The applications of machine learning to predict the forming of chemically stable amorphous solid dispersions prepared by hot-melt extrusion

Junhuang Jiang, Anqi Lu, Xiangyu Ma, Defang Ouyang, Robert O. Williams

PII: S2590-1567(23)00008-7

DOI: https://doi.org/10.1016/j.ijpx.2023.100164

Reference: IJPX 100164

To appear in:

Received date: 16 January 2023

Accepted date: 17 January 2023

Please cite this article as: J. Jiang, A. Lu, X. Ma, et al., The applications of machine learning to predict the forming of chemically stable amorphous solid dispersions prepared by hot-melt extrusion, (2023), https://doi.org/10.1016/j.ijpx.2023.100164

The Applications of Machine Learning to Predict the Forming of Chemically Stable Amorphous Solid Dispersions Prepared by Hot-Melt Extrusion

Junhuang Jiang [a], Anqi Lu [a], Xiangyu Ma [b], Defang Ouyang [c] and Robert O. Williams III [a, *]

[a] Division of Molecular Pharmaceutics and Drug Delivery, College of Pharmacy, The University of Texas at Austin, Austin, TX, 78712, USA

[b] Global Investment Research, Goldman Sachs, NY, USA

[c] State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences (ICMS), University of Macau, Macau, China

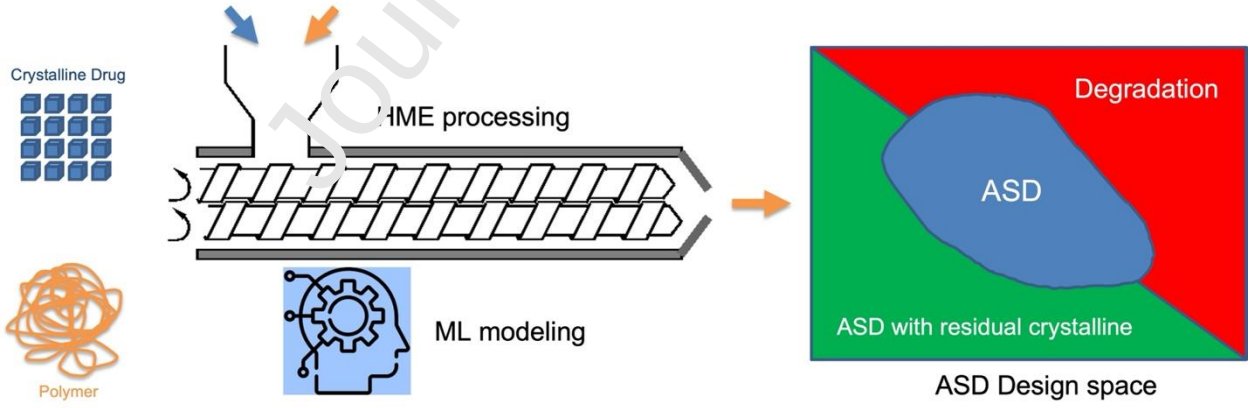
[*]Corresponding author.

E-mail address: bill.williams@austin.utexas.edu (R. O. Williams III).

Abstract

Amorphous solid dispersion (ASD) is one of the most important strategies to improve the solubility and dissolution rate of poorly water-soluble drugs. As a widely used technique to prepare ASDs, hot-melt extrusion (HME) provides various benefits, including a solvent-free process, continuous manufacturing, and efficient mixing compared to solvent-based methods, such as spray drying. Energy input, consisting of thermal and specific mechanical energy, should be carefully controlled during the HME process to prevent chemical degradation and residual crystallinity. However, a conventional ASD development process uses a trial-and-error approach, which is laborious and time-consuming. In this study, we have successfully built multiple machine learning (ML) models to predict the amorphization of crystalline drug formulations and

the chemical stability of subsequent ASDs prepared by the HME process. We utilized 760 formulations containing 49 active pharmaceutical ingredients (APIs) and 38 excipients. By evaluating the built ML models, we found that ECFP-LightGBM was the best model to predict amorphization with an accuracy of 92.8%. Furthermore, ECFP-XGBoost was the best in estimating chemical stability with an accuracy of 96.0%. In addition, the feature importance analyses based on SHapley Additive exPlanations (SHAP) and information gain (IG) revealed that several processing parameters and material attributes (i.e., drug loading, polymer ratio, drug's Extended-connectivity fingerprints (ECFP) fingerprints and polymer's properties) are critical for achieving accurate predictions for the selected models. Moreover, important API's substructures related to amorphization and chemical stability were determined, and the results are largely consistent with the literature. In conclusion, we established the ML models to predict formation of chemically stable ASDs and identify the critical attributes during HME processing. Importantly, the developed ML methodology has the potential to facilitate the product development of ASDs manufactured by HME with a much reduced human workload.



Graphical abstract

Keywords: Amorphous Solid Dispersion; Artificial Intelligence; Machine Learning; Hot-Melt Extrusion

Abbreviations

ASD, Amorphous solid dispersion; HME, Hot-melt extrusion; AI, Artificial intelligence; ML, Machine learning; RF, Random Forest; SVM, Support vector machine; SHAP, SHapley Additive exPlanations; ECFP; Extended-connectivity fingerprints; IG, Information gain

1. Introduction

Poor aqueous solubility is a common issue for many drugs at different stages, including pipeline candidates in development and commercial products, leading to lower bioavailability. According to the biopharmaceutical classification system (BCS), approximately 40% of the marketed products and 90% of drugs in development can be classified as poorly water-soluble (Jermain et al., 2018). Prior research has demonstrated that ASDs can effectively improve the solubility of poorly water-soluble drugs and subsequently improve their bioavailability (Pandi et al., 2020; Schittny et al., 2019). ASDs are solid dispersions in which the amorphous drug is dispersed in an excipient matrix such as polymers (Chiou et al., n.d.). The purpose of forming an ASD is to minimize this energy component by disrupting the drug crystal lattice (Jermain et al., 2018). By breaking the crystal lattice and hindering the lattice formation, the crystalline drug converts into an amorphous state, resulting in higher chemical potential, improved solubility, and bioavailability (Alonzo et al., 2010). HME is one of the widely used techniques to prepare ASDs with several benefits, such as a continuous manufacturing process, efficient and highly automated, and solvent-free compared with other conventional techniques, including spray drying and antisolvent precipitation (Huang and Williams, 2018). However, thermal/chemical degradation of drug substance during HME processing is one of its most important limitations and should be carefully considered during formulation development (Lu et al., 2014). It has been

reported that the melting point depression approach through intermolecular interaction by forming such as co-crystal and salt can effectively reduce the HME processing temperature, avoiding thermal degradation (Haser et al., 2018a; Liu et al., 2012). Wang et al. successfully developed an integrated ML-based platform, PharmDE, to estimate the compatibility between drugs and excipients for the preformulation evaluation of solid dispersions (Wang et al., 2021). This system immediately identifies the potential chemical degradation of a solid dispersion but does not include information on processing instruments such as HME and spray-drying (Wang et al., 2021). On the other side, sufficient energy (i.e., thermal energy and specific mechanical energy) is required during the HME process to fully convert crystalline API into amorphous and prevent the risk of residual crystallinity (Ma et al., 2019; Moseson and Taylor, 2018). Therefore, it's important to identify an optimal design space for forming an ASD using the HME process to prevent both chemical degradation and residual crystallinity.

Machine learning (ML) as a cutting-edge technique has gained more interest in the pharmaceutical industry, especially in drug formulation development. For example, ML models have successfully been applied to predict the aerosol performance of dry powder for inhalation (Jiang et al., 2022), detect tablet defects in XRCT images (Ma et al., 2020), and predict dissolution and storage stability of solid dispersions (Dong et al., 2021). In addition, multiple ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), XGBoost, LightGBM, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) have been widely used for different tasks during formulation development (Han et al., 2019; Jiang et al., 2022). RF is an ensemble algorithm consisting of multiple decision trees, which exhibits improved predictive performance and reduced over-fitting (Breiman, 2001). SVM is a linear classifier based on the margin maximization principle, and the hyperplane can optimally separate

the data into two or multiple categories (Adankon and Cheriet, 2009). XGBoost and LightGBM are two gradient-boosting algorithms developed recently and have demonstrated higher efficiency, flexibility, and portability (Chen et al., 2016; Ke et al., n.d.). As tree-based algorithms, RF and XGBoost performed well in predicting a solid dispersion's physical stability and dissolution rate (Dong et al., 2021). As a widely used traditional ML model, SVM has also demonstrated exemplary performance in determining the glass forming ability (GFA) of pharmaceutical compounds (Alhalaweh et al., 2014). Therefore, multiple supervised classification ML models, including RF, SVM, XGBoost, and LightGBM, will be implemented in this study.

Conventional ASD development involves several processes, including preformulation studies, processing optimization, and characterization. In addition, comprehensive understandings of APIs' and polymers' physiochemical properties are necessary, including glass forming ability, melt viscosity, miscibility, glass transition temperature, degradation temperature, and melting point, before conducting HME experiments. Therefore, a conventional ASD development approach requires many trial-and-error experiments, which are time-consuming and highly laborious. ML provides opportunities to design ASD by potentially reducing the human workload of conventional approaches. For example, Han et al. successfully applied multiple machine learning techniques to predict the physical stability over time from 646 formulations (Han et al., 2019). In this study, the random forest was identified as the best model, with the highest accuracy of 82.5%. In addition, experimental validation using estradiol-polyvinylpyrrolidone formulations and molecular modeling techniques was performed to further evaluate the model (Han et al., 2019). Moreover, Lee et al. successfully predicted the physical stability of amorphous solid dispersion using a deep neural network (Lee et al., 2022). The

researchers first applied a hybrid data sampling method and principal component analysis (PCA) to process the initial dataset, then the processed data was fed into a deep neural network for modeling (Lee et al., 2022). However, most of the published literature regarding ML applications in ASD focuses on storage stability, and limited research focuses on the HME process and the forming of an ASD. Therefore, we hypothesize ML models can accurately predict the forming of chemically stable ASDs and identify critical attributes during the HME process to reduce thermal degradation and residual crystallinity.

2. Methods and Materials

2.1. Data collection

In this study, we first conducted data mining using PubMed with the terms "hot melt-extrusion" and "amorphous solid dispersion," and the articles ranging between January 1, 2012, and January 31, 2022. We obtained a dataset containing 49 APIs and 38 excipients from 158 selected publications by literature mining. Pie charts displayed a brief description of the dataset in Fig. 1. Input variables consist of critical material attributes (CMAs) (i.e., APIs, excipients, drug loading (w/w), and excipient ratio (w/w)) and critical processing parameters (CPPs) (i.e., hot melt-extruder configuration, barrel temperature, screw speed, and feed rate). Histograms showed the distributions of different CPPs in Fig. 2. Amorphization and chemical stability were treated as two separate outputs for ML modeling. The amorphization of HME formulations was determined by the solid-state characterization results such as differential scanning calorimetry (DSC), X-ray powder diffraction (XRPD), and polarized light microscopy (PLM) described in the articles. The chemical stability of the HME formulations was characterized by high-performance liquid chromatography (HPLC), and 95% of drug content was set as a threshold for data classification (i.e., chemical stable: drug content > 95%; chemical unstable: drug content ≤ 95%). For

amorphization, amorphous formulations were labeled as "1," and crystalline formulations were as "0". In addition, chemically stable and unstable formulations were labeled as "1" and "0", respectively. After reorganizing the dataset, we obtained 760 and 495 formulation data points for amorphization and chemical stability modeling tasks, respectively. By reviewing the collected dataset, we observed that it is imbalanced concerning the portions of the targets' categories. Specifically, 16.3% (124) and 83.7% (636) formulations were determined as "crystalline" and "amorphous" among all amorphization data points, respectively. Moreover, 17.6% (87) and 82.4% (408) formulations were defined as "chemically unstable" and "chemically stable" in the chemical stability datasets. Therefore, data processing techniques such as class weights, upsampling, downsampling, and evaluation metrics, including F1 score and receiver operating characteristic and accuracy, must be considered for modeling imbalanced data.

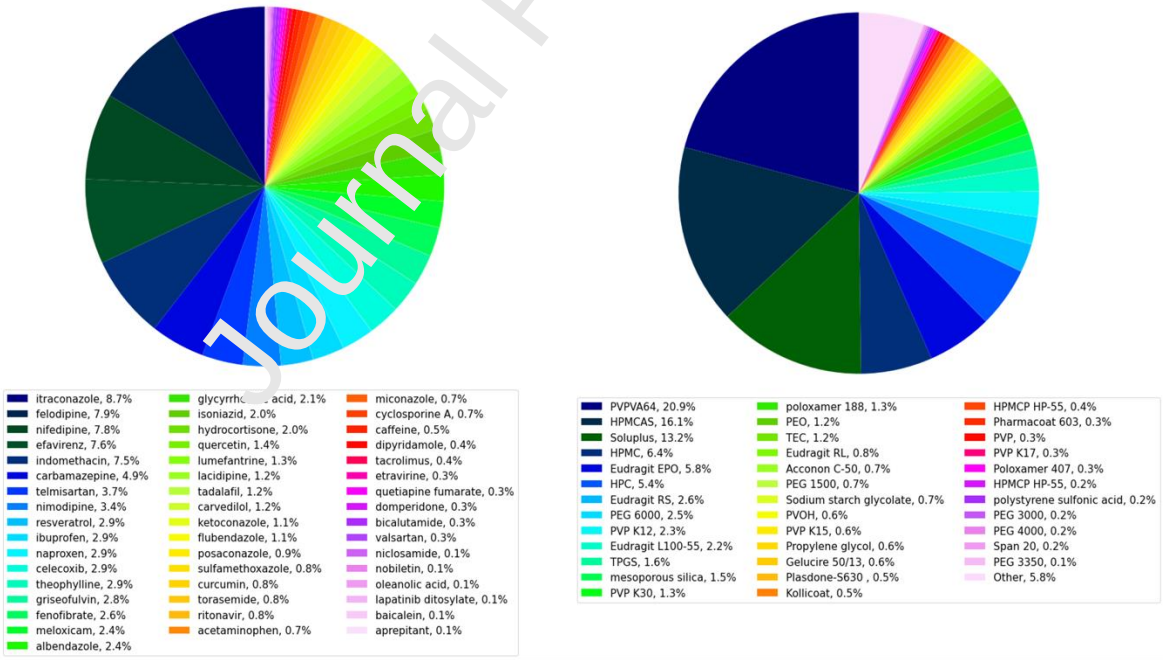


Fig. 1. A brief description of the APIs and excipients used in this study. The pie charts show the categories and the corresponding proportions of APIs (Left) and excipients (Right) used for ML modeling. According to the data exploratory analysis by the pie charts, itraconazole, felodipine, and nifedipine are three of the most widely used drugs in the dataset collected from the published literature with portions of 8.7%, 7.9%, and 7.8%, respectively. For excipients, polyvinylpyrrolidone (PVPVA64), hypromellose

acetate succinate (HPMCAS), and polyvinyl caprolactam-polyvinyl acetate-polyethylene glycol graft copolymer (Soluplus) were the top 3 most popular excipients in the obtained dataset.

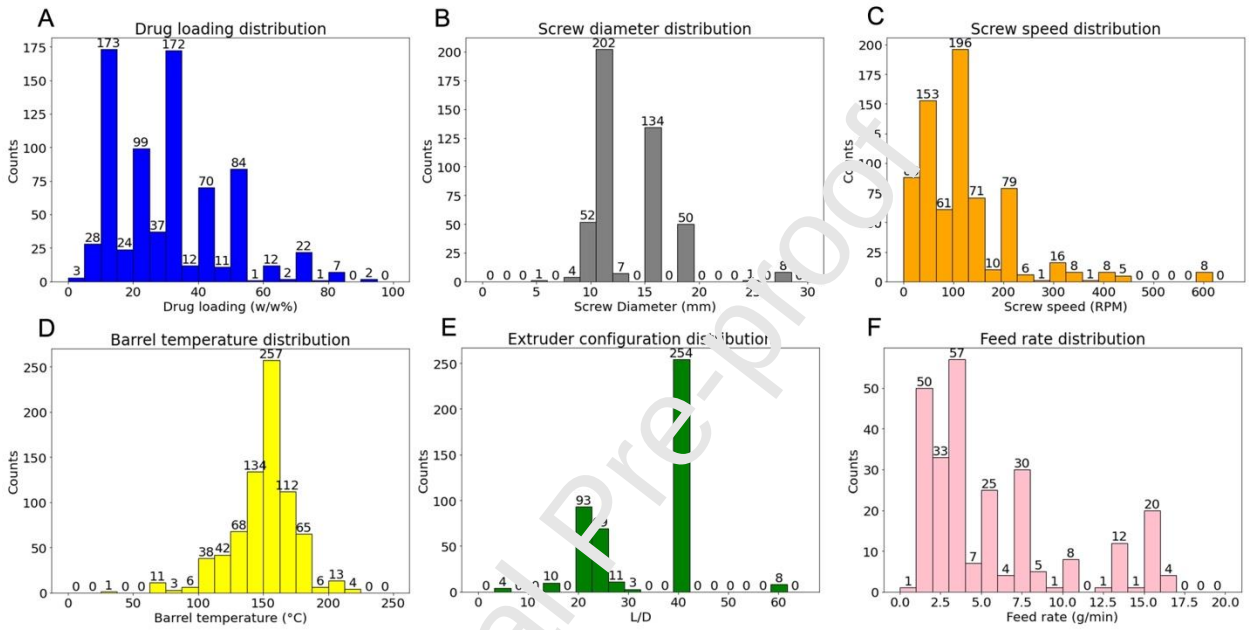


Fig. 2. Data distribution of 960 HME formulations. Initial dataset distribution of drug loading (w/w %) (A), screw diameter (mm) (B), screw speed (RPM) (C), barrel temperature (°C) (D), extruder configuration (length/diameter ratio (L/D)) (E), and feed rate (g/min) (F). The histogram plots provide a visual description of the distributions for input variables.

2.2. Data processing

Train-test split is a data processing method to evaluate the model performance. The training subset of the data will be used to fit the model and feature learning, and a testing subset is the other portion to evaluate the model unbiasedly. An 80%: 20% ratio is a widely used train-test split method for ML modeling (Dong et al., 2021). Therefore, for amorphization, all HME formulation data points were split into the training subset (608) and testing subset (152) by 80%: 20%. In addition, all chemical stability-related data points were split into the training subset (396)

and testing subset (99) by 80%: 20%. In addition, five-fold cross-validation was performed to ensure the model's generalization ability and prevent over-fitting.

After the data visualization, we observed that some input variables, mostly extruder configuration and processing parameters, are missing (Fig. 3). This is because researchers: (1) didn't mention the specific configurations of the instruments and (2) didn't mention or use the feeder for HME process in the publications. In addition, those parameters, such as feed rate and extruder configuration, are not symmetrically distributed. In this situation, the mean substitution of the missing values is not appropriate for processing the dataset. Therefore, to make the best use of the dataset and obtain a robust trained model, we fixed the missing values by median substitution instead of removing the whole data point.

To convert API and excipient molecules into computer-readable formats, multiple molecular representation methods, including extended-connectivity fingerprints (ECFP) and 2D molecular descriptors, were applied (Jiang et al., 2022). Molecular descriptors contained physical and chemical properties of the APIs and excipients were generated by computer (Raghunathan and Priyakumar, 2022). Some published literature compared the model's predictive performance using different molecular representation methods and found that ECFP-based models outperformed 2D molecular descriptors-based ones (Dong et al., 2021). Therefore, we will use both methods for just the representation of API since some formulations contain multiple excipients, and the dataset's dimension will be extremely high if using ECFP for all excipients. Specifically, excipients were represented by 208 2D molecular descriptors in RDKit, and APIs were described by either 208 2D molecular descriptors or ECFP fingerprints with a length of 1024 and radius of 3 in RDKit (version: RDDkt 2020.09.10) ("RDKit," n.d.). Finally, a dataset containing formulation compositions (i.e., drug loading and excipient ratio), drug and excipient

properties by molecular descriptors, processing parameters (i.e., barrel temperature, screw speed, and feed rate), and extruder configuration (i.e., screw diameter and L/D) were obtained for ML modeling.

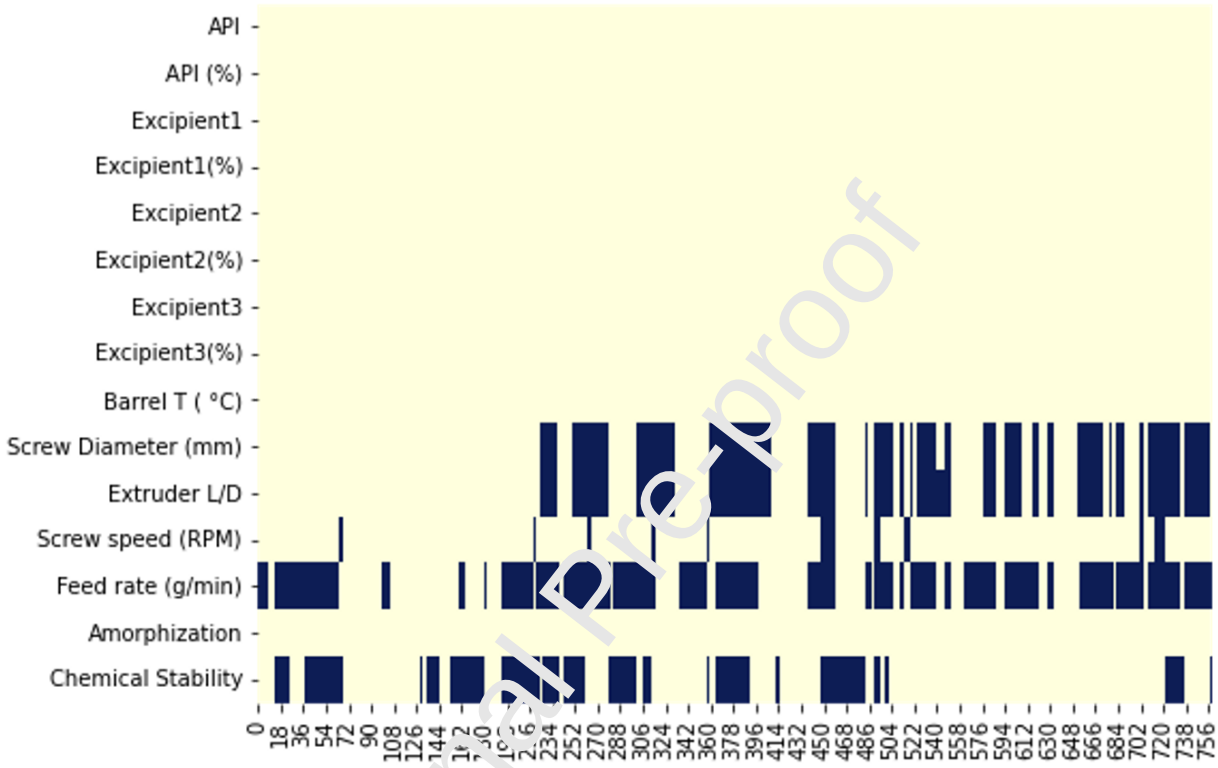


Fig. 3. Visualization of the missing values of the dataset containing 760 formulations. Background (light-yellow) indicates information was available in the literature, whereas dark-blue bars indicate the data was missing. It shows that most missing values in the input variables were the extruder's configuration, including screw diameter (mm), extruder L/D, and processing parameters consisting of screw speed (RPM) and feed rate (g/min). For the targets, all literature references contain amorphization information, whereas 495 formulations contain information on chemical stability. This graph provides an overview of the dataset regarding the missing values for specific variables, and the missing values of input variables will be fixed before ML modeling.

2.3. Machine learning algorithms

Multiple ML algorithms, including XGBoost, LightGBM, RF, and SVM, were applied to predict amorphization and chemical stability separately. Two sets of ML models were trained respectively for the two outputs because (1) the chemical stability dataset (760) has fewer observations than the amorphization dataset (495), and (2) the same ML algorithms may perform

differently in another dataset. In addition, two different molecular representation methods, namely 2D-descriptors (2D) and ECFP were used for drug molecules, resulting in 8 different types of corresponding ML models for each target. The ML models were finely tuned by adjusting the hyperparameters of different algorithms. The hyperparameters of the models were tuned in by both grid search and random search. The hyperparameters of XGBoost (subsample, minimum child weight, maximum depth, learning rate, gamma, colsample_bytree, and colsample_bylevel), LightGBM (learning rate, minimum child weight, number of estimators, and number of leaves), RF (bootstrap, maximum depth, maximum features, minimum samples leaf, minimum samples split, and the number of estimators), and SVM (the kernel function, the penalty parameter C, and the γ) are listed in Table 1. To solve the imbalanced data issue, the class weight of different categories was applied in all ML models. The implementation of the ML models was conducted by Scikit-Learn 1.0.1, and Python 3.9.7 as the programming language. Pandas 1.3.4 and NumPy 1.23.0 were open-source packages to process tabular data in ML. Matplotlib 3.5.0 and seaborn 0.11.2 were used as plotting libraries to visualize the dataset and the modeling results.

Table 1. ML model hyperparameter configurations. The prefix of each ML model is the molecular representation method of API. For example, 2D-XGBoost is the ML model using 2D-descriptors for API's molecular representation as input, whereas ECFP-XGBoost is the ML model using ECFP for API's molecular representation as input.

ML model hyperparameter configurations		
ML algorithms	**Amorphization model**	**Chemical stability model**
2D-XGBoost	0.6; 1; 4; 0.25; 0.5; 0.8; 0.3	0.8; 1; 5; 0.25; 0; 0.8; 1
2D-LightGBM	0.1; 6; 1000; 30	1; 1; 1000; 30
2D-RF	True; 10; auto; 1; 2; 50	True; 30; sqrt; 2; 2; 50
2D-SVM	rbf; 100; 0.01	rbf; 1000; 0.01
ECFP-XGBoost	0.6; 1; 3; 0.5; 1; 0.8; 0.5	1; 1; 4; 0.1; 0; 0.1; 1
ECFP-LightGBM	0.01; 1; 1000; 20	0.25; 1; 20; 20
ECFP -RF	True; 30; sqrt; 1; 2; 200	True; 30; sqrt; 2; 2; 50
ECFP -SVM	rbf; 1000; 0.1	rbf; 1000; 0.01

";" Separates different hyperparameters

2.4. Model evaluation metrics and interpretability

To evaluate the model's predictive performance properly, multiple metrics, namely accuracy (ACC), F1 score (F1), and receiver operating characteristic (ROC) area under the curve (AUC), were used in this study. ACC and F1 were defined by the following Equations (1), (2):

Equation 1

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Equation 2

$$F1 = \frac{2 \times precision \times recall}{precisioin + recall}$$

ACC and F1 were calculated from the confusion matrix, where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. The ROC curve shows the sensitivity and specificity results at every possible threshold.

Furthermore, SHapley Additive exPlanations (SHAP) and information gain (IG) were employed to further evaluate the model's interpretability and feature importance for the best prediction models, respectively. Interpretability is the process by which humans understand the ML model's decision instead of treating it as a "black box", and it contains local and global interpretability (Kopitar et al., 2019). Local interpretability refers to the explanation of each individual prediction, while global interpretability provides insights into the model in the whole dataset (Kopitar et al., 2019). In addition, feature importance analysis was conducted by IG, which is a method that measures the reduction in entropy by transforming a dataset (Ke et al., n.d.). Therefore, we will use SHAP for a local explanation of the ML models and IG for global feature importance analysis. In addition, it's critical to identify the important structural features indicated by ECFP fingerprints that relate to the model's output. Previous studies have demonstrated that the important substructures associated with the corresponding models were extracted by

implementing IG feature importance analysis. For example, ML has successfully been applied to predict lipid nanoparticle (LNP)-based mRNA vaccine, and the important ionized lipid's substructures were extracted based on IG analysis (Wang et al., 2022). Furthermore, the substructures of compounds and solvents were studied from the ML solubility prediction model (Ye and Ouyang, 2021). Therefore, the top 12 most important substructures of API calculated concerning IG values will be analyzed.

3. Model performance

3.1. Machine learning modeling results

3.1.1. Amorphization prediction results

Multiple ML models were successfully applied to predict the amorphization of the crystalline drug during the HME process. The summary of amorphization model prediction results is shown in Table 2. Overall, all ML models except for 2D-SVM performed well in the training subsets with evaluation metrics values higher than 0.95. In the cross-validation sets, 2D-XGBoost, 2D-LightGBM, ECFP-XGBoost, and ECFP-LightGBM showed similar prediction performance with ACC, F1, and AUC values of higher than 0.91, approximate 0.95, and approximate 0.92, respectively. 2D-SVM showed relatively poor performance with the lowest ACC, F1, and AUC of 0.895, 0.938, and 0.824, respectively. In the testing subsets, we found that two LightGBM-based ML models outperformed others. Specifically, ECFP-LightGBM performed best among all ML models, with the highest ACC, F1, and AUC of 0.928, 0.958, and 0.932, respectively. 2D-ECFP also predicted well with the highest ACC and F1 but a relatively lower AUC value of 0.895 compared to ECFP-LightGBM. AUC is a more representative metric for evaluating the model, especially for an imbalanced dataset in this study. Therefore, by analyzing and comparing the results of different ML models, ECFP-LightGBM performed well in both cross-validation

and testing subsets with good generalization ability and lower variance when feeding in new formulation data. Therefore, ECFP-LightGBM was selected as the best amorphization prediction model and will be further studied.

Table 2. ML prediction model performance of amorphization. The evaluation metrics (i.e., accuracy (ACC), F1-score (F1), and receiver operating characteristic (ROC) area under the curve (AUC)) were calculated based on the confusion matrix results implemented in Scikit-Learn. This table describes the prediction performance of eight ML models in training, cross-validation, and testing subsets, respectively.

	Training set			**Cross-validation set**			**Testing set**		
ML Algorithms	**ACC**	**F1**	**AUC**	**ACC**	**F1**	**AUC**	**ACC**	**F1**	**AUC**
2D-XGBoost	0.985	0.991	0.983	0.916	0.951	0.926	0.914	0.950	0.879
2D-LightGBM	0.984	0.990	0.982	0.911	0.947	0.927	0.928	0.958	0.895
2D-RF	0.979	0.987	0.948	0.906	0.944	0.911	0.908	0.945	0.839
2D-SVM	0.934	0.962	0.922	0.895	0.938	0.884	0.875	0.928	0.792
ECFP-XGBoost	0.982	0.989	0.973	0.914	0.950	0.916	0.914	0.951	0.897
ECFP-LightGBM	**0.990**	**0.994**	**0.986**	**0.913**	**0.949**	**0.926**	**0.928**	**0.958**	**0.932**
ECFP-RF	0.992	0.995	0.976	0.895	0.938	0.899	0.914	0.950	0.879
ECFP-SVM	0.987	0.992	0.976	0.906	0.944	0.923	0.914	0.950	0.866

3.1.2. Degradation prediction results

Eight chemical stability ML models were successfully trained, and the prediction results are shown in Table 3. All ML models showed excellent prediction performance in the training subsets with metrics values higher than 0.950. In the cross-validation subsets, most ML models performed well with ACC, F1, and AUC values higher than 0.950. However, both RF-based models showed relatively poor performance, with an ACC of 0.917 in the cross-validation set. In addition, evaluating the results in testing is important because it reflects the model's prediction variance when inputting new HME formulations. According to the ML model performance

summary in the testing subsets, ECFP-XGBoost exhibited the highest ACC, F1, and AUC of 0.960, 0.976, and 0.944, respectively. 2D-XGBoost also performed relatively well, with all metrics results higher than 0.90. SVM and LightGBM models exhibited good ACC and F1 values higher than 0.90 but relatively lower AUC ranging from 0.828 to 0.882. Interestingly, 2D-RF and ECFP-RF showed relatively lower performance among all ML models in the testing set. Therefore, ECFP-XGBoost was chosen for further analysis because it offered excellent prediction performance in cross-validation and testing subsets.

Table 3. ML prediction model performance of chemical stability. The evaluation metrics (i.e., accuracy (ACC), F1-score (F1), and receiver operating characteristic (ROC) area under the curve (AUC)) were calculated based on the confusion matrix results implemented in Scikit-Learn. This table describes the prediction performance of eight ML models in training, cross-validation, and testing subsets, respectively.

	Training set			Cross-validation set			Testing set		
ML Algorithms	**ACC**	**F1**	**AUC**	**ACC**	**F1**	**AUC**	**ACC**	**F1**	**AUC**
2D-XGBoost	0.997	0.998	0.993	0.962	0.977	0.974	0.949	0.970	0.911
2D-LightGBM	0.997	0.998	0.993	0.950	0.969	0.970	0.909	0.947	0.828
2D-RF	0.975	0.984	0.940	0.917	0.947	0.952	0.838	0.899	0.720
2D-SVM	0.995	0.997	0.992	0.956	0.976	0.961	0.939	0.964	0.882
ECFP-XGBoost	**0.992**	**0.995**	**0.990**	**0.952**	**0.972**	**0.965**	**0.960**	**0.976**	**0.944**
ECFP-LightGBM	0.995	0.997	0.992	0.965	0.979	0.983	0.929	0.958	0.857
ECFP-RF	0.970	0.981	0.934	0.917	0.939	0.954	0.869	0.919	0.753
ECFP-SVM	0.995	0.997	0.992	0.962	0.977	0.966	0.929	0.958	0.857

3.2. Feature importance analysis

3.2.1. SHAP analysis results

ECFP-LightGBM and ECFP-XGBoost were selected as the best models to predict amorphization and chemical stability during the HME process. SHAP was employed to investigate the model interpretability for these two models. The SHAP summary plot for ECFP-LightGBM is shown below in Fig. 4.

The SHAP results displayed the top 20 most important features contributing to the model prediction. These features contain critical processing parameters such as barrel temperature, feed rate, screw speed, and screw diameter and critical material attributes, including the first dominant excipient's loading, drug loading, excipient's properties represented by 2D descriptors, and drug's substructures which are indicated by ECFP. More importantly, the summary plot also provides an overview of the correlation between each feature and amorphization. For example, barrel temperature, first dominant excipient's loading, and screw speed exhibited positive correlations with the model output, while drug loading, some drug's substructures (e.g., ECFP_506, ECFP_121, and ECFP_361), and screw diameter were negatively correlated with the output.

The SHAP summary plot for the ECFP-XGBoost chemical stability prediction model is displayed in Fig. 5. The top 20 most important features include critical processing parameters, drug's properties, and excipient's properties. Among all features, barrel temperature was the most critical, and it negatively affected the chemical stability of the drug during the HME process. In addition, the screw diameter of the extruder showed a negative correlation with the model output. Drug substructures such as ECFP_875, ECFP_974, ECFP_440, and ECFP_721 are critical for the model to make the decision and will be discussed in the later session. The increased excipient one's ratio will contribute to the drug's chemical degradation.

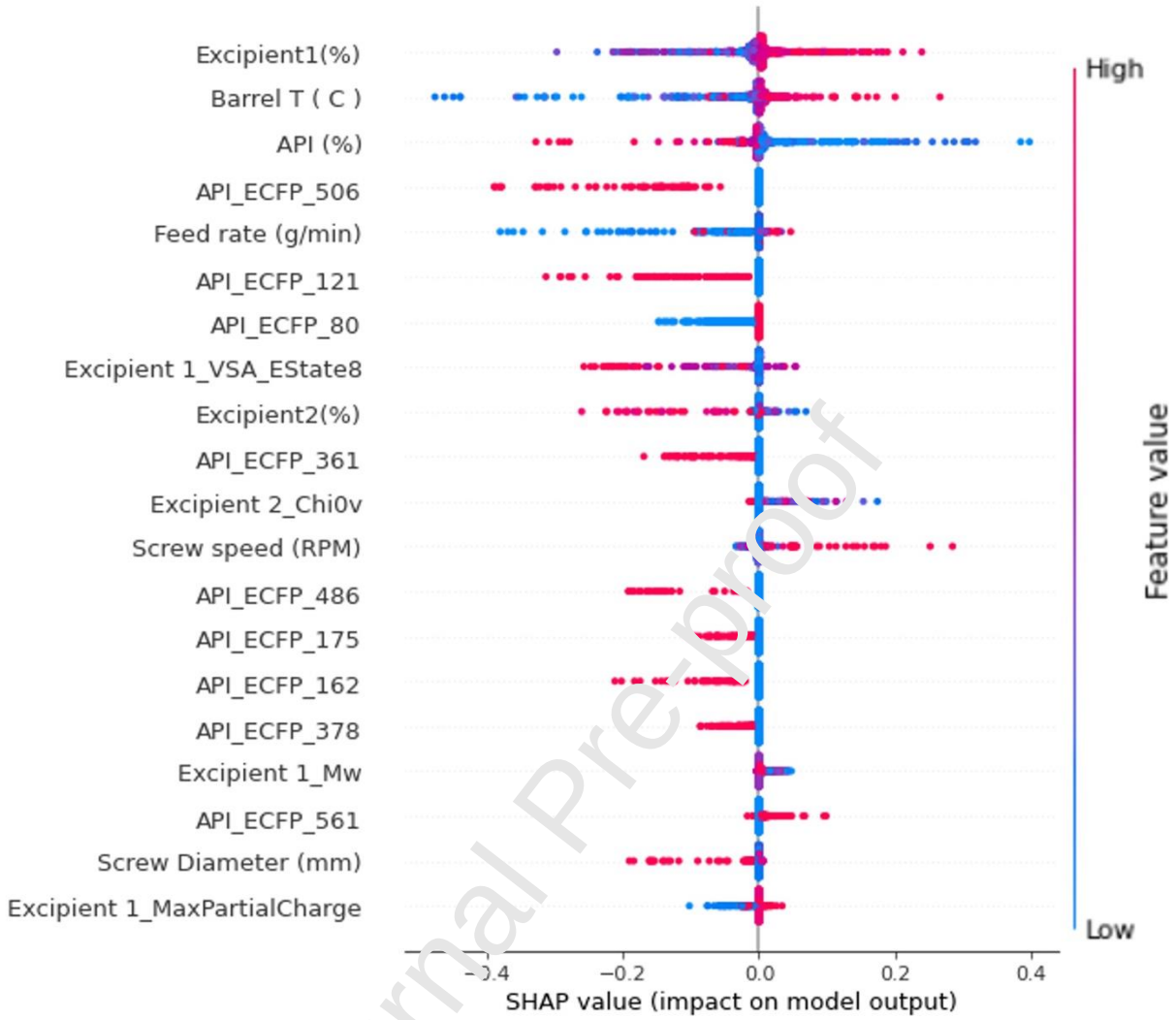


Fig. 4. SHAP summary plot on top 20 features for ECFP-LightGBM. It sorts all features by the sum of SHAP values and shows the impact distribution on the model output. Excipient1(%), first dominant excipient's loading; Barrel T(C), barrel temperature; API (%), drug loading; API_ECFP_506, drug's substructure at ECFP 506; Feed rate (g/min); API_ECFP_121; API_ECFP_80; Excipient 1_VSA_Estate8, first dominant excipient's MOE VSA descriptor; Excipient2(%), second dominant excipient's loading; API_ECFP_361; Excipient 2_Chi0v, second dominant excipient's topological descriptor; Screw speed (RPM); API_ECFP_486; API_ECFP_175; API_ECFP_162; API_ECFP_378; Excipient 1_Mw; first dominant excipient's molecular weight; API_ECFP_561; Screw Diameter (mm); Excipient 1_MaxPartialCharge, first dominant excipient's maximum partial charge. The color bar depicts the value of each feature; blue indicates a higher value, while red indicates a lower value. The higher the SHAP values, the more probability of generating an ASD.

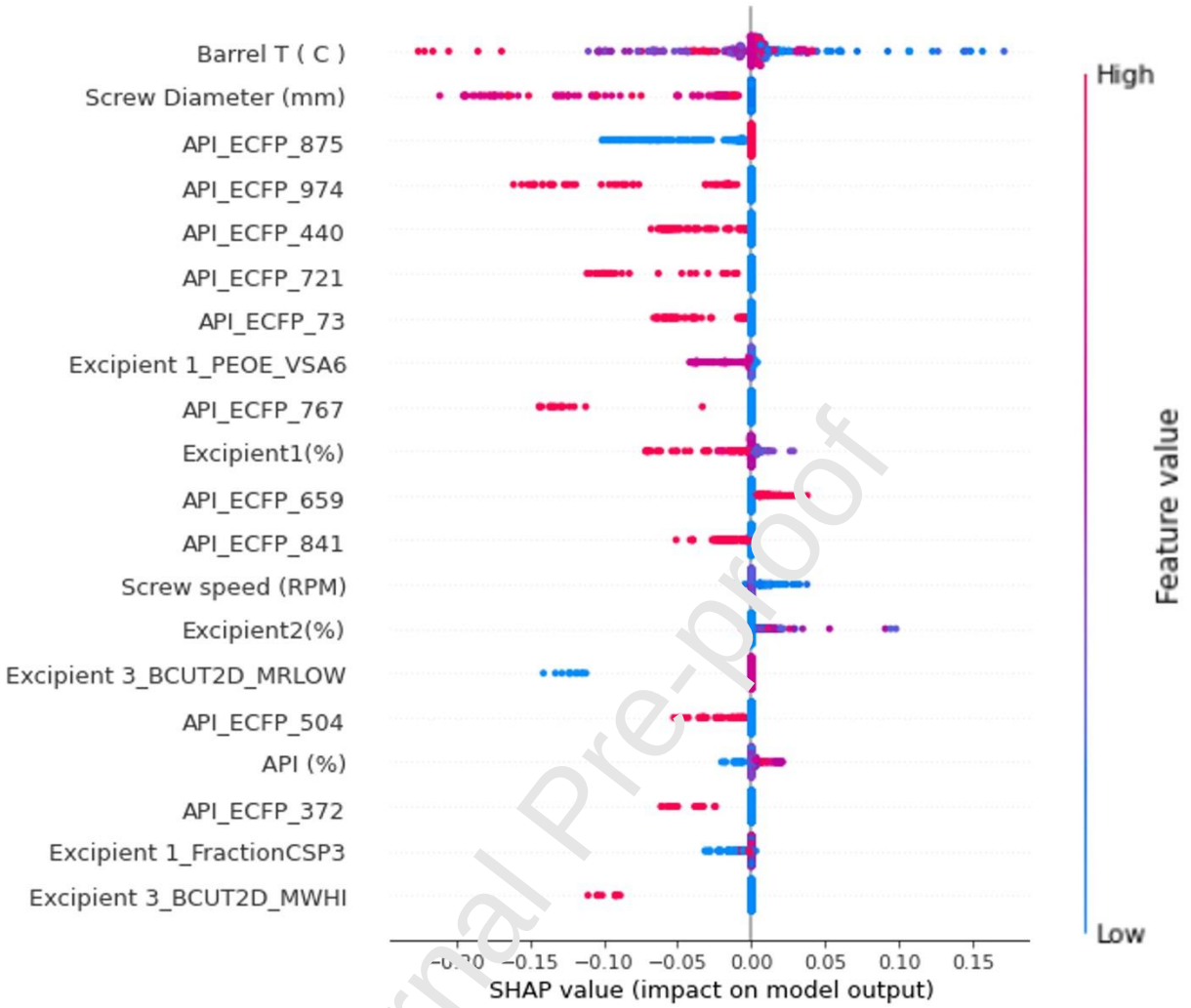


Fig. 5. SHAP summary plot of top 20 features for ECFP-XGBoost. It sorts all features by the sum of SHAP values and shows the impact distribution on the model output. Barrel T(C), barrel temperature; Screw Diameter (mm); API_ECFP_875, drug's substructure at ECFP 875; API_ECFP_974; API_ECFP_440; API_ECFP_721; API_ECFP_73; Excipient 1_PEOE_VSA6, first dominant excipient's molecular surface area descriptor; API_ECFP_767; Excipient1(%), first dominant excipient's loading; API_ECFP_659; API_ECFP_841; Screw speed (RPM); Excipient2(%), second dominant excipient's loading; Excipient 3_BCUT2D_MRLOW, third dominant excipient's topological descriptor; API_ECFP_504; API (%), drug loading; API_ECFP_372; Excipient 1_FractionCSP3, the fraction of C atoms that are SP3 hybridized from first dominant excipient's; Excipient 3_BCUT2D_MWHI, third dominant excipient's topological descriptor. The color bar depicts the value of each feature; blue indicates a higher value, while red indicates a lower value. The higher the SHAP values, the more probability of being chemically stable during the HME process.

3.2.2. Information gain results

To further evaluate the model interpretability and outputs-related structural features, information gain (IG) for each feature was calculated in two of the selected ML models. For the best

amorphization prediction model ECFP-XGBoost, the sorted IG values for the top 20 important features were shown in Fig. 6. Like the SHAP results, barrel temperature and first dominant excipient's loading ranked top two with regards to IG values among all features. First dominant excipient's properties such as VSA_Estate8 (an MOE VSA descriptor), MaxPartialCharge (maximum partial charge), and PEOE_VSA8 (a molecular surface area descriptor), other processing parameters including feed rate, screw speed, and extruder configuration (L/D), and drug's structural features are significant to predicting the amorphization.

In addition, the IG results summary for the chemical stability model ECFP-XGBoost was shown in Fig. 7. Barrel temperature was determined as the most critical feature with the highest IG value. Interestingly, some third dominant excipient's topological descriptors (e.g., BCUT2D_MWHI, BCUT2D_MRLOW, and BCUT2D_LOGPLOW) showed high importance. Screw speed and screw diameter were two other processing parameters that significantly affected chemical stability prediction, as shown in Fig. 7. Moreover, multiple API-related structure features such as ECFP_440, ECFP_792, ECFP_767, ECFP_721, and ECFP_875 are critical to predicting chemical stability and will be further investigated.

Furthermore, we sorted the top 12 important API's ECFP fingerprints calculated by IG analysis in Fig. 8 and Fig. 9. According to the results of the IG values, the critical structural features attributing to the amorphization and chemical stability predictions were identified. For the amorphization prediction model, API substructures: chlorine atom (Fig. 8, 1), benzene (Fig. 8, 2), nitrogen-containing heterocycles (Fig. 8, 3 and 8), pyrimidine (Fig. 8, 4), amide (Fig. 8, 5), benzenediol (Fig. 8, 6), tertiary amine (Fig. 8, 7), nitrogen atom (Fig. 8, 9), phenyl chloride (Fig. 8, 10), carbonyl group (Fig. 8, 11), and sulfur-containing heterocycles (Fig. 8, 12) were found to impact the amorphization model's prediction significantly.

For the chemical stability prediction model, API's substructures: dihydropyridine (Fig. 9, 1, 7), secondary amine (Fig. 9, 2), carbamate (Fig. 9, 3), nitrogen-containing heterocycles (Fig. 9, 4, 12), benzene (Fig. 9, 5, 9), aryl sulfide (Fig. 9, 6, 8), chlorine atom (Fig. 9, 10), and propylthio benzene (Fig. 9, 11) were critical for predicting the output.

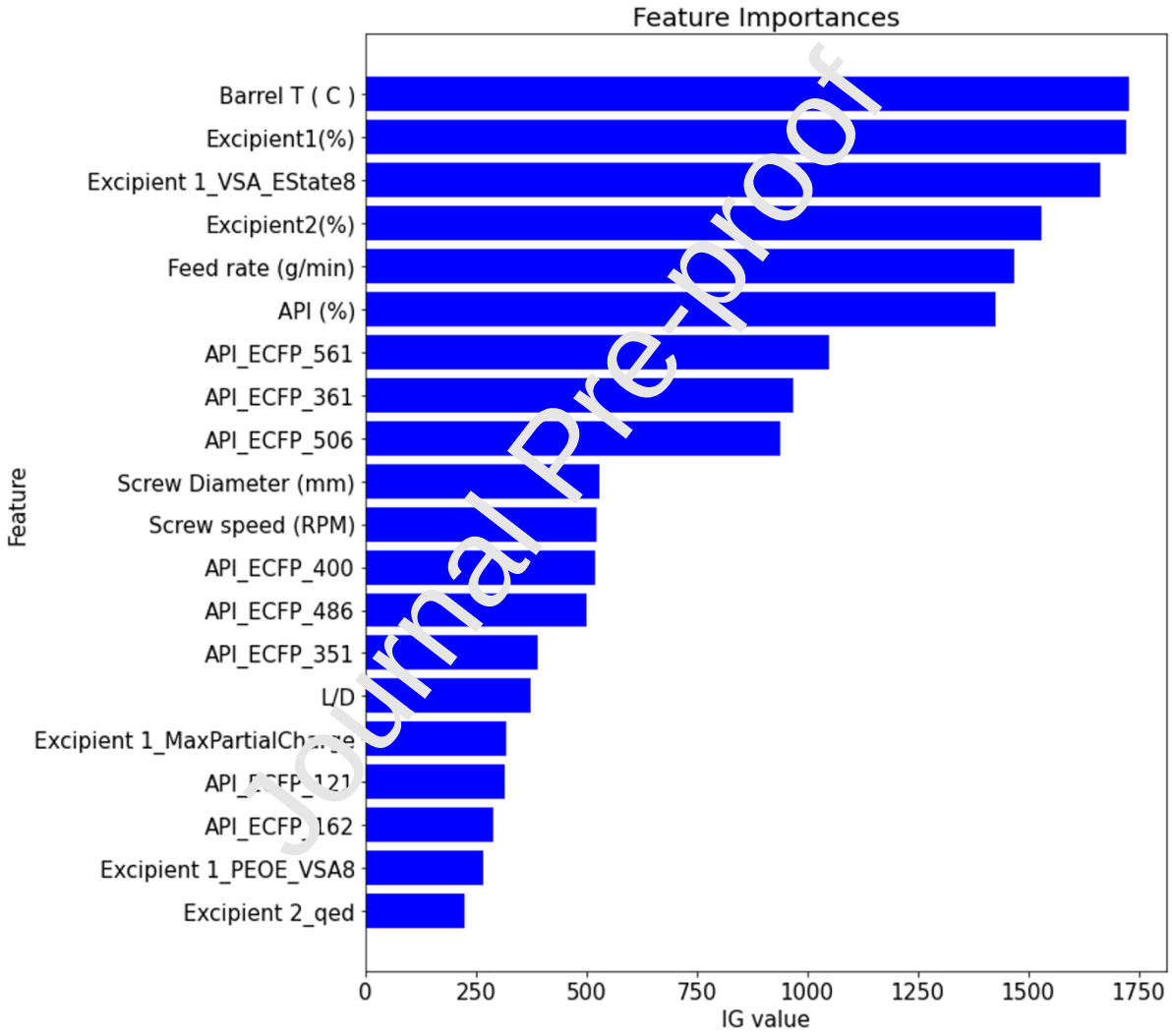


Fig. 6. IG results of the top 20 important features related to the ECFP-LightGBM model for amorphization prediction. Barrel T (C), barrel temperature; Excipient1(%), excipient one's ratio; Excipient 1_VSA_Estate8, first dominant excipient's MOE VSA descriptor; Excipient2(%), excipient two's ratio; Feed rate (g/min); API (%), drug loading; API_ECFP_561, drug's substructure at ECFP 561; API_ECFP_361; API_ECFP_506; Screw Diameter (mm); Screw speed (RPM); API_ECFP_400; API_ECFP_486; API_ECFP_351; L/D; Excipient 1_MaxPartialCharge, first dominant excipient's maximum partial charge; API_ECFP_121; API_ECFP_162; Excipient 1_PEOE_VSA8, first dominant

excipient's molecular surface area descriptor; Excipient 2_qed, second dominant excipient's weighted sum of ADS mapped properties. This graph provides a global explanation of the ML model and summarizes critical features attributed to the model prediction based on the IG values.

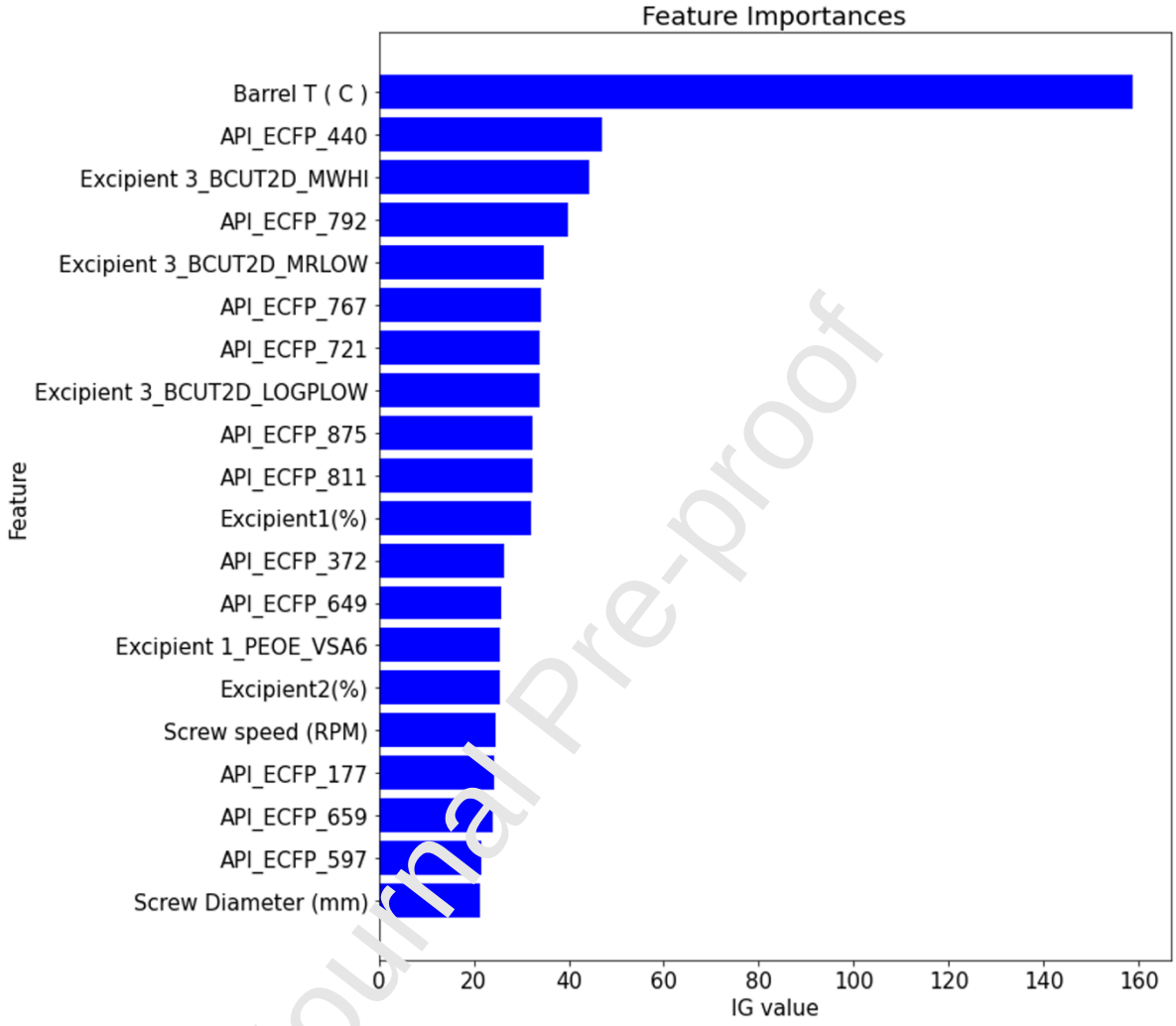


Fig. 7. IG results of the top 20 important features related to the ECFP-XGBoost model for chemical stability prediction. Barrel T (C), barrel temperature; API_ECFP_440, drug's substructure at ECFP 440; Excipient 3_BCUT2D_MWHI, third dominant excipient's topological descriptor; API_ECFP_792; Excipient 3_BCUT2D_MRLOW, third dominant excipient's topological descriptor; API_ECFP_767; API_ECFP_721; Excipient 3_BCUT2D_LOGPLOW, third dominant excipient's topological descriptor; API_ECFP_875; API_ECFP_811; Excipient1(%), excipient one's ratio; API_ECFP_372; API_ECFP_649; Excipient 1_VSA_Estate8, first dominant excipient's MOE VSA descriptor; Excipient2(%), excipient two's ratio; Screw speed (RPM); API_ECFP_177; API_ECFP_659; API_ECFP_597; Screw Diameter (mm). This graph provides a global explanation of the ML model and summarizes critical features attributed to the model prediction based on the IG values.

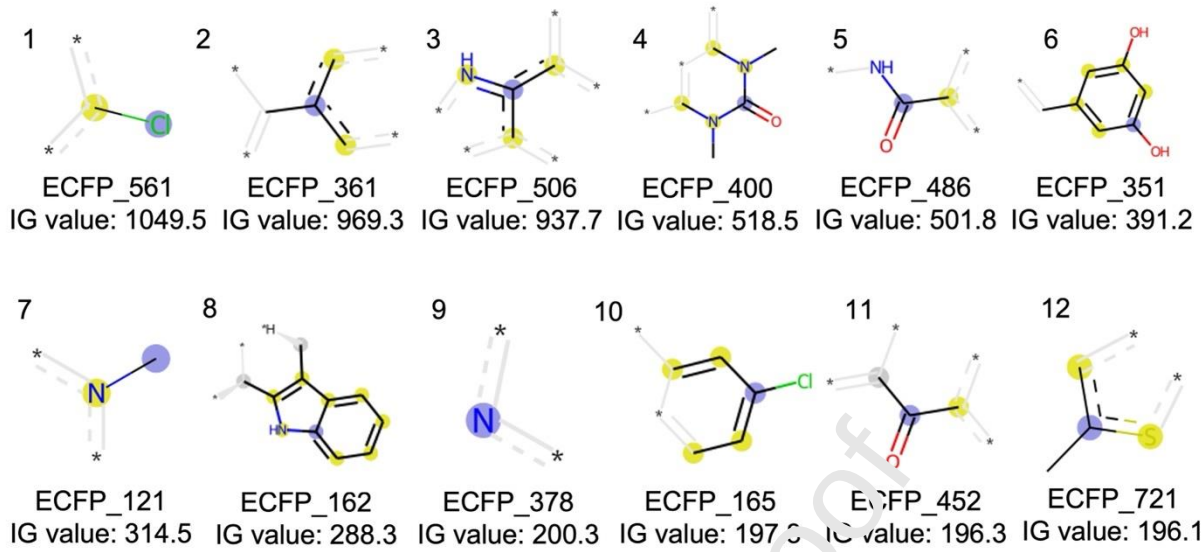


Fig. 8. Top 12 important API substructures and IG values for amorphization. The highlight color in the chemical structures indicates blue: the central atom in the environment; yellow: aromatic atoms; gray: aliphatic ring atoms.

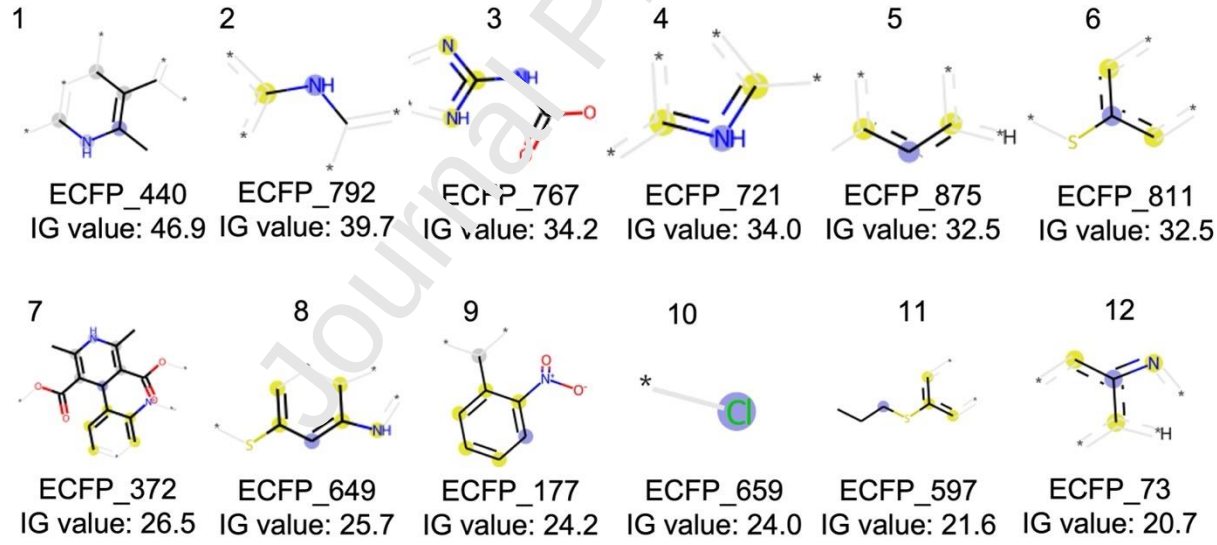


Fig. 9. Top 12 important API substructures and IG values for chemical stability. The highlight color in the chemical structure indicates blue: the central atom in the environment; yellow: aromatic atoms; gray: aliphatic ring atoms.

4. Discussion

4.1. Model development and selection

In this study, we first obtained a dataset containing 760 amorphization and 495 chemical stability HME datapoints by literature mining from 158 publications. We note that some articles did not provide the drug content information after HME processing from HPLC, so the chemical stability data points were relatively less than the amorphization data points. During the data processing, two molecular representation methods (i.e., 2D-descriptors and ECFP) were employed to compute API's properties and evaluate the molecular description's effect on the model performance. We observed that ECFP-based models showed a slightly better performance overall than the 2D-based model in the testing set. These two types of models performed similarly in the cross-validation sets. Dong et al. also found that ECFP-based models were superior to 2D-based models when evaluating the dissolution types of solid dispersion (Dong et al., 2021). According to the evaluation metrics results in Table 2 and Table 3, RF showed relatively poor performance with lower ACC, F1, and AUC in cross-validation and testing subsets. This is because the RF algorithm is a collection of multiple decision trees in a bagging ensemble method and makes the decision based on the majority of the trees, which would be problematic when having an imbalanced dataset (Fig. 10) (Aman Gupta, 2021). In addition, RF fails to maintain its performance when the data are sparse (Mohammed Shammeer, 2021). SVM-based models showed a moderate predictive performance for chemical stability but poor performance for amorphization. Specifically, 2D-SVM exhibited the lowest ACC (0.875), F1 (0.928), and AUC (0.792) in predicting amorphization among other ML models. This is likely because (1) it's sensitive to noise (i.e., target classes are overlapping), and (2) it doesn't perform well when having a large dataset and high dimensional features (Dhiraj K, 2019). Among all ML algorithms, LightGBM and XGBoost perform well and are suitable for both amorphization and chemical stability. One of the advantages of gradient boosting models such as XGBoost and

LightGBM is that it gives more importance to the misclassified categories and will minimize the loss by adding weak classifiers using gradient descent (Fig. 10). Then, a robust classifier was obtained through a gradient optimization process by all weak classifiers, which leads to high accuracy and prevents overfitting. Based on their specific structure, XGBoost and LightGBM are more robust than random forest, especially for imbalanced data in this study. Based on the metrics results, ECFP-LightGBM and ECFP-XGBoost performed well in the cross-validation and testing subsets are selected as the best models for amorphization and chemical stability, respectively.

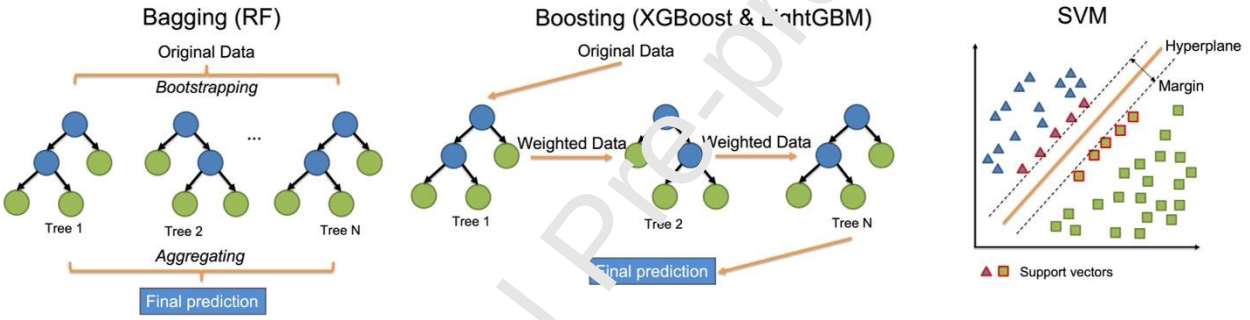


Fig. 10. Schematic representation of ML algorithms (i.e., RF, XGBoost, LightGBM, and SVM) used in this study. RF is a bagging algorithm that first constructs data subsets from the original data using the bootstrapping method. Then each decision tree will be trained, and the ensemble classifier will aggregate the results and make a prediction from each tree using majority voting. XGBoost and LightGBM are two Boosting algorithms that utilize the tree-growing concept sequentially. This boosting approach can construct new classifiers from previous ones and learn errors, resulting in lower bias. Finally, SVM is a supervised algorithm aiming to find a hyperplane in an N-dimensional space that can separate the data points.

4.2. Effect of input variables on amorphization

We successfully applied SHAP and IG analysis to further investigate the model's interpretability. According to the results, excipient one's ratio and barrel temperature are the top 2 important features in both methods. (Fig. 4 and Fig. 6) Barrel temperature, also regarded as thermal energy during the HME process, is one of the most critical factors when preparing ASDs (Ma et al., 2019). The thermal energy is typically generated by heat conduction from the screw elements

and barrel. It has been reported that an increased barrel temperature would increase molecular motion and diffusivity, resulting in facilitating the drug solubilization into the polymer matrix (Ma et al., 2019). It has been demonstrated that increased barrel temperature can effectively convert the crystalline drug to amorphous in several studies concerning multiple drug molecules such as nifedipine, carbamazepine, and gliclazide (Huang et al., 2017, 2016; Yang et al., 2016). Most importantly, Ma et al. studied the effect of energy input (i.e., thermal energy and specific mechanical energy) on ASDs and illustrated that the amorphization process is more triggered by thermal energy (Ma et al., 2019). Specifically, the increase in thermal energy would reduce the required specific mechanical energy, and the amorphization process is only attributed to thermal energy when the temperature is higher than 140 °C (Ma et al., 2019). Therefore, barrel temperature is one of the most important features and the model's explanation of results correspond to the literature. In addition, according to feature importance results, drug loading or excipient's ratio is also significant for the model's prediction. Based on SHAP local explanation, drug loading negatively correlates with amorphization. Drug loading must be carefully considered during ASD development, which is relevant to drug-polymer solubility and miscibility. Qian et al. proposed a hypothetical diagram of drug-polymer solubility, miscibility, and glass transition temperature (T_g) of an ASD system (Qian et al., 2010). This diagram studied the effect of temperature and drug loading on the miscibility and solubility between the drug and polymer. In addition, the diagram indicated that a supersaturated and immiscible mixture would form with high drug loading, which corresponds to the SHAP results in our study (Qian et al., 2010). Moreover, Tian et al. also studied the phase diagram of the drug-polymer system based on Flory-Huggins' interaction parameters (Tian et al., 2019). The diagram demonstrated that phase separation happened within both the binodal and spinodal boundaries in which the drug loading

is relatively high (Tian et al., 2019). In addition, according to SHAP and IG analysis, processing parameters, including feed rate (g/min) and screw speed (RPM), play important roles in the drug amorphization process. Changes in feed rate would impact the barrel fill rate, residence time, melt viscosity, and mechanical energy during the HME process. Screw speed also influences the residence time and specific mechanical energy (Butreddy et al., 2021). Specifically, a higher screw speed will lead to a shorter residence time and higher specific mechanical energy. In comparison, lower screw speed will contribute to a longer residence time and lower specific mechanical energy (Butreddy et al., 2021; Thompson and Williams, 2021). Extruder configurations, which can be indicated by screw diameter and L/D ratio, are also critical attributes, especially when scaling up for ASD manufacturing. It's typical to use a small-scale extruder with a screw diameter of 16-20 mm, which yields a production rate of 1-10 kg/h (Brown et al., 2014). 24-30 mm-diameter extruder is typically used for intermedia scale manufacturing (10-50kg/h) (Brown et al., 2014). However, it is difficult for intermedia scale extruders to conduct heat transfer, venting, and devolatilization compared to small scale units due to the lower barrel surface area (Brown et al., 2014). Haser et al. studied the scale-up development process of meloxicam ASD from Nano-16 to Micro-18 twin-screw extruder and found that full amorphous conversion of the crystalline drug was challenging due to the reduced peak shear of the bilobal geometry in Micro-18 (Haser et al., 2018b). Besides the above-mentioned CMAs and CPPs, other attributes, such as particle size distribution of the physical blends and die temperature, also affect the forming of ASDs. A reduced drug particle size or increased surface area can facilitate the dissolution rate of a drug in the rubbery polymer matrix, resulting in a higher degree of amorphization (Hempel et al., 2020). However, most of the literature does not provide information on the particle size distribution of drugs and polymers, so this variable was

not included in the dataset for ML modeling. In addition, die temperature is another critical parameter during HME processing, and the literature has shown that a slightly higher die temperature is important to reduce the die pressure (LaFountaine et al., 2016). Unfortunately, some literature failed to provide the information, which is likely because some instruments do not contain a separate heating element for the die. Therefore, we used barrel temperature as an HME processing parameter.

The top 12 important structural features of API that contributed to amorphization prediction were identified by IG in Fig. 8. Among all features, the chlorine atom (ECFP_561) was the most critical for building the model. Suzuki et al. reported that the chlorine atom was covalently bound to the benzene ring in indomethacin and had unique intermolecular interactions and halogen bonds with oxygen atoms, which probably contributed to the amorphization process (Suzuki et al., 2021). Kawakami comprehensively reviewed factors such as chemical structure, processing methods, and storage conditions that will affect the crystallization tendency of pharmaceutical glasses (Kawakami, 2019). In this review, a good glass former, synonymous with an amorphous solid, should have the following chemical-structural features: large molecular weight, low symmetry, low number of benzene rings, many rotatable bonds, and many more electronegative atoms, and high branching degree (Kawakami, 2019). According to IG results for structural features (Fig. 8), structures containing benzene rings (ECFP_361, ECFP_400, ECFP_351, and ECFP_165), nitrogen-containing heterocycles (ECFP_506 & ECFP_162), tertiary amine (ECFP_121), nitrogen atom (ECFP_378), and chlorine atom (ECFP_561) were the critical substructures for amorphization prediction, which mostly agrees with the literature above. Therefore, the amorphization model (ECFP-LightGBM) has demonstrated good interpretability based on SHAP and IG analysis.

4.3. Effects of input variables on chemical degradation

When analyzing the critical features for the selected chemical stability model (ECFP-XGBoost), barrel temperature was the most significant one, and it showed a negative correlation with chemical stability Fig. 5 and Fig. 7. Huang et al. stated that the effect of temperature on drug degradation rate could be described by Arrhenius kinetics (Equation):

Equation 3

$$k_T = k_{ref}\exp[-\frac{E_A}{R}(\frac{1}{T}-\frac{1}{T_{ref}})]$$

where k_T is the drug degradation rate at temperature T (K), k_{ref} is the drug degradation rate at the reference temperatureT_{ref} (K), E_A is the activation energy (J/mol), and R is the gas constant value (8.3145 J/mol·K) (Huang et al., 2017). According to Equation , the drug degradation rate is proportional to the processing temperature. In addition, chemical degradation during HME consists of oxidation and hydrolysis (Haser et al., 2017; Huang et al., 2017; Surasarang et al., 2016). Multiple thermally labile drugs such as albendazole, meloxicam, and gliclazide have shown severe chemical degradation when increasing barrel temperature (Haser et al., 2017; Huang et al., 2017; Surasarang et al., 2016). Therefore, barrel temperature is the most important feature for building the chemical stability model. In addition, screw diameter, an important indicator of extruder configuration, is a critical feature and showed a negative correlation with chemical stability. Mati´c et al. studied the effect of extruder configuration on the API degradation and found that the ZSE18 extruder (screw diameter = 18 mm) tended to have a higher degree of drug degradation than the ZSE12 extruder (screw diameter = 12 mm) (Mati´cmati´c et al., 2021). Haser et al. also observed significant chemical degradation (75.5%

purity) when scaling up from Nano-16 to Micro-18 extruder with no additional process change (Haser et al., 2018b). Moreover, other processing parameters, such as screw speed, play an important role in chemical stability during the HME process. The increase in screw speed generally increases specific mechanical energy and ultimately affects the chemical stability of an ASD (Thompson and Williams, 2021). In addition, higher screw speed may contribute to viscous heating, especially when extruding low-conductivity polymers, which can be indicated by a low Nahme-Griffith number (Marschik et al., n.d.). The excess thermal energy generated through high screw speed will further lead to chemical degradation. According to SHAP and IG results, drug loading or excipient's ratio is important to the chemical stability model. This is likely because some vinyl polymers, such as copovidone and povidone with residual peroxides, can trigger oxidative degradation (Iyer et al., 2021). Some excipient features, including Excipient 1_PEOE_VSA6 (first dominant excipient's molecular surface area descriptor), Excipient 3_BCUT2D_MWHI (third dominant excipient's topological descriptor), and Excipient 3_BCUT2D_MRLOW (third dominant excipient's topological descriptor), are also critical to building the prediction model, and further analysis must be conducted for evaluation.

By analyzing and comparing the IG results, we identified the top 12 important structural features related to the drug's chemical stability during the HME process (Fig. 9). Dihydropyridine was identified as the most important substructure with an IG value of 46.9 that significantly affected chemical stability. Literature has demonstrated that drug compounds containing dihydropyridine structures tend to be chemically unstable and will degrade through oxidization. For example, Damian et al. observed the photo-degradation of nifedipine under UV and formed nitroso-nifedipine derivative as a chemical degradant (Damian et al., n.d.). Dattatray et al. stated that meloxicam experienced oxidative degradation and formed PV and MIV as degradation products

(Modhave et al., 2011). In addition, amide in meloxicam (ECFP_792) will undergo chemical degradation through the hydrolysis pathway and form acid and an amine (Haser et al., 2017). And further degradation will happen to the acid through decarboxylation (Haser et al., 2017). Carbamate (ECFP_767) was also identified as a critical structural feature that affects chemical stability with an IG value of 34.2 (Fig. 9). API that contains carbamate structure may suffer chemical degradation during the HME process. For example, albendazole degrades into albendazole impurity A and methanol through basic hydrolysis (Surasarang et al., 2016). Moreover, drug compounds containing aryl sulfide (ECFP_811) are also vulnerable to chemical degradation. Surasarang et al. observed the chemical degradation of albendazole under high processing temperature or hydrogen peroxide (H_2O_2) and will form albendazole sulfoxide through the oxidation pathway as a result (Surasarang et al., 2016). Nitrogen-containing heterocycles (ECFP_721 & ECFP_73) are also chemically unstable, and the drug containing this substructure may degrade during the HME process (Focante et al., 2006). Overall, the IG results of the drug's structural features are mostly consistent with those reported in the literature, demonstrating the model's interpretability. Most importantly, the feature importance analysis will provide guidance on developing ASD formulation by simply inputting tabular data (i.e., API, excipients such as polymer, and processing parameters), and it can identify the potential failures due to the chemical degradation of drugs.

5. Conclusion

This study describes a novel method for applying multiple ML models to predict a drug's amorphization and chemical stability during the HME process. We first obtained a dataset to build up the ML models by literature mining from recent publications. Then, data processing

steps such as train-test split, molecular representation, and solving missing values were performed. Multiple metrics and feature importance tools were applied to evaluate the model prediction performance and interpretability. ECFP-LightGBM and ECFP-XGBoost were the best models for predicting amorphization and chemical stability, respectively. More importantly, the selected models showed good interpretability based on SHAP and IG results. Several important features, including barrel temperature, drug loading, extruder configuration, and API's chemical substructures, were identified and need to be carefully considered during ASD development in the future. This study used the ASD data generated by HME to build up ML models, which may render it applicable to ASD prepared by other techniques such as spray drying, KinetiSol, and antisolvent precipitation. By utilizing ML techniques to predict the forming of chemically stable ASDs, we may significantly reduce the workload of preliminary experiments and potentially facilitate the product development process of ASD.

CRediT authorship contribution statement

Junhuang Jiang: Conceptualization, Methodology, Data curation, Software, Visualization, Writing-original draft. **Angu Lu:** Data curation. **Xiangyu Ma:** Conceptualization, Writing-review & editing. **Defang Ouyang:** Conceptualization, Validation, Writing-review & editing. **Robert O. Williams III:** Conceptualization, Writing-review & editing, Supervision.

Declaration of Competing Interest

The authors declare no financial interests/personal relationships that may be considered as potential competing interests.

References

Adankon, M.M., Cheriet, M., 2009. Support Vector Machine. Encyclopedia of Biometrics 1303–1308. https://doi.org/10.1007/978-0-387-73003-5_299

Alhalaweh, A., Alzghoul, A., Kaialy, W., Mahlin, D., Bergström, C.A.S., 2014. Computational predictions of glass-forming ability and crystallization tendency of drug molecules. Mol Pharm 11, 3123–3132. https://doi.org/10.1021/MP500303A/ASSET/IMAGES/LARGE/MP-2014-00303A_0008.JPEG

Alonzo, D.E., Zhang, G.G.Z., Zhou, D., Gao, Y., Taylor, L.S., 2010. Understanding the behavior of amorphous pharmaceutical systems during dissolution. Pharm Res 27, 608–618. https://doi.org/10.1007/S11095-009-0021-1

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brown, C., DiNunzio, J., Eglesia, M., Forster, S., Lamm, M., Lowinger, M., Marsac, P., McKelvey, C., Meyer, R., Schenck, L., Terife, G., Troup, G., Smith-Goettler, B., Starbuck, C., 2014. HME for Solid Dispersions: Scale-Up and Late-Stage Development 231–260. https://doi.org/10.1007/978-1-4939-1598-9_7

Butreddy, A., Bandari, S., Repka, M.A., 2021. Quality-by-design in hot melt extrusion based amorphous solid dispersions: An industrial perspective on product development. European Journal of Pharmaceutical Sciences 158, 105655. https://doi.org/10.1016/J.EJPS.2020.105655

Chen, T., … C.G. sigkdd international conference on knowledge, 2016, undefined, 2016. Xgboost: A scalable tree boosting system. dl.acm.org 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785

Chiou, W., sciences, S.R.-J. of pharmaceutical, 1971, undefined, n.d. Pharmaceutical applications of solid dispersion systems. Elsevier.

Damian, G., Schmutzer, G., and, D.P.-J. of O., 2007, undefined, n.d. Investigation of light-induced free radicals in nifedipine. academia.edu.

Dong, J., Gao, H., Ouyang, D., 2021. PharmSD: A novel AI-based computational platform for solid dispersion formulation design. Int J Pharm 604, 120705. https://doi.org/10.1016/J.IJPHARM.2021.120705

Focante, F., Mercandelli, P., Sironi, A., Resconi, L., 2006. Complexes of tris(pentafluorophenyl)boron with nitrogen-containing compounds: Synthesis, reactivity and metallocene activation. Coord Chem Rev 250, 170–188. https://doi.org/10.1016/J.CCR.2005.05.005

Han, R., Xiong, H., Ye, Z., Yang, Y., Huang, T., Jing, Q., Lu, J., Pan, H., Ren, F., Ouyang, D., 2019. Predicting physical stability of solid dispersions by machine learning techniques. Journal of Controlled Release 311–312, 16–25. https://doi.org/10.1016/J.JCONREL.2019.08.030

Haser, A., Cao, T., Lubach, J.W., Zhang, F., 2018a. In Situ Salt Formation during Melt Extrusion for Improved Chemical Stability and Dissolution Performance of a Meloxicam−Copovidone Amorphous Solid Dispersion. https://doi.org/10.1021/acs.molpharmaceut.7b01057

Haser, A., Haight, B., Berghaus, A., Machado, A., Martin, C., Zhang, F., 2018b. Scale-Up and In-line Monitoring During Continuous Melt Extrusion of an Amorphous Solid Dispersion. AAPS PharmSciTech 19, 2818–2827. https://doi.org/10.1208/S12249-018-1162-5

Haser, A., Huang, S., Listro, T., White, D., Zhang, F., 2017. An approach for chemical stability during melt extrusion of a drug substance with a high melting point. Int J Pharm 524, 55–64. https://doi.org/10.1016/J.IJPHARM.2017.03.070

Hempel, N.J., Knopp, M.M., Berthelsen, R., Zeitler, J.A., Löbmann, K., 2020. The influence of drug and polymer particle size on the in situ amorphization using microwave irradiation. European Journal of Pharmaceutics and Biopharmaceutics 149, 77–84. https://doi.org/10.1016/J.EJPB.2020.01.019

Huang, S., O'Donnell, K.P., Delpon de Vaux, S.A., O'Brien, J., Stutzman, J., Williams, R.O., 2017. Processing thermally labile drugs by hot-melt extrusion: The lesson with gliclazide. European Journal of Pharmaceutics and Biopharmaceutics 119, 56–67. https://doi.org/10.1016/J.EJPB.2017.05.014

Huang, S., O'Donnell, K.P., Keen, J.M., Rickard, M.A., McGinity, J.W., Williams, R.O., 2016. A New Extrudable Form of Hypromellose: AFFINISOL™ HPMC HME. AAPS PharmSciTech 17, 106–119. https://doi.org/10.1208/S12249-015-0395-9/FIGURES/13

Huang, S., Williams, R.O., 2018. Effects of the Preparation Process on the Properties of Amorphous Solid Dispersions. AAPS PharmSciTech 19, 1971–1984. https://doi.org/10.1208/S12249-017-0861-7/TABLES/1

Iyer, R., Jovanovska, V.P., Berginc, K., Jaklič, M., Fabiani, F., Harlacher, C., Huzjak, T., Sanchez-Felix, M.V., 2021. Amorphous Solid Dispersions (ASDs): The Influence of

Material Properties, Manufacturing Processes and Analytical Technologies in Drug Product Development. Pharmaceutics 13. https://doi.org/10.3390/PHARMACEUTICS13101682

Jermain, S. v., Brough, C., Williams, R.O., 2018. Amorphous solid dispersions and nanocrystal technologies for poorly water-soluble drug delivery – An update. Int J Pharm 535, 379–392. https://doi.org/10.1016/J.IJPHARM.2017.10.051

Jiang, J., Peng, H.H., Yang, Z., Ma, X., Sahakijpijarn, S., Moon, C., Ouyang, D., Williams, R.O., 2022. The applications of Machine learning (ML) in designing dry powder for inhalation by using thin-film-freezing technology. Int J Pharm 626, 122179. https://doi.org/10.1016/J.IJPHARM.2022.122179

Kawakami, K., 2019. Crystallization Tendency of Pharmaceutical Glasses: Relevance to Compound Properties, Impact of Formulation Process, and Implications for Design of Amorphous Solid Dispersions. Pharmaceutics 2019, Vol. 11, Page 202 11, 202. https://doi.org/10.3390/PHARMACEUTICS11050202

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., n.d. Lightgbm: A highly efficient gradient boosting decision tree. papers.nips.cc.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., n.d. LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

Kopitar, L., Cilar, L., Kocbek, P., Stiglic, G., 2019. Local vs. Global Interpretability of Machine Learning Models in Type 2 Diabetes Mellitus Screening. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11979 LNAI, 108–119. https://doi.org/10.1007/978-3-030-37446-4_9/FIGURES/3

LaFountaine, J.S., Prasad, L.K., Brough, C., Miller, D.A., McGinity, J.W., Williams, R.O., 2016. Thermal Processing of PVP- and HPMC-Based Amorphous Solid Dispersions. AAPS PharmSciTech 17, 120–132. https://doi.org/10.1208/S12249-015-0417-7

Lee, H., Kim, J., Kim, S., Yoo, J., Choi, G.J., Jeong, Y.S., 2022. Deep Learning-Based Prediction of Physical Stability considering Class Imbalance for Amorphous Solid Dispersions. J Chem 2022. https://doi.org/10.1155/2022/4148443

Liu, X., Lu, M., Guo, Z., Huang, L., Feng, X., Wu, C., 2012. Improving the chemical stability of amorphous solid dispersion with cocrystal technique by hot melt extrusion. Pharm Res 29, 806–817. https://doi.org/10.1007/S11095-011-0605-4/FIGURES/9

Lu, M., Guo, Z., Li, Y., Pang, H., Lin, L., Liu, X., Pan, X., Wu, C., 2014. Application of Hot Melt Extrusion for Poorly Water-Soluble Drugs: Limitations, Advances and Future Prospects. Curr Pharm Des 20, 369–387. https://doi.org/10.2174/13816128113199990402

Ma, X., Huang, S., Lowinger, M.B., Liu, X., Lu, X., Su, Y., Williams, R.O., 2019. Influence of mechanical and thermal energy on nifedipine amorphous solid dispersions prepared by hot melt extrusion: Preparation and physical stability. Int J Pharm 561, 324–334. https://doi.org/10.1016/J.IJPHARM.2019.03.014

Ma, X., Kittikunakorn, N., Sorman, B., Xi, H., Chen, A., Marsh, M., Mongeau, A., Piché, N., Williams, R.O., Skomski, D., 2020. Application of Deep Learning Convolutional Neural Networks for Internal Tablet Defect Detection: High Accuracy, Throughput, and Adaptability. J Pharm Sci 109, 1547–1557. https://doi.org/10.1016/J.XPHS.2020.01.014

Marschik, C., Roland, W., Polymers, J.M.-, 2018, undefined, n.d. A network-theory-based comparative study of melt-conveying models in single-screw extrusion: A. isothermal flow. mdpi.com.

Mati´cmati´c, J., Alva, C., Eder, S., Reusch, K., Paudel, A., Khinast, J., 2021. Towards predicting the product quality in hot-melt extrusion: Pilot plant scale extrusion. Int J Pharm X 3, 100084. https://doi.org/10.1016/j.ijpx.2021.100084

Modhave, D.T., Handa, T., Shah, R.P., Singh, S., 2011. Successful characterization of degradation products of drugs using LC-MS tools: Application to piroxicam and meloxicam. Analytical Methods 3, 2864–2872. https://doi.org/10.1039/C1AY05493G

Moseson, D.E., Taylor, L.S., 2018. The application of temperature-composition phase diagrams for hot melt extrusion processing of amorphous solid dispersions to prevent residual crystallinity. Int J Pharm 553, 454–466. https://doi.org/10.1016/J.IJPHARM.2018.10.055

Pandi, P., Bulusu, R., Kommineni, N., Khan, W., Singh, M., 2020. Amorphous solid dispersions: An update for preparation, characterization, mechanism on bioavailability, stability, regulatory considerations and marketed products. Int J Pharm 586, 119560. https://doi.org/10.1016/J.IJPHARM.2020.119560

Qian, F., Huang, J., Hussain, M.A., 2010. Drug–Polymer Solubility and Miscibility: Stability Consideration and Practical Challenges in Amorphous Solid Dispersion Development. J Pharm Sci 99, 2941–2947. https://doi.org/10.1002/JPS.22074

Raghunathan, S., Priyakumar, U.D., 2022. Molecular representations for machine learning applications in chemistry. Int J Quantum Chem 122, e26870. https://doi.org/10.1002/QUA.26870

Random Forest Fails. The serious mess that random forest… | by Mohammed Shammeer | The Startup | Medium [WWW Document], n.d. URL https://medium.com/swlh/random-forest-fails-a8ca2d46c312 (accessed 9.28.22).

RDKit [WWW Document], n.d. URL https://www.rdkit.org/ (accessed 7.12.22).

Schittny, A., Huwyler, J., Puchkov, M., Huwyler, O., 2019. Mechanisms of increased bioavailability through amorphous solid dispersions: a review. https://doi.org/10.1080/10717544.2019.1704940

Surasarang, S.H., Keen, J.M., Huang, S., Zhang, F., Mcginity, J.W., Iii, R.O.W., 2016. Drug Development and Industrial Pharmacy Hot melt extrusion versus spray drying: hot melt extrusion degrades albendazole Hot melt extrusion versus spray drying: hot melt extrusion degrades albendazole. https://doi.org/10.1080/03639045.2016.1220577

Suzuki, H., Iwata, M., Ito, M., Noguchi, S., 2021. X-ray Absorption Near-Edge Spectroscopy Analysis of Indomethacin in Crystalline Forms and in Amorphous Solid Dispersions. Mol Pharm 18, 3475–3483. https://doi.org/10.1021/ACS.MOLPHARMACEUT.1C00405/ASSET/IMAGES/LARGE/MP1C00405_0010.JPEG

Thompson, S.A., Williams, R.O., 2021. Specific mechanical energy – An essential parameter in the processing of amorphous solid dispersions. Adv Drug Deliv Rev 173, 374–393. https://doi.org/10.1016/J.ADDR.2021.03.006

Tian, Y., Qian, K., Jacobs, E., Amstad, E., Jones, D.S., Stella, L., Andrews, G.P., 2019. The Investigation of Flory–Huggins Interaction Parameters for Amorphous Solid Dispersion Across the Entire Temperature and Composition Range. Pharmaceutics 2019, Vol. 11, Page 420 11, 420. https://doi.org/10.3390/PHARMACEUTICS11080420

Top 4 advantages and disadvantages of Support Vector Machine or SVM | by Dhiraj K | Medium [WWW Document], n.d. URL https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107 (accessed 9.28.22).

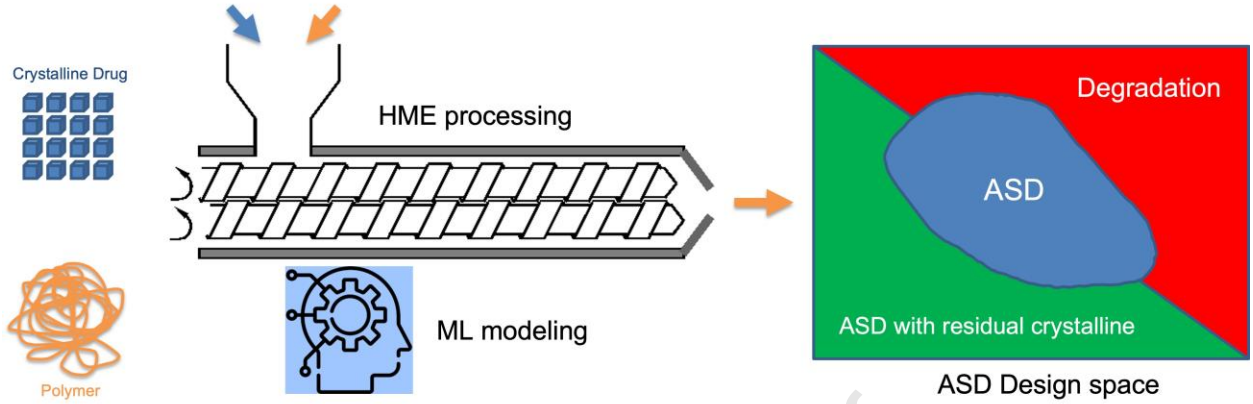
Wang, N., Sun, H., Dong, J., Ouyang, D., 2021. PharmDE: A new expert system for drug-excipient compatibility evaluation. Int J Pharm 607, 120962. https://doi.org/10.1016/J.IJPHARM.2021.120962

Wang, W., Feng, S., Ye, Z., Gao, H., Lin, J., Ouyang, D., 2022. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. Acta Pharm Sin B 12, 2950–2962. https://doi.org/10.1016/J.APSB.2021.11.021

XGBoost versus Random Forest. This article explores the superiority … | by Aman Gupta | Geek Culture | Medium [WWW Document], n.d. URL https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30 (accessed 9.28.22).

Yang, F., Su, Y., Zhang, J., Dinunzio, J., Leone, A., Huang, C., Brown, C.D., 2016. Rheology Guided Rational Selection of Processing Temperature To Prepare Copovidone−Nifedipine Amorphous Solid Dispersions via Hot Melt Extrusion (HME). https://doi.org/10.1021/acs.molpharmaceut.6b00516

Ye, Z., Ouyang, D., 2021. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. Ye and Ouyang Journal of Cheminformatics 13, 98. https://doi.org/10.1186/s13321-021-00575-3



Graphical Abstract