The Journal of Chemical Physics

Dimensionality reduction in machine learning for nonadiabatic molecular dynamics: Effectiveness of elemental sublattices in lead halide perovskites

Cite as: J. Chem. Phys. **156**, 054110 (2022); https://doi.org/10.1063/5.0078473
Submitted: 12 November 2021 • Accepted: 16 January 2022 • Accepted Manuscript Online: 17 January 2022 • Published Online: 02 February 2022

Wei Bin How, Bipeng Wang, Weibin Chu, et al.

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Perspective on integrating machine learning into computational chemistry and materials science
The Journal of Chemical Physics **154**, 230903 (2021); https://doi.org/10.1063/5.0047760

A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu
The Journal of Chemical Physics **132**, 154104 (2010); https://doi.org/10.1063/1.3382344

Analysis of the classical trajectory treatment of photon dynamics for polaritonic phenomena
The Journal of Chemical Physics **156**, 054101 (2022); https://doi.org/10.1063/5.0079379

AIP Publishing

The Journal of Chemical Physics

ARTICLE

scitation.org/journal/jcp

Dimensionality reduction in machine learning for nonadiabatic molecular dynamics: Effectiveness of elemental sublattices in lead halide perovskites

Cite as: J. Chem. Phys. 156, 054110 (2022); doi: 10.1063/5.0078473
Submitted: 12 November 2021 • Accepted: 16 January 2022 •
Published Online: 2 February 2022

View Online Export Citation CrossMark

Wei Bin How,[1] Bipeng Wang,[2] Weibin Chu,[3] Sergiy M. Kovalenko,[4,5] Alexandre Tkatchenko,[6] and Oleg V. Prezhdo[2,3,7,a)]

AFFILIATIONS

[1] Division of Chemistry and Biological Chemistry, School of Physical and Mathematical Sciences, Nanyang Technological University, 637371, Singapore

[2] Department of Chemical Engineering, University of Southern California, Los Angeles, California 90089, USA

[3] Department of Chemistry, University of Southern California, Los Angeles, California 90089, USA

[4] Department of Organic Chemistry, V. N. Karazin Kharkiv National University, Kharkiv 61022, Ukraine

[5] I. M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), Moscow 119991, Russian Federation

[6] Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg

[7] Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90089, USA

[a)] Author to whom correspondence should be addressed: prezhdo@usc.edu

ABSTRACT

Supervised machine learning (ML) and unsupervised ML have been performed on descriptors generated from nonadiabatic (NA) molecular dynamics (MD) trajectories representing non-radiative charge recombination in $CsPbI_3$, a promising solar cell and optoelectronic material. Descriptors generated from every third atom of the iodine sublattice alone are sufficient for a satisfactory prediction of the bandgap and NA coupling for the use in the NA-MD simulation of nonradiative charge recombination, which has a strong influence on material performance. Surprisingly, descriptors based on the cesium sublattice perform better than those of the lead sublattice, even though Cs does not contribute to the relevant wavefunctions, while Pb forms the conduction band and contributes to the valence band. Simplification of the ML models of the NA-MD Hamiltonian achieved by the present analysis helps to overcome the high computational cost of NA-MD through ML and increase the applicability of NA-MD simulations.

I. INTRODUCTION

The field of chemistry was transformed dramatically with the introduction of computational tools and methods. Chemists now have access to a wide array of approaches to predict nearly every observable property for molecules and materials.[1–6] However, even with the powerful computing resources we have today, the immense computational cost of these tools renders it impractical to apply rigorous quantum-chemical treatment for large systems, limiting the quality of predictions for such systems.[7–10]

The recent integration of machine learning (ML) into quantum and theoretical chemistry aims to circumvent the high computational load.[11–13] ML seeks to uncover non-trivial patterns and trends from large datasets for the prediction of some properties in the dataset. One of the first applications of ML in quantum chemistry was the use of a neural network to learn the relationship between the atomic positions and the potential energy surface.[14] Since then, fueled by the recent explosion of data generated by molecular simulations, there have been significant developments in ML techniques or algorithms in atomistic simulations.[15–19]

Lately, it has been shown that ML can be applied to nonadiabatic (NA) molecular dynamics (MD) simulations.[20–31] NA-MD is a powerful tool for the study of excited-state dynamics, involving quantum transitions between states, in a wide range of chemical systems.[32–37] NA-MD simulations have been used to characterize the ultrafast response of molecules to external electromagnetic fields and to rationalize the results of time-resolved spectroscopic experiments.[38–43] The reliability of the NA-MD simulations depends heavily on the accuracy of the geometry-dependent forces and energies and the NA coupling (NAC) between the ground and excited states. Traditionally, these values are obtained via *ab initio* calculations performed with system geometries along MD trajectories. However, these calculations tend to be extremely computationally intensive, limiting the application of NA-MD for large systems and long timescales. To address this drawback, ML has been applied to predict the energies and NAC between ground and excited states, significantly reducing the computational load for NA-MD simulations.[44–51] Aside from the prediction of physical observables, such as bandgap, ML has also been applied on the trajectories of NA-MD simulations to discover structural factors that influence the physical properties of materials.[21,52–54] One such technique is the use of mutual information (MI), which acts as a measure of the mutual dependence between two variables. The advantage of MI is that it is supported by an information theoretic background, insensitive to the size of the dataset, and its results are relatively easy to interpret.[55,56]

We focus on $CsPbI_3$, which is a well-studied representative of metal halide perovskites (MHPs). Due to their relatively low cost and unique optoelectronic properties, MHPs have led to significant developments in photovoltaics and have shown great promise as candidates for light emitting diodes and other applications.[57–67] There exists a strong demand for the development of methods to perform low-cost NA-MD on MHP for the streamlining of the MHP design process, integrating theoretical approaches to predict the important physical properties of potential candidates,[68–71] including bandgap, electron-vibrational coupling, and charge carrier lifetimes. The geometric structure and projected density of states (PDOS) of $CsPbI_3$ can be seen in Fig. 1. The energy gap between the valence band maximum (VBM) and the conduction band maximum (CBM) is 1.67 eV. The VBM is primarily supported by Pb and I atoms, while the CBM is supported by Pb atoms.

We apply MI to study a variety of ML models of the NA-MD Hamiltonian with the goals of understanding the relationships between system geometries and the Hamiltonian and using this understanding to construct minimal ML models capable of achieving accurate NA-MD simulation results. We investigate the extent in which the sublattices of the individual elements in $CsPbI_3$ encapsulate the quantum mechanics of the charge carriers. We quantify the accuracy of the prediction of the bandgap and the scalar NAC by a kernel ridge regressor (KRR) model based on the geometric properties of the sublattice. Then, using different sets of descriptors, we perform NA-MD simulations based on the KRR models of the NA-MD Hamiltonian and compare the results with the simulations based on the *ab initio* density functional theory (DFT) Hamiltonian. We demonstrate a significant reduction of the dimensionality of the standard ML model used in the development of ML force-fields. In particular, descriptors generated from the iodine sublattice of $CsPbI_3$ are sufficient to predict the bandgap and the scalar NAC and to obtain an accurate charge carrier lifetime. Surprisingly, KRR models based on the cesium sublattice perform better than those of the lead sublattice, even though Cs does not contribute to the relevant wavefunctions, while Pb determines the CBM and contributes to the VBM, Fig. 1. These findings highlight the complex interplay of different structural components in MHPs.

II. METHODS

The *ab initio* nonadiabatic (NA) molecular dynamics (MD) simulation of the pristine $CsPbI_3$ system was performed using the Pyxaid software.[1,72] The tetragonal lattice was represented by a $\sqrt{2} \times \sqrt{2} \times 2$ unit cell with the lattice constants of $9.02 \times 9.02 \times 12.76$ Å^3. The nonradiative charge recombination process was simulated using the decoherence-induced surface hopping (DISH) approach under the classical path approximation (CPA) to reduce the computational load of the calculations.[73] A hybrid quantum–classical approach is used, whereby the slower and heavier ion cores are handled classically, while the electrons are treated quantum mechanically using real-time time-dependent density functional theory (TF-DFT). The decoherence time was approximated using the pure-dephasing time, calculated via the second order cumulant approximation of the optical response theory.[74,75] As the decoherence time is significantly shorter than the charge carrier lifetime, the effects of decoherence are significant and should be accurately represented. Hence, the DISH approach was used for the NA-MD simulation, as it can accurately represent decoherence effects.

The geometric structure of $CsPbI_3$ was obtained via *ab initio* calculations using the Vienna *Ab Initio* Simulation Package (VASP)

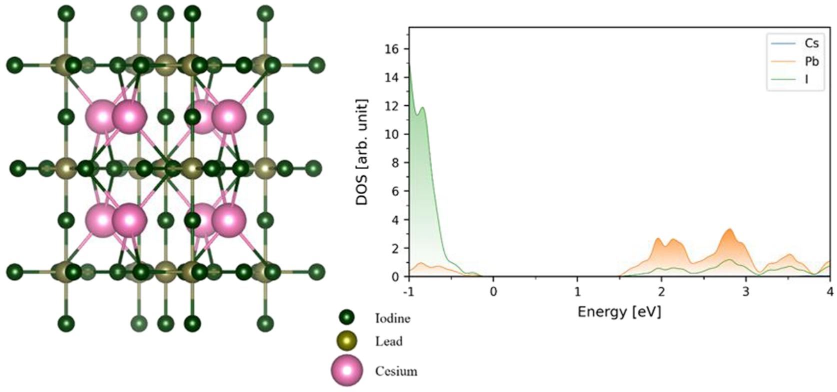


FIG. 1. (Left) Geometric structure of pristine $CsPbI_3$. (Right) Projected density of states (DOS) for pristine $CsPbI_3$, with the VBM energy set to zero.

and the Perdew–Burke–Ernzerhof (PBE) functional.[76–78] The structure shown in Fig. 1 was first optimized, then heated up, and equilibrated at room temperature. Subsequently, a 7 ps trajectory was generated using a time step of 1 fs in the microcanonical ensemble. Under this approach, we can precompute a trajectory and then obtain the NAC along the precomputed trajectory. The bandgap and NA coupling (NAC) values were also obtained using VASP and the PBE exchange–correlational functional. The CA (Concentric Approximation)-NAC package was employed to compute the scalar NAC values from the overlap between two wavefunctions at adjacent timesteps using the expression as follows:[79,80]

$$\begin{aligned} d_{ji} &= -i\hbar\langle\varphi_j(\boldsymbol{r},\boldsymbol{R}(t))|\nabla_{\boldsymbol{R}}|\varphi_i(\boldsymbol{r},\boldsymbol{R}(t))\rangle\frac{d\boldsymbol{R}}{\mathrm{d}t} \\ &= -i\hbar\frac{\langle\varphi_j(\boldsymbol{r},\boldsymbol{R}(t))|\nabla_{\boldsymbol{R}}H(\mathbf{R}(t))|\varphi_i(\boldsymbol{r},\boldsymbol{R}(t))\rangle}{E_i - E_j}\frac{d\boldsymbol{R}}{\mathrm{d}t} \\ &= -i\hbar\left\langle\varphi_j(\boldsymbol{r},\boldsymbol{R}(t))\left|\frac{\partial}{\partial t}\right|\varphi_i(\boldsymbol{r},\boldsymbol{R}(t))\right\rangle \\ &\approx -\frac{i\hbar}{2\Delta t}\{\langle\varphi_j(\boldsymbol{r},\boldsymbol{R}(t))|\varphi_i(\boldsymbol{r},\boldsymbol{R}(t+\Delta t))\rangle \\ &\quad - \langle\varphi_j(\boldsymbol{r},\boldsymbol{R}(t+\Delta t))|\varphi_i(\boldsymbol{r},\boldsymbol{R}(t))\rangle\}. \end{aligned} \tag{1}$$

From the trajectory obtained, 4% of the dataset, representing 280 datapoints equally spaced along the trajectory, was selected as the training set, while the remainder was designated as the test set. Interpolating NAC along the precomputed trajectory has been shown to be efficient in reducing the number of required NACs by more than one order of magnitude.[44] As NAC exhibits more complex dependence on nuclear geometry than energy, interpolating NAC along a precomputed trajectory is much easier than developing NAC models that work for all relevant geometries, similar to the machine learning force field (ML-FF) models. Currently, we are pursuing a strategy by developing ML-FFs and then performing NA-MD calculations under the CPA by sampling a small fraction of energy gaps and NAC and interpolating the remaining values. Such an approach has uncovered important rare events that change NA-MD.[81]

Additionally, the autocorrelation functions (ACFs) of the energy gap and NAC were computed, and the results are presented in Fig. 2. The ACFs show that the bandgap and NAC ACFs decay to 0 within 200–300 fs and then oscillate over a range of frequencies, indicating that the bandgap and NAC values do not exhibit strong correlations over long timescales and in particular over the 7 ps trajectory.

In contrast to the input to the Schrödinger equation, nuclear charges and atomic positions are not good inputs for ML as they lack certain desirable properties, such as roto-translational invariance. Forcefully implementing roto-translational invariance for atomic positions will lead to a significant increase in required training data, resulting in lengthy training.[16] Hence, the geometric information of the sublattices obtained from the trajectory were first converted into descriptors before they were used as inputs for the ML model. As there are three times as many iodine atoms as cesium and lead atoms, only every third iodine atom was chosen systematically, taken to be to the right of every lead atom in the simulation cell, to define the iodine sublattice. This allowed us to treat all three elemental sublattices on equal footing. Additionally, we also compared the performance of these descriptors against that of the descriptors obtained from the full iodine sublattice, containing all 12 iodine atoms. To conserve the total size of the dataset, the model trained on the descriptors obtained from the full iodine sublattice used 1/3 of the original training set, comprising 93 datapoints equally spaced along the trajectory. The descriptors were extracted from the dataset via the use of a symmetry function adapted from the work of Smith, Isayev, and Roitberg,[82] whereby an adapted version of Behler and Parrinello's symmetry function was used to include radial and angular information.[14] The symmetry function used is as follows:

$$\begin{aligned} G_i^{\mathrm{mod}} &= 2^{1-\zeta}\sum\nolimits_{j,k\neq i}^{\mathrm{atoms}}\left(1+\cos\left(\theta_{ijk}-\theta_s\right)\right)^{\zeta} \\ &\quad\times e^{\left[-\eta\left(\frac{\mathrm{R}_{ij}+\mathrm{R}_{ik}}{2}-\mathrm{R}_s\right)^2\right]}\mathrm{f}_{\mathrm{C}}(\mathrm{R}_{ij})\mathrm{f}_{\mathrm{C}}(\mathrm{R}_{ik}). \end{aligned} \tag{2}$$

Given atoms i, j, and k, θ_{ijk} represents the angle centered on atom i, while R_{ij} and R_{ik} refer to the distance between atoms i and j and atoms i and k, respectively. This symmetry function probes both the local radial information of atom i via the Gaussian terms and the local angular environment via the cosine terms. The parameters θ_s and R_s tune the centers, while the parameters ζ and η tune the widths of the angular and Gaussian terms, respectively. In this work, ζ and η were defined as 1 and 0.15, respectively, in order to maintain the angular and radial features at similar magnitudes for an appropriate representation of both features in the dataset. The R_s and θ_s values

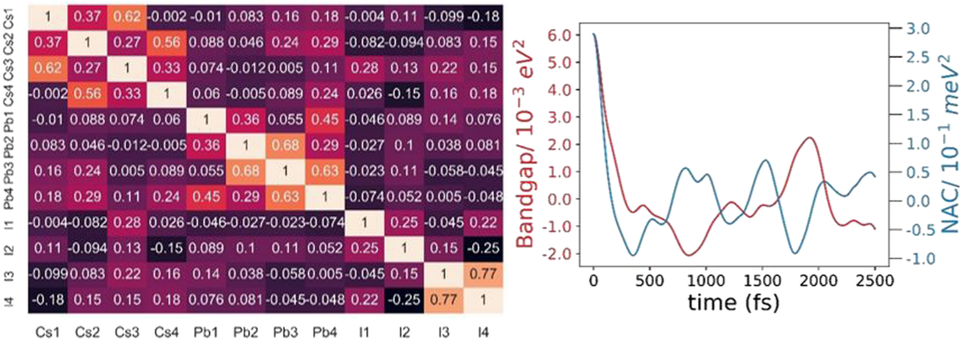
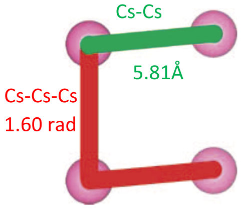
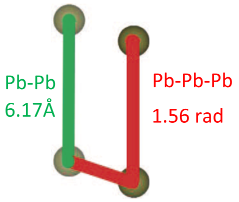
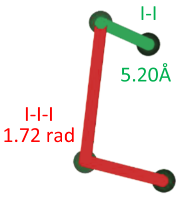


FIG. 2. (Left) Correlation matrix for the descriptors of the cesium, lead, and iodine (four atoms) sublattices. (Right) Autocorrelation functions for the bandgap and NAC.

TABLE I. Illustration of the angles and distances for the computation of θ_s [Eq. (1)]. The R_s and θ_s values were set to the average minimum distance and maximum angle between the atoms in the particular sublattice.

Cesium sublattice	Lead sublattice	Iodine sublattice
		

were set to the minimum distance and maximum angle between the atoms in the sublattice averaged over the entire duration of the NA-MD simulation. Illustrations of the angles and distances used for the calculation of θ_s and R_s are shown in Table I.

$f_C(R)$ represents a piecewise cutoff function to approximate the local chemical environment of the atom and is defined as follows:

$$f_C(R) = \begin{cases} 0.5 \times \cos\left(\frac{\pi R_{ij}}{R_C}\right) + 0.5 & \text{for } R_{ij} \le R_C, \\ 0.0 & \text{for } R_{ij} > R_C. \end{cases} \tag{3}$$

The radial cutoff R_C determines the size of the local chemical environment of the atom. It was set to 9.1 Å, corresponding to the distance between the center of the rectangular simulation cell and its vertex (Fig. 1).

The pairwise MI, $I(X, Y)$, between two features was calculated by estimating entropy from k-nearest neighbor distances. First, the supports of X and Y are partitioned into bins of finite size based on the Chebyshev distance. Then, taking R to be the Chebyshev distance between a point, z_i, and its kth neighbor, we obtain the total number of points, n_x and n_y, whose distance from z_i is strictly less than R in the X and Y subspaces, respectively. The mutual information can then be estimated via the following formula:

$$I(X,Y) = \langle \psi(k) - \psi(n_x+1) + \psi(n_y+1)\rangle + \psi(N). \tag{4}$$

Here, $\psi(x)$ represents the digamma function of both variables, while N represents the total number of datapoints. In this study, k was set to 3, as it has been demonstrated to be the optimal parameter for halide perovskites.[55,56,83]

Ridge regression is a regression method that tackles the issue of overfitting by including an additional variable for regularization, $\|w\|^2$, representing the square of the coefficients of the model. This term adds a penalty to complex models with high regression coefficients, thus encouraging simpler models with small coefficients. Using the ordinary least squares cost function as an example, the total cost function is as follows:

$$C = \frac{1}{2}\sum_i (y_i - w^T x_i)^2 + \frac{1}{2}\lambda\|w\|^2, \tag{5}$$

where λ represents the regularization factor and y_i and x_i represent the target variable and the predictor variable of the ith data point. Kernel ridge regression extends ridge regression to a nonlinear case by learning from a non-linear feature space induced by the kernel and the data. In this way, the predictor variables are replaced by their corresponding feature vectors: $x_i \to \Phi_i = \Phi(x_i)$, induced by a kernel, whereby $k(x_i, x_j) = \Phi(x_i)^T\Phi(x_j)$. Using the kernel trick, it is possible to work with the inner product of the feature vectors rather than the vectors themselves, allowing for significant computational savings during training and prediction, as the dimensionality of the feature vectors can be extremely high.

Kernel trick,

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}. \tag{6}$$

Prediction of x,

$$w = (\lambda I_d + \Phi\Phi^T)^{-1}\Phi y = \Phi(\Phi^T\Phi + \lambda I_n)^{-1} y, \tag{7}$$

$$y = w^T\Phi = y(\Phi^T\Phi + \lambda I_n)^{-1}\Phi^T\Phi(x) = y(K + \lambda I_n)^{-1}\kappa(x), \tag{8}$$

where $K(bx_i, bx_j) = \Phi(x_i)^T\Phi(x_j)$ and $\kappa(x) = K(x_i, x_j)$.

The Laplacian kernel was employed, $K(x,y) = e^{-\gamma\|x-y\|_1}$, whereby x and y represent two separate input vectors and $\|x-y\|_1$ represents the Manhattan distance between them.

The feature generation of descriptors, the calculation of MI, and the training and evaluation of the model were performed with the Scikit-learn package using Python.[84] All the models used KRR based on the Laplacian kernel with the L2 penalty set to 0.0001.

III. RESULTS AND DISCUSSION

Table II shows the mean MI for the descriptors of each elemental sublattice with the bandgap and NAC. It is expected that the MI value with the NAC is smaller than that with the bandgap since the calculation for the bandgap involves nuclear positions, which are represented by the descriptors, while the NAC also depends explicitly on nuclear velocity [Eq. (1)], which is not captured by the descriptors. In addition, the NAC is a more complex function of nuclear geometry than the bandgap, and therefore, it may be more challenging for the simple descriptors to predict the NAC than the bandgap. It is natural that the descriptors for the iodine sublattice have a higher MI than those of cesium since the VBM is partially supported by iodine atoms, while cesium contributes to neither the VBM nor the CBM, as seen in the PDOS plot in Fig. 1. However, it is surprising that the MI for lead is the lowest, considering that both the

TABLE II. Mean mutual information (MI) values [Eq. (4)] and adjusted R^2 score for the descriptors [Eq. (2)] for each elemental sublattice in $CsPbI_3$.

Elemental sublattice	MI with bandgap	Bandgap model R^2 score	MI with NAC	NAC model R^2 score
Cesium	1.38	0.86	1.35	0.88
Lead	1.04	0.47	1.03	0.47
Iodine (4 atoms)	1.66	0.99	1.60	0.96
Iodine (12 atoms; 1/3 of training data)	1.58	0.82	1.54	0.88

VBM and the CBM are supported by lead. One would expect *a priori* that lead descriptors should have the highest MI with the bandgap and the NAC. The above facts might be rationalized by the importance of the iodine octahedral structure in determining the values for the bandgap and the NAC, as indicated in the previous analyses[21,52] based on the traditional bond angles and lengths rather than the current ML descriptors. It is likely that the cesium sublattice can provide information regarding octahedral tilt because cesium atoms interact with iodines from different octahedra, while leads are nearest neighbors to iodines from the same octahedron, and therefore, the descriptors associated with leads cannot characterize the octahedral tilts. The iodine sublattice can provide information regarding the iodine octahedral structure as well.

The environment of a particular atom can have a strong influence on the properties of the atom. In particular, such a situation arises in liquids in which the properties of the ion depend significantly on the surrounding solvent. Interestingly, the properties of polarons in MHPs can be compared to solvated charges in liquids,[85,86] in particular since MHPs are softer than traditional inorganic semiconductors, undergo large scale anharmonic motions,[87,88] and can contain components with asymmetric charge distribution, such as $CH_3NH_3^+$, which can rotate and "solvate" the charges. In order to establish the extent of correlation between the different elemental sublattices, we computed correlation coefficients between the descriptors for cesium, lead, and iodine (four atoms). The results are presented in Fig. 2. The correlation coefficients, r, between two variables x and y, are calculated using the formula as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \tag{9}$$

where x_i and y_i represent the value of the x and y variables of the ith data point, while $\bar{x}$ and $\bar{y}$ represent the means of the x and y variables, respectively. It should be noted that the symmetries of the perfect tetragonal perovskite lattice are perturbed by thermal fluctuations. Due to the relatively short length of the trajectory, the symmetries are not fully recovered by the ensemble averaging, and the data shown in the correlation coefficient matrix should be interpreted in a semi-quantitative way. The data demonstrate that the atoms of a particular type exhibit notable correlation (4 × 4 blocks around the diagonal). However, the atoms of different types show little correlation, indicating that the sublattices are quite independent.

Figures 3 and 4 show the results of the prediction of both bandgap and NAC based on the individual element sublattice descriptors. As expected, the performance of the models follows the same ranking as that for MI, but it is interesting to note that the

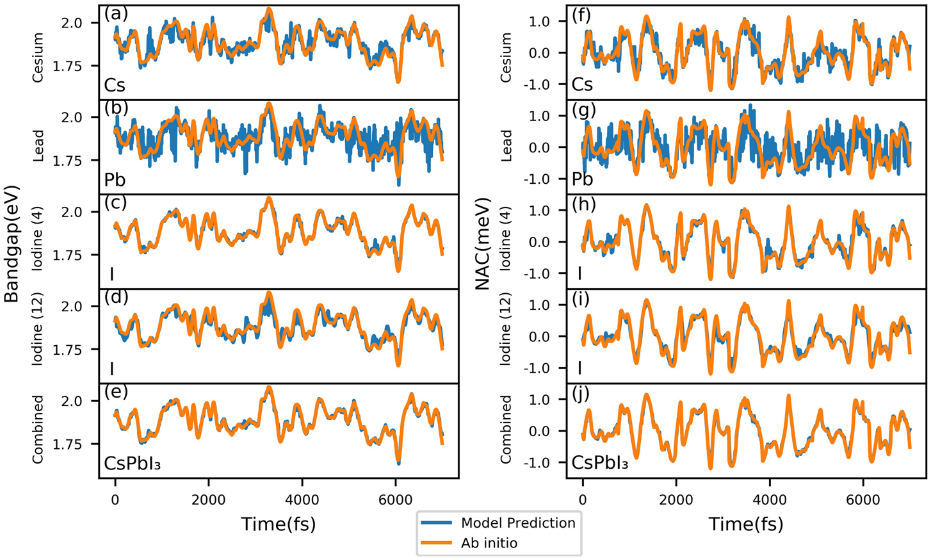


FIG. 3. Individual element sublattice prediction vs *ab initio* values for bandgap and NAC. (a)–(d) refer to models trained on cesium, lead, and iodine sublattice descriptors, respectively, for the prediction of bandgap. (f)–(i) refer to models trained on cesium, lead, and iodine sublattice descriptors for the prediction of NAC. (e) and (j) refer to models trained on the descriptors of all three sublattices to predict bandgap and NAC, respectively.

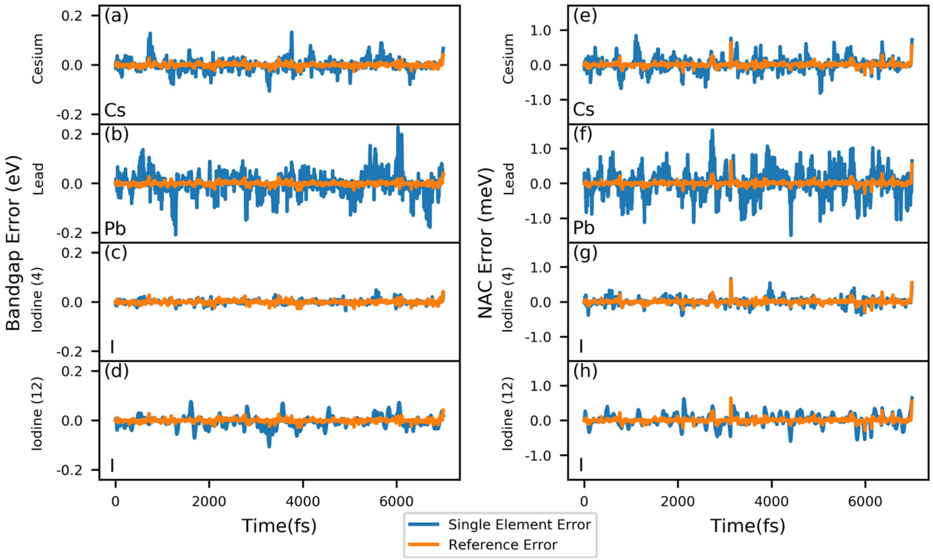


FIG. 4. Individual element sublattice prediction errors plotted against the reference error, which is the error made by the model trained on the descriptors of all three elements. (a)–(d) refer to models trained on descriptors obtained from cesium, lead, and iodine sublattices, respectively, for the prediction of bandgap. (e)–(h) refer to models trained on descriptors obtained from cesium, lead, and iodine sublattices, respectively, for the prediction of NAC.

iodine sublattice alone can provide a satisfactory fit for both bandgap and NAC despite only partially supporting the VBM and having no contributions to the CBM even though both VBM and CBM properties are essential for the calculation of the bandgap and the NAC. Meanwhile, the lead sublattice, despite supporting both the VBM and the CBM, provides the worst fit for both the bandgap and the NAC. The performance of the model based on the iodine sublattice alone is very similar to that of the model based on the combined descriptors from all three sublattices. Additionally, Table II also presents the adjusted R^2 score for the respective models. The adjusted R^2 score was chosen as the metric to account for the different numbers of predictors in the models. The correlation coefficients confirm the conclusion obtained based on the mutual information. Iodines and cesiums provide significantly better predictions of the bandgap and nonadiabatic coupling than Pb atoms. Furthermore, training on 1/3 of the training data on the descriptors obtained from the full iodine sublattice (12 atoms) provides a viable alternative to training based on 1/3 of iodine atoms (four atoms) and all training data. The descriptors obtained from this alternative strategy have lower MI values (Table II), and the corresponding KRR models perform slightly worse for both the bandgap and the NAC. Interestingly, this indicates that the breaking of the rotational symmetry of the lattice, caused by decreasing the size of the iodine sublattice from 12 atoms to 4, can provide better results.

Afterward, we used the bandgap and NAC values obtained from the ML models to perform NA-MD simulations. Figure 5 and Table III show the NA-MD simulation results. It is interesting that using descriptors from the lead sublattice leads to a slightly better NA-MD performance for NA-MD than the cesium sublattice despite its significantly worse performance for the prediction of bandgap and NAC (Figs. 3 and 4). The descriptors from the iodine sublattice and the combined descriptors still retain the top positions for both the prediction of bandgaps and NACs and NA-MD simulations. It is noteworthy that the performance of the NA-MD simulations using only the iodine sublattice is rather satisfactory, providing a

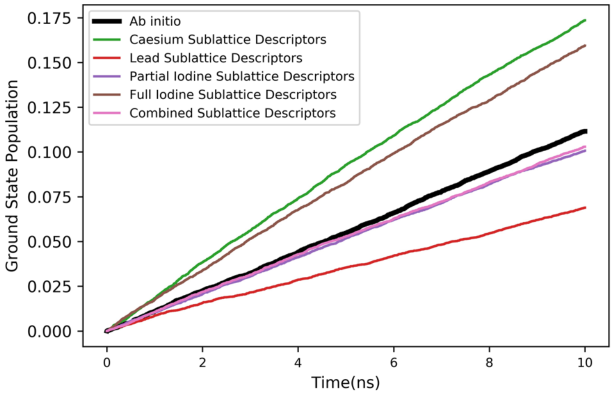


FIG. 5. NA-MD results based on bandgap and NAC calculated via *ab initio* methods or predicted via the KRR models. The simulation calculates the population of the ground state over 10 ns.

TABLE III. Results of the NA-MD simulations represented via the gradient line of the best fit of the data shown in Fig. 5. The results are ordered according to the deviation from the *ab initio* line.

NA-MD data	Gradient (10^{-1} ns)
Ab initio	1.11
Combined sublattices	1.04
Partial iodine sublattice	1.02
Full iodine sublattice	1.63
Lead sublattice	0.69
Cesium sublattice	1.79

good agreement with the *ab initio* NA-MD simulations. The inclusion of descriptors based on the other sublattices only provides a marginal benefit and is unnecessary for a reliable NA-MD simulation. The latter observation allows a significant reduction in the complexity of the ML model for the NA Hamiltonian. In particular, since we use every third iodine atom of $CsPbI_3$, and hence, every fifth atom overall, we reduce the number of descriptors by a factor of 5, compared to traditional models that employ descriptors arising from all atoms. Such a reduction in the model complexity may play particularly important roles for more complex systems involving large numbers of atoms and in ML models based on neural networks, which involve highly nonlinear searches for optimal model parameters.

IV. CONCLUSION

Focusing on nonradiative charge recombination in a popular solar cell and optoelectronic material, $CsPbI_3$, we applied unsupervised and supervised ML to analyze the NA-MD Hamiltonian, determine which geometric descriptors are most suitable for building ML models of the NA-MD Hamiltonian, reduce the complexity of the standard ML models, and test the NA-MD performance of the reduced models against the *ab initio* NA-MD results. We demonstrated that descriptors extracted from every third atom of the iodine sublattice are sufficient for the prediction of the bandgap and NAC values that lead to satisfactory NA-MD results. Additionally, we have uncovered an unusual trend for the performance of the individual elemental sublattices in $CsPbI_3$, with the lead sublattice performing extremely poorly, especially for the prediction of the bandgap and NAC, despite being involved in the *ab initio* calculation of these values. This has been rationalized by the significance of the iodine octahedral structure in determining the bandgap and NAC values. Cesium interacts with iodines from different octahedra, and therefore cesium descriptors can reflect octahedral tilting angles. In comparison, lead interacts with iodines from the same octahedron, and therefore the properties of the lead sublattice do not reflect the octahedral tilting. At the same time, the ML model based on the lead sublattice gives better NA-MD simulation results than the model based on the cesium sublattice even though the quality of the NA-MD Hamiltonian shows the opposite trend. The original ML model of the NA-MD Hamiltonian was simplified both quantitatively by reducing the number of descriptors five-fold and qualitatively by using one instead of three types of atoms. The significant simplification of the ML model helps to overcome the high computational cost of NA-MD simulations through ML and increase the applicability of NA-MD simulations to large complex systems and longer time-scales.

ACKNOWLEDGMENTS

This work was supported by the U.S. National Science Foundation under Grant No. CHE-1900510.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. V. Akimov and O. V. Prezhdo, J. Chem. Theory Comput. **10**, 789 (2014).

[2] W. Somogyi, S. Yurchenko, and A. Yachmenev, J. Chem. Phys. **155**, 214303 (2021).

[3] Y. She, Z. Hou, O. V. Prezhdo, and W. Li, J. Phys. Chem. Lett. **12**, 10581 (2021).

[4] F. Tran, J. Doumont, P. Blaha, M. A. L. Marques, S. Botti, and A. P. Bartók, J. Chem. Phys. **151**, 161102 (2019).

[5] R. Long and N. J. English, Chem. Mater. **22**, 1616 (2010).

[6] T. A. Barckholtz and T. A. Miller, J. Phys. Chem. A **103**, 2321 (1999).

[7] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Annu. Rev. Phys. Chem. **71**, 361 (2020).

[8] T. A. Profitt and J. K. Pearson, Phys. Chem. Chem. Phys. **21**, 26175 (2019).

[9] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C **121**, 511 (2017).

[10] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, and A. Tkatchenko, Chem. Rev. **121**, 9816 (2021).

[11] M. Ceriotti, C. Clementi, and O. Anatole von Lilienfeld, J. Chem. Phys. **154**, 160401 (2021).

[12] E. Cuierrier, P.-O. Roy, and M. Ernzerhof, J. Chem. Phys. **155**, 174121 (2021).

[13] C.-I. Wang, I. Joanito, C.-F. Lan, and C.-P. Hsu, J. Chem. Phys. **153**, 214113 (2020).

[14] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[15] M. Liu and J. R. Kitchin, J. Phys. Chem. C **124**, 17811 (2020).

[16] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Comput. Phys. Commun. **247**, 106949 (2020).

[17] J. R. Moreno, J. Flick, and A. Georges, Phys. Rev. Mater. **5**, 083802 (2021).

[18] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).

[19] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, Nat. Commun. **10**, 5024 (2019).

[20] P. O. Dral, M. Barbatti, and W. Thiel, J. Phys. Chem. Lett. **9**, 5660 (2018).

[21] S. M. Mangan, G. Zhou, W. Chu, and O. V. Prezhdo, J. Phys. Chem. Lett. **12**, 8672 (2021).

[22] J. Westermayr and P. Marquetand, Chem. Rev. **121**, 9873 (2021).

[23] J. Westermayr, M. Gastegger, M. F. S. J. Menger, S. Mai, L. González, and P. Marquetand, Chem. Sci. **10**, 8100 (2019).

[24] J. Westermayr, F. A. Faber, A. S. Christensen, O. A. von Lilienfeld, and P. Marquetand, Mach. Learn.: Sci. Technol. **1**, 025009 (2020).

[25] R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, J. Chem. Phys. **143**, 084111 (2015).

[26] J.-K. Ha, K. Kim, and S. K. Min, J. Chem. Theory Comput. **17**, 694 (2021).

[27] W.-K. Chen, W.-H. Fang, and G. Cui, Phys. Chem. Chem. Phys. **21**, 22695 (2019).

[28] C.-K. Lee, C. Lu, Y. Yu, Q. Sun, C.-Y. Hsieh, S. Zhang, Q. Liu, and L. Shi, J. Chem. Phys. **154**, 024906 (2021).

[29] Z. Zhang, Y. Zhang, J. Wang, J. Xu, and R. Long, J. Phys. Chem. Lett. **12**, 835 (2021).

[30] P. O. Dral and M. Barbatti, Nat. Rev. Chem. **5**, 388 (2021).

[31] K. Lin, J. Peng, F. L. Gu, and Z. Lan, J. Phys. Chem. Lett. **12**, 10225 (2021).

[32] X. Wang and R. Long, J. Phys. Chem. Lett. **12**, 7553 (2021).

[33] L. Qiao, W.-H. Fang, R. Long, and O. V. Prezhdo, J. Am. Chem. Soc. **143**, 9982 (2021).

[34] R. Shi, W.-H. Fang, A. S. Vasenko, R. Long, and O. V. Prezhdo, Nano Res. (published online 2021).

[35] S. Mukherjee and S. A. Varganov, J. Chem. Phys. **155**, 174107 (2021).

[36] D. Zanuttini, J. Douady, E. Jacquet, E. Giglio, and B. Gervais, J. Chem. Phys. **134**, 044308 (2011).

[37] T. W. Kim, S. Jun, Y. Ha, R. K. Yadav, A. Kumar, C.-Y. Yoo, I. Oh, H.-K. Lim, J. W. Shin, R. Ryoo, H. Kim, J. Kim, J.-O. Baeg, and H. Ihee, Nat. Commun. **10**, 1873 (2019).
[38] A. D. Wright, L. R. V. Buizza, K. J. Savill, G. Longo, H. J. Snaith, M. B. Johnston, and L. M. Herz, J. Phys. Chem. Lett. **12**, 3352 (2021).
[39] A. Stolow, A. E. Bragg, and D. M. Neumark, Chem. Rev. **104**, 1719 (2004).
[40] S. Banerjee, J. Kang, X. Zhang, and L.-W. Wang, J. Chem. Phys. **152**, 091102 (2020).
[41] R. Long, O. V. Prezhdo, and W. Fang, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **7**, e1305 (2017).
[42] W. Stier and O. V. Prezhdo, J. Phys. Chem. B **106**, 8047 (2002).
[43] V. Zobač, J. P. Lewis, and P. Jelínek, Nanotechnology **27**, 285202 (2016).
[44] B. Wang, W. Chu, A. Tkatchenko, and O. V. Prezhdo, J. Phys. Chem. Lett. **12**, 6070 (2021).
[45] J. Westermayr, M. Gastegger, and P. Marquetand, J. Phys. Chem. Lett. **11**, 3828 (2020).
[46] D. Hu, Y. Xie, X. Li, L. Li, and Z. Lan, J. Phys. Chem. Lett. **9**, 2725 (2018).
[47] W.-K. Chen, X.-Y. Liu, W.-H. Fang, P. O. Dral, and G. Cui, J. Phys. Chem. Lett. **9**, 6702 (2018).
[48] G. W. Richings and S. Habershon, J. Phys. Chem. A **124**, 9299 (2020).
[49] M. F. S. J. Menger, J. Ehrmaier, and S. Faraji, J. Chem. Theory Comput. **16**, 7681 (2020).
[50] E. Posenitskiy, F. Spiegelman, and D. Lemoine, Mach. Learn.: Sci. Technol. **2**, 035039 (2021).
[51] A. Farahvash, C.-K. Lee, Q. Sun, L. Shi, and A. P. Willard, J. Chem. Phys. **153**, 074111 (2020).
[52] G. Zhou, W. Chu, and O. V. Prezhdo, ACS Energy Lett. **5**, 1930 (2020).
[53] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, Chem. Rev. **121**, 9722 (2021).
[54] P. Tavadze, G. Avendaño Franco, P. Ren, X. Wen, Y. Li, and J. P. Lewis, J. Am. Chem. Soc. **140**, 285 (2018).
[55] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).
[56] B. C. Ross, PLoS One **9**, e87357 (2014).
[57] S. D. Stranks and H. J. Snaith, Nat. Nanotechnol. **10**, 391 (2015).
[58] N.-G. Park, J. Phys. Chem. Lett. **4**, 2423 (2013).
[59] M. Grätzel, Nat. Mater. **13**, 838 (2014).
[60] M. A. Green, A. Ho-Baillie, and H. J. Snaith, Nat. Photonics **8**, 506 (2014).
[61] T. C. Sum, M. Righetto, and S. S. Lim, J. Chem. Phys. **152**, 130901 (2020).
[62] V. N. Tuoc and T. D. Huan, J. Chem. Phys. **152**, 014104 (2020).
[63] A. S. Kshirsagar and A. Nag, J. Chem. Phys. **151**, 161101 (2019).
[64] J. K. Yamamoto and A. S. Bhalla, J. Appl. Phys. **70**, 4469 (1991).
[65] T. Matsushima, M. R. Leyden, T. Fujihara, C. Qin, A. S. D. Sandanayaka, and C. Adachi, Appl. Phys. Lett. **115**, 120601 (2019).
[66] M. Wang, Z. Ni, X. Xiao, Y. Zhou, and J. Huang, Chem. Phys. Rev. **2**, 031302 (2021).
[67] L. Xu, R. Molaei Imenabadi, W. G. Vandenberghe, and J. W. P. Hsu, APL Mater. **6**, 036104 (2018).
[68] J. Li, B. Pradhan, S. Gaur, and J. Thomas, Adv. Energy Mater. **9**, 1901891 (2019).
[69] M. Kar and T. Körzdörfer, J. Chem. Phys. **149**, 214701 (2018).
[70] X.-G. Li, B. Blaiszik, M. E. Schwarting, R. Jacobs, A. Scourtas, K. J. Schmidt, P. M. Voyles, and D. Morgan, J. Chem. Phys. **155**, 154702 (2021).
[71] J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, J. Chem. Phys. **154**, 230903 (2021).
[72] A. V. Akimov and O. V. Prezhdo, J. Chem. Theory Comput. **9**, 4959 (2013).
[73] H. M. Jaeger, S. Fischer, and O. V. Prezhdo, J. Chem. Phys. **137**, 22A545 (2012).
[74] A. V. Akimov and O. V. Prezhdo, J. Phys. Chem. Lett. **4**, 3857 (2013).
[75] B. F. Habenicht, H. Kamisaka, K. Yamashita, and O. V. Prezhdo, Nano Lett. **7**, 3260 (2007).
[76] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
[77] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[78] G. Kresse and J. Hafner, Phys. Rev. B **49**, 14251 (1994).
[79] W. Chu, Q. Zheng, A. V. Akimov, J. Zhao, W. A. Saidi, and O. V. Prezhdo, J. Phys. Chem. Lett. **11**, 10073 (2020).
[80] W. Chu and O. V. Prezhdo, J. Phys. Chem. Lett. **12**, 3082 (2021).
[81] W. Chu, W. A. Saidi, and O. V. Prezhdo, ACS Nano **14**, 10608 (2020).
[82] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).
[83] L. F. Kozachenko and N. N. Leonenko, Probl. Inf. Transm. **23**, 95 (1987).
[84] 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 É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).
[85] F. Wang, Y. Fu, M. E. Ziffer, Y. Dai, S. F. Maehrlein, and X.-Y. Zhu, J. Am. Chem. Soc. **143**, 5 (2021).
[86] P. Kambhampati, J. Phys. Chem. C **125**, 23571 (2021).
[87] W. Chu, W. A. Saidi, J. Zhao, and O. V. Prezhdo, Angew. Chem., Int. Ed. **59**, 6435 (2020).
[88] W. Li, A. S. Vasenko, J. Tang, and O. V. Prezhdo, J. Phys. Chem. Lett. **10**, 6219 (2019).