Two worlds collide: Image analysis methods for quantifying structural variation in cluster molecular dynamics

K. G. Steenbergen and N. Gaston

Citation: The Journal of Chemical Physics **140**, 064102 (2014); doi: 10.1063/1.4864753
View online: http://dx.doi.org/10.1063/1.4864753
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/140/6?ver=pdfcov
Published by the AIP Publishing

Two worlds collide: Image analysis methods for quantifying structural variation in cluster molecular dynamics

K. G. Steenbergen[1,a] and N. Gaston[2]
[1]*Physikalische und Theoretische Chemie, Freie Universität Berlin, Takustraße 3, 14195 Berlin, Germany*
[2]*MacDiarmid Institute for Advanced Materials and Nanotechnology, Victoria University of Wellington, P.O. Box 600, 6140 Wellington, New Zealand*

(Received 15 November 2013; accepted 27 January 2014; published online 11 February 2014)

Inspired by methods of remote sensing image analysis, we analyze structural variation in cluster molecular dynamics (MD) simulations through a unique application of the principal component analysis (PCA) and Pearson Correlation Coefficient (PCC). The PCA analysis characterizes the geometric shape of the cluster structure at each time step, yielding a detailed and quantitative measure of structural stability and variation at finite temperature. Our PCC analysis captures bond structure variation in MD, which can be used to both supplement the PCA analysis as well as compare bond patterns between different cluster sizes. Relying only on atomic position data, *without* requirement for *a priori* structural input, PCA and PCC can be used to analyze both classical and *ab initio* MD simulations for any cluster composition or electronic configuration. Taken together, these statistical tools represent powerful new techniques for quantitative structural characterization and isomer identification in cluster MD. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4864753]

I. INTRODUCTION

One of the greatest challenges of small cluster molecular dynamics (MD) is extracting meaningful information from hundreds of thousands of time steps. The significant volume of data requires methods that allow the researcher to reduce or summarize, often by using averages over many time steps. For quantities such as the potential, kinetic, or total energy, averages or expectation values provide useful summaries in order to extract quantities such as specific heat or temperature. As these thermodynamic measurables are inherently defined by average behavior, little or no information is lost in the summary.

In contrast, geometric structural measures present a challenge. At the most basic level, MD geometric data consist of a large number of bond lengths and angles configured in a particular manner. For low-temperature MD, these changes may consist only of small bond vibrations, for which averaging may still be an adequate measure. However, at higher temperature, structural averages lose more and more information – particularly for clusters with complex potential energy landscapes, where isomerization is frequent.

The tools for analyzing geometric structure at finite temperature have remained rather limited. Structural measures commonly employed for larger clusters, such as common neighbor analysis[1–3] or the Steinhardt order parameter[4,5] are less reliable for so few atoms where a bulk-like environment is not well-defined. An algorithm based on structural alignment has been used for the identification and tracking of sodium isomers, which aides in exploring the potential energy surface of metals.[6,7] Using a set of random rotations to compare cluster structures, this algorithm requires that a set of possible isomers are known *a priori*, against which structural data can be compared. As molecular similarity has been a long-studied problem to aid in drug design, a number of other structural matching algorithms have also been proposed,[8–11] each of which require the same *a priori* knowledge of either the full or partial structural landscape.

In recent applications of the graph theoretic approach, two powerful software packages for structural analysis have been introduced: *moleculaRnetworks*[12,13] and *ChemNetworks*.[14] With a wide variety of examples applied to solute-solvent analysis, each package supplies a wide-range of topological analysis tools which could be applied to cluster structural analysis. The *moleculaRnetworks* software package is based on PageRank (used by Google to evaluate website importance), and can be used to evaluate the polyhedral arrangement of structures; however, this requires an input of known polyhedra for comparison. The program library of polyhedra have between 4–10 vertices, although the list can be user-appended. The *ChemNetworks* package can be used to generate a list of all geodesic and euclidean path lengths between vertices, which have been defined by the user, providing insight into the geometric shape. The output of the overall shape of an individual cluster would be limited by the library of polyhedra to "recognizable" or previously-determined shapes.

How, then, does one go about identifying isomerization in a simple, straight-forward manner without *a priori* or "library" structural data? In the process of analyzing molecular dynamics simulations for small gallium clusters (9–36 atoms),[15–17] we required methods allowing for the direct comparison of structural similarity over a range of isomers, across multiple cluster sizes and temperatures, without any *a priori* knowledge of cluster structure. We noted similarities to the challenges encountered in the field of remote sensing image analysis, where the analysis methods typically involve a large amount of data measured under dissimilar conditions.

[a]kgsteen@gmail.com

Borrowing techniques from the field of remote sensing image analysis, here we present two novel methods of geometric structural analysis for cluster MD which require no *a priori* knowledge of cluster structure: the principal components analysis (PCA)[18,19] and the Pearson Correlation Coefficient (PCC).[20] These straightforward analysis techniques represent a unique approach to extracting meaningful and quantitative geometric information from cluster MD simulations. Each method complements or extends previous analysis methods, with a simple but powerful approach that allows for isomer comparisons both within and between cluster sizes. Here, we present an overview of the PCA and PCC analysis techniques applied to small cluster MD, using representative examples from our simulations on small gallium clusters sized 9, 12, and 20 atoms.[15,16]

II. PRINCIPAL COMPONENTS ANALYSIS

The PCA[18,19] is a standard statistical tool that identifies the axes (dimensions) of greatest variance in a data set. It is most commonly used to reduce high-dimensional data to a subset of orthogonal dimensions that capture the greatest variance: the principal components. Although far from the typical application of PCA, when applied to three-dimensional *xyz* coordinates, it identifies the three longest, orthogonal dimensions of the coordinates.

For MD structural data, the *xyz* coordinates are the atomic positions at a particular time step. PCA first identifies the axis of maximum variation of atomic positions, or first principal component ($\mathbf{P}_1$). The 2nd principal component ($\mathbf{P}_2$) is the axis that captures the next greatest variation of the atomic positions, while being entirely uncorrelated (orthogonal) with the first axis. The third principal component ($\mathbf{P}_3$) will then capture the largest remaining variation of the atomic positions uncorrelated with the first two axes.

The direction of each principal axis ($\mathbf{P}_j$) is found in the eigenvectors of the PCA, which allow for the direct measurement of the orthogonal axis lengths. To best visualize the method, it may be helpful to imagine "shrink wrapping" the atoms to create a 3D-volume which entirely, but minimally, contains the data. The distance between where each axis pierces the shrink wrap is the length of the principal component axis. Denoted here as ℓ_1, ℓ_2, and ℓ_3, the three orthogonal axis-lengths yield an excellent picture of shape changes in the course of a finite temperature MD simulation. We can also glean information from the three principal component eigenvalues, p_1, p_2, and p_3. The eigenvalues measure the variance (σ^2) in the position data along each principal component axis, which often provides a clearer representation of structural changes.

In Fig. 1, we contrast the PCA results between an elongated and nearly-spherical Ga_{20}^+ structure in order to demonstrate the effectiveness of the PCA in capturing cluster shape. One strength of the PCA is that it can be efficiently applied at every time step of a MD simulation, yielding a simple method for capturing overall geometric changes as a function of time. As previously mentioned, this analysis requires no *a priori* knowledge of structure, shape, or composition, relying only on the *xyz* atomic coordinates at each time step.

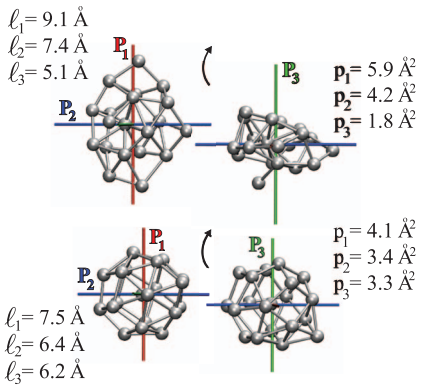

FIG. 1. Representative examples of PCA for two distinctly-shaped Ga_{20}^+ structures:[15] (top) two views of an elongated structure and (bottom) two views of a nearly spherical cluster (direction of rotation between views indicated by the arrow). Each principal component axis ($\mathbf{P}_j$) is illustrated. The corresponding principal component eigenvalues (p_j) and axis lengths (ℓ_j) are annotated beside each structure.

A. Overview of PCA for MD

In the PCA for MD, we represent the coordinates of a cluster at time step t as a matrix, $\mathbf{X}$, with dimensions NxJ. There are 3 columns, J, numbered as $j = \{1,2,3\}$, which respectively represent the x, y, and z coordinates of each ion. The number of rows, N, is equal to the number of atoms in each cluster. The data are preprocessed by performing a simple spatial translation so that the mean of each column is zero, i.e., the center of the cluster coordinates is (0,0,0). This yields the centered matrix $\mathbf{X}_0$.

The PCA generally follows the well-known singular value decomposition (SVD) algorithm. For atomic position data, we compute the SVD of $\mathbf{X}_0$ as

$$\mathbf{X}_0 = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}, \tag{1}$$

where $()^{\mathbf{T}}$ is the transpose of the matrix within the parentheses. $\mathbf{U}$ is an NxN matrix representing the normalized eigenvectors of $\mathbf{X}_0\mathbf{X}_0^{\mathbf{T}}$. $\mathbf{V}$ is a JxJ matrix representing the normalized eigenvectors of $\mathbf{X}_0^{\mathbf{T}}\mathbf{X}_0$. The three eigenvalues of $\mathbf{X}_0\mathbf{X}_0^{\mathbf{T}}$ are equal to those of $\mathbf{X}_0^{\mathbf{T}}\mathbf{X}_0$, represented here as λ_j. The middle matrix, Σ, is an NxJ diagonal matrix of the square roots of the eigenvalues,

$$\Sigma_{jj} = \sqrt{\lambda_j}. \tag{2}$$

The standard deviation of each principal component, which for our application relate the standard deviation of the ionic positions along each principal component axis, can then be represented as

$$\sigma_j = \sqrt{\frac{\lambda_j}{(N-1)}} = \frac{\Sigma_{jj}}{\sqrt{(N-1)}}. \tag{3}$$

The principal component eigenvalues, p_j, are simply the variances

$$p_j = \sigma_j^2 = \frac{\lambda_j}{(N-1)} = \frac{(\Sigma_{jj})^2}{(N-1)}. \tag{4}$$

The principal component eigenvectors, $\mathbf{P_j}$, are represented by the rows of $\mathbf{V}$. We can then project the cluster coordinates onto principal component axes by taking the matrix product

$$\mathbf{B} = \mathbf{X_0 V}. \tag{5}$$

Each of the three columns in $\mathbf{B}$ now represents the x, y, and z coordinates of the cluster atoms projected so that the axis of maximum variance aligns with the x-axis, the second maximum variance aligns with the y-axis, and the third maximum variance falls along the z-axis. The lengths of each principal component axis, ℓ_j, can be easily determined by subtracting the minimum from maximum coordinate in each column of $\mathbf{B}$.[21]

We note that previous research has utilized the eigenvalues of the $\mathbf{X_0^T X_0}$ in order to characterize the deformation (from spherical) of global minimum sodium clusters.[22,24] Solov'yov *et al.* further determine the atomic distribution through the standard deviation, which differs from Eq. (3) only in the denominator, N, relating their use of a statistical population.[22] Referred to as the principal value tensor[22] or quadrupole tensor,[23,24] it is important to discern $\mathbf{X_0^T X_0}$ from the matrix $\mathbf{V}$, which consists of the normalized eigenvectors of $\mathbf{X_0^T X_0}$. This additional step of the full PCA analysis allows for the projection onto principal component axes (Eq. (5)) and the determination of the cluster axis lengths (ℓ_j).

B. PCA results for MD

Fig. 2 demonstrates the PCA results for the 450 K (average temperature) MD simulation of Ga_{20}^+ (computational details cited in our previous work).[15] The top panels illustrate the axis lengths and PCA eigenvalues as a function of MD simulation time. From the axis lengths (top panel), we can easily identify transitions between two distinct structural isomers. However, the PCA eigenvalues highlight a third isomer between 25 and 30 ps. The time periods for each isomer are annotated by the **A**, **B**, and **C** in the eigenvalue plot, which correspond to each of the structures given at the bottom row of Fig. 2.

We have tested the PCA on clusters as small as 9 atoms, where it still effectively identifies structural transitions. Fig. 3 demonstrates the PCA applied to the Ga_9 simulation at both 445 K and 535 K.[16] PCA identifies two structural isomers, each with the same cubic base and differing only in the relative positioning of the adatom. Such a small variation would have been difficult to systematically identify visually or by other methods of analysis. However, PCA clearly distinguishes the two, allowing for an accurate determination of the relative stability of each isomer at each finite temperature. The adatom-up structure is less stable at 445 K, persisting for only 40% of the simulation time; however, at 535 K, the adatom-up structure persists for 74% of the simulation time, surprisingly becoming the more stable configuration at higher finite temperatures.

While PCA analysis is a simple, powerful tool for MD structural analysis, there is at least one limitation: we cannot directly compare results between cluster sizes. Each addi-

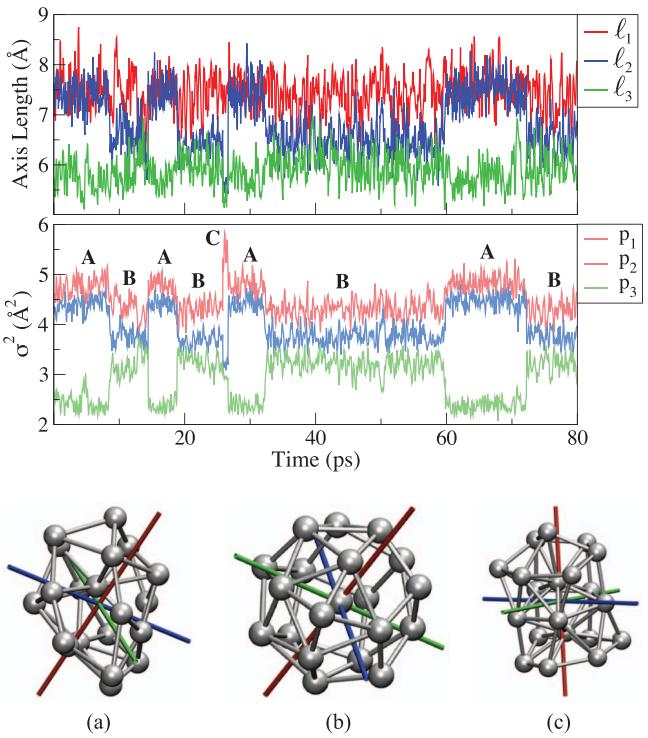

FIG. 2. (Top) Three principal axis lengths, ℓ_j and (middle) the three principal eigenvalues, p_j are given for each MD time step of the 450 K simulation of Ga_{20}^+.[15] From the eigenvalues, it is easy to discern the three distinct structural motifs **A**, **B**, and **C**, which correspond (respectively) to the representative structures shown in (a) obtained from ~66 ps, (b) from ~12 ps, and (c) from ~26 ps.

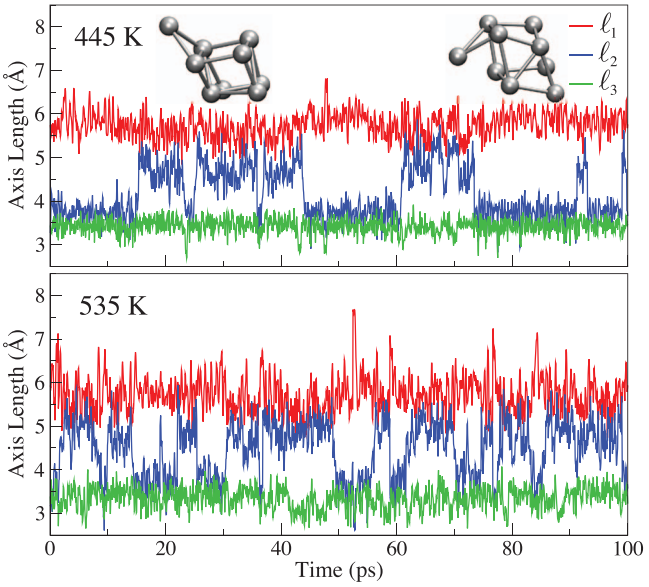

FIG. 3. Three principal axis lengths, ℓ_j are given for each MD time step of the (top) 445 K and (bottom) 535 K simulation of Ga_9.[16] From the distinct changes in the ℓ_2 over the course of the simulation, we can easily identify two structural isomers differing only by the position of the adatom: adatom-up (top, left) and adatom-down (top, right). Noting the axis-length patterns, we can additionally quantify and compare the structural stability of two isomers between finite temperatures: the up/down simulation time ratios are 40%/60% at 445 K and 74%/36% at 535 K.

tional atom naturally changes the lengths of the axes. In order to compare structures between cluster sizes, we utilize the PCC.[20]

III. PEARSON CORRELATION COEFFICIENT

PCC[20] is a simple correlation measure, calculated as the covariance of two data sets divided by the product of their standard deviations. For two data sets x and y, both of the same size n, this easily simplifies to

$$PCC(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (6)$$

where $\bar{x}$ and $\bar{y}$ represent the average of each data set and x_i and y_i are the individual data points. PCC(x,y) measures the degree of linear correlation between x and y, yielding a value of (+1) when the two data sets are perfectly correlated, (0) for perfectly **un**correlated data, and (−1) for perfect inverse-correlation.

A. Overview of PCC for MD

In order to apply the PCC to MD for the purpose of capturing non-shape related variation, we utilize a pair distribution function (PDF) to summarize the bond structure at each time step. We then average the PDFs over a short time to obtain a time-average pair distribution function (taPDF). Although some detail is lost in averaging, we found that the single-time step PDF's were too noisy for meaningful PCC analysis. Each taPDF then represents a short-time average cluster structure, which can be thought of as a structural signature. These structural signatures are compared using the PCC analysis, which provides a quantitative measure of structural similarity based on bond patterns.

After extensive testing, we determined that an average over 40 fs (20 time steps with dt = 2 fs) adequately reduces the finite temperature noise while still capturing structure-specific bond patterns. We utilize a PDF bin size of 0.05 Å, where n is the total number of bins. The quantities $\bar{x}$ and $\bar{y}$ are then the average of each of the correlated taPDFs. The individual data points, x_i and y_i, are the bond frequencies of each bin.

B. PCC results for MD

For ease in illustrating the concept of the PCC, we utilize the structural motifs **A** and **B** represented in Fig. 2. We average the PDFs for clusters within 40 fs of structures (A) and (B), yielding taPDF's g_a and g_b (shortened from $g_a(r)$ and $g_b(r)$). We also calculate g_{a_2}, another **A**-motif taPDF obtained from ∼17 ps (Fig. 2). The top panel of Fig. 4 illustrates g_a compared with g_{a_2}, resulting in the relatively high PCC(g_a,g_{a_2}) = 0.94. The distinct structural signatures from g_a and g_b are contrasted in the second panel, with the appropriately lower PCC(g_a,g_b) = 0.84. These two PCC values are obtained by measuring a regression distance from linear (perfect) correlation, as illustrated in the bottom panel of Fig. 4. By this example, it is easy to identify the origin of the higher PCC, as the g_a–g_{a_2} data (red circle) is more closely-clustered around its linear regression line (red dashed).

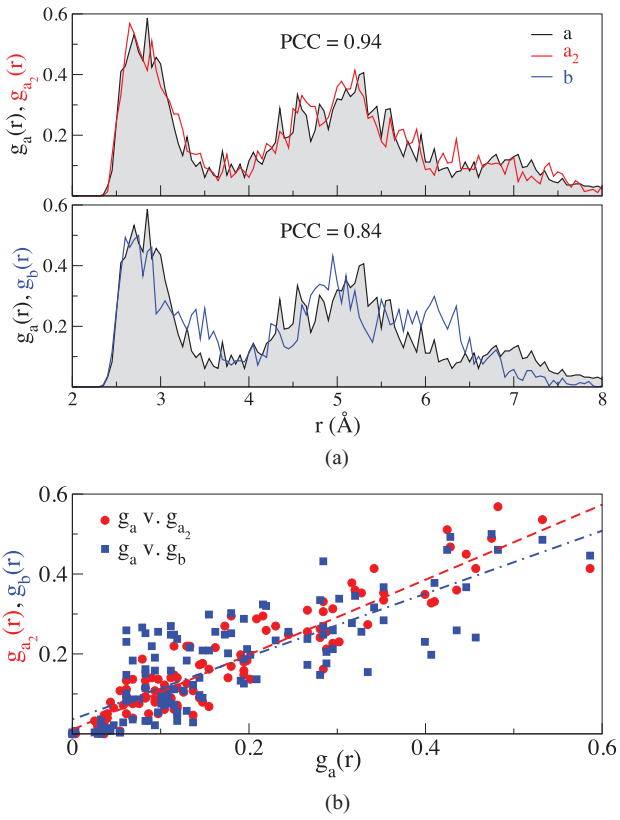

FIG. 4. (a) The reference taPDF, g_a, is correlated with two other taPDF's: (top panel) g_{a_2} and (bottom panel) g_b. Both g_a and g_{a_2} are averaged over structures of motif **A**, and their similar taPDF's yield a high correlation (PCC) value. The taPDF for g_b was obtained from an average over structures of motif **B**, explaining the distinct taPDF's and low correlation value. (b) A plot illustrating the PCC measure, with g_a plotted against both g_{a_2} (red, circle) and g_b (blue, square). PCC perfect correlations of (+1) are demonstrated by the respective linear regression lines (red dashed and blue dashed-dotted). The PCC is calculated by a regression distance, graphically showing that g_a is less linearly correlated with g_b, as these data points (blue squares) are less closely clustered around their regression line (blue dashed-dotted).

Extending this analysis to MD, we refer to the illustrative example in the top panel of Fig. 5. We first calculate a taPDF for every 40 fs window of the simulation, annotated as g_{t_n}. Selecting a reference taPDF of g_a, we calculate the PCC between g_a and each time-window's taPDF, PCC(g_a,g_{t_n}). When calculated for an entire MD simulation, the correlation trends yield a detailed picture of structural variation as measured by bonding patterns. Exemplified in the second panel of Fig. 5, we select both g_a and g_b as reference taPDFs for the Ga_{20}^{+} simulation at 450 K. With the highest PCC values illustrating the best bond-pattern correlation, we observe that the PCC patterns exactly match those demonstrated by the PCA for this same simulation (Fig. 2). It is noted that the PCC does not distinguish structural motif **C**, which may arise from either the short duration of this structural motif (∼5 ps) or due to a bond structure closely resembling that of motif **A**.

An additional strength of the PCC is that it can be applied across a range of cluster sizes. We select a reference

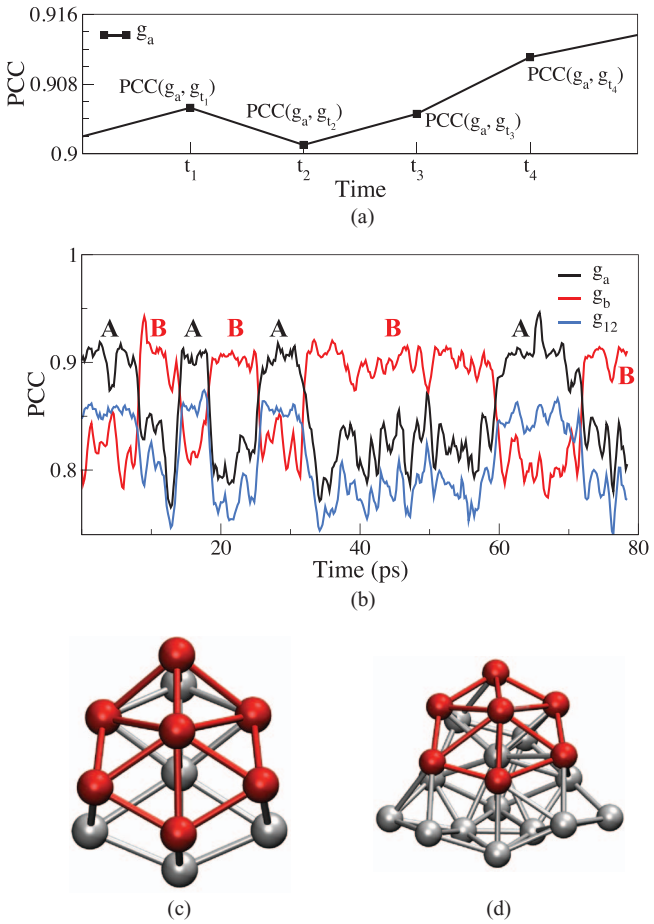

FIG. 5. (a) An illustrative example of PCC analysis applied to a MD simulation. The reference taPDF, g_a, is correlated with each time window taPDF, g_{t_n}, yielding the series of correlation values PCC(g_a,g_{t_n}). (b) PCC analysis for the Ga_{20}^+ simulation at 450 K, using reference taPDF's g_a (black) and g_b (red). The pattern closely matches the PCA for the same simulation (Fig. 2), as demonstrated here by the motif labels **A** and **B**. We also include a PCC comparison between this Ga_{20}^+ simulation data and a reference taPDF from the Ga_{12} simulation at 450 K (g_{12}, blue). We note g_{12} has maximum PCC when compared with taPDF's of motif **A** and closely follows the PCC trend of g_a, illustrating how PCC can be used to correlate bond patterns across cluster sizes.

taPDF from the Ga_{12} simulation[16] at 450 K (g_{12}) and compare it to the same-temperature Ga_{20}^+ taPDFs. Illustrated by the blue curve in Fig. 5, although the correlation values are notably lower due to the significant reduction in cluster size, the PCC pattern closely follows g_a: exhibiting maximums when correlated with structures of motif **A** and minimums for motif **B**. From this PCC analysis, we can discern that the bond structure of this Ga_{12} cluster best matches the 20-atom **A**-motif structures. Figs. 5(c) and 5(d) compare a representative structure from $g_{12}(r)$ with the Ga_{20}^+ **A**-motif structure (different perspective from Fig. 2(a)). The double-row, hexagonal ring structure is common to both clusters, although it is only highlighted for the uppermost ring for visual clarity.

While the PCC analysis can also be applied across a range of finite temperatures, it is particularly sensitive to any x-stretching, which arises in taPDF data as the temperature is raised (bonds lengthen). The PCC values will, therefore, be notably lower when correlating between different temperatures, which greatly affects the clarity of the results. For our simulations, we noted that the correlations were still instructive for average temperature differences up to 100 K (not crossing the phase transition); however, this would likely be unique for each different system.

IV. SUMMARY

We have demonstrated the PCA and PCC analyses for representative examples of the 9, 12, and 20-atom gallium simulations. With no *a priori* structural input, we have illustrated that PCA can be used to easily discern changes in the overall shape of a cluster, while the PCC can be used to identify both changes to the bond structure as well as common structural motifs between different cluster sizes. Since both methods rely only on the atomic positions, they can be applied to both classical and *ab initio* MD data across a wide-range of sizes and compositions. While each analysis method has weaknesses, coupling the PCA and PCC analyses creates a powerful tool for quantifying structural variation at finite temperature. Future work will include an investigation of x-stretching models that may allow PCC comparison across a wider temperature range, as well as additional correlation methods that may enhance or extend the analysis.

ACKNOWLEDGMENTS

This work has been supported by the Marsden Fund of the Royal Society of New Zealand under Contract No. IRL0801. We thank the New Zealand eScience Infrastructure (NeSI), particularly the BlueFern (University of Canterbury, nesi67) and Pan (University of Auckland, nesi7) supercomputer teams for computational time and support. K.G.S. extends a special thanks Dr. Ronald B. Lockwood, Dr. James Gardner, and Dr. Peter Armstrong, for their excellent instruction and years of encouragement and support.

[1]A. Clarke and H. Jónsson, Phys. Rev. E **47**, 3975 (1993).
[2]S. Hendy and B. Hall, Phys. Rev. B **64**, 085425 (2001).
[3]S. Hendy and J. Doye, Phys. Rev. B **66**, 235402 (2002).
[4]P. Steinhardt, D. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[5]J. Neirotti, F. Calvo, D. Freeman, and J. Doll, J. Chem. Phys. **112**, 10340 (2000).
[6]J. Vásquez-Pérez, G. Martínez, A. Köster, and P. Calaminici, J. Chem. Phys. **131**, 124126 (2009).
[7]S. Kearsley, Acta Crystallogr., Sect. A: Found. Crystallogr. **45**, 208 (1989).
[8]M. Barakat and P. Dean, J. Comput.-Aided Mol. Des. **5**, 107 (1991).
[9]J. Mestres, D. Rohrer, and G. Maggiora, J. Comput. Chem. **18**, 934 (1997).
[10]J. Nissink, M. Verdonk, J. Kroon, T. Mietzner, and G. Klebe, J. Comput. Chem. **18**, 638 (1997).
[11]X. Gironés, D. Robert, and R. Carbó-Dorca, J. Comput. Chem. **22**, 255 (2001).
[12]B. Mooney, L. Corrales, and A. Clark, J. Comput. Chem. **33**, 853 (2012).
[13]B. Mooney, L. Corrales, and A. Clark, J. Phys. Chem. B **116**, 4263 (2012).
[14]A. Ozkanlar and A. Clark, J. Comput. Chem. **35**, 495 (2014).
[15]K. Steenbergen, D. Schebarchov, and N. Gaston, J. Chem. Phys. **137**, 144307 (2012).
[16]K. Steenbergen and N. Gaston, Phys. Chem. Chem. Phys. **15**, 15325 (2013).
[17]K. Steenbergen and N. Gaston, Phys. Rev. B **88**, 161402(R) (2013).
[18]K. Pearson, Philos. Mag. **2**, 559 (1901).
[19]H. Abdi and L. Williams, Wiley Interdiscip. Rev. Comput. Stat. **2**, 433 (2010).

[20]J. L. Rodgers and W. Nicewander, Am. Stat. **42**, 59 (1988).
[21]This PCA mathematical development for MD data assumes $N \geq 3$: that each cluster has a minimum of 3 atoms.
[22]I. Solov'yov, A. Solov'yov, and W. Greiner, Phys. Rev. A **65**, 053203 (2002).
[23]S. Zorriasatein, M. Lee, and D. Kanhere, Phys. Rev. B **76**, 165414 (2007).
[24]K. J. Jose, S. Khire, and S. Gadre, "A density functional investigation on the structures, energetics, and properties of sodium clusters through electrostatic guidelines and molecular tailoring," in *Aromaticity and Metal Clusters*, edited by P. K. Chattaraj (CRC Press, 2010), pp. 205–226.