

Gaussian approximation potential for amorphous Si : H

Davis Unruh^{1,*}, Reza Vatan Meidanshahi², Stephen M. Goodnick², Gábor Csányi³, and Gergely T. Zimányi¹¹Physics Department, University of California, Davis, California 95616, United States²School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287-5706, USA³Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom

(Received 5 January 2022; revised 15 May 2022; accepted 1 June 2022; published 27 June 2022)

Hydrogenation of amorphous silicon (a -Si : H) is critical for reducing defect densities, passivating midgap states and surfaces, and improving photoconductivity in silicon-based electro-optical devices. Modeling the atomic-scale structure of this material is critical to understanding these processes, which in turn is needed to describe c -Si/ a -Si : H heterojunctions that are at the heart of modern solar cells with world-record efficiency. Density functional theory (DFT) studies achieve the required high accuracy but are limited to moderate system sizes of 100 atoms or so by their high computational cost. Simulations of amorphous materials have been hindered by this high cost because large structural models are required to capture the medium-range order that is characteristic of such materials. Empirical potential models are much faster, but their accuracy is not sufficient to correctly describe the frustrated local structure. Data-driven, machine-learned interatomic potentials have broken this impasse and have been highly successful in describing a variety of amorphous materials in their elemental phase. Here, we extend the Gaussian approximation potential (GAP) for silicon by incorporating the interaction with hydrogen, thereby significantly improving the degree of realism with which amorphous silicon can be modeled. We show that our Si : H GAP enables the simulation of hydrogenated silicon with an accuracy very close to DFT but with computational expense and run times reduced by several orders of magnitude for large structures. We demonstrate the capabilities of the Si : H GAP by creating models of hydrogenated liquid and amorphous silicon and showing that their energies, forces, and stresses are in excellent agreement with DFT results, and their structure as captured by bond and angle distributions are in agreement with both DFT and experiments.

DOI: [10.1103/PhysRevMaterials.6.065603](https://doi.org/10.1103/PhysRevMaterials.6.065603)

I. INTRODUCTION

Hydrogenated amorphous silicon (a -Si : H) is a widely used material, with applications ranging from thin-film transistors [1] to solar cells [2,3]. Created using plasma-enhanced chemical vapor deposition (PECVD) of silane gas (SiH_4) [4], a -Si : H has far fewer defects than pure amorphous silicon due to the passivating role of the H atoms, the concentration of which typically ranges from 6 to 18 at. %. As a result, a -Si : H has demonstrably superior electrical transport properties and photoconductivity [5,6] in comparison with pure amorphous silicon. Therefore, the role of H in a -Si : H films is crucial for obtaining effective devices.

Remarkably, several aspects of a -Si : H remain the focus of research: understanding the local atomic interactions, the classes of defects and their formation and statistics, and the long-term structural dynamics all pose questions that are unsettled. Of particular interest is the light-induced defect formation, known as the Staebler-Wronski effect [7], which is directly relevant for photovoltaic applications. Another important issue is the degradation of hydrogen passivation at the amorphous/crystalline interface in silicon heterojunctions (HJs) [3,8,9], which is probably a key driver of the enhanced

degradation [10] of the efficiency world record holder: Si HJ solar cells [11]. This enhanced degradation is one of the main factors that slows down the market acceptance of these extremely promising Si HJ solar cells.

All of the above challenges demonstrate the need for theoretical and simulation-based research efforts to provide crucial guidance and insight. Atomistic simulation methods are the primary tools for materials modeling. The most accurate of these methods are those based on first principles, in this case, electronic-structure theory. The most prominent method is density functional theory (DFT) [12], which is highly accurate but at the cost of significant computational expense and unfavorable scaling behavior (typically cubic, although linear scaling versions exist with large prefactors [13]). With rare and heroic exceptions, DFT techniques are typically used to simulate systems with no more than a few hundred atoms. Larger-scale simulations of many hundreds or a few thousand atoms or disordered systems that require simulating many thousands of samples are typically undertaken using parameterized empirical interatomic potentials within molecular dynamics (MD) methods [14]. These potentials typically have simple functional forms and are usually optimized to reproduce a few key material characteristics on a limited set of structures with good accuracy. These functional forms are parameterized to strike a balance between obtaining good accuracy for the selected observables for the selected

*dgunruh@ucdavis.edu

structures and transferability to other observables and/or other structures for which the potential was not fitted to. The limited number of parameters of these models makes it very hard to satisfy both of these goals.

It is also important to note that an alternative exists in the form of tight-binding (TB) models, which represent a middle ground between DFT and interatomic potentials. The TB approach enables a minimal description of the electronic structure and is capable of using inputs from DFT, yielding a qualitatively robust description of the electron behavior in a wide range of solids with a computational cost that is significantly reduced compared with DFT [15]. These models scale similarly to interatomic potentials [16,17], and substantial effort has gone into developing accurate TB models for a wide range of materials [18–20]. However, their utility remains primarily limited to materials where the inclusion of charge transfer is of the utmost priority [21,22]. In other cases, it is not clear that they are more accurate or transferable than classical interatomic potentials, despite being significantly more expensive [23].

Several interatomic potentials have been developed for MD simulations of pure silicon phases and achieved reasonable accuracy [24–26]. However, as discussed further below, even for Si, there was clearly room for improving the correspondence with experimental and DFT results. There have been notably fewer papers on Si-H interatomic potentials, even though these are essential for accurate modeling of solar cells. Reference [27] reported limited correlation between *a*-Si : H MD simulations using the Stillinger-Weber interatomic potential for the Si-H pair-correlation function with experiments. Reference [28] reported moderate correlation between *a*-Si : H MD simulations of the Si-H pair correlation function using the Tersoff potential with experiments. Reference [29] reported discrepancies between *a*-Si : H MD simulations using various MD potentials and experiments regarding the radial distribution function (RDF) and vibrational frequencies. Reference [30] reported that the hydrogen pairing tendencies observed in Ref. [31] were not adequately captured by MD simulation. Finally, Ref. [32] reported that MD simulations of reaction processes were inadequate with existing MD potentials and went on to suggest improvements.

In the last decade, the adoption of machine-learning (ML) methods [23,33–45] has changed the above picture, particularly for the simulation of hard materials. These ML-based methods construct highly flexible nonparametric interatomic potentials by training them on DFT-computed energies and gradients of a wide variety of structures and can deliver near-DFT-level accuracy for similar structures at a computational cost that is thousands of times less than that of DFT for moderate sizes systems. Since interatomic potentials scale linearly with system size, the savings can reach factors of millions for large structures or when using expensive hybrid correlation functionals. Some of these ML models additionally include descriptions of charge transfer, enabling more accurate treatment of systems with strong nonlocal dependencies in the electronic structure [46–48]. For recent reviews of all of these methods and their application to materials simulation, see Refs. [49,50].

The advantage of ML models in general and ML-based potentials in particular is the flexibility of their functional

form, which enables them to accurately interpolate the potential energy of structures in a broad training database, covering a wide variety of structures and phases. However, the nature of this flexibility also leads ML models to be increasingly less accurate for structures that are farther away from the training dataset. Thus, the disadvantage of ML-based potentials is their limited transferability or extrapolation. The key to success is to transform the problem of fitting the total energy of large systems into smaller subtasks for which interpolation suffices: for extended materials, this is achieved by the ansatz that the total energy is a sum of local terms, *site or atomic energies*, which are only functions of the local environment. Creating a truly general-purpose potential thus requires a very large training database that covers all relevant *local environments* well. When the potential is used, it will be accurate if the configurations encountered consist of local environments that are near those in the database. If radically new local environments are encountered (e.g., because of phase unanticipated transitions or the system is taken to a pressure or temperature range far outside where the training structures came from), accuracy will be severely compromised. Note that the notion of *nearness* used above in practice is defined by the architecture of the ML model and the types of regularization that are used in obtaining the optimal parameters.

In this paper, we use the Gaussian approximation potential (GAP) framework [33,34]. The GAP is based on the Gaussian process regression (also known as kernel regression) methodology, typically using the smooth overlap of atomic positions (SOAP) kernel to describe the similarity between local environments in materials [34]. One advantage of kernel regression is that the flexibility of the functional form naturally grows with the amount of training data. The general approach, which has recently been reviewed in detail [50], is to fit a potential to an initial training set of total energies, forces, and stresses calculated using a given electronic structure method (typically DFT) and then use this potential to explore a wide range of structures using a variety of algorithms including MD and geometry optimization. The training dataset is then iteratively broadened, particularly with configurations where the fit is not yet good enough, until the desired accuracy is reached for the configurations and observables of interest. The ultimate validation of the potential is by testing macroscopic observables using samples generated by the potential and not explicitly used in the training.

A GAP model has already been developed to describe pure Si [23], including many crystalline and amorphous phases, and has been used to study the structure of *a*-Si [51] and its phase transitions under pressure in unprecedented detail and precision [52]. However, *a*-Si/*c*-Si HJ solar cells contain a significant amount of hydrogen, most often in the 10–15% range. It is well known that existing interatomic potentials of Si : H are unable to capture experimentally measurable quantities with sufficient precision, as discussed earlier [27–32], including partial RDFs, pairing tendencies, and reaction processes. Therefore, the effort to describe, analyze, and then mitigate the problematic excess degradation of these HJ cells requires the development of a GAP model for Si : H, which we report on in this paper. We validate our Si : H GAP on hydrogenated liquid Si and amorphous Si, which are the two most relevant phases since *a*-Si : H is typically modeled using

melt-quench simulations of liquid Si : H. To gauge the utility of our Si : H GAP, we show how it qualitatively outperforms state-of-the-art parameterized interatomic potentials by delivering excellent agreement with DFT calculations and experimental results. This Si : H GAP will be key for future research by qualitatively enhancing the precision of MD simulations of hydrogenated silicon solids to DFT levels, enabling a wide variety of electro-optical simulations, including simulations of *a*-Si : H/*c*-Si HJ solar cells and degradation simulations which may eventually clarify long-term structural dynamics problems such as the Staebler-Wronski effect.

II. THE GAP PLATFORM

In this paper, we train a GAP on a set of structures and accompanying energies, forces, and stresses, all computed using DFT. The GAP is constructed as the sum of a purely repulsive core potential and a SOAP kernel. The repulsive potential is built with cubic splines that are fitted to the interaction of pairs of atoms in vacuum as computed by DFT. These pair potentials are built for all three relevant atomic pairs: Si-Si, Si-H, and H-H. Including these repulsive core pair potentials serves a dual purpose. First, a large fraction of the interaction energy between atoms can in fact be described by a pair potential, which describes exchange repulsion at close atomic distances and some effects of chemical bonding at far distances. Second, the repulsive portion of the potential is difficult to capture using the same kernel function that is appropriate to describe valence bonding due to the steepness of the energy curve at close approach, and thus, capturing it with a pair potential with a piece-wise spline enhances overall numerical efficiency significantly.

The total GAP energy for the system is generated as a sum of the repulsive pair potential and the many-body SOAP kernel which is constructed as a linear sum over kernel basis functions:

$$E = \sum_{i < j} V^{(2)}(r_{ij}) + \sum_i \sum_s^M \alpha_s K(R_i, R_s), \quad (1)$$

where i and j index the atoms in the system, $V^{(2)}$ is the two-body repulsive pair potential, r_{ij} is the distance between atoms i and j , K is the SOAP kernel basis function, and R_i is the collection of relative position vectors corresponding to the neighbors of atom i which is termed a *neighborhood*. The s sum ranges over the set of M representative atoms, selected from the input dataset, whose neighborhoods are chosen to serve as a basis in which the potential is expanded. The coefficients α_s are determined by a regularized linear regression of the energies, forces, and stresses of the system computed with Eq. (1) that are parameterized by these α_s values and compared with the corresponding energies, forces, and stresses computed by DFT for the same structures. This comparison results in a set of linear equations that are solved for the coefficients α_s using Tikhonov regularization. The representative neighborhoods over which Eq. (1) is evaluated are selected by choosing basis neighborhoods which are maximally dissimilar from each other (using CUR matrix decomposition of the matrix corresponding to all the neighborhoods in the training dataset), such that the variety of the entire set of possible

neighborhoods can be well represented by interpolation over a small number of these basis neighborhoods. The details of all these algorithms and their general application to fitting a GAP model have recently been reviewed in Ref. [50].

The SOAP kernel $K(R_i, R_j)$ characterizes the similarity between two neighborhoods: it is maximal when the two neighborhoods are identical, and it is smaller and smaller for more and more different neighborhoods. To quantify the similarity of two neighborhoods, each neighborhood R_i of atom i is represented by a set of corresponding neighbor densities, a separate one built for each kind of element in the neighborhood:

$$\rho_i^z(\mathbf{r}) = \sum_{i'} \delta_{zz'} f_{\text{cut}}(r_{i'i'}) \exp\left(-\frac{\mathbf{r} - \mathbf{r}_{i'i'}}{2\sigma_{\text{atom}}^2}\right), \quad (2)$$

where z corresponds to the element whose density around atom i is constructed, and the $\delta_{zz'}$ factor selects neighbors corresponding to element z to be included in the sum. In the present case, there are two elements (Si and H), so two neighbor densities are built around each atom. The calculation of the SOAP kernel starts by integrating the overlap of the two neighbor densities:

$$\tilde{K}(R_i, R_j) = \sum_{zz'} \int_{\hat{R} \in \text{SO}(3)} d\hat{R} \left| \int d\mathbf{r} \rho_i^z(\mathbf{r}) \rho_j^{z'}(\hat{R}\mathbf{r}) \right|^2. \quad (3)$$

Next, the integral \tilde{K} is normalized by the self-overlaps. Finally, it is raised to an integer power and has a suitable hyperparameter prefactor assigned:

$$K(R_i, R_j) = \delta^2 \left| \frac{\tilde{K}(R_i, R_j)}{\sqrt{\tilde{K}(R_i, R_i) \tilde{K}(R_j, R_j)}} \right|^\xi. \quad (4)$$

In practice, the neighbor densities are represented by an expansion in a basis of spherical harmonics and radial functions, and the integral in Eq. (3) is efficiently evaluated as a scalar product. A detailed review of SOAP and similar neighbor density based representations, descriptors, and kernels is in Ref. [53]. The application to pure Si was comprehensively described in Ref. [23], which we refer to for further details and validation.

III. Parameterizing THE SI : H GAP

A. Outline of training and reference structures

ML models in general and GAP models in particular are trained using three sets of structures. First, the GAP is trained by fitting to DFT-calculated energies, forces, and stresses on the structures of an adaptively created training set. This training set is iteratively expanded by creating new structures with simple MD simulations (such as a structural relaxation or a brief thermal protocol) that use the current version of the GAP model, starting from structures already contained within the training set. We then carry out DFT and GAP tests on the new structures and determine the differences between them. Structures for which the comparative fit is poor are added to the next-generation training set, and the Si-H GAP model is retrained.

Second, the results of the training procedure are regularly evaluated using a reference set. The GAP is not trained on this reference set of structures. Instead, the energies, forces, and stresses are calculated with both the GAP and with DFT on the structures within the validation set, and the difference is used to quantify and monitor the progress of the training.

Third, the model undergoes final testing on an additional validation set of structures. This set of data is not used to inform the progress of the training but is instead utilized as a final benchmark for performance, augmenting the reference set.

The structures of the training set and the reference set were assembled by adding H to pure Si structures of representative Si phases and structures. The initial training set included about 150 structures. The reference set included about 110 structures. These pure Si structures were taken from our previous work on *a*-Si/*c*-Si interface degradation [9], from the reference database on which the published Si-only GAP was trained [23], or were generated using the atomic simulation environment [54]. The representative phases of Si were (1) amorphous silicon, (2) liquid silicon, (3) diamond silicon with a vacancy, (4) diamond silicon with a divacancy, (5) diamond silicon with an interstitial Si, (6) amorphous/crystalline silicon interface structures, and (7) diamond surface structures with (100) and (111) orientation. The atomic concentration of added H was in the 6–12 at. % range for the liquid and amorphous phases and for the interface structures, in the 4–8 at. % range for the bulk diamond phases, and in the 12–20 at. % range for the diamond surfaces. In each of these structures, a sufficient amount of H was added to fully passivate all dangling or highly strained bonds. Furthermore, to train the Si-H GAP to accurately model hydrogen-related defects, structures with an H atom added or taken away from the fully passivated structures were also added both to the training set and to the reference set.

The additional structures of the validation set were assembled by using the new Si-H GAP to perform MD simulations (see Sec. III E below for details of the MD procedures). The validation set consisted of 150 additional hydrogenated structures. The representative phases were (1) amorphous silicon, (2) liquid silicon, and (3) amorphous/crystalline silicon interface structures.

Finally, it is noted that the training set also contains an isolated Si atom and an isolated H atom as structures. Isolated here means the structure possesses a large enough unit cell that the atom is effectively in isolation, even though its energy is computed using periodic boundary conditions and the same DFT parameter settings for consistency. Including these isolated atoms as training structures is essential because the GAP is fit to the binding energy, not the total energy (i.e., the binding energy plus the energy of isolated atoms).

B. DFT calculation details

The DFT calculations were performed using the QUANTUM ESPRESSO 6.2.1 software package [55–57], with the key parameters as follows. The Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional was used with periodic boundary conditions [58]. The core and valence electron interactions were described by the norm-conserving

TABLE I. Si and H SOAP kernel parameter values.

Element	n_{\max}	l_{\max}	δ	ζ	r_{cut}	w	σ_{atom}
Si	10	6	3	4	5.0	1.0	0.5
H	9	6	1	4	3.5	0.5	0.4

pseudopotential function. The calculations were performed with Marzari-Vanderbilt electronic smearing [59]. This smearing method was chosen, as it ensures that the DFT energies and forces are consistent, and the electronic free energy was used as the energy training target because its derivatives with respect to the atomic positions are reported as *forces* by the DFT code (according to the Hellman-Feynman theorem). An energy cutoff of 42 Ry was employed for the plane-wave basis set, and a Monkhorst-Pack grid method was used to define the k-point mesh which samples the Brillouin zone. The k-point spacing was chosen to be 0.2 \AA^{-1} .

We also made sure that the energy/atom values calculated for each structure were accurate within 1 meV/atom. This was tested by verifying that the energy/atom values did not change $> \pm 1$ meV/atom when the cutoff was pushed to exceptionally large values or the k-point density was substantially increased.

A separate set of parameters was used to perform DFT Born-Oppenheimer MD (DFT-BOMD) simulations [60]. DFT-BOMD was employed to compare the performance of GAP-driven MD simulations vs DFT-driven MD simulations. The DFT-BOMD simulations utilized the PWSCF module of the QUANTUM ESPRESSO software, using the same PBE exchange-correlation functional as before. An energy cutoff of 36 Ry was used for the plane-wave basis set, and the first Brillouin zone was sampled using only the Γ point. A Gaussian smearing width of 0.01 Ry was implemented to the density of states to avoid convergence problems with metallic configurations.

C. Fitting the Si-H GAP

Several hyperparameters were needed to define the SOAP kernels centered on Si and H atoms. These included (1) n_{\max} and l_{\max} , the maximum number of radial and angular indices for the spherical harmonic expansion of the neighbor densities; (2) δ , a hyperparameter that set the energy scale of the many-body term in the SOAP kernel; (3) ζ , the exponent used to construct the SOAP kernel; (4) the cutoff radius r_{cut} that characterized the radius beyond which the cutoff function within the neighbor densities converges to zero, and its associated transition width w that set the rate of this convergence; and (5) σ_{atom} , the smearing parameter for the neighbor density function. Table I summarizes the values of these parameters.

The fitting also required choosing regularization parameters for each kind of target data: σ_{energy} for energies, σ_{force} for force components, and σ_{virial} for virial stress components. These regularization parameters can be set at different values for different parts of the training set. These parameters represented the accuracy of the GAP to be reached with the fitting procedure. The relative values of these same parameters for different parts of the training set also determined the relative weight each phase/structure has in the loss function. It is possible to input these values in the potential in a number of

TABLE II. σ_{energy} values for each structure type.

Phase	σ_{energy} (eV/atom)
Amorphous silicon	0.0015
Liquid silicon	0.003
<i>a</i> -Si/ <i>c</i> -Si interface	0.003
Diamond Si phases	0.001
Isolated atom	0.0001

different ways, including using the same value for all structures or as the same value for all structures in a single phase, or structure type, for instance, liquid Si. We ended up using a combination of these methods: we assigned the regularization values on a structure-by-structure basis. The σ_{energy} values were chosen to be the same for all structures in a given phase, see Table II. In some detail, σ_{force} was defined on a per-atom basis as

$$\sigma_{\text{force}} = \begin{cases} 0.1, & |F| < 2.0 \text{ eV \AA}^{-1} \\ 0.05|F|, & |F| \geq 2.0 \text{ eV \AA}^{-1}, \end{cases} \quad (5)$$

where $|F|$ was the magnitude of the force vector on that respective atom. Here, σ_{virial} was defined on a per-structure basis as

$$\sigma_{\text{virial}} = \begin{cases} 0.025N_{\text{atoms}}, & \max(|\tau_i|) < 1.0 \text{ eV} \\ 0.025N_{\text{atoms}} \max(|\tau_i|), & \max(|\tau_i|) \geq 1.0 \text{ eV}, \end{cases} \quad (6)$$

where N_{atoms} was the number of atoms in the structure, and $\max(|\tau_i|)$ was the maximal norm of the virial stress tensor.

D. Adaptive training of the Si-H GAP

As described in the Introduction, the conventional wisdom is that a ML model needs to be trained on a broad database of structures, containing as many structures as possible, to be usable in a general sense. Such an approach requires training on enormous training sets, and its product can end up overfitted with an excessive number of parameters. As an efficient alternative, we adopted an adaptive training procedure, wherein first the GAP was trained on the structures contained in the training set/database. Subsequently, the GAP was validated on structures outside the training set with DFT measurements of energies, forces, and virial stresses. The reference structures for which the quality of the fits fell below a threshold were then added to the training set, and the GAP was then re-trained. Such adaptive procedures grow the training database in an iterative manner. Using this procedure, it was not necessary to add thousands of structures indiscriminately to the training set, as the adaptive training procedure on its own preferentially gravitated only toward structures that needed to be included into the training set without any external input, thereby avoiding the problem of overfitting.

We performed all MD simulations using the LAMMPS software package built with QUIP package support [61–63] that can use GAP models. We evaluated the accuracy of the GAP compared with DFT with the error metric of the weighted root mean square error (RMSE). As the targeted accuracy of the GAP varied from structure to structure, set by their regularization parameters σ , the natural measure of accuracy is not

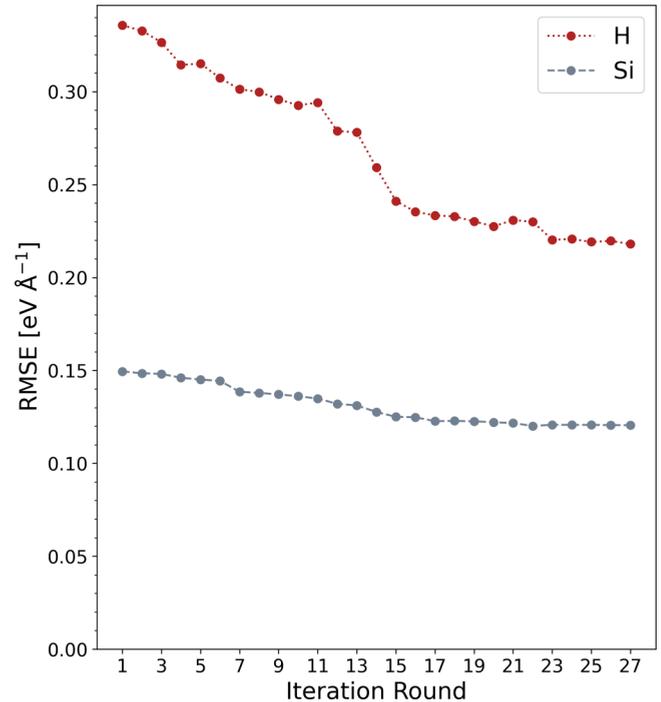


FIG. 1. Iterative progression of the average of the root mean square error (RMSE) of the Si (gray dashed) and H (red dotted) Cartesian force components, measured with respect to density functional theory (DFT) on the reference structure set.

the absolute RMSE value but its ratio to the regularization parameter σ . We captured this by weighting the RMSE by σ^2 as follows:

$$\text{RMSE}_{\text{weighted}} = \left[\frac{\sum_{i=1}^N \frac{(x_{i,\text{GAP}} - x_{i,\text{DFT}})^2}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}} \right]^{1/2}. \quad (7)$$

Here, N is the number of data points, x_{GAP} is the value measured by GAP, and x_{DFT} is the value measured by DFT, where x can be an energy, a force, or a virial stress.

We conducted 27 full rounds of iterative training of the Si-H GAP. Figure 1 shows the iteration-by-iteration progress of the average of the (unweighted) RMSE of the force components, measured by comparing the Si-H GAP forces with the DFT forces on the reference database, separately for Si and for H. In iterations 1–11 of the above-described adaptive procedure, we added new structures, shown in the right column of Table III, distinguished by the phase and the preparation method, such as heating or annealing, and doing so across different temperature ranges. Beginning with iteration 12, we started to add new structures to the adaptive training set that were distinguished by their different *a*-Si : H architecture or structure, such as *c*-Si/*a*-Si : H interfaces and different vacancy structures. A few sets of such structures from various phases were added to the training set. Once the new structures were added to the training set, the training of the GAP continued with the newly added structures. Table III lists the structures added to the training set over the 27 iterations. The results of this 27-iteration training are discussed next.

TABLE III. Structure types added in each iteration.

Iteration	Structure type
1	Optimized structures (all phases)
2	Optimized structures (all phases)
3	Low- T anneal of a -Si : H
4	High- T anneal of liq-Si : H
5	High- T anneal of liq-Si : H
6	Med- T anneal (1100 K) of a -Si : H
7	Heating a -Si : H from 500 to 800 K at 10^{13} K/s
8	Heating a -Si : H from 800 to 1100 K at 10^{13} K/s
9	Heating a -Si : H from 1100 to 1400 K at 10^{13} K/s
10	Heating a -Si : H from 1100 to 1400 K at 10^{13} K/s
11	Heating a -Si : H from 800 to 1400 K at 10^{12} K/s
12	Added new a -Si : H structures
13	Added new a -Si : H structures
14	Added c -Si/ a -Si : H interface structures
15	Added c -Si/ a -Si : H interface structures
16	Added new c -Si divacancy structures
17	Added new liq-Si : H structures
18	Added new c -Si vacancy structures
19	Added new c -Si interstitial structures
20	Low- T anneal of c -Si/ a -Si : H interface structures
21	Optimization of c -Si/ a -Si : H interface structures
22	NPT high- T anneal of liq-Si : H structures
23	NVT high- T anneal of liq-Si : H structures
24	Quenching liq-Si : H from 2000 to 1500 K at 10^{13} K/s
25	Annealing quenched liq-Si : H structures at 1500 K
26	Quenching liq-Si : H from 1500 to 1400 K at 10^{12} K/s
27	Added hydrogen passivated c -Si surface (100) and c -Si surface (111) structures

Figure 1 shows how the force RMSE on the Si and H atoms evolved over the 27 training iterations. Broadly speaking, every iteration successfully reduced the force RMSE. The force RMSE improved notably in iteration 7 because we changed the method of assigning the regularization parameters σ_{force} . In the first six iterations, the regularization parameters σ_{force} were assigned the same value for all structures in a given phase. From iteration 7, we switched to assigning the σ_{force} values for each structure individually.

In addition, in the same iteration we changed the fitting protocol for the regularization parameters σ_{force} and σ_{virial} to depend on the magnitude of the atomic force $|F|$, or stress, measured by DFT, as shown in Eq. (5). We changed the regularization because it was natural to relax the fitting for atoms, structures, and phases where the forces and stresses were larger. These two changes reduced the force RMSE and thus increased the accuracy of the GAP considerably, as visible in Fig. 1.

To form an overall picture about the efficiency of the training, we recall that the fitting of each quantity x was guided by the corresponding regularization parameter σ_x . As captured by the concept of the weighted RMSE, the GAP-simulated quantities were fitted to the DFT-simulated quantities only with a σ_x accuracy, or tolerance. As such, the σ_x values represent the

lowest error thresholds typically attainable with this regularization procedure. We recall that the regularization parameter σ_{force} was set to $\sigma_{\text{force,min}} = 0.100 \text{ eV \AA}^{-1}$ for weaker forces $F < F_{\text{th}} = 2.0 \text{ eV \AA}^{-1}$ and proportionally higher for stronger forces $F > F_{\text{th}}$. Visibly, the training successfully reduced the (unweighted) RMSE for the forces on the Si atoms from 50% above $\sigma_{\text{force,min}}$ to only 20% above $\sigma_{\text{force,min}}$. Moreover, since the forces F often exceeded the threshold value F_{th} , the effective $\sigma_{\text{force,eff}}$ was in fact higher than $\sigma_{\text{force,min}}$. Therefore, the training brought the accuracy of the RMSE even closer than 20% to the effective $\sigma_{\text{force,eff}}$. Thus, the training of the Si-H GAP substantially improved the precision of the calculated force values and made the Si-H GAP very close to reaching the theoretical limit of its precision.

The same training also reduced the RMSE for the forces on the H atoms from $\sim 0.34 \text{ eV \AA}^{-1}$ down to 0.22 eV \AA^{-1} . Just like for the Si atoms, we place this training into context. For the H atoms, the forces F exceeded F_{th} much more often than for Si atoms. The reason for the higher force RMSE values for H are likely due to locality, as explained and evidenced below.

We recall that a locality limit applies to the Si-H GAP because it does not include long-range interactions beyond 5 Å for Si and 3.5 Å for H. For a discussion of this aspect, see Ref. [64]. This locality limit provides a natural procedure to determine an accuracy threshold the Si-H GAP forces can possibly achieve [33,64]. We can quantify this limit by a locality test procedure: a central atom is selected in a simulation cell, and a sphere is defined around it with a radius r_{fix} . The atoms inside this sphere are kept fixed. The atoms outside this sphere are assigned a velocity corresponding to a high temperature of, e.g., $T = 2000 \text{ K}$, and evolved with MD using the Si-H GAP for 1 ps. Representative snapshot structures are gathered over the course of this evolution. The force acting on the central atom in these snapshot structures is calculated using DFT. The standard deviation of the distribution of the forces on the central atom over all snapshots is a natural measure of a threshold accuracy attainable by any interatomic potential with the given finite cutoff since it captures how much the DFT-calculated forces can vary due to configuration changes outside the cutoff. Note that this test is entirely independent of the interatomic potential, and the locality that it measures is a quantum mechanical property of the system.

We performed this locality test on the H atoms in the most disordered phases: the amorphous and liquid structures. We found that, for the appropriate cutoff radius of $r_{\text{fix}} = 3.5 \text{ \AA}$, the above-defined force standard deviation was 0.20 eV \AA^{-1} . Adopting this result of the locality test as the attainable value of the force RMSE, we conclude that the 27-round training of the Si-H GAP reduced the force RMSE of the H atoms from 70% above the locality limit to only 10% above.

Next, Fig. 2 shows the correlation between the energies/atom computed by the Si-H GAP and by DFT, as measured on the reference and validation sets after 27 iterations. Our adaptive method reached an RMSE of the reference correlation curve to be as little as 4 meV/atom, close to the precision of our DFT calculations, 1 meV/atom. The RMSE of the validation correlation curve, which only contains higher excess energy liquid, amorphous, and interface structures and covers slightly different regions of the potential energy surface, was as little as 6 meV/atom,

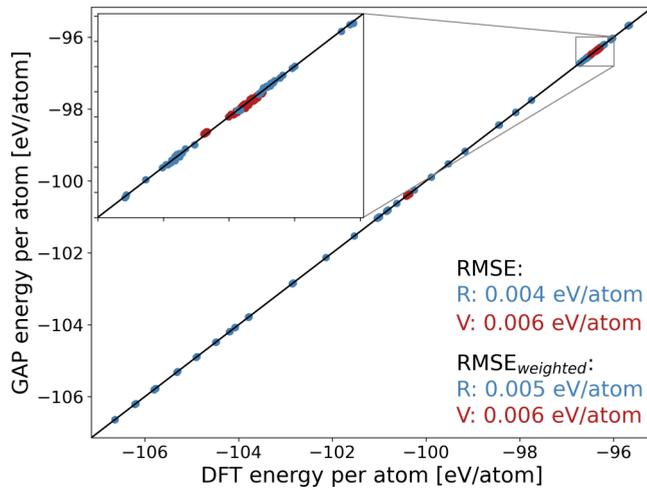


FIG. 2. Correlation between Gaussian approximation potential (GAP) and density functional theory (DFT) measurements of the energies/atom on both the reference (blue) and validation (red) sets of structures, after iteration 27.

still close to DFT precision. To account for the difference in excess energies of the structures in the two datasets, a weighted RMSE was also included. This weighted RMSE used the average excess energy of each type of configuration as inputs, computed across both datasets, assigning the weights [the σ values in Eq. (7)] to be the inverse average excess energy multiplied by the highest of these averages. This weighted RMSE is 0.005 meV/atom for the reference set and 0.006 meV/atom for the validation set, indicating the model is performing as expected.

E. Analysis and validation of the Si-H GAP

We will now analyze the usability of our GAP model by conducting MD simulations with it, calculating experimentally observable quantities, and comparing the results to those obtained by using DFT and by experiments. While our broad goal was to create a universally useful Si-H GAP, our primary motivation and intended use for this Si-H GAP was to create highly accurate a -Si : H structures via a melt-quench procedure from the liquid Si : H and to subsequently use it to measure defect creation/annihilation energy barriers via the nudged elastic band method. Therefore, we will focus our analysis of the Si-H GAP on the liquid phase liq-Si : H, on the amorphous phase a -Si : H, and on some representative defect structures.

There are several different metrics for evaluating the realism of the resulting structures that can be experimentally observed and tested. These include (1) the RDF for mono-atomic systems or the partial pair correlation function for multi-atomic systems; (2) the bond length distributions and the bond angle distribution; (3) the excess energy/atom, measured relative to c -Si; and (4) the vibrational spectra. The relative importance of these experimental metrics is the subject of ongoing debate, with different proponents arguing in favor of either the RDF or the vibrational spectra [65]. It is significantly easier to measure the RDF computationally, but its experimental determination is more difficult, as it requires

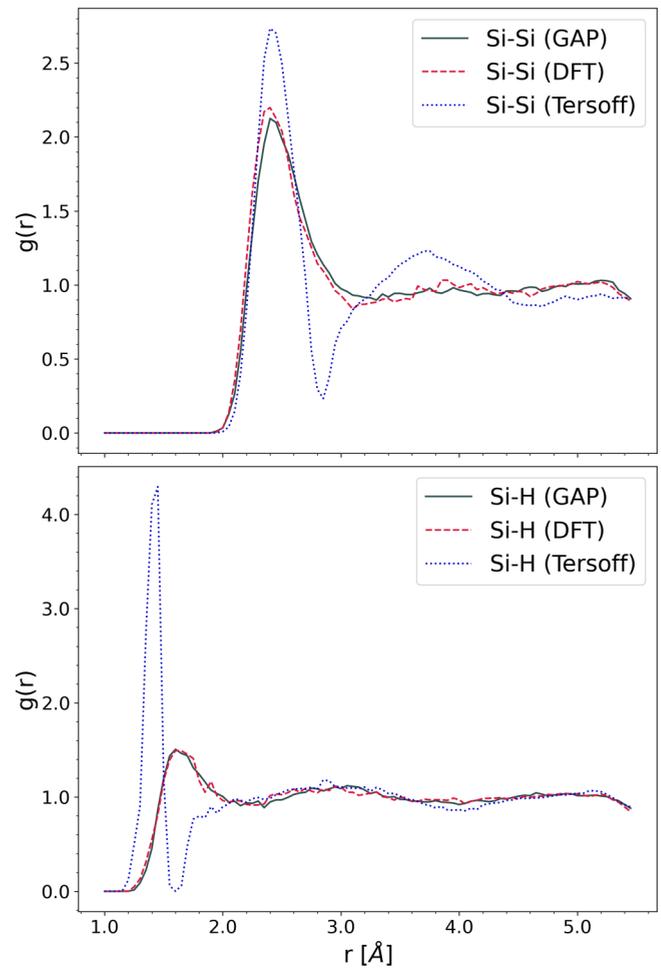


FIG. 3. Partial pair correlation functions of equilibrated liquid Si infused with H, with a density of 2.58 g/cm^3 and $T = 2000 \text{ K}$. Top: Si-Si; bottom: Si-H. Comparison provided between the Gaussian approximation potential (GAP), a Si-H Tersoff potential, and density functional theory (DFT).

x-ray diffraction measurements of the structure factor $S(Q)$ out to at least 40 \AA^{-1} [66]. Conversely, it is much easier to measure the vibrational spectra experimentally using Raman and Fourier-transform infrared spectroscopy [67], but it is a substantial challenge computationally. In this paper, we report the computation and analysis of the following observables: the partial pair correlation functions, bond angle distributions, coordination statistics, and some defect characteristics, while we leave the study of the vibrational spectra to future work.

1. Liquid phase

To simulate the structure of liquid Si infused with H, we performed constant number/volume/temperature (NVT ensemble) MD simulations as implemented in the LAMMPS software package, built with QUIP package support [61–63]. Separate simulations were carried out using the Si-H GAP and a Si-H Tersoff potential [68,69] to provide a comparison basis. Each simulation started with a cubic supercell of side length 5.26 \AA , containing 64 Si atoms and 8 H atoms placed at random, corresponding to a density of 2.58 g/cm^3 and to an approximate H at. concentration of 12%. The locations of the

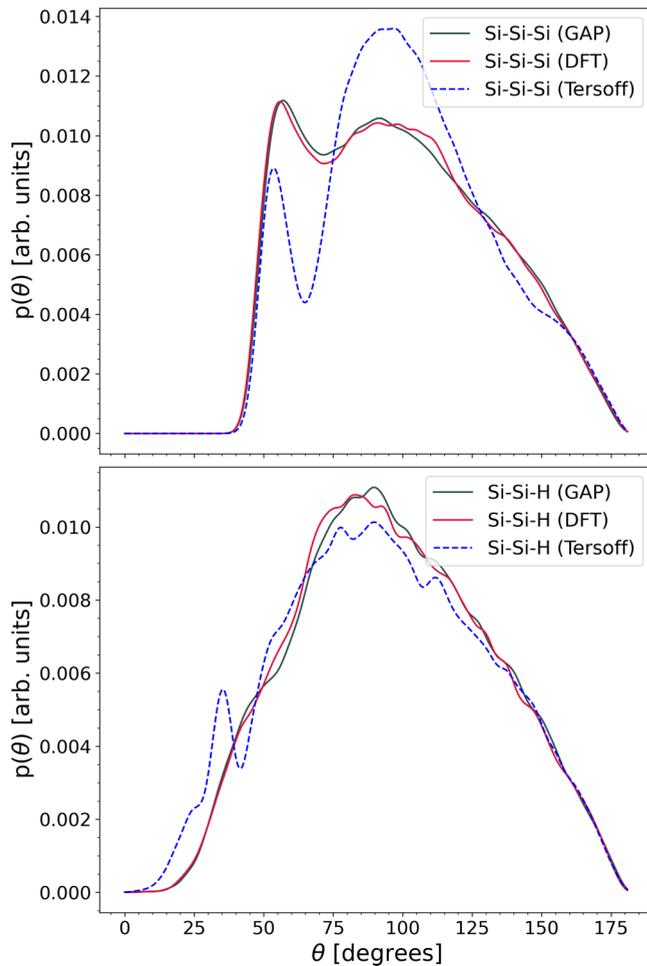


FIG. 4. Angular distribution functions, or $g_3(r, \theta)$, of equilibrated liquid Si infused with H, with a density of 2.58 g/cm^3 and $T = 2000 \text{ K}$. Top: Si-Si-Si; bottom: Si-Si-H. Comparison provided between the Si-H Gaussian approximation potential (GAP), a Si-H Tersoff potential, and density functional theory (DFT).

atoms were optimized with the interatomic potential, before equilibrating at $T = 2000 \text{ K}$ for 150 000 time steps of 0.5 fs duration. Structural data were gathered over an additional 10 000 time steps of 0.25 fs.

The same observables were also determined by DFT-BOMD simulations using QUANTUM ESPRESSO or gathered from the literature where available. Just as above, the starting configuration consisted of 64 Si and 8 H atoms located at random nonoverlapping positions in a cubic supercell of side length 5.26 \AA . Simulations consisted of equilibrating the structure at $T = 2000 \text{ K}$ over 100 000 time steps of 0.25 ps, using the Verlet algorithm, rescaling the velocities at every step to keep the temperature fixed at $T = 2000 \text{ K}$. After equilibration, structural data were gathered over an additional 6000 time steps of 0.25 ps.

Figure 3 shows the Si-Si and Si-H partial pair correlation functions $g(r)$ computed by these three methods, sometimes also referred to as RDFs. Visibly, the partial pair correlation functions $g(r)$ computed with our Si-H GAP are in excellent agreement with the results of the DFT computation: both the peak locations and peak heights exhibit high-quality match-

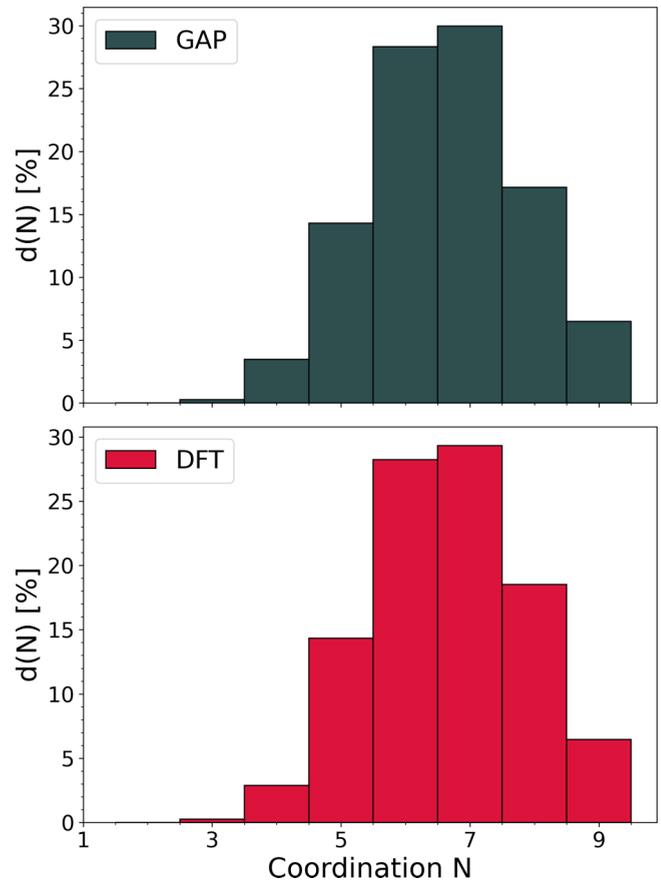


FIG. 5. Averaged coordination statistics of Si atoms in equilibrated liquid Si infused with H, with a density of 2.58 g/cm^3 and $T = 2000 \text{ K}$. Top: Gaussian approximation potential (GAP); bottom: density functional theory (DFT).

ing. In contrast, the partial pair correlation functions $g(r)$ determined with the Tersoff potential deliver much poorer correspondence with DFT.

Figure 4 shows the bond angle distribution functions $p(\theta) = g_3(r_{\text{cutoff}}, \theta)$, also known as the angular distribution functions. The cutoff bond length r_{cutoff} for $g_3(r_{\text{cutoff}}, \theta)$ is taken to be the radius corresponding to the first minimum of the partial pair correlation functions $g(r)$ beyond their initial peaks, 3.1 \AA for Si-Si and 2.2 \AA for Si-H. Just like with the partial pair correlation functions $g(r)$, the computational results with our Si-H GAP for both the Si-Si-Si bond angle distribution and the Si-Si-H bond angle distribution are in excellent agreement with the DFT computations. They reproduce all notable features of the DFT results with high quality. As before, the calculations using the Tersoff potential track the DFT results with a substantially inferior quality.

Finally, Fig. 5 shows the coordination statistics of the Si atoms. Results are only shown for the Si-H GAP and DFT. Here, the coordination includes both the neighboring Si and H atoms. The coordination shell for each atomic species is again defined using the first minimum of the corresponding partial pair correlation function. On this front, the Si-H GAP once again delivers excellent agreement with DFT. To place these observables in context, we refer to Ref. [70], which provides these same observables for pure liquid Si.

2. Amorphous phase

Next, we simulated the structure of a -Si : H. To this end, we performed both constant volume (NVT ensemble) and constant pressure (NPT) MD simulations, again as implemented in the LAMMPS software package with QUIP package support. As before, we performed the calculation using our Si-H GAP and then performed the same calculation using a Si-H Tersoff potential [68,69] to provide a point of comparison. Each simulation started with the same general procedure as the previous liquid phase section, except with a larger supercell containing 216 Si atoms and 28 H atoms and an initial density of 2.3 g/cm^3 . Once the liquid structures were equilibrated, the same general procedure was followed for each potential, except the Tersoff calculations used cooling rates which were reduced by a factor of two for further accuracy.

The details of the GAP simulation procedure were as follows. We formed a liquid Si : H phase structure at $T = 2000 \text{ K}$. Once the liquid structure was equilibrated, it was cooled in the NVT ensemble down from $T = 2000$ to 1500 K at a rate of 10^{13} K/s with a time step of 1 fs . Further equilibration was then performed at $T = 1500 \text{ K}$ for 100 ps before the structure was cooled down to $T = 500 \text{ K}$ at a rate of 10^{12} K/s . Both of these steps were also performed in the NVT ensemble. To collect structural data, the structure was relaxed to the local energy minimum in atomic positions, then equilibrated over $20\,000$ time steps of 1 fs in the NPT ensemble at 0 pressure and $T = 500 \text{ K}$, followed by gathering the structural data over an additional $10\,000$ time steps. This procedure resulted in a -Si : H with a mass density of $\sim 2.22 \text{ g/cm}^3$, consistent with PECVD films containing 12 at. \% H [71].

It should be noted that these choices of cooling rates are important for matching to experiment. While in DFT-BOMD studies the lowest achievable cooling rates are on the order of 10^{14} K/s , further lowering the cooling rate used in classical MD simulations results in a closer match to experimental atomic structures. Our choice of cooling rates is consistent with previous melt-quench studies of pure a -Si [72–76]. The choice of cooling rate was discussed explicitly in Ref. [51], where it was found that, in pure a -Si, using a GAP and a variable cooling rate between 10^{11} and 10^{13} K/s was capable of obtaining reliable structural models of a -Si. In our system, we found that using a constant 10^{12} K/s cooling rate was sufficient to yield reliable structural models of a -Si : H.

For this comparative analysis, the DFT-BOMD data were gathered from Ref. [77]. This paper performed a very extensive and therefore computationally expensive study, thereby producing very high-quality data. Finally, the comparison base was extended by including experimental data acquired by neutron scattering measurements [78].

The partial pair coordination functions $g(r)$ are presented in Fig. 6. Results are given for the Si-Si and the Si-H partial pair correlation functions $g(r)$. The key features of the Si-Si $g(r)$ are the three strong peaks at ~ 2.4 , 3.8 , and 5.6 \AA , corresponding to the first, second, and third neighbor peaks. Similarly, the key features of the Si-H $g(r)$ are the four peaks at ~ 1.5 , 3.1 , 4.9 , and 6.5 \AA , corresponding to preferential Si-H separations. Remarkably, the Si-Si and Si-H partial pair correlation functions produced with our Si-H GAP both achieved excellent agreement with the reference $g(r)$ of the

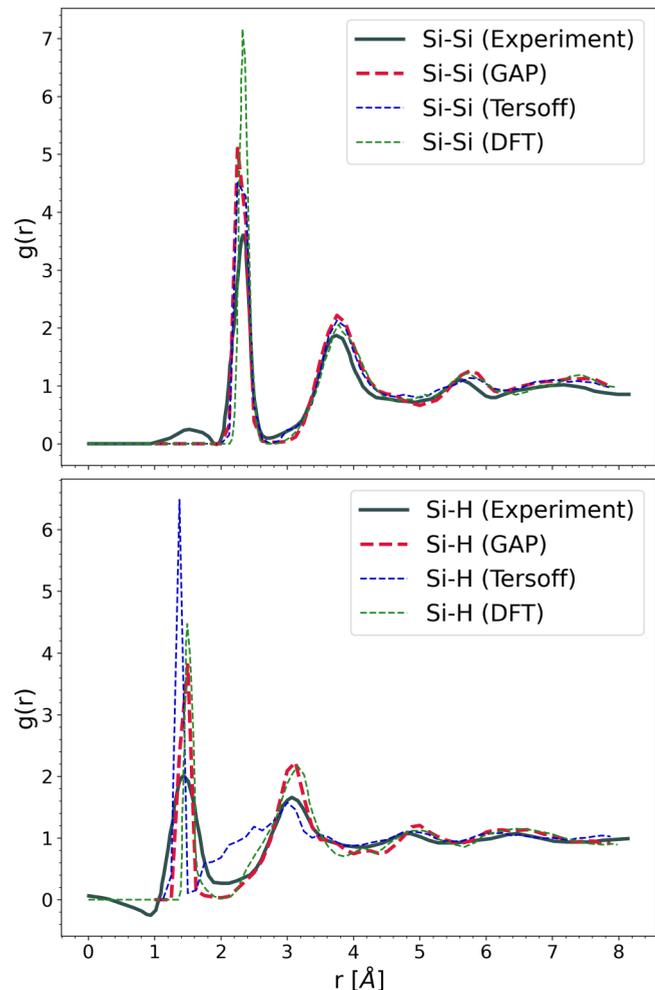


FIG. 6. Partial pair correlation functions of a -Si : H. Top: Si-Si; bottom: Si-H. Results provided for Gaussian approximation potential (GAP) and Tersoff simulations and reference results generated by a neutron scattering experiment (reproduced as published) [78] and density functional theory (DFT) calculations [77] are shown for comparison.

DFT calculations once again, matching peak locations and heights. In fact, the Si-H GAP correlation functions achieved slightly better agreement with the experiments than the DFT correlation functions. The Si-Si partial pair correlation function of the Tersoff potential notably improved compared with the liquid-Si : H case, but the Si-H partial pair correlation function continued to provide only a poor fit to the DFT and experimental results.

We extended and strengthened our validation of the Si-H GAP by performing further structural measurements. Table IV shows time-averaged measurements of the following quantities: the average coordination of the Si atoms N_c , the mean Si-Si bond length r and its standard deviation σ_r , and the mean angle $\theta_{\text{Si-Si-Si}}$ of the Si-Si-Si bond angle distribution and its standard deviation $\sigma_{\theta_{\text{Si-Si-Si}}}$. (The Si-Si-Si bond angle distribution is Gaussian in the case of a -Si : H, so its standard deviation is a well-defined quantity.) When taking these measurements, the cutoff bond lengths were taken to be the first minimum values of the partial pair correlation functions

TABLE IV. Short-range order of the a -Si : H produced by the Si-H GAP compared with Tersoff, DFT, and PECVD and sputtering (experiment) methods. N_c is the average coordination of the Si atoms, r and σ_r are the mean and the deviation of the Si-Si bond length. $\theta_{\text{Si-Si-Si}}$ and $\sigma_{\theta_{\text{Si-Si-Si}}}$ are the mean and the deviation of the Si-Si-Si bond angle distribution. The coordination and means are gathered at $T = 300$ K, while the deviations were measured on a snapshot of an optimized structure.

a -Si : H structure (H at. %)	N_c	r (Å)	σ_r (Å)	$\theta_{\text{Si-Si-Si}}$	$\sigma_{\theta_{\text{Si-Si-Si}}}$
Si-H GAP (11.5%)	3.98 ± 0.01	2.376	0.047	109.2	10.7
Tersoff (11.5%)	4.14 ± 0.01	2.392	0.077	109.0	13.2
Jarolimek (DFT, 11.1%) [77]	3.89 ± 0.03	2.377	0.049	108.9	13.6
Filipponi (PECVD, 8%) [79]	3.88 ± 0.12	2.35 ± 0.01			
Vignoli (PECVD, 12%) [80]	3.71 ± 0.07	2.37 ± 0.04			8.7
Wakagi (sputtering, 12.8%) [81]	4.0	2.363 ± 0.004	0.038 ± 0.008		9.3
Kail (various, 12-15%) [82]					9.0–9.7

in Fig. 6 after their initial peaks, 2.75 Å for Si-Si and 1.9 Å for Si-H.

As before, the validation and utility of the Si-H GAP is demonstrated by a comparison with corresponding measurements with three relevant techniques: MD using Tersoff potentials, DFT, and experimental measurements of the PECVD and sputtering type. The atomic concentrations of H in the referenced a -Si : H measurements varied from 8 to 12.8%. In all cases, the coordination and means were calculated on atomic configurations held at $T = 300$ K. Also shown in Table IV are measurements of the static disorder, namely, the deviation of the Si-Si bond lengths and Si-Si-Si bond angles, measured on a snapshot of an optimized structure.

The key message of Table IV is that MD calculations with the Si-H GAP reached the precision of DFT and thus produced results consistent with the experiments for 4 of the 5 observables, whereas Tersoff-based MD showed notable variance relative to DFT and experiments. Intriguingly, for the fifth observable, the width of the bond angle distribution $\sigma_{\theta_{\text{Si-Si-Si}}}$ of the DFT results deviated from the experiments considerably, and in this case, the Si-H GAP result reproduced the experiments even better than DFT. We also measured the Si-Si-H bond angle distribution, where the mean bond angle was found to be $105.9 \pm 6.0^\circ$, and the average Si-H bond length was found to be 1.53 ± 0.03 Å.

We also note that the coordination of the Si atoms N_c was found to be very close to 4, which indicated that very few dangling bonds were present in the GAP supercell even at $T = 300$ K. This is compelling evidence that even this limited concentration of H of 12.5% could passivate Si dangling bonds so efficiently that only 2% of all bonds remained unsatisfied.

Next, we calculated the lowest achievable excess energy of the Si : H supercell. This excess energy was defined as the energy difference of the supercell compared with a same-size supercell containing hydrogenated crystalline silicon with the same number of Si and H atoms, where the H atoms were placed at the tetrahedral interstitial sites. When both energies were computed with Si-H GAP alone, the excess energy was 0.20 eV/atom. When both structures were additionally relaxed with DFT, the excess energy was 0.09 eV/atom. This latter excess energy is well within the measured a -Si : H experimental range of 0.06–0.13 eV/atom [83]. As expected, it is lower than the excess energies of 0.13–0.14 eV/atom of pure amorphous silicon structures that we generated using the

Si-only GAP [23] for a previous work [9], as the presence of the H atoms relieved some of the atomic strain.

As a final demonstration of the utility of our ML-based Si-H interatomic potential GAP, we conducted a study of extended defects, as those play an important role in the description of a -Si. Using the same procedure as above, we created a large a -Si : H supercell that contained 4096 Si atoms and 558 H atoms. This system size is prohibitively expensive for DFT-BOMD since the computational resources required by DFT-BOMD scale with N^3 , while this size is comfortably attainable for MD utilizing the Si-H GAP. We created this large supercell using the same melt-quench procedure as described for the smaller supercells. A ball-and-stick rendering of the large supercell is presented in Fig. 7(a). We checked that the partial pair correlation functions and other short-range order statistics are almost identical for $\text{Si}_{216}\text{H}_{28}$ and $\text{Si}_{4096}\text{H}_{558}$ supercells, so we do not separately show them here. Instead, we highlight features that manifest on medium- to long-range scales, and thus can only be studied in such large supercells: the void structure of a -Si : H.

We characterized the void structure. First, we used the ZEO++ package [84] to sweep a set of Monte Carlo sampled points in our structural model and constructed the largest spheres which could be centered on each point without touching any atoms [85]. We then switched to the OVITO package to analyze whether these individual spheres overlapped/joined with other spheres. When this happened, we constructed the union of these spheres as a definition of larger voids. An example is visible at the lower-front of Figs. 7(a) and 7(b). Figure 7(b) offers an alternative visualization, where we only show the hydrogen atoms (in red) and the larger voids (in gray).

Figure 8 shows the void radius histogram, or pore-size distribution function, for the structural model of Fig. 7. Conspicuously, the majority of the voids present in the structure have radii between 2 and 4 Å, which is consistent with the void structure observed in WWW-generated a -Si : H configurations [86]. This range of void radii approximately corresponds to the size of mono- or divacancies in c -Si. There are also a few larger voids present in the structure with radii between 5 and 6 Å. Figure 7(b) shows that these larger-radius voids can and do have irregular shapes, often referred to as *nanovoids* [71]. In the studied H concentration range $\sim 12\%$, hydrogen released the structural strains very efficiently, and this relaxation caused most of the defects to shrink into

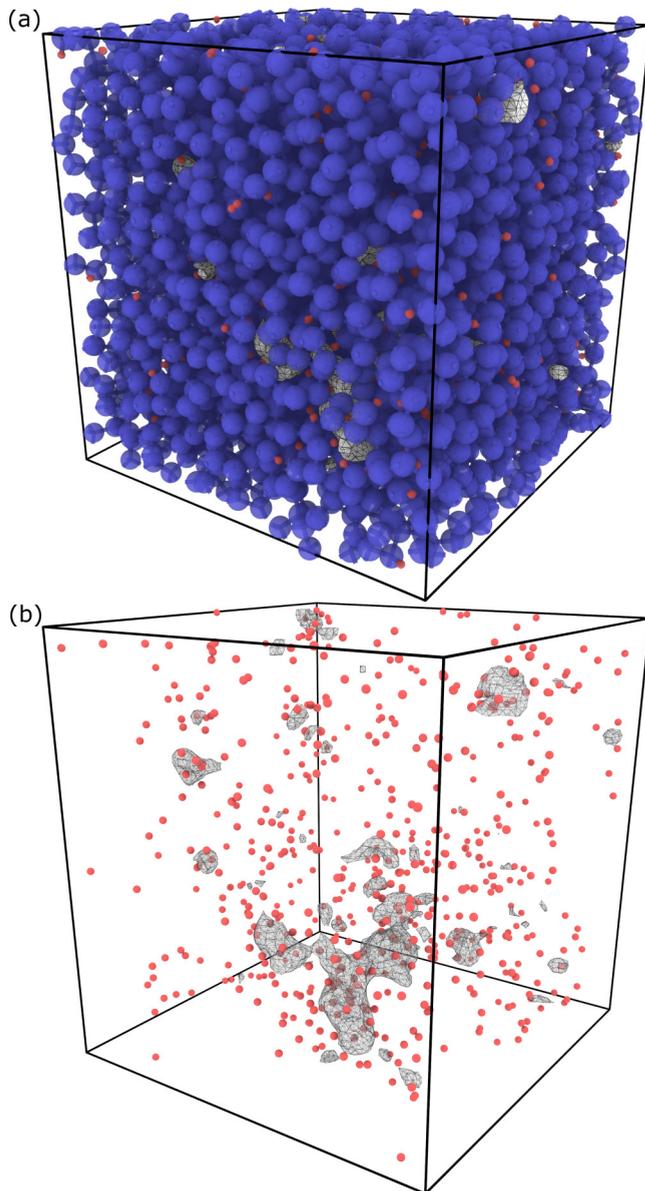


FIG. 7. (a) Ball-and-stick rendering of a structural model of a -Si : H, containing 4096 Si atoms and 558 H atoms. Si atoms are presented in blue, and H atoms are presented in red. The larger voids which are present in the atomic configuration are highlighted here with white surface meshes. (b) A visualization of the distribution of H atoms and voids within the supercell of the structural model. The voids here are represented with transparent white meshes.

mono- and divacancies, with only a few nanovoids left. Experimenters studying the nature of voids in a -Si : H have found that nanovoids only begin to dominate $> 14 \pm 2$ at. % H [71].

IV. CONCLUSIONS

In this paper, we reported on the development of a silicon-hydrogen GAP with a SOAP kernel, Si-H GAP. We trained this potential over 27 iterative rounds, simulating a wide range of phases, structures, and thermal protocols, while fitting energies, forces, and stresses to the results of DFT

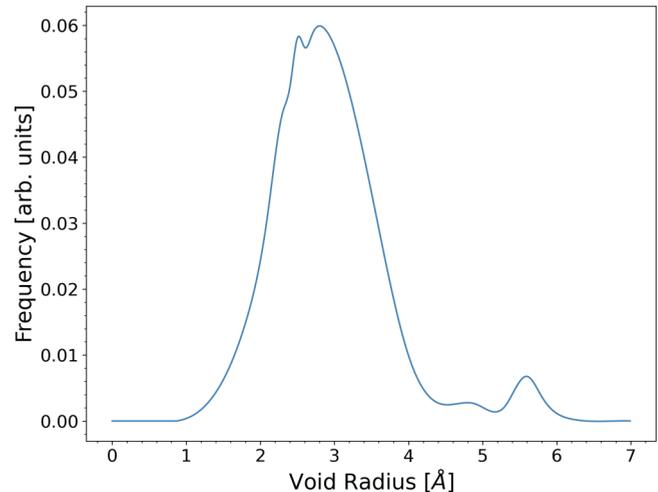


FIG. 8. The void radius histogram of the a -Si : H structural model shown in Fig. 7, as calculated by ZEO++.

calculations in each round. The training set was expanded in each round only as required by an adaptive protocol. This adaptive training enables the economic use of computational resources.

The Si-H GAP was not only able to closely match DFT measurements of microscopic quantities such as the energies, forces, and virial stresses, but it was also able to reproduce structural characteristics including partial pair correlation functions, bond angle distributions, bond length distributions, and average coordination numbers.

For context and comparison, we also performed MD simulations based on the Tersoff potential and calculated the same quantities. The comparison of the parametric Tersoff and nonparametric ML-based Si-H GAP results demonstrated the improvement that is obtained by adopting the nonparameteric ML model. While the impressive increase of accuracy of the Si-H GAP comes at a cost of computational complexity and speed over traditional parameteric interatomic potentials, the Si-H GAP achieves DFT-level accuracy while still being many orders of magnitude faster than DFT.

The Si-H GAP developed here is limited in a few ways. (i) It was not developed to be a general-purpose interatomic potential and thus did not include structural phases beyond the key phases which are needed to reliably produce experimentally accurate amorphous Si : H. For instance, it did not include all of the various surface reconstructions of diamond Si. Fortunately, future training could simply add additional structures to the training set, should this be desired, and thus, the potential can be adapted as needed in future work. (ii) Long-range interactions beyond 5 Å for Si and 3.5 Å for H were not included. This limits the maximum accuracy of the potential, as it runs into locality limits. For a discussion of this phenomenon, see Ref. [64]. Properly including long-range interactions and integrating them with the short-range interactions is still an outstanding problem within the GAP framework, although simple dispersion and Coulomb interactions are easy to add on, as has been done before in

Refs. [33,87,88]. (iii) The training database was assembled by hand and did not include any automated procedures for constructing the database. Using an active learning approach to create and manage the training set would reduce the cost of adaptively expanding the training set, as DFT would only be performed on an as-needed basis and would speed up the training of the potential.

Those issues aside, the merits of the presented potential are clear. Since in electro-optical applications and in most experimental studies amorphous silicon structures typically contain some amount of hydrogen, our Si-H GAP can prove to be very useful to enable simulation studies with unprecedented accuracy and utility.

The Si-H GAP [89,90], the GAP suite of programs (for noncommercial use) [91], the QUANTUM ESPRESSO software package [92], the LAMMPS software package [62], and ZEO++ [93] are freely available.

ACKNOWLEDGMENTS

We acknowledge useful and inspiring discussions with Chase Hansen, Mariana Bertoni, and Salman Manzoor. This paper was supported by US Department of Energy Solar Energy Technologies Office Grants No. DE-EE0008979 and No. DE-EE0009835.

-
- [1] M. J. Thompson and H. C. Tuan, Amorphous silicon electronic devices and their applications, *IEEE Aerosp. Electron* **2**, 25 (1987).
- [2] D. E. Carlson and C. R. Wronski, Amorphous silicon solar cell, *Appl. Phys. Lett.* **28**, 671 (1976).
- [3] S. Bernardini and M. I. Bertoni, Insights into the degradation of amorphous silicon passivation layer for heterojunction solar cells, *phys. status solidi (a)* **216**, 1800705 (2019).
- [4] R. C. Chittick, J. H. Alexander, and H. F. Sterling, The preparation and properties of amorphous silicon, *J. Electrochem. Soc.* **116**, 77 (1969).
- [5] P. G. Le Comber and W. E. Spear, Electronic Transport in Amorphous Silicon Films, *Phys. Rev. Lett.* **25**, 509 (1970).
- [6] W. Spear, R. Loveland, and A. Al-Sharbaty, The temperature dependence of photoconductivity in *a*-Si, *J. Non-Cryst. Solids* **15**, 410 (1974).
- [7] D. L. Staebler and C. R. Wronski, Reversible conductivity changes in discharge-produced amorphous Si, *Appl. Phys. Lett.* **31**, 292 (1977).
- [8] J. Holovsky, S. M. De Nicolás, S. De Wolf, and C. Ballif, Amorphous/crystalline silicon interface stability: Correlation between infrared spectroscopy and electronic passivation properties, *Adv. Mater. Interfaces* **7**, 2000957 (2020).
- [9] D. Unruh, R. V. Meidanshahi, C. Hansen, S. Manzoor, S. M. Goodnick, M. I. Bertoni, and G. T. Zimanyi, From femtoseconds to gigaseconds: The SOLDEG platform for the performance degradation analysis of silicon heterojunction solar cells, *ACS Appl. Mater. Interfaces* **13**, 32424 (2021).
- [10] C. Deline, S. Johnston, S. R. Rummel, B. Sekulic, D. C. Jordan, S. R. Rummel, P. Hacke, S. R. Kurtz, K. O. Davis, E. J. Schneller *et al.*, Silicon heterojunction system field performance, *IEEE J. Photovoltaics* **8**, 177 (2018).
- [11] K. Masuko, M. Shigematsu, T. Hashiguchi, D. Fujishima, M. Kai, N. Yoshimura, T. Yamaguchi, Y. Ichihashi, T. Mishima, N. Matsubara *et al.*, Achievement of more than 25 crystalline silicon heterojunction solar cell, *IEEE J. Photovolt* **4**, 1433 (2014).
- [12] P. Geerlings, F. De Proft, and W. Langenaeker, Conceptual density functional theory, *Chem. Rev.* **103**, 1793 (2003).
- [13] J. C. A. Prentice, J. Aarons, J. C. Womack, A. E. A. Allen, L. Andrinopoulos, L. Anton, R. A. Bell, A. Bhandari, R. K. Charlton, G. A. Bramley *et al.*, The ONETEP linear-scaling density functional theory program, *J. Chem. Phys.* **152**, 174111 (2020).
- [14] M. Finnis, *Interatomic Forces in Condensed Matter* (Oxford University Press, Oxford, 2003).
- [15] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon, *Phys. Rev. B* **51**, 12947 (1995).
- [16] S. Goedecker and L. Colombo, Efficient Linear Scaling Algorithm for Tight-Binding Molecular Dynamics, *Phys. Rev. Lett.* **73**, 122 (1994).
- [17] J. D. Kress, S. R. Bickham, L. A. Collins, B. L. Holian, and S. Goedecker, Tight-Binding Molecular Dynamics of Shock Waves in Methane, *Phys. Rev. Lett.* **83**, 3896 (1999).
- [18] T. Frauenheim, F. Weich, T. Köhler, S. Uhlmann, D. Porezag, and G. Seifert, Density-functional-based construction of transferable nonorthogonal tight-binding potentials for Si and SiH, *Phys. Rev. B* **52**, 11492 (1995).
- [19] M. Gaus, Q. Cui, and M. Elstner, Density functional tight binding: Application to organic and biological molecules, *WIREs Comput. Mol. Sci.* **4**, 49 (2014).
- [20] M. Elstner and G. Seifert, Density functional tight binding, *Phil. Trans. R. Soc. A* **372**, 20120483 (2014).
- [21] J. P. Lewis, P. Jelínek, J. Ortega, A. A. Demkov, D. G. Trabada, B. Haycock, H. Wang, G. Adams, J. K. Tomfohr, E. Abad *et al.*, Advances and applications in the fireball *ab initio* tight-binding molecular-dynamics formalism, *Phys. Status Solidi B* **248**, 1989 (2011).
- [22] A. N. Andriotis and M. Menon, Self-consistent tight-binding molecular-dynamics method for cluster studies, *Phys. Rev. B* **59**, 15942 (1999).
- [23] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Machine Learning a General-Purpose Interatomic Potential for Silicon, *Phys. Rev. X* **8**, 041048 (2018).
- [24] H. Balamane, T. Halicioglu, and W. A. Tiller, Comparative study of silicon empirical interatomic potentials, *Phys. Rev. B* **46**, 2250 (1992).
- [25] S. J. Cook and P. Clancy, Comparison of semiempirical potential functions for silicon and germanium, *Phys. Rev. B* **47**, 7686 (1993).
- [26] A. van Duin, A. Strachan, S. Stewman, Q. Zhang, X. Xu, and W. Goddard, ReaxFF_{SiO} reactive force field for silicon and silicon oxide systems, *J. Phys. Chem. A* **107**, 3803 (2003).
- [27] I. Kwon, R. Biswas, and C. M. Soukoulis, Molecular-dynamics simulations of defect formation in hydrogenated amorphous-silicon, *Phys. Rev. B* **45**, 3332 (1992).

- [28] T. Ohira, O. Ukai, M. Noda, Y. Takeuchi, M. Murata, and H. Yoshida, Molecular-dynamics simulations of hydrogenated amorphous silicon thin-film growth, *MRS Proceedings* **408**, 445 (1995).
- [29] B. Tuttle and J. B. Adams, Structure of α -Si : H from Harris-functional molecular dynamics, *Phys. Rev. B* **53**, 16265 (1996).
- [30] M. V. Ramana Murty and H. A. Atwater, Empirical interatomic potential for Si-H interactions, *Phys. Rev. B* **51**, 4889 (1995).
- [31] J. Boland, Scanning-tunneling microscopy of the interaction of hydrogen with silicon surfaces, *Adv. Phys.* **42**, 129 (1993).
- [32] Y. Sakiyama, Y. Iga, H. Yamaguchi, S. Takagi, and Y. Matsumoto, Multiscale analysis for dissociative adsorption of SiH₄ on Si(100) surface, *Surf. Coat. Technol.* **200**, 3385 (2006).
- [33] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [34] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, *Phys. Rev. B* **87**, 184115 (2013).
- [35] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, *J. Comput. Phys.* **285**, 316 (2015).
- [36] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, Accurate force field for molybdenum by machine learning large materials data, *Phys. Rev. Materials* **1**, 043603 (2017).
- [37] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, *Phys. Rev. Lett.* **98**, 146401 (2007).
- [38] J. Behler, Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations, *Phys. Chem. Chem. Phys.* **13**, 17930 (2011).
- [39] J. Behler, Perspective: Machine learning potentials for atomistic simulations, *J. Chem. Phys.* **145**, 170901 (2016).
- [40] E. Kocer, J. K. Mason, and H. Erturk, A novel approach to describe chemical environments in high-dimensional neural network potentials, *J. Chem. Phys.* **150**, 154102 (2019).
- [41] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, *Multiscale Model. Simul.* **14**, 1153 (2016).
- [42] R. Jinnouchi, F. Karsai, and G. Kresse, On-the-fly machine learning force field generation: Application to melting points, *Phys. Rev. B* **100**, 014105 (2019).
- [43] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, *Sci. Adv.* **3**, e1603015 (2017).
- [44] J. S. Smith, O. Isayev, and A. E. Roitberg, ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost, *Chem. Sci.* **8**, 3192 (2017).
- [45] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B* **99**, 014104 (2019).
- [46] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network, *Phys. Rev. B* **92**, 045131 (2015).
- [47] S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker, and M. Amsler, High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride, *Phys. Rev. B* **95**, 104105 (2017).
- [48] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, General-purpose machine learning potentials capturing nonlocal charge transfer, *Acc. Chem. Res.* **54**, 808 (2021).
- [49] V. L. Deringer, M. A. Caro, and G. Csányi, Machine learning interatomic potentials as emerging tools for materials science, *Adv. Mater.* **31**, 1902765 (2019).
- [50] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, Gaussian process regression for materials and molecules, *Chem. Rev.* **121**, 10073 (2021).
- [51] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics, *J. Phys. Chem. Lett.* **9**, 2879 (2018).
- [52] V. L. Deringer, N. Bernstein, G. Csányi, C. B. Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, Origins of structural and electronic transitions in disordered silicon, *Nature (London)* **589**, 59 (2021).
- [53] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Physics-inspired structural representations for molecules and materials, *Chem. Rev.* **121**, 9759 (2021).
- [54] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus *et al.*, The atomic simulation environment—a Python library for working with atoms, *J. Phys.: Condens. Matter* **29**, 273002 (2017).
- [55] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials, *J. Phys.: Condens. Matter* **21**, 395502 (2009).
- [56] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. Buongiorno Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni *et al.*, Advanced capabilities for materials modelling with QUANTUM ESPRESSO, *J. Phys.: Condens. Matter* **29**, 465901 (2017).
- [57] P. Giannozzi, O. Baseggio, P. Bonfà, D. Brunato, R. Car, I. Carnimeo, C. Cavazzoni, S. de Gironcoli, P. Delugas, F. Ferrari Ruffino *et al.*, QUANTUM ESPRESSO toward the exascale, *J. Chem. Phys.* **152**, 154105 (2020).
- [58] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [59] N. Marzari, D. Vanderbilt, A. De Vita, and M. C. Payne, Thermal Contraction and Disordering of the Al(110) Surface, *Phys. Rev. Lett.* **82**, 3296 (1999).
- [60] R. Car and M. Parrinello, Unified Approach for Molecular Dynamics and Density-Functional Theory, *Phys. Rev. Lett.* **55**, 2471 (1985).
- [61] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, *J. Comput. Phys.* **117**, 1 (1995).
- [62] <http://lammps.sandia.gov>.
- [63] <https://github.com/libAtoms/QUIP>.
- [64] V. L. Deringer and G. Csányi, Machine learning based interatomic potential for amorphous carbon, *Phys. Rev. B* **95**, 094203 (2017).

- [65] F. Gaspari, I. M. Kupchak, A. I. Shkrebtii, and J. M. Perz, Modeling noncrystalline materials: Use of vibrational spectra as a protocol for validation, *Phys. Rev. B* **79**, 224203 (2009).
- [66] K. Laaziri, S. Kycia, S. Roorda, M. Chicoine, J. L. Robertson, J. Wang, and S. C. Moss, High Resolution Radial Distribution Function of Pure Amorphous Silicon, *Phys. Rev. Lett.* **82**, 3460 (1999).
- [67] K. O. Bugaev, A. A. Zelenina, and V. A. Volodin, Vibrational spectroscopy of chemical species in silicon and silicon-rich nitride thin films, *Int. J. Spectrosc.* **2012**, 281851 (2012).
- [68] J. Tersoff, Modeling solid-state chemistry: Interatomic potentials for multicomponent systems, *Phys. Rev. B* **39**, 5566 (1989).
- [69] F. de Brito Mota, J. F. Justo, and A. Fazio, Hydrogen role on the properties of amorphous silicon nitride, *J. Appl. Phys.* **86**, 1843 (1999).
- [70] I. Štich, R. Car, and M. Parrinello, Structural, bonding, dynamical, and electronic properties of liquid silicon: An *ab initio* molecular-dynamics study, *Phys. Rev. B* **44**, 4262 (1991).
- [71] A. H. M. Smets, W. M. M. Kessels, and M. C. M. van de Sanden, Vacancies and voids in hydrogenated amorphous silicon, *Appl. Phys. Lett.* **82**, 1547 (2003).
- [72] R. V. Meidanshahi, S. Bowden, and S. M. Goodnick, Electronic structure and localized states in amorphous Si and hydrogenated amorphous Si, *Phys. Chem. Chem. Phys.* **21**, 13248 (2019).
- [73] M. Legesse, M. Nolan, and G. Fagas, A first principles analysis of the effect of hydrogen concentration in hydrogenated amorphous silicon on the formation of strained Si-Si bonds and the optical and mobility gaps, *J. Appl. Phys.* **115**, 203711 (2014).
- [74] M. Legesse, M. Nolan, and G. Fagas, Revisiting the dependence of the optical and mobility gaps of hydrogenated amorphous silicon on hydrogen concentration, *J. Phys. Chem. C* **117**, 23956 (2013).
- [75] M. Nolan, M. Legesse, and G. Fagas, Surface orientation effects in crystalline-amorphous silicon interfaces, *Phys. Chem. Chem. Phys.* **14**, 15173 (2012).
- [76] A. Pedersen, L. Pizzagalli, and H. Jónsson, Optimal atomic structure of amorphous silicon obtained from density functional theory calculations, *New J. Phys.* **19**, 063018 (2017).
- [77] K. Jarolimek, R. A. de Groot, G. A. de Wijs, and M. Zeman, First-principles study of hydrogenated amorphous silicon, *Phys. Rev. B* **79**, 155206 (2009).
- [78] R. Bellisent, A. Menelle, W. Howells, A. C. Wright, T. Brunier, R. Sinclair, and F. Jansen, The structure of amorphous Si : H using steady state and pulsed neutron sources, *Phys. B: Condens. Matter* **156-157**, 217 (1989).
- [79] A. Filipponi, F. Evangelisti, M. Benfatto, S. Mobilio, and C. R. Natoli, Structural investigation of *a*-Si and *a*-Si : H using x-ray-absorption spectroscopy at the Si *K* edge, *Phys. Rev. B* **40**, 9636 (1989).
- [80] S. Vignoli, P. Mélinon, B. Masenelli, P. R. I. Cabarrocas, A. M. Flank, and C. Longeaud, Over-coordination and order in hydrogenated nanostructured silicon thin films: Their influence on strain and electronic properties, *J. Phys.: Condens. Matter* **17**, 1279 (2005).
- [81] M. Wakagi, K. Ogata, and A. Nakano, Structural study of *a*-Si and *a*-Si : H films by EXAFS and Raman-scattering spectroscopy, *Phys. Rev. B* **50**, 10666 (1994).
- [82] F. Kail, J. Farjas, P. Roura, C. Secouard, O. Nos, J. Bertomeu, F. Alzina, and P. R. I. Cabarrocas, Relaxation and derelaxation of pure and hydrogenated amorphous silicon during thermal annealing experiments, *Appl. Phys. Lett.* **97**, 031918 (2010).
- [83] F. Kail, J. Farjas, P. Roura, C. Secouard, O. Nos, J. Bertomeu, and P. R. I. Cabarrocas, The configurational energy gap between amorphous and crystalline silicon, *Phys. Status Solidi RRL* **5**, 361 (2011).
- [84] T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, and M. Haraczkyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials, *Microporous Mesoporous Mater.* **149**, 134 (2012).
- [85] M. Pinheiro, R. L. Martin, C. H. Rycroft, A. Jones, E. Iglesia, and M. Haraczkyk, Characterization and comparison of pore landscapes in crystalline porous materials, *J. Mol. Graphics Modell.* **44**, 208 (2013).
- [86] E. Guerrero and D. A. Strubbe, Computational generation of voids in *a*-Si and *a*-Si : H by cavitation at low density, *Phys. Rev. Materials* **4**, 025601 (2020).
- [87] V. L. Deringer, M. A. Caro, and G. Csányi, A general-purpose machine-learning force field for bulk and nanostructured phosphorus, *Nat. Commun.* **11**, 5461 (2020).
- [88] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, and A. Michaelides, An accurate and transferable machine learning potential for carbon, *J. Chem. Phys.* **153**, 034702 (2020).
- [89] <https://github.com/dgunruh/Si-H-GAP>.
- [90] <https://libatoms.github.io/GAP/data.html>.
- [91] <http://github.com/libatoms/GAP>.
- [92] www.quantum-espresso.org.
- [93] www.zeoplusplus.org.