Prediction of Melting Point and Fusion Enthalpy of Cocrystal by Machine Learning combined with molecular informatics

Yingquan Hao, Yusuke Shimoyama*

Department of Chemical Science and Engineering, Tokyo Institute of Technology
2 Chome-12-1 Ookayama, Meguro, Tokyo 152-8550, Japan
* To whom correspondence should be addressed.
E-mail: yshimo@chemeng.titech.ac.jp Tel: +81-3-5734-3285 FAX: +81-3-5734-3285

Abstract

In the past decades, the solubility of the pharmaceuticals become lower. To solve this problem, the cocrystal have been considerate. By forming a new crystal structure with an additive which is called as coformer (CF), the thermodynamic property like melting point, fusion enthalpy of the pharmaceutical crystal can be modified. But by now, the melting point, fusion enthalpy of the cocrystal is mostly collected by experiments. So, in this research, we will try to build a machine learning model which can predict melting point and fusion enthalpy of cocrystal only from the molecular information of pharmaceutical (API) and coformer.

To make the molecular understandable for computer, we use the statistical charge distribution calculated by quantum calculation based on Conductor-like Screening Model (COSMO) as the descriptor for API and coformer. In additional, because the fusion enthalpy and melting is strongly related with energy, the sigma-profile of the API and coformer is input as a heat map matrix. The fusion enthalpy and melting point of the API and coformer are also inputted. For machine learning, random forest regression is applied.

About 165 data sets are collect from the published paper. As the final result, the correlation coefficient R, the mean relative error (MRE), and mean absolute error (MAE) for each test datasets are as below. For the melting point prediction, random forest regression model gives out the R = 0.57, and the MRE =5.9 %, MAE=24.8 K in cross-validation data. And for fusion enthalpy prediction, random forest regression model shows best performance, give out a R = 0.61, MRE =35.92% and the MAE=13.3 kJ mol^{-1} in cross-validation data. For both cases, the R is higher than 0.5 which indicating correlation between prediction and experiment. And it is much higher than the published prediction model only based on the molecular weight, fusion enthalpy, melting point of API and CF, indicating the importance of the datasets and the consideration about the physical rule in machine learning model.

Keywords

Cocrystal, Machine Learning, Chemical Calculation

1. Introduction

In 2017, Maleki and his coworkers reported about 90% of the candidate for active pharmaceutical ingredients (API) is classified as poor water soluble during the development. [1]This indicates the dissolution behavior of the API may become a bottle neck in pharmaceutical development. Because the solubility and the dissolution rate of API is a key factor in absorption process to human body, and poor dissolution properties always lead to low bioavailability. [2] To solve this problem, cocrystallization have been considerate. By introduction of new component which is normally called as coformer (CF) into the lattice of the API, a kind of new crystal structure can be obtained, and the dissolution property such as the dissolution rate and solubility can be modified. [3, 4]. But for some cases, the solubility of the API also can be lowered by the cocrystal formation. [5] Based on the general understand of the solid-liquid equilibrium showed as following equation (1), besides of the interaction of API and CF with the solvent, the thermal properties like fusion enthalpy (H_{fus}) and the melting point (T_m) of the cocrystal also play an important role of the dissolution behavior. [6] But by now, there is only few models have been reported for prediction of the H_{fus} and T_m for cocrystal.

$$lnx_i = \frac{H_{fus}}{R}\left(\frac{1}{T_m} - \frac{1}{T}\right) - ln\gamma_i \qquad (1)$$

In 2018, Gamidi and his coworker use the artificial neural networks (ANN) successfully predict the T_m of the cocrystal based on the molecular weight, binding energy, and melting point of API and coformer with a mean relative error (MRE) about 6.26 %. [7] Following with this research, in 2020, Gamidi report a new method which can use the ANN to predict the T_m and H_{fus} only from the molecular weight, fusion enthalpy, and melting point of the API/CF, and finally get a R^2 over 0.98. [6]. But for the both of the research, the dataset used is quite limited, 51 kinds of compounds and 66 pairs of cocrystals for the research in 2018, and only 8 compounds and 30 pairs of cocrystals for the research in 2020. Moreover, the model in 2020 only input the molecular weight, fusion enthalpy, and melting point of the API/CF for the prediction of property of cocrystal, indicating there is no interaction need to be considerate between API and CF in cocrystalization. This actually do not follow the general understanding of physical chemistry rule.

So, in this research, we enlarge the database of H_{fus} and T_m of cocrystal to a data size with 110 kinds of compounds and 165 pairs of cocrystals. And the model reported by Gamidi in 2020 (NEAT6) is reproduced on the enlarged database to recheck the performance. [6] In addition, we think the interaction between the API and CF play an important role in the thermal properties of the cocrystal. Based on this, the sigma-profile calculate by COnductor-like Screening MOdel (COSMO) of the API and CF is added as a heat map to the input of the machine learning model for T_m and H_{fus} prediction of cocrystal to improve the precision and general ability of the machine learning based prediction model for the thermal properties of the cocrystal.

2. Methods & Modeling

2.1. Data extraction from literatures

H_{fus} and T_m of 165 pairs of cocrystal and 110 kinds of single compounds were collected from 138 literatures with following rules. 1). The value for most stable polymorph is used, if polymorph is reported. 2). The value is averaged from several literature, if it is reported more than once.

2.2. Calculation of sigma-profile by COSMO

A quantum calculation based on conductor-like screening model (COSMO) [8] was applied for the calculation of molecular spatial screening charge density and molecular surface area of API and coformer molecules. COSMO calculation was conducted by a software TURBOMOLE 6.5. In brief, the molecular calculated in COSMO was putted into a cage built by solvent-accessible surfaces of the molecular. Then the geometry of the molecular will be optimized with b3-lyp DFT function in the environment with set dielectric constant, in this case it is infinite. Next, the induced charges by the molecular in cage will be calculated. Finally, these resulting surface charges will be transformed to one-dimensional distribution, sigma-profile, which shows the information about the area distribution of the surface segment within specific charge density range on the solvent-accessible surfaces of calculated molecular. This sigma-profiles calculated from quantum chemistry have showed its good performance in calculation of chemical properties, screening of cocrystal. Because it successfully captures the molecular feature such as hydrogen bond donor and acceptor, and hydrophobic interactions. [9]

2.3. Implement of machine learning model

The prediction model based on molecular weight, fusion enthalpy, melting point of API/CF reported by Gamidi in 2020 is reproduced following the setup as the literature. [6]

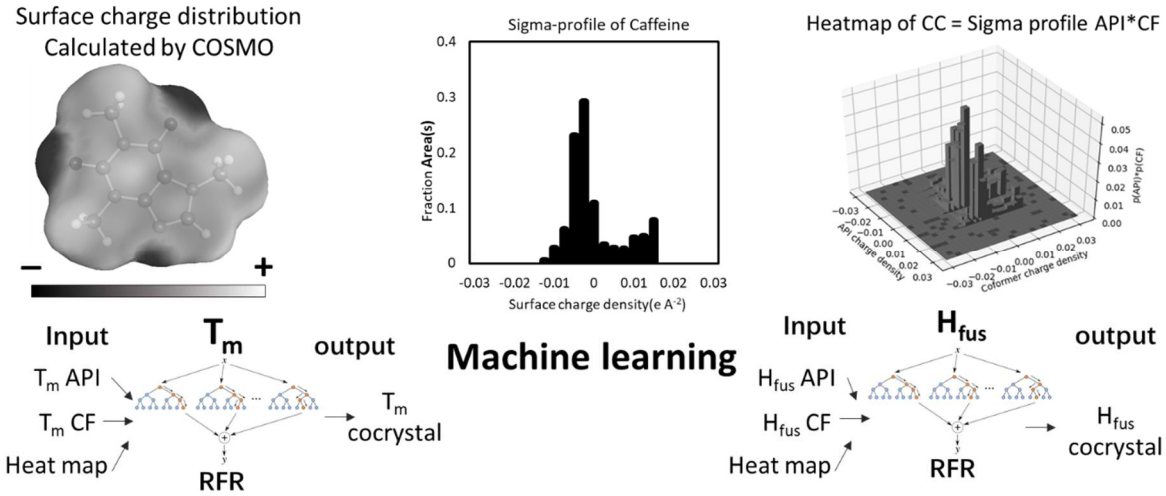

Fig. 1. Sigma-profile form COSMO and the model flow of the COSMO-RFR

For our model, COSMO-RFR, random forest regression (RFR) is used. Random forest is a kind of representative ensemble machine learning algorithms. It constructs a strong regressor

by an ensemble of individual decision trees under the frame of bagging. [10] In additional, it is chosen because its strong advantages in processing the nonlinear relationship between input and output. [11]

As for the input of COSMO-RFR, the H_{fus} of API and CF are inputted after multiplied by the mol ratio in the formed cocrystal for H_{fus} prediction. And for the melting point prediction of the cocrystal, T_{m} of API and CF are also multiplied by the mol ratio in the formed cocrystal. In addition to the property of pure API and CF, for the both of the prediction model, a heat map of sigma-profile from API and CF is inputted to describe the interaction between API and CF, with the following formulation. The detail of the model flow is showed in Fig.1.

$$Heatmap_{cocrystal}[i][j] = area_{API}[i] * area_{CF}[j] \qquad (2)$$

3. Result and discussion

The model is evaluated with 10-fold cross validation in the training and test datasets for the T_{m} and H_{fus} by both of the model (NEAT6 and COSMO-RFR). The graphs plotting the prediction versus the experimental value for H_{fus} and T_{m} are showed in Figs. 2H and 2T, respectively. Also, the mean absolute error (MAE), mean relative error (MRE) are used to check the value error of the prediction. And the relative coefficient R is used to check if the prediction model catches the overall tendency of the thermal properties of the cocrystal. The result is showed in Table.1.

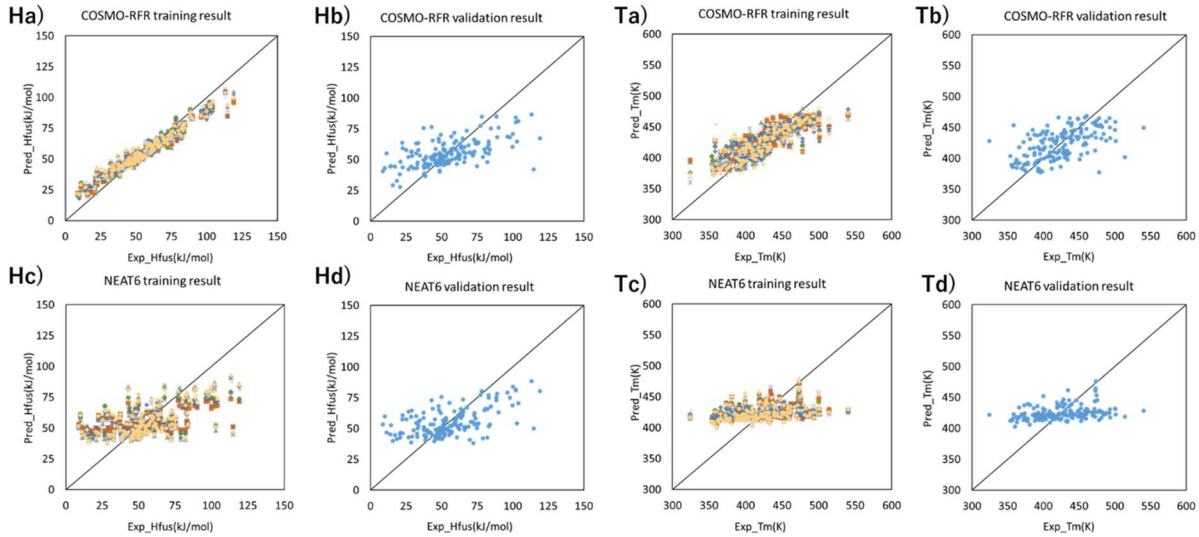

Fig. 2. The prediction value versus experimental value for Ha). H_{fus} by COSMO-RFR in training datasets. Hb). H_{fus} by COSMO-RFR in validation datasets. Hc). H_{fus} by NEAT6 in training datasets. Hd). H_{fus} by NEAT6 in validation datasets. Ta), Tb), Tc), Td) same prediction for T_{m}.

In Figs. 2Ha and 2Hc, it can be seen obviously that the COSMO-RFR model can fit well with the training data, while the NEAT6 model fit the training data much poorer. This may due to that the NEAT6 only offer six inputs for the prediction and also the model is too shallow to

simulate the function link the input and output. As mentioned in Table.1, the MAE is 5.28 kJ mol^{-1}, MRE is 14.04 % and *R* is 0.97 for the COSMO-RFR in training datasets which is much better than NEAT6, while MAE is 14.10 kJ mol^{-1}, MRE is 39.25 % and *R* is 0.57 for NEAT6. Moreover, the validation result of COSMO-RFR and NEAT6 for H_{fus} prediction are showed in Figs. 2Hb 2Hd. the distribution of the data points of NEAT6 in validation data sets is very similar to the training data sets, while the prediction precision of the COSMO-RFR clearly decrease in validation. But the COSMO-RFR still show higher precision because of considering the physics chemical interaction between the API/CF pair. The MAE is 13.30 kJ mol^{-1}, the MRE is 35.92 % and the *R* is 0.61 for the COSMO-RFR while the MAE is 14.35 kJ mol^{-1}, the MRE is 40.20 and the *R* is 0.57 for the NEAT6.

Table 1

Performance of the NEAT6 and COSMO-RFR in prediction of H_{fus} and T_m

Fusion enthalpy (H_{fus})						
	train-MAE	train-MRE	train-R	validation-MAE	validation-MRE	validation-R
COSMO-RFR	5.28 kJ mol^{-1}	14.04 %	0.97	13.30 kJ mol^{-1}	35.92 %	0.61
NEAT6	14.10 kJ mol^{-1}	39.25 %	0.57	14.35 kJ mol^{-1}	40.20 %	0.57
Melting point (T_m)						
	train-MAE	train-MRE	train-R	validation-MAE	validation-MRE	validation-R
COSMO-RFR	16.50 K	3.90%	0.85	24.80 K	5.90 %	0.57
NEAT6	30.30 K	7.20%	0.34	30.08 K	7.10 %	0.32

Figs. 2Ta 2Tc show the prediction result of T_m in training datasets versus the experimental for both of two models. Different from the fusion entropy case, the COSMO-RFR model fit the training datasets poorer, while the NEAT6 model nearly give out an average value for the prediction of the T_m. This indicate it is even impossible to training the model for the NEAT6 with only six inputs. As for the COSMO-RFR, even the interaction is considerate, but since the T_m is not only related to the interaction energy, but also related to the inconsiderate entropy change corresponding with the flexibility of the molecular structure according to the following equation. So, the prediction precision dropped compared with the H_{fus} modeling.

$$\Delta S = \frac{\Delta H}{T} \qquad (3)$$

As mentioned in Table.1, the MAE is 16.50 K, MRE is 3.90 % and R is 0.85 for the COSMO-RFR in training datasets which is still better than NEAT6, while MAE is 30.03K, MRE is 7.20 % and R is 0.34 for NEAT6. Moreover, the validation result of COSMO-RFR and NEAT6 for T_m prediction are showed in Figs. 2Tb 2Td. The distribution of the data points of NEAT6 in validation data sets is very similar to the training data sets. For the both of the training and

validation datasets, NEAT6 prefer to give an average prediction instead of a value related to the experimental data. As for the COSMO-RFR, even the performance dropped due to no consideration about the entropy change during the melting, the COSMO-RFR still show higher precision because of considering the physics chemical interaction between the API/CF pair than NEAT6. The MAE is 24.80 K, the MRE is 5.90 % and the R is 0.57 for the COSMO-RFR while the MAE is 30.08 K, the MRE is 7.10 % and the R is 0.32 for the NEAT6.

4. Conclusion

In this research we enlarge the database for prediction of the T_m and H_{fus} of the cocrystal, and compare our model COSMO-RFR with the reproduce model NEAT6. By introducing the interaction between API and CF by heat map of the its sigma profile from COSMO in COSMO-RFR, we successfully improve the performance of the machine learning based prediction model for the thermal properties of the cocrystal.

References

[1] A. Maleki, H. Kettiger, A. Schoubben, J.M. Rosenholm, V. Ambrogi, M. Hamidi, Journal of Controlled Release, 262 (2017) 329-347.

[2] M. Guo, X. Sun, J. Chen, T. Cai, Acta Pharmaceutica Sinica B, (2021).

[3] A. Shevchenko, L.M. Bimbo, I. Miroshnyk, J. Haarala, K. Jelínková, K. Syrjänen, B. van Veen, J. Kiesvaara, H.A. Santos, J. Yliruusi, International journal of pharmaceutics, 436 (2012) 403-409.

[4] A. Shayanfar, K. Asadpour-Zeynali, A. Jouyban, Journal of Molecular Liquids, 187 (2013) 171-176.

[5] N.R. Goud, R.A. Khan, A. Nangia, Modulating the solubility of sulfacetamide by means of cocrystals, CrystEngComm, 16 (2014) 5859-5869.

[6] R.K. Gamidi, Å.C. Rasmuson, Crystal Growth & Design, 20 (2020) 5745-5759.

[7] G. Rama Krishna, M. Ukrainczyk, J. Zeglinski, Å.C. Rasmuson, Crystal Growth & Design, 18 (2018) 133-144.

[8] A. Klamt, G. Schüürmann, Journal of the Chemical Society, Perkin Transactions 2, (1993) 799-805.

[9] A. Klamt, F. Eckert, M. Hornig, M.E. Beck, T. Bürger, Journal of computational chemistry, 23 (2002) 275-281.

[10] I. Miyazato, S. Nishimura, L. Takahashi, J. Ohyama, K. Takahashi, The journal of physical chemistry letters, 11 (2020) 787-795.

[11] L. Auret, C. Aldrich, Minerals Engineering, 35 (2012) 27-42.