ELSEVIER

Machine learning-guided synthesis of advanced inorganic materials

Bijun Tang [1], Yuhao Lu [2], Jiadong Zhou [1], Tushar Chouhan [2], Han Wang [1], Prafful Golani [1], Manzhang Xu [1], Quan Xu [3,*], Cuntai Guan [2,*], Zheng Liu [1,4,5,*]

[1] School of Materials Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
[3] State Key Laboratory of Heavy Oil Processing, China University of Petroleum (Beijing), Beijing, China
[4] CINTRA CNRS/NTU/THALES, UMI 3288, Research Techno Plaza, 50 Nanyang Drive, Border X Block, Level 6, Singapore 637553, Singapore
[5] Chemistry and Materials Centre, Nanyang Environment and Water Research Institute, Singapore 637141, Singapore

Synthesis of materials with minimum number of trials is of paramount importance towards the acceleration of advanced materials development. The enormous complexity involved in existing multi-variable synthesis methods leads to high uncertainty, numerous trials and exorbitant cost. Recently, machine learning (ML) has demonstrated tremendous potential for material discovery and property enhancement. Here, we extend the application of ML to guide material synthesis process through the establishment of the methodology including model construction, optimization, and progressive adaptive model (PAM). Two representative multi-variable systems are studied. A classification ML model on chemical vapor grown MoS_2 is developed, capable of optimizing the synthesis conditions to achieve a higher success rate. And a regression model is constructed on the hydrothermal-grown carbon quantum dots, to enhance the process-related properties such as the photoluminescence quantum yield. The importance of synthesis parameters on experimental outcomes is particularly extracted from the constructed ML models. Furthermore, off-line analysis shows that enhancement of the experimental outcome with minimized number of trials can be achieved with the effective feedback loops in PAM, suggesting the great potential of involving ML to guide new material synthesis at the beginning stage. This work serves as a proof of concept for using ML in facilitating the synthesis of inorganic materials, thereby revealing the feasibility and remarkable capability of ML in opening up a new promising window for accelerating material development.

Introduction

Material synthesis is always a challenging problem hindering the development of advanced inorganic materials. Complex synthesis not only entails large uncertainties but is also costly and time-consuming [1]. For example, two-dimensional (2D) materials have received substantial research interests in recent years attributed to their unique and fascinating properties [2–5], and chemical vapor deposition (CVD) is considered as one of the most promising methods to realize the controllable and scalable synthesis of these intriguing materials [6–8]. However, CVD process contains numerous variables like reaction temperature, chamber pressure, carrier gas flow rate, etc., significantly aggravating its unpredictability. Especially, early exploration of the optimal synthesis condition was solely driven by a laborious trial-and-error process, rendering extremely long development cycles. Additionally, a large number of trials

* Corresponding authors.
E-mail addresses: Xu, Q. (xuquan@cup.edu.cn), Guan, C. (CTGuan@ntu.edu.sg), Liu, Z. (z.liu@ntu.edu.sg).

 https://doi.org/10.1016/j.mattod.2020.06.010

together with expensive precursors and high energy consumption result in exorbitant research and development costs. Not only CVD, other multi-variable synthesis methods including hydrothermal, chemical vapor transport (CVT), atomic layer deposition (ALD) and molecular beam epitaxy (MBE), also have such issues. Therefore, an effective learning strategy towards optimizing and accelerating the synthesis of advanced inorganic materials, is urgently required.

Recently, machine learning (ML) methods have demonstrated great potential in substantially accelerating materials development, as shown in Fig. 1a. For instance, ML models have been applied for the discovery of new materials (perovskite halides [9,10], metallic glasses [11], shape memory alloys [12], inorganic–organic hybrid materials [13], etc.) and prediction of material properties (electronic properties of inorganic materials [14], grain boundary energies of crystalline materials [15], material phase transition [16], crystal structures [17], etc.), which are primarily within phase 2 of the material development. While phase 3, material synthesis, the critical step towards the final application of materials, remains less studied. Moreover, along with development of high-throughput first-principles computations, the need for efficient and controllable synthesis becomes even more pressing to cope with the dramatically growing volume of predicted and screened materials. Among the few pioneering studies of ML-guided synthesis, most of them focus on exploring the space and underlying mechanism of specific chemical reactions, instead of the synthesis of materials for practical applications [18–20]. Thus, it is timely to explore the capability of ML to guide the synthesis process of advanced materials.

In this work, to demonstrate the feasibility of optimizing and accelerating the synthesis process of materials through ML, we implement supervised ML on the CVD synthesis of 2D MoS_2, which is a promising candidate for numerous applications [2,3,21–23]. The paradigm is schematically depicted in Fig. 1b. The synthesis data are retrieved from archived laboratory notebooks from our laboratory. Our goal is to achieve the (1) quantitative understanding of the synthesis systems, (2) improvement of the experimental outcome and (3) acceleration of material

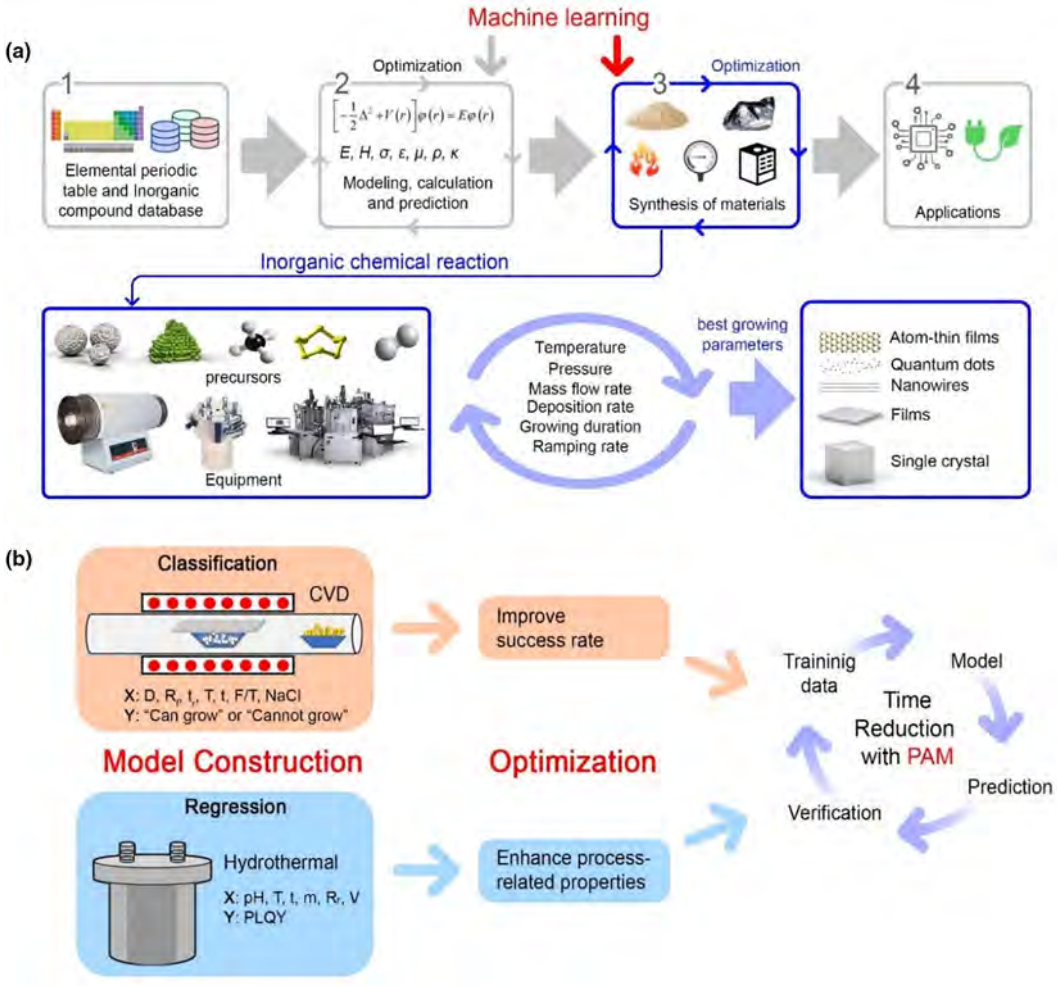

FIGURE 1

Schematic illustration of a paradigm for ML-guided synthesis of advanced inorganic materials. (a) The life cycle of materials development includes four phases: elements and compound database preparation, property prediction and optimization, materials synthesis, as well as practical application. As ML has demonstrated its great potential in phase 2, its feasibility in material synthesis scenario (i.e., phase 3) is investigated in this work. (b) Workflow to achieve the optimization and acceleration of inorganic material synthesis. Model construction, optimization and PAM are the three key steps, applicable to both classification and regression material synthesis scenarios.

synthesis with ML. A classification model is constructed, from which the influence of each synthesis parameter on the experimental outcome can be extracted and quantified, providing the experimentalists general guidance on parameter tuning for future experiments. The trained ML model is also capable of predicting the probability of successful synthesis given a set of CVD parameters and recommending the most favorable conditions. Progressive adaptive model (PAM) is further introduced to accelerate the development of new materials, which can maximize the experimental outcome and effectively reduce the number of trials. Most importantly, the principle demonstrated on CVD synthesis can be extended to other multi-variable synthesis methods, not only to improve the success rate, but also to enhance the process-related properties. A regression model is also successfully constructed on the hydrothermal-grown carbon quantum dots (CQDs), to enhance the property of photoluminescence quantum yield (PLQY).

ML-guided CVD synthesis with high success rate

In order to realize the controllable synthesis of advanced inorganic materials, the ambiguous relationship between various synthesis parameters and outcomes need to be understood. ML has great potential for unveiling such relationships through learning existing synthesis data, and then recommending optimal growth conditions with high success rate. Here, the CVD-grown MoS_2 is targeted not only because of its fascinating properties, but more importantly, the information obtained and methodology established with MoS_2 can be potentially extended to a full spectrum of CVD-grown materials including carbon nanotubes [24], graphene [25], oxides [26], nitrides [27], and transitional metal dichalcogenides (TMDs, with a general formula of MX_2 where M refers to transition metal and X refers to chalcogen atoms) [6]. They are synthesized in a similar manner: (1) feeding precursors (e.g., gas, liquid or solid based), (2) controlling/optimizing the macroscopic parameters (e.g., precursor flowing rate, temperature, pressure) and (3) obtaining the materials, as shown in Fig. 1.

Dataset. The CVD-grown MoS_2 dataset, containing 300 experimental data points, is collected from our archived laboratory notebook. The detailed synthesis process is presented in Methods. Among them, MoS_2 is successfully obtained in 183 experiments (61%), whereas the rest 117 experiments show negative results (39%). A binary classification problem in ML is hence formulated by defining "Can grow" as positive class and "Cannot grow" as negative class. Sample size of 1 μm is adopted as the boundary to classify the "Can grow" and "Cannot grow" categories. Growth of MoS_2 with sample size larger than 1 μm is considered as "Can grow" and smaller than 1 μm is considered as "Cannot grow". This criterion is based on the fact that it is hard to determine whether the point of interest is the sample or a nucleation site when its size is below 1 μm owing to the resolution limit of optical microscope (OM). Also, there is no practical use for such small sized MoS_2. Characterizations of a typical "Can grow" sample including OM, Raman spectroscopy and scanning transmission electron microscopy are shown in Fig. S1 in the Supporting Information.

Feature engineering. 19 features including gas flow rate, reaction temperature, reaction time, etc. are initially identified to describe the CVD process collectively (Table S1, Supporting Information). The initial feature set consists of two parts: synthesis process-related (i.e., CVD furnace parameters) as well as reaction-related (i.e., reactant information). After eliminating the fixed parameters and those with missing data, 7 features with complete records are retained and constitute the final feature set, which are also empirically considered as essential parameters for CVD-grown MoS_2 by the experimentalists [6,28,29]. The new feature set consists of distance of S outside furnace (D), gas flow rate (R_f), ramp time (t_r), reaction temperature (T), reaction time (t), addition of NaCl, and boat configuration (F/T) where F and T represent flat and tilted and boat refers to the container in which precursors are placed for CVD reaction. Detailed feature overview and histogram of the dataset over each feature are shown in Table S2 and Fig. S2, respectively, in the Supporting Information. Pearson's correlation coefficients are calculated to quantify the mutual information content between all pairwise features. For ML applications, it is desired to have features with minimum redundancy in information (Fig. 2a). Low linear correlations for most of the features indicate that independent and

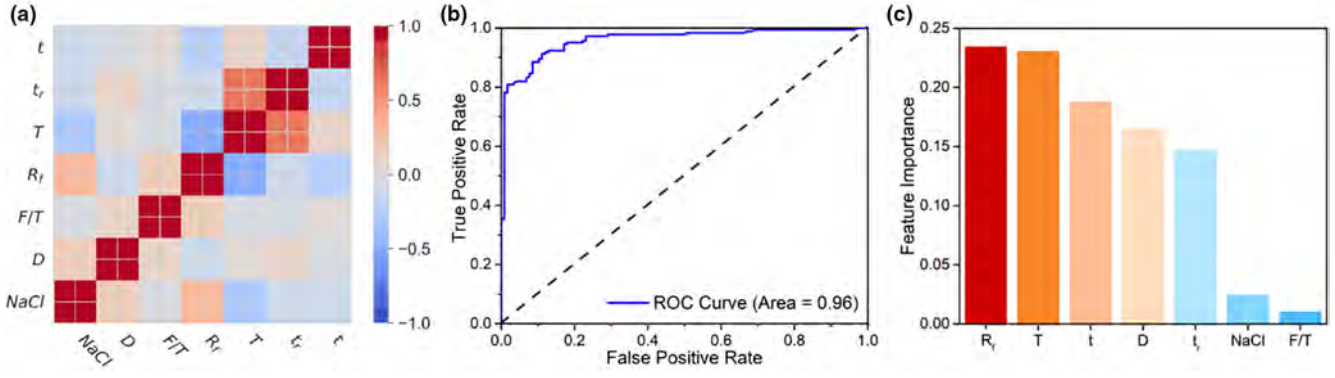

FIGURE 2

Model evaluation and interpretation of the CVD-grown MoS_2 dataset. (a) The heat map of the Pearson's correlation coefficient matrix among the selected features of CVD-grown MoS_2. (b) Receiver operating characteristic (ROC) curve of XGBoost-C. High AUROC unveils the great capability of the model to distinguish between two classes. (c) Feature importance retrieved from XGBoost-C that learns from all 300 data samples, with unique and consistent SHapley Additive exPlanations (SHAP) method. R_f and T are the two most important features.

informative features have been selected to form the essential feature set [10].

Model selection. Based on "no-free-lunch theorem" [30], there is no universally optimal algorithm for all problems. Thus, in this work, XGBoost classifier (XGBoost-C, a more powerful variant of gradient boosting decision tree; see Methods) [31], support vector machine classifier (SVM-C) [32], Naïve Bayes classifier (NB-C) [33], and multilayer perceptron classifier (MLP-C) [34] are employed on MoS_2 dataset for selecting the best model. These models have been successfully applied or considered in many material science problems, especially for those with small datasets [10,12,35,36]. Notably, considering the small dataset but possibly intricate relationship between the features and the outcome, both simple and complex models are considered. Each candidate model is evaluated with ten runs of nested cross validation to avoid overfitting in model selection [37]. Detailed working principle is illustrated in Fig. S3 in the Supporting Information. The whole dataset is shuffled in each run of nested cross validation, with the outer loop assessing the performance of the models on unseen datasets (ten-fold outer cross validation), and the inner loop conducting hyperparameter search and model fitting (ten-fold inner cross validation).

Detailed discussion of the characteristics of each model and the quantitative comparisons among models is provided in Fig. S4 in the Supporting Information, indicating that XGBoost-C reproduces the best agreement to the true synthesis outcomes and generalizes well to unseen data. Receiver operating characteristic (ROC) curve of XGBoost-C is presented in Fig. 2b, which reports the prediction performance of positive class (correctly versus incorrectly predicted) with all possible prediction thresholds (see Methods) [38]. Large area under ROC curve (AUROC) of 0.96 reflects the model's effectiveness in distinguishing between "Can grow" and "Cannot grow" classes. Moreover, the learning curve displayed in Fig. S5 in the Supporting Information shows the performance of the model with increasing number of epochs during training. The consistent performance of the model with a narrow gap between training and validation denotes that XGBoost-C is not overfitted to the training data, ensuring its good performance on unseen data. Therefore, XGBoost-C is chosen to learn the nonlinear mapping from CVD synthesis parameters to experimental outcome from the whole MoS_2 dataset, and subsequently make predictions on unexplored conditions.

Optimization of synthesis condition for higher success rate. Optimization involves distilling the importance of synthesis parameters on experimental outcome from the trained ML model, and applying the same model to obtain the optimal experimental conditions. The schematic of optimization process is provided in Fig. S6 in the Supporting Information. SHapley Additive exPlanations (SHAP) is used to quantitatively understand the CVD synthesis system through unveiling the intricate relationship between features and output captured in the obtained best model, XGBoost-C (see Methods). It is a unified approach to interpret ML models by using additive feature importance measures that is proven to be unique and consistent with human intuition [39]. As shown in Fig. 2c, the gas flow rate (R_f) plays the most important role in determining whether MoS_2 can be synthesized, followed by the reaction temperature (T) and reaction time (t). To interpret from the perspective of laboratory experiments, R_f is a very important growth parameter, which affects the exposure time and sulfur source controlling. Low R_f will dramatically decrease the deposition rate of precursors on the substrate, and thus make it difficult to grow MoS_2. On the other hand, MoS_2 can hardly be synthesized at very high R_f either, because high flow rate may cause instability during crystal growth and atoms do not get enough time to move into the right lattice position [40]. T is critical in determining the vapor pressure of the reactants and dominating the nucleation rate and the growth rate of MoS_2 grains. Moreover, the thickness of formed sample normally possesses a positive correlation with reaction time t. Thus, the appropriate selection of R_f, T and t is of great importance for the synthesis of atomic-layer 2D MoS_2 from both ML and experimental points of view [6,28,29]. This can also serve as a general guidance for the synthesis of other 2D materials, especially TMDs, due to their similar growing conditions as discussed above.

The optimal synthesis conditions of 2D MoS_2 are further identified with XGBoost-C in the unexplored search space, whose distribution is illustrated in Fig. S7 in the Supporting Information. In order to achieve this, the possible input range of each critical parameter is firstly defined as in Table S3 in the Supporting Information, resulting in 2,112,000 possible combinations in total. Next, XGBoost-C is applied to predict the "Can grow" probability of all the conditions. 10 synthesis conditions with the highest predicted probabilities are then tested in the laboratory with results shown in Table S4 in the Supporting Information. Detailed feature analysis of the recommended 10 conditions is provided in Fig. S8 in the Supporting Information. 2D MoS_2 is successfully synthesized under all 10 conditions, which substantially exceeds the 61% success rate in the original MoS_2 dataset, verifying the validity and effectiveness of the ML model.

Acceleration with progressive adaptive model (PAM). With the proven effectiveness of ML guidance in material synthesis, it is hypothesized that the early intervention of ML might lead to an enhanced experimental outcome and time reduction. Therefore, PAM is further proposed, which starts from a small initial dataset and evolves with iterative feedback loops. In another study, a similar adaptive design strategy has been successfully applied to identify the material composition with the optimal targeted property, in which case the authors focus on the best outcome achieved [12]. However, the proposed PAM aims to enhance the total experimental outcomes of continuous synthesis loops. Thus, the whole searching process is evaluated instead of assessing the best result achieved by any single loop. To directly compare the performance of human strategy-based and ML-guided synthesis, PAM is evaluated with the same CVD-grown MoS_2 dataset through off-line analysis. The existing dataset serves as a baseline representing human strategy.

The schematic of PAM is provided in Fig. 3a. Initially, N_1 synthesis conditions are randomly chosen and labeled by their respective synthesis outcomes extracted from the dataset. N_1 is determined such that there are at least ten samples in each class to draw the boundary between classes, in order to perform the ten-fold cross validation. XGBoost-C model is first trained on N_1 data and then used to predict the "Can grow" probability of the rest ($300 - N_1$) synthesis conditions, assuming the experi-

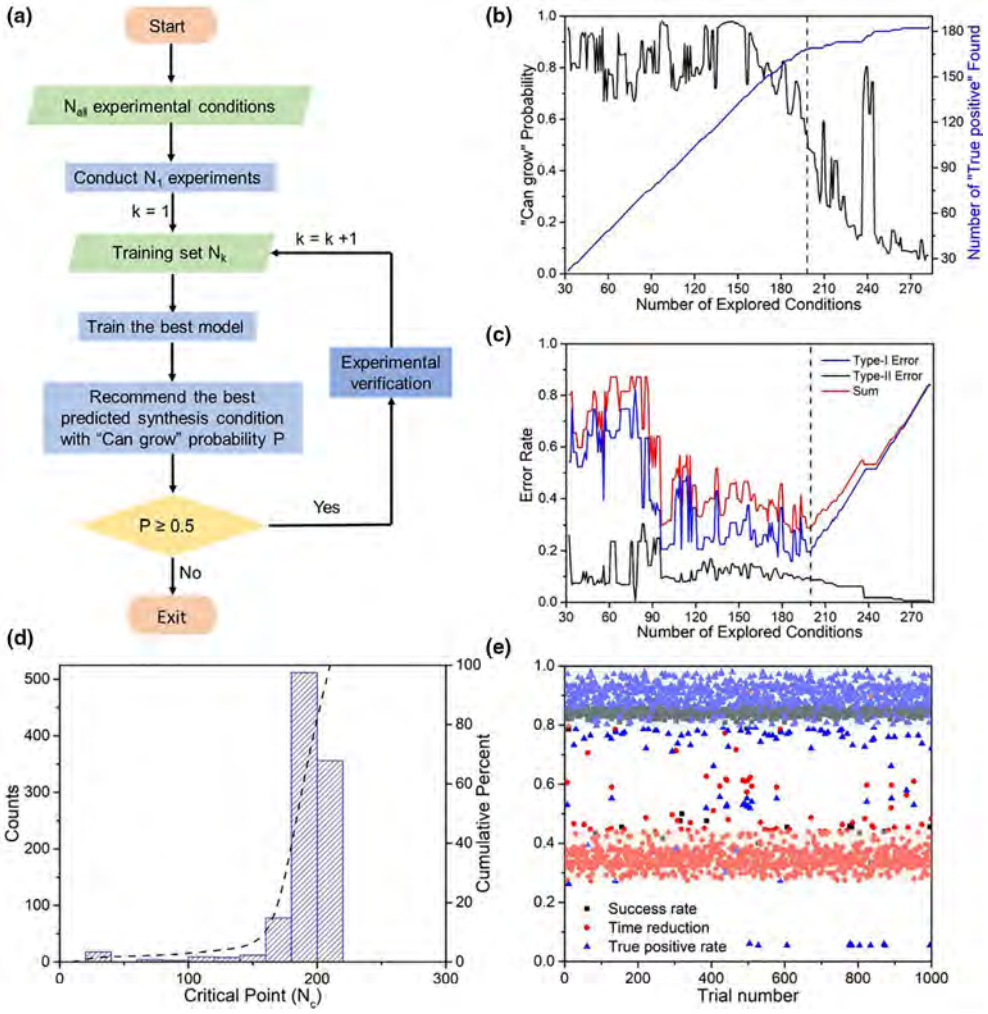

FIGURE 3

Schematic of PAM for accelerating inorganic synthesis and its overall performance on CVD-grown MoS_2. (a) Outline of the PAM workflow, displaying feedback loops and exiting condition. (b) Plot of the highest predicted "Can grow" probability versus the number of explored conditions, smoothed with median filter of window size 3 to highlight the trend. Blue line represents the number of "Can grow" samples found versus the number of explored conditions. (c) Plot of error rate of PAM versus the number of explored conditions, smoothed with median filter of window size 3 to highlight the trend. The vertical dashed lines in b and c indicate the critical point. (d) Distribution of the critical points of 1000 PAM trials. The critical points densely distribute around the mean of 189.28. (e) Plot of the success rate, time reduction and true positive rate on the whole dataset achieved in each PAM trial. Together with (b), it shows that PAM performs stably and consistently produces high success rate.

ments are yet to be conducted. One condition with the highest probability together with its true label is then augmented to the training set. The same steps are repeated in the subsequent loops. PAM stops at the critical point, N_C, where the "Can grow" probabilities of all (300 – N_C) conditions are predicted to be smaller than 50.0% for the first time (i.e., PAM predicts all the remaining conditions as "Cannot grow"). The success rate, true positive rate and time reduction based on all the experiments conducted are evaluated to assess the performance of PAM-guided materials synthesis (see Methods).

One typical trial of PAM is visualized and analyzed in Fig. 3b and c. The highest predicted "Can grow" probability of each trial is plotted versus the number of explored conditions, as shown in Fig. 3b. The highest probability drops below 50.0% at 193, thus N_C = 193. Among the 193 experiments, 166 produce "Can grow" samples, resulting in a success rate of 86.01%, with ~25% improvement from the original MoS_2 dataset. In addition, 35.67% time reduction with 90.71% true positive rate is achieved with PAM, at the cost of an error margin, where type-I and type-II errors equal to 15.89% and 10.56%, respectively (see Methods). The sum of type-I and type-II error reaches the lowest around N_C, verifying the validity of the choice of the critical point.

To further investigate whether the randomly selected initial training set affects the model's performance, PAM is repeated 1000 times on shuffled MoS_2 dataset. Validation of the selection of 1000 trials is provided in Fig. S9 in the Supporting Information. 1000 trials of PAM result in a distribution of N_C as shown in Fig. 3d, where N_C mainly clusters to the mean of 189.28 (±28.89). The respective success rate, time reduction and true positive rate of 1000 trials are calculated and presented in Fig. 3e. The success rate remains stably high at around 83.60%, with small variance (±5.57). Average of time reduction and true positive rate are 36.90% (±9.63) and 87.13% (±13.78), respectively. It is thus seen that PAM can help the experimentalist iden-

tify the optimal synthesis conditions quickly, and thus considerably reduce the time on empirical trials.

Based on the CVD-grown MoS_2 dataset, through model construction, optimization and PAM, our results above demonstrate that our proposed ML methodology can achieve high success rate and time reduction, and has great advantages in navigating complex multi-variable synthesis systems of inorganic materials.

To investigate the generalizability of the proposed method within the CVD system, 255 experimental data of WTe_2 are then retrieved from the archived laboratory notebook and fed to the XGBoost-C model. WTe_2 is another important member of 2D TMDs family, which is also one of the most promising type-II Weyl semimetals and topological materials [41,42]. Corresponding results are provided in Fig. S10 in the Supporting Information. As indicated in Fig. S10c, XGBoost-C performs very well on the WTe_2 dataset with high AUROC of 0.93. Feature importance extracted from the model (Fig. S10d), suggests that H_2 flow rate (i.e., gas flow rate, R_f) and reaction temperature (T) are the top two features for the CVD synthesis of WTe_2, which are in good agreement with the results of MoS_2. A detailed feature comparison between MoS_2 and WTe_2 systems has been summarized in the Table S5 in the Supporting Information. Based on the feature importance, morphology control of WTe_2 is achieved and a rational growth mechanism is further proposed, which are discussed in detail in another study [43]. One typical trial of PAM on WTe_2 dataset is visualized and analyzed in Fig. S10e, with N_C equals to 142. It suggests that 44.31% time reduction with 92.11% true positive rate could be achieved with the proposed PAM.

ML-guided hydrothermal synthesis with enhanced targeted property

To further verify the generalizability of our established methodology across synthesis methods, we have extended its application to the hydrothermal system, a well-known multi-variable synthesis method to obtain inorganic materials [44]. aiming to enhance the process-related properties as shown in Fig. 1b. Recently, carbon quantum dots (CQDs) obtained by hydrothermal method have gained substantial attention for their tunable low toxicity, high biocompatibility and robust surface engineering capacity and thus have been widely used in diverse fields including light emitter, sensors, catalysis, bio-imaging, and energy harvesting etc [45,46]. Therefore, improving the properties of CQDs with ML is of great research interest. Most importantly, it showcases the feasibility of our methodology in addressing regression problems on top of classification problems (CVD-grown MoS_2 dataset).

Dataset and model construction. For the growth of CQDs, the experimental setup and detailed synthesis process are provided in Fig. S11 in the Supporting Information and Methods. Empirically, six hydrothermal parameters are identified as significant input features: pH value (pH), reaction temperature (T), reaction time (t), mass of precursor A (M), ramp rate (R_r) and solution volume (V). Detailed feature overview is shown in Table S6 in the Supporting Information. Feature correlation is presented in Fig. S12 in the Supporting Information, with low linear correlations verifying the effectiveness of feature selection. As high photoluminescence quantum yield (PLQY) is a key property of quantum dots desired for applications, it is targeted in this work for further enhancement. 467 experimental records are retrieved from our archived laboratory notebooks, with different growth parameters and respective PLQY ranging from 0 to 1 clearly labeled.

In order to best infer PLQY from the features, several regression algorithms are evaluated with nested cross validation mentioned above, including XGBoost regressor (XGBoost-R) [31], support vector machine regressor (SVM-R),[32] Gaussian process regressor (GP-R) [47], and multilayer perceptron regressor (MLP-R) [34]. Coefficient of determination (R^2) is adopted as the primary performance indicator, which measures the proportion of variance of the outcome (i.e., PLQY) that is predictable from the features. Detailed model comparison results are provided in Fig. S13 in the Supporting Information. XGBoost-R outperforms the rest by a large margin with its R^2 equals to 0.8402, where approaching one is desirable; and is thus selected as the best model.

Optimization for higher PLQY. After obtaining the trained XGBoost-R model with the full dataset, feature importance of the hydrothermal system is studied as well. As shown in Fig. 4a, pH value plays the most important role in determining the value of PLQY, followed by reaction temperature and reaction time. This coincides with our expectation: (1) pH will affect the formation of CQDs, as small stable CQDs would dissolve in the acidic and basic solutions. (2) Optimal reaction temperature is required for the formation of CQDs. Higher temperature will result in higher average kinetic energy of molecules and more collisions per unit time, damaging the formation of stable CQDs; and lower temperature would retard or even prevent the formation of CQDs because of insufficient chemical reaction energy. (3) Reaction time is an important factor controlling the size of CQDs, which then affects their photoluminescence properties owing to the quantum confinement effect. Inadequate time will not lead to the formation of CQDs, while prolonged time will result in large-sized CQDs. When the size is larger than the exciton Bohr radius, the quantum confinement effect of CQDs will be impaired and PLQY will be reduced.

The trained XGBoost-R model is then applied to predict the PLQY of 1,555,840 possible synthesis conditions resulting from the combinations of different values of features shown in Table S7 in the Supporting Information. Eleven synthesis conditions are recommended by the model attributed to their highest predicted PLQY. Experiments are then carried out in the lab with results provided in Table S8 in the Supporting Information. High PLQY of 55.5% (vs. 52.8%, the highest PLQY in the training set) is achieved surprisingly, which is one of the highest PLQY reported with such ultra-low heteroatom doping precursor ratio [45]. Characterizations of the obtained CQDs are provided in Fig. 4b and Fig. S14 in the Supporting Information. Moreover, the average PLQY in the recommendation set reaches 53.56%, more than twice the average value of the training set. Comparison of the performances of the training set and the ML-provided recommendation set is presented in Fig. S15 in the Supporting Information, indicating great effectiveness of the ML model.

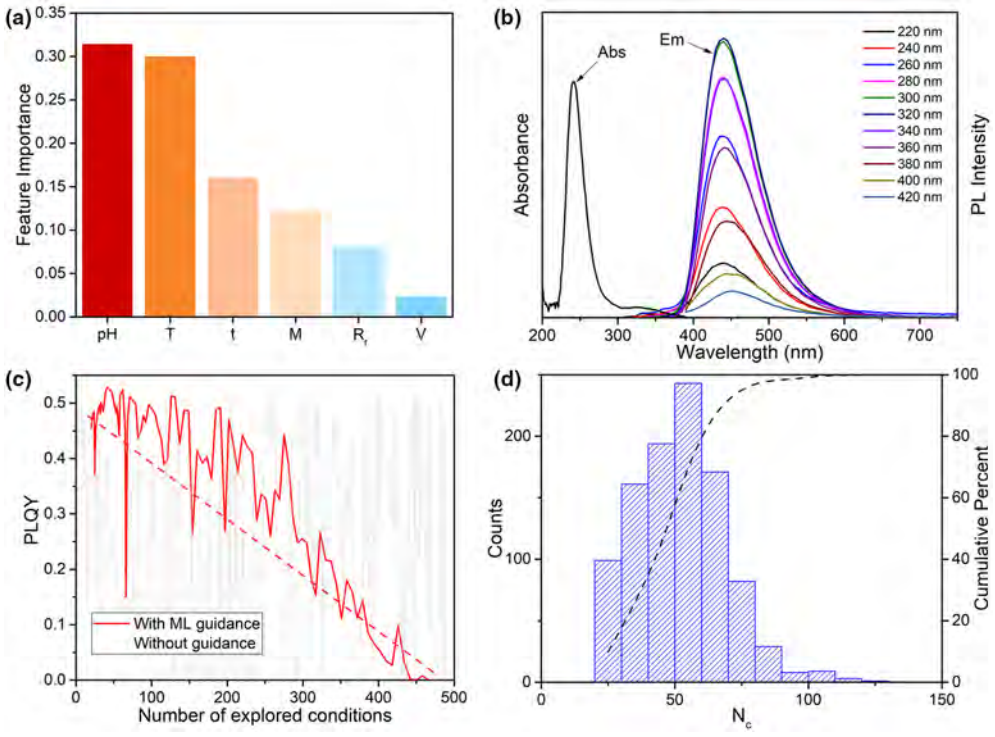

FIGURE 4

Optimization and acceleration of hydrothermal-grown CQDs with XGBoost-R and PAM. (a) Feature importance retrieved from XGBoost-R that learns from the full dataset. The most important features are pH and T. (b) The UV–Vis absorption wavelength of the as prepared S,N-CQDs, and the fluorescence emission spectra at different excitation wavelengths. (c) Plot of PLQY achieved with or without ML guidance versus number of explored conditions of a typical PAM trial, smoothed with median filter of window size 3 to highlight the trend. (d) Distribution of the critical points of 1000 PAM trials, suggesting that PAM model is 99.9% confident to find the best condition of this confined dataset within 115 experiments.

Acceleration with PAM. The limitation with traditional experimental exploration arises from the heuristic choice of experimental conditions due to the lack of guidance. Specifically, the optimal synthesis condition within the pre-defined search space needs to be explored through a large number of experiments. In the CQDs dataset, without ML guidance, the probability of finding the best synthesis condition is evenly distributed among the full dataset of 467 experiments, leading to excessive waste of time. To tackle such problems as well as to test the generalizability of PAM proposed above, the performance of PAM on CQDs regression dataset is carefully examined, aiming to efficiently identify the best synthesis condition with minimum number of trials.

In a typical run of PAM, along with the increase of explored conditions, the corresponding true yield shows a clear declining trend, suggesting that the PAM model is capable of identifying the best synthesis conditions at the early stage of PAM loops (see Fig. 4c). In contrast, the original empirical exploration demonstrates a much more random nature. In order to verify that PAM can perform stably with varying initial training sets, PAM on CQDs dataset is repeated 1000 times with randomly chosen initial training sets. In each trial, the loop number where the best synthesis condition is found, denoted by N_c, is recorded. The results of 1000 trials are summarized in Fig. 4d, from which we can see that the PAM model is able to find the best condition of this confined dataset within 115 experiments with 99.9% confidence. Compared to the 467 experiments that need to be conducted through a heuristic approach, our proposed PAM-based approach helps to achieve 75.37% reduction in time for identifying the optimal synthesis conditions for WTe_2.

Conclusion and outlook

In summary, this study demonstrates the successful application of ML in guiding the synthesis of inorganic materials, through the establishment of the methodology including model construction, optimization and PAM. High AUROC of 0.96 is achieved with XGBoost-C for the CVD system in predicting the synthesis result of 2D MoS_2, thereby optimizing its CVD synthesis condition. PAM, whose active feedback loop renders ML capable of guiding new material synthesis at the beginning stage, is next used to enhance the experimental outcome as well as minimize the number of trials. More importantly, we also demonstrate that the proposed methodology could be extended to any type of multi-variable synthesis method across various material categories. In this paper, this is achieved by applying the proposed ML strategy to a hydrothermal system to effectively improve the process-related properties (i.e., PLQY) of CQDs. Our results corroborate the potential of ML to optimize and accelerate the material synthesis process, thereby encouraging the development of advanced inorganic materials for practical applications in terms of time reduction and property enhancement.

In this study, we have focused on using ML models for the synthesis of a single type of material with a few important features, to

simplify the complex problem of material synthesis through bypassing the chemistry factors behind such processes. However, to further exploit the useful information contained in historical trials and guide material synthesis more effectively, a more comprehensive model involving chemistry-related features such as the vapor pressure, solubility, reactivity, etc., as well as various types of material is very much required. It may not only produce more accurate predictions and guidance, but also reveal new information or hypotheses regarding the fundamental mechanism of successful synthesis by inverting the model. Additionally, the establishment of an integrated database across inorganic material synthesis systems is another promising perspective. Standardization of data recording protocols and construction of a universal feature list are two essential premises. Despite huge amount of efforts and joint interdisciplinary collaboration demanded, the established database will contribute to the mining of unidentified relationships between synthesis parameters and experimental outcomes, paving the way for data-driven intelligent synthesis of advanced inorganic materials with ML.

Methods

Synthesis of MoS_2: Sulfur (S) and molybdenum trioxide (MoO_3) are used as precursors. Si wafer with a 280 nm SiO_2 top layer is used as substrate. The MoO_3 powder is put into the boat, and Si/SiO_2 substrate is put on the boat with the polished surface down. The boat is then placed in the middle of the 1-in. diameter quartz tube. Sulfur powder is positioned a few centimeters away from the furnace mouth in the upstream and Argon (Ar) gas is used as the carrier gas. The system is heated to the growth temperature with designated ramping rate and maintained for a few minutes for the growth of MoS_2.

XGBoost: XGBoost derived from Gradient Boosting Decision Tree (GBDT) [31,48], is a typical class of gradient boosting that employs decision trees as base estimators. It makes decision through an ensemble of M base estimators $h_m, m = 1, \cdots, M$:

$$\widehat{y}_i = \sum_{m=1}^{M} h_m(x_i)$$

Given N training data $\{(x_i, y_i)\}_{i=1}^{N}$, the objective is to minimize:

$$obj(\theta) = \sum_{i=1}^{N} l(y_i, \widehat{y}_i) + \sum_{m=1}^{M} \Omega(h_m)$$

where $\sum_{i=1}^{N} l(y_i, \widehat{y}_i)$ is the training loss and $\sum_{m=1}^{M} \Omega(h_m)$ is a regularization term which penalizes complexity of the base estimators. Additive training strategy adds one new tree at a time, by choosing the tree that optimizes the objective at step t: $obj(\theta)^t = \sum_{i=1}^{N} l\left(y_i, \widehat{y}_i^{\,t}\right) + \sum_{m=1}^{t} \Omega(h_m)$whereas $\widehat{y}_i^{\,t} = \sum_{m=1}^{t} h_m(x_i) = \widehat{y}_i^{\,t-1} + h_t(x_i)$.

ROC curve: To plot the ROC curve of XGBoost-C model [38], nested cross validation is employed to generate predicted probabilities on 300 data samples respectively. In ten-fold outer cross validation, nine folds are used as the model development set, while the predicted probabilities of the remaining 30 samples are recorded. In the inner cross validation, the best hyperparameters are determined on the model development set with stratified ten-fold cross validation. True positive rate, which makes up the *y*-axis of ROC curve, indicates the percentage of true positive samples that are correctly predicted. False positive rate, *x*-axis of ROC curve, is the percentage of true negative samples that are falsely predicted as positive.

SHapley Additive exPlanations (SHAP): SHAP is a unified approach for additive feature attribution, which produces theoretically sound and unique solutions [39]. The explanation model $g(z')$ satisfies $g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$, where $z' \in \{0,1\}^M$, M is the number of input features, and $\phi_i \in \mathbb{R}$. $z_i' = 1$ indicates a feature that is being observed, otherwise it is denoted by 0. ϕ_i represents the feature importance value.

To compute SHAP values, $f_x(S) = E[f(x)|x_S]$ is defined where$f$ is the function, i.e., ML model, to be explained, S is the set of non-zero indexes in z', and $E[f(x)|x_S]$ is the expected value of the function conditioned on the subset S of the input features. Using these conditional expectations, SHAP value is assigned to each feature:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)$$

where N is the set of all input features.

Success rate, time reduction and true positive rate on the whole dataset: Success rate is defined as the number of "Can grow" conditions explored divided by the total number of explored conditions (i.e., N_C), while time reduction is calculated as $\frac{|N_{all} - N_c|}{|N_{all}|}$. True positive rate is defined as the number of correctly predicted positive samples over the total number of true positive samples. The true labels are obtained from experimental results, with positive and negative classes referring to "Can grow" and "Cannot grow", respectively. At N_C of each trial of PAM, the true positive is computed as the number of "Can grow" conditions found divided by the total number of true conditions in the whole dataset (i.e., 183 for the MoS_2 dataset).

Type-I and Type-II Error Rate: Type-I error rate indicates the percentage of falsely predicted positive samples (False Positive, as it is falsely predicted as positive but actually in negative class) over the total number of true negative samples. Similarly, type-II error rate is defined as the percentage of falsely predicted negative samples (False Negative, as it is false predicted as negative but actually in positive class) over the number of true positive samples. For PAM, type-I and type-II error rates are calculated as follows. N_1 refers to the initial training set. N_K represents the training set in the k^{th} loop (inclusive of N_1), and thus the test set is $(N_{all} - N_K)$. Since $(N_K - N_1)$ are the experiments conducted under the guidance of PAM, which are regarded as the most probable "Can grow" conditions in each loop, they are predicted as positive samples by PAM. The prediction results of the test set are produced by the model trained on N_K. In the k^{th} loop:

$$\text{Type} - \text{I error rate} = \frac{(N_k - N_1)^- + (N_{all} - N_k)^{FP}}{(N_{all} - N_1)^-}$$

$$\text{Type} - \text{II error rate} = \frac{(N_{all} - N_k)^{FN}}{(N_{all} - N_1)^+}$$

whereas for a random set N, N^+ and N^- represent the number of samples in true "Can grow" and "Cannot grow" class respectively

in N, N^{FP} stands for number of false positive samples in N, N^{FN} indicates the number of false negative samples in N.

Synthesis of carbon quantum dots (CQDs): 10–60 mL 0.01 M sulfamide solution and 0.2–20 g sodium citrate is added into a 100 mL Teflon-lined stainless-steel autoclave. Then, the autoclave is kept in an oven at 80–300 °C for 0.1–12 h. After the reaction, the resulting product is filtered using a 0.22 mm membrane filter followed by concentrating using rotary evaporator to obtain the purified (S, N)-CQDs. The filtrate is dialyzed in a 500 Da dialysis bag for 2 days to obtain the final S, N-CQDs, against ultrapure water which is renewed every 10–12 h, until almost no Na + (below detective limit) is detected in DI water.

Data and code availability

The raw data required to reproduce these findings are available to download from https://github.com/MSwML/ML-guided-material-synthesis.git. The source code used to perform the ML tests and to generate the plots used here, is available under MIT License.

CRediT authorship contribution statement

Bijun Tang: Conceptualization, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Yuhao Lu:** Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Jiadong Zhou:** Investigation, Writing - review & editing. **Tushar Chouhan:** Writing - review & editing. **Han Wang:** Software, Formal analysis. **Praffful Golani:** Investigation, Writing - review & editing. **Manzhang Xu:** Investigation, Visualization. **Quan Xu:** Investigation, Supervision, Writing - review & editing. **Cuntai Guan:** Conceptualization, Supervision, Writing - review & editing. **Zheng Liu:** Conceptualization, Supervision, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

B. T and Y. L contributed equally to this work. This work was supported by National Research Foundation-Competitive Research Program (NRF-CRP21-2018-0007). This work was also supported from the Singapore Ministry of Education Tier 3 Programme "Geometrical Quantum Materials" (MOE2018-T3-1-002), AcRF Tier 2 (2016-T2-2-153, 2016-T2-1-131), AcRF Tier 1 (RG7/18 and RG161/19). The authors also gratefully acknowledge the support from the National Natural Science Foundation of China (Grant No. 61974120). Q. X acknowledges the financial support from National Key Research and Development Plan (2019YFA0708300), Science Foundation of China University of Petroleum (2462019QNXZ02, 2462018BJC004).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mattod.2020.06.010.

References

[1] J.P. Correa-Baena et al., Joule 2 (8) (2018) 1410.
[2] B. Radisavljevic et al., Nat. Nanotechnol. 6 (3) (2011) 147.
[3] O. Lopez-Sanchez et al., Nat. Nanotechnol. 8 (7) (2013) 497.
[4] Q.H. Wang et al., Nat. Nanotechnol. 7 (11) (2012) 699.
[5] Y. Chen et al., Chem. Rev. 118 (13) (2018) 6409.
[6] Z.Y. Cai et al., Chem. Rev. 118 (13) (2018) 6091.
[7] J.D. Zhou et al., Nature 556 (7701) (2018) 355.
[8] H. Li et al., Chem. Rev. (2017).
[9] G. Pilania et al., Front. Mater. (2016) 3.
[10] S.H. Lu et al., Nat. Commun. (2018) 9.
[11] F. Ren et al., Sci. Adv. 4 (2018) 4.
[12] D.Z. Xue et al., Nat. Commun. (2016) 7.
[13] P. Raccuglia et al., Nature 533 (7601) (2016) 73.
[14] O. Isayev et al., Nat. Commun. (2017) 8.
[15] C.W. Rosenbrock et al., Npj Comput. Mater. (2017) 3.
[16] L.L. Li et al., Sci. Adv. 4 (2018) 3.
[17] A. Ziletti et al., Nat. Commun. (2018) 9.
[18] J.M. Granda et al., Nature 559 (7714) (2018) 377.
[19] C.W. Coley et al., Acs Central Sci. 3 (5) (2017) 434.
[20] S.V. Ley et al., Angew. Chem. Int. Ed. 54 (11) (2015) 3449.
[21] Z.Y. Yin et al., ACS Nano 6 (1) (2012) 74.
[22] D. Voiry et al., Nano Lett. 13 (12) (2013) 6222.
[23] C. Mai et al., Nano Lett. 14 (1) (2014) 202.
[24] A.M. Cassell et al., J. Phys. Chem. B 103 (31) (1999) 6484.
[25] A. Reina et al., Nano Lett. 9 (1) (2008) 30.
[26] X. Li et al., J. Vacuum Sci. Technol. A 21 (4) (2003) 1342.
[27] K.K. Kim et al., Nano Lett. 12 (1) (2011) 161.
[28] Q. Fu et al., RSC Adv. 5 (21) (2015) 15795.
[29] J.Y. Chen et al., Adv. Sci. 3 (2016) 8.
[30] D.H. Wolpert, W.G. Macready, IEEE Trans. Evol. Comput. 1 (1) (1997) 67.
[31] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, p. 785.
[32] C.C. Chang, C.J. Lin, Acm T. Intel. Syst. Tec. 2 (2011) 3.
[33] N. Friedman et al., Mach. Learn. 29 (2–3) (1997) 131.
[34] D.E. Rumelhart et al., Nature 323 (6088) (1986) 533.
[35] R.H. Yuan et al., Adv. Mater. 30 (2018) 7.
[36] S. Sun et al., Joule (2019).
[37] G.C. Cawley, N.L.C. Talbot, J. Mach. Learn. Res. 11 (2010) 2079.
[38] J.A. Hanley, B.J. Mcneil, Radiology 143 (1) (1982) 29.
[39] S.M. Lundberg, S.I. Lee, Adv. Neur. In. (2017) 30.
[40] S. Wang et al., Chem. Mater. 26 (22) (2014) 6371.
[41] S.-Y. Xu et al., Nat. Phys. 14 (9) (2018) 900.
[42] Y. Shi et al., Sci. Adv. 5 (2) (2019) eaat8799.
[43] M. Xu, et al., Machine learning driven synthesis of few-layered WTe2. In arXiv e-prints, (2019).
[44] W. Shi et al., Chem. Soc. Rev. 42 (13) (2013) 5714.
[45] Q. Xu et al., J. Mater. Chem. B 4 (45) (2016) 7204.
[46] S.Y. Lim et al., Chem. Soc. Rev. 44 (1) (2015) 362.
[47] C.E. Rasmussen, C.K.I. Williams, Adapt. Comput. Mach. Le (2005) 1.
[48] J.H. Friedman, Ann. Stat. 29 (5) (2001) 1189.