictability, identify influential variables, and support design-space visualization. However, model selection must be appropriate to the size and structure of the dataset. Many formulation datasets generated through planned experimental designs are small. In such cases, deep learning models are generally unsuitable because they require larger datasets and can easily overfit. Small-data regression methods, such as Gaussian Process Regression, Support Vector Regression, tree-based ensemble methods, k-nearest neighbors, and regularized polynomial regression, are more appropriate when combined with rigorous validation [5–7].

Validation is especially important when machine learning is applied to small experimental datasets. A random train-test split can be unstable when only a few dozen observations are available, because model performance may depend strongly on which runs are placed in the test set. Leave-one-out cross-validation is a practical alternative for such datasets. In this approach, each experimental run is used

once as an unseen test observation, while the remaining runs are used for training. This provides an out-of-sample prediction for every run and allows performance to be assessed using metrics such as mean absolute error, root mean square error, and cross-validated coefficient of determination. These metrics help determine whether the model provides useful predictive value beyond a simple mean-response baseline [6,8].

Interpretability is also essential for the use of machine learning in formulation science. A predictive model is useful only when its outputs can be related to pharmaceutical understanding. Model-agnostic approaches such as permutation importance and partial dependence analysis provide a transparent way to examine the contribution of formulation and process variables. Permutation importance estimates how strongly model performance depends on each input variable, whereas partial dependence plots show the average modeled effect of numeric variables such as compression pressure or formulation ratio. These tools help translate model outputs into formulation-relevant conclusions and reduce the risk of treating machine learning as a black-box exercise [7,9].

The published D-optimal design dataset for metformin hydrochloride orally disintegrating tablets provides a suitable case for machine-learning modeling. The dataset contains 30 experimental runs with defined formulation and process inputs, including co-processed metformin hydrochloride ratio, co-processed excipient ratio, compression pressure, and excipient brand. The reported responses include angle of repose, Hausner's ratio, Carr's index, hardness, and disintegration time [10]. This structure allows response-wise modeling of both powder-flow and tablet-performance attributes. The present work does not claim new formulation development or new experimental optimization. Instead, it uses the published design data as a controlled dataset to evaluate how small-data machine-learning models perform across different critical quality attributes.

Accordingly, the objective of this study was to apply machine-learning modeling to D-optimal design data for metformin hydrochloride orally disintegrating tablets. Multiple small-data regression algorithms were compared using leave-one-out cross-validation. The analysis focused on identifying which responses could be predicted reliably from the available formulation and process variables and which responses remained difficult to model. Permutation importance and partial dependence analysis were then used to interpret the relative influence of formulation ratio, compression pressure, and excipient brand. By keeping the analysis within the original experimental domain, this work aims to show how machine learning can provide a complementary digital layer for interpreting Quality by Design formulation data without overextending beyond the available evidence.

# 2. Materials and Methods

### 2.1 Data source

This study used a published D-optimal design dataset for metformin hydrochloride orally disintegrating tablets as the source data for machine-learning modeling [10]. In the source study, orally disintegrating tablets were prepared using spray-dried co-processed metformin hydrochloride, commercially available

co-processed excipients, and direct compression. The experimental design included mixture variables, a continuous process variable, and a categorical excipient variable. This structure made the dataset suitable for response-wise computational modeling of formulation and process effects.

No new tablet preparation, laboratory testing, or experimental optimization was performed in the present work. The analysis was restricted to the reported D-optimal design data and was conducted to examine whether small-data machine-learning models could predict and interpret the relationships between formulation/process variables and measured powder and tablet responses. The overall computational workflow is shown in Fig. 1.

Only the 30-run D-optimal design table was used for model training and validation. Additional data reported in the source study, including solid-state characterization, scanning electron microscopy, dissolution testing, taste evaluation, and stability testing, were not used as machine-learning inputs because they were not part of the 30-run design matrix. These data were considered only as formulation context during interpretation.

### 2.2 Dataset preparation

The numerical dataset was prepared by extracting the values from the published D-optimal design table [10]. Each experimental run was entered as one row, and each formulation variable or measured response was entered as a separate column. The final analytical dataset contained 30 rows, four input variables, and five response variables. The extracted data were checked for consistency of numerical values, excipient names, response units, and completeness before model development.

Formulation numbers were retained only as row identifiers and were not used as predictive variables. Run-pattern codes were also excluded because they are experimental labels and do not represent independent formulation or process information. The final machine-learning dataset therefore included only variables with direct formulation meaning: co-processed metformin hydrochloride ratio, co-processed excipient ratio, compression pressure, and excipient brand.

The full extracted dataset is provided as Supplementary Table S1. A summary of the variables used for modeling is provided in Table 1. No experimental run was removed from the dataset, and no response value was imputed.

### 2.3 Variables and responses

Four variables were used as model inputs. The first input was the ratio of co-processed metformin hydrochloride in the formulation, and the second input was the ratio of co-processed excipient. These two variables represented the mixture composition of the tablet blend. The third input was compression pressure, expressed in psi, which represented the main process variable during tableting. The fourth input was co-processed excipient brand, which was treated as a categorical material variable. The excipient brands included Pharmaburst, Ludiflash, F-Melt, Prosolv, and SmartEX-QD [10].

Five measured responses were modeled separately. Three responses described powder-flow behavior: angle of repose, Hausner's ratio, and Carr's index. Two responses described tablet performance: hardness and disintegration time. These responses were selected because they were the critical responses reported in the D-optimal design dataset and are relevant to orally disintegrating tablet development. Powder-flow responses are important for blend handling and die filling, whereas hardness and disintegration time describe the balance between mechanical integrity and rapid tablet breakup.

Each response was treated as a separate single-output regression problem. This approach was selected because the responses may be controlled by different formulation mechanisms and may not have the same degree of predictability from the available input variables. The complete list of inputs and responses, including their type and studied range, is summarized in Table 1.

### 2.4 Data preprocessing

The dataset was inspected before modeling to identify missing values, inconsistent entries, and nonnumeric response values. All 30 experimental runs contained complete values for the four input variables and five responses; therefore, no row deletion or imputation was required. Numeric variables were retained in their original units to preserve formulation meaning during interpretation.

The categorical excipient-brand variable was converted into machine-readable form using one-hot encoding. This approach allowed the models to distinguish between excipient brands without assigning an artificial numerical order to the categories. For algorithms sensitive to the scale of input variables, including Support Vector Regression, Gaussian Process Regression, k-nearest neighbors, and Polynomial Ridge Regression, numeric variables were standardized. For tree-based algorithms, including Random Forest, Extra Trees, and Gradient Boosting, numeric scaling was not required.

All preprocessing steps were embedded within the machine-learning pipeline. During cross-validation, scaling and one-hot encoding were fitted only on the training portion and then applied to the held-out test observation. This procedure was used to avoid data leakage and to provide a more reliable estimate of out-of-sample prediction performance.

### 2.5 Model development

A model-zoo approach was used to compare multiple regression algorithms suitable for a small formulation dataset. The tested models included Polynomial Ridge Regression, Random Forest Regression, Extra Trees Regression, Gradient Boosting Regression, Support Vector Regression with a radial basis function kernel, Gaussian Process Regression, and k-nearest neighbors regression. A mean-response baseline model was also included to determine whether the machine-learning models provided improvement over simple average prediction. The complete model set is summarized in Table 2.

Polynomial Ridge Regression was included as a regularized polynomial comparator because conventional design-of-experiments analysis commonly uses polynomial response models. Ridge regularization was used to reduce instability caused by correlated variables and the small sample size.

Random Forest and Extra Trees were included because they can capture nonlinear relationships and interaction effects without requiring a predefined equation. Gradient Boosting was included as another ensemble method that can improve prediction by sequentially reducing residual error.

Support Vector Regression was selected because it is suitable for small nonlinear regression problems. Gaussian Process Regression was included because it is well suited to structured experimental-design datasets and can represent smooth response-surface-like behavior. k-nearest neighbors regression was used as a simple nonparametric comparator. Artificial neural networks and deep learning models were not used because the dataset contained only 30 experimental runs, which would create a high risk of overfitting and unstable prediction.

### 2.6 Model validation

Model validation was performed using leave-one-out cross-validation. This strategy was selected because the dataset contained only 30 experimental runs, making a conventional fixed train-test split unstable and dependent on which runs were assigned to the test set. In each validation cycle, one experimental run was held out as the test observation, and the remaining 29 runs were used for model training. The trained model then predicted the response value for the held-out run.

This process was repeated 30 times for each response and each algorithm so that every formulation run served once as an unseen test observation. The cross-validated predictions were then compared with the observed values to calculate model performance. This approach allowed all available observations to contribute to model evaluation while still generating out-of-sample predictions.

The same validation procedure was applied to all algorithms and all responses. Because preprocessing was included within the modeling pipeline, each cross-validation cycle performed scaling and encoding using only the training data. The held-out observation remained unseen until prediction. The overall validation workflow is summarized in Fig. 1.

### 2.7 Performance metrics

Model performance was assessed using mean absolute error, root mean square error, and the leave-one-out cross-validated coefficient of determination. Mean absolute error was used to estimate the average magnitude of prediction error in the original units of each response. Root mean square error was also calculated because it gives greater weight to larger prediction errors and is useful for comparing the overall predictive performance of different algorithms.

The cross-validated coefficient of determination was used to evaluate the proportion of response variability explained by each model during leave-one-out prediction. Positive $R^2$ values indicated that the model performed better than a mean-response prediction, whereas values close to zero or negative values indicated weak predictive utility. Because the dataset was small, $R^2$ was interpreted together with mean absolute error and root mean square error rather than used as the only basis for model selection.

For each response, the best model was selected according to the lowest root mean square error obtained during leave-one-out cross-validation. Observed-versus-predicted plots were generated for the best model of each response to visually assess prediction agreement. The complete performance results are reported in Table 3, and the best-model summary is reported in Table 4.

2.8 Model interpretation

Model interpretation was performed using permutation importance and partial dependence analysis. These methods were selected because they provide transparent, model-agnostic information that can be related to formulation understanding. Permutation importance was used to estimate the relative contribution of each input variable to model prediction. In this method, the values of one input variable are randomly permuted while the other variables are kept unchanged. The resulting increase in prediction error reflects the importance of that variable for the selected response.

Permutation importance was calculated for the best-performing model of each response after fitting the selected model to the complete 30-run dataset. The interpreted input variables were co-processed metformin hydrochloride ratio, co-processed excipient ratio, compression pressure, and excipient brand. The results were expressed as mean importance values with variability across repeated permutations. These values were used to compare the relative influence of formulation composition, process condition, and excipient type.

Partial dependence plots were generated for the numeric variables to examine the average modeled effect of each factor on the predicted response. These plots were prepared for co-processed metformin hydrochloride ratio, co-processed excipient ratio, and compression pressure. Because the two mixture variables are complementary within the design, their partial dependence trends were interpreted cautiously and were not considered independent causal effects. The interpretation outputs are presented in Fig. 3 and Fig. 4.

2.9 Design-space prediction

Design-space prediction was performed as an exploratory analysis after selecting the best model for each response. The purpose was to visualize how the model predicted tablet performance across the studied formulation and process region. The analysis focused on hardness and disintegration time because these responses represent the key balance required for orally disintegrating tablets.

Prediction grids were generated across the studied range of co-processed metformin hydrochloride ratio and compression pressure. The co-processed excipient ratio was adjusted according to the mixture relationship so that the total formulation ratio remained within the original design domain. Separate prediction maps were generated for each excipient brand. The trained model was then used to predict response values at each grid point.

The resulting maps were used to identify regions predicted to provide adequate hardness and rapid disintegration. For interpretation, hardness of at least 70 N and disintegration time of 60 s or less were

used as practical performance criteria. The predicted maps are presented in Fig. 5. These maps were considered exploratory and were not treated as experimentally confirmed design spaces because no new validation batches were prepared in the present study.

### 2.10 Software

All analyses were performed using Python in a Jupyter Notebook environment. Data handling was performed using pandas and NumPy. Machine-learning models, preprocessing pipelines, leave-one-out cross-validation, performance metrics, permutation importance, and partial dependence analysis were implemented using scikit-learn. Figures were generated using matplotlib and exported at 300 dpi for manuscript preparation.

The computational workflow was designed to be reproducible. The dataset was loaded from a single Excel file, and all preprocessing steps were incorporated into model pipelines. Random seeds were fixed where applicable for stochastic models, including Random Forest, Extra Trees, and Gradient Boosting. The same validation method and performance metrics were applied to all models and responses.

The final workflow generated model-performance tables, observed-versus-predicted values, best-model summaries, permutation-importance outputs, partial dependence plots, and design-space prediction maps. These outputs were saved automatically from the notebook and used for manuscript tables and figures.

## 3. Results

### 3.1 Dataset overview

The dataset used for machine-learning modeling contained 30 experimental runs from the D-optimal formulation design. Each run included four input variables and five measured responses. The input variables were co-processed metformin hydrochloride ratio, co-processed excipient ratio, compression pressure, and excipient brand. The metformin hydrochloride ratio ranged from 0.62 to 0.80, while the co-processed excipient ratio ranged from 0.20 to 0.38. Compression pressure was evaluated at 800, 1200, and 1600 psi. The categorical excipient variable included five brands: Pharmaburst, Ludiflash, F-Melt, Prosolv, and SmartEX-QD. The variables used for modeling are summarized in Table 1.

The response variables represented both powder-flow and tablet-performance attributes. Angle of repose ranged from 26.03° to 36.55°, Hausner’s ratio from 1.25 to 1.45, and Carr’s index from 19.85% to 29.99%. Tablet hardness ranged from 28.50 to 106.69 N, while disintegration time ranged from 17.00 to 272.33 s. This range of values indicated that the dataset contained sufficient response variability to compare model performance across powder-flow and tablet-performance endpoints. The modeling workflow applied to this dataset is shown in Fig. 1.

### 3.2 Model comparison

All models were evaluated using leave-one-out cross-validation for each of the five responses. The complete model-comparison results are presented in Table 3. Model performance differed across responses. In general, powder-flow responses were predicted with lower and more variable accuracy than tablet-performance responses. Among the powder-flow responses, angle of repose showed the weakest prediction, whereas Hausner's ratio and Carr's index showed moderate predictability.

Tablet hardness and disintegration time showed the strongest predictive performance. For hardness, the best model achieved a mean absolute error of 6.80 N, root mean square error of 8.73 N, and cross-validated $R^2$ of 0.840. For disintegration time, the best model achieved a mean absolute error of 18.97 s, root mean square error of 26.84 s, and cross-validated $R^2$ of 0.858. These results indicate that the available formulation and process variables were more informative for tablet-level responses than for angle of repose. The observed-versus-predicted plots for the best models are shown in Fig. 2.

### 3.3 Powder-flow prediction

Prediction of powder-flow responses was limited to moderate. Angle of repose was the least predictable response in the dataset. Support Vector Regression with a radial basis function kernel provided the lowest prediction error for this response, with a mean absolute error of 2.17°, root mean square error of 2.61°, and cross-validated $R^2$ of 0.158. Although the absolute error was not large, the low $R^2$ indicated that the model explained only a small part of the observed variability. The observed-versus-predicted plot also showed broad scatter around the identity line (Fig. 2A).

Hausner's ratio was better predicted than angle of repose. Gaussian Process Regression gave the best performance for Hausner's ratio, with a mean absolute error of 0.022, root mean square error of 0.028, and cross-validated $R^2$ of 0.601. Carr's index was also best predicted by Gaussian Process Regression, with a mean absolute error of 1.16%, root mean square error of 1.90%, and cross-validated $R^2$ of 0.486. These findings suggest that density-derived flow indices were more amenable to prediction than angle of repose. However, the moderate $R^2$ values indicate that powder-flow prediction remained less robust than tablet-response prediction in the present dataset.

### 3.4 Tablet-response prediction

The tablet-performance responses showed the strongest model performance. Hardness was best predicted by Gaussian Process Regression, with a mean absolute error of 6.80 N, root mean square error of 8.73 N, and cross-validated $R^2$ of 0.840. The observed-versus-predicted plot showed good agreement across the hardness range, with most predictions close to the identity line (Fig. 2D). This indicates that the selected formulation and process variables captured the main hardness trend across the studied D-optimal design space.

Disintegration time was also best predicted by Gaussian Process Regression. This response showed the highest cross-validated $R^2$ among all outputs, with a value of 0.858. The mean absolute error was 18.97 s and the root mean square error was 26.84 s. The observed-versus-predicted plot showed reasonable

agreement across the response range, although some deviation was observed for formulations with longer disintegration times (Fig. 2E). Overall, these results show that machine-learning modeling was most useful for tablet-level responses directly related to compaction and orally disintegrating tablet performance.

### 3.5 Best models

The best model for each response was selected using the lowest leave-one-out cross-validated root mean square error. Support Vector Regression was selected only for angle of repose. Gaussian Process Regression was selected for the remaining four responses: Hausner's ratio, Carr's index, hardness, and disintegration time (Table 4). This pattern indicates that Gaussian Process Regression was the most consistent algorithm for the present D-optimal design dataset.

Compared with the mean-response baseline, the selected models improved prediction for most responses. For hardness, the root mean square error decreased from 22.58 N for the baseline model to 8.73 N for Gaussian Process Regression. For disintegration time, the root mean square error decreased from 73.60 s to 26.84 s. The improvements were smaller for powder-flow responses, particularly angle of repose. Therefore, best-model selection confirmed that the predictive benefit of machine learning was strongest for hardness and disintegration time, while prediction of angle of repose remained limited.

### 3.6 Variable importance

Permutation importance analysis was performed using the best model selected for each response. The results are summarized in Table 5 and shown in Fig. 3. For the powder-flow responses, excipient brand showed the highest contribution to model prediction. For angle of repose, excipient brand was the most important variable, followed by compression pressure. A similar pattern was observed for Hausner's ratio and Carr's index, where excipient brand again contributed most strongly to prediction. These results indicate that differences among the co-processed excipient systems were important for modeling flow-related responses.

For tablet hardness, compression pressure was the dominant variable, followed by excipient brand. For disintegration time, compression pressure also showed the highest importance, with excipient brand as the second major contributor. These findings indicate that tablet-level responses were mainly influenced by compaction conditions and excipient type. The co-processed metformin hydrochloride ratio and co-processed excipient ratio also contributed to model prediction, but their individual effects should be interpreted cautiously because the two mixture variables are complementary within the design.

### 3.7 Partial dependence

Partial dependence plots were generated to examine the average modeled effect of numeric variables on each response. The plots are presented in Fig. 4. For angle of repose, the predicted response changed only modestly across the studied ranges of metformin hydrochloride ratio, excipient ratio, and compression pressure. This result was consistent with the weak cross-validated performance observed

for this response and suggests that the selected numeric variables did not strongly explain angle-of-repose variability.

For hardness, compression pressure showed a clear positive effect. Predicted hardness increased as compression pressure increased from 800 to 1600 psi. Disintegration time also increased with compression pressure, indicating that stronger compaction was associated with slower tablet breakup. These paired trends show the main formulation trade-off observed in the dataset: higher compression pressure improved predicted hardness but also tended to prolong predicted disintegration time. The partial dependence results therefore supported the variable-importance findings and provided a formulation-based interpretation of the machine-learning models.

### 3.8 Design-space prediction

Design-space prediction was performed for hardness and disintegration time because these responses define the key performance balance required for orally disintegrating tablets. The predicted response maps are shown in Fig. 5. The maps were generated within the original experimental region and were interpreted as exploratory machine-learning predictions rather than experimentally confirmed design spaces.

Compression pressure strongly influenced both predicted hardness and predicted disintegration time. At lower compression pressure, predicted disintegration time was generally shorter, but hardness was more likely to remain below the practical target of 70 N. At higher compression pressure, predicted hardness increased, but disintegration time also tended to increase. This pattern was consistent with the partial dependence analysis.

When both criteria were considered together, Pharmaburst-containing formulations showed the most favorable predicted region within the explored domain. This region was associated with predicted hardness of at least 70 N and predicted disintegration time of 60 s or less. These results suggest that machine-learning maps can support preliminary visualization of formulation regions that may provide a favorable balance between mechanical strength and rapid disintegration. However, this region should be considered exploratory because no additional confirmatory batches were prepared in the present study.

### 3.9 Optimized formulation prediction

The selected best models were used to predict the responses for the optimized formulation condition reported in the source design study. The input condition consisted of a co-processed metformin hydrochloride ratio of 0.68, a co-processed excipient ratio of 0.32, compression pressure of 1600 psi, and Pharmaburst as the excipient brand. The predicted responses are summarized in Table 6.

For this formulation, the models predicted an angle of repose of 32.37°, Hausner’s ratio of 1.354, Carr’s index of 25.65%, hardness of 75.96 N, and disintegration time of 57.25 s. The predicted hardness was consistent with the intended mechanical performance of the optimized tablet. The predicted

disintegration time also remained within the practical criterion of 60 s, although it was more conservative than the experimentally reported disintegration behavior in the source study.

The optimized-formulation prediction should therefore be interpreted as supportive model-based estimation rather than experimental confirmation. The results indicate that the models captured the general performance direction of the optimized formulation, particularly the high-hardness and rapid-disintegration region associated with Pharmaburst at higher compression pressure. However, the prediction does not replace experimental validation and should be considered only within the original design domain.

# 4. Discussion

The present study applied machine-learning modeling to D-optimal design data for metformin hydrochloride orally disintegrating tablets. The objective was to evaluate whether small-data regression models could predict and interpret powder-flow and tablet-performance responses using formulation and process variables from the published design dataset. The findings showed a clear response-dependent pattern. Tablet-level responses, particularly hardness and disintegration time, were predicted with good leave-one-out cross-validation performance, whereas powder-flow responses showed weaker or moderate predictability (Table 4; Fig. 2). This distinction is important because hardness and disintegration time are central to the performance of orally disintegrating tablets, where mechanical integrity must be balanced with rapid breakup in the oral cavity [2,3].

Gaussian Process Regression was the most consistent algorithm in this dataset. It was selected as the best model for Hausner's ratio, Carr's index, hardness, and disintegration time, while Support Vector Regression was selected only for angle of repose. This outcome is reasonable because the dataset was generated from a structured D-optimal design rather than from a large heterogeneous dataset. In such experimental designs, response behavior is expected to change within a defined formulation and process region. Gaussian Process Regression can represent smooth response-surface-like relationships and is therefore suitable for small structured datasets when validation is carefully applied [6,7,11]. However, this result should not be interpreted as general superiority of Gaussian Process Regression for all formulation problems. Its performance in the present study reflects the structure of the dataset, the number of experimental runs, and the response patterns available for modeling.

Hardness was predicted with strong performance, with a cross-validated $R^2$ of 0.840 and a root mean square error of 8.73 N (Table 4). This suggests that the selected inputs captured the main sources of variation in tablet strength. The importance analysis showed compression pressure as the dominant contributor to hardness, followed by excipient brand (Table 5; Fig. 3). This agrees with established tablet compaction principles. Increasing compression pressure promotes particle rearrangement, densification, deformation, and interparticulate bonding, which generally increases mechanical strength [2,6]. The effect of excipient brand is also formulation-relevant because co-processed orally disintegrating tablet excipients can differ in composition, particle engineering, compressibility,

disintegrant content, and dilution capacity [2,5]. Therefore, different excipient systems may produce different compact strength even at similar formulation ratios and compression pressure.

Disintegration time also showed strong predictability and had the highest cross-validated $R^2$ among all modeled responses. The importance analysis identified compression pressure and excipient brand as the main contributors to disintegration time (Table 5; Fig. 3). Partial dependence analysis further showed that disintegration time increased with compression pressure (Fig. 4). This trend is consistent with the expected effect of compaction on tablet porosity and liquid penetration. Higher compression pressure may strengthen the compact but can reduce pore volume and slow tablet breakup [2,3]. At the same time, the excipient system may counterbalance this effect through wicking, swelling, and rapid water uptake [2,13]. The findings therefore reflect the central formulation trade-off in orally disintegrating tablets: increasing compression pressure improves hardness but may delay disintegration.

The powder-flow responses were less consistently predicted. Angle of repose was the weakest response, with a cross-validated $R^2$ of only 0.158 despite selection of the best-performing model (Table 4; Fig. 2A). This indicates that the available input variables explained only a small portion of the observed variability. Angle of repose can be influenced by factors that were not included in the modeling dataset, such as particle-shape distribution, surface roughness, residual moisture, cohesiveness, electrostatic behavior, and experimental handling conditions [4]. Therefore, weak prediction of angle of repose does not necessarily indicate failure of the modeling workflow. Rather, it suggests that this response requires additional material descriptors for reliable prediction. Hausner's ratio and Carr's index were predicted more effectively than angle of repose, although their performance remained moderate. These density-derived indices may be more directly related to powder packing than angle of repose, but they are still influenced by powder microstructure and interparticulate interactions [4,7].

The design-space prediction provided useful visual interpretation of the hardness–disintegration balance (Fig. 5). The maps showed that lower compression pressure generally favored shorter disintegration time but could compromise hardness, whereas higher compression pressure improved hardness while tending to prolong disintegration. When both criteria were considered together, Pharmaburst-containing formulations showed the most favorable predicted region, with hardness of at least 70 N and disintegration time of 60 s or less within the studied domain. This result supports the role of excipient selection in maintaining tablet strength without excessive loss of disintegration performance [2,5]. However, these maps should be interpreted as exploratory model-based outputs rather than experimentally confirmed design spaces, because no new validation batches were prepared in the present study.

Prediction of the optimized formulation further illustrates both the usefulness and limitation of the approach. For the reported optimized condition, the model predicted hardness of 75.96 N and disintegration time of 57.25 s (Table 6). These values indicate that the model identified the optimized region as mechanically strong and rapidly disintegrating within the practical criterion used in this analysis. However, the predicted disintegration time was more conservative than the experimentally

reported optimized-batch behavior in the source study [10]. This difference may reflect the limited number of training points and the strong pressure–disintegration relationship learned from the dataset. It may also indicate that excipient-specific disintegration mechanisms are not fully represented by the four input variables used in the model.

From a Quality by Design perspective, the study shows that machine learning can provide a complementary layer for interpreting formulation data. The models helped distinguish responses that were well predicted from those that remained uncertain. This is useful because it prevents overgeneralization of AI outputs and supports response-wise interpretation. For hardness and disintegration time, the models provided meaningful predictive and explanatory information. For angle of repose, the results indicated limited predictability and the need for richer material descriptors. Thus, machine learning should be used as an additional decision-support tool rather than as a replacement for experimental design, formulation knowledge, or confirmatory testing [5,7].

Several limitations should be recognized. First, the analysis used only 30 published experimental runs and did not include new laboratory validation [10]. Second, the two mixture variables were complementary, which limits independent interpretation of their individual effects. Third, excipient brand was treated as a categorical variable, but detailed physicochemical descriptors such as particle size, porosity, moisture content, and composition were not included. Fourth, the predicted maps were generated only within the original design range and should not be extrapolated beyond that domain. Future work should include additional confirmatory batches, independent test points, and richer material descriptors, particularly for powder-flow prediction [4,9]. Overall, the findings indicate that small-data machine learning can complement D-optimal formulation modeling, particularly for tablet-performance responses, when validation is rigorous and conclusions remain within the limits of the available experimental evidence.

# 5. Conclusion

This study applied small-data machine-learning models to D-optimal design data for metformin hydrochloride orally disintegrating tablets. The analysis showed that model performance was response dependent. Tablet-level responses were predicted more reliably than powder-flow responses, with hardness and disintegration time showing the strongest leave-one-out cross-validation performance. Gaussian Process Regression was the most consistent algorithm across the dataset and provided the best performance for Hausner’s ratio, Carr’s index, hardness, and disintegration time, while Support Vector Regression performed best for angle of repose.

The interpretation analysis indicated that compression pressure and excipient brand were the main contributors to tablet-performance responses. This finding supports the formulation relevance of compaction conditions and co-processed excipient selection in achieving the required balance between mechanical strength and rapid disintegration. The design-space prediction further suggested that Pharmaburst-containing formulations provided the most favorable model-predicted region within the

studied domain. However, these predictions should be considered exploratory and not experimentally confirmed.

Overall, the study demonstrates that machine learning can provide a useful complementary layer for interpreting D-optimal formulation data, particularly for tablet-performance attributes. At the same time, the limited predictability of angle of repose highlights the need for richer material descriptors when modeling powder-flow behavior. Future work should include independent validation batches and additional physicochemical descriptors to strengthen model reliability and support broader application in Quality by Design-based formulation development.

# Declarations

### Author Contributions

S.K., S.Y., D.S.S., and N.S. conceived the study and developed the overall machine-learning modeling framework for the D-optimal formulation dataset. S.K. and S.Y. contributed to dataset organization, secondary data extraction, preliminary interpretation, and manuscript drafting. V.K. contributed to statistical and computational interpretation, model-validation strategy, and review of the analytical workflow. D.S.S. and N.S. supervised the formulation interpretation, QbD context, machine-learning workflow, and final manuscript refinement. S.Sr. and Y.S. assisted in interpretation of tablet-performance responses, model outputs, and formulation-related findings. N.A., J.M.K., S.T., K.K., K.D., S.S.M., N.R., and S.M. contributed to literature review, scientific interpretation, manuscript editing, and critical review of the final article. All authors reviewed and approved the final manuscript.

**Funding:** No external funding was received for this work.

**Ethics Declarations:** This study was based on secondary computational analysis of a previously published D-optimal formulation dataset. No new human participants, animal experiments, clinical samples, biological specimens, or patient data were used in the present work.

**Clinical Trial Number:** Not applicable.

**Conflict of Interest:** The authors declare that they have no competing interests.

### Data Availability

The present study is based on secondary computational analysis of the D-optimal design dataset reported in the previously published study by Roslan et al., entitled “Formulation Optimization and Evaluation of Metformin Hydrochloride Orally Disintegrating Tablets Using Spray Drying and D-Optimal Design of Experiments.” No new experimental batches, biological experiments, animal studies, clinical samples, or patient data were generated in the present work. The processed dataset, model outputs, and computational workflow are available from the corresponding author upon reasonable request.

**Supplementary Material**

The processed dataset, machine-learning model outputs, Python/Jupyter Notebook workflow, leave-one-out cross-validation results, observed-versus-predicted values, permutation-importance outputs, partial-dependence plots, and exploratory design-space prediction maps are available from the corresponding author upon reasonable request.

# References


1. Agiba AM, Ahmed MA. Insights into formulation technologies and novel strategies for the design of orally disintegrating dosage forms: a comprehensive industrial review. Int J Pharm Pharm Sci. 2019;11(9):8–20. doi:10.22159/ijpps.2019v11i9.34828
2. Aodah AH, Fayed MH, Alalaiwe A, Alsulays BB, Aldawsari MF, Khafagy ES. Design, optimization, and correlation of in vitro/in vivo disintegration of novel fast orally disintegrating tablet of high-dose metformin hydrochloride using moisture activated dry granulation process and quality by design approach. Pharmaceutics. 2020;12(7):598. doi:10.3390/pharmaceutics12070598
3. Momeni M, Afkanpour M, Rakhshani S, Mehrabian A, Tabesh H. A prediction model based on artificial intelligence techniques for disintegration time and hardness of fast disintegrating tablets in pre-formulation tests. BMC Med Inform Decis Mak. 2024;24(1):88. doi:10.1186/s12911-024-02485-4
4. Diaz LP, Brown CJ, Ojo E, Mustoe CL, Florence AJ. Machine learning approaches to the prediction of powder flow behaviour of pharmaceutical materials from physical properties. Digit Discov. 2023;2(3):692–701. doi:10.1039/d2dd00106c
5. 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
6. Kim SH, Han SH, Seo DW, Kang MJ. Evaluation of prediction models for the capping and breaking force of tablets using machine learning tools in wet granulation commercial-scale pharmaceutical manufacturing. Pharmaceuticals. 2025;18(1):23. doi:10.3390/ph18010023
7. Ye Z, Yang W, Yang Y, Ouyang D. Interpretable machine learning methods for in vitro pharmaceutical formulation development. Food Front. 2021;2(2):195–207. doi:10.1002/fft2.78
8. Schmitt J, Baumann JM, Morgen MM. Predicting spray dried dispersion particle size via machine learning regression methods. Pharm Res. 2022;39(12):3223–3239. doi:10.1007/s11095-022-03370-3
9. Bannigan P, Bao Z, Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A, et al. Machine learning models to accelerate the design of polymeric long-acting injectables. Nat Commun. 2023;14(1):35. doi:10.1038/s41467-022-35343-w
10. Roslan MF, Thiruvarselva K, Kanakal MM, Chik Z, Widodo RT. Formulation optimization and evaluation of metformin hydrochloride orally disintegrating tablets using spray drying and D-optimal design of experiments. J Pharm Innov. 2026;21:202. doi:10.1007/s12247-026-10386-4

11. Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42. doi:10.1038/s41524-023-01000-z
12. Bounab Y, Antikainen O, Sivén M, Juppo A, et al. Advancing direct tablet compression with AI: a multi-task framework for quality control, batch acceptance, and causal analysis. Eur J Pharm Sci. 2025;212:107142. doi:10.1016/j.ejps.2025.107142
13. Ghazwani M, Hani U. Determination of disintegration time using formulation data for solid dosage oral formulations via advanced machine learning integrated optimizer models. Sci Rep. 2025;15(1):30118. doi:10.1038/s41598-025-15996-5

# Tables

**Table 1. Variables used for machine-learning modeling**

| Variable code | Variable name | Type | Range or categories | Role in model |
|---|---|---|---|---|
| X1 | Co-processed metformin hydrochloride ratio | Numeric mixture variable | 0.62–0.80 | Input |
| X2 | Co-processed excipient ratio | Numeric mixture variable | 0.20–0.38 | Input |
| X3 | Compression pressure | Numeric process variable | 800, 1200, 1600 psi | Input |
| X4 | Co-processed excipient brand | Categorical material variable | Pharmaburst, Ludiflash, F-Melt, Prosolv, SmartEX-QD | Input |
| Y1 | Angle of repose | Numeric response | 26.03–36.55° | Output |
| Y2 | Hausner’s ratio | Numeric response | 1.25–1.45 | Output |
| Y3 | Carr’s index | Numeric response | 19.85–29.99% | Output |
| Y4 | Hardness | Numeric response | 28.50–106.69 N | Output |
| Y5 | Disintegration time | Numeric response | 17.00–272.33 s | Output |

The dataset contained 30 experimental runs extracted from the published D-optimal design table [10]. Each response was modeled separately as a single-output regression task.

**Table 2. Machine-learning models used in the study**

| Model | Model type | Purpose in the analysis | Preprocessing applied |
|---|---|---|---|
| Mean baseline | Naïve comparator | Used to determine whether machine-learning models improved over average-response prediction | One-hot encoding for categorical input |
| Polynomial Ridge Regression | Regularized polynomial model | Used as a QbD-like polynomial comparator with regularization | Standardization and one-hot encoding |
| Random Forest Regression | Tree-based ensemble | Used to capture nonlinear relationships and interaction effects | One-hot encoding |
| Extra Trees Regression | Randomized tree-based ensemble | Used to assess nonlinear prediction with additional randomization | One-hot encoding |
| Gradient Boosting Regression | Boosted ensemble | Used to sequentially reduce residual prediction error | One-hot encoding |
| Support Vector Regression | Kernel-based regression | Used for nonlinear small-data modeling with radial basis function kernel | Standardization and one-hot encoding |
| Gaussian Process Regression | Probabilistic kernel-based regression | Used for smooth response-surface-like prediction in a structured experimental dataset | Standardization and one-hot encoding |
| k-Nearest Neighbors Regression | Nonparametric regression | Used as a simple distance-based comparator | Standardization and one-hot encoding |

**Table 3. Summary of best model performance for each response**

| Response | Best model | MAE | RMSE | LOOCV $R^2$ | Interpretation |
|---|---|---|---|---|---|
| Angle of repose | Support Vector Regression | 2.17° | 2.61° | 0.158 | Weak prediction |
| Hausner's ratio | Gaussian Process Regression | 0.022 | 0.028 | 0.601 | Moderate prediction |
| Carr's index | Gaussian Process Regression | 1.16% | 1.90% | 0.486 | Moderate prediction |
| Hardness | Gaussian Process Regression | 6.80 N | 8.73 N | 0.840 | Good prediction |
| Disintegration time | Gaussian Process Regression | 18.97 s | 26.84 s | 0.858 | Good prediction |

MAE: mean absolute error; RMSE: root mean square error; LOOCV: leave-one-out cross-validation.

**Table 4. Improvement of selected models over mean-response baseline**

| Response | Baseline RMSE | Best model | Best-model RMSE | Direction of improvement |
|---|---|---|---|---|
| Angle of repose | Not shown | Support Vector Regression | 2.61° | Limited improvement; weak predictability retained |
| Hausner's ratio | 0.0466 | Gaussian Process Regression | 0.028 | Improved prediction |
| Carr's index | 2.75% | Gaussian Process Regression | 1.90% | Improved prediction |
| Hardness | 22.58 N | Gaussian Process Regression | 8.73 N | Marked improvement |
| Disintegration time | 73.60 s | Gaussian Process Regression | 26.84 s | Marked improvement |

**Table 5. Main variables identified by permutation importance**

| Response | Best model | Most important variable | Second important variable | Key interpretation |
|---|---|---|---|---|
| Angle of repose | Support Vector Regression | Excipient brand | Compression pressure | Flow prediction was mainly associated with excipient type, but overall predictability was weak. |
| Hausner's ratio | Gaussian Process Regression | Excipient brand | Compression pressure | Density-derived flow behavior was influenced mainly by excipient type and compaction-related formulation context. |
| Carr's index | Gaussian Process Regression | Excipient brand | Compression pressure | Carr's index showed stronger dependence on excipient brand than on mixture ratios. |
| Hardness | Gaussian Process Regression | Compression pressure | Excipient brand | Tablet strength was mainly governed by compaction pressure and excipient functionality. |
| Disintegration time | Gaussian Process Regression | Compression pressure | Excipient brand | Tablet breakup behavior was mainly influenced by compaction pressure and excipient type. |

**Table 6. Machine-learning prediction for the reported optimized formulation**

| Response | Best model used | Predicted value | Practical interpretation |
|---|---|---|---|
| Angle of repose | Support Vector Regression | 32.37° | Model-based estimate of powder-flow behavior |
| Hausner's ratio | Gaussian Process Regression | 1.354 | Model-based estimate of density-derived flow behavior |
| Carr's index | Gaussian Process Regression | 25.65% | Model-based estimate of compressibility-related flow index |
| Hardness | Gaussian Process Regression | 75.96 N | Predicted to meet the intended mechanical-strength region |
| Disintegration time | Gaussian Process Regression | 57.25 s | Predicted to remain within the practical criterion of ≤60 s |

**Input condition:** Co-processed metformin hydrochloride ratio = 0.68; co-processed excipient ratio = 0.32; compression pressure = 1600 psi; excipient brand = Pharmaburst.

# Figures

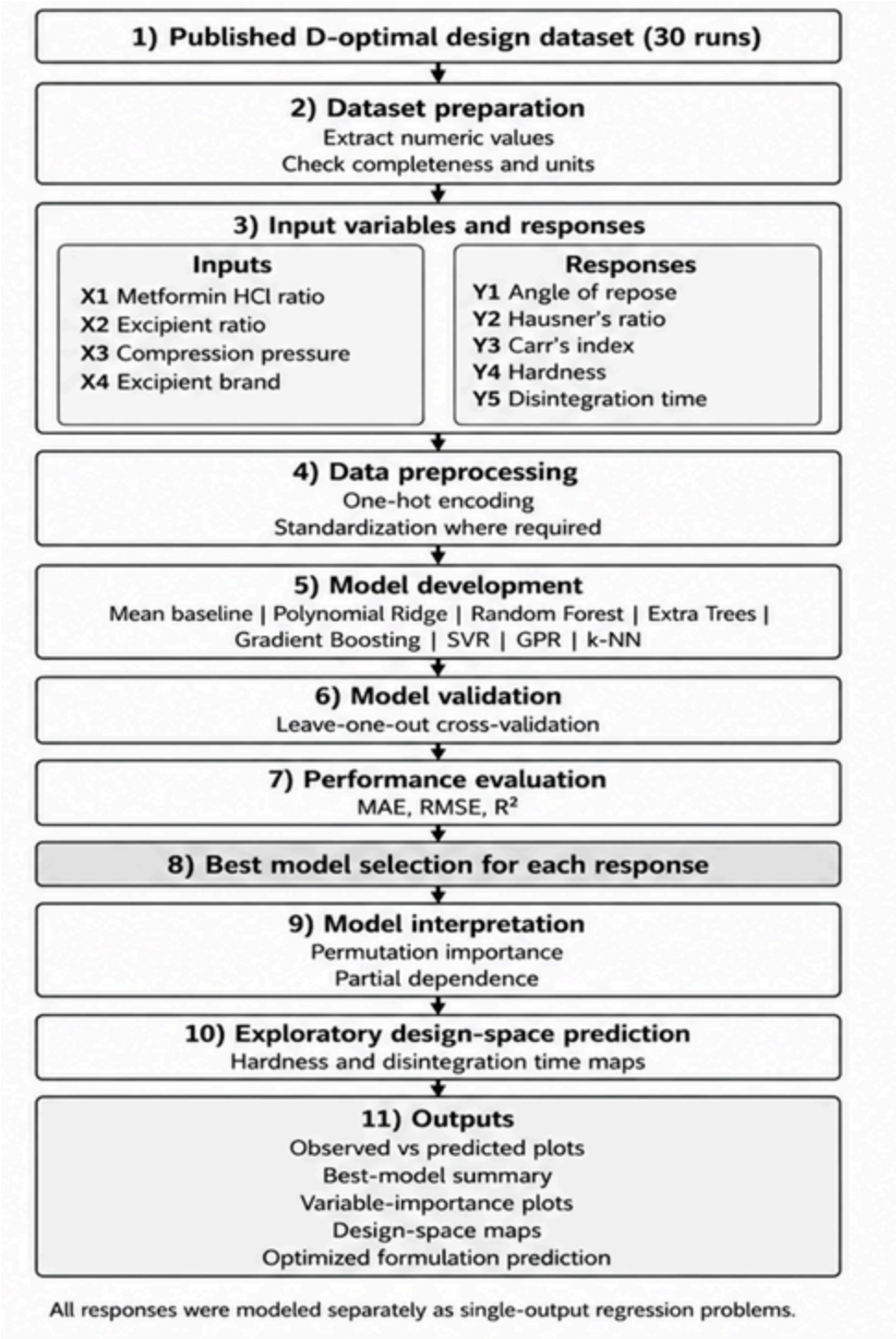
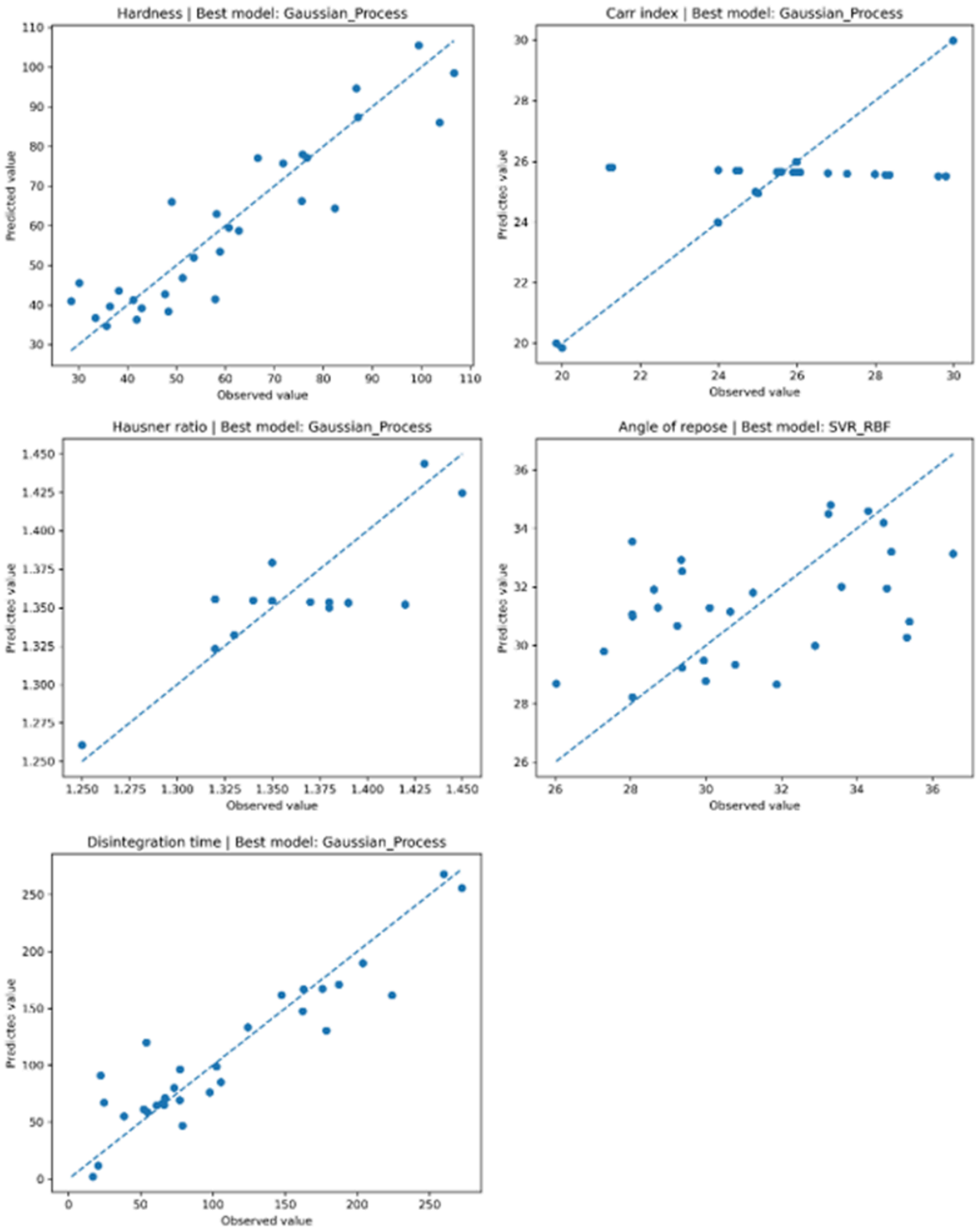


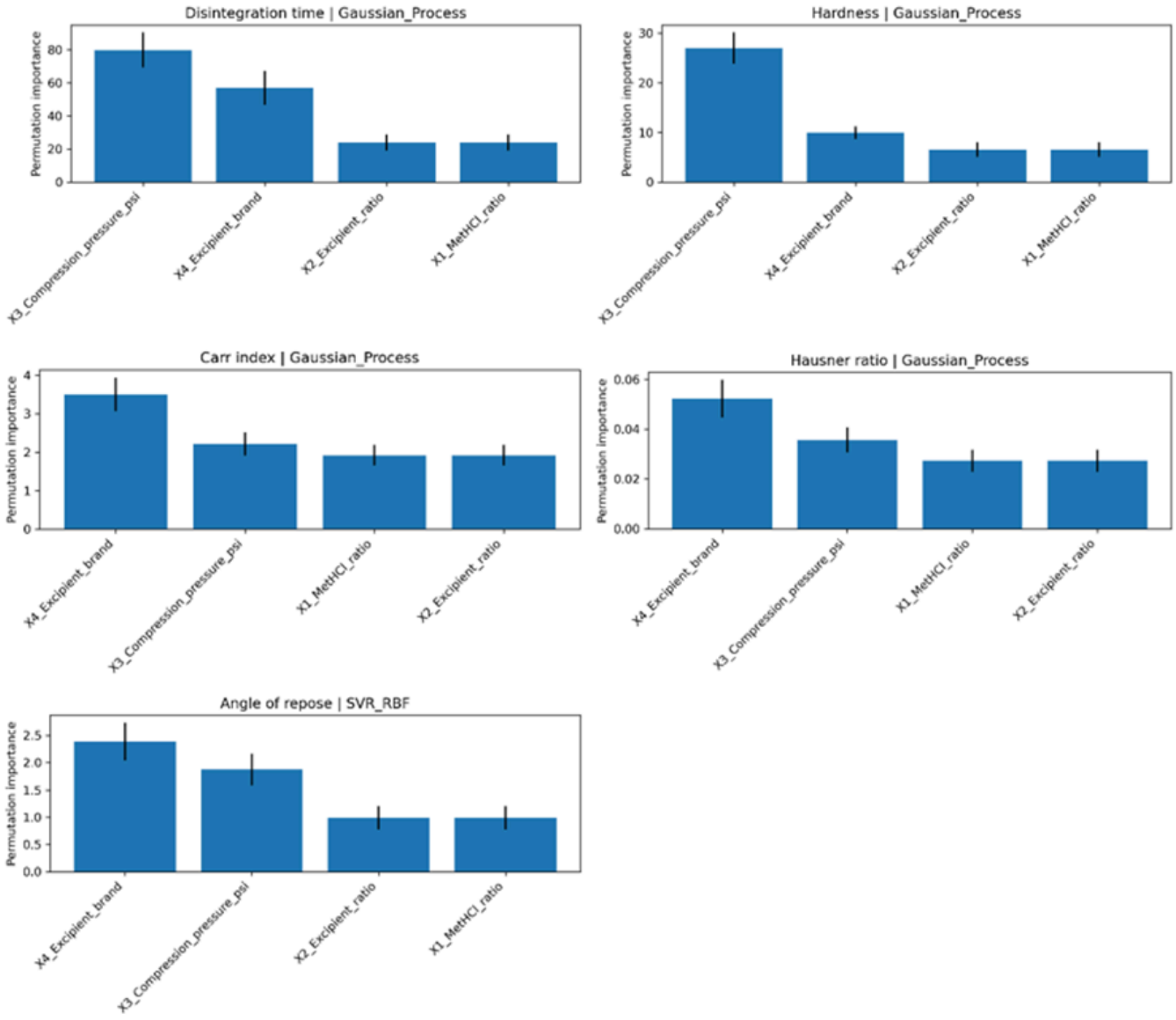

Figure 1

**Workflow of machine-learning modeling of D-optimal design data for metformin hydrochloride orally disintegrating tablets.**



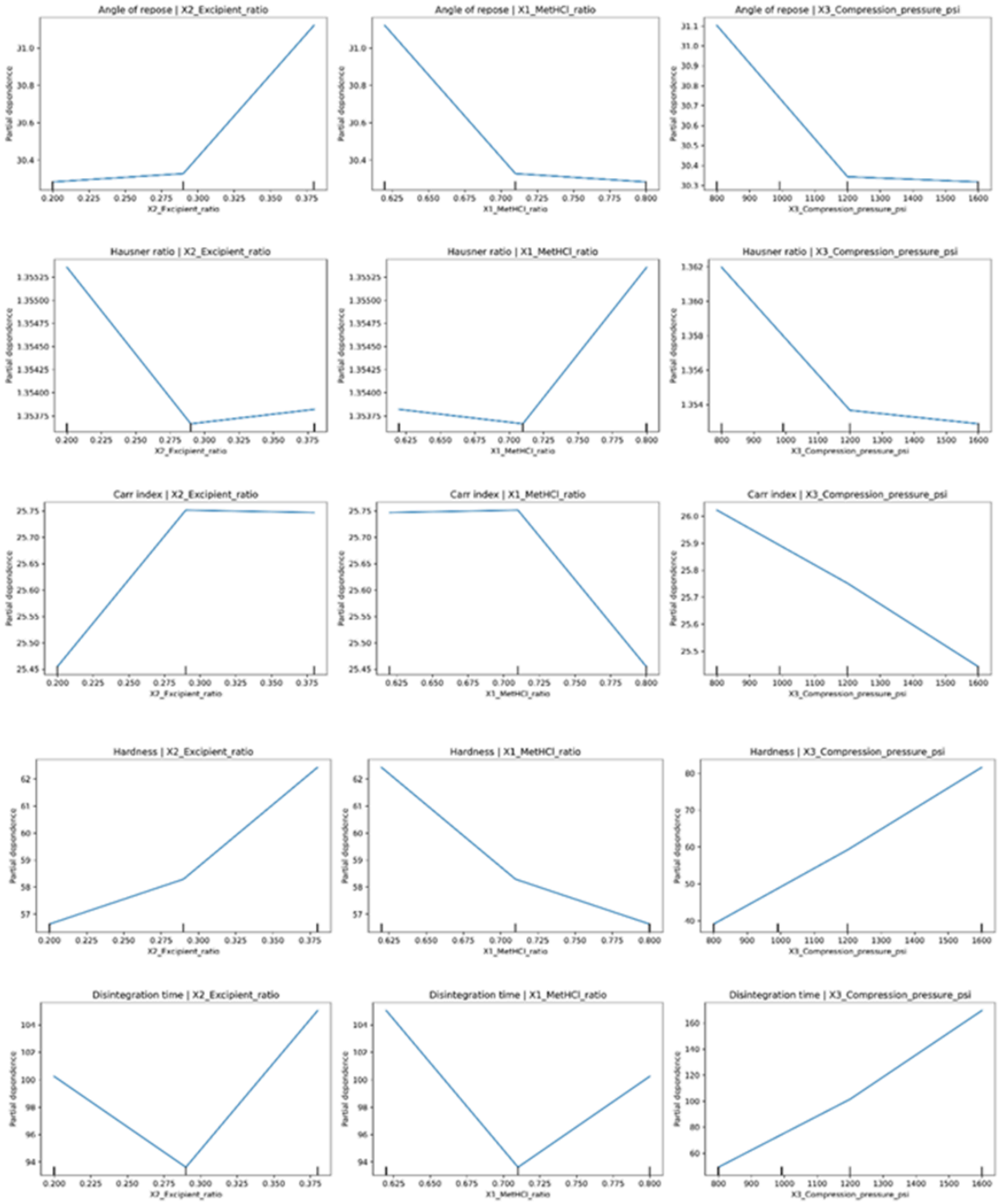

## Figure 2

Observed-versus-predicted plots for the best-performing models of each response.



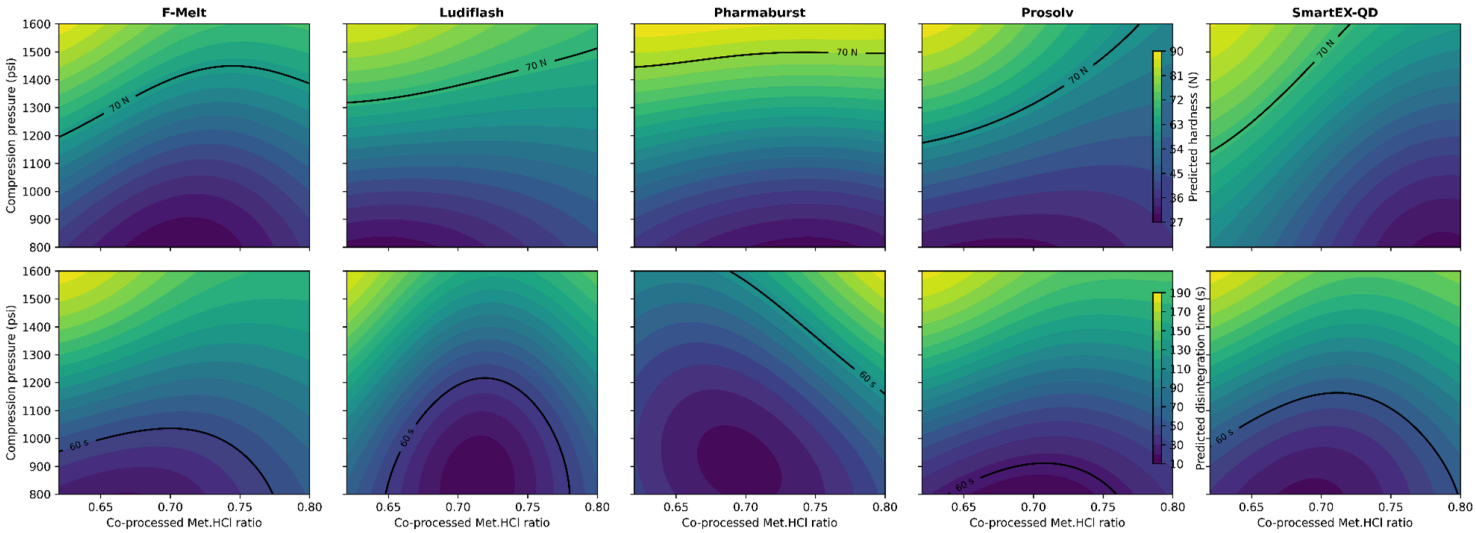

## Figure 3

Permutation importance of input variables for the best-performing models.




## Figure 4

Partial dependence plots of numeric variables for the best-performing models.




## Figure 5

Exploratory design-space prediction maps for hardness and disintegration time.