Supplementary Material: Machine Learning to Instruct Single Crystal Growth by Flux Method*

Tang-Shi Yao(姚唐适)[1,2§], Cen-Yao Tang(唐岑瑶)[1,2§], Meng Yang(杨萌)[1,2§], Ke-Jia Zhu(朱恪嘉)[1,2], Da-Yu Yan(闫大禹)[1,2], Chang-Jiang Yi(伊长江)[1,2], Zi-Li Feng(冯子力)[1,2], He-Chang Lei(雷和畅)[4], Cheng-He Li(李承贺)[4], Le Wang(王乐)[1,2], Lei Wang(王磊)[1**], You-Guo Shi(石友国)[1,2**], Yu-Jie Sun(孙煜杰)[1,3,5**], Hong Ding(丁洪)[1,3,5]

[1]Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing 100190

[2]University of Chinese Academy of Sciences, Beijing 100049

[3]CAS Centre for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing 100049

[4]Department of Physics and Beijing Key Laboratory of Opto-electronic Functional Materials and Micro-nano Devices, Renmin University of China, Beijing 100872

[5]Songshan Lake Materials Laboratory, Dongguan 523808

* Supported by Ministry of Science and Technology and Ministry of Education of China (2016YFA0401000, 2015CB921000), the National Natural Science Foundation (11574371, 11774399, 11774398), the National Key Research and Development Program of China (No.2017YFA0302901), Beijing Natural Science Foundation(Z180008), and the Strategic Priority Research Program of Chinese Academy of Sciences(No. XDB28000000).

§Tang-Shi Yao, Cen-Yao Tang and Meng Yang contributed to this work equally.

**To whom correspondence should be addressed. Email: wanglei@iphy.ac.cn; ygshi@iphy.ac.cn; yjsun@iphy.ac.cn

(Received 30 April 2019)

Text A: Dataset and Features selection.

The data comes from two laboratories that grow single crystals. The first group of data comes from the Shi Youguo's group of Institute of Physics of Chinese Academy of Sciences, and the second group comes from the Lei Hechang group of Renmin University of China. The Group-I contains ternary compounds and the Group-II contains binary and ternary compounds. After excluding reactions with incomplete laboratory notebook entries, we got 775 complete experimental data from Group-I, and 649 data with 272 different kinds of samples remained after removing duplicate conditions. Group-I contains 64 elements. For further verification, we also obtained 163 data from Group-II, and 115 remained after removing duplicates. Crystal sizes were coded with the labels 0 for no solid product, 1

for single crystals having standard single-crystal x-ray diffraction data. The data recorded in the laboratory are the composition and ratio of the raw materials, the heating program for growth and the quality of the grown samples. The features we extracted from the experimental data are listed in Table S3. We have selected three characteristics: (i) the physicochemical properties and the constituent elements of the sample, including electronegativity, atomic radius, atomic mass, the position of the element in the periodic table, and the space group of the sample, as shown in Table S1, (ii) features extracted from the phase diagrams according to the selected flux and ratio as listed in Table S2, (iii) characteristics extracted from the experimental features as listed in Table S3.

There are 107 original features. In order to match the data volume and reduce the noise, we need to reduce the dimension of the features: univariate feature selection; correlation analysis between features, reducing some features; using model features to reduce unimportant characteristics, principal component analysis.

Text B: Model evaluation and reliability analysis.

We use machine learning methods such as SVM, decision trees, RF and GDBT to study this problem. We used accuracy, f1-scores, recall rates for successful samples, and accuracy for successful sample predictions to evaluate this model. The accuracy of the model and the f1-score represent the learning ability of the model. The higher the scores of these two indicators, the stronger the model's analysis ability. A high recall rate for successful sample predictions indicates that it is not easy for a model to misjudge a successful condition as a failure condition. The precision on successful samples represents success rate for conditions provided by the model.

The data we used to test the validity of the model is outside the dataset which was used to construct the model. And types of samples included in the test dataset are different from types of samples used to construct the model. This ensures that the scores of our outcomes can assess the migration ability of the model in a new sample. The model was tested against the known data using a standard 1/4-test and 3/4-training data split. The scale of our test set is large enough to make the statistical results confident. Moreover, a 10-fold cross-validation was used to analyze the model. The single SVM model used to predict experimental results has an accuracy of 80% in describing all the reaction types in its test-set data. The average accuracy over 10 training/validation split is 73%. As shown in Fig. S2, the larger the amount of data we have, the better the model performs. Getting a larger amount of data is one way we can improve the performance of our models. The credibility of the statistical results of our predictions on the new sample is higher as the amount of data increases. According to our current scale of data, the statistical results of our model on the new sample are highly reliable.

Code availability Part of the code for this project is available at supplementary materials.

Table S1. Physicochemical properties of samples and sample constituents.

Property name	Description
vapor_100t	vapor pressure of flux M when temperature
vapor_kt	vapor pressure of flux N when temperature is
num_A	atomic number of element A
mass_A	atomic mass of element A
melting_point_A	melting point of element A
boiling_point_A	boiling point of element A
electronegativity_A	atomic electronegativity of element A
metalcrystal_A	crystal structure of element A
row_A	the number of cycles of element A in the
cycle_A	the family of cycles of element A in the
radius_A	atomic radius of element A
num_B	atomic number of element B
mass_B	atomic mass of element B
melting_point_B	melting point of element B
boiling_point_B	boiling point of element B
electronegativity_B	atomic electronegativity of element B
metalcrystal_B	crystal structure of element B
row_B	the number of cycles of element B in the
cycle_B	the family of cycles of element B in the
radius_B	atomic radius of element B
num_C	atomic number of element C
mass_C	atomic mass of element C
melting_point_C	melting point of element C
boiling_point_C	boiling point of element C
electronegativity_C	atomic electronegativity of element C
metalcrystal_C	crystal structure of element C
row_C	the number of cycles of element C in the
cycle_C	the family of cycles of element C in the
radius_C	atomic radius of element C
num_M	atomic number of flux M
mass_M	atomic mass of flux M
melting_point_M	melting point of flux M
boiling_point_M	boiling point of flux M
electronegativity_M	atomic electronegativity of flux M
metalcrystal_M	crystal structure of flux M
row_M	the number of cycles of flux M in the periodic
cycle_M	the family of cycles of flux M in the periodic
radiu_M	atomic radius of flux M
num_N	atomic number of flux N
mass_N	atomic mass of flux N
melting_point_N	melting point of flux N
boiling_point_N	boiling point of flux N

Property name	Description
electronegativity_N	atomic electronegativity of flux N
metalcrystal_N	crystal structure of flux N
row_N	the number of cycles of flux N in the periodic
cycle_N	the family of cycles of flux N in the periodic
radius_N	atomic radius of flux N
radius_A_M	atomic radius difference between element A
radius_B_M	atomic radius difference between element B
radius_C_M	atomic radius difference between element C
radius_A_N	atomic radius difference between element A
radius_B_N	atomic radius difference between element B
radius_C_N	atomic radius difference between element C
electronegativity_A_M	atomic electronegativity difference between
electronegativity_B_M	atomic electronegativity difference between
electronegativity_C_M	atomic electronegativity difference between
electronegativity_A_N	atomic electronegativity difference between
electronegativity_B_N	atomic electronegativity difference between
electronegativity_C_N	atomic electronegativity difference between
density_A	density of element A
density_B	density of element B
density_C	density of element C
density_M	density of element M
density_N	density of element N
density_A_M	atomic density difference between element A
density_B_M	atomic density difference between element B
density_C_M	atomic density difference between element C
density_A_N	atomic density difference between element A
density_B_N	atomic density difference between element B
density_C_N	atomic density difference between element C
density_mean	average density
density_M_std	the standard deviation of the density
density_N_std	The standard deviation of the density
density_std	standard deviation of all elemental densities
electronegativity_A_flux	atomic electronegativity difference between
electronegativity_B_flux	atomic electronegativity difference between
electronegativity_C_flux	atomic electronegativity difference between
radius_A_flux	atomic radius difference between element A
radius_B_flux	atomic radius difference between element B
radius_C_flux	atomic radius difference between element C
Space_Group	space group of samples

Table S2. Features extracted from binary phase diagrams.

Property	Description
ratio_phase_temh_A_M	In the binary phase diagram, when the temperature is the highest temperature, the ratio of M dissolved A
ratio_phase_teml_A_M	In the binary phase diagram, when the temperature is the lowest temperature, the ratio of M dissolved A
ratio_phase_temh_B_M	In the binary phase diagram, when the temperature is the highest temperature, the ratio of M dissolved B
ratio_phase_teml_B_M	In the binary phase diagram, when the temperature is the lowest temperature, the ratio of M dissolved B
ratio_phase_temh_C_M	In the binary phase diagram, when the temperature is the highest temperature, the ratio of M dissolved C
ratio_phase_teml_C_M	In the binary phase diagram, when the temperature is the lowest temperature, the ratio of M dissolved C
ratio_phase_temh_A_M	In the binary phase diagram, when the temperature is the highest temperature, the ratio of N dissolved A
ratio_phase_teml_A_M	In the binary phase diagram, when the temperature is the lowest temperature, the ratio of N dissolved A
ratio_phase_temh_B_M	In the binary phase diagram, when the temperature is the highest temperature, the ratio of N dissolved B
ratio_phase_teml_B_M	In the binary phase diagram, when the temperature is the lowest temperature, the ratio of N dissolved B
ratio_phase_temh_C_M	In the binary phase diagram, when the temperature is the highest temperature, the ratio of N dissolved C
ratio_phase_teml_C_M	In the binary phase diagram, when the temperature is the lowest temperature, the ratio of N dissolved C
phase_A	the lowest temperature at which element A is completely dissolved
phase_B	The lowest temperature at which element B is completely dissolved
phase_C	The lowest temperature at which element C is completely dissolved

Table S3. Experimental conditions for growing single crystal.

Property	Description
ratio_A	molar ratio of raw material A
ratio_B	molar ratio of raw material B
ratio_C	molar ratio of raw material C
ratio_M	molar ratio of raw material M
ratio_N	molar ratio of raw material N
tem_h	maximum temperature
tem_l	centrifugal temperature
v	cooling rate
time	highest temperature residence time
temd	cooling rate

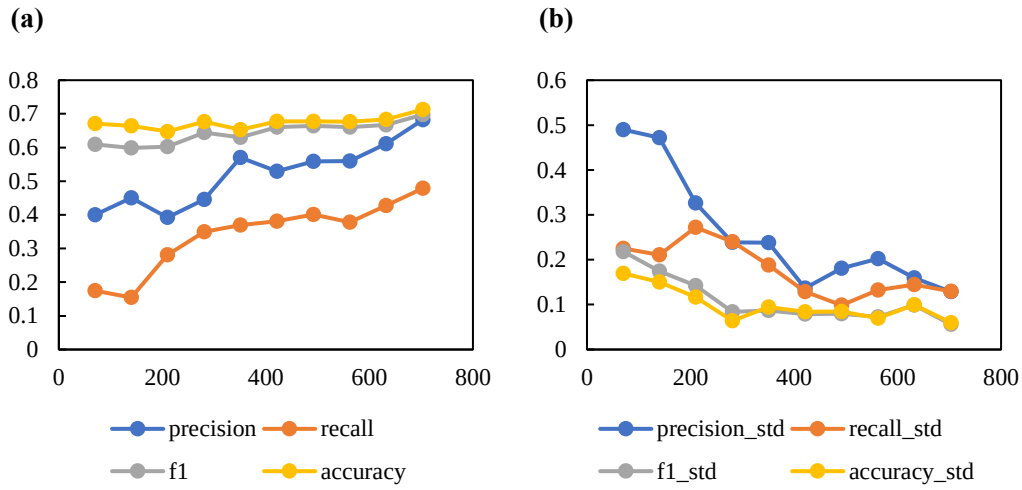

Fig. S1. The impact of data volume on the SVM models on Group I. (a) SVM learning curve plot on Group. Reported test set accuracy, f1-scores, recall on success samples and precision on success samples, averaged over 10 training/test splits where each split contains different kinds of samples. (b) The variance of accuracy, f1-scores, recall on success samples and precision on success samples, averaged over 10 training/test splits where each split contains different kinds of samples.