# Supporting Information for "An Artificial Neural Network Reveals the Nucleation Mechanism of a Binary Colloidal $AB_{13}$ Crystal"

Gabriele M. Coli* and Marjolein Dijkstra*

*Soft Condensed Matter, Debye Institute for Nanomaterials Science, Department of Physics, Utrecht University, Princetonplein 1, 3584 CC Utrecht, The Netherlands.*

E-mail: g.m.coli@uu.nl; m.dijkstra@uu.nl

## Multiple bond order parameters: the Shape Matching algorithm

The exemplary 2D projection of the (averaged) bond order parameters in Fig. 3 of the main text reveals that two bond order parameters are not sufficient to achieve a satisfactory classification. An artificial neural network (ANN) is an efficient and optimised way of using multiple descriptors in the classification. In order to provide a more local description of a particle, we employed 36 *non-averaged* bond order parameters (BOP) as descriptors, even though they offer a poorer characterisation of the particles with respect to their *averaged* counterparts. An ANN combines these parameters by giving them different weights in the classification and adding non-linearity if one or more hidden layers are present in the chosen ANN architecture. However, there are also other techniques, outside the realm of machine learning, that are based on more than 2 descriptors in the classification. One of these methods is the shape matching technique (SM).[1]

The SM technique requires a set of *shape descriptors* or order parameters that can identify on a single-particle level the different phases we wish to classify. For each particle $i$ in the system, we build a vector $\vec{\mathbf{S}}_i$ consisting of all the values of the shape descriptors for particle $i$. Subsequently, we determine the reference vector $\vec{\mathbf{S}}_A$ of all the shape descriptors for the respective phases, *e.g.* phase $A$, we like to identify. By employing a *similarity metrics*, we find the reference vector that is most similar to the one determined for particle $i$, and particle $i$ will be classified accordingly.

In order to compare the performance of the SM algorithm with the ANN used in this work, we define the shape descriptor vector to be the same as the input vector of the ANN, which is an array of 36 non-averaged BOPs (Eq. 2 of the main text).

Subsequently, we determine the reference vectors for the four classes that we wish to identify. Instead of using the values of the shape descriptors for the perfect crystalline structures, we take the average values of the BOPs for each class. This approach corresponds to selecting the central point of the cloud in the 36-dimensional BOP space of each phase. We find that this choice leads to better results in the classification of particles than using the perfect crystalline structures. Moreover, it can be straightforwardly be extended to non-crystalline structures—the fluid phase in this case.

Finally, a similarity metrics has to be defined that quantifies the similarity of the shape descriptor vector $\vec{\mathbf{S}}_i$ of particle $i$ and the reference vector $\vec{\mathbf{S}}_A$ of phase $A$, $M(\vec{\mathbf{S}}_i, \vec{\mathbf{S}}_A)$. The most common metrics is the Euclidean distance

$$M_{dist}(\vec{\mathbf{S}}_i, \vec{\mathbf{S}}_A) = 1 - [(\vec{\mathbf{S}}_i - \vec{\mathbf{S}}_A)/(|\vec{\mathbf{S}}_i| + |\vec{\mathbf{S}}_A|)]. \tag{1}$$

Another choice is based on the projection of vector $\vec{\mathbf{S}}_i$ onto the reference vector $\vec{\mathbf{S}}_A$

$$M_{proj}(\vec{\mathbf{S}}_i, \vec{\mathbf{S}}_A) = 1/2[1 + (\vec{\mathbf{S}}_i \cdot \vec{\mathbf{S}}_A)/(|\vec{\mathbf{S}}_i||\vec{\mathbf{S}}_A|)] \tag{2}$$

We find that $M_{proj}$ gives the best results. The accuracy of the SM method based on $M_{proj}$ is reported in Table 1 along with the accuracy of the ANN used in this work.

Table 1: Accuracies of the SM technique and the ANN for all four classes.

| Class | SM | ANN |
|---|---|---|
| $AB_{13}$ - Large | 100.0% | 100.0% |
| $AB_{13}$ - Small | 89.8% | 98.1% |
| Fluid | 91.8% | 97.8% |
| fcc | 93.2% | 99.4% |

Although the accuracy is high for *e.g.* the large species of the $AB_{13}$ crystal, the overall accuracy is not satisfactory enough for the purpose of this work. The reason is that in the SM technique each BOP is equally important as the classification is simply based on a dot-product. Consequently, if a BOP is very noisy in the training set, it will obfuscate the classification. The remedy for this problem is to employ different weights for different BOPs, based on their importance. This is exactly what the ANN learns during the training phase, when the weights are tuned.

# Details on the Artificial Neural Network

## Choice of architecture

In this work, we employ an ANN composed of two hidden layers, in addition to the input and output layer. In principle, multiple architectures are possible and the choice of the number of hidden layers requires a trial-and-error tuning.[2,3] In particular, adding hidden layers to the neural network means more hyperparameters (namely weights and biases) that have to be tuned during the training phase, which allow for a better total score on the training set, but may also result in a higher chance of *overfitting* the data. The latter is signaled by the score on the validation set, which is not as high as the score on the training set.

In order to select the architecture of the ANN, we trained multiple ANNs, differing from each other solely by the number of hidden layers. In Table 2 we show the accuracy of each

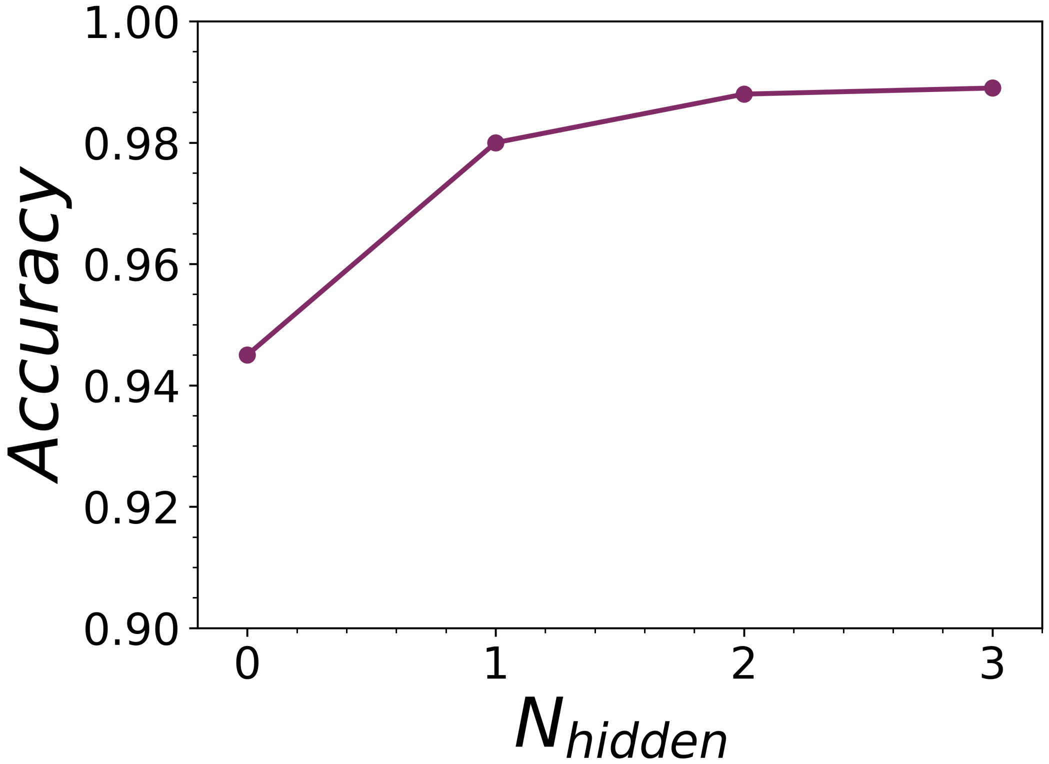

Figure 1: Total accuracy of the ANNs as a function of the number of hidden layers $N_{hidden}$. The total accuracy increases significantly up to 3 hidden layers, but the addition of the third hidden layer does not constitute a significant improvement in the performance of the ANN.

of these ANNs on the test set, while in Fig. 1 we display the total accuracy as a function of the number of hidden layers $N_{hidden}$.

We observe from Fig. 1 that, even though the total accuracy of the ANN increases with the addition of up to 3 hidden layers, the third layer does not constitute a significant improvement in the performance of the ANN. We therefore decided to employ an ANN with 2 hidden layers as a compromise between efficiency and accuracy.

Table 2: Accuracies for all four classes of the ANNs with varying number of hidden layers $N_{hidden}$.

| Class | $\mathrm{N}_{hidden} = 0$ | $\mathrm{N}_{hidden} = 1$ | $\mathrm{N}_{hidden} = 2$ | $\mathrm{N}_{hidden} = 3$ |
|---|---|---|---|---|
| $AB_{13}$ - Large | 100.0% | 100.0% | 100.0% | 100.0% |
| $AB_{13}$ - Small | 95.9% | 97.4% | 98.1% | 99.2% |
| Fluid | 94.1% | 96.3% | 97.8% | 96.9% |
| fcc | 97.9% | 99.1% | 99.4% | 99.6% |

## Fluid particles

When building the training and validation sets to train the ANN, one of the first questions to ask is how many classes the ANN should identify. This choice depends also on the features that are used in the input layer, as they must be able to distinguish the different classes in order to achieve a successful score at the end of the training phase.

In our work, we have selected 4 target classes: large species of the $AB_{13}$ crystal, small species of the $AB_{13}$ crystal, fcc particles and fluid particles. Regarding the fluid particles, one can make a further distinction based on species. This choice would split the fluid particles into large-species and small-species fluid particles, leading to a total of 5 different classes.

This different choice is valid based on the assumption that in the binary fluid with stoichiometry $x_L = 1/14$, large and small species have different local environments. This assumption is indeed correct, as can be seen from Fig. 2, where we plot the probability distribution functions of 4 exemplary BOPs, namely $q_4$, $q_6$, $w_4$ and $w_6$ for the large and small species of the binary fluid phase separately. We clearly see from Fig. 2 that almost all probability distribution functions overlap considerably, hampering the distinction of the two classes. Nevertheless, the two $q_6$ distribution functions are well separated for the large and small species, allowing for the possibility of selecting an output layer with 5 different classes.

To this end, we build a second training set with $10^5$ large-species and $10^5$ small-species fluid samples. We note that in the previous training set, the ratio between large-species fluid samples and small-species fluid samples was approximately equal to the stoichiometry of the system under investigation, and hence the number of large-species fluid samples was much lower than for the small-species counterpart. In Table 3 we show the accuracy of the ANN when the fluid particles are distinguished in large and small species.

We observe from Table 3 that the ANN correctly classifies almost all the large species of the fluid phase, whereas the accuracy of the large species of the $AB_{13}$ crystal and of the fcc crystal remains unchanged. The similarity of the local environments of the small species of $AB_{13}$ crystal and of the small species of the fluid phase is still signaled by the relatively low

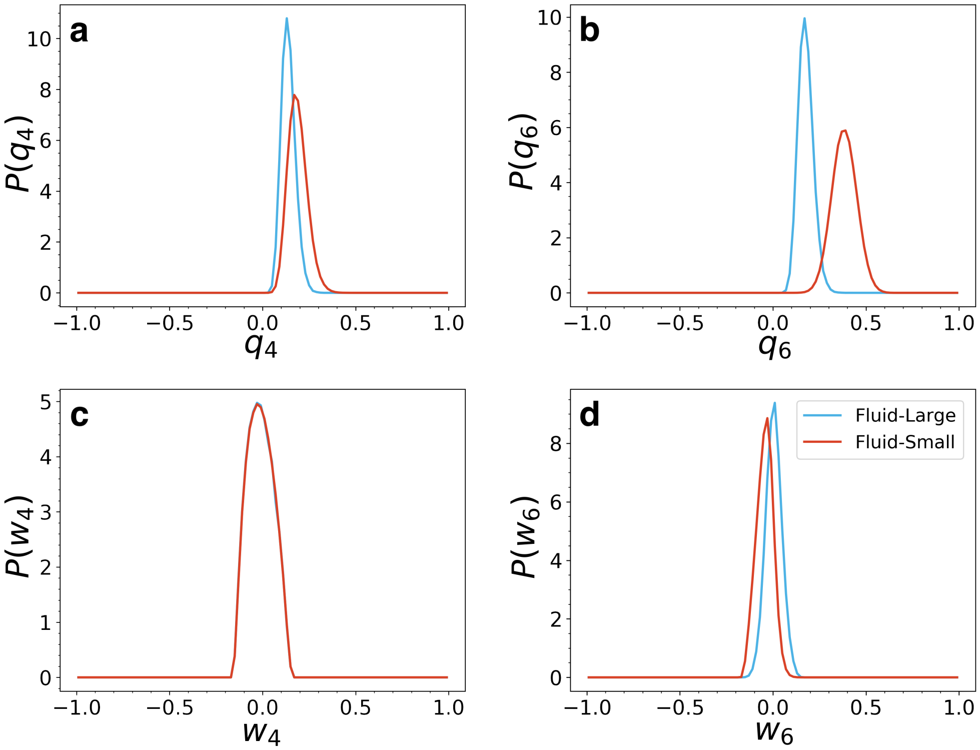

Figure 2: Probability distribution functions of 4 representative BOPs for large species of the fluid phase (light blue curves) and small species of the fluid phase (red curves). In particular, we plot a) $P(q_4)$, b) $P(q_6)$, c) $P(w_4)$, and d) $P(w_6)$. The majority of the BOPs show overlapping distributions (in the case of $w_4$ they are hardly distinguishable), but the fact that there is at least one parameter ($q_6$) for which the distributions are well separated guarantees that the distinction of small-species and large-species fluid particles is possible.

score of the ANN, and hence a correct classification of the small species remains a challenge.

In conclusion, we have shown that an ANN with five target classes involving a further distinction of the fluid particles is equally accurate as the ANN with four classes as employed in the present study, and represents a suitable alternative choice.

## Interface detection

One of the main difficulties common to all nucleation studies is to correctly identify the location of the interface between the solid nucleus and the surrounding fluid phase. In this section we show the performance of the ANN in carrying out this task.

A common approach to classify particles according to the different thermodynamic phases

Table 3: Accuracies of the ANN trained to distinguish the small-species and large-species binary fluid particles.

| Class | ANN |
|---|---|
| $AB_{13}$ - Large | 100.0% |
| $AB_{13}$ - Small | 98.3% |
| Fluid - Large | 99.9% |
| Fluid - Small | 97.3% |
| fcc | 99.3% |

is to determine the typical BOP values of all the reference phases and to select "by hand" a threshold value, called a *decision boundary.* Subsequently, the BOP values can be measured for each particle in the system, and classified according to these thresholds. In the case of a two-phase coexistence, the particles at the interface of the two coexisting phases will have BOP values which are close to these decision boundaries. Hence, the choice of these thresholds is fundamental from a quantitative point of view.

In the case of an ANN, the classification is again performed through a decision boundary. The main difference is that in this case the decision boundary is not selected manually but by the machine learning algorithm through the minimization of a loss function evaluated for a training set.

To verify the performance of the ANN in identifying the interface, we consider a system with an approximately spherical nucleus of an $AB_{13}$ crystal surrounded by a metastable fluid as obtained from a seeding simulation at a pressure where the crystalline seed has grown out and almost doubled its original size. For each particle, we determine the probability of belonging to the $AB_{13}$ crystal phase or the probability of belonging to the fluid phase. These probabilities are calculated through the trained ANN. In Fig. 3b, we show a cut-through image of the grown crystal nucleus along with the ANN output for 5 exemplary particles, *i.e.* the probability that a particle belongs to the $AB_{13}$-Large, $AB_{13}$-Small, fluid, and fcc phase. From the image, we observe that particles well-inside the crystalline nucleus are correctly classified with great certainty (probabilities around 99.9%) as belonging to the $AB_{13}$ crystal, while particles at the interface, *i.e.* where the crystal transforms into the disordered fluid

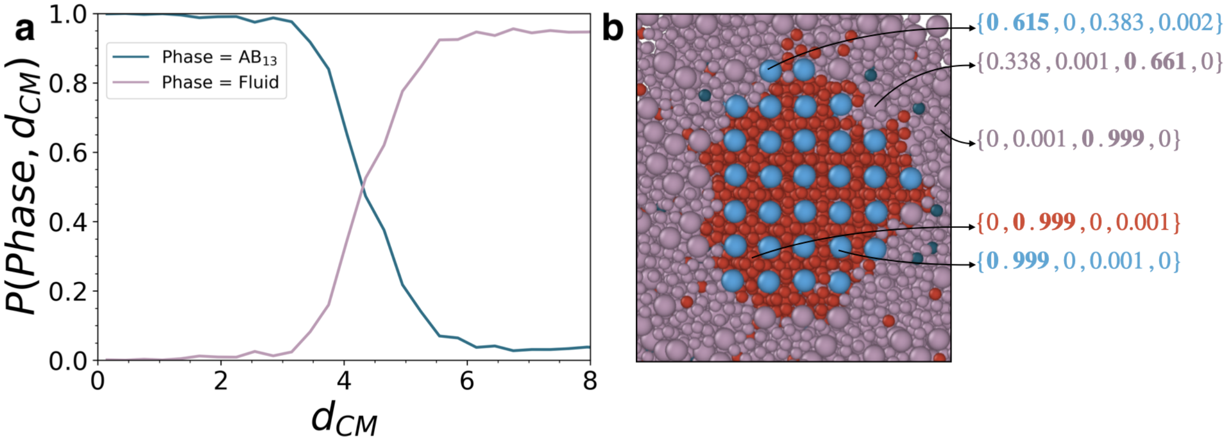

Figure 3: a) Probability that a particle belongs to one of the two classes of the $AB_{13}$ crystal (dark blue curve) or to the fluid phase (light purple curve) as a function of the radial distance $d_{CM} = |\mathbf{r}_i - \mathbf{r}_{CM}|/\sigma_L$ from the centre-of-mass position of the nucleus as predicted by the ANN. The analysis has been carried out on a single snapshot of a seeding simulation performed at $\beta P \sigma_L^3 = 51.3$. In b), we show a cut-through image of the grown nucleus along with the ANN output, *i.e.* the probability to belong to the $AB_{13}$-Large, $AB_{13}$-Small, fluid, and fcc phase, for 5 exemplary particles.

phase, the ANN gives similar scores to the probabilities of being crystalline or fluid-like. In the fluid phase, the ANN correctly identifies the fluid particles with again probabilities around 99.9%.

Subsequently, we calculate the radial distance of each particle $i$ with respect to the center-of-mass of the nucleus, $d_{CM} = |\mathbf{r}_i - \mathbf{r}_{CM}|/\sigma_L$, where $\mathbf{r}_i$ is the position of particle $i$, and $\mathbf{r}_{CM}$ is the center-of-mass position of the nucleus. In Figure 3a, we plot the radial averaged probability that a particle belongs to one of the two classes of the $AB_{13}$ crystal (dark blue curve) or to the fluid phase (light purple curve) as a function of the radial distance $d_{CM} = |\mathbf{r}_i - \mathbf{r}_{CM}|/\sigma_L$ from the centre-of-mass position of the nucleus. We make the following observations. Figure 3a shows that particles with a radial distance $d_{CM} < 3$, which corresponds to the solid bulk of the nucleus, are indeed classified by the ANN as particles belonging to one of the two classes of the $AB_{13}$ phase with a probability close to 1. The probability that these particles belong to the fluid phase is nearly zero. Upon approaching the interface at $d_{CM} \simeq 4.3$, the probability that a particle belongs to the $AB_{13}$ phase gradually decreases, while the probability that a particle belongs to the fluid phase increases. The probability that a particle belongs to the fluid phase is about equal to the probability that a particle belongs to the $AB_{13}$ phase at $d_{CM} \simeq 4.3$, and exceeds the probability of the $AB_{13}$ phase for $d_{CM} > 4.3$. Almost all particles are identified as fluid-like for $d_{CM} > 6$, which corresponds to the surrounding fluid phase.

In conclusion, the ANN effectively recognises the solid core of the $AB_{13}$ phase and the surrounding fluid, as well as the interface between the two bulk phases present in the system. The interfacial width is about $2\sigma_L$ as expected. We remark that the *non-averaged* BOPs enable us to locate the interface much more accurately than the averaged BOPS as the non-averaged BOPs only take into account the first shell of neighboring particles. Hence, we use the non-averaged BOPs in our nucleation study.

## Information provided by the network

As shown by our previous analysis, we clearly observe that our trained ANN constitutes a useful tool to distinguish with high accuracy the $AB_{13}$, binary fluid, and fcc phase on a single-particle level by using input vectors composed of 36 (non-averaged) BOPs. Hence, the ANN can straightforwardly be used as an order parameter in nucleation studies. However, such a black box implementation of the ANN can be unsatisfactory in terms of understanding how it makes decisions. We therefore investigate in more detail which features the ANN prioritises to distinguish the different classes. To this end, we apply the improved stepwise[4,5] and the input perturbation[6,7] techniques in order to determine the relative importance (RI) of all features.

In the improved stepwise technique, we determine the relative importance (RI) of a feature by replacing each value in the data set with its arithmetic mean, while keeping the remaining 35 features unchanged. Subsequently, we calculate the accuracy of the ANN on this data set. If the change in the accuracy is large, the relative importance of this feature is high in making a prediction.

The input perturbation method works in a similar manner. Instead of replacing each

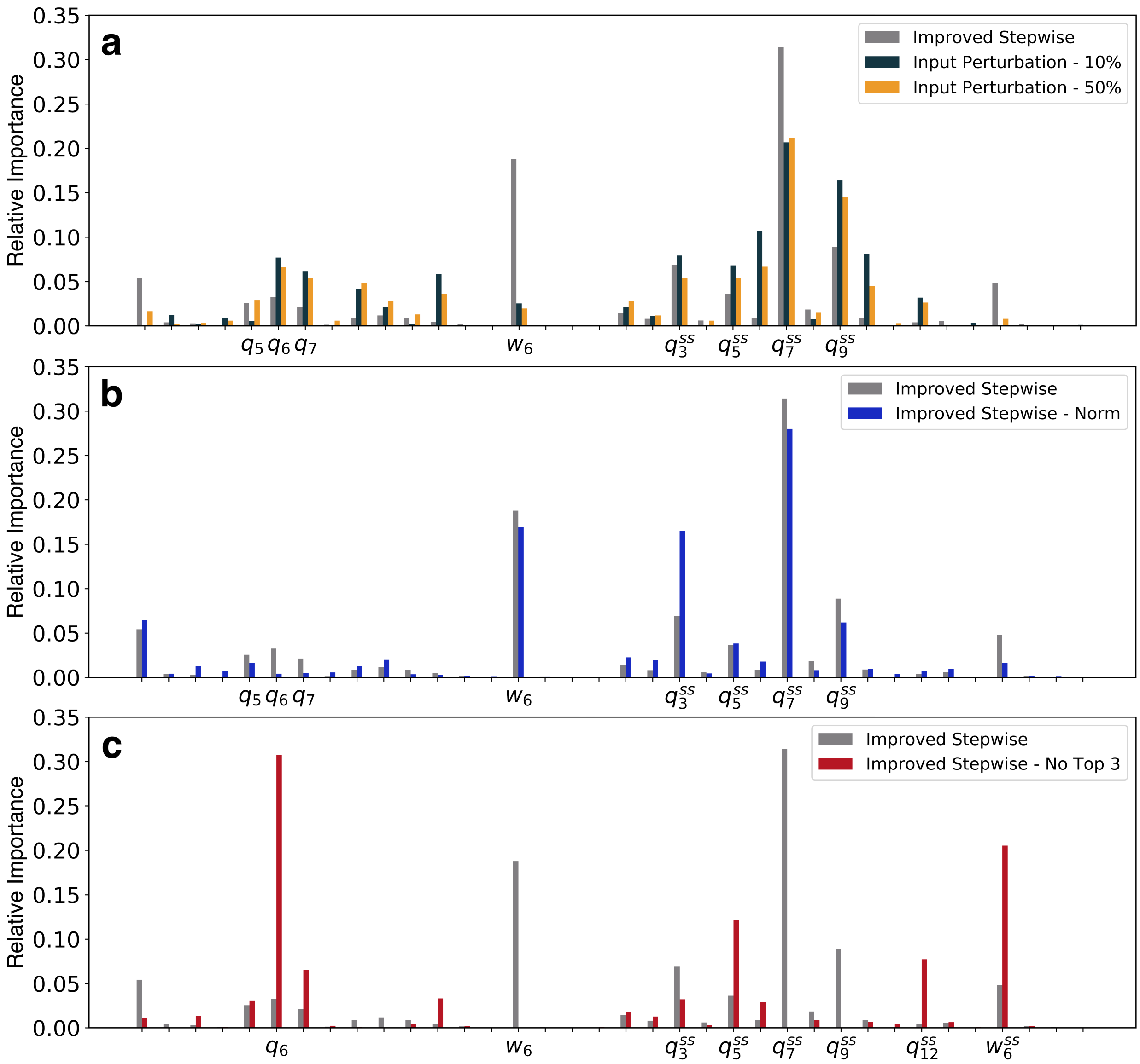

Figure 4: a) The relative importance RI($k$) of feature $k$ of our input vector (Eq. 2 in the main text) describing the local particle environment as determined by the improved stepwise technique and the input perturbation method using Gaussian white noise with a 10% and 50% variance. b) Application of the improved stepwise method on the original data set (grey) and a normalised data set (blue) such that each feature has a distribution with zero mean and unit variance. c) Application of the improved stepwise method of the original ANN (grey) and a second one trained on all the features with the exception of the top 3 most important ones (red).

value of a particular feature with its mean, we add a fixed amount of white noise, and then again assess the change in the total accuracy. The noise is sampled from a uniform distribution, centered in zero and whose interval is equal to twice a fixed percentage of the input value. We test this method twice, selecting the above percentage to be equal first to 10% and then to 50% of the input value.

In both methods, we repeat this process for all 36 input features, and calculate the relative importance of the $k^{th}$ feature RI($k$) using

$$\mathrm{RI}(k) = \frac{\Delta\mathrm{Acc}(k)}{\sum\limits_{k=1}^{36} \Delta\mathrm{Acc}(k)}, \tag{3}$$

where $\Delta$Acc($k$) is the difference in accuracy after applying one of the two methods to the $k^{th}$ feature.

In Fig. 4a, we show our results as obtained by applying the two methods on the entire data set. It is interesting to note that the ANN does not solely prioritise even symmetries, which often play a decisive role in crystal structure recognition. In fact, odd symmetries - fivefold, sevenfold and ninefold - heavily determine the learning process of the neural network, as shown by the high RI values of the corresponding order parameters. We attribute this behavior to the unique crystal structure of the $\mathrm{AB}_{13}$ phase consisting of icosahedral small-sphere clusters.

However, due to the different nature of the two algorithms, the RI($k$)'s vary - sometimes considerably - for the two methods. This analysis is therefore useful solely to understand which features the ANN relies the most on. As further proof, we train another ANN with the same architecture but such that the probability distribution of each feature has zero mean and unit standard deviation. Usually, this pre-processing is applied in cases where the input vector is composed of features that vary by orders of magnitude, facilitating in this way the training of the model.[3] In this work, this pre-processing has not been applied because all the BOPs vary approximately within the same range. We again apply the improved stepwise method and compare the results with the improved stepwise method applied on the original features in Fig. 4b. We clearly observe that the values of the relative importance vary due to differences in the numerical values of the normalized and unnormalized features, but the features that ANN recognises as the most important for the classification remain the same.

Finally, one may ask the question whether a smaller set of BOPs can also provide a satisfactory classification. We therefore trained two additional ANNs, one for which we use only the top three features picked up by the improved stepwise method as the most relevant ones, namely $w_6$, $q_7^{ss}$, and $q_9^{ss}$, as input vector, and one ANN for which we employ all the other features apart from these top three features as input.

In the first case using only the top 3 features we get a total accuracy of about 93%. This result demonstrates the importance of these 3 BOPs in order to obtain a satisfactory classification, even though it cannot replace the ANN trained on all 36 features.

In the second case using the remaining 33 features, the trained ANN achieves an accuracy almost identical to that of the original ANN (98.7% instead of 98.8%). This result shows that there are several ways to obtain a satisfactory classification of the different phases. By removing the top three relevant features, the ANN will give priority to other features, equally valid, as shown in Fig. 4c. This different set of relevant features can sometimes be highly correlated with the removed ones, as in the case of $w_6^{ss}$, which has a Pearson correlation value with $w_6$ equal to 0.87.

## Comparison with Random Forest

In the previous sections we have demonstrated that an ANN successfully classifies particles according to the different thermodynamic phases in the system. We have also compared the performance of the ANN with a technique outside the realm of machine learning, namely the SM technique. The performance of the ANN is superior to that of the SM technique. However, the variety of machine learning techniques is vast, and the ANN is only one of them. We therefore also compare the performance of the ANN with another machine learning method.

In this section we study a second state-of-the-art machine learning algorithm, namely

Random Forest (RF), and evaluate its accuracy on the same data set used for the ANN. This allows us to determine which technique works best for the purpose of this work.

RF is a machine learning algorithm with a philosophy that is very different from that of an ANN. RF, in its classification setting, consists of an ensemble of smaller classifiers called Decision Trees (DT).[3] Each of these DTs takes a given input, and applies a series of *if-else* conditions based on the value of the input vector features. These conditions allow the tree to branch out and finally return a result, the predicted class of the sample. In an RF, multiple DTs are trained using different parts of the training set and of the different features. For each sample to be classified, each DT votes for a certain class based on its prediction. The RF returns the class that receives the most votes. RF often succeeds not only in providing satisfactory accuracy results, but also on shedding light on the predictive ability of individual features.

To offer a comparison with the ANN used in this study, we train a RF model using 100 independent DTs. An important parameter that must be tuned is the depth of each tree, *i.e.* the number of consecutive *if-else* statements prior to a classification outcome. In fact, if this parameter is too high, RF is likely to overfit the data. We find that the maximum depth of the model for which we do not overfit the training set is equal to 9. Below we show the accuracy for each class obtained using RF, together with the results of the ANN used in this work.

Table 4: Accuracies of the RF and the ANN for all four classes.

| Class | RF | ANN |
|---|---|---|
| $AB_{13}$ - Large | 100.0% | 100.0% |
| $AB_{13}$ - Small | 96.6% | 98.1% |
| Fluid | 97.4% | 97.8% |
| fcc | 98.1% | 99.4% |

Table 4 shows that the RF produces excellent results, although slightly lower than the ANN, particularly in the classification of fluid particles and small species of the $AB_{13}$ crystal.

Depending on the specific application, RF may be preferred when accuracy is not the

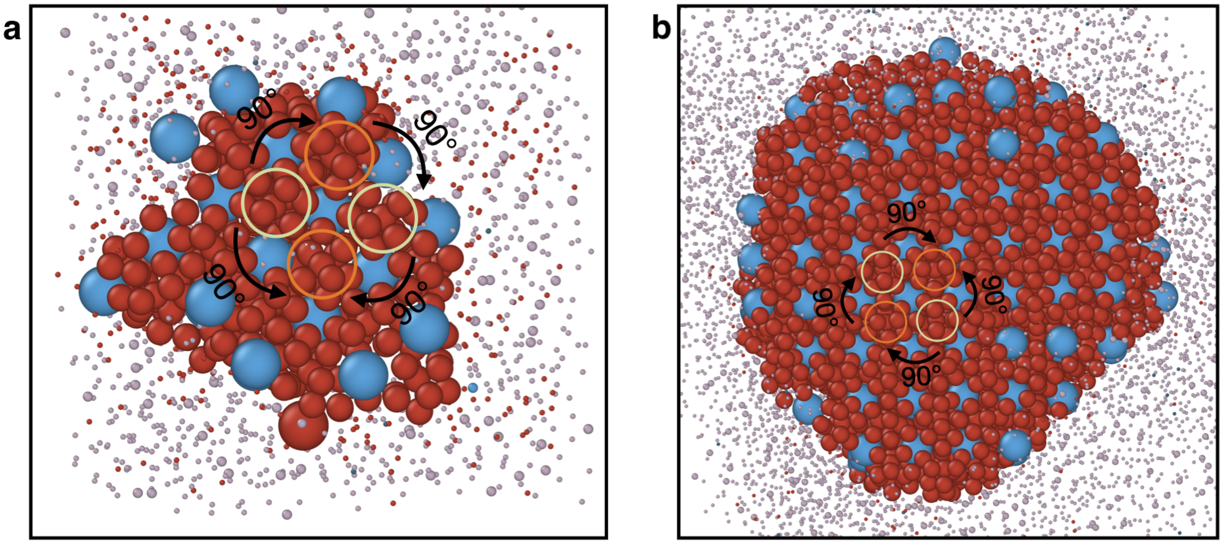

Figure 5: Cut-through of an $AB_{13}$ crystal formed *via* a) spontaneous nucleation and b) the seeding approach, demonstrating that adjacent icosahedral clusters of small particles are rotated by 90 degrees with respect to each other.

only parameter to take into account, *e.g.* with RF it is possible to obtain a very fast training of the model and a better interpretability of the classification. In this work, we require a method that is as accurate as possible as the number of solid particles in the crystal nucleus as predicted by the machine learning method determines the thermodynamics of the nucleation process. We therefore decided to employ the ANN in our nucleation study.

## Icosahedral clusters

In this section, we show in Fig. 5 that adjacent icosahedral clusters of small species in the $AB_{13}$ crystal formed *via* spontaneous nucleation and the seeding approach are rotated by 90 degrees with respect to each other.

## References

1. Keys, A. S.; Iacovella, C. R.; Glotzer, S. C. Characterizing Complex Particle Morphologies through Shape Matching: Descriptors, Applications, and Algorithms. *Journal of Compu-*

*tational Physics* **2011**, *230*, 6438–6463.

2. Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer Science+Business Media, New York, NY, 2006.

3. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer series in statistics New York, 2001; Vol. 1.

4. Gevrey, M.; Dimopoulos, I.; Lek, S. Review and Comparison of Methods to Study the Contribution of Variables in Artificial Neural Network Models. *Ecological modelling* **2003**, *160*, 249–264.

5. Olden, J. D.; Joy, M. K.; Death, R. G. An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data. *Ecological modelling* **2004**, *178*, 389–397.

6. Yao, J.; Teng, N.; Poh, H.-L.; Tan, C. L. Forecasting and Analysis of Marketing Data Using Neural Networks. *J. Inf. Sci. Eng.* **1998**, *14*, 843–862.

7. Scardi, M.; Harding Jr, L. W. Developing an Empirical Model of Phytoplankton Primary Production: A Neural Network Case Study. *Ecological modelling* **1999**, *120*, 213–223.