THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

Cite This: J. Phys. Chem. C 2019, 123, 15859−15866

pubs.acs.org/JPCC

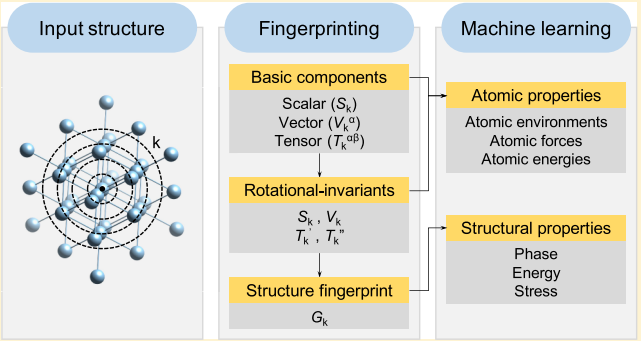
General Atomic Neighborhood Fingerprint for Machine Learning-Based Methods

Rohit Batra, Huan Doan Tran, Chiho Kim, James Chapman, Lihua Chen, Anand Chandrasekaran, and Rampi Ramprasad*

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, United States

Supporting Information

ABSTRACT: To facilitate chemical space exploration for material screening or to accelerate computationally expensive first-principles simulations, inexpensive surrogate models that capture electronic, atomistic, or macroscopic materials properties have become an increasingly popular tool over the last decade. The most fundamental quantity common across all such machine learning (ML)-based methods is the *fingerprint* used to numerically represent a material or its structure. To increase the learning capability of the ML methods, the common practice is to construct fingerprints that satisfy the same symmetry relations as displayed by the target material property of interest (for which the ML model is being developed). Thus, in this work, we present a general, simple, and elegant fingerprint that can be used to learn different electronic/atomistic/structural properties, irrespective of their scalar, vector, or tensorial nature. This fingerprint is based on the concept of multipole terms and can be systematically increased in sophistication to achieve a desired level of accuracy. Using the examples of Al, C, and hafnia (HfO_2), we demonstrate the applicability of this fingerprint to easily classify different atomistic environments, such as phases, surfaces, point defects, and so forth. Furthermore, we demonstrate the generality and effectiveness of this fingerprint by building an accurate, yet inexpensive, ML-based potential energy model for the case of Al using a reference data set that is obtained from density functional theory computations. Finally, we note that the fingerprint definition presented here has applications in fields beyond materials informatics, such as structure prediction, identification of defects, and detection of new crystal phases.



INTRODUCTION

Machine learning (ML)-based surrogate models have been employed in the field of materials science to efficiently predict a diverse set of materials properties,[1−5] suggest potential synthesis routes,[6−9] classify crystal structures,[10,11] map potential energy surfaces,[12,13] or to accelerate first-principles computations.[14−19] At the heart of many such ML models lies the fundamental principle of the structure−property relationship, that is, a material behavior is an outcome of its underlying structure or atomic arrangement. Thus, *fingerprinting*, or effectively capturing information about the atomic neighborhood, is one of the key ingredients of these ML-based methods.[20−24] Depending on the target property of interest, this information may be required at a global scale (or structure level) or at a local point. For example, identifying crystal phases of materials and/or their associated defects belongs to the former class and is dealt with using a *structure* fingerprint. On the other hand, obtaining knowledge of microscopic quantities, such as electronic charge distribution or atomic forces falls under the latter class and is approached using an *atomic or local* fingerprint. Irrespective of the scale, fingerprinting atomic neighborhood is a quintessential problem frequently encountered in chemical and materials science and finds repercussions far beyond materials informatics, with applications in structure prediction for drug discovery, identifying defects during molecular simulations, and detecting new phases using X-ray diffraction or high-throughput computation data.

Many prescriptions have, therefore, been proposed in the past to fingerprint or numerically represent the atomic neighborhood around an atom, that is, the atomic or local fingerprint, or to capture the overall structure of a material using a structure fingerprint. The fingerprint definitions must satisfy certain mathematical properties to allow construction of accurate, generalizable, and efficient models. For example, the fingerprints should uniquely represent the atomic neighborhood and vary smoothly with atomic displacements. If possible, they should be highly correlated to and follow the same symmetry relations as that of the target property. The radial distribution function (RDF) is perhaps the most basic

Received: April 26, 2019
Revised: June 4, 2019
Published: June 6, 2019

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

fingerprinting scheme satisfying these relations. Although very informative, the RDF presents a practical challenge of representing a continuous function. To counter this, Ziletti et al.[10] devised a 2-D structure fingerprint based on simulated diffraction patterns capable of easily identifying average crystal symmetries of a material. Their fingerprint definition is not only invariant to structural rotations or translations but also, more importantly, is insensitive to the presence of small defects or minor atomic perturbations, which are often present in structural databases (such as Inorganic Crystal Structure Database,[25] Materials Project,[26] etc.). Other examples of structure fingerprints include a Coulomb matrix,[27] many-body tensor representation,[28] deep tensor neural network (NN),[13] Voronoi tessellation,[29] combination of RDF with atomic charges,[30] cross-correlation of powder diffraction pattern,[31] and so forth.[32,33]

For the case of atomic fingerprints, several schemes motivated from the field of first-principles electronic structure methods or classical potentials have been proposed. These include symmetry functions,[34] bispectrum coefficients,[35] smooth overlap of atomic positions,[36] group-theoretical high-order rotational invariants,[37] and AGNI,[14,16] among others.[38] These local or atomic fingerprints can either be directly used to learn local/spatial quantities, such as charge density, exchange−correlation potential, and so forth, or be collected/averaged over a group of atoms to learn global quantities like total energy. Furthermore, a relatively new approach is to use convolution or graph NN to completely eliminate the need of manually constructed fingerprints but directly use ML to arrive at the most efficient fingerprint (or latent) space.[39]

In this work, we present a simple and mathematically elegant fingerprint that can be used to capture the local atomic neighborhood around an arbitrary point in space (or about an atom). The fingerprint is general and can be used to learn scalar (energies), vector (forces), or tensorial (stresses) properties, while efficiently preserving their symmetry requirements. Indeed, in recent work,[16] this fingerprint was used to learn and predict the electronic charge density and the electronic density of states for Al and polyethylene, two diverse materials. It is based on multipole expansion and thus can be systematically improved in terms of prediction accuracy. Further, using simple transformations it can be made rotationally invariant for learning properties with such constraints. Here, we demonstrate the use of this fingerprint to an important material classification problem, that is, the characterization of different atomic environments, including phases, surfaces, grain boundaries, vacancies, clusters, and so forth. Further, using this atomic fingerprint, we derive a simple definition of structural fingerprint, which is then used to quantify (dis)similarity between different structures of a material. Taking the example of C structures obtained from Materials Project database, we showcase how our fingerprint could be used to identify structurally similar or "duplicate" entries in large databases and also assist structure search methods for efficient screening. To further highlight the significance of the structure fingerprint, we build a comprehensive potential energy model—an example of a regression problem—for the case of Al under various conformational/morphological settings. The accuracy of the energy model clearly suggests that the fingerprint is generalizable to a large conformational space and yet is able to easily discern between configurations with subtle atomic changes. Thus, we believe that the fingerprint presented here can be useful for a diverse range of materials properties for which it is important to efficiently capture the atomic neighborhood information.

METHODS

Fingerprint: Basic Components. In this work, we present a grid-based representation of local atomic environment that can be used for various classifications and regression problems in the field of material science. The representation consists of a hierarchy of features capturing different aspects of atomic neighborhood and involves components resembling scalar, vector, and tensorial definitions as described below. The simplest term is the scalar component that captures the radial information of atoms (discerning their chemical identity) around a grid-point g using a predefined set of Gaussian functions (k) of varying widths σ_k, defined as

$$S_{k\Omega} = c_k \sum_{i=1}^{N_\Omega} \exp\left(\frac{-r_{gi}^{\,2}}{2\sigma_k^{\,2}}\right) f_c(r_{gi}) \tag{1}$$

where r_{gi} is the distance between the atom i of species Ω and the reference grid-point g. $f_c(r_{gi})$ is the cutoff function defined as $0.5\left[\cos\left(\frac{\pi r_{gi}}{R_c}\right) + 1\right]$, which smoothly decays to zero for atoms at distance larger that R_c from the reference grid-point. The coefficient c_k is the normalization constant given by $\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right)^3$. The overall dimensionality of the scalar component is the product of number of Gaussians k and the type of elemental species Ω. Similarly, the vector and tensorial components of the fingerprint are defined by

$$V_{k\Omega}^{\,\alpha} = c_k \sum_{i=1}^{N_\Omega} \frac{r_{gi}^{\,\alpha}}{r_{gi}} \exp\left(\frac{-r_{gi}^{\,2}}{2\sigma_k^{\,2}}\right) f_c(r_{gi}) \tag{2}$$

$$T_{k\Omega}^{\,\alpha\beta} = c_k \sum_{i=1}^{N_\Omega} \frac{r_{gi}^{\,\alpha} r_{gi}^{\,\beta}}{r_{gi}^{\,2}} \exp\left(\frac{-r_{gi}^{\,2}}{2\sigma_k^{\,2}}\right) f_c(r_{gi}) \tag{3}$$

where α and β represent each of the x, y, or z directions. The prefactors $r_{gi}^{\,\alpha}/r_{gi}$ and $\frac{r_{gi}^{\,\alpha} r_{gi}^{\,\beta}}{r_{gi}^{\,2}}$ distinguish the vector and tensorial components from their scalar counterpart. They help to capture the angular information of atomic neighborhood through projections of atomic distances and their cross-products, in all directions. Further, these prefactors allow for the preservation of essential rotation-dependent relations present in different vectorial or tensorial properties. Thus, for component-by-component learning of directionally dependent quantities, such as atomic forces or stresses, one can directly use the appropriate fingerprint definitions provided in eqs 2 and 3. In fact, in our previous works,[14,15,40] we have successfully demonstrated the use of just the vector components to develop force fields for elemental Al, Cu, C, and more. Similarly, the individual tensorial components of the fingerprint can be used to learn the elements of the stress tensor matrix.

Three points are worth mentioning here. First, although only a general definition to capture atomic neighborhood around a grid point is presented, it could be easily used to derive atomic fingerprints by locating the grid points at the atomic positions.

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

Table 1. Summary of References Data Set Used for Al, C, and Hafnia[a]

configuration class	Al	C	HfO_2
defect-free bulk	fcc, bcc, and hcp (w and w/o strain)	dc, graphite, bc8, lonsdaleite, and 57 bulk phases from Materials Project	M, T, O1, O2, OA, and C
point defects	supercell with 1, 2, 6 random vacancies		
planar defects	(100), (110), (111), (200), (333) surfaces; (111), (210), (310), (320), (510) grain boundaries; (111) stacking fault	graphene	(111) and (110) surfaces of M, T, O1, and OA phases
point and planar defects	adatom(s) on (100), (110), and (111) surfaces		
clusters	radius 3, 5, 8, 10, and 12 Å clusters	fullerenes with 60, 70, and 180 atoms; SWCNTs	

[a]The data set is categorized into different classes of configurations based on the types of defects.

Second, the derived atomic fingerprint definition, being based on distances, is invariant to translation or permutations of atoms of the same elements and, if needed, can be easily made rotationally invariant (will be discussed next). Third, although in eqs 1−3 we presented Gaussian functions centered at the reference grid-point, these definitions can be easily extended to include Gaussians with varying mean values.

We also note that the general grid-based fingerprint definition could be used to learn many spatially continuous field quantities, such as electronic charge density, and its restricted atom-centered version or atomic fingerprint, could be used to learn atomic properties, such as atomic forces, energies, and so forth. Additionally, in contrast to other approaches[36,37] that project the atomic neighbor density function onto a basis set involving the spherical harmonics for capturing the angular distribution and a set of radial functions for capturing the radial distribution of the neighbors, in our approach three separate Cartesian components of the atomic neighbor density function are first projected onto a set of radial functions, and then, rotational invariants are constructed. This results in relatively simpler fingerprint with lower dimensionality.

Fingerprint: Rotationally Invariant Components. As noted earlier, the prefactors in eqs 2 and 3 render the vector (V_k^{α}) and tensorial components of the fingerprint $(T_k^{\alpha\beta})$ directionally dependent, in contrast to the scalar component which is rotationally invariant. Nonetheless, for problems involving directionless quantities, these components can also be transformed to rotationally invariant representations. The vector component can be transformed using

$$V_{k\Omega} = \sqrt{(V_{k\Omega}{}^{x})^2 + (V_{k\Omega}{}^{y})^2 + (V_{k\Omega}{}^{z})^2} \tag{4}$$

while for the case of the tensorial component, two rotationally invariant representations can be derived

$$\begin{aligned} T'_{k\Omega} &= T_{k\Omega}{}^{xx}T_{k\Omega}{}^{yy} + T_{k\Omega}{}^{yy}T_{k\Omega}{}^{zz} + T_{k\Omega}{}^{xx}T_{k\Omega}{}^{zz} - (T_{k\Omega}{}^{xy})^2 \\ &\quad - (T_{k\Omega}{}^{yz})^2 - (T_{k\Omega}{}^{zx})^2, \text{ and} \\ T''_{k\Omega} &= \det T_{k\Omega}{}^{\alpha\beta} \end{aligned} \tag{5}$$

We note that $T'''_{k\Omega} = T_{k\Omega}{}^{xx} + T_{k\Omega}{}^{yy} + T_{k\Omega}{}^{zz}$ is yet another rotationally invariant representation that can be derived using tensorial fingerprint but is dismissed owing to its equivalence to the scalar component $S_{k\Omega}$. Thus, the total number of rotationally invariant components of the fingerprint for a system with n_Ω species represented using k Gaussians is $4kn_\Omega$ (i.e., 1 S, 1 V and 2 T components).

In this work, using the problem of classifying atomic environments in Al, C, and hafnia (HfO_2), we show that each of these scalar, vector, and tensorial components add distinct information about the local neighborhood. Thus, systematic improvements in model accuracy can be achieved by incorporating more complex vector and tensorial components, in addition to simple scalar terms. To classify distinct atomic environments in the aforementioned systems, we derive atomic fingerprints by setting the position of the reference atom as the value of the grid-point g in the above equations. Further, because atomic environments should be independent of the orientation of the system, we use rotationally invariant fingerprint components $S_{k\Omega}$, $V_{k\Omega}$, $T'_{k\Omega}$ and $T''_{k\Omega}$. Twenty Gaussians with width varying from 0.75 to 8 Å on a logarithmic scale, along with the cutoff parameter R_c = 8 Å were used.

Structure Fingerprint. The presented atomic fingerprint definition can be used to arrive at a structure fingerprint (G) to numerically represent a given configuration using

$$G_{k\theta\Phi}{}^{n} = \{M_\theta{}^{n}(S_{k\Phi});\ M_\theta{}^{n}(V_{k\Phi});\ M_\theta{}^{n}(T_{k\Phi})\} \tag{6}$$

where the scalar, vector, and tensor components are collected together after their transformation through the function $M^n(.)$, which represents the nth moment. To capture atomic neighborhood information of all possible elemental pairs, both θ and Φ loop over all of the elemental species Ω while respecting their order. For example, in case of hafnia, the different combinations will be $G_{kOO}{}^{n}$, $G_{kOHf}{}^{n}$, $G_{kHfO}{}^{n}$, and $G_{kHfHf}{}^{n}$. Further, if we put n = 1 in eq 6, the scalar component reduces to

$$G_{k\theta\Phi}{}^{1,\text{scalar}} = M_\theta{}^{1}(S_{k\Phi}) = \text{Avg}\left[\sum_{\text{atoms of specie }\theta}(S_{k\Phi})\right] \tag{7}$$

Similarly, other vector and tensorial components can also be obtained and collected together with the scalar part to obtain the structure fingerprint. The first moment-based structure fingerprint can be interpreted as the average atomic environment of the constituting species in the system. Further, it should be noted that by construction, structure fingerprint G is rotationally invariant and thus is suitable for distinguishing structures/conformers of a material or learning its energies (or any other rotationally invariant global quantity, such as pressure). We, thus, evaluate the performance of this fingerprint first for a classification problem of distinguishing different structures of C and then for a regression problem of learning potential energy model for Al. We restrict ourselves to structure fingerprint with just the first moment, that is, n = 1, as accurate results were obtained with such a simple definition.

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

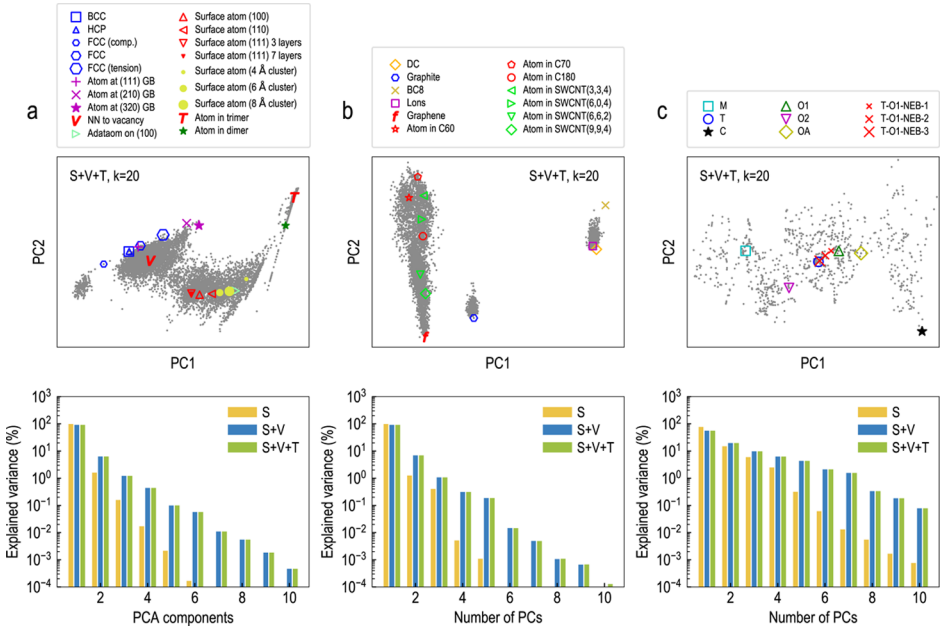


Figure 1. Ability of the presented fingerprint to characterize different atomic environments in (a) Al, (b) C, and (c) hafnia. While important atomic environments are highlighted using large colored symbols, thermal fluctuations in atomic configurations obtained from DFT-based MD simulations are presented using small gray symbols. The panels in bottom row demonstrate the systematic increase in the information captured by the fingerprint on addition of more complex vector and tensorial components.

Similar to the atomic fingerprints, R_c = 8 Å and Gaussian widths varying from 0.75 to 8 Å on a logarithmic scale were selected. The number of Gaussians was chosen to be 10 and 20.

Data Set. Comprehensive reference data sets, summarized in Table 1, were prepared for Al, C, and hafnia and are available for download at our online repository https://khazana.gatech.edu. To have enough diversity, configurations in different phases, surfaces, planar and point defects, and clusters and involving different levels of strains were incorporated. For each case, the commonly observed phases and configurations were included: (1) face-centered cubic (fcc) ($Fm\bar{3}m$), body-centered cubic (bcc) ($Im\bar{3}m$), and hexagonal close-packed (hcp) ($P6_3/mmc$) phases for Al, (2) dc ($Fd\bar{3}m$), graphite and bc8 ($Ia\bar{3}$) phases, and fullerenes and nanotubes for C, and (3) different polar and nonpolar phases in hafnia. To further add to the atomic diversity introduced due to thermal fluctuations, density functional theory (DFT)-based molecular dynamics (MD) simulations were performed for each configuration using the Vienna Ab initio simulation package (VASP).[41] While a substantial portion of the data set was obtained from our past studies,[15,18,42−44] a small subset was created as part of this work to further diversify the available atomic environments. Consistent with our previous works, we used the Perdew, Burke, and Ernzerhof exchange−correlation functional[45] and projector-augmented wave methodology.[46] The energy cutoff and Monkhorst−Pack k-point mesh[47] were carefully calibrated to ensure numerical convergence in both energy and atomic forces. For the case of C, about 57 additional structures belonging of several bulk phases were also obtained from Materials Project.[26] These 57 structures were exclusively used to assess the capability of the structure fingerprint to distinguish and quantify (dis)similarity between different conformations of a material. Among Al, C, and hafnia, the data set for Al was the most comprehensive and thus was used to train the potential energy model. Finally, we note that to validate the accuracy and transferability of the Al energy model, a few more configurations, distinct from the ones used for training, were also generated and will be discussed later.

RESULTS AND DISCUSSION

Classification: Atomic Environment. Figure 1 demonstrates the performance of our fingerprint for the classification problem of distinguishing different atomic environments included in the Al, C, and hafnia data sets. Because hafnia contains two elemental species, here we present results corresponding to the case of O atomic fingerprint only (similar results were achieved using Hf environment as well). Although the overall fingerprint size is much larger (Al: 80, C: 80, and hafnia: 160), we use the first two principal components (containing the maximum variance in data) to plot our results. Further for each system, we include three fingerprint scenarios: only scalar (S), both scalar and vector ($S + V$), and scalar, vector, and tensorial ($S + V + T$). Also, because we have a large number of atomic configurations obtained from MD simulations, we highlight only a few interesting cases in large colored symbols. The remaining atomic configurations (mostly thermal fluctuations and minor disorder) are denoted using small gray markers. Additionally, for this analysis, we restrict the number of C bulk phases to a few well-known phases only.

From Figure 1a,b, it can be clearly seen that the first principal component (PC 1) carries information discerning between the bulk-like and surface environments. For example, for Al, moving from left to right across the PC 1 axis, one encounters pure bulk-like fcc, bcc, and hcp, to vacancy nearest neighbors, to (111), (110), and (100) slabs, to clusters, to trimer, and finally to dimer environments. Similarly, for C, one can see bulk-like dc and bc8 phases on extreme right, while atomic environments with large surface areas, such as fullerenes and nanotubes [single-walled carbon nanotubes (SWCNTs)], on the extreme left. Graphite is understandably

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

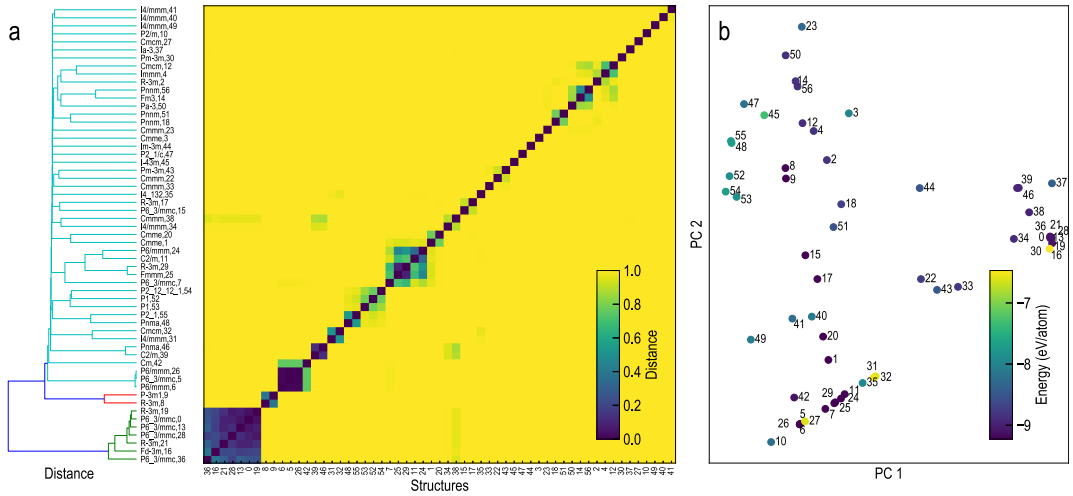


Figure 2. Ability of the presented structure fingerprint to characterize different phases of C. The (dis)similarity between structure fingerprints of various C configurations is illustrated using (a) dendrogram and heat map, and (b) first two principal components obtained through principal component analysis.

somewhere in between them. The PC 2 axis also carries important information, illuminating the effects of strain in Al and separating fullerenes, graphene, and SWCNTS in C. For the case of hafnia, we present data only for the six different bulk phases (see Figure 1c), which our fingerprint is easily able to differentiate. Interestingly, the polar O1 phase, which is argued to be responsible for the recent ferroelectric observations in hafnia, is very close to the parent T phase, suggesting structural similarity between the two phases. Snapshots from nudged elastic band computations connecting the T to O1 phase transformations are also visualized.[48] As suggested by Figure S1 in the Supporting Information, all of the above-mentioned trends were found to be more dominant for $S + V$ or $S + V + T$ definitions, in contrast to just the S definition of the fingerprint, suggesting that both $S + V$ and $S + V + T$ definitions capture more information. This is revealed explicitly in the scree plots presented in the bottom-row panels, wherein for both $S + V$ and $S + V + T$ definitions, larger number of components capture the variance in the data. Thus, both $S + V$ and $S + V + T$ fingerprint definitions are good at classification of atomic neighborhood.

Classification: Material Configuration. Next, we look at the performance of the structure fingerprint to distinguish different crystal phases of C using a data set of 57 structures. The (dis)similarity between configurations can be defined based on the L2 norm distance between the structure fingerprints of these configurations. Further, here, we normalize the L2 norm distance using a tanh() function to allow easy identification of similar structures. Using this (dis)similarity definition, a hierarchical dendrogram and heat map of the distance matrix are presented in Figure 2a, while the spread of this data set in the fingerprint space is illustrated using the first two principal components in Figure 2b. The dominance of the light regions in Figure 2a and segregation of phases in Figure 2b clearly suggest the ability of the fingerprint to differentiate the majority of these structures; note that the diagonal of the distance matrix represents the difference between same structures and thus is equal to zero. However, a few off-diagonal cases with close to zero difference between fingerprints can also be seen. On a closer inspection, these cases were found to belong to either the same structure or phases with extremely similar bonding. For example, three structures (# 5, 6, and 26) of graphene with different lengths of the vacuum region, and four structures (# 0, 13, 28, and 36) of $P6_3/mmc$ space group with only subtle changes in the bonding pattern were found; see Figures S2 and S3 in the Supporting Information. The former case reflects the significance of the fingerprint to assist identification of "duplicate" structures in large databases; it should be noted that even simulated X-ray diffraction would fail to classify these structures as identical because of different lengths along the axis normal to graphene layer, as elucidated in Figure S2 of the Supporting Information. A few cases with distinct structure fingerprints but belonging to the same space group can also be noticed in Figure 2a. Such cases were verified to be distinct by examining their energies and visualizing their structures (e.g., see notes on structures # 40, 41, and 49 with $I4/mmm$ space group in the Supporting Information). Thus, overall, this structure fingerprint was found to provide a good quantitative measure to describe (dis)similarity between configurations and could be utilized to characterize different phases of a material, find duplicates in large structural databases, or to identify previously sampled configurations during structure prediction.

Regression: Energy Model. The ability of the fingerprint to classify distinct material configurations already suggests its capability in developing accurate structure fingerprint-based regression models because learning—either classification or regression—happens in the same feature space fabricated by the fingerprint. To build a regression model that learns the energy of a system, two general strategies can be used. In the Behler and Parrinello approach,[34] the atomic fingerprints are mapped to atomic energies, the sum of which is fit to the total energy of the system. In another approach which was utilized in this work, the atomic fingerprints are combined to generate a structure fingerprint which is then mapped to the total energy. See Figure S5 in the Supporting Information for further discussion on the two approaches. Using the aforementioned Al data set and kernel ridge regression as the ML algorithm, we mapped the structure fingerprints (G) to the DFT-computed potential energies to create an Al energy model. The details on the procedure adopted to sample the training data and screen the best model and the behavior of the learning curves are provided in section S2 of the Supporting Information.

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

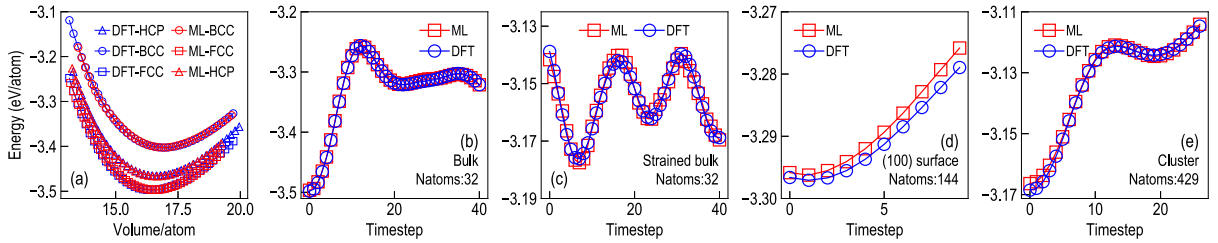


Figure 3. Prediction accuracy of the Al energy model developed in this work for various (a) static runs and (b–e) configurations obtained from DFT-based MD trajectories. The atomization energy is set as the reference in each case.

Figure 3 illustrates the performance of the energy model for various configurations in comparison to DFT. The energy model clearly captures the energy versus volume trends for fcc, bcc, and hcp phases as evident from Figure 3a, thereby, accurately predicting properties derived from this curve, such as the lattice constant, cohesive energy, and so forth. Further, our model performs well for snapshots obtained from DFT-MD trajectories on various Al conformations. A few such examples for fcc bulk, fcc bulk under compressive strain (a = 3.75 Å), (100) surface, and a cluster are shown in Figure 3b–e. We, however, note that these trajectories were also part of the larger reference data set from which the training set was sampled during the model construction. Thus, it is possible that the model has "seen" some of these configurations during the training. Nonetheless, the chances of seeing all configurations are negligible, as only ∼10% of the data was used for training (with some stochasticity in sampling as describe in the Supporting Information).

To have a more rigorous validation for our energy model, we constructed a few distinct configurations of Al twin boundaries, (100), (110), and (111) slabs and clusters of various sizes, which were different from the training set. The accurate prediction of the energy model (see Figure 4) for these

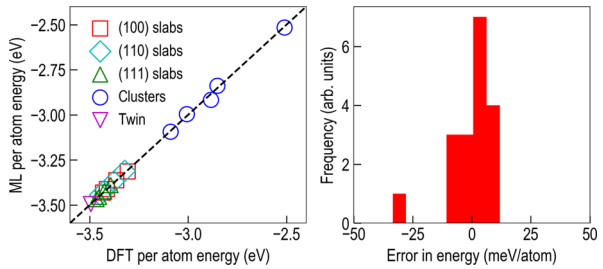


Figure 4. Prediction accuracy of the Al energy model developed in this work for configurations different from the training set, demonstrating the transferability of the model for new cases.

"unseen" cases suggests its generalizability and applicability to new cases. The root mean square error (RMSE) in prediction was found to be about 10 meV/atom, which is quite comparable to other ML-based models developed in the past.[18] Further, we note that this error is on completely "new" data, which is different from even the test set present during model construction for which a RMSE of about 3–4 meV/atom was achieved. Using this model, we also computed some general physical properties of Al, results for which are shown in Table 2. The model is able to accurately reproduce various bulk properties and defect formation energies of Al as computed using DFT. Thus, we believe that our energy model could provide inexpensive and quick estimates for

Table 2. Overall Performance of the Energy Model to Capture Various Physical Properties of Elemental Al, in Comparison to DFT

#	property	phase/name	DFT	ML	% error
1	cohesive energy (eV/atom)	fcc	−3.497	−3.497	−0.016
		bcc	−3.403	−3.403	−0.006
		hcp	−3.466	−3.467	0.024
2	V_0 (Å^3/atom)	fcc	16.492	16.582	0.548
		bcc	16.942	16.922	−0.119
		HCP	16.648	16.619	−0.172
3	bulk modulus (GPa)	fcc	76.072	78.441	3.115
		bcc	69.002	66.557	−3.543
		hcp	72.913	73.419	0.693
4	vacancy formation energy (eV)	fcc	0.742	0.730	−1.600
5	surface energy (J/m^2)	(100)	1.190	1.184	−0.438
		(110)	1.726	1.693	1.693
		(111)	0.869	0.853	−1.911
6	stacking fault (J/m^2)	(111)/[1−10]	0.136	0.144	5.805
7	grain boundary energies (J/m^2)	(210)	0.787	0.848	7.659
		(310)	0.957	0.991	3.590
8	self-diffusion energy (J/m^2)	fcc	0.865	0.711	−17.834

energies of diverse Al configurations, at least in the general domain it has been trained on.

To make our models easily accessible, we have also developed an online AGNI platform which can be accessed from https://khazana.gatech.edu, where the prediction models constructed here can be used for quick energy estimates. An atomic force model, which is based on the vector component of the fingerprint and developed in our previous work,[15] has also been included to allow quick atomic force predictions. The implementation of both the force and energy models will also be made available through the official version of large-scale atomic/molecular massively parallel simulator (LAMMPS).[49]

In a related recent work, we have used the grid-based local fingerprint definition along with NNs to learn and predict electronic charge density and local density of states.[16] Thus, future efforts will be directed toward using the same fingerprint definition (with relevant scalar, vector, and tensorial components) to learn electronic, atomic, or structural properties of a variety of materials. We also note that as with any other ML problem, the active learning scheme can be implemented using this fingerprint to continuously update and

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

improve the performance of the ML model for more diverse configurations.[50,51]

Finally, the atomic or structure fingerprint definitions presented here find applications in fields much beyond materials informatics. For example, the structure fingerprint can be used to answer a crucial question within structure search methods, that is, whether the newly sampled configuration is distinct from the list of previously explored configurations? Refining large material structural databases to remove duplicates or exploring their configurational diversity could be another application of this fingerprint. Similarly, the atomic fingerprint definition could be utilized to find atoms with specific defects or atomic neighborhood from large-scale MD simulations.

CONCLUSIONS

In conclusion, we presented a general fingerprint to capture the atomic neighborhood around an arbitrary point in space or about an atom. This fingerprint consists of scalar, vector, and tensorial components, each of which can be appropriately combined with existing ML approaches to learn diverse materials properties, such as energies, atomic forces, stresses, electronic charge density, electronic density of states, and so forth, each with different symmetry requirements. Moreover, to learn rotationally invariant properties (e.g., energies), the direction-dependent vector and tensorial components can be transformed into their respective rotationally invariant representations. Each of the scalar, vector, and tensor components was demonstrated to carry distinct information, allowing systematic improvement in the performance of fingerprint with the addition of more complex vector and tensorial components, besides the scalar component. The presented fingerprint definition was also used to formulate a structure fingerprint representation.

In the context of materials informatics, we demonstrated the ability of our fingerprint toward two general problems of classification and regression. For the former case, the fingerprint was used to classify different atomic environments in Al, C, and hafnia data sets, including phases, surfaces, planar and point defects, clusters, and so forth, thereby providing a reliable metric to identify structural defects or find duplicate phases. For the regression problem, we used structure fingerprint to build a general Al potential energy model, performance of which was validated for a diverse set of Al configurations, going well-beyond the training set. The accuracy of the energy model was validated for various physical properties of Al and was found to be in good agreement with DFT. Thus, overall, the simple and elegant fingerprint presented here is believed to be suitable for a large class of chemistry and materials science problems that involve numerically representing the atomic neighborhood information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcc.9b03925.

Discussion on the classification ability of the presented fingerprint and the overall strategy adopted to build the Al energy model, including the machine approach, training set sampling, model screening and behavior of the learning curves and additional results on Al elastic constants obtained from the energy model (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: rampi.ramprasad@mse.gatech.edu.

ORCID

Rohit Batra: 0000-0002-1098-7035
Huan Doan Tran: 0000-0002-8093-9426
Chiho Kim: 0000-0002-1814-4980
James Chapman: 0000-0001-8451-0275
Rampi Ramprasad: 0000-0003-4630-1565

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The financial support of this work through grant no. N00014-18-1-2113 from the Office of Naval Research and partial computational support through a Extreme Science and Engineering Discovery Environment (XSEDE) allocation number TG-DMR080058N are acknowledged.

REFERENCES

(1) Lookman, T.; Balachandran, P. V.; Xue, D.; Hogden, J.; Theiler, J. Statistical Inference and Adaptive Design for Materials Discovery. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 121−128.

(2) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* **2017**, *129*, 156−163.

(3) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122*, 17575−17585.

(4) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction Model of Band Gap for Inorganic Compounds by Combination of Density Functional Theory Calculations and Machine Learning Techniques. *Phys. Rev. B* **2016**, *93*, 115104.

(5) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**, *2*, 16028.

(6) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual Screening of Inorganic Materials Synthesis Parameters with Deep Learning. *npj Comput. Mater.* **2017**, *3*, 53.

(7) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436−9444.

(8) Segler, M. H. S.; Waller, M. P. Neural-symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem.—Eur. J.* **2017**, *23*, 5966−5971.

(9) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction using Neural Sequence-to-sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103−1113.

(10) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful Classification of Crystal Structures using Deep Learning. *Nat. Commun.* **2018**, *9*, 2775.

(11) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput Machine-learning-driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324−7331.

(12) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(13) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866

(14) Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R. A Universal Strategy for the Creation of Machine Learning-based Atomistic Force Fields. *npj Comput. Mater.* **2017**, *3*, 37.
(15) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2016**, *121*, 511−522.
(16) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the Electronic Structure Problem with Machine Learning. *npj Comput. Mater.* **2019**, *5*, 22.
(17) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.
(18) Pun, G. P. P.; Batra, R.; Ramprasad, R.; Mishin, Y. Physically-informed Artificial Neural Networks for Atomistic Modeling of Materials. *Nat. Commun.* **2019**, *10*, 2339.
(19) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2018**, *5*, 57−64.
(20) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.
(21) Mueller, T.; Kusne, A. G.; Ramprasad, R. *Reviews in Computational Chemistry*; John Wiley & Sons, Inc, 2016; pp 186−273.
(22) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.
(23) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; et al. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(24) Grisafi, A.; Wilkins, D. M.; Willatt, M. J.; Ceriotti, M. Atomic-scale Representation and Statistical Learning of Tensorial Properties. **2019**, arXiv:1904.01623.
(25) Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)-Present and Future. *Crystallogr. Rev.* **2004**, *10*, 17−22.
(26) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.
(27) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(28) Huo, H.; Rupp, M. Unified Representation for Machine Learning of Molecules and Crystals. **2017**, arXiv:1704.06439; pp 13754−13769.
(29) Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including Crystal Structure Attributes in Machine Learning Models of Formation Energies via Voronoi Tessellations. *Phys. Rev. B* **2017**, *96*, 024104.
(30) Willighagen, E. L.; Wehrens, R.; Verwer, P.; De Gelder, R.; Buydens, L. M. C. Method for the Computational Comparison of Crystal Structures. *Acta Crystallogr., Sect. B: Struct. Sci.* **2005**, *61*, 29−36.
(31) de Gelder, R.; Wehrens, R.; Hageman, J. A. A Generalized Expression for the Similarity of Spectra: Application to Powder Diffraction Pattern Classification. *J. Comput. Chem.* **2001**, *22*, 273−289.
(32) Zhu, L.; Amsler, M.; Fuhrer, T.; Schaefer, B.; Faraji, S.; Rostami, S.; Ghasemi, S. A.; Sadeghi, A.; Grauzinyte, M.; Wolverton, C.; et al. A Fingerprint Based Metric for Measuring Similarities of Crystalline Structures. *J. Chem. Phys.* **2016**, *144*, 034203.
(33) Valle, M.; Oganov, A. R. Crystal fingerprint space—a novel paradigm for studying crystal-structure sets. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2010**, *66*, 507−517.
(34) Behler, J.; Parrinello, M. Generalized Neural-network Representation of High-dimensional Potential-energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(35) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(36) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(37) Seko, A.; Togo, A.; Tanaka, I. Group-theoretical high-order rotational invariants for structural representations: Application to linearized machine learning interatomic potential. **2019**, arXiv:1901.02118.
(38) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. **2019**, arXiv:1904.08875.
(39) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. **2018**, arXiv:1812.05055.
(40) Chapman, J.; Batra, R.; Uberuaga, B. P.; Pilania, G.; Ramprasad, R. A Comprehensive Computational Study of Adatom Diffusion on the Aluminum (1 0 0) Surface. *Comput. Mater. Sci.* **2019**, *158*, 353−358.
(41) Kresse, G.; Furthmüller, J. Efficient iterative schemes forab initiototal-energy calculations using a plane-wave basis set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169−11186.
(42) Batra, R.; Tran, H. D.; Ramprasad, R. Stabilization of Metastable Phases in Hafnia Owing to Surface Energy Effects. *Appl. Phys. Lett.* **2016**, *108*, 172902.
(43) Batra, R.; Huan, T. D.; Rossetti, G. A.; Ramprasad, R. Dopants Promoting Ferroelectricity in Hafnia: Insights from a Comprehensive Chemical Space Exploration. *Chem. Mater.* **2017**, *29*, 9102−9109.
(44) Batra, R.; Huan, T. D.; Jones, J. L.; Rossetti, G.; Ramprasad, R. Factors Favoring Ferroelectricity in Hafnia: A First-Principles Computational Study. *J. Phys. Chem. C* **2017**, *121*, 4139−4145.
(45) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(46) Blöchl, P. E. Projector Augmented-wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953−17979.
(47) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-zone Integrations. *Phys. Rev. B: Solid State* **1976**, *13*, 5188−5192.
(48) Huan, T. D.; Sharma, V.; Rossetti, G. A., Jr.; Ramprasad, R. Pathways Towards Ferroelectricity in Hafnia. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 064111.
(49) Plimpton, S. Fast Parallel Algorithms for Short-range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.
(50) Huan, T. D.; Batra, R.; Chapman, J.; Kim, C.; Chandrasekaran, A.; Ramprasad, R. Iterative-learning Strategy for the Development of Application-specific Atomistic Force Fields **2019**, under review.
(51) Kim, C.; Chandrasekaran, A.; Jha, A.; Ramprasad, R. Active-Learning and Materials Design: The Example of High Glass Temperature Polymers. *MRS Commun.* **2019**, 1.

DOI: 10.1021/acs.jpcc.9b03925
J. Phys. Chem. C 2019, 123, 15859−15866