Virtual coformer screening by a combined machine learning and physics-based approach†

Jiuchuang Yuan,‡ Xuetao Liu,‡ Simin Wang, Chao Chang, Qiao Zeng, Zhengtian Song, Yingdi Jin, Qun Zeng, Guangxu Sun, Shigang Ruan, Chandler Greenwell and Yuriy A. Abramov *cd

Cocrystals as a solid form technology for improving physicochemical properties have gained increasing popularity in the pharmaceutical, nutraceutical, and agrochemical industries. However, the list of potential coformers contains hundreds of molecules; far more than can be routinely screened and confirmed. Cocrystal screening experiments require significant amounts of active ingredients at an early project stage, and are expensive and time-consuming. Physics-based models and machine learning (ML) models have both been used to perform virtual cocrystal screening to guide experimental screening efforts, but both have certain limitations. Here, we present a combined ML/COSMO-RS fast virtual cocrystal screening method that proves to be significantly better than the sum of its parts in application to internal and external validation sets. To achieve that, we have defined the optimal threshold values of ML cocrystallization probability and COSMO-RS excess enthalpy of drug/coformer mixing for the combined coformer ranking. An approach to determine an applicability domain (AD) of the ML model has been implemented. The speed and accuracy of the new combined model allow it to be a good alternative to the physics-based CSP-based approach to support pharmaceutical projects with tight timeline and budget constraints.

1. Introduction

The vast majority of drug products, marketed worldwide, are formulated in oral solid dosage forms, most of which are crystalline. Formulation of drug substances in a crystalline form tends to provide superior chemical stability, purity, and manufacturability relative to amorphous and liquid form formulations. However, crystallization of a free form of a drug molecule may be challenging, or the physical properties of the form may be unacceptable. In that case, a frequently employed strategy in the pharmaceutical industry involves a solid form change from a monocomponent (free base or acid) into a multicomponent one (salt or cocrystal).1

Cocrystallization is becoming a major tool for inducing crystallization and/or property optimization via solid form change for poorly ionizable drug molecules. Pharmaceutical cocrystals can be defined as a homogeneous solid phase containing a neutral drug molecule with one or more coformers in a crystal lattice with a defined stoichiometry.2 Cocrystals are crystalline in their pure form under ambient conditions. Cocrystallization may allow optimization of various properties of a drug substance,3,4 including aqueous solubility and dissolution rate,5–9 mechanical properties,10–12 hydration stability13–17 and chemical stability.18,19

The number of coformers for screening in the pharmaceutical industry is potentially in the hundreds, and more than one technique is often applied to confirm cocrystal formation.20 Therefore, experimental coformer screening greatly benefits from guidance provided by computational methods.21 Computational screening serves to guide selection of a subset of the most promising coformers for an API cocrystallization experimental follow up.

Current approaches to rational coformer screening may be roughly divided into three groups: (1) knowledge-based (KB),22–27 (2) physics-based (PB)28–36 and (3) machine learning37 approaches. The physics-based methods adopt a molecular level consideration and include synthonic engineering,28 Hansen solubility parameters,29,30 surface site interaction points (SSIPs),31–33 and COSMO-RS34,35 and CSP-
based\textsuperscript{36} approaches. There are two contributions to cocrystal formation.\textsuperscript{36} One is related to miscibility between the drug molecule and coformer in the amorphous phase. It can be measured by mixing (or excess) free energy and reflects contributions of short-range interactions only. The crystallinity contribution to cocrystal formation is measured by a change of free energy of fusion during cocrystallization. It reflects long-range order effects. An overall energy lowering should be provided by the formation of cocrystals.\textsuperscript{36,38} Most of the PB methods are focused predominantly on the prediction of the miscibility between a drug molecule and coformers in an amorphous or supercool liquid phase. It was recently demonstrated that the crystallinity contribution to paracetamol and indomethacin cocrystal formation is the dominant one, and therefore is a major source of error for current virtual screening calculations.\textsuperscript{36} Therefore, the CSP-based approach should be the most attractive one for an accurate virtual coformer screening.

However, despite various algorithmic simplifications, the CSP-based approach is the most time-consuming model and may not be applicable to support projects with very tight timeline constraints. Therefore, further development of less accurate but faster approaches to rational coformer selection is of high interest. In this study, we report a combined ML and COSMO-RS method for fast coformer screening.

Statistical machine learning algorithms are broadly used to produce predictive empirical models from available experimental data.\textsuperscript{39} Various ML models were previously reported in application to drug development topics, including drug–β-cyclodextrin complexation,\textsuperscript{40} melting points,\textsuperscript{41–44} free energy and enthalpy of sublimation,\textsuperscript{39,41,45} solubility prediction,\textsuperscript{46} etc. Recently, the support vector machine (SVM) ML algorithm was proposed to guide selection of coformers for API cocrystallization.\textsuperscript{37}

The limitation of accuracy of ML models in application to crystalline state properties (including coformer selection) is related to the lack of descriptors that account for the long-range order (crystallinity) phenomenon.\textsuperscript{36} In addition, statistical models are known to provide poor predictions when extrapolated beyond the chemical space of the training set. A combined ML and physics-based model can potentially overcome these limitations via complementarity.

The COSMO-RS approach to coformer selection uses fluid thermodynamics to predict the miscibility of the drug molecule and coformer in a supercooled liquid phase as measured by excess enthalpy, \( H_{ex} \), or the \( f_{fit} \)coformer property.\textsuperscript{34,35} \( H_{ex} \) describes the change in enthalpy of the system upon mixing of the pure components in the stoichiometric composition. Crystallinity effects are not described by the COSMO-RS model.

2. Building the dataset

The dataset of experimental coformer screening was collected based on a combination of published reports\textsuperscript{32,37} (1160 observations) and in-house coformer screening studies for 9 APIs (151 observations). In total the dataset consists of 1311 cocrystallization experiments for 63 APIs and 198 coformers where approximately one third are experimentally observed cocrystals (positive hits) and two thirds of the results are from experiments where no cocrystal was found (negative hits, Table 1). A distribution of common molecular properties across the entire data set is presented in Fig. 1. The dataset was split randomly into training (80%) and testing (20%) sets, which allowed selection of the representative subsets of the original datasets.

An external validation set was created from reported cocrystallization studies of ibuprofen.\textsuperscript{47–50} It consists of screened coformers, of which 7 successfully formed cocrystals.

3. Computational approaches

3.1. ML approach

1D and 2D standard molecular descriptors and fingerprints (Pubchem and Morgan ECFP) were utilized for each coystal pair (1 API and 1 coformer), using RDKit,\textsuperscript{51} PaDEL\textsuperscript{52} and Mordred\textsuperscript{53} cheminformatics toolkits. After the removal of the zero-variance and highly correlated features, a total of 969 features per API-coformer pair remained. The experimentally observed cocrystals were labelled “1”. Label “0” was assigned to experimental results where cocrystallization was not observed.

The classification random forest (RF) ML algorithm,\textsuperscript{54} as implemented in version 0.23.2 of the scikit-learn package, was utilized to predict the probability of cocrystallization between API–coformer pairs.\textsuperscript{55} RF is one of the most popular and widely used machine learning algorithms because it provides reliable performance across a wide range of predictive modeling problems.\textsuperscript{56} RF is an ensemble of \( n \) estimators unpruned decision trees created by using max_samples bootstrap samples of the training data and a random subset of max_features variables to define the best split at each node. A balanced class weighting was used to account for the potential bias introduced by having a greater number of unsuccessful cocrystal experiments in the training set. The best-performing parameters for the RF model were

<table>
<thead>
<tr>
<th>Data points</th>
<th>Grecu et al.\textsuperscript{33}</th>
<th>Wicker et al.\textsuperscript{37}</th>
<th>In-house</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive hits</td>
<td>141 (28.0%)</td>
<td>254 (38.7%)</td>
<td>28 (18.5%)</td>
<td>423 (32.3%)</td>
</tr>
<tr>
<td>Negative hits</td>
<td>362 (72.0%)</td>
<td>403 (61.3%)</td>
<td>123 (81.5%)</td>
<td>888 (67.7%)</td>
</tr>
</tbody>
</table>
determined by a 10-fold cross-validation (CV) on the entire training dataset. The RF prediction of cocrystallization probability (“prob”) for an API–coformer pair was made by counting the fraction of trees in the forest that vote for the cocrystal formation class.

An approach to determine an applicability domain of the ML model has been implemented based on the mean Euclidean distance of 5 nearest training set points in the chemical space of the descriptors and fingerprints used in ML model building. Only features with higher than zero importance were utilized for AD determination. They were normalized to avoid the effect of the scale. The AD allows defining whether the ML model can be safely used for a reliable prediction of the cocrystallization probability of a new API–coformer pair. About 80% of compounds in the test set and all compounds in the external validation set were determined to be inside the AD.

3.2. COSMO-RS approach

COSMO-RS screening was performed using COSMOTHERM software.57 Up-to 10 solvent conformers were generated by the RDKit toolkit with an RMSD cutoff value of 0.5 Å. The generated conformations were further optimized in a perfect conductor medium with the Turbomole package58 using the BP86 density functional59–61 with the TZVP62 basis set (BP/TZVP/COSMO-RS level of theory).

3.3. Model performance estimation

The model performance was evaluated using the predictions made for the test and external validation sets. A performance measurement was performed based on receiver operator characteristic (ROC) curves. A ROC curve plots the true positive vs. false-positive (FP) prediction rate for a binary classifier as its discrimination threshold is varied from small to higher values. The area under the curve (AUC) measures the overall performance of the model. The AUC should always be higher than 0.5, indicating that the model is better than random selection. An ideal performance corresponds to AUC = 1.

4. Results and discussion

The best RF model parameters were estimated to be: class_weight = ‘balanced’, n_estimators = 400, bootstrap = true, random_state = 10, min_samples_split = 4, min_samples_leaf = 2, max_depth = 25, max_features = 20 and max_samples = none. The class_weight = balanced mode is used to mitigate an impact of training set class imbalances. It uses weights which are inversely proportional to the frequency of class observation to penalize differently a false classification from the minority and majority class. The random_state parameter controls the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking for the best split at each node. Minimum_samples_split defines the minimum number of samples required to split a node. Max_feature defines the number of features to consider when looking for the best split. Minimum_samples_leaf defines the minimum number of samples required to be at a leaf (terminal) node. Max_depth defines the maximum depth of a tree. Max_samples = none makes the sample size for bootstrapping the same as the training set.

The list of the most important descriptors is presented in Fig. S1.†

The 10-fold cross validation of the training set resulted in a balanced accuracy of predictions of 0.79 with a 95% confidence interval of 0.72–0.85. The balanced accuracy is defined as: $\frac{1}{2} \left( \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right)$, where TP, TN, FP and FN denote the number of true positive, true negative, false positive and false negative predictions, respectively. The threshold value of predicted probability of prob = 0.42 was selected based on the analysis of the balanced accuracy of the cross-validation results (Fig. S2†).
A good overall performance of the RF model is reflected by the ROC curve with AUC = 0.86 for the test set (Fig. 2).

The confusion matrix of the test set predictions is presented in Table 2. The predictive accuracy of the model is 79%, which is in line with the training set cross-validation performance. The precision and recall values correspond to 66% and 76%, respectively. This reflects the model tendency towards reducing a number of false negatives (observed co-crystals which are incorrectly predicted as unlikely to be formed).

In contrast to the ML model, the COSMO-RS model is physics-based and was not trained on the experimental results. Therefore, the comparison between the COSMO-RS and ML models should be performed on the data outside the ML training set. A side-to-side comparison of the ML and COSMO-RS models was performed based on the datapoints from the internal test set, as well as on the external validation set, as discussed below.

The ML and COSMO-RS models perform reasonably well on the test set as reflected by the ROC curves with AUC values of 0.86 and 0.75, respectively (Fig. 3). The predictive power of the ML model was found to be superior.

The performance drops for both pure models when applied to the external validation set (Fig. 4a). While the ML model remains acceptable (AUC = 0.70), the COSMO-RS approach has an AUC value of 0.44, which is below the random guessing result. However, despite a reasonably good coformer ranking by the ML model (Fig. 4b), only two (out of two) true positive hits would have been selected based on threshold of prob > 0.42. At the same time, the COSMO-RS approach could provide selection of additional coformers from the top of the $H_{ex}$ ranking. Depending on the $H_{ex}$ threshold, experimentally determined successful coformers #8 and #14, which are missed by the ML model, could be selected for an experimental follow up.

The above results demonstrate reasonably good performances of the ML and COSMO-RS methods for rational coformer selection with some advantages of the former approach. The ML method may not be applicable to systems outside the AD, which would require unreliable extrapolation of the predictions beyond the chemical space of the training set. In contrast, the PB COSMO-RS model does not suffer from the AD limitations and focuses only on the prediction of the miscibility between a drug molecule and coformer. However, large and not uniform crystallinity contributions to co-crystallization are completely ignored by the model.36

Table 2  Confusion matrix of ML model predictions for the test set

<table>
<thead>
<tr>
<th>Confusion matrix</th>
<th>Pred. 1</th>
<th>Pred. 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>True 1</td>
<td>61</td>
<td>19</td>
</tr>
<tr>
<td>True 0</td>
<td>32</td>
<td>151</td>
</tr>
</tbody>
</table>

Fig. 3  Receiver operating characteristic curves and AUC values for the ML, COSMO-RS and ML + COSMO-RS predictions for the test set.

Fig. 4  a) ROC curves and AUC values for the ML, COSMO-RS and ML + COSMO-RS predictions for the ibuprofen external validation set; b) coformer ranking by the ML (top) and COSMO-RS (bottom) models for the external validation set. Red dots and crosses indicate successful and unsuccessful co-crystallization observations, respectively. The coformer numbering is presented in Table S1.†
Therefore, to mitigate the above limitations, we propose a combined ML + COSMO-RS approach that prioritizes ranking based on the prob values. The goal of the combined approach is to improve the selection of the top ranked coformers for the experimental follow up.

The combined model thresholds and the ranking scheme ordering of the predictions were simultaneously scanned to maximize an average AUC value of the 10-fold CV of the training set. Top coformer ranking is performed based on the ML cocrystallization probability higher than 0.62. This is followed by the coformers with prob > 0.42 and $H_{\text{ex0}}$ lower than a threshold value of $H_{\text{ex0}} = -4.4$ kJ mol$^{-1}$, as ranked by decreasing prob values. After that, the coformers are ranked according to the scheme presented in Table S2. Virtual coformer ranking for cocrystal pairs which are outside the AD chemical space is performed by the COSMO-RS method.

The ML threshold of $\text{prob} = 0.62$ reflects the superior performance of the ML model at high prob values. The significantly negative $H_{\text{ex0}}$ value could be required to counterbalance the typically destabilizing crystallinity contribution, which is ignored in the COSMO-RS approach.$^{36}$

Successful validation of the ML + COSMO-RS approach is demonstrated by the ROC curves for the internal and external validation sets in Fig. 3 and 4a. A modest improvement of the ML + COSMO-RS model performance relative to the ML approach is reflected by an AUC value of 0.87 vs. 0.86 when applied to the test set (Fig. 3). The ML + COSMO-RS method exhibits superior performance on the ibuprofen coformer screening test set with an AUC value of 0.79 (Fig. 4).

5. Conclusions

In summary, we have presented a novel ML model for virtual coformer screening which provides reasonably good performance when applied to the test and external validation sets. To mitigate the potential limitations of the statistical approaches (such as poor extrapolation beyond the training set chemical space), we proposed a combination of the ML model with the physics-based COSMO-RS approach. Similarly, the statistical approach can “learn” some of the contributions that are poorly described by the COSMO-RS approach. In this way the two techniques are complementary to one another. The ML + COSMO-RS model demonstrated superior performance when applied to the test and external validation sets. The new combined model is relatively fast and accurate and should provide a good alternative to the physics-based CSP-based approach to support pharmaceutical projects with tight timeline and budget constraints.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the Shenzhen Special Fund for the Development of Strategic Emerging Industries.

Notes and references
