How to cite: *Angew. Chem. Int. Ed.* **2024**, e202403842
doi.org/10.1002/anie.202403842

Amorphous Materials *Hot Paper*

Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning

Joe D. Morrow, Chinonso Ugwumadu, David A. Drabold, Stephen R. Elliott, Andrew L. Goodwin, and Volker L. Deringer**

Abstract: The structure of amorphous silicon (a-Si) is widely thought of as a fourfold-connected random network, and yet it is defective atoms, with fewer or more than four bonds, that make it particularly interesting. Despite many attempts to explain such "dangling-bond" and "floating-bond" defects, respectively, a unified understanding is still missing. Here, we use advanced computational chemistry methods to reveal the complex structural and energetic landscape of defects in a-Si. We study an ultra-large-scale, quantum-accurate structural model containing a million atoms, and thousands of individual defects, allowing reliable defect-related statistics to be obtained. We combine structural descriptors and machine-learned atomic energies to develop a classification of the different types of defects in a-Si. The results suggest a revision of the established floating-bond model by showing that fivefold-bonded atoms in a-Si exhibit a wide range of local environments–analogous to fivefold centers in coordination chemistry. Furthermore, it is shown that fivefold (but not threefold) coordination defects tend to cluster together. Our study provides new insights into one of the most widely studied amorphous solids, and has general implications for understanding defects in disordered materials beyond silicon alone.

Introduction

Amorphous silicon (a-Si) is the textbook example of a disordered material, with a structure that overall resembles Zachariasen's concept of a continuous random network (CRN) of covalently bonded atoms.[1–3] Many interesting properties and phenomena relating to a-Si have been discussed over the years: the hyperuniform nature of the disordered network,[4] the transition between the high-coordinated metallic liquid and the fourfold-connected structure of a-Si,[5] its complex phase behavior under pressure,[6–8] and a tension–compression asymmetry in its mechanical properties.[9] Many of those phenomena originate on the atomic scale, and fully understanding their origins in terms of chemical structure and bonding has long been a central research goal.

While the overall structure of a-Si is based on fourfold-connected atoms ($N=4$, where N denotes the number of bonded nearest neighbors), there is particular interest in those "defective" atoms that have fewer ($N=3$) or more ($N=5$) neighbors. In well-relaxed structural models of a-Si, N is easily determined: there is a clear minimum in the radial distribution function, separating the first (bonded) peak from the second (non-bonded) one, at about 2.85 Å. Defining neighbors, and therefore coordination defects, is straightforward if the minimum in the distribution approaches zero; it becomes more ambiguous otherwise, which is most relevant to $N=5$ atoms. There is experimental evidence—for example, from spectroscopy—for the presence of point defects in a-Si, emphasizing that its structure is more nuanced than a simplified all-fourfold CRN description.[10–15]

Computer simulations have long played a key role in studying amorphous networks, and a-Si has been a prominent example.[16] In the 1980s, models of a-Si were created via melt-quench molecular dynamics (MD) with the empirical Stillinger–Weber potential.[17] The large concentration of coordination defects in the models obtained this way (≈20 %) permitted a discussion of the average structure of $N=5$ defects, as well as an analysis of the energetics predicted by the potential. Later studies, with increasingly fast computers, extended system sizes to hundreds of thousands of atoms with empirical potentials.[18–20] Equally, quantum-mechanically based (density-functional theory, DFT) MD studies on much smaller systems have provided important insights.[22–24] However, predictive DFT simulations of a-Si beyond the few-nm length scale remain out of

[*] J. D. Morrow, Prof. A. L. Goodwin, Prof. V. L. Deringer
Inorganic Chemistry Laboratory
Department of Chemistry
University of Oxford, Oxford OX1 3QR, United Kingdom
E-mail: andrew.goodwin@chem.ox.ac.uk
volker.deringer@chem.ox.ac.uk

C. Ugwumadu, Prof. D. A. Drabold
Department of Physics and Astronomy
Nanoscale and Quantum Phenomena Institute (NQPI)
Ohio University, Athens, Ohio 45701, United States

Prof. S. R. Elliott
Physical and Theoretical Chemistry Laboratory
Department of Chemistry
University ofOxford, OX1 3QZ, United Kingdom

reach, due to the long simulation times required and the cubic scaling of computational cost with system size. For example, a current "optimal" DFT-MD-based a-Si structure contains 215 atoms,[23] and the current limit for such types of MD simulations appears to be about 1,000 atoms.[24] There is debate over the relative abundance of $N=3$ and $N=5$ defects, with spectroscopic evidence ambiguous[25,26] and computational methodologies, such as ab initio MD[21] and empirical potentials,[27] tending to predict more 5-fold defects.

Recent developments in machine-learning (ML) based interatomic potentials have made it possible to prepare realistic, DFT-accurate structural models of a-Si of much larger size,[8,28–31] with a published million-atom model reaching a cell length of about 27 nm.[30] We have previously established that slow quenching from the simulated melt using ML potentials yields structural models whose characteristics agree well with existing experimental data.[8,28,30] There is increasing evidence that the local, per-atom, energy predictions obtainable from atomistic ML (but not normally from DFT) are amenable to post hoc chemical interpretation: for example, we have shown that ML atomic energies in amorphous graphene can be used both to drive structural exploration and to analyze the resulting structures.[32] Neural-network models have been interrogated with regards to local energy contributions as well.[33–35] A recent study has shown that the density of ML atomic-energy contributions can yield insights into complex solid-state ion conductors.[36]

In a previous Communication in the present journal, we have shown that ML atomic energies can be used to discriminate three- and five-coordinated atoms in a-Si.[29] We now build upon those pilot studies but expand on them vastly—for example, by analyzing a simulation cell containing a million atoms, of which several thousand are defective. We identify three structural prototypes for over-coordinated atoms in a-Si, allowing us to develop a "taxonomy" of defects in this canonical disordered material, and we explain the tendency for fivefold defects to aggregate via the strain that they induce on their atomic neighbors. Beyond silicon, our study paves the way for routine quantum-accurate, million-atom scale, ML-driven simulations of rare events such as defect formation in functional materials.

Results and Discussion

Our simulations of a-Si are based on a computational approach which we call "indirect learning" (Figure 1a),[30] and which corresponds to teacher–student models for knowledge distillation that are more commonly used in ML research. The approach has been validated in our previous, more technical work,[30,38] and we use it here to set the stage for an in-depth analysis of the defects in a-Si. Indirect learning involves using an accurate, trusted, but computationally slow ML potential (the teacher model) to train a second, much faster one (the student model). The latter enables simulations of accuracy and reliability on a par with that of the teacher (of the order of 10 meV per atom for amorphous Si vs. DFT), whilst requiring much less computational time, by a factor of about 1,000 in this case. A visual comparison between 100,000-atom and 1 M-atom models, drawn to scale in Figure 1b, illustrates the advantage of the approach. The structure factor and radial distribution functions are in very close agreement amongst experiment, teacher-derived, and student-derived structural models and the energetics are similar between student and teacher structural models for defects and 4-fold atoms (Figure S1). Million-atom simulation cells are necessary to systematically study defects that occur at the few-percent level, and, crucially, to investigate "second-order" processes, such as the interaction of defects with one another. Here, the simulation cell contains tens of thousands of examples of threefold- and fivefold connected defective atoms (Figure 1c). This simulation was affordable only because of the increased computational efficiency of the student model.

Perhaps the most serious assumption made with most current ML potentials is that the total energy of a system of atoms (a quantum-mechanical observable) can be decomposed into a sum of local atomic energies (which are not observables).[39,40] The major benefit of making this locality assumption is the ability to "machine-learn" energetics independent of the system size, as well as the resulting linear scaling of computational cost with the number of simulated atoms. Although initially conceived for this purpose alone, there is increasing evidence that local energies have a physical relevance that can be useful for analysis beyond the mere construction of ML potentials,[29,36,41] and that their robustness can be quantified.[39] We note that the idea of considering local contributions to the total energy in silicon has been pioneered based on empirical potentials,[17] and our present ML-based approach extends this type of thinking and places it on a more quantitative, DFT-accurate basis.

Figure 1d shows the ML atomic energies in the 1 M-atom model from Ref. [30], separated according to coordination numbers for the central atom. It confirms our earlier findings that $N=3$ atoms have distinctly higher energies on average than do $N=5$ ones.[29] Note, however, that the distributions in Figure 1d result from a much larger dataset than in Ref. [29] and therefore do not include any broadening.

Figure 1e presents an analysis of the defects in a-Si from an alternative, and somewhat orthogonal, purely structural perspective, by examining the distances to individual neighbors for each atom in the structure. As the coordination number of the central atom increases from 3 to 5, so too do the distances between bonded atoms, as is reasonable from a chemical perspective. The well-defined 5-th peak, and the even spacing between peaks for the 5-fold distribution, indicate that a majority of such defects are best described as truly 5-coordinated, rather than alternatively as [4+1] with a mostly non-bonded 5-th neighbor; the latter is what one would expect in tetrahedral amorphous carbon.[46] The longer tail observed for the $n=5$ peak contains those more distant 5-th neighbors that indeed are in this minor group of [4+1] environments. The typical bond lengths for $N=3$ and $N=5$ defects in our a-Si model differ strongly from those for the non-defective 4-fold atoms in the rest of the structure, consistent with the observation that defects

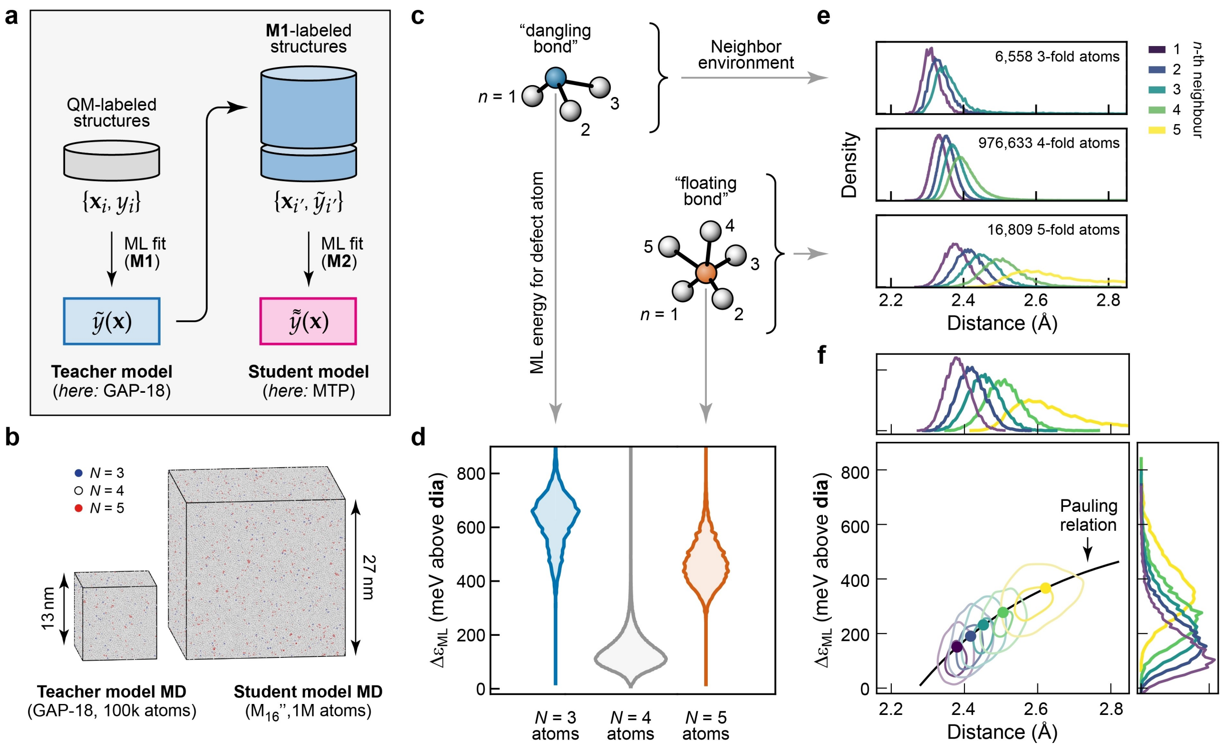


Figure 1. Defects in amorphous silicon from million-atom simulations. (**a**) Schematic of the teacher–student approach to ML potential fitting described in Ref. [30]. A reliable, but relatively slow teacher model (**M1**) is used to generate a set of many more structures, to which the more specialized, but faster student model (**M2**) is then fitted. (**b**) Visualization of 100,000- (*left*) and million-atom (*right*) structural models of a-Si, drawn to scale for direct comparison. Panels (a) and (b) are adapted from Ref. [30], which was originally published under a CC BY licence (https://creativecommons.org/licenses/by/4.0/). (**c**) Schematic drawing of 3- and 5-fold-connected atoms, with neighboring atoms numbered in the order of their distance from the central atom. (**d**) Distributions of the ML-predicted atomic energies, $\varepsilon_{\mathrm{ML}}$, of defects in the million-atom model, shown separately for different nearest-neighbor coordination numbers, N. All values are referenced to crystalline diamond-type silicon, which is set as the energy zero. (**e**) Neighbor distributions, resolved according to 3-, 4-, and 5-fold-connected central atoms (separate axes) and their individual immediate neighbors, sorted by distance. (**f**) 2D correlation plot of the neighbor density for 5-coordinated defects (as in panel e) versus energy (as in panel d), given separately for the immediate neighbors ($n=1$–5, purple to yellow). The black line describes a fit to a Pauling-like relation,[37] $\log(\ n(E)\ ) = A(r_{\mathrm{c}} - r)$, where $n(E)$ is the bond strength defined in terms of neighbor local energies as $(E_{\max} - E)/(E_{\max} - E_{\min})$, $r_{\mathrm{c}} = 2.38$ Å is a characteristic radius (taken to be the mean minimum bond length in the structure), r is the bond length, and A is a fitting parameter with optimal value 1.17. Further details on the fitting function and choice of parameters are given in Figure S9.

have significantly higher energies than do most bulk-like atoms (Figure 1d).

In Figure 1f, we examine the correlation between those energetic and structural indicators, *viz.* the local energy of the neighbors of the 5-fold atom on the *y*-axis, and the distance of the respective neighbor from the central atom on the *x*-axis. We find a striking, logarithmic relationship between these two quantities, which shows that the elevated local energy at 5-fold defect centers is also delocalized onto the surrounding atoms with longer bonds to the defect. It is comforting that ML local energies, the physical meaning of which has been a source of debate, reproduce a well-founded result from valence theory: the exponential dependence of bond strength on bond length. The equivalent correlation plots for $N=3$ and $N=4$ atoms (Figure S8) show a much tighter distribution of energies for the respective atomic neighbors.

The presence of some relatively short fifth-neighbor contacts at distances below 2.6 Å (Figure 1e) invites the question whether there is a "typical" geometric structure of 5-fold defects in a-Si. In fact, this question has been studied for a long time. The term "floating bonds" has been used to describe defects consisting of silicon atoms with five bonded neighbors, initially introduced by Pantelides in 1986 as part of the first recognition of the importance of over-coordinated atoms in a-Si.[25] The canonical structure of such a defect was described as a perfect tetrahedron with an extra fifth atom bonded directly opposite to one of the equivalent existing bonds. However, it was already noted that, in the amorphous phase, bond lengths and angles can vary considerably from such an idealized geometry.

In Figure 2, we address this question with reference to the common prototypes for 5-coordinated atoms that are known from structural inorganic chemistry. At the top of Figure 2a, we sketch the trigonal bipyramidal (TBP) and square pyramidal environments that one would expect from the valence shell electron-pair repulsion (VSEPR) model[42,43] that is frequently discussed in undergraduate chemistry textbooks, and also an idealized floating-bond

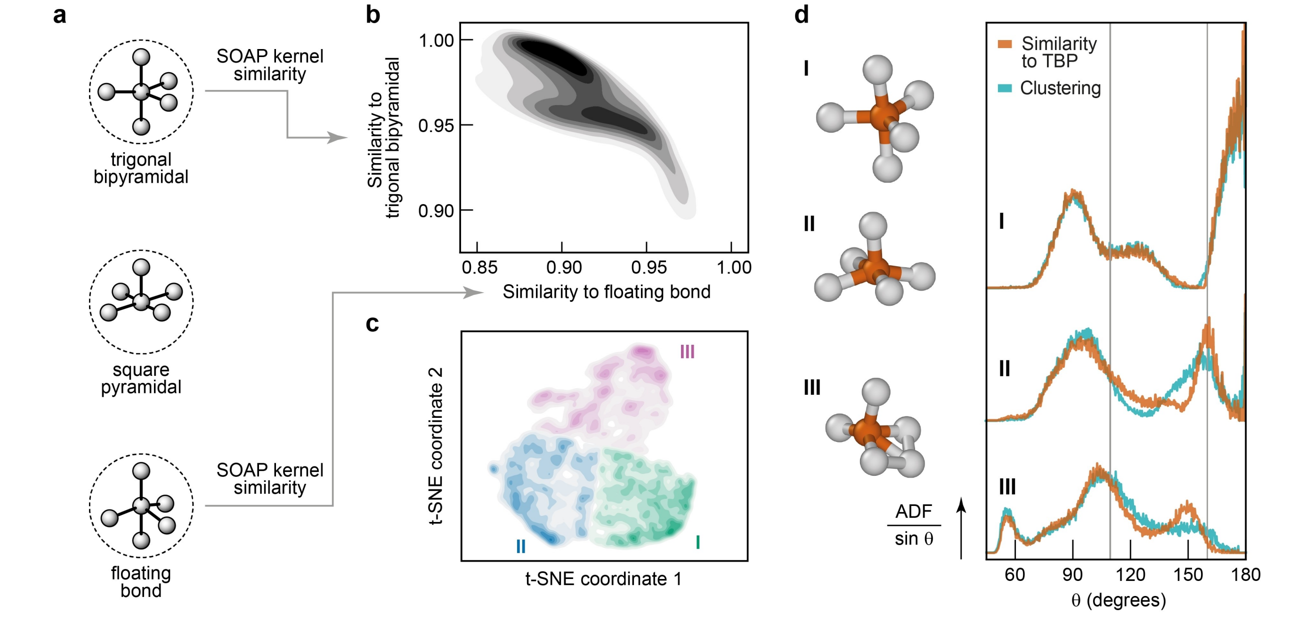


Figure 2. Categories of fivefold-connected defects. (**a**) Schematic drawing of idealized trigonal bipyramidal (TBP), square pyramidal, and "floating-bond" environments. The first two are shown in line with the established valence shell electron-pair repulsion (VSEPR) model;[42,43] the third is a tetrahedral environment with a fifth atom directly opposite a bond.[25] (**b**) 2D plot of the SOAP kernel similarity of $N = 5$ atoms in the a-Si model to the idealized TBP and floating-bond environments, respectively. The distribution of values for individual atoms is shown as a heat map. (**c**) Unsupervised classification of 5-fold atoms. The full distance matrix, $\mathbf{D} = \sqrt{2 - 2\mathbf{K}}$, is embedded in 2D with the dimensionality reduction algorithm t-SNE,[44] where $\mathbf{K}$ is the kernel matrix built from the similarity of each 5-fold atom with every other 5-fold atom. The bisecting *k*-means algorithm is used to identify clusters **I–III**.[45] See Figure S11 for more details. (**d**) Bond-angle distribution function (BADF), scaled by sin θ, for atomic triples centered on all 5-fold coordinated atoms. The BADFs are plotted separately for the three distinct categories of $N = 5$ defects related to idealized structures respectively from top to bottom (as illustrated with selected examples of such configurations from the a-Si model). The 5-fold atoms are separated into these categories via two methods but with similar results: by comparison to the idealized structures of panels **a** and **b** via SOAP similarity (orange) and from unsupervised clustering (cyan). Vertical lines at the tetrahedral angle (109.5°) and at 160° are guides for the eye.

environment with a fifth bond directly opposite one of the bonds in a tetrahedral environment (bottom of Figure 2a).[25]

We took the 5-fold defects in the 1 M-atom structural model and evaluated their structural similarity to the three respective prototypes on a scale of 0 (dissimilar) to 1 (identical) via the Smooth Overlap of Atomic Positions (SOAP) kernel.[47] In this analysis, we aimed to discriminate environments based on their bond angles and their respective similarity to the prototypical VSEPR configurations; we therefore re-scaled all nearest-neighbor distances to 2.5 Å, such that we examine only the angular distribution of atoms around each defect. A multimodal distribution of similarity values results, as displayed in Figure 2b. By dividing this distribution into three regions, we arrive at a classification of the 5-fold defects in a-Si as TBP-like (**I**), square-pyramidal-like (**II**), or "floating bond"-like (**III**).

A very similar classification is obtained using unsupervised ML[48,49] in Figure 2c: dimensionality reduction followed by clustering. In this plot, the distance between points corresponds to their structural dissimilarity as measured by SOAP. The three-lobed distribution, with higher densities of points at the outer edges, supports the identification of exactly three major classes of 5-fold atoms.

We show bond-angle distribution functions (BADFs) and accompanying representative example structures taken from the model in Figure 2d. We note that the BADFs in a-Si have been closely linked to the Raman transverse-optic peak width,[50] to the exponential tails in optical-absorption band edges, and to the through-bond (topological) distance to over- and under-coordinated atoms.[51] Our million-atom simulations allow us to derive finely resolved BADF plots,[30] revealing immediately that the different categories have distinctly different shapes. The BADF for category **I** defects in Figure 2c directly reflects the angles in an idealized TBP environment: 180° between the axial atoms, 120° between equatorial atoms, and 90° between axial and equatorial atoms. The square-pyramidal BADF (category **II**) has a fairly sharp distribution of angles at 160° and a broader peak at ≈100° angles. The designation of a "floating bond" (**III**), maximally dissimilar from the TBP (**I**), is less well-defined structurally, but can be understood as originating from an ideal tetrahedron with an additional 5-th atom approaching a face or edge, as originally suggested.[25] In the BADF, this arrangement manifests in the following features: (i) the occurrence of ≈60° angles between the 5-th atom and those neighbors that it approaches most closely; (ii) a corresponding closing of the ideal tetrahedral angle as these neighbors are displaced away from the 5-th atom; and (iii) larger angles between the 5-th atom and those on the opposite side of the tetrahedron. Fewer than half of all $N=5$ defects adopt a structure similar to the floating-bond description.

The boundaries between the classes of 5-fold defects, however, are somewhat fuzzy and this suggests an overlap of types of defect structures, rather than three entirely separate categories, as expected for such a strongly disordered amorphous structure. This observation is consistent with the low energy barriers between structural minima that are observed for 5-fold complexes in molecular inorganic chemistry, which often exhibit fluxionality.[52–54] Further discussion of where we draw the boundaries between classes, along with partial pair correlation functions for defects, can be found in Figures S2 and S3.

The correlation between bond length and local energy in Figure 1f had already suggested that the mechanical strain associated with 5-fold defects extends over a longer range than the radius of the defect center itself. Figure 3 now provides a more detailed analysis of the atomic energies and their degree of "locality"–that is, their dependence on nearest-, next-nearest-, and further neighbors. The locality of physical properties is of general interest for the development of atomistic ML models, because it determines directly the extent to which information can be captured by finite-range models.[55]

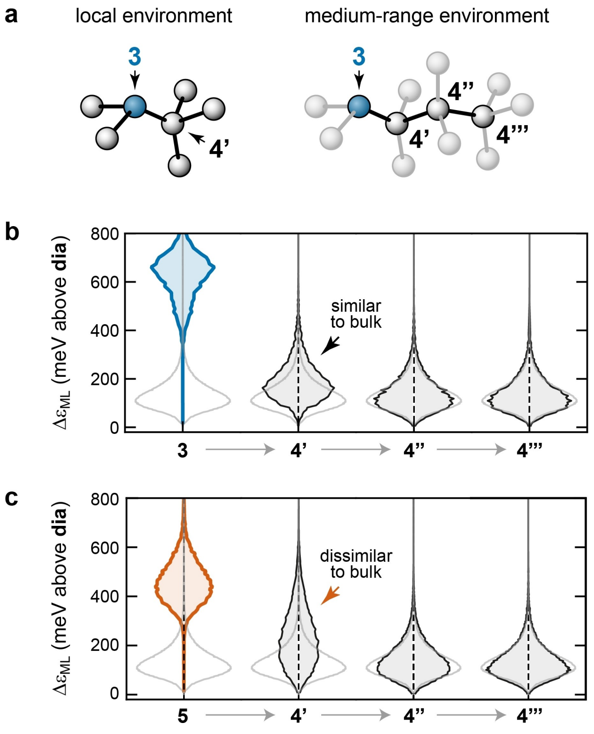


Figure 3. Locality of defect environments. (**a**) Schematic of the labeling system for local and medium-range environments, based on the bond topology. A fourfold atom directly connected to a defect (here, to a 3-coordinate atom) is labeled as 4′, a fourfold atom connected to that one is labeled as 4″, and so on. (**b**) Distribution of ML local energies for $N=3$ atoms and their surroundings. The distribution for bulk a-Si is shown in light gray. (**c**) Same but for $N=5$ atoms and their surroundings.

We introduce some notation to describe the topological neighborhood of defective atoms, as sketched schematically in Figure 3a: 4-fold coordinated atoms that are directly bonded to a defect center (i.e., to an under- or over-coordinated atom) are referred to as 4′; 4-coordinate atoms directly bonded to a 4′ atom, but not to a defect center, are referred to as 4″; those bonded to a 4″ atom, but not to any closer defect, are referred to as 4‴. Atoms which are more topologically distant than the third neighbor shell, 4‴, we merely call "bulk".

Figure 3b indicates how, upon moving away from an $N=3$ defect atom, the atomic-energy distributions quickly approach those of the bulk. The directly adjacent atoms, 4′, have a distribution that is similar to that of the bulk atoms (shown in light grey)—in other words, an $N=3$ atom does not seem to notably affect the atomic energies of its directly bonded neighbors. In contrast, Figure 3c shows that the 4′ atoms connected to an $N=5$ defect are much higher in energy on average. This suggests that the structural disturbance of defects in this case is less localized, and has a longer-range effect. In both cases, the energy distributions for the 4″ atoms and beyond are not notably affected by the presence of nearby defects.

The difference in the locality of the 5-fold and 3-fold defects has an interesting consequence: when the environment of each defect is considered, the locally averaged energy for 3-fold and 5-fold defects is 726 meV and 749 meV above the bulk, respectively, which indicates that 3-fold defects are predicted to have a marginally *lower* formation energy, in contrast to what was found previously for the energies of the defect atoms themselves.[29] Note that these sums take into account the clustering of 5-fold defects described in Figure 5, which reduces the number of 4′ atoms per 5-fold defect.

Why, then, are more 5-fold defects observed in MD, despite their higher energy? This seemingly contradictory result can be rationalized as a consequence of either entropy or kinetics. The 5-fold centers have considerably more flexibility than the relatively rigid pyramidal structure of 3-fold defects, which implies 5-fold defects may have higher entropy and hence a lower *free energy* of formation. 5-fold defects are more similar to the structure of molten Si than are 3-folds, so are likely to be more kinetically accessible during quenching. The difference in the local strain amongst 5-folds and 3-folds also leads us to advise care in interpreting a single local energy in the context of defects, as highlighted previously in the context of amorphous graphene.[32] Local averages, which include physical strain induced by the defect and smooth out the assignment of local energies, appear to be more appropriate in drawing physical interpretations.

The structural and energetic landscapes of defects, even for seemingly simple crystalline materials, can be highly complex, and the exploration and understanding of those defects requires advanced computational techniques. For example, it was shown recently that relaxations of small-scale defect models often determine incorrect (metastable) defect geometries, and that a more comprehensive exploration of the associated structural and energetic landscape is required even for seemingly simple inorganic crystals.[56–58]

ML techniques have begun to be applied to defects in crystalline materials as well.[59–61] Here, we investigate the question whether there is a connection between defects in crystalline silicon, which are well-studied, and the atomic environments of defects in the amorphous phase.

We have previously shown that 2D scatter plots of structural properties (specifically, the "diamond-likeness") and atomic energies are useful for understanding defects in a-Si.[29] We now extend this methodology by using it to compare defects in the amorphous form of silicon to different types of defects in the crystalline phase, as shown in Figure 4. The results are broadly in line with expectations: the 3-fold defects are most similar to vacancies (blue in Figure 4a), 5-fold defects are most similar to crystalline interstitials (orange in Figure 4c), and neither defect is very similar to the perfect crystal, as measured by the SOAP-kernel similarity on the horizontal axis. The bimodal distribution in the similarity of 5-fold defects to vacancies suggests that a minority of such defects, likely with two long bonds, could be considered as being related to vacancies. Note that the interstitial example was produced by equilibrating an idealized 10-fold coordinated interstitial site in diamond-type Si at 500 K. These simulations show that the structure of the crystalline interstitial, like the amorphous 5-fold defect, is highly fluxional and is therefore difficult to understand via even several "typical" structures. Fluxionality of 5-coordinate complexes is common in molecular systems because of the similar energies of different conformations, so it is perhaps unsurprising to observe a similar effect in an extended solid. In turn, this observation highlights the advantage of having many examples of individual defects in the million-atom model available for comparison, as we have discussed in the context of Figure 3.

Our analysis in Figure 4 extends the long-standing earlier assumption that amorphous materials contain structural building blocks of the corresponding crystalline phases.[1,62] We therefore suggest that amorphous phases can be thought of as containing building blocks of crystalline phases *and of defects* in crystals.

Although the majority of over-coordinated defects in a-Si exist as isolated $N=5$ atoms, we find a stronger tendency for those atoms to occur close together than would be expected for a random placement of defects in a CRN—i.e., the clustering that would be observed if 5-fold defects were placed randomly without the influence of energetics. The energy distributions in Figure 5a, now evaluated for defects and their local environments up to $4'''$ sites, show that such clustering reduces the total average energy associated collectively with defect atoms and their immediate neighborhoods. This is largely a result of clustered defects having fewer nearest neighbors per 5-fold defect. Neighbors of defects, on average, have an elevated energy compared to bulk a-Si. To avoid the extensivity of the excess energy with surface area when referred to the ground-state crystal energy, we reference the energies in Figure 5a to the average value for bulk a-Si. In this way, the neighbors of a defect will only increase the energy of the defect cluster if they have an average energy above that of bulk a-Si. Hence, we can converge the prediction of the cluster excess energy, so that it becomes independent of the number of neighbors considered, by including neighbors that are sufficiently distant from the defect center as to be typical of the bulk. In practice, this is achieved by including the $4'''$ atoms, *viz.* $\Delta E=\sum_i(E_i-\bar{E})/n_5$, where i indicates the topological classification of atom i and runs over 5, $4'$, $4''$, and $4'''$; and n_5 is the number of 5-fold atoms in the cluster. The widths of the distributions in Figure 5a get narrower with cluster size because larger clusters have a greater total number of $4'$, $4''$ *etc.* neighbors (although fewer per 5-fold atom). The relationship between probability and cluster size is approximately exponential (Figure 5b), with the exception of clusters of size 2, which are disfavored compared to 3-membered clusters. 3-fold defects are far less likely to cluster than 5-fold defects: only 19 examples of 3–3 bonds

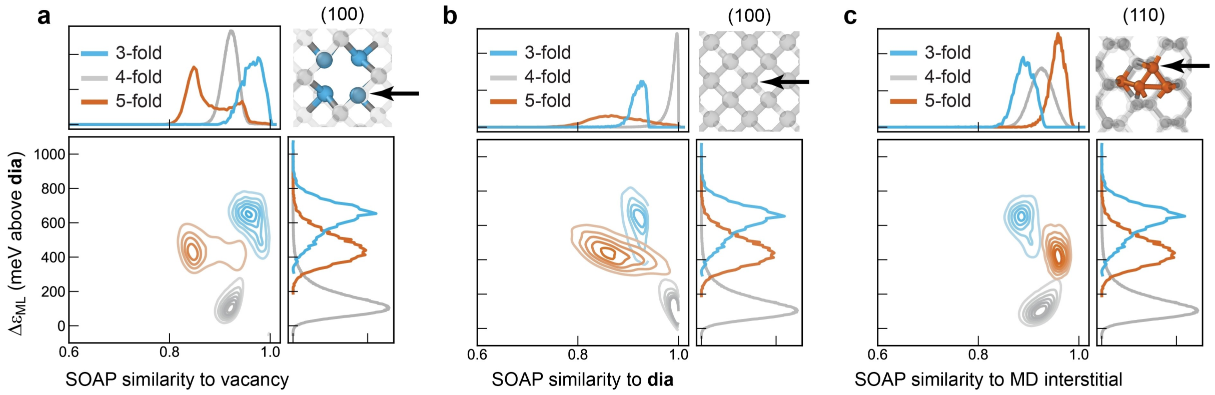


Figure 4. Connection with defects in crystals. (**a**) A 2D SOAP similarity–energy map for the structure obtained at a quench rate of 10^{11} K s^{-1}, analyzing separately 3-fold, 4-fold, and 5-fold coordinated atoms. The horizontal axis shows the structural similarity to the relaxed vacancy; the $N=3$ atoms (*blue*) are structurally the most similar to the vacancy; the $N=5$ atoms (*orange*) the least so. The vertical axis gives the ML atomic energy. The structure is visualized in the top right part, with the viewing direction given, and the arrow indicating the atom used for comparison. (**b**) Same for the perfect crystalline diamond structure, as in Ref. [29]. (**c**) Same for a snapshot from an MD simulation of an interstitial in the crystalline structure.

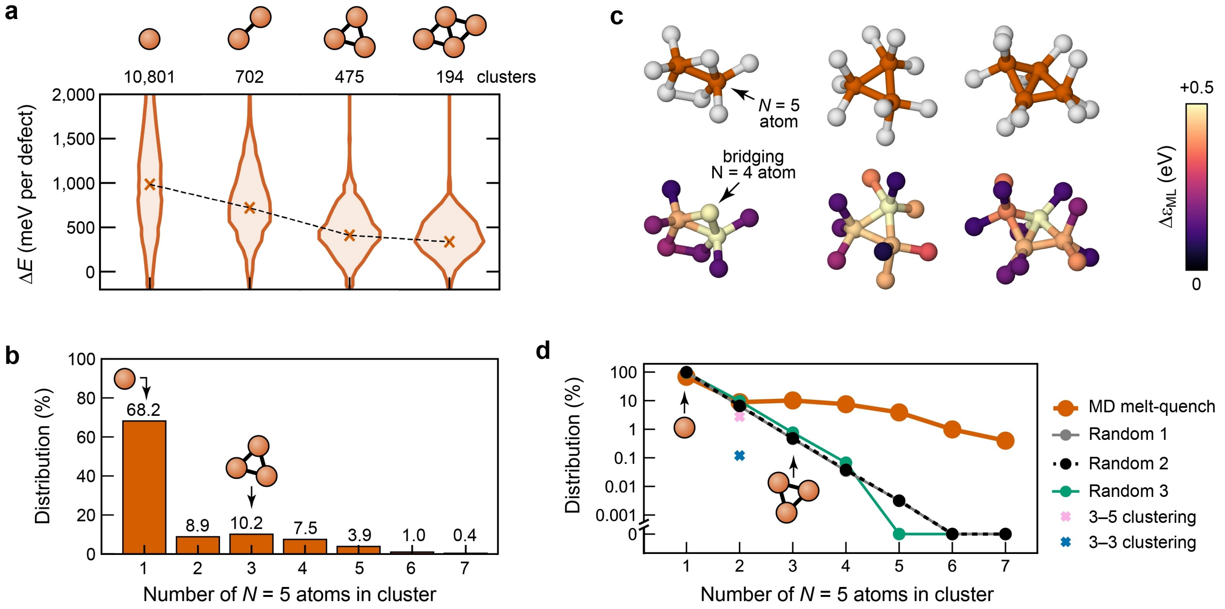


Figure 5. Clustering of defects in amorphous silicon. (**a**) ML-predicted energy distributions for the most common 5-fold defect clusters, with the number of occurrences of each structure directly above each histogram. Defect cluster energies are calculated by summing the individual atomic energies of defect cores and their immediate topological neighbors up to 3 bonds away relative to the mean a-Si energy and are reported per coordination defect, as described in the main text. Crosses indicate the mean of each distribution; the dashed line is a guide for the eye. (**b**) Statistics for the number of occurrences of clustered 5-coordination defects. (**c**) Examples of clustered $N=5$ defects of different sizes, color-coded by coordination numbers of the atoms (*upper row*) and by their ML local energy above diamond-type Si, $\triangle\varepsilon_{ML}$ (*lower row*). Note the high atomic energy of the bridging $N=4$ atom in the left part of the panel, indicated by an arrow. (**d**) Comparison of observed clustering of 5-fold defects in the MD-derived a-Si model (orange, same data as in **b**) with three different idealizations of a random clustering distribution and the occurrence of 3-fold-5-fold bonded pairs (pink cross) and 3-fold-3-fold pairs (blue cross). Protocol 1 (gray) randomly labels 1.68 % of atoms (the 5-fold defect concentration) in a perfect diamond structure as 'defective'. Protocol 2 (black) randomly labels 1.68 % of atoms on the 4-fold network of the a-Si model as 'defective'. Protocol 3 employs a bond-switching algorithm to equilibrate the MD-derived clustering distribution by moving 5-fold centers randomly across the graph formed by atoms (nodes) and bonds (edges). This accounts for an increased likelihood of clustering purely due to the 5 bonds at 5-fold-coordinate atoms compared to the 4 bonds for most atoms. More details of the idealizations are given in the Supporting Information.

were found (cf. Figure S12) and no clusters were found containing more than one 3–3 bond.

A possible qualitative explanation for this difference is provided by the examples in Figure 5c: in the first case, a pair of $N=5$ atoms is connected via a bridging $N=4$ atom which in itself has a rather high atomic energy (highlighted by an arrow), consistent with the strain formed in the associated three-membered ring. In the larger fragments shown in Figure 4c, there are still three-membered rings, but they are now formed by three and four $N=5$ atoms clustering together, respectively. In these cases, the directly-connected $N=4$ atoms have lower energies. The clustering of 5-coordinate defects can be interpreted as causing a reduction in the defective "surface area".

Conclusions

Our analyses support a comprehensive picture of defects in amorphous silicon. On the one hand, 3-fold connected "dangling-bond" defects are high in energy and do not strongly affect their surroundings. On the other hand, 5-fold connected defects are associated with a broad range of possible structures[29] which we can understand, at one extreme, as being similar to a trigonal bipyramid with an even distribution of bond lengths, and at the other extreme, as being similar to the floating-bond description applied previously. 5-fold defects also have an extended influence on their surroundings, reaching beyond their immediate atomic neighbor environment, which explains their observed tendency to cluster together.

Beyond silicon, our study has more general implications for materials modeling. Defects and impurities in solids occur at a concentration of a couple of percent at most, yet they are highly consequential for the performance of functional materials. The structure and chemical bonding of these defects likely differ markedly from the bulk, requiring a quantum-mechanically accurate description of the potential-energy surface associated with defect formation. In this work, we showcased how ML interatomic potentials can be used to study rare events in a systematic way, and how ML-predicted atomic energies can explain the stability of defect environments in one of the canonical amorphous materials, *viz.* a-Si. Qualitatively, this means a step away from idealized CRN models or small-scale simulation systems which (necessarily) contain only a handful of defects, moving towards a fully realistic description of the amorphous state.[63] In the future, we envisage similar ML-driven studies of defects and defect complexes in a wide range of functional materials.

Acknowledgements

J.D.M. acknowledges funding from the EPSRC Centre for Doctoral Training in Inorganic Chemistry for Future Manufacturing (OxICFM), EP/S023828/1. S.R.E. acknowledges financial support from the Leverhulme Trust (UK) in the form of an Emeritus Fellowship. A.L.G. gratefully acknowledges financial support from the E.R.C. (Grant 788144). The authors acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (https://doi.org/10.5281/zenodo.22558), as well as support from the John Fell Oxford University Press (OUP) Research Fund. We are grateful for computational support from the UK national high performance computing service, ARCHER2, for which access was obtained via the UKCP consortium and funded by EPSRC grant ref EP/X035891/1. D.A.D. and C.U. employed BRIDGES-2 at the Pittsburgh Supercomputing Center through allocation DMR190008P and PHY230007P from the Advanced Cyberinfrastructure Coordination Ecosystem, supported by the U. S. National Science Foundation.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data supporting the present study are publicly available at https://doi.org/10.5281/zenodo.10794053.

Keywords: amorphous materials · computational chemistry · coordination defects · machine learning · solid-state structures

[1] W. H. Zachariasen, *J. Am. Chem. Soc.* **1932**, *54*, 3841–3851.

[2] F. Wooten, K. Winer, D. Weaire, *Phys. Rev. Lett.* **1985**, *54*, 1392–1395.

[3] S. R. Elliott, *Physics of Amorphous Materials*, Longman, Harlow, **1990**.

[4] R. Xie, G. G. Long, S. J. Weigand, S. C. Moss, T. Carvalho, S. Roorda, M. Hejna, S. Torquato, P. J. Steinhardt, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13250–13254.

[5] A. Hedler, S. L. Klaumünzer, W. Wesch, *Nat. Mater.* **2004**, *3*, 804–809.

[6] S. K. Deb, M. Wilding, M. Somayazulu, P. F. McMillan, *Nature* **2001**, *414*, 528–530.

[7] P. F. McMillan, M. Wilson, D. Daisenberger, D. Machon, *Nat. Mater.* **2005**, *4*, 680–684.

[8] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, S. R. Elliott, *Nature* **2021**, *589*, 59–64.

[9] Y. Wang, J. Ding, Z. Fan, L. Tian, M. Li, H. Lu, Y. Zhang, E. Ma, J. Li, Z. Shan, *Nat. Mater.* **2021**, *20*, 1371–1377.

[10] J. H. Stathis, *Phys. Rev. B* **1989**, *40*, 1232–1237.

[11] S. Roorda, W. C. Sinke, J. M. Poate, D. C. Jacobson, S. Dierker, B. S. Dennis, D. J. Eaglesham, F. Spaepen, P. Fuoss, *Phys. Rev. B* **1991**, *44*, 3702–3725.

[12] X. Urli, C. L. Dias, L. J. Lewis, S. Roorda, *Phys. Rev. B* **2008**, *77*, 155204.

[13] A. Polman, D. C. Jacobson, S. Coffa, J. M. Poate, S. Roorda, W. C. Sinke, *Appl. Phys. Lett.* **1990**, *57*, 1230–1232.

[14] G. N. van den Hoven, Z. N. Liang, L. Niesen, J. S. Custer, *Phys. Rev. Lett.* **1992**, *68*, 3714–3717.

[15] S. Knief, W. von Niessen, T. Koslowski, *Phys. Rev. B* **1998**, *58*, 4459–4472.

[16] L. J. Lewis, *J. Non-Cryst. Solids* **2022**, *580*, 121383.

[17] W. D. Luedtke, U. Landman, *Phys. Rev. B* **1989**, *40*, 1164–1174, and references therein.

[18] R. Atta-Fynn, P. Biswas, *J. Chem. Phys.* **2018**, *148*, 204503.

[19] D. Dahal, R. Atta-Fynn, S. R. Elliott, P. Biswas, *J. Phys. Conf. Ser.* **2019**, *1252*, 012003.

[20] D. Dahal, S. R. Elliott, P. Biswas, *Phys. Rev. B* **2022**, *105*, 115203.

[21] R. Car, M. Parrinello, *Phys. Rev. Lett.* **1988**, *60*, 204–207.

[22] I. Štich, R. Car, M. Parrinello, *Phys. Rev. B* **1991**, *44*, 11092–11104.

[23] A. Pedersen, L. Pizzagalli, H. Jónsson, *New J. Phys.* **2017**, *19*, 063018.

[24] Y. Xu, Y. Zhou, X.-D. Wang, W. Zhang, E. Ma, V. L. Deringer, R. Mazzarello, *Adv. Mater.* **2022**, *34*, 2109139.

[25] S. T. Pantelides, *Phys. Rev. Lett.* **1986**, *57*, 2979–2982.

[26] D. K. Biegelsen, M. Stutzmann, *Phys. Rev. B* **1986**, *33*, 3006–3011.

[27] R. L. C. Vink, G. T. Barkema, W. F. van der Weg, N. Mousseau, *J. Non-Cryst. Solids* **2001**, *282*, 248–255.

[28] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, G. Csányi, *J. Phys. Chem. Lett.* **2018**, *9*, 2879–2885.

[29] N. Bernstein, B. Bhattarai, G. Csányi, D. A. Drabold, S. R. Elliott, V. L. Deringer, *Angew. Chem. Int. Ed.* **2019**, *58*, 7057–7061.

[30] J. D. Morrow, V. L. Deringer, *J. Chem. Phys.* **2022**, *157*, 104105.

[31] Y. Wang, Z. Fan, P. Qian, M. A. Caro, T. Ala-Nissila, *Phys. Rev. B* **2023**, *107*, 054303.

[32] Z. El-Machachi, M. Wilson, V. L. Deringer, *Chem. Sci.* **2022**, *13*, 13720–13731.

[33] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.

[34] X. Chen, M. S. Jørgensen, J. Li, B. Hammer, *J. Chem. Theory Comput.* **2018**, *14*, 3933–3942.

[35] M. Eckhoff, J. Behler, *J. Chem. Theory Comput.* **2019**, *15*, 3793–3809.

[36] S. Wang, Y. Liu, Y. Mo, *Angew. Chem. Int. Ed.* **2023**, *62*, e202215544.

[37] L. Pauling, *J. Am. Chem. Soc.* **1947**, *69*, 542–553.

[38] J. D. Morrow, J. L. A. Gardner, V. L. Deringer, *J. Chem. Phys.* **2023**, *158*, 121501.

[39] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[40] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.

[41] S. Chong, F. Grasselli, C. B. Mahmoud, J. D. Morrow, V. L. Deringer, M. Ceriotti, *J. Chem. Theory Comput.* **2023**, *19*, 8020–8031.

[42] R. J. Gillespie, R. S. Nyholm, *Q. Rev. Chem. Soc.* **1957**, *11*, 339–380.

[43] R. J. Gillespie, *Chem. Soc. Rev.* **1992**, *21*, 59–69.

[44] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res* **2008**, *9*, 2579–2605.

[45] M. Steinbach, G. Karypis, V. Kumar, *KDD Workshop Text Min.* **2000**, *400*, 525–526.

[46] M. A. Caro, G. Csányi, T. Laurila, V. L. Deringer, *Phys. Rev. B* **2020**, *102*, 174201.

[47] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.

[48] M. Ceriotti, *J. Chem. Phys.* **2019**, *150*, 150901.

[49] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, A. Laio, *Chem. Rev.* **2021**, *121*, 9722–9758.

[50] D. Beeman, R. Tsu, M. F. Thorpe, *Phys. Rev. B* **1985**, *32*, 874–878.

[51] Y. Pan, F. Inam, M. Zhang, D. A. Drabold, *Phys. Rev. Lett.* **2008**, *100*, 206403.

[52] R. S. Berry, *J. Chem. Phys.* **1960**, *32*, 933–938.

[53] E. P. A. Couzijn, J. C. Slootweg, A. W. Ehlers, K. Lammertsma, *J. Am. Chem. Soc.* **2010**, *132*, 18127–18140.

[54] C. Moberg, *Angew. Chem. Int. Ed.* **2011**, *50*, 10290–10292.

[55] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073–10141.

[56] S. R. Kavanagh, A. Walsh, D. O. Scanlon, *ACS Energy Lett.* **2021**, *6*, 1392–1398.

[57] I. Mosquera-Lois, S. R. Kavanagh, *Matter* **2021**, *4*, 2602–2605.

[58] I. Mosquera-Lois, S. R. Kavanagh, A. Walsh, D. O. Scanlon, *npj Comput. Mater.* **2023**, *9*, 25.

[59] N. C. Frey, D. Akinwande, D. Jariwala, V. B. Shenoy, *ACS Nano* **2020**, *14*, 13406–13417.

[60] M. Arrigoni, G. K. H. Madsen, *npj Comput. Mater.* **2021**, *7*, 71.

[61] M. D. Witman, A. Goyal, T. Ogitsu, A. H. McDaniel, S. Lany, *Nat. Comput. Sci.* **2023**, *3*, 675–686.

[62] J. Mavračić, F. C. Mocanu, V. L. Deringer, G. Csányi, S. R. Elliott, *J. Phys. Chem. Lett.* **2018**, *9*, 2985–2990.

[63] C. Chang, V. L. Deringer, K. S. Katti, V. Van Speybroeck, C. M. Wolverton, *Nat. Rev. Mater.* **2023**, *8*, 309–313.

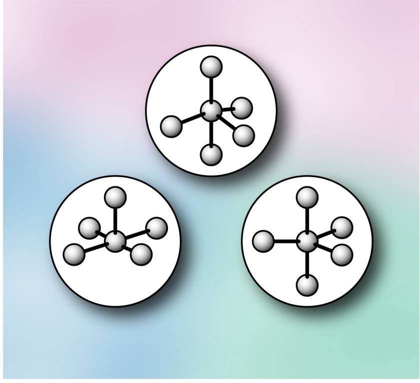
Manuscript received: February 23, 2024
Accepted manuscript online: March 22, 2024
Version of record online: ■■, ■■

Research Articles

Amorphous Materials

J. D. Morrow, C. Ugwumadu, D. A. Drabold, S. R. Elliott, A. L. Goodwin,* V. L. Deringer* ______ e202403842

Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning



Machine-learning methods and computational chemistry are combined to develop a comprehensive understanding of coordination defects in amorphous silicon. Fivefold-connected atoms fall into three categories based on their chemical structure, and are found to cluster together, which can be explained based on energetic arguments.

Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning

J. D. Morrow, C. Ugwumadu, D. A. Drabold, S. R. Elliott, A. L. Goodwin, V. L. Deringer**

Supporting Information for

Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning

Joe D. Morrow,[a] Chinonso Ugwumadu,[b] David A. Drabold,[b] Stephen R. Elliott,[c] Andrew L. Goodwin[a,]* & Volker L. Deringer[a,]*

[a] *Inorganic Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3QR, UK*

[b] *Department of Physics and Astronomy, Nanoscale and Quantum Phenomena Institute (NQPI), Ohio University, Athens, Ohio 45701, United States*

[c] *Physical and Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3QZ, UK*

* andrew.goodwin@chem.ox.ac.uk; volker.deringer@chem.ox.ac.uk

Computational Methods

Teacher–student ML potentials

We generate fast and robust interatomic potential models by using one ML model to "teach" another. This idea is related to knowledge distillation in neural networks, albeit defined more generally. In Ref. [S1], we demonstrated it for a kernel-based method, *viz.* the Gaussian approximation potential (GAP) framework,[S2] as the teacher model, and a linear-fitting technique, viz. the moment tensor potential (MTP) approach,[S3] as the student model. We used the indirectly-learned M_{16}'' model[S1] to generate structures via MD, and we used local energies predicted by the teacher GAP model[S4] for analysis. Using the M_{16}'' local energies gave qualitatively similar results. The difference in total energy between teacher and student methods is 3 meV per atom, with a root-mean-square difference of 48 meV for local energies. This difference in local energies is of a similar size to that achieved by training directly on local energies derived from GAP models for carbon and silicon.[S5,S6]

Molecular dynamics

MD simulations and local-energy evaluations were carried out using the LAMMPS software,[S7] interfaced to the MTP[S3] and GAP[S2] codes, respectively. The timestep was 1 fs. In Ref. [S1], melt–quench simulations were conducted using the same variable-rate protocol as described in Ref. [S8], that is:

1. Melting a random initial structure at 2500 K for 20 ps
2. Equilibrating the resulting liquid at 1500 K for 100 ps
3. Quenching at a variable rate: 10^{13} K s^{-1} between 1500–1250 K, 10^{11} K s^{-1} between 1250–1050 K over which vitrification takes place, back to 10^{13} K s^{-1} between 1050–500 K for a total cooling time of 2.08 ns, before a relaxation with conjugate-gradient (CG) descent.

The same type of analysis of defects performed herein on the structure cooled at 10^{11} K s^{-1} (taken from Ref. [S1]) was performed on a separate 1M-atom a-Si structure produced by annealing a rapidly quenched structure (10^{13} K s^{-1} throughout the temperature range 1500–500 K) at 850 K for 3 ns before cooling back close to 0 K at 10^{13} K s^{-1} and relaxing by CG descent. The results for the annealed structure, which are equivalent to those in the main article, may be found in Figures S6–S7 and show that our conclusions are robust with respect to the precise number of defects and the path to reaching them.

Electronic-structure analysis for validation

We analyzed with DFT the electronic structure of a small-scale (512 atom) structural model made in the same way as the million-atom model to further validate structures produced with the ML potential. A clear energy gap between valence and conduction bands and exponentially shaped Urbach tails are indicators of high-quality structural models of a-Si.[S9] In our model, Urbach tails are evident and dangling bonds produce electronic states near the middle of the gap (Figure S4). Floating bonds in these models, which are thought to play a role in charge-carrier transport,[S10] do not produce localized states in the gap, and as such would not be observed in an electron spin resonance experiment, similar to observations in Ref. [S11]. However, we observe that tail states are distributed among 4′ atoms in the vicinity of the floating bonds (Figure S5). This analysis suggests that only the dangling-bond sites would yield an electronic signature in the gap,[S12] for which the concentration is low compared to other MD-derived structures at 0.7%.[S13] Hence, the electronic properties of the million-atom model are consistent with other high-quality structural models. Further details of the electronic-structure calculations, including analysis of the ultra-large structure with a tight-binding Hamiltonian, are provided in Figures S4–S5 and their captions below. As an aside, we note that regions of strain defined by a 5-fold-coordinated atom and its immediate 4′ neighborhood are rather more mobile than the atoms themselves, which allows the strained regions to diffuse through the structure at an appreciable rate even after vitrification has slowed atomic diffusion. The time-average of this motion could further contribute to the exponential Urbach tails.

Calculation of local averages for defect formation energy

A rough estimate of the energy of a 5-fold defect, including its environment, can be made using the mean values of distributions in Figure 3, after subtracting the bulk mean energy, and counting the approximate number of neighbors (neglecting the possibility of shared neighbors):

331 meV (5-fold) + 5 × 104 meV (4′) + 15 × 7 meV (4″) + 45 × 0.07 meV (4‴) = 964 meV

A similar calculation for 3-fold defects gives 772 meV.

Summations of the local-energy data using the following pair of methods are consistent with this estimate. Figure 5a of the main text includes 5-fold defects among the 10,801 'isolated' cases provided there are no direct bonds between two 5-folds. An isolated 5-fold's environment is then constructed by searching for all surrounding 4′ atoms whilst ignoring any other defects, then 4″s, and finally 4‴s. This encompasses cases such as a bonded 5–4′–5 triple, where the second 5-fold would be ignored. The average energy of 'isolated' 5-fold atoms under this definition is 984 meV.

An alternative definition of 'isolated' considers only 5-fold atoms for which no other defects occur up to the surrounding 4‴ shell so no other 5-folds are found within 3 bonds. There are fewer (3,953) of such 5-fold defects in the 1M-atom structure, with an average energy of 950 meV.

Across almost all defects (> 97%), using the second definition, the average energy of 5-folds and their environment, including those that cluster, is 749 meV. This value compares with average energies for 3-fold atoms of 726 meV. The remaining 3% were excluded because these rare cases involve, for example, 2-fold and 6-fold connected atoms, or more complicated clusters involving 3- and 5-fold coordination which are complex to handle systematically.

Idealizations of random clustering distributions

The clustering distributions of the following models are included in Figure 5d of the main text to illustrate the increased probability of 5-folds clustering.

Random protocol 1: 16,809 atoms in a 1M-atom perfect diamond structure (the same defect concentration as in the MD structure) were labeled as ‘defects’, without disturbing their positions or bonding topology.

Random protocol 2: 16,809 random atoms in the amorphous 1M-atom structure were relabeled as ‘defects’ using the existing bonding topology defined by a 2.85 Å cutoff. The majority of these are, statistically, 4-fold connected atoms.

Random protocol 3: a bond creation/deletion procedure takes the original clustering distribution of the MD-derived model and randomizes it by walking 5-fold defects across the graph initialized by the bonds in the structure (2.85 Å cutoff). The atomic positions are ignored, with only the bonding topology modified at each step as follows:

1. A 5-fold connected atom is selected at random for movement (atom $\boldsymbol{i}$)
2. A 4-fold connected target is selected at random to become the new 5-fold one under the constraint that it must be exactly 3 bonds away from atom $\boldsymbol{i}$ (atom $\boldsymbol{l}$)
3. Atoms $\boldsymbol{j}$ and $\boldsymbol{k}$ are selected randomly to form a path from atom $\boldsymbol{i}$ to atom $\boldsymbol{l}$ under the constraint that both must not be neighbors of either $\boldsymbol{i}$ or $\boldsymbol{l}$

The bond $\boldsymbol{i}$–$\boldsymbol{j}$ is deleted and bond $\boldsymbol{j}$–$\boldsymbol{l}$ is formed, which moves the 5-fold across the graph by 3 bonds. This procedure can be repeated until the cluster distribution is equilibrated, giving a random distribution in green in Figure 5 after 1.65 million moves (approx. 100 moves per 5-fold center).

Supplementary Figures

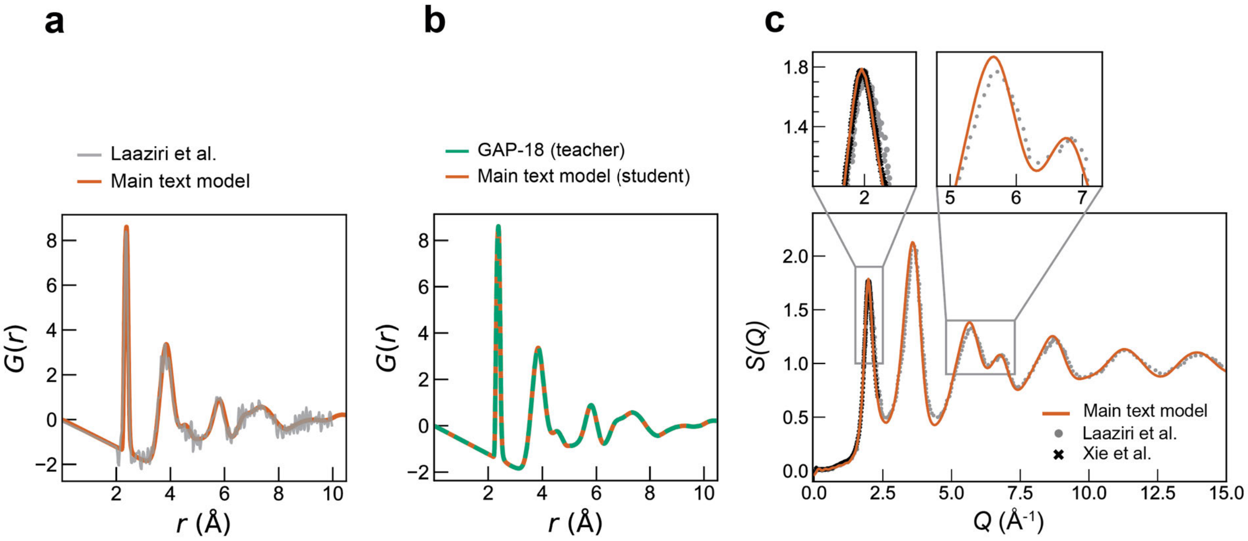


Figure S1. Comparison of structural indicators for 1M-atom model with experiment. (**a**) Total radial distribution function, *G*(r), for the 1M-atom model of the main text compared to the experiment from Ref. [S14]. (**b**) *G*(r) for the same structural model compared to the 100,000-atom model produced by GAP-18,[S4] which is the "teacher" potential as described in Ref. [S1]. (**c**) Structure factor, *S*(Q), for the 1M-atom model (orange) compared with experiments from Refs. [S14] (gray) and [S15] (black). The difference in total energy between teacher and student methods is 3 meV per atom, with a root-mean-square difference of 48 meV for local energies.

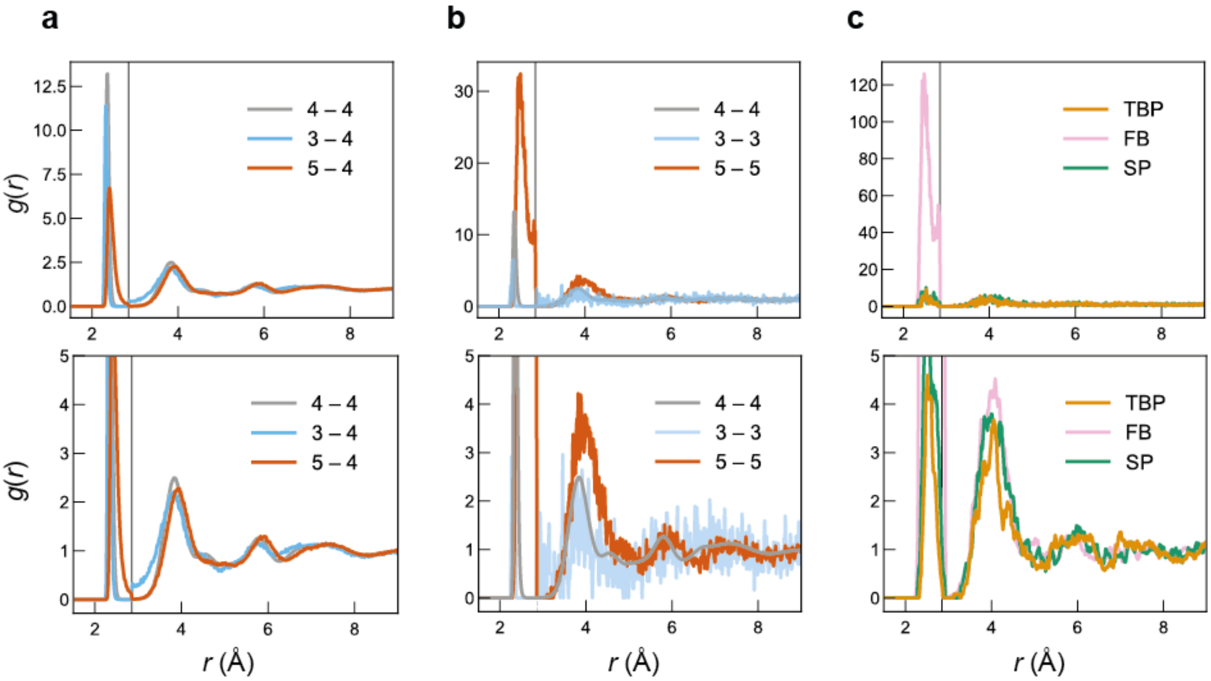


Figure S2. Pair correlations for defect atoms. In all columns, we show radial distribution functions, $g(r)$, in full (upper panels) and magnified (lower panels). (**a**) Correlations between each of 3-, 4-, and 5-fold connected atoms and 4-fold connected atoms. (**b**) Correlations between atoms of the same connectivity. (**c**) Correlations between 5-fold connected defects, separated by classification into trigonal-bipyramid-like (TBP), “floating bond”-like (FB), and square-pyramidal-like (SP) as defined in Figure 2 of the main text. Vertical lines indicate the radial cut-off used to define coordination numbers.

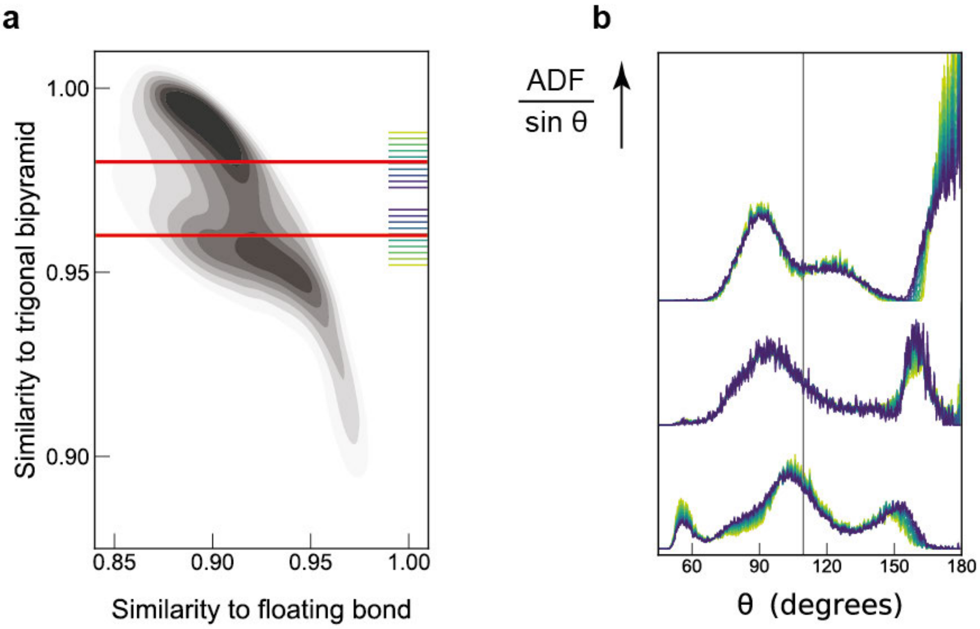


Figure S3. Classification methods for defects. (**a**) SOAP similarity values used for defining the three classes of fivefold defect (red lines) in Figure 2 of the main text. Structures with a similarity greater than the upper red line are categorized as trigonal-bipyramid-like, those between the lines as square-pyramidal-like, and those below as “floating bond”-like. (**b**) Sensitivity of ADFs to similarity cut-off values. ADFs are shown in colors purple to yellow corresponding to a range of similarity values in panel a of the same color. The form of the ADF is quite insensitive over a wide range of similarity values; therefore the results in Figure 2 are not strongly affected by the exact cut-off value for the similarity to distinguish classes.

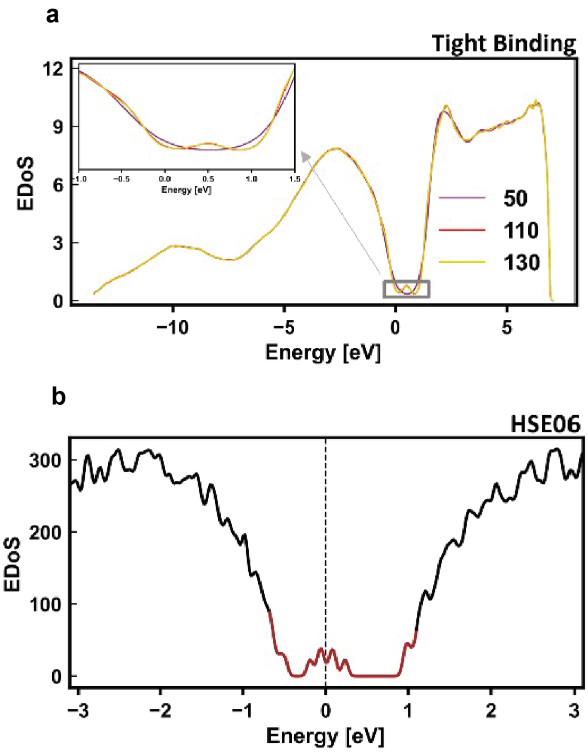


Figure S4. Electronic-structure analysis for validation of a-Si models. (**a**) Electronic density of states (EDoS) for the 1-million atom a-Si model discussed in the main text, computed using the tight-binding methodology of Ref. [S16]. The coloring of the lines refers to the number of moments used in the maximum-entropy reconstruction. See Ref. [S17] for a summary of the methodology and an example of a DOS computation on a similarly large (albeit crystalline) Si structure. (**b**) EDoS for a 512-atom a-Si model computed with the HSE06 hybrid functional.[S18,S19] The 512-atom model is constructed using the same protocol as for the 1-million atom model, and the two models exhibit gaps of similar width. The Fermi level is shifted to *E* = 0 eV. The mid-gap region is highlighted in brown and discussed further in Figure S5.

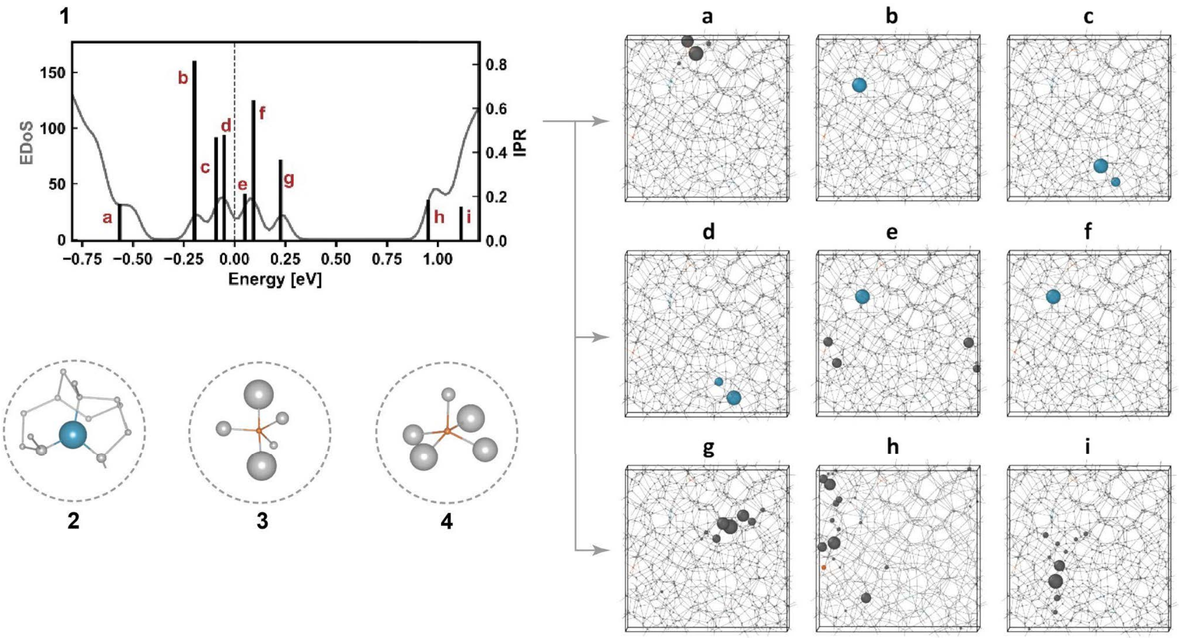


Figure S5. (**1**) Localized states within the energy-gap region (brown region in Figure S4b) obtained for the 512-atom structural model. (**a–i**) Projection of localized states onto the 512-atom model, labeled **a–i** in panel **1**. (**2–4**) Localization environment for mid-gap and tail states of defects and their neighbors. The sizes of the spheres centered on atoms indicate the degree of localization, with 3-, 4-, and 5- fold atoms shown in gray, blue, and orange colors respectively, for both **2–4** and **a–i**. States near mid-gap are localized exclusively on dangling bonds, consistent with earlier electronic-structure computations in Ref. [S12]. The environment of the localized 3-fold atom in **b**, **e**, and **f** is shown up close in **2**. The valence tail state **a** is distributed among 4-fold atoms that are directly connected to floating bonds in a tetrahedral environment with a fifth atom directly opposite a bond (as described in the main text). The environment of **a** is illustrated in more detail in **3**. The electronic state immediately above the Fermi level, denoted as **e** in panel **1**, consists of a linear combination involving a 3-fold atom and 4-fold atoms directly connected to a 5-fold atom in a square pyramidal environment—also discussed elsewhere.[S20] The localization environment of **e** is essentially a mixture of environments **2** and **4**. The conduction tail state **h** is primarily localized on 4-fold coordinated atoms near the same 5-fold atom in the square pyramidal environment of **4**. Our 512-atom cell may be expected to produce representative defects but is not necessarily exhaustive due to its modest size.

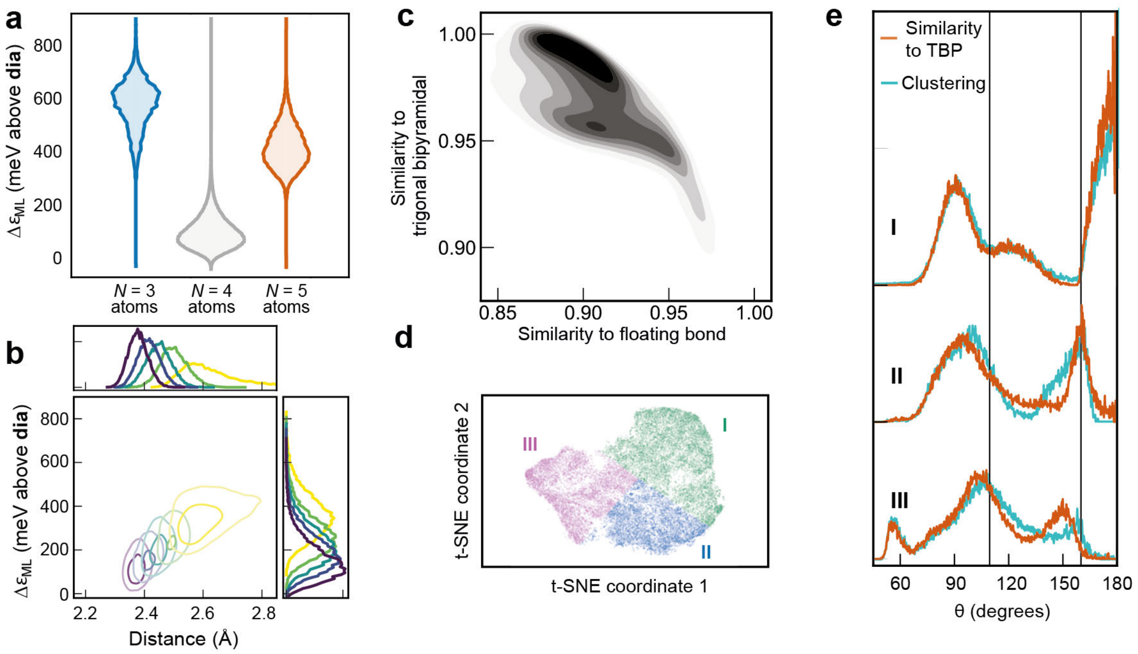


Figure S6. An alternative 1-million atom structure from rapid quenching and annealing. We repeat analyses as in the main text for a separate, entirely uncorrelated 1M atom structure, which was produced with a slightly different protocol: rapid cooling at 10^{13} Ks^{-1} followed by a long annealing period of 2 ns at 840 K. This structure has a similar number of 5-fold atoms as the structure discussed in the main text, which was derived from slower cooling at 10^{11} K s^{-1}. (**a**) Distributions of the ML atomic energies, ε_{ML}, of defects in the model, shown separately for different nearest-neighbor coordination numbers, *N*. (**b**) Two-dimensional correlation plot of the neighbor density for 5-coordinated defects (as in panel **e**). (**c**) 2D plot of SOAP kernel similarity of *N* = 5 atoms in the a-Si model to the idealized TBP and floating-bond environments, respectively. The distribution of values for individual atoms is shown as a heat map. (**d**) Unsupervised classification of 5-fold atoms. The full distance matrix, $\mathbf{D} = \sqrt{2 - 2\mathbf{K}}$, is embedded in 2D with the dimensionality-reduction algorithm t-SNE,[S21] where **K** is the kernel matrix built from the similarity of each 5-fold atom with every other 5-fold atom. Bisecting *k*-means is used to identify clusters **I**–**III**. (**e**) Bond-angle distribution function (BADF) plots, scaled by $\sin\theta$, for atomic triples centered on all 5-fold coordinated atoms. The BADFs are plotted separately for the three distinct categories of 5-fold defects related to idealized structures respectively from top to bottom (as illustrated with selected examples of such configurations from the a-Si model). The 5-fold atoms are separated into these categories via two methods but with similar results: by comparison to the idealized structures of panels **a** and **b** via SOAP similarity (orange) and from unsupervised clustering (cyan). Vertical black lines at the ideal tetrahedral angle (109.5°) and at 160° are guides for the eye.

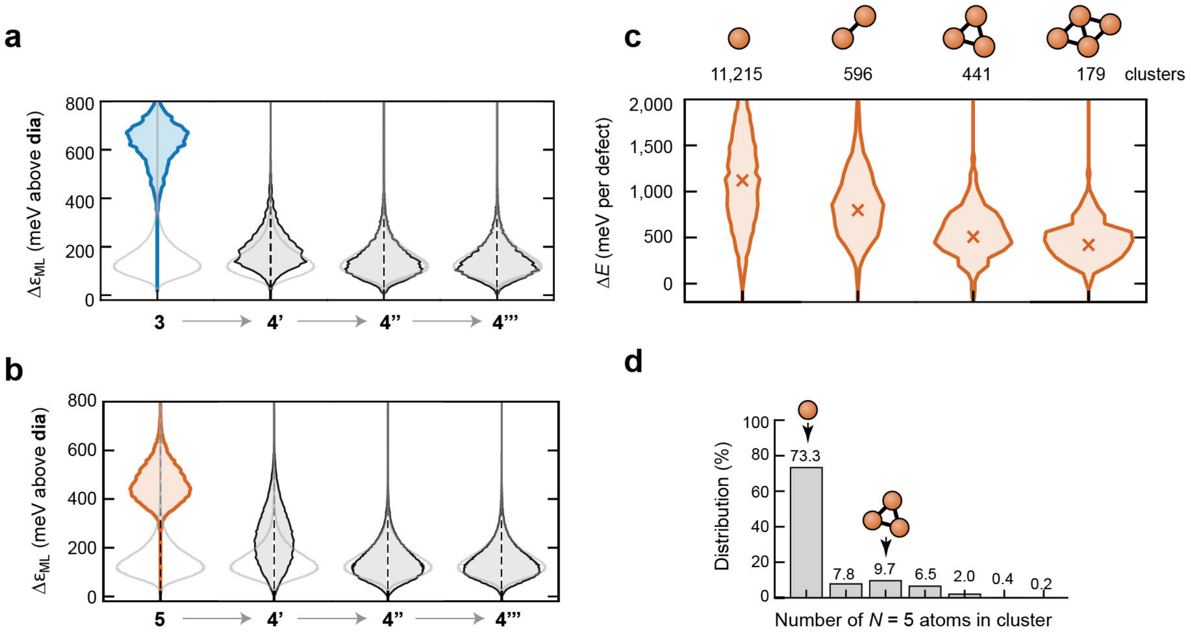


Figure S7. Locality effects in the alternative 1-million atom structure. Further analysis of the locality of defects for the annealed 1-million atom structure of Figure S6. (**a**) Distribution of ML local energies for *N* = 3 atoms and their surroundings. The distribution for bulk a-Si is shown in light grey. (**b**) Same but for *N* = 5 atoms and their surroundings. (**c**) ML-predicted energy distributions for the most common 5-fold defect clusters, with the number of occurrences of each structure directly above each bar. Defect cluster energies are calculated by summing the individual atomic energies of defect cores and their immediate topological neighbors up to 3 bonds away relative to the mean a-Si energy and are reported per-coordination defect: viz. $\Delta E = \sum_i (E_i - \bar{E})\,/n_5$ where i indicates the topological classification of atom i and runs over 5, 4′, 4″, and 4‴; and n_5 is the number of 5-fold atoms in the cluster. Crosses indicate the mean of each distribution. (**d**) Statistics for the number of occurrences of clustered 5-coordination defects.

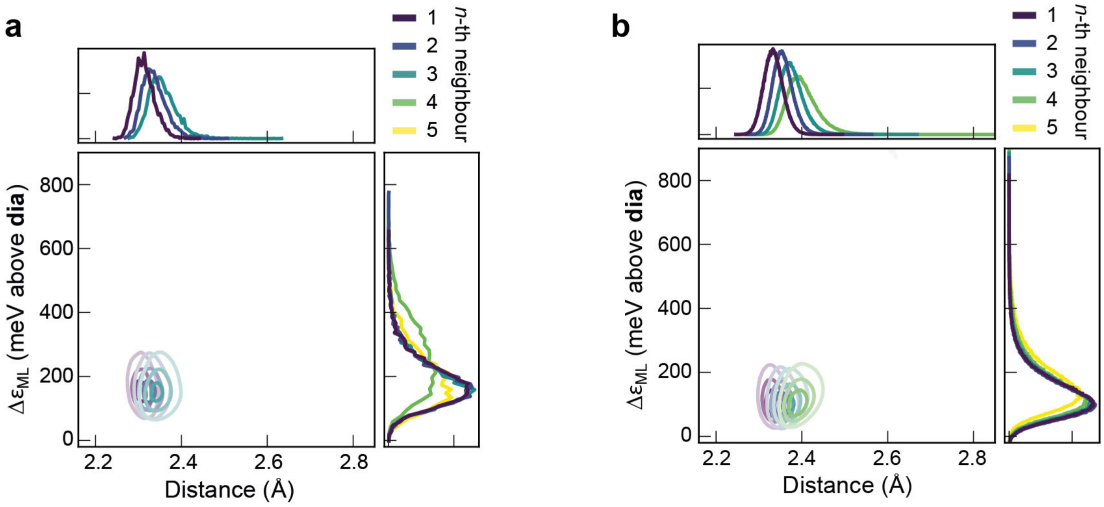


Figure S8. Correlation plots for 3-fold and 4-fold coordinated atoms. Two-dimensional correlation plot of the neighbor density for (**a**) 3-fold defects and (**b**) 4-fold coordinated atoms versus energy, given separately for the immediate neighbors (*n* = 1–5, purple to green to yellow). Note that the 4-th and 5-th neighbors are off the *x*-axis scale for 3-fold connected atoms. Likewise for the 5-th neighbors of the 4-fold connected atoms. The energies of 5-th neighbors for 3-fold and 4-fold atoms are like that of bulk a-Si, which indicates that these more distant atoms are typical of the bulk, in contrast to the 5-th neighbors of 5-fold defects.

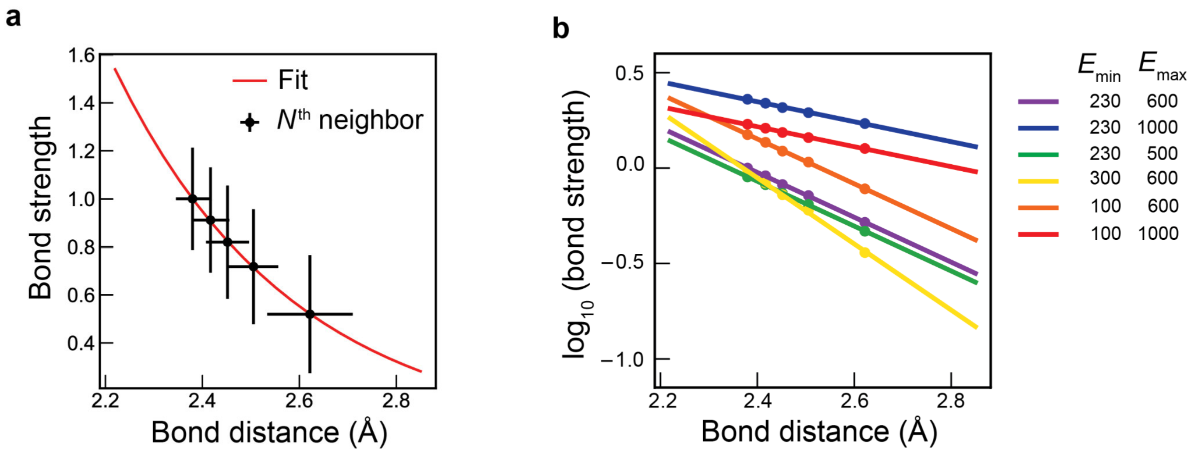


Figure S9. Logarithmic dependence of local-energy-derived bond strength with distance. (**a**) The mean energy (circle) and standard deviation (bars) of each of the N^{th} neighbors in Figure 1e is fitted to an equation similar to Pauling's relation between atomic radii and interatomic distances in metals[S22]. (**b**) We show fitted equations for a range of energy values for normalization to demonstrate the insensitivity to these values. The modified equation to produce these fits is

$$\log(\, n(E)\,) = A(r_c - r)$$

with the bond strength is defined in terms of local energies as

$$n(E) = \frac{E_{max} - E}{E_{max} - E_{min}}$$

where r_c = 2.38 Å is a characteristic radius (taken to be the mean minimum bond length in the structure), r is the bond distance, and A is a fitting parameter with optimal value 1.17 from linear regression. In the definition of bond strength, we use the approximate highest and lowest energy neighbors of 5-folds to transform local energies to a dimensionless bond-strength measure between 0 (no bond) and 1 (strongest bond). The quality of the fit is not sensitive to the precise values used for E_{min} over the range 100–300 meV and E_{max} over 500–1000 meV. In panel **a**, E_{max} = 600 meV and E_{min}= 230 meV. The sum of the mean bond strengths for the 5 bonds to a fivefold atom is 3.97 using this definition, which suggests that the additional bonds do not make up for their increased weakness.

Directly fitting to the bond-energy data with an additional parameter, B, gives a similar quality of fit using the equation

$$\log(\, E\,) = A(r_c - r) + B.$$

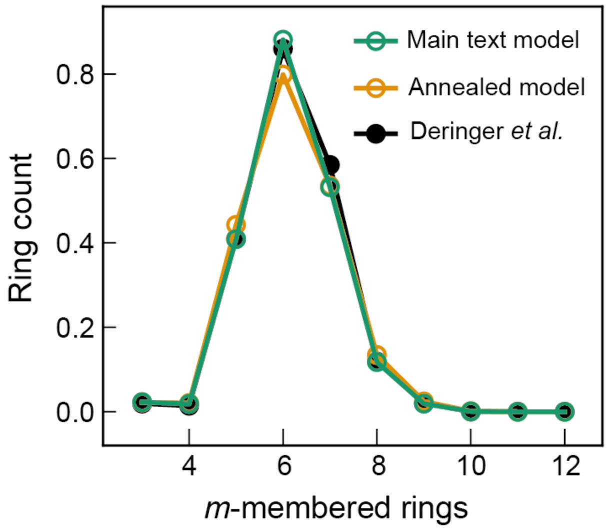


Figure S10. Ring statistics for 1M atom models. Ring counts per atom are displayed for both 1M-atom structural models: the one discussed in the main text (green) and the alternative, newly created annealed structure introduced in Figure S6 (orange). The results are compared to those for a 4,096-atom reference structure (black) from Deringer et al. (Ref. [S8]; obtained with the GAP-18 potential[S4]). The structure discussed in the main text was prepared using the same MD protocol as the 4,096-atom reference and they are here shown to have very similar ring statistics, providing further validation of the quality of the 1M atom model. The algorithm used to obtain ring counts is described in Ref. [S23] (see Computational Methods section above).

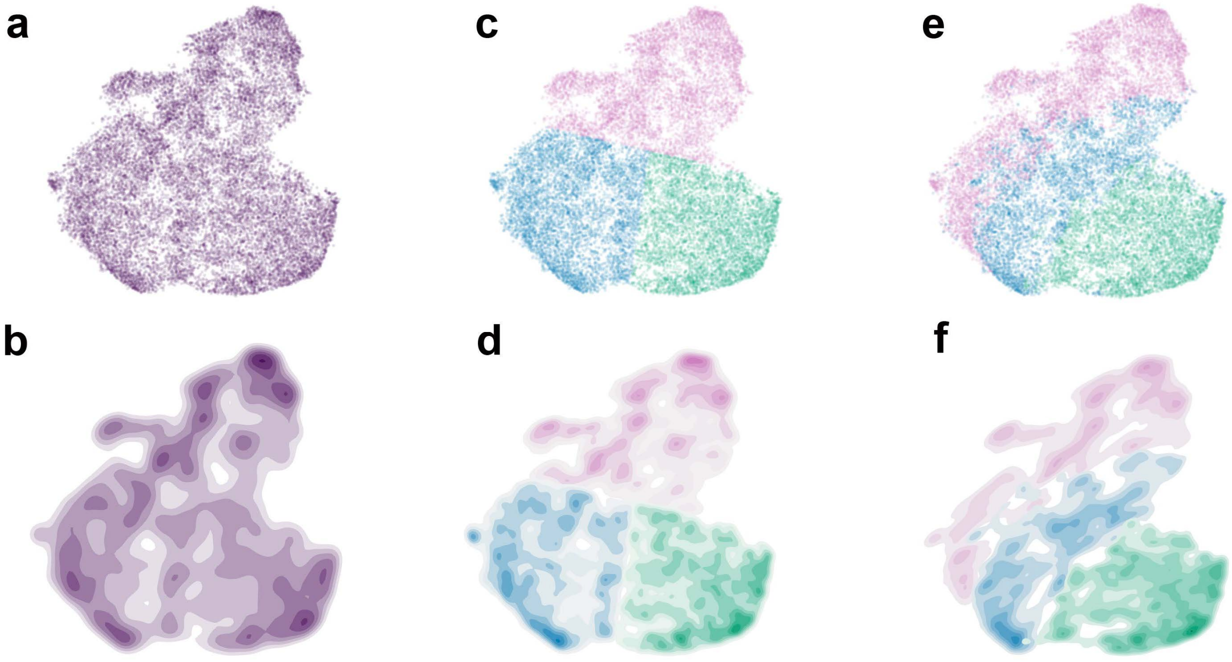


Figure S11. Comparison of clustering methods. Scatter (upper) and gaussian KDE (lower) plots are shown for each of the following: (**a**–**b**) the t-SNE projection of the kernel matrix between SOAP vectors for all 5-fold atoms; (**c**–**d**) coloring by classification label derived from bisecting k-means clustering – an automated method to draw the class boundaries; (**e**–**f**) coloring by classification label derived from comparison with idealized trigonal bipyramidal and floating bond geometries as described in Figure 2. The three most densely populated regions of the map are consistently identified by the supervised and unsupervised classification approaches. The areas throughout the center of the map and in the lower left corner are more uncertain. 74% of environments are put in the same category by both methods.

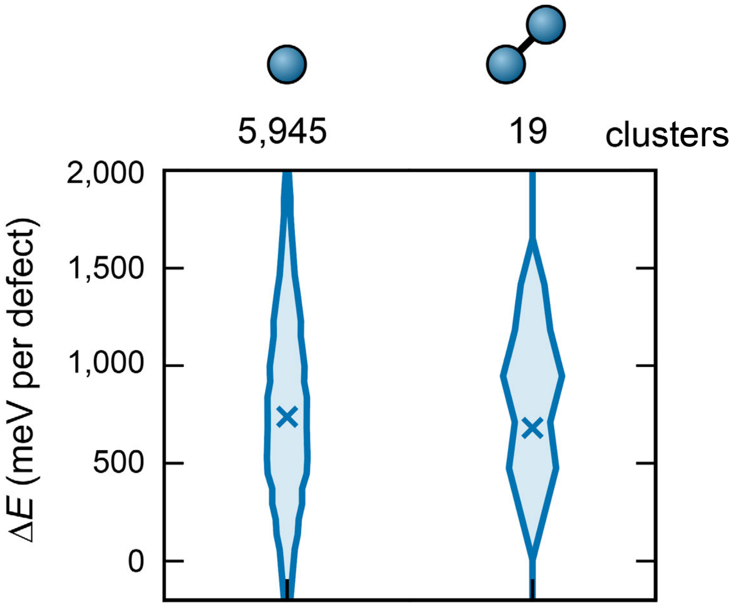


Figure S12. Absence of substantial clustering of 3-fold defects in amorphous silicon. Similar to Figure 5 of the main text, local energy distributions for the largest defect cluster containing only 3-fold atoms are displayed next to the corresponding distribution for an isolated 3-fold defect. As with Figure 5, the defect cluster energies are calculated by summing the individual atomic energies of defect cores and their immediate topological neighbors up to 3 bonds away relative to the mean a-Si energy (excluding other defects) and are reported per-coordination defect. Crosses indicate the mean of each distribution.

Supplementary References

[S1] J. D. Morrow, V. L. Deringer, *J. Chem. Phys.* **2022**, *157*, 104105.

[S2] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.

[S3] A. V. Shapeev, *Multiscale Model. Simul.* **2016**, *14*, 1153–1173.

[S4] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, *Phys. Rev. X* **2018**, *8*, 041048.

[S5] J. L. A. Gardner, Z. F. Beaulieu, V. L. Deringer, *Digital Discovery* **2023**, *2*, 651–662.

[S6] S. Chong, F. Grasselli, C. Ben Mahmoud, J. D. Morrow, V. L. Deringer, M. Ceriotti, *J. Chem. Theory Comput.* **2023**, *19*, 8020–8031.

[S7] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, S. J. Plimpton, *Comput. Phys. Commun.* **2022**, *271*, 108171.

[S8] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, G. Csányi, *J. Phys. Chem. Lett.* **2018**, *9*, 2879–2885.

[S9] Y. Pan, F. Inam, M. Zhang, D. A. Drabold, *Phys. Rev. Lett.* **2008**, *100*, 206403.

[S10] D. A. Drabold, Y. Li, B. Cai, M. Zhang, *Phys. Rev. B* **2011**, *83*, 045201.

[S11] P. A. Fedders, D. A. Drabold, S. Klemm, *Phys. Rev. B* **1992**, *45*, 4048–4055.

[S12] N. Bernstein, B. Bhattarai, G. Csányi, D. A. Drabold, S. R. Elliott, V. L. Deringer, *Angew. Chem. Int. Ed.* **2019**, *58*, 7057–7061.

[S13] R. Atta-Fynn, P. Biswas, *J. Chem. Phys.* **2018**, *148*, 204503.

[S14] K. Laaziri, S. Kycia, S. Roorda, M. Chicoine, J. L. Robertson, J. Wang, S. C. Moss, *Phys. Rev. B* **1999**, *60*, 13520–13533.

[S15] R. Xie, G. G. Long, S. J. Weigand, S. C. Moss, T. Carvalho, S. Roorda, M. Hejna, S. Torquato, P. J. Steinhardt, *Proc. Natl. Acad. Sci.* **2013**, *110*, 13250–13254.

[S16] D. A. Drabold, O. F. Sankey, *Phys. Rev. Lett.* **1993**, *70*, 3631–3634.

[S17] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, S. R. Elliott, *Nature* **2021**, *589*, 59–64.

[S18] J. Heyd, G. E. Scuseria, M. Ernzerhof, *J. Chem. Phys.* **2003**, *118*, 8207–8215.

[S19] J. Heyd, G. E. Scuseria, M. Ernzerhof, *J. Chem. Phys.* **2006**, *124*, 219906.

[S20] J. Dong, D. A. Drabold, *Phys. Rev. Lett.* **1998**, *80*, 1928–1931.

[S21] L. van der Maaten, G. Hinton, *J Mach Learn Res* **2008**, *9*, 2579–2605.

[S22] L. Pauling, *J. Am. Chem. Soc.* **1947**, *69*, 542–553.

[S23] X. Yuan, A. N. Cormack, *Comput. Mater. Sci.* **2002**, *24*, 343–360.