

# Deep Learning Accelerated Gold Nanocluster Synthesis

Jiali Li, Tiankai Chen, Kaizhuo Lim, Lingtong Chen, Saif A. Khan, Jianping Xie, and Xiaonan Wang\*

The understanding of inorganic reactions, especially those far from the equilibrium state, is relatively limited due to the inherent complexity. Poor understanding of the underlying synthetic chemistry constrains the design of efficient synthesis routes toward the desired final products, especially those at atomic precision. Using the synthesis of atomically precise gold nanoclusters as a demonstration platform, a deep learning framework for guiding material synthesis is successfully developed to accelerate the workflow. With only 54 examples, the graph convolutional neural networks (GCNN) plus siamese neural networks (SNN) classification model is trained. The prediction capability is demonstrated with the successful prediction of literature-reported protocols. In addition, understanding of the synthesis process can be acquired from a decision tree trained by plentiful generated data from a well-trained classification model. This study not only provides a data-driven method accelerating gold nanocluster synthesis, but also sheds light on understanding complex inorganic material synthesis with low data.

## 1. Introduction

Sub-2 nm gold nanoparticles, or gold nanoclusters (Au NCs), attracted much research interest in the past two decades.<sup>[1,2]</sup> Undoubtedly, the growing research interest is motivated by the unique properties of Au NCs, such as discrete electronic states, defined molecular formula and structure, quantized charging, molecular chirality, and strong photoluminescence.<sup>[3–8]</sup> These properties are not observed in bulk gold or gold nanoparticles with core sizes larger than 2 nm. As a result, the Au NCs are extensively studied and can be potentially applied in various fields. More interestingly, these properties are greatly determined by the size and the structure of the Au NCs, or to be


specific, by their molecular formulas. Being ultrasmall in size, Au NCs usually consist of several to one hundred gold atoms in the core, which are stabilized by a shell of ligands (e.g., thiolates and phosphines). Therefore, an atomically precise Au NC can be represented by a formula of  $[M_nL_m]^q$ , where  $n$ ,  $m$ , and  $q$  are the number of metal atoms, ligand molecules, and the net charge in one NC, respectively. A small difference in the values of  $n$ ,  $m$ , and  $q$  can greatly affect the properties of the corresponding Au NC.<sup>[9]</sup> Moreover, the separation of mix-sized Au NCs requires high-resolution separation techniques such as high-performance liquid chromatography due to the inherent tiny difference in their sizes and structures. Therefore, obtaining Au NCs at atomic precision after the synthesis will greatly promote the use of their properties for

potential applications. The synthesis of Au NCs typically adopts from the Brust method.<sup>[10]</sup> It is a two-step reduction where the ligand is first mixed with an Au(III) salt (typically  $\text{HAuCl}_4$ ), followed by the addition of another reducing agent (a schematic illustration is shown in **Figure 1**). In the first step, Au(III) is reduced into Au(I), which is then coordinated with the thiolate ligand to form a mixture of Au(I–ligand complex of various sizes (e.g.,  $\text{Au}_4\text{L}_4$ ,  $\text{Au}_6\text{L}_6$ , and  $\text{Au}_{10}\text{L}_{10}$ ). After the addition of a reducing agent (e.g.,  $\text{NaBH}_4$ ), a mixture of Au NCs of different sizes will form during the initial stage of reduction. Whether these NC species can grow into a single size (becoming monodispersed at atomic precision) is highly dependent on the size distribution of the initial mixture and the reaction conditions (such a process is also known as “size-focusing”).<sup>[11]</sup> As a result, the direct synthesis of atomically precise Au NCs is challenging without careful experimental design.

A computer-assisted design can accelerate the synthesis process and help chemists develop the chemical synthesis route in a more effective way. From the Materials Genome initiative, the first principle simulations along with the quantitative structure–property relationship method as a data-driven tool are built to predict the properties of the materials efficiently without having to conduct the actual experiment.<sup>[12,13]</sup> Based on this effort, more researchers are looking at improving the synthesis process itself. In these studies, both successful and unsuccessful experiments are used to train machine learning (ML) models. These ML approaches do not require solid understanding of a specific domain as required by the first principle simulation approaches. Instead, they are able to gain experience from the past

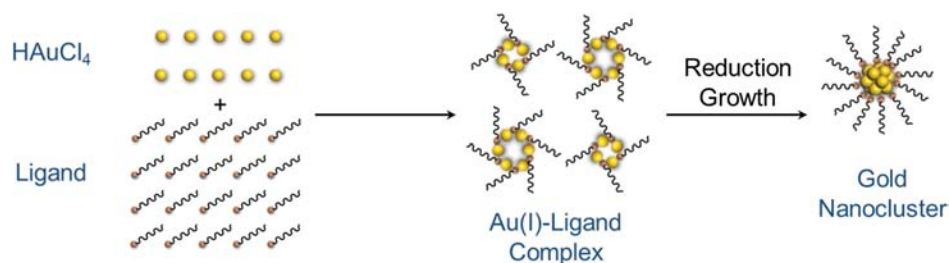
J. Li, Dr. T. Chen, K. Lim, Prof. S. A. Khan, Prof. J. Xie, Prof. X. Wang  
Department of Chemical and Biomolecular Engineering  
National University of Singapore  
4 Engineering Drive 4, Singapore 117585, Singapore  
E-mail: chewxia@nus.edu.sg

L. Chen  
Department of Computer Science  
University of Southern California  
941 Bloom Walk, Los Angeles, CA 90089, USA

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.201900029>.

© 2019 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.201900029



**Figure 1.** Synthesis of Au NCs. Schematic illustration of the modified Brust method in the synthesis of Au NCs.

experimental data to guide future experiments. Such data-driven screening approaches have been proved successful in organic chemistry, where a large amount of tabulated synthetic data are available.<sup>[14–17]</sup> Similarly, for inorganic materials, the identification of synthesis parameters driven by ML has also been studied, including the synthesis of a desired phase state and obtaining a certain range of properties.<sup>[18–20]</sup> However, the synthetic chemistry of the reactions in most of these studies involves only one or limited types of reactions (e.g., with only coordination reactions). On the other hand, the synthesis of Au NCs is more complex. The first stage (mixing of Au(III) salt with ligands) involves the reduction from Au(III) to Au(I) and the coordination between Au and the ligand, while reduction from Au(I) to Au(0) and the aggregation behavior of Au(0) appear in the second-stage reaction. Such a complex reaction system has not yet been explored by ML or other data-driven methods. Nevertheless, ML is able to deal with complex systems in principle from its strong capability of classification tasks in nature. Due to the complexity of the reactions involved in their synthesis, there are very limited number of reports on the direct synthesis of atomically precise Au NCs, resulting in low data amount for the ML training process. Under such conditions, a simple ML model will most likely be unable to learn much information.<sup>[21]</sup> However, one-shot learning methods such as Siamese Neural Networks (SNN) have the potential to perform well with low data amount. This deep neural network has shown strong capability in image recognition with limited examples and promising prediction ability in low-data drug discovery with molecular structure information as an input.<sup>[21–23]</sup>

Herein, we present an ML model that is able to accelerate the understanding in the synthesis of atomically precise Au NCs under a low-data condition by training all the parameters together in this complex reaction system. Although some mechanistic understandings have been gained for the synthetic chemistry, their capabilities to predict the synthesis across the diverse reaction systems are still weak.<sup>[24–26]</sup> For example, a protocol for the synthesis of atomically precise Au<sub>25</sub> NCs works only for specific ligands, while the same Au NC species is unlikely to be synthesized when other ligands are adopted. We use ML to conduct a classification task based on studying the relationship between reaction conditions, molecular properties, and the final monodispersity in the product (i.e., whether they are atomically precise Au NCs). The ML model uses an SNN stacked with the Graph Convolutional Neural Networks (GCNN). With the trained classification model, we can generate new sets of synthetic experimental conditions that are more likely to be successful and can

be carried out in the laboratory, allowing the acceleration of the material discovery process. This first task is also known as synthesis parameters recommendation. In the second part, a “model-of-model” is created to map the black-box SNN onto a human-interpretable decision tree, which can provide more chemical insights despite low amounts of data in the training dataset. With our GCNN + SNN model, we are able to learn insights such as temperature trends, while the other models tested in this work failed to learn such insights due to the low-data constraint.

## 2. Results and Discussion

We extracted the synthesis conditions of 27 examples from reported literatures and 27 examples from our own laboratory. The dataset includes synthesis conditions that were able to obtain atomically precise Au NCs in the product (i.e., successful examples, only one Au NC species after the reaction), as well as those not able to obtain atomically precise Au NCs in the product (i.e., unsuccessful examples, more than one Au NC species after the reaction).

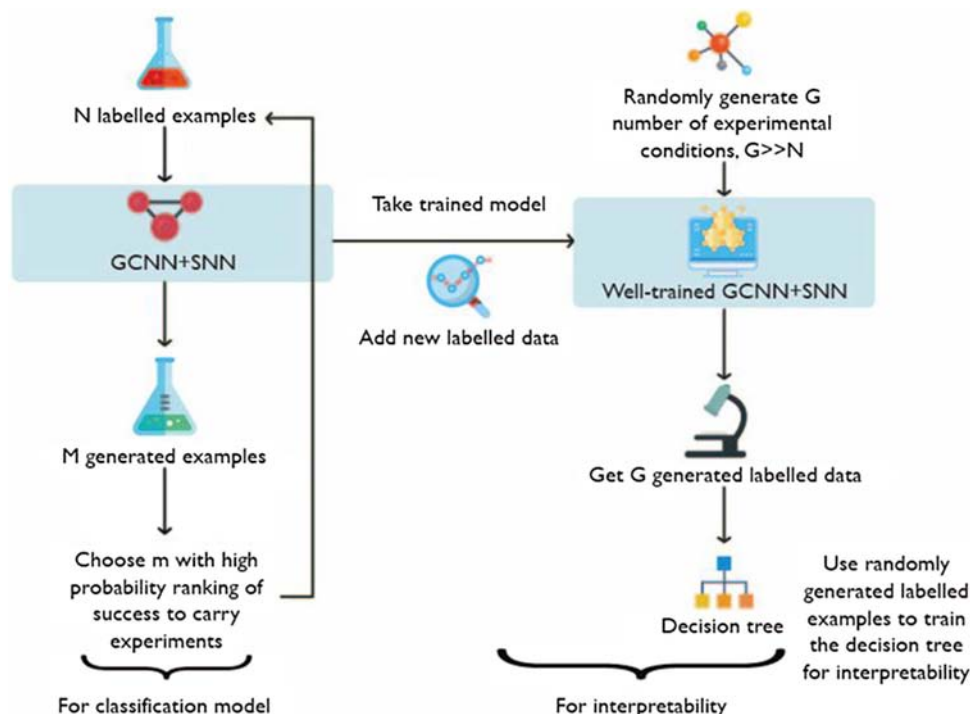
To demonstrate the performance of the proposed framework, we have evaluated the synthesis of atomically precise Au NCs in two aspects. The first aspect is the process where the key reaction components (including the ligands, solvent, and reducing agent) for our experiment can be varied. The second aspect is an optimization process of reaction conditions without changing the key reaction components. The former aspect is defined as an explorative process and the latter is defined as an optimization process. To aid in the explorative process, the proposed ML models should be capable of learning the molecular information, which denotes the information based on the molecular physicochemical and structural similarity between the key reaction components. Meanwhile, to perform well in the optimization task, the ML models should learn synthetic condition information, which is related to how the continuous variations in reaction conditions (e.g., temperature and pH) affect the final results.

As a first attempt, some well-studied ML methods, such as Support Vector Machine (SVM) and Dense Neural Network (DNN), were used with the domain-related descriptors as features (Table S1, Supporting Information). Such simple strategies have been proved to be insufficient for this low-data-amount case, as both the key statistical evaluation indicators and probability distributions for experimental conditions are not satisfactory, which will be discussed in detail below. It indicates that a simple ML approach with basic descriptors is not strong enough

to learn either molecular information or synthetic condition information effectively for aiding in an explorative or optimization process from a small dataset. Thus, a deep learning framework with a closed learning loop based on SNN is constructed as shown in **Figure 2**. The use of one-shot learning method, SNN, is motivated by the fact that only a limited amount of ground-true experimental data in the inorganic material synthesis field are available and the failure in simple ML strategies such as SVM. The aim of this framework is to help material scientists accelerate the understanding of the atomically precise Au NC synthesis through two major approaches: first, increasing the synthesis and characterization rates of new atomically precise Au NCs by the key classification model; and secondly, mapping the classification model into an interpretable decision tree to gain chemical insights. The key classification model is an SNN with the GCNN stacked on the top of it. With the randomly initialized conditions, the model is able to identify relative promising experimental conditions with high success probability for carrying out successful experiments with either an explorative or an optimization approach. The outcomes from both successful and unsuccessful experiments can provide feedback to improve this key classification model with a closed learning loop. With the aid of the well-trained key classification model, an arbitrarily large

number (about 10 000) of new “synthetic” examples can be generated for building the decision tree, which can provide further information to make the ML model a “white box” with chemical insights. These insights, in turn, will help in understanding the synthesis process better. The construction of the key classification models, the comparison between different ML methods, and the chemical insights gained from the proposed approach are discussed later.

**Key classification model:** The dataset we have extracted is heterogeneous from diverse sources. It consists of various examples using different ligands, solvents, and reducing agents aiming to produce Au NCs (such as Au<sub>25</sub> or Au<sub>38</sub>), providing the molecular information variations. However, it is worthy to note that typically only one or a few reaction conditions are altered from the view of parameter optimization, leading to insufficient information on synthetic condition variations. We labeled an experimental set as 1 if atomically precise Au NCs were obtained in the product and 0 conversely. According to the different features in the 54 examples, 2 dataset groups were created for training: 1) 54 examples from the full dataset (54 dataset I sheet in the Supporting Information data); and 2) 35 examples of aqueous synthesis of Au<sub>25</sub> NCs (35 dataset II in the Supporting Information data). The performance of different ML models



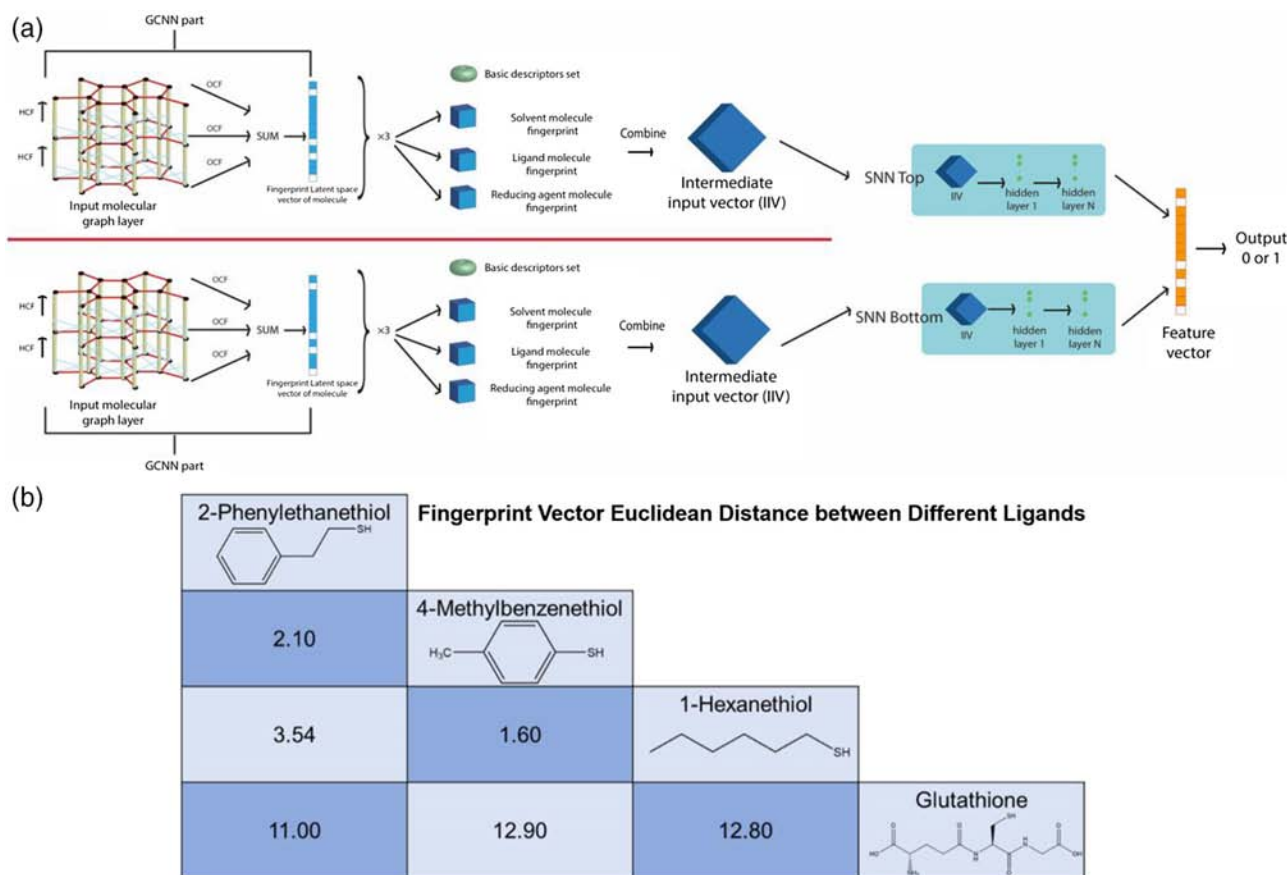
**Figure 2.** A deep learning framework with a closed learning loop based on SNN. This framework is divided into two parts: 1. Classification model (left), and 2. Interpretation process (right). For the first part: initially,  $N$  (a small number in our low data case) labelled examples (i.e., binary classification of atomic precision or not) from literature or laboratory are collected. These examples are used to train the key classification models, GCNN + SNN in this figure. Although it is feasible to substitute this key classification model with other ML models (e.g., SVM and DNN), the GCNN + SNN model illustrated here is the best-performing one which will be discussed later. By using the well-trained model, one can generate  $M$  arbitrary number of examples with any arbitrary reaction conditions. These  $M$  examples will come out with probability of successful synthesis where scientists can pick the ones with high successful probability to carry out experiments. Both successful and unsuccessful experiments will provide feedback for classification learning to improve the classification performance. For the second part: a large number  $G$  (about 10 000) of experimental conditions examples are generated randomly. Then the well-trained classification model in this figure, GCNN + SNN, is taken to predict the outcomes of these examples to label them. These  $G$  generated labelled examples are then used to train a decision tree for getting chemical insights.

on the dataset group I as our key focus is compared. A “model-of-model” is built on the training of the dataset group II to learn the chemical insights based on all possible parameters and especially the effects of both pH and ligand-to-Au molar ratio. It should be noted that the subdataset (i.e., dataset group II) is chosen because pH information is not available for the synthesis of Au NCs conducted in organic solvents and the ligand-to-Au molar ratio may be in different ranges for the synthesis of differently sized Au NCs.

Among the 17 basic descriptors chosen, 12 features are related to the physicochemical properties or the structure information of the key reaction components (e.g., the dielectric constant of the solvent, the number of rotatable bonds of the ligand, and the  $\chi_{\text{logP3}}$ <sup>[27]</sup>). These descriptors are used to distinguish the key reaction components (e.g., ligands, solvent, and reducing agent) and find how they affect the final classified state. These descriptors can provide molecular information, which contributes to the ability of the ML models to run an explorative process classification. In addition, there are five more descriptors on the key operation

variables (e.g., temperature and pH) of the entire process, which are related to the optimization process. The detailed descriptor information is listed in Table S1, Supporting Information. Apart from these descriptors, a GCNN is trained together with classification models in order to account for the rich molecular structure information of the key reaction components. The entire model is illustrated in **Figure 3a** using the SNN with a GCNN stacked on the top.

The SNN is a matching neural network that takes in a pair of input, which is named as an intermediate input vector (IIV) here as shown in **Figure 3a**. Each IIV is passed through an identical half (top or bottom of the SNN) with the same densely connected hidden layer weights and biases. The L1 distance between the last layers of each half of the SNN is taken and connected to one final output node that will be between 0 and 1 in value. A distance value close to 1 means that the two IIV are likely to belong to the same classified state (i.e., both leading to successful synthesis or both leading to unsuccessful synthesis), while the value close to 0 means that they are likely belonging to different classified



**Figure 3.** Model illustration and Fingerprint Vector Euclidean distance calculation. a) GCNN was used to produce Fp for solvent molecules, ligand molecules, and reducing agent molecules. The SNN is trained by giving the models batches of pairs from a set of labeled examples (the training dataset) with the half being of the same class and the other half being a different class. In our model, the GCNN is trained as part of the SNN and its weights and biases are updated together with the SNN instead of being pre-trained. This enables the GCNN to map discrete molecules to their Fp vector latent space in a manner that is more suited for the matching neural network. For SVM and DNN, the input structures are the same as SNN, but have no such separation of the top and bottom parts of the classification model. b) Fingerprint vector Euclidean distance between four chosen ligands are calculated. If the structures of two ligands are similar, their fingerprint vector output from the GCNN will have a small value of Euclidean distance between them and vice versa. This shows the capability of GCNN to produce structure similarity information that enhances the performance of classification models by learning molecular information.



states. The IIV consists of four parts, three molecular fingerprint vectors (Fp vectors) that contain structure information, and the vector of a basic descriptor set (Table S1, Supporting Information) which contains other molecular information such as the electronegativity of the ligands. There is one Fp vector each for the solvent, the ligand, and the reducing agent, respectively. Those four parts are concatenated together to form the IIV which provides both structural and physicochemical information that a neural network can learn from (rather than discretized molecules as a one-hot vector input to the neural network).

The Fp vector is the output of a GCNN which takes the Simplified Molecular-Input Line-Entry System (SMILES) canonical name<sup>[28]</sup> of a molecule as an input. GCNN originates from the idea of convolutional neural networks (CNN)<sup>[29]</sup> which have the ability to extract multiscale localized spatial features and compose them into more expressive representations. However, CNN can only take Euclidean data structures (e.g., images as a 2D grid). Generalizing it to nonEuclidean data structures (e.g. any graph with nodes and edges), GCNN has the same properties as a CNN with a local connection, shared weights and multilayer structure.<sup>[30]</sup> In the case of a molecular fingerprint, SMILES contains the structural information of a molecule. It is converted to a tensor as a high-dimensional representation of data (e.g., 1D tensors as vectors and 2D tensors as matrix) with RDKit python library to store the molecular structure, so that each atom is considered as a node and the molecular bonds are edges between the nodes. This is the base layer of the GCNN (as shown by the top left in red color in Figure 3a, with vertices representing individual atoms and edges representing bonds). A Hidden Convolutional Filter (HCF) is applied to the layer which calculates the information flow from neighbors for each node (as indicated by the sky blue lines in Figure 3a molecular graph in red color). The output from HCF is the same size as the initial layer and a HCF with different weights is applied again. This process is done for  $P$  times, so there are  $P + 1$  number of layers in total. Each layer is further input into an Output Convolutional Filter to get the output column vectors with the same length. All these output column vectors are finally summed up to get the Fp vector.

After training the GCNN + SNN on a training dataset, the trained model can be used to predict whether an example that has not been observed before belongs to the success or failure class. This is achieved by splitting the training dataset into two support sets, each containing either all the success or the failure data, respectively. The unseen example is input to the top half of the SNN and an example from the success support set is given to the bottom half of the SNN. This is looped until the unseen example is compared to every other success example from the support set. The output node value is averaged, and a value closer to 1 implies a higher probability of success. This is repeated by comparing the unseen example with the failure support set and the output node value is averaged again, and a value closer to 1 implies a higher probability of failure this time. If the success support set has a higher average output value than the failure support set, it implies that the unseen example is more likely to be a success case and vice versa.

To illustrate the ability of the GCNN in providing useful structural information and show it is well trained in our work, we have constructed the adjacency matrix with some sample molecules.

Figure 3b shows the Euclidean distance between the calculated vectors of three ligand molecules. The shorter the Euclidean distance between two ligand molecules, the more similar they are in structure and property. This aligns with the chemical similarity between molecules with smaller Euclidean distance. For example, glutathione is a tripeptide featuring functional groups such as amine groups and amide groups, while the other three thiols in Figure 3b are simple thiols with hydrocarbon chains. Therefore, glutathione has the longest Euclidean distances with all three other thiols while the Euclidean distances between the three thiols are much shorter. Thus, the GCNN allows the neural network to generalize to an unseen set of molecules by considering how similar the unseen molecules are to the set of molecules in the training dataset via the Fp vector output. This is in contrast to training a neural network with molecules that are represented as one-hot vectors (no ability to generalize to new molecules) or with basic descriptors only (might not incorporate as much structural information as a GCNN, see Table 1).

We have experimented with six different ML classification models, including SVM, DNN, SNN, and all of them combined with GCNN, respectively, on dataset group I. Table 1 presents the three key statistical performance indicators of the classification models. The Matthews Correlation Coefficient (MCC) is used as the main indicator as it compares the model's prediction ability to a random guess. It calculates the Pearson Correlation Coefficient for a two-class confusion matrix and provides a measure of model performance that is unaffected by class imbalance. During the training, tenfold cross validation is used to validate the model. As most examples in the dataset feature a different set of key reaction components (leading to molecular information heterogeneity), each fold would likely have a test set that has a different set of reaction components from the training set. Thus, for the classification model to perform well during a tenfold cross validation, it has to generalize the information it learns from the training set onto the test set with different molecular information, achieving the first goal of performing the explorative task.

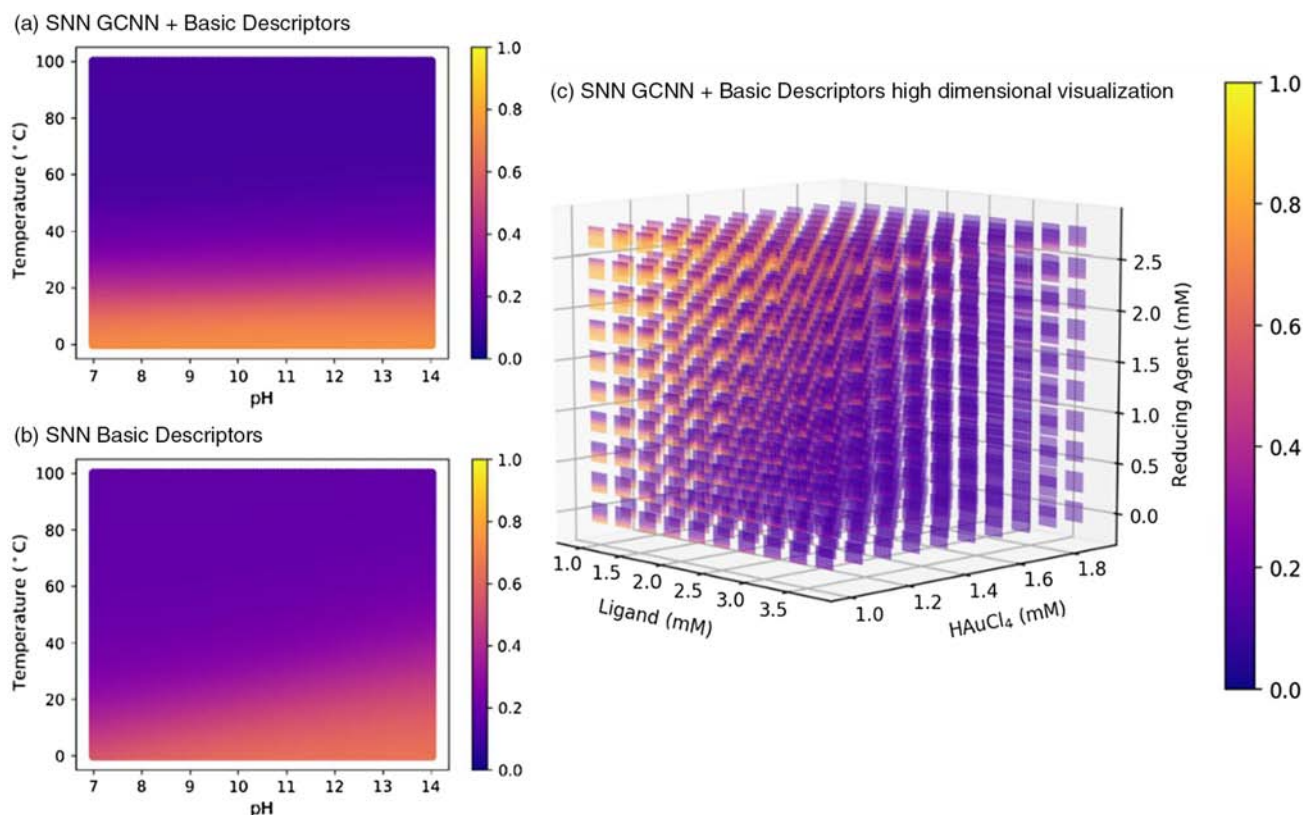
From Table 1, the combined GCNN models perform better in all three statistical indicators than models without GCNN. This is

**Table 1.** Key statistical performance indicators. Performance of six key classification models by using tenfold validation is described by three statistical indicators. Accuracy is a simple indicator of the proportion of correct predictions. The  $F_1$  score is the harmonic average of both recall and precision, where an  $F_1$  score reaches its best value at 1 and worst at 0. MCC compares the prediction ability of a model to a random guess. If the value is positive, it means it is better than random guess. The higher the MCC, the better the prediction ability with 1 being the highest value.

Models	Accuracy	$F_1$ Score	MCC
GCNN + SNN	0.81	0.79	0.65
GCNN + DNN	0.83	0.83	0.67
GCNN + SVM	0.80	0.80	0.60
SNN	0.72	0.65	0.46
DNN	0.69	0.65	0.37
SVM	0.48	0.39	−0.05

likely due to the GCNN ability in providing rich structural information, which enables the model to learn and use more molecular information regardless of the type of classification model used. This is important as it potentially allows for an easy method to incorporate molecular structure information by stacking a GCNN at the top and using its output as a feature vector on top of any type of general model. When comparing the models combined with GCNN, it seems that their performance is comparable to one another. However, the indicators in Table 1 mainly show the models' ability of molecular information learning and its proficiency in the explorative task. The second goal is to develop a model that can learn synthetic condition information, that is, how the reaction condition affects the probability of a successful synthesis. To investigate whether such learning has been achieved by the six models, we used the trained model as an "oracle" to generate 10 000 synthetic data, each with the same key reaction component (6-mercaptopurine acid-protected Au NCs by  $\text{NaBH}_4$  reduction) but with variations in pH and temperature. The 2D probability map of success for the models is plotted in Figure 4 and Figure S1, Supporting Information. The color at each point indicates the probability of success for that particular pH and temperature. Only, SNN-based models in Figure 4 show variations in probability when the reaction condition varies.

This suggests that the SNN-based models have learned the synthetic condition information that if the temperature exceeds a certain value (around  $50^\circ\text{C}$ ), the probability of a successful synthesis of atomically precise Au NCs will diminish to zero as the NCs start to decompose. This agrees with the conclusion from the literature and other experiments.<sup>[31]</sup> However, for the other four SVM- and DNN-based models (Figure S1a–d, Supporting Information), there is no obvious color variation in the 2D probability map, which indicates no synthetic condition learning achieved by these two models. Lastly, due to the lack of variations in the pH within our dataset, the SNN could not learn as much insights as temperature about how pH affects the synthesis. This makes the probability change not obvious along the pH axis; however, there is still pH-related probability variation. Apart from analyzing in only two dimensions, a 5D high-dimensional visualization is shown in Figure 4c showing the variations in the concentrations of all three key reaction components together with the variation in the temperature and the pH. This figure shows that our proposed model is capable of investigating the synthesis of atomically precise Au NCs, which is a rather complex system involving a number of variations, through simultaneous multidimensional study (a higher dimension can also be achieved but not suitable for visualization here).



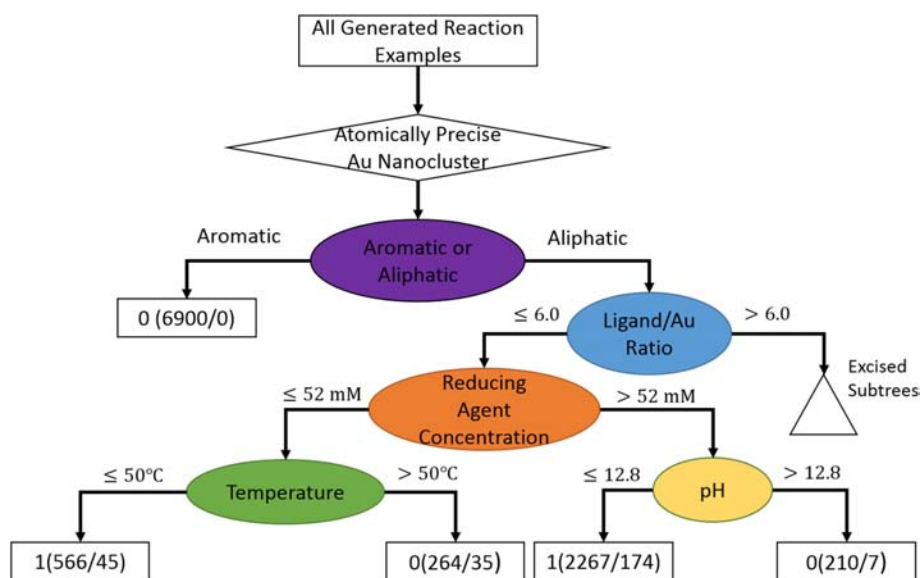
**Figure 4.** Two 2D probability heatmaps generated from six ML classification models and a 5D probability visualization. a,b) 2D probability heatmaps with the variations in pH and temperature predicted by SNN + GCNN model and SNN model, respectively. The x-axis is pH starting from 7 due to the lack of examples of the synthesis carried in acidic/neutral conditions. The y-axis is temperature with Degree Celsius as the unit. The color-scale indicates the probability of successful synthesis of atomically precise Au NC with yellow as 1 and dark blue as 0. c) 5D probability visualization showing the variations in the concentrations of all three key reaction components together with the variation in temperature and pH. The x, y and z axes are concentrations of ligand,  $\text{HAuCl}_4$ , and reducing agent in mM, respectively. At each point, there is a heatmap in pH and temperature similar to the one described above but with less data points (The center of the heat map is the coordinates for concentrations).

To validate the realizability of our proposed ML model, we use the best-performing GCNN + SNN model to predict the synthesis protocols that did not appear in our training dataset. Using previously published literature results, we have predicted the probabilities of obtaining atomically precise Au NCs from six different synthesis conditions.<sup>[32–36]</sup> These protocols were experimentally proved to be successful in synthesizing atomically precise Au NCs of various sizes (Au<sub>23</sub>, Au<sub>25</sub>, Au<sub>29</sub>, Au<sub>38</sub>, and Au<sub>44</sub>). As shown in Table S2, Supporting Information, the probabilities of successfully obtaining atomically precise Au NCs are high from the model prediction (with a range between 0.646 and 0.849). The high predicted probabilities indicate that our GCNN + SNN model is capable of identifying the optimized reaction conditions even though the model has not been exposed to those synthesis conditions. In addition, our ML model can handle predictions for various ligand molecules, reaction conditions, and solvent scenarios (including situations with two different solvents mixed together), leading to the possible synthesis of various-sized NCs. This suggests our two-step approach with the efficient classification model can be a useful platform for using ML strategy in accelerating synthesis process in the nanocluster field.

Deep learning methods such as GCNN + SNN are normally opaque to simple examination as “black-box” models. To gain chemical insights, we developed a “model of the model” by using the best-trained GCNN + SNN as a generative model to generate sufficiently large amounts of synthetic data, followed by using the synthetic data to train a decision tree. Firstly, the GCNN + SNN is trained and used to generate a large number of random examples (11 095 in total, as a random number around 10 000) by initializing all the descriptors randomly. We generated these examples because the original dataset is too small to train a decision tree, but this can be overcome by generating a large

synthetic dataset. Although this decision tree model will perform no better than the SNN model itself, the interpretations and implications can play an important role in understanding the Au NC synthesis and accelerating the domain development. As a proof of concept, we used the decision tree approach to generate chemical insights in synthesis of atomically precise Au<sub>25</sub> NCs in aqueous phase. Thus, dataset group II is adopted here. The calculated *F*<sub>1</sub> score and MCC value of the decision tree are 0.95 and 0.90, respectively, calculated by evaluating another 1000 randomly generated examples from the GCNN + SNN. Such high *F*<sub>1</sub> test value and MCC value indicate that the building of a “model of model” is promising as the decision tree well maps the well-trained “black-box” SNN model.

The decision tree is shown in the flowchart in Figure 5. From this flowchart, we can generate some synthetic chemistry guidelines to assist in designing the synthesis route. The decision tree examines the five reaction conditions in the synthesis of Au NCs in the aqueous phase including the chain of the ligand (aromatic or aliphatic), the ligand-to-Au molar ratio, the concentration of the reducing agent, pH of the reaction solution, and the reaction temperature. The probabilities of the successful synthesis are given based on the combination of all five conditions. The tree shows that firstly, when water is used as the solvent for the synthesis of Au NCs, using aliphatic ligands will have a much higher chance of obtaining atomically precise Au NCs compared with using aromatic ligands. This is consistent with the fact that the aromatic ligands are overall less soluble in water, and good ligand solubility is critical for a well-controlled synthesis. Secondly, the ligand-to-Au molar ratio is found to be an important factor for the aqueous-phase synthesis of Au<sub>25</sub> NCs. The model predicts that for successful synthesis, the ligand-to-Au molar ratio should be less than 6.0. The knowledge again matches with our understanding that Au(I)–ligand complexes



**Figure 5.** Decision tree in the prediction of the synthesis of atomically precise Au<sub>25</sub> NCs in aqueous phase. Ovals represent decision nodes and rectangles represent reaction-outcome bins. Triangles mean excised subtrees due to both extra small examples in that branch and chemical intuition. The numbers on the arrows correspond to decision attributing test values. Each reaction-outcome bin (rectangle) corresponds to a specific reaction-outcome value (1 success, 0 failure). The number in the parentheses is the number of reaction examples correctly assigned to that bin (any incorrectly classified reactions are given after a slash).

of different size and structure will form at different ligand-to-Au molar ratios. The short Au(I)-ligand complexes formed at high ligand-to-Au molar ratios favor the formation of large Au nanoparticles instead of NCs.<sup>[12]</sup> Moreover, the effects of the reducing agent concentration have been learned and investigated. On one hand, if the reducing agent concentration falls below 52 mM, the reaction temperature should be kept below 50 °C for the formation of atomically precise Au<sub>25</sub>. This is probably because preventing the decomposition of Au NCs with low reaction temperature is important in this mild reduction environment.<sup>[32]</sup> On the other hand, a value of pH below 12.8 is found critical for the atomically precise Au<sub>25</sub> NC synthesis in the cases of reducing agent concentration above 52 mM. It should be noted that only alkaline conditions (pH > 7.0) have been trained for the pH values due to the lack of examples of the synthesis carried in acidic/neutral conditions. The weakly alkaline condition is important for simultaneously tuning the formation kinetics and thiol etching abilities in the growth of Au<sub>25</sub> NCs.<sup>[35]</sup>

### 3. Conclusion

To conclude, we have shown that ML accelerates the synthesis of atomically precise Au NCs by incorporating all parameters in the synthesis into consideration instead of focusing on one or two parameters in the experimental discoveries. Our purely data-driven results show that there is a quantitative probability pattern for the successful synthesis of atomically precise Au NCs based on the combinations of reagents and the reaction conditions. Moreover, the successful prediction of six new experimentally proven successful synthesis recipes showed the reliability of the model proposed. Within this two-step ML framework, chemical insights for this complex reaction system (comprising of various types of reactions toward an atomically precise product) have been effectively generated. The combinations of a few key parameters including ligand type, ligand-to-Au molar ratio, pH, and reaction temperature have been identified for the successful synthesis of atomically precise Au NCs from the model-of-model decision tree. In addition, our GCNN + SNN approach works well with low data (only 54 examples, instead of >1000 typically used in ML studies on chemical synthesis) which is a common situation where obtaining the desired product is challenging. Our work has provided a framework for classifying the diverse parameters in a chemical synthesis and elucidated ML applications in a complex chemical synthesis system with limited number of successful examples.

Despite the good explorative process performance of the proposed model, the performance in optimizing the reaction conditions needs further development. Currently, from the heterogeneous property of the available dataset, the key reaction components that lead to successful synthesis of atomically precise Au NCs can be identified. However, predicting the exact composition in the atomically precise Au NCs is not maturely available yet. This is due to the lack of both successful and unsuccessful examples for machines to classify. We anticipate the prediction ability and interpretation ability of the proposed framework with available models can be improved by further high-throughput experiments focused on a specific system with

fixed Au atom number. We identify this as a promising direction for future investigation.

### 4. Experimental Section

**Machine Learning:** The training of the six key classification models was constructed using Keras library. The GCNN was implemented using Keras Neural Graph Fingerprint codes and RDkit (<https://github.com/rdkit/rdkit>). SVM and DNN were well-known models; therefore, we briefly introduced our SNN model.

The GCNN considered the reaction molecules as inputs and gave a fingerprint vector  $V$ . It was then concatenated with the vector  $W$  representing the selected features from the reaction using a direct sum of the vector space to produce the intermediate input vector (IIV)  $X$ .

$$X = V \oplus W \quad (1)$$

Pairs of IIV were then put into the SNN. This model learned a measure of similarity. For an input  $X$ , a half model SNN parameterized by  $W$  will return result  $G_w(X)$ . The similarity metric for a pair of input  $X_1$  and  $X_2$  was<sup>[27]</sup>

$$E_w(X_1, X_2) = \|G_w(X_1) - G_w(X_2)\| \quad (2)$$

which was the L1 distance between the two outputs. The parameters  $W$  of the model need to be trained so that if  $X_1$  and  $X_2$  were in the same class (either success or failure), the similarity metric was small and vice versa.

Therefore, the loss function had contrastive terms for the input pair with the same classes and the different classes. We used binary cross entropy function and the general term for  $N$  inputs were written as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N L(W, (Y, X_1, X_2)^i) \quad (3)$$

$$L(W, (Y, X_1, X_2)^i) = (1 - Y) \log(E_w(X_1, X_2)) + Y \log(1 - E_w(X_1, X_2)) \quad (4)$$

Here,  $(Y, X_1, X_2)^i$  was the  $i$ -th example with inputs  $X_1$ ,  $X_2$ , and label  $Y$ .  $Y$  is 1 if the inputs were in the same class or 0 if they were in different classes. The log function was a smooth monotonically increasing function so that the minimization of the loss function maximized the L1 distance between the pair in different classes and minimized it for the same-class pairs.

The hyperparameters of these models are optimized by running 50 trials of Gaussian process optimization (gp minimize) using scikit opt library. The goal is to minimize MCC. The best hyperparameter based on the lowest MCC value out of the 50 trial runs was chosen (details in Supporting Information). The fingerprint Euclidean distance is calculated by taking the GCNN portion of the trained model. The SMILES code of selected molecules were transferred by RDkit into the graph tensor and then input into the GCNN portion to get the fingerprint vector output. Since we used tenfold cross validation, one trained model had ten instances each with a different training dataset. The fingerprint calculation steps for all ten instances were repeated and the average output was taken. The norms between various ligands to get the Euclidean distance adjacency matrix were taken. The 2D probability heatmap was constructed by repeating the following steps for all six key classification models. Ten thousand new examples using a mesh grid for pH and temperature were generated while the other experimental conditions and key reaction components were kept constant. The 10 000 examples were input into the trained model to get an evaluated output. The above two steps were repeated for all ten instances from tenfold validation and the average output was taken. A scatter plot was plotted for the 10 000 points with the pH and temperature as  $x$  and  $y$  axis and the colorscale as the probability of success. For 5D visualization, the same approach was adopted; however in each dimension, ten evenly spacing coordinates in a reasonable range similar to the experimental conditions were chosen (e.g. for pH we chose 7–14) and these synthesized



experimental conditions were predicted by the trained model. The three space coordinates represented the concentration of a reaction component and at each point a heat map similar to the one mentioned above was plotted. The details of the experimental validation prediction method are given in, Supporting Information. The model-of-model decision tree was constructed by randomly generating about 10 000 examples varying all variables. The generated examples were input into a trained model to classify into success or failure and use the output examples to train the decision tree using sklearn library with details in Supporting Information.

**Synthesis of Gold Nanoclusters:** The synthesis of gold nanoclusters was modified from Brust method. In general, a ligand solution was first added to the solvent, followed by the addition of HAuCl<sub>4</sub> solution. After that, the pH of the solution was adjusted to the desired value (for aqueous-phase synthesis). A reducing agent solution is then mixed with the reaction mixture for the reduction into Au NCs. The detailed parameters for the synthesis of the laboratory examples are listed in Supporting Information.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

J.L. and T.C. contributed equally to this work. J.L. and T.C. developed the database. J.L., K.L., and L.C. developed the machine learning models and statistical analyses, supervised by X.W. T.C. performed the experimental preparation and validation, supervised by J.X. S.A.K., J.X. and X.W. conceived and executed the project. J.L. and T.C. wrote the manuscript with input from all the co-authors. All authors discussed the results and commented on the manuscript. The authors thank Y. Zhang in designing Figure 2 and 3. The authors acknowledge the Singapore RIE2020 Advanced Manufacturing and Engineering (AME) Programmatic grant "Accelerated Materials Development for Manufacturing" by the Agency for Science, Technology and Research under Grant No. A1898b0043.

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

atomic precision, deep learning, gold nanocluster, low data, synthesis

Received: May 24, 2019  
Published online: July 8, 2019

- [1] I. Chakraborty, T. Pradeep, *Chem. Rev.* **2017**, *117*, 8208.
- [2] R. C. Jin, C. J. Zeng, M. Zhou, Y. X. Chen, *Chem. Rev.* **2011**, *116*, 10346.
- [3] A. Venzo, S. Antonello, J. A. Gascon, I. Guryanov, R. D. Leapman, N. V. Perera, A. Sousa, M. Zamuner, A. Zanella, F. Maran, *Anal. Chem.* **2011**, *83*, 6355.
- [4] P. D. Jadzinsky, G. Calero, C. J. Ackerson, D. A. Bushnell, R. D. Kornberg, *Science* **2007**, *318*, 430.
- [5] I. Dolamic, S. Knoppe, A. Dass, T. Burgi, *Nat. Commun.* **2012**, *3*, 798.
- [6] Z. T. Luo, X. Yuan, Y. Yu, Q. B. Zhang, D. T. Leong, J. Y. Lee, J. P. Xie, *J. Am. Chem. Soc.* **2012**, *134*, 16662.
- [7] Y. Yu, Z. T. Luo, D. M. Chevrier, D. T. Leong, P. Zhang, D. E. Jiang, J. P. Xie, *J. Am. Chem. Soc.* **2014**, *136*, 1246.
- [8] C. M. Aikens, *J. Phys. Chem. Lett.* **2011**, *2*, 99.
- [9] Y. Negishi, K. Nobusada, T. Tsukuda, *J. Am. Chem. Soc.* **2005**, *127*, 5261.
- [10] M. Brust, M. Walker, D. Bethell, D. J. Schiffrin, R. Whyman, *J. Chem. Soc., Chem. Commun.* **1994**, 801.
- [11] R. C. Jin, H. F. Qian, Z. K. Wu, Y. Zhu, M. Z. Zhu, A. Mohanty, N. Garg, *J. Phys. Chem. Lett.* **2010**, *1*, 2903.
- [12] T. Le, V. C. Epa, F. R. Burden, D. A. Winkler, *Chem. Rev.* **2012**, *112*, 2889.
- [13] T. Kalil, C. Wadia, presented at NSTC, Materials Genome Initiative for Global Competitiveness, Washington, D.C., USA, June 2011.
- [14] B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.* **2009**, *1*, 31.
- [15] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem., Int. Ed.* **2016**, *55*, 5904.
- [16] S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Myers, *Angew. Chem. Int. Ed.* **2015**, *54*, 3449.
- [17] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434.
- [18] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.
- [19] S. M. Copp, A. Gorovits, S. M. Swasey, S. Gudiband, P. Bogdanov, E. G. Gwinn, *ACS Nano* **2018**, *12*, 8240.
- [20] P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, *Nat. Commun.* **2018**, *9*, 1668.
- [21] H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 283.
- [22] G. R. Koch, R. Zemel, R. Salakhutdinov, presented at *ICML Deep Learning Workshop, Siamese Neural Networks for One-Shot Image Recognition*, Lille, France, July 2015.
- [23] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Adv. Neural Information Process. Systems* **2015**, *28*, 2224.
- [24] T. K. Chen, Z. T. Luo, Q. F. Yao, A. X. H. Yeo, J. P. Xie, *Chem. Commun.* **2016**, *52*, 9522.
- [25] Y. Yu, X. Chen, Q. F. Yao, Y. Yu, N. Yan, J. P. Xie, *Chem. Mater.* **2013**, *25*, 946.
- [26] J. F. Parker, J. E. F. Weaver, F. McCallum, C. A. Fields-Zinna, R. W. Murray, *Langmuir* **2010**, *26*, 13650.
- [27] T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang, *J. Chem. Information Model.* **2007**, *47*, 2140.
- [28] D. Weininger, *J. Chem. Information Model.* **1988**, *28*, 31.
- [29] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Adv. Neural Information Process. Systems* **2012**, 1097.
- [30] S. Chopra, R. Hadsell, Y. LeCun, in *Proc. of Computer Vision and Pattern Recognition*, IEEE, San Diego, CA **2005**, p. 539.
- [31] T. K. Chen, Q. F. Yao, X. Yuan, R. R. Nasaruddin, J. P. Xie, *J. Phys. Chem. C* **2017**, *121*, 10743.
- [32] Q. F. Yao, X. Yuan, V. Fung, Y. Yu, D. T. Leong, D. E. Jiang, J. P. Xie, *Nat. Commun.* **2017**, *8*, 927.
- [33] Y. Yu, Q. F. Yao, K. Cheng, X. Yuan, Z. T. Luo, J. P. Xie, *Part. Part. Syst. Char.* **2014**, *31*, 652.
- [34] Q. F. Yao, V. Fung, C. Sun, S. D. Huang, T. K. Chen, D. E. Jiang, J. Y. Lee, J. P. Xie, *Nat. Commun.* **2018**, *9*, 1979.
- [35] X. Yuan, B. Zhang, Z. T. Luo, Q. F. Yao, D. T. Leong, N. Yan, J. P. Xie, *Angew. Chem., Int. Ed.* **2014**, *53*, 4623.
- [36] Y. Yu, Z. T. Luo, Y. Yu, J. Y. Lee, J. P. Xie, *ACS Nano* **2012**, *6*, 7920.