PHYSICAL REVIEW E **96**, 011301(R) (2017)

Machine-learning approach for local classification of crystalline structures in multiphase systems

C. Dietz, T. Kretz, and M. H. Thoma

I. Physikalisches Institut, Justus Liebig Universität Giessen, Heinrich-Buff-Ring 16, D 35392 Giessen, Germany

(Received 30 January 2017; revised manuscript received 20 June 2017; published 19 July 2017)

Machine learning is one of the most popular fields in computer science and has a vast number of applications. In this work we will propose a method that will use a neural network to locally identify crystal structures in a mixed phase Yukawa system consisting of fcc, hcp, and bcc clusters and disordered particles similar to plasma crystals. We compare our approach to already used methods and show that the quality of identification increases significantly. The technique works very well for highly disturbed lattices and shows a flexible and robust way to classify crystalline structures that can be used by only providing particle positions. This leads to insights into highly disturbed crystalline structures.

DOI: 10.1103/PhysRevE.96.011301

I. INTRODUCTION

Identifying crystalline structures is a common task in physics. The behavior of solids strongly depends on the underlying structural properties. Various analysis methods have been developed to retrieve reliable information about crystals. Some of the most popular methods are bond-orientational order parameter (BOOP) [1], bond angle analysis (BAA) [2], common neighbor analysis (CNA) [3], centrosymmetry parameter (CSP) [4], common neighborhood parameter (CNP) [5], and topological cluster classification (TCC) [6].

In the case of BOOP, BAA, and CSP the algorithms implement an order parameter derived from the location of neighboring particles, while methods like CNA and TCC identify topological symmetries of the neighborhood of a particle. CNP combines the features of CSP and CNA.

These methods all have in common that they try to find certain symmetries to identify crystalline structures. In other words, all of these methods solve classification problems, because they assign particles into categories. This work will consider the structures fcc, hcp, and bcc.

Fortunately, classification problems are a large part of the thriving field of machine-learning algorithms, with powerful libraries like Tensorflow [7] and scikit-learn [8] being freely available. For example, these algorithms have been successfully applied to find structural flow defects in disordered solids [9] and for characterizing the "softness" of particles in glassy liquids [10]. A crystal analysis method capable of finding structures autonomously using machine-learning techniques was recently published by Reinhart *et al.* [11].

It is also possible to calculate physical quantities in theoretical physics using machine learning. The Curie temperature of an Ising model could be predicted with surprising accuracy using a neural network [12]. A combination of unsupervised feature extraction and a neural network was able to predict phase transitions in a Kitaev chain [13].

Machine-learning algorithms are designed for a general purpose, which means that they have to "learn" the properties of a specific classification problem from labeled data. Because of this, we train such an algorithm from artificial fcc, hcp, and bcc lattices in three dimensions using only the particle positions as input. By restricting the method only to the positions, it is possible to apply it to a wide area of structural analysis problems, for example, plasma crystals [14], molecular-dynamical simulations [15], or colloidal crystals [16].

However, it is required to choose a well-defined representation of the structural properties in which the different structures can be well separated. This representation is called "feature vector."

II. FEATURE VECTOR

The features of a particle which is in a crystalline structure can be described by the relative coordinates of a particle i and its neighbors j. In the case of 12 neighbors, this would lead to a 36-dimensional feature vector, which could be used for classification. However, this would not be a robust method. The rotations of the crystal would have to be considered and trained, leading to an inflated dimensionality of the problem. Scaling or translating the crystal would be an analogous challenge, which needs to be addressed. This is why we propose a feature vector that is translational, rotational, and scale invariant.

We use several proven methods and generalize them in a fixed neighborhood to create such a feature vector, which we call "crystal signature."

First of all, it is important to choose the definition of neighborhood to be used. Because we intend to calculate Voronoi cells later on, the most natural neighborhood is the corresponding Delaunay neighborhood [17]. Alternatives would be the use of a cutoff radius or n-nearest neighbors, however, they are not automatically well defined for mixed phase systems, because they do not adapt to the local structure (unless additional information for the structure is used [15]). Also they require an additional input by the user, which can significantly influence the quality of identification and unnecessarily complicates the process.

After the neighborhood is obtained for a particle i, the signature can be defined. For the first part of the signature, we determine all of the distances between neighbors j and k:

$$d_{jk}(i) = \frac{1}{d_0(i)}\|\mathbf{r}_j - \mathbf{r}_k\| \quad (j \neq k), \tag{1}$$

$$d_0(i) = \frac{1}{6}\sum_{j=1}^{6}\|\mathbf{r}_i - \mathbf{r}_j\|, \quad \|\mathbf{r}_i - \mathbf{r}_j\| \leqslant \|\mathbf{r}_i - \mathbf{r}_{j+1}\|. \tag{2}$$

Scale invariance can be achieved by normalizing the distances with a characteristic length d_0, which is the average distance between a particle and its nearest six neighbors. We choose the six nearest neighbors, because this is the minimal number of neighbors we can expect in a three-dimensional Delaunay neighborhood.

Because the number of neighbors N may vary for every particle and there are $N/2(N-1)$ distances d_{jk}, it is practical to calculate a histogram from all d_{jk} with a certain number of equidistant bins, in this case $N_{\text{bins}} = 12$, as feature space. This number of bins works best for the method proposed here and is determined by trial and error.

We will not only consider distances, but also the angles between neighbors. The so-called "bond angles" between the particle i and a pair of neighbors j and k can be calculated as follows [2]:

$$\cos[\theta_{jk}(i)] = \frac{(\mathbf{r}_i - \mathbf{r}_j)\cdot(\mathbf{r}_i - \mathbf{r}_k)}{\|\mathbf{r}_i - \mathbf{r}_j\|\|\mathbf{r}_i - \mathbf{r}_k\|} \quad (j \neq k). \tag{3}$$

Analogous to the neighbor distances, we calculate a histogram with $N_{\text{bins}} = 8$ from the bond angles (borrowed from Ackland and Jones [2]) and extend the signature vector with it.

Because a clear distinction of fcc and hcp is challenging [2], a method based on the well-known BOOP of a particle i will be helpful. An improved and more robust implementation based on Voronoi cells, called Minkowski structure metric (MSM) by Mickel *et al.* [17], will be used throughout this work:

$$q'_{lm}(i) = \sum_{f\in\mathcal{F}(i)} \frac{A(f)}{A} Y_{lm}(\theta_f,\phi_f), \tag{4}$$

$$q'_l(i) = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{l} \left|q'_{lm}\right|^2}. \tag{5}$$

The parameter q'_l is the second order rotational invariant and will be calculated for $l = 4,6$ [1]. The azimuthal angle is denoted by θ_f and the polar angle by ϕ_f. The spherical harmonics Y_{lm} are weighted by the Voronoi facet area $A(f)$ for all facets $f \in \mathcal{F}(i)$. The sum of all facet areas is denoted as $A = \sum_{f\in\mathcal{F}(i)} A(f)$ [17].

We will also calculate the third order rotational invariant w'_l based on the MSM [18]:

$$w'_l(i) = \sum_{\substack{m_1,m_2,m_3\\ m_1+m_2+m_3=0}} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \times \frac{q'_{lm_1}(i)q'_{lm_2}(i)q'_{lm_3}(i)}{q'_l(i)^3}. \tag{6}$$

The Wigner 3-j symbol [19] is denoted by the expression within large parentheses.

The MSM possesses some drawbacks for phase mixtures of bcc and fcc (or hcp), such that an improved implementation, called BCCMSM, has been recently proposed [20]. However, this is not needed in our signature, because it is easy to distinguish bcc from fcc or hcp using neighbor distances and bond angles. On the contrary, the latter methods are not well suited to distinguish between fcc and hcp, which is where the MSM does particularly well. This means that the different methods will complement their weaknesses and add to their strengths.

TABLE I. Composition of the crystal signature used for classification.

Name	Symbol	Dimensions
Neighbor distances	d_{ij}	12
Bond angles	$\cos(\theta_{ijk})$	8
Minkowski structure metric	q'_l, w'_l	4
Minkowski tensor	$\zeta_1,\ldots,\zeta_6$	6
Number of neighbors	N	1

A method based on Minkowski tensors has been proposed to distinguish between fcc and hcp [21], which will be used to further improve the classification. These tensors can precisely quantify the shape of a convex surface. To be able to distinguish the isotropic cells of fcc and hcp, a translational and scale invariant tensor of fourth rank is needed [21]:

$$\left(W_1^{0,4}\right)_{\alpha\beta\gamma\delta} = \sum_f \frac{A(f)}{A} n_\alpha n_\beta n_\gamma n_\delta. \tag{7}$$

The Cartesian components of the triangulated facet normals are denoted by n_α for $\alpha = 1,2,3$. Similar to the MSM, the facet normals are weighted by the facet area $A(f)$. Writing the tensor as a symmetric 6×6 matrix using the Voigt notation [22,23] enables us to calculate the eigenvalues. The eigenvalues $(\zeta_1,\ldots,\zeta_6)$ are rotational invariant and will be included in the crystal signature [21].

Last but not least, the number of Delaunay neighbors itself depends on the local crystal structure (12 for undisturbed fcc/hcp and 14 for undisturbed bcc). Although this is easily disturbed with noise, this quantity is naturally available and can be included in the signature resulting in a small increase of the quality of classification. Table I shows an overview of the different components of the signature.

Topological methods like CNA or TCC are not used. The CNA is (compared to MSM) sensitive to noise and cannot be calculated using only one Voronoi cell, because next-to-nearest neighbors have to be taken into account [20]. This contradicts the goal of a strictly local method using only the direct Voronoi neighbors. The TCC requires a modified Voronoi algorithm to work reliably in thermally dislocated structures [6]. However, a core idea of the MSM is the use of a standard Voronoi neighborhood to mitigate inconsistencies due to different definitions of neighborhood [17], which is why the TCC will not be considered. A recent machine-learning approach from Reinhart *et al.* [11] exclusively relies on topological methods, which is in contrast to the methods proposed in this work.

The CSP has been tested as part of the signature, but it did not significantly improve the precision of classification.

Overall, the mixed crystal signature (MCS) consists of 31 dimensions, which allows for a reliable classification of different crystalline structures. Instead of using hard coded decisions [2,15] or polygonial regions in a diagram [20] to classify the structures, we will apply a machine-learning algorithm to this problem. With this approach, the classification is utterly separated from the signature, which adds flexibility to the method. Also, this means that both parts of the method can be optimized separately.

III. MACHINE LEARNING

We choose to only identify the crystalline structure of the particle, not whether the particle is in a solid or disordered state. To achieve this, we filter the disordered particles out by using a modified scalar product of BOOPs [24,25], which can be analogously defined for MSM [14]:

$$S(i) = \frac{1}{N}\sum_{j=1}^{N}\sum_{m=-6}^{6}\tilde{q}'_{6m}(i)\tilde{q}'^{*}_{6m}(j), \tag{8}$$

$$\tilde{q}'_{6m}(i) = \frac{q'_{6m}(i)}{\left(\sum_{m=-6}^{6}|q'_{6m}(i)|^2\right)^{1/2}}. \tag{9}$$

The number of neighbors is denoted by N and the scalar product is averaged over all neighbors j. This method allows us to set a single threshold for a particle i to decide whether it is in the solid or disordered phase. We consider a particle solid if $S(i) < 0.55$, which works best for our case.

With the filtering in place, we will then use a multilayer perceptron (MLP) with one hidden layer consisting of 250 neurons to classify fcc, bcc, and hcp using the scikit-learn library [8].

The feature space behaves nonlinearly and due to the artificially created datasets, it is possible to generate as much data as is needed to prevent the neural network from overfitting.

The MLP network has to be trained with an appropriate dataset to learn how to classify the crystal structures in our signature. To achieve this, we compute artificial data of fcc, hcp, and bcc lattices, called "training data." These three datasets will be disturbed by different levels of Gaussian noise, which resembles a Brownian motion of the particles. The noise is controlled by the width σ of the Gaussian and is displayed in percent of the characteristic length d_0.

The training data has been created for a noise of 1%–20% with roughly 38 000 particles for the different crystal structures (a volume of 34×34×34 with $d_0 = 1$ was used). Particles at the border of the volume are not considered. The trend for every structure with filtering is shown in Fig. 1.

We normalize the number of detected particles to 100% in the unperturbed case. It is apparent that the structures gradually decrease due to excluding "disordered" particles using the scalar product of the MSM. Also, it is visible that hcp particles are more affected by noise than fcc or bcc particles. Although this is an unwanted behavior, it could not be mitigated by choosing different thresholds of $S(i)$. We will use the 10% mark to quantify the performance of the method because it is the highest noise level which can realistically occur on crystalline systems (such as plasma crystals).

At a noise of 10% there is still ∼64% of hcp present. The numbers for fcc (∼84%) and bcc (∼88%) are even higher.

The MLP has then been trained with 3%–15% noise to achieve the best possible classification. The classification has to be verified which is done by using a new set of artificial data. This data is called "test data" and is identical to the training data but with a different random seed used for the Gaussian noise. The prediction of the MLP on the test data can be seen in Fig. 2.

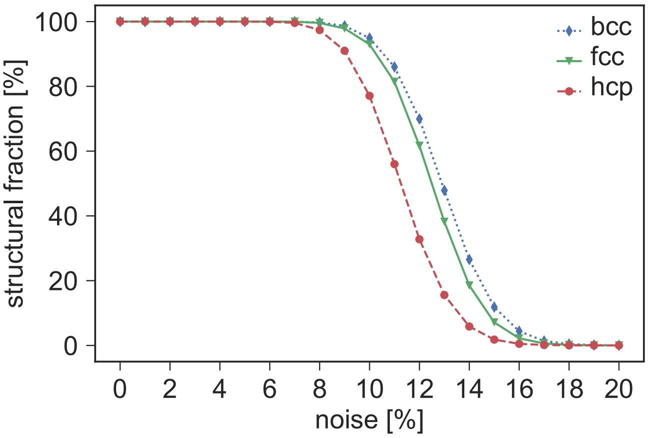

FIG. 1. Trend of the different crystal structures, which is used to train a MLP network.

The prediction shows no falsely classified particles for up to 6% of noise. The number of false classifications increases to 11% at 10% noise. We are able to detect ∼79% of particles at a noise of 10% while having considerably lower false positives compared to similar methods. Comparing with BCCMSM [20], we are able to detect 98% more particles with 39% less false classification at 10% noise. In [20] it was also demonstrated that a-CNA [15] and BOOP perform worse than BCCMSM regarding the yield of particles. The artificial dataset used in this work differs slightly from the data in [20], because we only consider particles in the bulk crystal.

The falsely classified particles can be resolved for the different structures, which is shown in Table II.

It is clearly visible that distinguishing fcc from hcp is the most difficult part of the classification. Overall, identifying the hcp structure is the most challenging. This means that hcp will be under-represented in highly disturbed lattices.

The method shows excellent performance at levels of high noise. Fortunately, the number of falsely classified particles can be easily adapted to specific requirements by adjusting the

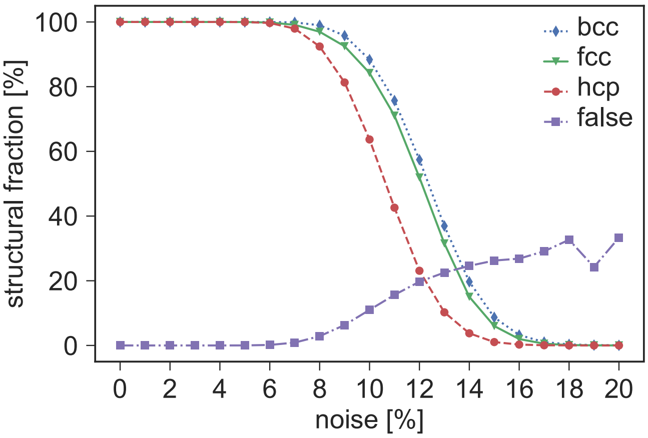

FIG. 2. Prediction of the MLP network to classify hcp, fcc, and bcc including falsely classified particles.

TABLE II. Percentage of false classification for every crystal structure at a noise of 10%.

Structure	False fcc	False hcp	False bcc
fcc		2.4%	1.0%
hcp	3.3%		2.0%
bcc	0.9%	1.3%	

threshold of $S(i)$. Thus, a trade-off between yield and false classification can be made.

However, using only artificial data is not enough to verify the advantages of such a method. Because of this, we simulate a Yukawa system using LAMMPS [26], which crystallizes as a true multiphase system consisting of fcc, hcp, and bcc domains. The system consists of charged microparticles (spherical, 3 μm diameter), which are levitated in a harmonic potential while being compressed by gravity. This is a simple model for a plasma crystal under laboratory conditions, which has the properties needed to verify the performance of the method [20]. The prediction from MCS for this system compared with the BCCMSM method is shown in Fig. 3.

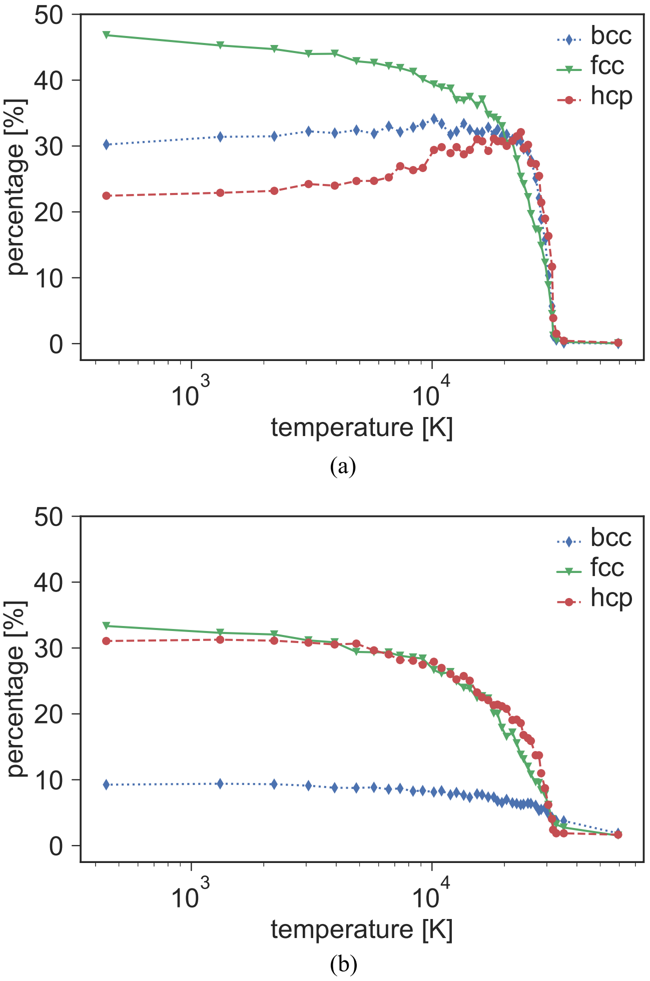

FIG. 3. Structural composition of a slowly cooled down Yukawa system for MCS using machine learning (a) and BCCMSM using manual classification (b).

It is visible that the method performs well for simulated data. The phase transition at 3×10^4 K is more pronounced and considerably steeper using the MCS method. Interestingly, the numbers of hcp and fcc are significantly higher than the results from BCCMSM indicates. With MCS we can accurately classify 99% of the particles as solid at low temperatures and are able to get the complete picture of our data, while BCCMSM is only able to detect 73% of the particles. Also, the BCCMSM does not drop down to 0% for high temperatures, which indicates falsely classified particles. This does not occur applying MCS.

Because we are now able to classify in structures at higher temperatures, it is observed that the numbers of bcc and hcp are regressing after the phase transition at 10^4 K, which suggests that rearrangements from hcp and bcc to fcc on lower temperatures took place. We cannot observe this behavior using BCCMSM and are now able to investigate the phase transition with complete structural information.

IV. CONCLUSION

We have presented an approach for crystal analysis which works very well for highly disturbed lattices while using already proven methods which can be easily implemented. Also, it could be verified that this method is able to successfully classify structures in true multiphase systems consisting of small fcc, hcp, and bcc clusters. Because the method only uses a Delaunay neighborhood around a particle, it is strictly local. This will be particularly useful for applications such as three-dimensional plasma crystals which often is a multiphase system with high noise and small crystalline clusters. Also, we could see that a much higher number of hcp and rearrangements of hcp and bcc particles is detected for crystalline mixed phase systems. Having the complete structural information will make it possible to efficiently compare our simulations of Yukawa systems with the results from plasma crystals and use them to make assumptions about particle charge and screening length.

With appropriate training data the method could be taught to detect phase transitions in amorphous materials, defects in crystals, or anisotropic structures in complex plasmas without having to change the algorithm. Only the particle locations and a set of labeled data has to be provided, which makes the method suitable for a broad variety of applications.

Additionally, the accuracy of MCS can be matched to specific requirements by adjusting a single threshold, which makes the method flexible and user-friendly. Because we rely on standard methods (Voronoi and MLP), the MCS can be implemented using well-established libraries.

Overall, the method significantly improves the state of three-dimensional crystal analysis and leads to a more complete picture where only static data is available.

ACKNOWLEDGMENT

This work was supported by DLR (Grant No. 50 WM 1442).

[1] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).

[2] G. J. Ackland and A. P. Jones, Phys. Rev. B **73**, 054104 (2006).

[3] J. D. Honeycutt and H. C. Andersen, J. Phys. Chem. **91**, 4950 (1987).

[4] C. L. Kelchner, S. J. Plimpton, and J. C. Hamilton, Phys. Rev. B **58**, 11085 (1998).

[5] H. Tsuzuki, P. S. Branicio, and J. P. Rino, Comput. Phys. Commun. **177**, 518 (2007).

[6] A. Malins, S. R. Williams, J. Eggers, and C. P. Royall, J. Chem. Phys **139**, 234506 (2013).

[7] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[9] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras, and A. J. Liu, Phys. Rev. Lett. **114**, 108001 (2015).

[10] S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu, Nat. Phys. **12**, 469 (2016).

[11] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, and A. Z. Panagiotopoulos, Soft Matter **13**, 4733 (2017).

[12] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).

[13] E. P. L. van Nieuwenburg,Y.-H. Liu, and S. D. Huber, Nat. Phys. **13**, 435 (2017).

[14] B. Steinmüller, C. Dietz, M. Kretschmer, and M. Thoma, Phys. Plasmas **24**, 033705 (2017).

[15] A. Stukowski, Modell. Simul. Mater. Sci. Eng. **20**, 045021 (2012).

[16] U. Gasser, E. R. Weeks, A. Schofield, P. Pusey, and D. Weitz, Science **292**, 258 (2001).

[17] W. Mickel, S. C. Kapfer, G. E. Schröder-Turk, and K. Mecke, J. Chem. Phys **138**, 044501 (2013).

[18] S. Winczewski, J. Dziedzic, and J. Rybicki, Comput. Phys. Commun. **198**, 128 (2016).

[19] E. P. Wigner, in *The Collected Works of Eugene Paul Wigner* (Springer, New York, 1993), pp. 608–654.

[20] C. Dietz and M. H. Thoma, Phys. Rev. E **94**, 033207 (2016).

[21] S. C. Kapfer, W. Mickel, K. Mecke, and G. E. Schröder-Turk, Phys. Rev. E **85**, 030301 (2012).

[22] S. C. Kapfer, W. Mickel, F. M. Schaller, M. Spanner, C. Goll, T. Nogawa, N. Ito, K. Mecke, and G. E. Schröder-Turk, J. Stat. Mech. (2010) P11010.

[23] M. M. Mehrabadi and S. C. Cowin, Q. J. Mech. Appl. Math. **43**, 15 (1990).

[24] P. R. ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, J. Chem. Phys. **104**, 9932 (1996).

[25] W. Lechner and C. Dellago, J. Chem. Phys. **129**, 114707 (2008).

[26] S. Plimpton, J. Comput. Phys. **117**, 1 (1995).