IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 045029 https://doi.org/10.1088/2632-2153/ac2cfe

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
19 May 2021

REVISED
25 August 2021

ACCEPTED FOR PUBLICATION
5 October 2021

PUBLISHED
22 October 2021

PAPER

Predicting polarizabilities of silicon clusters using local chemical environments

Mario G Zauchner[1], Stefano Dal Forno[2], Gábor Csányi[3], Andrew Horsfield[1,*] and Johannes Lischner[1]

[1] Department of Materials, Thomas Young Centre, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[2] Department of Physics, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[3] Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: mario.zauchner15@imperial.ac.uk

Keywords: machine learning polarizabilities, silicon cluster polarizabilities, predicting polarizabilities of nanoparticles, RPA polarizabilities of silicon clusters

Abstract

Calculating polarizabilities of large clusters with first-principles techniques is challenging because of the unfavorable scaling of computational cost with cluster size. To address this challenge, we demonstrate that polarizabilities of large hydrogenated silicon clusters containing thousands of atoms can be efficiently calculated with machine learning methods. Specifically, we construct machine learning models based on the smooth overlap of atomic positions (SOAP) descriptor and train the models using a database of calculated random-phase approximation polarizabilities for clusters containing up to 110 silicon atoms. We first demonstrate the ability of the machine learning models to fit the data and then assess their ability to predict cluster polarizabilities using k-fold cross validation. Finally, we study the machine learning predictions for clusters that are too large for explicit first-principles calculations and find that they accurately describe the dependence of the polarizabilities on the ratio of hydrogen to silicon atoms and also predict a bulk limit that is in good agreement with previous studies.

1. Introduction

Clusters and nanoparticles are used in a variety of scientific and industrial applications, including optoelectronics [1], photocatalysis [2], medical imaging [3, 4] or single electron transistors [5]. Electronic excitations often play a key role in these applications, but theoretical techniques for calculating excited-state properties of materials, such as the first-principles GW/Bethe–Salpeter method, are typically limited to very small systems. A key bottleneck of such excited-state calculations of clusters and nanoparticles is the determination of the static polarizability which is often calculated using a sum-over-states technique [6].

As an efficient alternative to first-principles techniques, machine learning (ML) based techniques have been explored in recent years. For example, ML has been employed to efficiently represent potential energy surfaces [7–10] or to predict electronic ground state densities [11–14]. Additionally, projects such as the Materials Project [15] and the Open Quantum Materials database [16, 17] have made an effort to make first-principles data of a wide range of materials publicly available. A key ingredient in ML methods is a descriptor which acts as a molecular fingerprint and encodes the structure and chemistry of a molecule. For example, the smooth-overlap of atomic positions (SOAP) descriptor [18] has been widely used for the comparison of different chemical environments of an atom. For this, the overlap of the corresponding neighbourhood densities (constructed as a sum of Gaussians centered on atoms in the local environment) is expressed in terms of the coefficients in a basis of spherical harmonics and radial basis functions [18]. Very recently, several groups have also started to explore the applicability of ML approaches to calculate molecular polarizabilities and dipole moments [19–22]. For example, Grisafi *et al* [20] introduced a symmetry adapted variant of the SOAP descriptor [18] to predict polarizability tensors of molecules. Similarly,

Wilkins and coworkers [19] used the symmetry-adapted SOAP kernel to predict polarizabilities and first hyperpolarizabilities of small organic molecules with high accuracy. Recently, Veit *et al* [22] used a combination of the symmetry adapted SOAP kernel and the scalar SOAP kernel to predict dipole moments of small molecules with close to DFT accuracy. However, the applicability of ML approaches to polarizabilities of clusters and nanoparticles, in particular the ability to make predictions about nanoparticles too large for first-principles calculations remains largely unexplored.

To assess the performance of ML approaches for cluster polarizabilities, we focus in this work on hydrogenated silicon clusters. These systems are well suited for this purpose because their polarizabilities have been studied in detail with a variety of modelling techniques. For example, simple empirical models, such as bond polarizability models, have been used to predict Raman spectra of silicon clusters in good agreement with experiment [23]. Empirical models can be extended beyond the assumption of additivity of atomic polarizabilities. A class of models that captures interactions between polarization centres are dipole interaction models [24], which have been successfully applied to the construction of polarizable force fields [25]. Highly accurate cluster polarizabilities can be obtained using *ab initio* approaches such as density functional theory (DFT) [26–31], Møller–Plesset perturbation theory [31], coupled-cluster theory [31–33] or the random phase approximation (RPA) [34–36]. For example, Mochizuki and Ågren [36] used the RPA and the second-order polarization propagator approximation to calculate the polarizabilities of spherical hydrogenated silicon clusters with up to 35 Si atoms and found that the polarizability per silicon atom approaches the bulk limit from below. In contrast, for unhydrogenated silicon clusters Jackson and coworkers found that the bulk value is approached from above as the size of the cluster increases [29, 30]. This behaviour was attributed to the presence of dangling bonds on the surface. Furthermore, it was observed that the polarizability depends sensitively on the shape of the cluster [30, 35]. Jansik *et al* [35] compared polarizabilities of three-dimensional (3D), two-dimensional (2D) and one-dimensional (1D) hydrogenated silicon structures and found that the presence of π-bonds in 2D systems leads to a much stronger increase in the polarizability as a function of cluster size when compared to 1D and 3D clusters [35]. A similar trend was observed when comparing prolate and compact clusters, with prolate structures showing a significantly larger polarizability per silicon atom than compact structures [30].

In the present work, we explore the ability of machine learning models based on the SOAP [18] descriptor to describe and predict static polarizabilities of hydrogenated silicon clusters calculated from RPA static density-density response functions. We chose the SOAP descriptor due it its widespread use for a variety of atomic scale regression tasks and its systematic nature. Previous work [19, 20] has already demonstrated the ability to predict isotropic scalar polarizabilities and also the full polarizability tensor using SOAP and generalizations thereof. The symmetry-adapted SOAP descriptor has also been used successfully in conjunction with physical insights [22] to predict molecular dipole moments. Furthermore, we note that SOAP is a generic 3-body descriptor of the neighbour density that obeys rotational and permutational symmetries [37], and as such encompasses simpler descriptors such as RDFs and ADFs [38–40] (which are particular projections of the neighbour density), and in the limit of no basis truncation equivalent to other 3-body descriptors such as Behler-Parrinello Atom Centered Symmetry Functions [41] and the FCHL [42] descriptors. To generate a data set, we first calculate scalar isotropic polarizabilities of a set of clusters containing between 10 and 110 silicon atoms using the RPA. We then investigate the ability of the ML approach to reproduce the calculated polarizabilities and find that almost perfect agreement can be obtained when the size of the local chemical environments is sufficiently large to contain the whole cluster. Importantly, the ML models already describe the qualitative behaviour of the average polarizability per atom if the local environment only contains nearest neighbour atoms. These findings establish the fittability of RPA scalar polarizabilities using local SOAP descriptors which—in contrast to mean-field DFT data—has not been explored to date. Next, we study the ability of ML to predict polarizabilities of clusters. Interestingly, we find that the predictive power of ML is strongest when the size of the chemical environment is relatively small. These insights enable the reliable prediction of polarizabilities of large clusters which are difficult to calculate with standard first-principles techniques and constitute a first step towards efficient ML approaches for excited-state properties of materials.

2. Methods

2.1. Random phase approximation polarizabilities

Scalar polarizabilities of molecules and clusters were calculated within the RPA in a linear response framework. The RPA was chosen because it is known to give an accurate description of the dielectric properties of bulk silicon [43]. The polarizability tensor α_{ij} relates the induced dipole moment with Cartesian components μ_i to the applied static electric field E_j according to

$$\mu_i = \sum_j \alpha_{ij} E_j. \tag{1}$$

To obtain an expression for α_{ij}, we express μ_i in terms of the induced electronic charge density $\Delta\rho(\mathbf{r})$ via

$$\mu_i = -e \int d\mathbf{r} \Delta\rho(\mathbf{r}) r_i, \tag{2}$$

where e denotes the proton charge and r_i is the Cartesian component of the position vector. The induced charge density is determined by the interacting density-density response function $\chi(\mathbf{r}, \mathbf{r}')$ according to

$$\Delta\rho(\mathbf{r}) = e \sum_j E_j \int d\mathbf{r}' \chi(\mathbf{r}, \mathbf{r}') r_j', \tag{3}$$

where we used that the potential associated with the applied electric field is given by $V(\mathbf{r}) = e\sum_j E_j r_j$. Combining these equations yields

$$\alpha_{ij} = -e^2 \int d\mathbf{r} d\mathbf{r}' \chi(\mathbf{r}, \mathbf{r}') r_i r_j'. \tag{4}$$

Finally, the scalar polarizability α is obtained by dividing the trace of α_{ij} by three.

To evaluate equation (4) the interacting density-density response function must be determined. In the RPA χ obeys the Dyson equation

$$\chi(\mathbf{r}, \mathbf{r}') = \chi_0(\mathbf{r}, \mathbf{r}') \tag{5}$$

$$+ \int d\mathbf{r}_1 d\mathbf{r}_2 \, \chi_0(\mathbf{r}, \mathbf{r}_1) v(\mathbf{r}_1 - \mathbf{r}_2) \chi(\mathbf{r}_2, \mathbf{r}'), \tag{6}$$

where $v(\mathbf{r} - \mathbf{r}')$ denotes the Coulomb interaction and χ_0 is the non-interacting density-density response function given by [44, 45]

$$\chi_0(\mathbf{r}, \mathbf{r}') = \sum_{ij} \frac{f_i(1 - f_j)}{\epsilon_i - \epsilon_j} \tag{7}$$

$$\times \left[\phi_i^*(\mathbf{r}) \phi_j(\mathbf{r}) \phi_j^*(\mathbf{r}') \phi_i(\mathbf{r}') + \text{c.c.} \right], \tag{8}$$

where f_i denotes an occupancy factor and ϕ_i and ϵ_i denote Kohn–Sham orbitals and eigenvalues, respectively. Note that the summation ranges over both occupied and unoccupied states resulting in the well-known difficulties of converging such sum-over-states expressions. To numerically calculate scalar polarizabilities, we employ a plane-wave/pseudopotential approach. Specifically, the BerkeleyGW programme package [46, 47] is used to calculate $\chi_{\mathbf{G}\mathbf{G}'}$ where $\mathbf{G}$ and $\mathbf{G}'$ denote reciprocal lattice vectors of the periodically repeated supercell. Note that interactions between images are avoided by using a truncated Coulomb interaction. The interacting density-density response function in real space is then given by

$$\chi(\mathbf{r}, \mathbf{r}') = \frac{1}{V} \sum_{\mathbf{G},\mathbf{G}'} e^{i\mathbf{G}\cdot\mathbf{r}} \chi_{\mathbf{G},\mathbf{G}'} e^{-i\mathbf{G}'\cdot\mathbf{r}'}, \tag{9}$$

where $V = L^3$ denotes the volume of the cubic supercell, with L being the side length. Finally, the scalar polarizability is found to be

$$\alpha = \frac{e^2}{3V} \sum_i \sum_{\mathbf{G},\mathbf{G}'} \chi_{\mathbf{G},\mathbf{G}'} \Delta_{\mathbf{G},i} \Delta_{\mathbf{G}',i}^*, \tag{10}$$

with

$$\Delta_{\mathbf{G},x} = \begin{cases} \frac{L^4}{2} \delta_{G_x,0} \delta_{G_y,0} \delta_{G_z,0} & \text{if } G_x = 0, \\ \frac{L^3}{iG_x} \delta_{G_y,0} \delta_{G_z,0} & \text{otherwise,} \end{cases} \tag{11}$$

and similar expressions for $\Delta_{\mathbf{G},y}$ and $\Delta_{\mathbf{G},z}$.

Finally, we note that other—more efficient—approaches than the one described above exist for the calculation of the static scalar polarizability, such as the finite field method [48]. However, our ultimate interest is in applying ML techniques to accelerate excited-state calculations and these methods require the full interacting density-density response function which cannot easily be obtained with other methods.

2.2. Environment descriptors

The ability to assess the similarity of different chemical environments plays a key role in machine learning of material properties. In this work, we use the SOAP approach [18] where the environment of atom i is described by the set of neighbourhood densities

$$\rho_i^{\nu}(\mathbf{r}) = \sum_{i \neq j}^{N_\nu} e^{-\gamma_\nu (\mathbf{r}-\mathbf{r}_{ij})^2}, \tag{12}$$

where ν denotes a specific element that is present in the atom's environment with N_ν being the number of such atoms up to a given cut-off radius r_{cut}. In addition, γ_ν is a hyperparameter describing the size of the neighbour atom.

The similarity of two chemical environments described by the neighbourhood densities $\rho_i = \{\rho_i^{\nu}\}_\nu$ and $\rho_j = \{\rho_j{}^{\nu}\}_\nu$ can be measured by the kernel [49]

$$k(\rho_i, \rho_j) = \int d\hat{R} \left| \sum_{\nu} \int d\mathbf{r} \rho_i^{\nu}(\mathbf{r}) \rho_j^{\nu}(\hat{R}\mathbf{r}) \right|^2, \tag{13}$$

where $\hat{R}$ denotes a rotation matrix. To evaluate the kernel integral, the angular dependence of the neighbourhood densities is expanded in a basis of spherical harmonics Y_{lm} and the radial part in a set of orthogonal radial basis functions $g_n(r)$ according to

$$\rho_i^{\nu}(\mathbf{r}) = \sum_{nlm} c_{i,nlm}^{\nu} g_n(r) Y_{\text{lm}}(\hat{\mathbf{r}}), \tag{14}$$

where $c_{i,nlm}^{\nu}$ is an expansion coefficient. Here, l ranges from zero to a cut-off value l_{max} and m ranges from $-l$ to l. As radial basis functions, the modified spherical Bessel functions of the first kind are used and n ranges from zero to a cut-off value n_{max}.

The similarity kernel equation (13) has the appealing property that the integrals can be carried out analytically yielding [18]

$$k(\rho_i, \rho_j) = \sum_{\nu \leqslant \nu'} \sum_{nn'l} d_{i,nn'l}^{\nu,\nu'} d_{j,nn'l}^{\nu,\nu'} \tag{15}$$

$$d_{i,nn'l}^{\nu,\nu'} = \sum_{m} c_{i,nlm}^{\nu} (c_{i,n'lm}^{\nu'})^*. \tag{16}$$

From the above expressions, it can be seen that the set of coefficients $\{d_{i,nn'l}^{\nu,\nu'}\}$ plays the role of a descriptor vector $\mathbf{d}_i$ for the environment of atom i. In practice, we calculate the descriptor vectors using the Quippy software package [50]. The kernel matrix $k(\rho_i, \rho_j)$ is then calculated according to

$$k(\rho_i, \rho_j) = \mathbf{d}_i \cdot \mathbf{d}_j. \tag{17}$$

Finally, we note that the sensitivity of the kernel to differences between atomic environments can be increased by defining the effective SOAP kernel [18, 49]

$$K(\rho_i, \rho_j) = \left(\frac{k(\rho_i, \rho_j)}{\sqrt{k(\rho_i, \rho_i) k(\rho_j, \rho_j)}} \right)^{\epsilon}. \tag{18}$$

In this work, we use $\epsilon = 2$.

2.3. Learning cluster polarizabilities

The SOAP descriptor allows the comparison of different environments of a given atom. However, the polarizability is calculated for an entire molecule or cluster consisting of many atoms. To harness the SOAP approach for the prediction of cluster polarizabilities, it is therefore necessary to relate atomic properties to cluster properties. One way to achieve this is by expressing the polarizability α_I of cluster I as the sum of atomic contributions α_i [49] according to

$$\alpha_I = \sum_{i=1}^{N_I} \alpha_i, \tag{19}$$

where N_I denotes the total number of atoms in the cluster. While the atomic contributions can provide some valuable intuition about the dielectric response of complex clusters, it is important to stress that these quantities are not directly measurable and should be interpreted with care [51].

Using standard kernel ridge regression, the atomic polarizabilities can be expressed as

$$\alpha_i = \sum_{j}^{N_{\text{train}}} K_{ij}\zeta_j, \tag{20}$$

where N_{train} denotes the total number of atoms in the training set (i.e. the total number of atoms contained in all training set clusters), ζ_j is a coefficient obtained from training the SOAP model, and $K_{ij} \equiv K(\rho_i, \rho_j)$. Inserting equation (20) into equation (19) yields

$$\alpha_I = \sum_{i}^{N_I}\sum_{j}^{N_{\text{train}}} K_{ij}\zeta_j = \sum_{j}^{N_{\text{train}}} K_{I,j}^{\text{sum}}\zeta_j, \tag{21}$$

where we defined the sum kernel $K_{I,j}^{\text{sum}} = \sum_i^{N_I} K_{ij}$.

Determining the coefficients ζ_j is difficult as the fit to the calculated cluster polarizabilities is strongly underdetermined (as the number of coefficients is the total number of atoms of all clusters in the training set). To make progress, the number of coefficients must be reduced. Intuitively, this should be possible as the atomic environments of many atoms in the training set are very similar. Practically, this sparsification is achieved by means of a singular value decomposition (SVD) of the descriptor matrix $\mathbf{D}$ whose rows contain the descriptor vectors from equation (16). Specifically, $\mathbf{D}$ is expressed as

$$\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{22}$$

where $\mathbf{U}$ and $\mathbf{V}^T$ contain the right and left singular vectors, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values. If many environments in $\mathbf{D}$ are similar, only a few singular values will have large magnitudes. We only retain those singular values which are larger than a given threshold and use the corresponding left singular vectors (which form a matrix $\tilde{\mathbf{V}}$) as a new basis to represent $\mathbf{D}$.

The elements of the SOAP kernel $\tilde{\mathbf{K}}$ corresponding to this new set of effective descriptors are obtained by projecting the descriptors $\mathbf{d}_i$ onto the rows $\tilde{\mathbf{v}}_j$ of the truncated matrix of singular vectors $\tilde{\mathbf{V}}$ according to

$$\tilde{K}_{ij} = \mathbf{d}_i \cdot \tilde{\mathbf{v}}_j. \tag{23}$$

Next, the effective sum kernel $\tilde{\mathbf{K}}^{\text{sum}}$ can be calculated using equation (21), but now the number of coefficients ζ_i is equal to the number of singular vectors whose singular values exceed the threshold. Finally, the vector of coefficients ζ is obtained from [49]

$$\zeta = \left[\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} + (\tilde{\mathbf{K}}^{\text{sum}})^T\mathbf{\Lambda}^{-1}\tilde{\mathbf{K}}^{\text{sum}}\right]^{-1}(\tilde{\mathbf{K}}^{\text{sum}})^T\mathbf{\Lambda}^{-1}\boldsymbol{\alpha}, \tag{24}$$

where $\mathbf{\Lambda} = \lambda\mathbf{I}$ with λ being a regularization parameter and $\boldsymbol{\alpha}$ denotes the vector of calculated cluster polarizabilities.

Alternatively, the cluster polarizability can be expressed as the number of silicon atoms multiplied by their average polarizability α^{av} (note that in this definition α^{av} also contains the smaller contribution from the hydrogen atoms)

$$\alpha = N_{\text{Si}}\alpha_{\text{Si}}^{\text{av}}. \tag{25}$$

To calculate the average polarizability, we average the SOAP kernel matrix over environments belonging to pairs of clusters [51]

$$K_{IJ}^{\mathrm{av}} = \frac{1}{N_I N_J} \sum_{i}^{N_I} \sum_{j}^{N_J} K_{ij}. \tag{26}$$

Using kernel ridge regression, the average polarizability of the silicon atoms in a given cluster is expressed as

$$\alpha_{\mathrm{Si}}^{\mathrm{av}} = \sum_{J}^{n_{\mathrm{train}}} K_J^{\mathrm{av}} \zeta_J, \tag{27}$$

where n_{train} denotes the number of clusters in the training set and the vector of coefficients ζ_I is determined by

$$\zeta = (\mathbf{K}^{\mathrm{av}} + \mathbf{\Lambda})^{-1} \boldsymbol{\alpha}^{\mathrm{av}}, \tag{28}$$

where $\boldsymbol{\alpha}^{\mathrm{av}}$ is the vector containing the average polarizabilities per atom of the training set clusters. As a consequence of the averaging, no additional sparsification procedure is required as in the case of the sum kernel.

This method has the advantage that the average polarizability can be written as a sum of atomic contributions, which allows one to assign polarizabilities to individual atoms. This can be achieved by omitting the average over the index i in equation (26), which yields a prediction for each silicon atom in a cluster

It is interesting to note that the polarizability obtained from the average kernel can also be expressed as a sum of atomic contributions given by

$$\alpha_i = \frac{1}{N_J} \sum_{J}^{n_{\mathrm{train}}} \sum_{j}^{N_J} K_{ij} \zeta_J. \tag{29}$$

Apart from the scaling factor $1/N_J$, the last equation is very similar to equation (20) of the sum kernel approach, with the additional constraint that the coefficients ζ_j on atoms in a cluster J are all equal, $\zeta_j = \zeta_J \; \forall j \in J$. The effect of the scaling factor is that while the sum kernel is extensive (its magnitude scales with the number of atoms in the cluster), the average kernel is intensive, independent of system size. As a consequence of this, large clusters get more heavily weighted in the solution of the least squares problem, equation (24), compared with that for the average kernel.

Finally, we also use the 'coherent average' kernel (denoted 'coh'), which is obtained as follows. Rather than computing a SOAP descriptor for each atomic environment, as in equation (15), we take the spherical harmonic coefficients c_{nlm}^{ν} and average them first to obtain, for cluster I

$$\bar{c}_{I,nlm}^{\nu} = \frac{1}{N_I} \sum_{i=1}^{N_I} c_{i,nlm}^{\nu}, \tag{30}$$

and then square these to form the averaged descriptor vector $\bar{\mathbf{d}}_I$ with components,

$$\bar{d}_{I,nn'l}^{\nu,\nu'} = \sum_{m} \bar{c}_{I,nlm}^{\nu} (\bar{c}_{I,n'lm}^{\nu'})^*, \tag{31}$$

and the coherent (unnormalized) kernel between clusters I and J as

$$k^{\mathrm{coh}}(I,J) = \bar{\mathbf{d}}_I \cdot \bar{\mathbf{d}}_J. \tag{32}$$

2.4. Physical models

We also use two simple physical-based models to fit the calculated RPA polarizabilities. In the first approach, the cluster polarizability is assumed to be proportional to the number of silicon atoms N_{Si} in the cluster, i.e.

$$\alpha = \alpha_{\mathrm{Si}}^{\mathrm{av}} N_{\mathrm{Si}}, \tag{33}$$

with $\alpha_{\mathrm{Si}}^{\mathrm{av}}$ denoting the average polarizability per silicon atom (again, any contributions from hydrogen atoms is included in $\alpha_{\mathrm{Si}}^{\mathrm{av}}$ in this definition). In contrast to the SOAP fitting with the average kernel, the average polarizability is assumed to be the same for all clusters.

The second model is a bond polarizability approach where the cluster polarizability is expressed as a sum of contributions from Si–Si bonds and Si–H bonds according to

$$\alpha = \alpha_{\mathrm{Si-Si}} N_{\mathrm{Si-Si}} + \alpha_{\mathrm{Si-H}} N_{\mathrm{Si-H}}, \tag{34}$$

where $\alpha_{\mathrm{Si-H}}$ and $\alpha_{\mathrm{Si-Si}}$ are the polarizabilities of Si–H and Si–Si bonds, respectively, and $N_{\mathrm{Si-Si}}$ and $N_{\mathrm{Si-H}}$ are the number of Si–Si and Si–H hydrogen bonds, respectively. While this model explicitly includes the contribution of the hydrogen atoms, it is also assumed that the bond polarizabilities are independent of the cluster size and shape.

In a hydrogenated silicon cluster with only sp^3 bonding, the number of Si–Si bonds and Si–H bonds can be expressed in terms of the number of silicon and hydrogen atoms. In particular, $N_{\mathrm{Si-Si}}$ is given by $(4N_{\mathrm{Si}} - N_{\mathrm{H}})/2$, and $N_{\mathrm{Si-H}}$ is equal to N_{H}. Substituting these expressions into equation (34) yields

$$\alpha = \frac{4N_{\mathrm{Si}} - N_{\mathrm{H}}}{2} \alpha_{\mathrm{Si-Si}} + N_{\mathrm{H}} \alpha_{\mathrm{Si-H}}. \tag{35}$$

Dividing both sides by N_{Si} yields the polarizability per silicon atom

$$\frac{\alpha}{N_{\mathrm{Si}}} = 2\alpha_{\mathrm{Si-Si}} + \left(\alpha_{\mathrm{Si-H}} - \frac{\alpha_{\mathrm{Si-Si}}}{2}\right) \frac{N_{\mathrm{H}}}{N_{\mathrm{Si}}}. \tag{36}$$

Interestingly, this shows that the polarizability per silicon atom is a function of the ratio of hydrogen and silicon atoms only.

2.5. Generation of clusters

To generate atomic structures of hydrogenated silicon clusters we follow a similar procedure as Barnard and Wilson [52] who carve spherical clusters from a perfect silicon crystal, terminate the dangling bonds on the surface with hydrogen atoms and then relax the atomic positions using DFT. Unfortunately, this approach only yields very few clusters with 100 or less silicon atoms. Because of the relatively large computational cost associated with the RPA polarizability calculations, we instead use the following approach to generate clusters: starting from the spherical $Si_{123}H_{100}$ cluster, we remove silicon atoms from the surface, terminating any dangling bonds with hydrogen atoms and relax the structure with DFT. In this way, a set of 100 hydrogenated silicon clusters containing between 10 and 110 Si atoms is obtained for which RPA polarizabilities are calculated. In addition, we include the spherical clusters with less than 123 Si atoms from Barnard and Wilson [52].

2.6. Computational details

The plane-wave/pseudopotential DFT code Quantum Espresso [53, 54] was used to obtain Kohn–Sham energies ϵ_n and wavefunctions $\phi_n(\mathbf{r})$. We employed the PBE exchange-correlation functional, norm-conserving pseudopotentials from the original Quantum Espresso Pseudopotential library [53, 54] and a plane-wave cut-off of 65 Ry. The clusters were placed in a cubic unit cell with sufficient vacuum to avoid interactions between periodically repeated images. Next, cluster polarizabilities were calculated with BerkeleyGW [46, 47] using a plane-wave cutoff of 6 Ry and a truncated Coulomb interaction. A total of 600 Kohn–Sham states were included in the summation for χ which was found to be sufficient to converge the scalar polarizabilities. SOAP descriptors were constructed with $l_{\mathrm{max}} = 9$ and $n_{\mathrm{max}} = 20$ and $\gamma_\nu = 2.0$ for $r_{\mathrm{cut}} \leqslant 10.0$ Å and $\gamma_\nu = 0.5$ for $r_{\mathrm{cut}} > 10.0$ Å. In all calculations, we only study local environments of silicon atoms. As all hydrogen atoms are bonded to silicon atoms, their contribution to the cluster polarizabilities can be captured indirectly through their influence on the silicon atoms.

3. Results and discussion

3.1. Fitting polarizabilities

Figure 1(a) shows the RPA polarizabilities of the hydrogenated silicon clusters as function of the number of silicon atoms in the cluster. We observe that the polarizability exhibits a linear behaviour which suggests that the Si atoms provide the dominant contribution.

Deviations from the linear behaviour become explicit when the cluster polarizability is divided by the number of silicon atoms, see figure 1(b). For clusters containing more than 80 Si atoms, the polarizability per

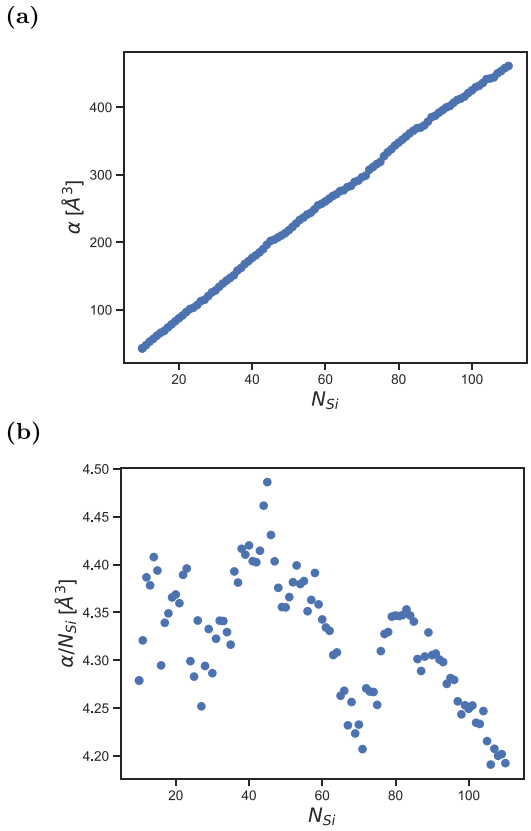

Figure 1. (a) RPA scalar polarizability α of hydrogenated silicon clusters versus number of silicon atoms N_{Si}. The polarizability increases approximately linearly with N_{Si}. (b) RPA polarizability divided by N_{Si} shows deviations from linearity.

silicon atom decreases. Interestingly, α/N_{Si} increases for cluster between 70 and 80 silicon atoms, but decreases again for clusters between 40 and 70 silicon atoms. For clusters with less than 40 Si atoms, there is a significant amount of scatter in the polarizabilities but overall α/N_{Si} tends to increase with increasing number of Si atoms. Overall, the polarizability per silicon atom has an M-like shape as function of the number of silicon atoms.

For very large clusters, α/N_{Si} should converge to the atomic RPA polarizability of bulk silicon which is 3.77 Å^3 (determined using the Clausius–Mossotti relation using a bulk dielectric constant of 12.2 [43]). This explains the observed decrease of α/N_{Si} for $N_{\mathrm{Si}} > 80$. Note that in our results the bulk value is not approached from below because we have not removed the hydrogen contributions from the cluster polarizabilities [35, 36].

To understand these findings, we first compare our results to two physical-based models: a model in which the cluster polarizability is assumed to be proportional to the number of Si atoms (denoted the linear N_{Si} model) and a bond polarizability model (see Methods). The parameters of both models were fitted to the calculated RPA data using a least squares optimization. The results are shown in figure 2(a). While the linear N_{Si} model cannot capture any dependence of α/N_{Si} on the number of silicon atoms, the bond polarizability model correctly describes several key features. In particular, it shows a decreasing trend for large clusters and a minimum near $N_{\mathrm{Si}} = 70$. For small clusters, the bond polarizability model predicts an increase in polarizability as the number of Si atoms is reduced in disagreement with the RPA data. Interestingly, the bond polarizability model also features a significant scatter for small clusters. As discussed in the methods section, α/N_{Si} in the bond polarizability model only depends on the ratio of hydrogen and silicon atoms $N_{\mathrm{H}}/N_{\mathrm{Si}}$ suggesting that this parameter is an important effective descriptor of the hydrogenated silicon clusters.

While the bond polarizability model captures several features, we note that neither of the two physical models can capture the full M-shape of the polarizability per Si atom in figure 2(a). Furthermore, from the least square fits of the linear N_{Si} model to the RPA data, we find $\alpha_{\mathrm{Si}}^{\mathrm{av}} = 4.29$ Å^3. This is significantly larger than the RPA value in bulk Si of 3.77 Å^3. The parameters of the bond polarizability model are found to be $\alpha_{\mathrm{Si-Si}} = 1.98$ Å^3 and $\alpha_{\mathrm{Si-H}} = 1.32$ Å^3. As the polarizability per Si atom is $2\alpha_{\mathrm{Si-Si}}$, the predicted bulk value is 3.96 Å^3 which is in better agreement with RPA results.

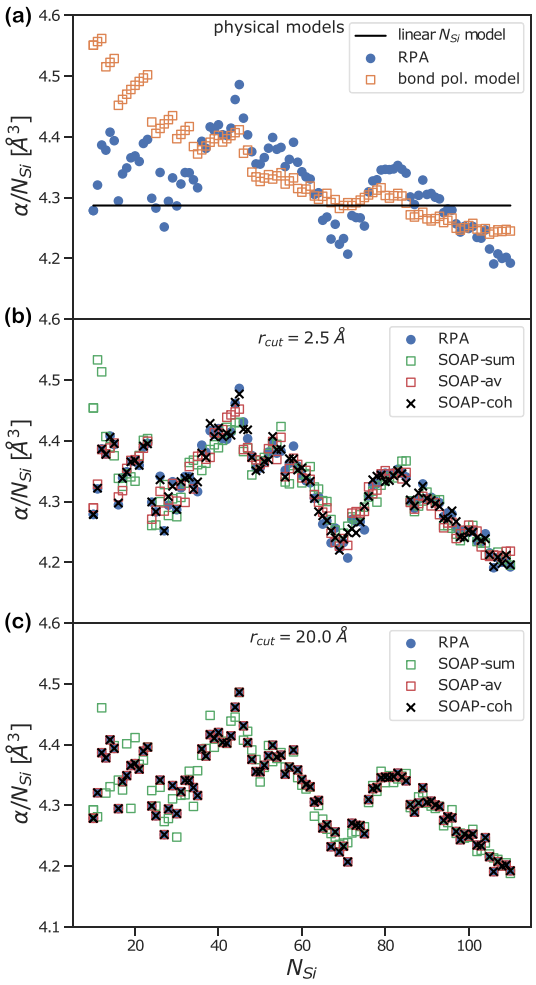

Figure 2. Comparison of α/N_{Si} (cluster polarizability divided by the number of silicon atoms) from RPA calculations (blue dots) with fits from various models. (a) The black line shows the optimal fit for the linear N_{Si} model, see equation (33). Orange triangles denote results from the bond polarizability model, see equation (34). (b) Results obtained using SOAP with the sum kernel (green squares), the average kernel (red squares) and the coherent kernel (black crosses). A cut-off of 2.5 Å was employed. (c) Same as (b) but with a cut-off of 20 Å.

The above analysis demonstrates that both physical-based models have several shortcomings. This is a consequence of two factors: (a) their parameters do not depend on the properties of the local chemical environment, i.e. bond lengths or bond angles. In particular for small clusters, significant atomic relaxations occur resulting in changes to the bond polarizabilites compared to the larger clusters which are not captured by the bond polarizability model. (b) The models do not capture the effects of interactions between the polarizable units. As a consequence, they cannot distinguish between clusters containing the same numbers of Si and H atoms and do not capture the dependence of the polarizability on the cluster shape. To overcome these problems, we now explore the ability of machine learning models to describe the polarizabilities of Si clusters.

Figures 2(b) and (c) show the results from the machine learning model using both the sum kernel, the average kernel and the coherent kernel (see methods). The real space cutoff that determines the size of the chemical environment of each atom is $r_{\mathrm{c}} = 2.5$ Å in figure 2(b) and $r_{\mathrm{c}} = 20$ Å in figure 2(c). In the fit, the regularization parameter λ was kept small (10^{-15} for the sum kernel model and 10^{-12} for the average and coherent kernels) in order to allow as much flexibility in the parameters as possible. For the smaller cut-off (where only nearest neighbour atoms are included in the local environment), all three kernels provide an improved description compared to the physical-based models. Specifically, they capture the M-shape of α/N_{Si} as function of N_{Si} and also reproduce the scatter for smaller clusters. The coherent kernel is slightly better than the averaged kernel, and significant deviations from the calculated polarizabilities are only observed for the smallest cluster sizes when the sum kernel is used. When r_{c} is increased to 20 Å, the agreement between the ML models and the calculated polarizabilities is significantly improved. In particular,

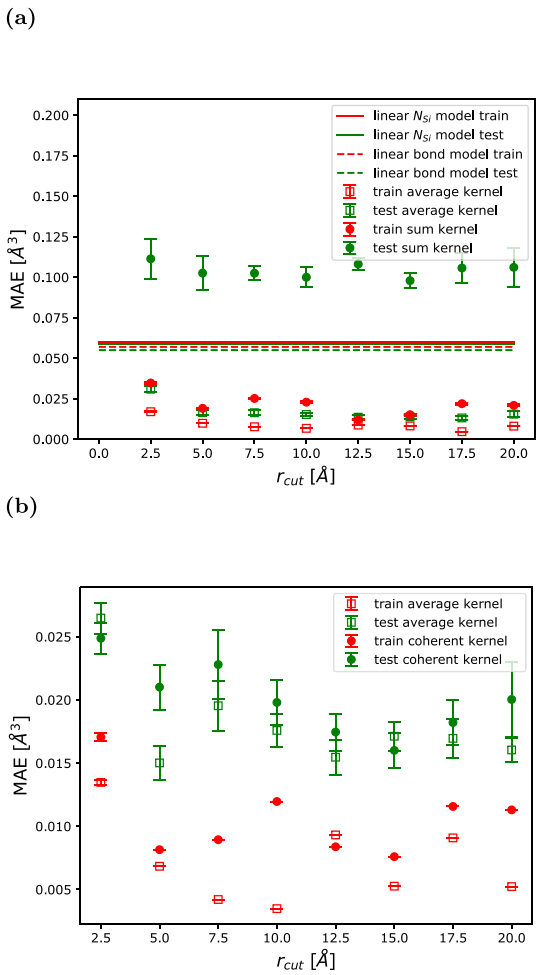

Figure 3. Average mean absolute error (MAE) of α/N_{Si} (the cluster polarizability divided by the number of silicon atoms) from various machine learning models and physical-based models versus the cut-off radius r_{cut} that determines the size of the local chemical environment. Optimal regularization parameters were determined using five-fold cross-validation. Error bars indicate the standard error of the average MAE across the five training and validation sets used in the cross validation procedure.

the results from the average and the coherent kernel are in almost perfect agreement with the data, while the sum kernel results show small deviations for smaller clusters. The good results obtained for the short cutoff indicate that polarizabilities are dominated by local chemical effects. However, long-range interactions also influence polarizabilities and this is captured when the cutoff radius is increased.

3.2. Predicting polarizabilities

Up to this point, we only considered the ability of the SOAP approach to fit the calculated cluster polarizabilities. To investigate SOAP's capacity to predict polarizabilities of clusters that it was not trained on, we use k-fold cross validation [55]. In this procedure, the clusters in the data set are randomly assigned to five sub-sets. Next, four sub-sets are used to train the ML approach and the fifth sub-set is used as the test set. This is done five times with each sub-set acting as test set once. We optimize the regularization parameter λ to minimize the mean average error (MAE). The optimal parameters are listed in the appendix. The resulting MAE and its standard deviation as function of r_{cut} are shown in figure 3(a). The average kernel and the coherent kernel yield very similar results and are compared in figure 3(b). Strikingly, the sum kernel model produces the largest MAE for the test set among all methods. In particular, the test set MAE is significantly larger than the training set MAE indicating poor capacity to predict polarizabilities. In contrast, the average kernel model yields the smallest test set MAE which is only slightly worse than the training set error. The coherent kernel model yields slightly worse predictions than the average kernel, with the biggest difference between the two occurring at $r_{\mathrm{cut}} = 5.0$ Å. The MAE of the two physical-based models lies between those of the sum kernel and the average kernel. The different performances of the sum kernel and the average kernels originate from the different training procedures: the sum kernel model is trained on total cluster polarizabilities, while the average kernel is trained on the average polarizability per silicon atom, see equation (28). As a consequence, the sum kernel model is biased towards more accurate predictions for large

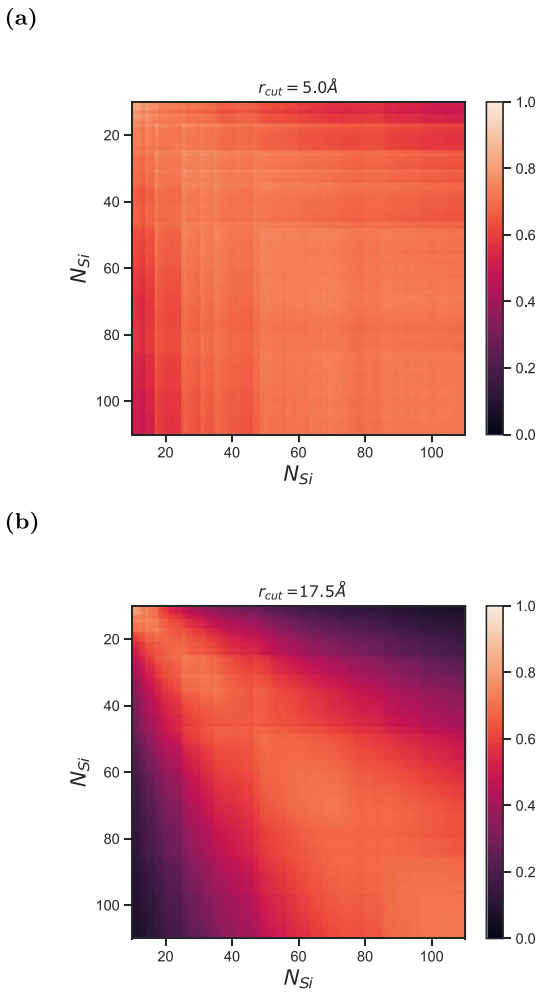

Figure 4. Matrix elements of the average SOAP kernel K_{IJ}^{av}, see equation (26), for cut-offs r_{cut} of 5.0 Å (a) and 17.5 Å (b). Note the rapid decay of the matrix elements along rows and columns when a large cut-off radius is used.

clusters and is less accurate for small clusters. This can also be seen in figure 2(c) which shows that the quality of the sum kernel fit improves for larger clusters. This has been observed before by Stocker *et al* [56] who argued that the intensive average kernel has the advantage of equally weighting small and large molecules, which is beneficial when learning quantities over a large range of cluster sizes. Interestingly, the average kernel performs somewhat better than the coherent kernel suggesting that a model of the cluster polarizability that can be expressed as a sum of atomic contributions constitutes a better representation of the system's dielectric response.

Figure 3 also shows that the minimum test set MAE for the average kernel and the coherent kernel is obtained around $r_{\text{cut}} = 12.5$ Å, while for the sum kernel the minimum is achieved for $r_{\text{cut}} = 15.0$ Å. Interestingly, neither kernel benefits significantly from increasing r_{cut} beyond 5 Å. To understand this finding, we compare the elements of the average kernel matrix for $r_{\text{cut}} = 5.0$ Å and $r_{\text{cut}} = 17.5$ Å, see figure 4. For the smaller cutoff, the kernel matrix decays slowly along the rows and columns of the kernel matrix. In contrast, the decay is significantly more pronounced for the larger cutoff suggesting that a smaller cutoff facilitates the recognition of similar chemical environments in clusters of different size. This is not surprising because for large cutoffs the chemical environment contains a significant amount of vacuum for small clusters, but not for large clusters.

Next, we explore the ability of the ML approach to predict polarizabilities of large clusters based on a training set of small clusters. For this, we train the average kernel on the 60, 70 or 80 smallest clusters and then predict the polarizabilities of the remaining large clusters in the data set. Figure 5 shows the resulting test set MAE as function of the cutoff radius. All curves exhibit a minimum at small cut-offs near $r_{\text{cut}} = 5$ Å and the smallest MAE is obtained for the largest training set. For the smaller training sets ($n_t = 60$ or 70) the MAE increases rapidly as the cutoff is increased, while for the largest training set the increase is mild (and another minimum is found at $r_{\text{cut}} = 17.5$ Å). Similar to our findings in the k-fold cross validation, this shows that it is not beneficial to increase the cut-off radius beyond a certain value.

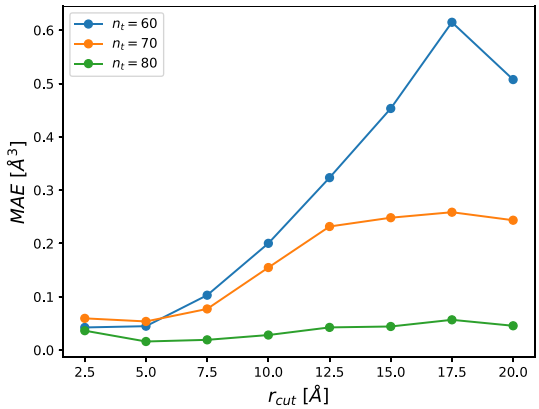

Figure 5. Average kernel test set error as a function of SOAP cut-off. The smallest n_t clusters were included in the training set for each curve. The test set consists of the remaining $100 - n_t$ clusters.

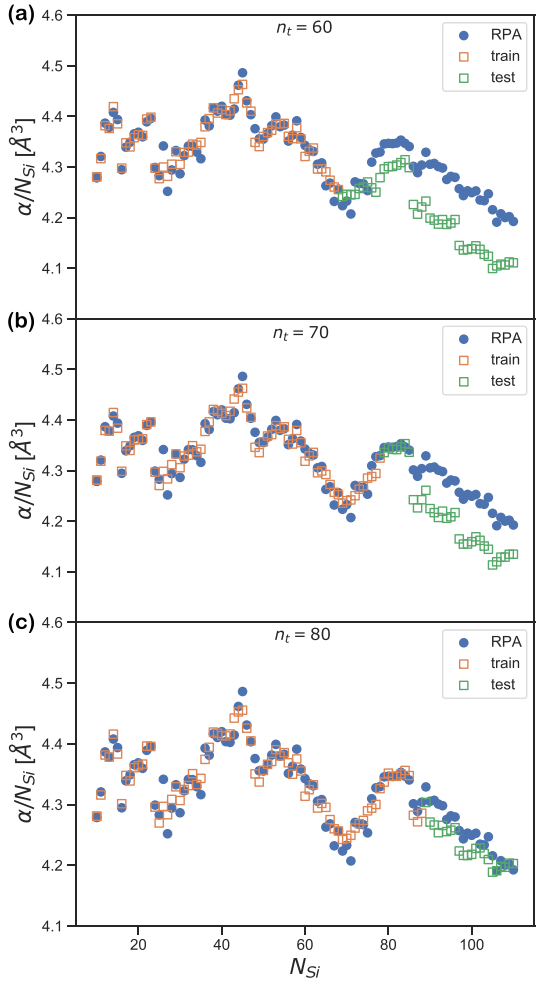

Figure 6. Comparison of RPA results for α/N_{Si} (cluster polarizability divided by the number of silicon atoms) and training and test set predictions of the average kernel model. The training set consists of the (a) $n_t = 60$, (b) $n_t = 70$ and (c) $n_t = 80$ smallest clusters and the test set contains the remaning $100 - n_t$ large clusters.

Figures 6(a)–(c) compare the predictions of the average kernel with $r_{\mathrm{cut}} = 5$ Å with the calculated RPA polarizabilities per silicon atom. For all three training set sizes, the ML model captures the qualitative trends. For $n_t = 60$, the average kernel correctly predicts the increase of α/N_{Si} at $N_{\mathrm{Si}} = 70$ and also the decrease starting at $N_{\mathrm{Si}} = 80$. While the ML models underestimate the polarizabilities per Si atom for large clusters when $n_t = 60$ and $n_t = 70$, good quantitative agreement is achieved for $n_t = 80$.

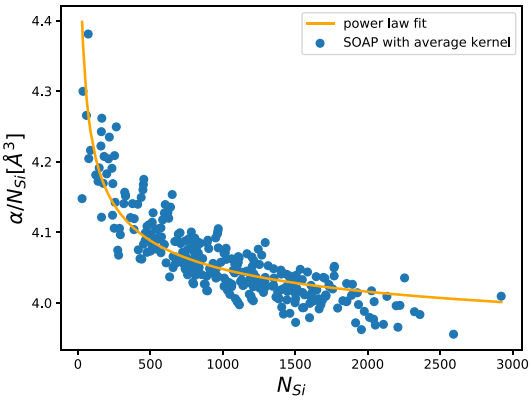

Figure 7. Cluster polarizability divided by the number of silicon atoms for all clusters of Silicon Quantum Dot database [52] from the average SOAP kernel model with $r_{\text{cut}} = 5.0$ Å.

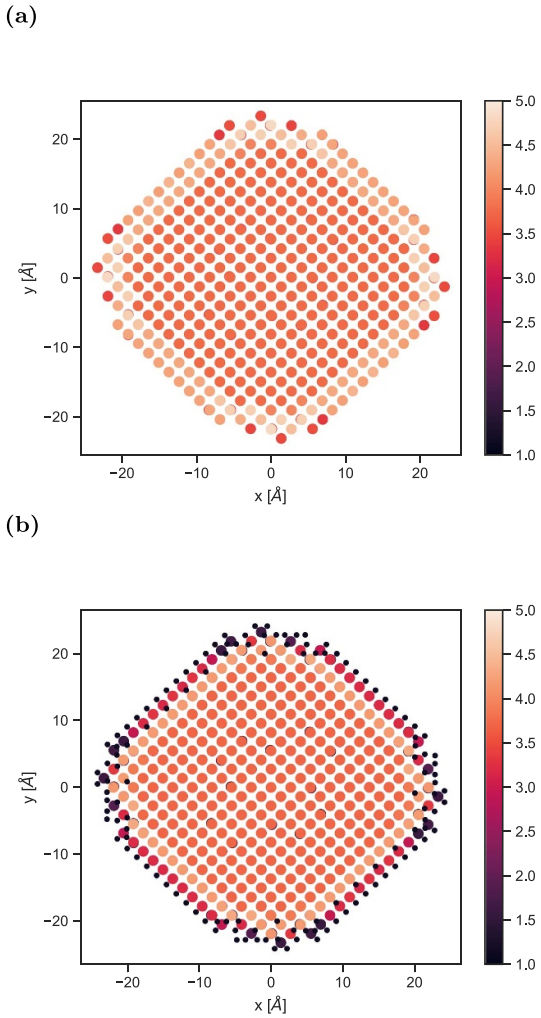

Figure 8. Atomic polarizabilities of the $Si_{2109}H_{604}$ cluster obtained from the SOAP average kernel method. Shown is a cross section through the center of the cluster. (a) Atomic polarizabilities when only silicon chemical environments are used. (b) Atomic polarizabilities when both silicon and hydrogen chemical environments are used. For hydrogen environments $r_{\text{cut}} = 1.6$ Å was used and for silicon environments $r_{\text{cut}} = 5.0$ Å was used. Large dots represent silicon atoms and small dots represent hydrogen atoms.

Finally, we train the average kernel model on the entire data set (using $r_{\text{cut}} = 5$ Å) and predict the average polarizabilities of the entire Silicon Quantum Dot data set containing clusters with up to 3000 silicon atoms [52]. The results are shown in figure 7. It can be observed that the polarizability per Si atom converges slowly to its bulk limit as N_{Si} increases and there is significant scatter in the results. The scatter in α/N_{Si} reflects

the different N_H/N_{Si} ratios and different environments present in the clusters. To understand the slow convergence to the bulk value, note that the number of silicon atoms scales with the cluster volume, while the number of hydrogen atoms is roughly proportional to the surface area. This suggests that α/N_{Si} should be proportional to the inverse radius of the cluster or, equivalently, to $1/N_{Si}^{1/3}$. Indeed, figure 7 shows that the ML predictions are well described by the function $a + b/N_{Si}^{1/3}$ with $a = 3.89$ Å^3 and $b = 1.55$ obtained from a least-squares fit. The value of a agrees well with the RPA atomic polarizability of bulk silicon of 3.77 Å^3 [43].

Additional insights can be obtained by analyzing the atomic polarizabilities obtained from the SOAP average kernel method, see equation (29). Figure 8 shows the atomic polarizabilities of a $Si_{2109}H_{604}$ cluster. In figure 8(a) only local chemical environments of silicon atoms are considered (and the effect of the hydrogen atoms is captured indirectly through their influence on the silicon chemical environments). Silicon atoms in the center of the cluster have a polarizability of 3.76 Å^3, in excellent agreement with value extracted from bulk calculations of 3.77 Å^3 [43]. The polarizability of the silicon atoms in the two surface layers is larger, sometimes as large as 5 Å^3. The reason for this increase is that the surface silicon atoms are bonded to hydrogen atoms and their atomic polarizability is effectively the sum of the silicon and hydrogen contributions. To disentangle contributions from silicon and hydrogen atoms to the cluster polarizability, figure 8 shows the atomic polarizabilities from a calculation that explicitly takes chemical environments of hydrogen atoms into account. Interestingly, the results suggest that the atomic polarizability of subsurface silicon atoms is larger than the bulk value, but the polarizability of surface silicon atoms (which are bonded to hydrogens) is smaller. The average atomic polarizability of the silicon atoms is found to be 3.63 Å^3. This is in agreement with the results of Mochizuki *et al* [36], who predicted that the bulk limit of the silicon atomic polarizability is approached from below.

4. Conclusions

In this work, we have demonstrated that machine learning models based on the SOAP descriptor can be used to accurately and efficiently predict polarizabilities of large hydrogenated silicon clusters. Using the random phase approximation, we calculated the polarizabilities of a set of hydrogenated silicon clusters containing between 10 and 110 silicon atoms. We then assessed the ability of three machine learning models (one using the sum kernel, one using the average kernel and one the coherent kernel) to fit the calculated polarizabilities and find that all three models perform well when the local environment includes nearest neighbour atoms only. Increasing the size of the environment improves the quality of the fit. Next, we investigated the ability of the machine learning models to predict polarizabilities of clusters that are not in the training set. Using k-fold cross validation, we find that the average kernel performs significantly better than the sum kernel and that the predictions only weakly depend on the size of the chemical environment. We also tested the predictive power of the average kernel when it is trained on small clusters only and find that quantitative accuracy can be achieved if the training set is sufficiently large. Finally, we use the average kernel approach to predict the polarizabilities of hydrogenated silicon atoms with up to 3000 silicon atoms and find that the results approach the correct bulk limit. The ability to efficiently calculate polarizabilities of large clusters paves the way towards using machine learning for excited-state properties of these systems. For example, the static density-density response function (from which the polarizability is calculated) is a key ingredient for calculating quasiparticle properties within the GW approach (typically when used in conjunction with a generalized plasmon-pole model) and also for calculating optical properties by solving the Bethe–Salpeter equation. Symmetry-adapted kernel regression could be used to straightforwardly generalise our models to predict the full polarisability tensor [20].

Data availability statement

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

Acknowledgments

This work was supported through a studentship in the Centre for Doctoral Training on Theory and Simulation of Materials at Imperial College London funded by the EPSRC (EP/L015579/1). We acknowledge the Thomas Young Centre under Grant Number TYC-101. This work used the ARCHER UK National Supercomputing Service (www.archer.ac.uk), and the Imperial College London High-Performance Computing Facility.

Appendix

Table 1. Regularization paramaters λ_{av} and λ_{sum} determined from k-fold cross validation [55] at different cut-off radii r_{cut}.

r_{cut} (Å)	λ_{av}	λ_{sum}	λ_{coh}
2.5	10^{-8}	10^{-8}	10^{-8}
5.0	10^{-5}	0.0001	10^{-6}
7.5	10^{-5}	0.01	10^{-5}
10.0	0.0001	0.01	0.0001
12.5	10^{-5}	10^{-6}	10^{-5}
15.0	10^{-5}	10^{-5}	10^{-5}
17.5	0.0001	0.001	0.0001
20.0	0.0001	0.001	0.0001

ORCID iDs

Mario G Zauchner https://orcid.org/0000-0002-0901-5642
Stefano Dal Forno https://orcid.org/0000-0002-9869-7306

References

[1] Wang Y Q, Wang Y G, Cao L and Cao Z X 2003 High-efficiency visible photoluminescence from amorphous silicon nanoparticles embedded in silicon nitride *Appl. Phys. Lett.* **83** 3474
[2] Curtis I S, Wills R J and Dasog M 2021 Photocatalytic hydrogen generation using mesoporous silicon nanoparticles: influence of magnesiothermic reduction conditions and nanoparticle aging on the catalytic activity *Nanoscale* **13** 2685
[3] Park J-H, Gu L, von Maltzahn G, Ruoslahti E, Bhatia S N and Sailor M J 2009 Biodegradable luminescent porous silicon nanoparticles for *in vivo* applications *Nat. Mater.* **8** 331
[4] O'Farrell N, Houlton A and Horrocks B R 2006 Silicon nanoparticles: applications in cell biology and medicine *Int. J. Nanomed.* **1** 451
[5] Fu Y, Dutta A, Willander M and Oda S 2000 Carrier conduction in a Si-nanocrystal-based single-electron transistor-I. Effect of gate bias *Superlattices Microstruct.* **28** 177
[6] Onida G, Reining L and Rubio A 2002 Electronic excitations: density-functional versus many-body Green's-function approaches *Rev. Mod. Phys.* **74** 601
[7] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics without the electrons *Phys. Rev. Lett.* **104** 136403
[8] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, Von Lilienfeld O A, Müller K-R and Tkatchenko A 2015 Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space *J. Phys. Chem. Lett.* **6** 2326
[9] Li Z, Kermode J R and De Vita A 2015 Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces *Phys. Rev. Lett.* **114** 096405
[10] Faber F A, Lindmaa A, Von Lilienfeld O A and Armiento R 2016 Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals *Phys. Rev. Lett.* **117** 135502
[11] Brockherde F, Vogt L, Li L, Tuckerman M E, Burke K and Müller K-R 2017 Bypassing the Kohn–Sham equations with machine learning *Nat. Commun.* **8** 872
[12] Alred J M, Bets K V, Xie Y and Yakobson B I 2018 Machine learning electron density in sulfur crosslinked carbon nanotubes *Compos. Sci. Technol.* **166** 3
[13] Grisafi A, Fabrizio A, Meyer B, Wilkins D M, Corminboeuf C and Ceriotti M 2019 Transferable machine-learning model of the electron density *ACS Cent. Sci.* **5** 57
[14] Chandrasekaran A, Kamal D, Batra R, Kim C, Chen L and Ramprasad R 2019 Solving the electronic structure problem with machine learning *npj Comput. Mater.* **5** 22
[15] Jain A *et al* 2013 The materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[16] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies *npj Comput. Mater.* **1** 15010
[17] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD) *JOM* **65** 1501
[18] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[19] Wilkins D M, Grisafi A, Yang Y, Lao K U, DiStasio R A and Ceriotti M 2019 Accurate molecular polarizabilities with coupled cluster theory and machine learning *Proc. Natl Acad. Sci.* **116** 3401
[20] Grisafi A, Wilkins D M, Csányi G and Ceriotti M 2018 Symmetry-adapted machine learning for tensorial properties of atomistic systems *Phys. Rev. Lett.* **120** 036002
[21] Tuan-Anh T and Zaleśny R 2020 Predictions of high-order electric properties of molecules: can we benefit from machine learning? *ACS Omega* **5** 5318
[22] Veit M, Wilkins D M, Yang Y, DiStasio R A and Ceriotti M 2020 Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles *J. Chem. Phys.* **153** 024113
[23] Povarnitsyn M E, Shcheblanov N S, Ivanov D S, Yu Timoshenko V and Klimentov S M 2020 Vibrational analysis of silicon nanoparticles using simulation and decomposition of Raman spectra *Phys. Rev. Appl.* **14** 014067

[24] Wang J, Cieplak P, Li J, Hou T, Luo R and Duan Y 2011 Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability *J. Phys. Chem.* B **115** 3091
[25] Wang Z-X, Zhang W, Wu C, Lei H, Cieplak P and Duan Y 2006 Strike a balance: optimization of backbone torsion parameters of amber polarizable force field for simulations of proteins and peptides *J. Comput. Chem.* **27** 781
[26] Vasiliev I, Öğüt S and Chelikowsky J R 1997 Ab initio calculations for the polarizabilities of small semiconductor clusters *Phys. Rev. Lett.* **78** 4805
[27] Deng K, Yang J and Chan C T 2000 Calculated polarizabilities of small Si clusters *Phys. Rev.* A **61** 025201
[28] Bazterra V E, Caputo M C, Ferraro M B and Fuentealba P 2002 On the theoretical determination of the static dipole polarizability of intermediate size silicon clusters *J. Chem. Phys.* **117** 11158
[29] Jackson K, Pederson M, Wang C-Z and Ho K-M 1999 Calculated polarizabilities of intermediate-size Si clusters *Phys. Rev.* A **59** 3685
[30] Jackson K A, Yang M, Chaudhuri I and Frauenheim T 2005 Shape, polarizability and metallicity in silicon clusters *Phys. Rev.* A **71** 033205
[31] Maroulis G and Pouchan C 2003 Assessing the performance of *ab initio* methods on static (hyper)polarizability predictions for silicon clusters. Si_4 as a test case *Phys. Chem. Chem. Phys.* **5** 1992
[32] Maroulis G, Begué D and Pouchan C 2003 Accurate dipole polarizabilities of small silicon clusters from *ab initio* and density functional theory calculations *J. Chem. Phys.* **119** 794
[33] Papadopoulos M G, Reis H, Avramopoulos A, Erkoç Ş and Amirouche L 2006 Polarizabilities and second hyperpolarizabilities of Zn_m Cd_n clusters *Mol. Phys.* **104** 2027
[34] Fetter A L and Walecka J D 2003 *Quantum Theory of Many-Particle Systems* (*Dover Books on Physics*) (New York: Dover)
[35] Jansik B, Schimmelpfennig B, Norman P, Mochizuki Y, Luo Y and Ågren H 2002 Size, order and dimensional relations for silicon cluster polarizabilities *J. Phys. Chem.* A **106** 395
[36] Mochizuki Y and Ågren H 2001 Polarizability of silicon clusters *Chem. Phys. Lett.* **336** 451
[37] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials *Chem. Rev.* **121** 9759
[38] Honrao S J, Xie S R and Hennig R G 2020 Augmenting machine learning of energy landscapes with local structural information *J. Appl. Phys.* **128** 085101
[39] Choudhary K, DeCost B and Tavazza F 2018 Machine learning with force-field-inspired descriptors for materials: fast screening and mapping energy landscape *Phys. Rev. Mater.* **2** 083801
[40] Rohrhofer F M, Saha S, Di Cataldo S, Geiger B C, von der Linden W and Boeri L 2021 Importance of feature engineering and database selection in a machine learning model: a case study on carbon crystal structures (arXiv:2102.00191)
[41] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401
[42] Faber F A, Christensen A S, Huang B and von Lilienfeld O A 2018 Alchemical and structural distribution based representation for universal quantum machine learning *J. Chem. Phys.* **148** 241717
[43] Hybertsen M S and Louie S G 1987 *Ab initio* static dielectric matrices from the density-functional approach. I. Formulation and application to semiconductors and insulators *Phys. Rev.* B **35** 5585
[44] Adler S L 1962 Quantum theory of the dielectric constant in real solids *Phys. Rev.* **126** 413
[45] Wiser N 1963 Dielectric constant with local field effects included *Phys. Rev.* **129** 62
[46] Hybertsen M S and Louie S G 1986 Electron correlation in semiconductors and insulators: band gaps and quasiparticle energies *Phys. Rev.* B **34** 5390
[47] Deslippe J, Samsonidze G, Strubbe D A, Jain M, Cohen M L and Louie S G 2012 BerkeleyGW: a massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures *Comput. Phys. Commun.* **183** 1269
[48] Calaminici P, Jug K and Köster A M 1998 Density functional calculations of molecular polarizabilities and hyperpolarizabilities *J. Chem. Phys.* **109** 7756
[49] Ceriotti M, Willatt M J and Csányi G 2018 Machine learning of atomic-scale properties based on physical principles *Handbook of Materials Modeling: Methods: Theory and Modeling* ed W Andreoni and S Yip (Cham: Springer International Publishing) pp 1–27
[50] Kermode J 2021 (available at: https://warwick.ac.uk/fac/sci/eng/staff/jrk) (Accessed 3 June 2020)
[51] Bartók A P, De S, Poelking C, Bernstein N, Kermode J R, Csányi G and Ceriotti M 2017 Machine learning unifies the modeling of materials and molecules *Sci. Adv.* **3** e1701816
[52] Barnard A and Wilson H 2015 *Silicon Quantum Dot Data Set. v2* (CSIRO) (https://doi.org/10.4225/08/5721BB609EDB0)
[53] Giannozzi P *et al* 2009 QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials *J. Phys.: Condens. Matter.* **21** 395502
[54] Giannozzi P *et al* 2017 Advanced capabilities for materials modelling with QUANTUM ESPRESSO *J. Phys.: Condens. Matter.* **29** 465901
[55] Rasmussen C E and Williams C K I 2005 *Gaussian Processes for Machine Learning Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press) (Available at: https://doi.org/10.7551/mitpress/3206.003.0008)
[56] Stocker S, Csányi G, Reuter K and Margraf J T 2020 Machine learning in chemical reaction space *Nat. Commun.* **11** 5505