Origins of structural and electronic transitions in disordered silicon

Volker L. Deringer,[1,*] Noam Bernstein,[2] Gábor Csányi,[3] Chiheb Ben Mahmoud,[4,5] Michele Ceriotti,[4,5] Mark Wilson,[6] David A. Drabold[7] & Stephen R. Elliott[8,9]

[1]*Department of Chemistry, Inorganic Chemistry Laboratory, University of Oxford, Oxford OX1 3QR, United Kingdom*

[2]*Center for Materials Physics and Technology, U.S. Naval Research Laboratory, Washington, DC 20375, United States*

[3]*Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

[4]*Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[5]*National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[6]*Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford OX1 3QZ, United Kingdom*

[7]*Department of Physics and Astronomy, Ohio University, Athens, Ohio 45701, United States*

[8]*Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom*

[9]*Present address: Trinity College, Cambridge CB2 1TQ, United Kingdom*

*volker.deringer@chem.ox.ac.uk

Structurally disordered materials continue to pose fundamental questions[1–4], including that of how different disordered phases ("polyamorphs") can coexist and transform from one to another[5–9]. As a widely studied case, amorphous silicon (*a*-Si) forms a fourfold-coordinated, covalent network at ambient conditions and much higher-coordinated, metallic-like phases under pressure[10–12]. However, a detailed mechanistic understanding of the structural transitions in disordered silicon has been lacking, due to intrinsic limitations of even the most advanced experimental and computational techniques. Here, we show how

atomistic machine-learning (ML) models can break through this long-standing barrier, describing liquid–amorphous and amorphous–amorphous transitions with quantum-mechanical accuracy for a system of 100,000 atoms (ten-nanometre length scale). Our simulations reveal a three-step transformation sequence for *a*-Si under increasing external pressure. First, polyamorphic low- and high-density amorphous (LDA and HDA) regions are found to coexist, rather than appearing sequentially. Then, we observe a structural collapse into a distinct, very-high-density amorphous (VHDA) phase. Finally, our simulations indicate the transient nature of this VHDA phase: it rapidly nucleates crystallites, ultimately leading to the formation of a poly-crystalline structure, consistent with experiments[13–15] but not seen in earlier simulations[11,16–18]. An ML model for electronic densities of states (DOS) confirms the onset of metallicity during VHDA formation and subsequent crystallisation. These results shed new light on liquid and amorphous states of silicon, and, in a wider context, they exemplify a holistic, ML-driven approach to predictive materials modelling.

The state-of-the-art in understanding structurally complex materials, including liquid and amorphous matter, has been reached in no small part by means of computer simulations. Still, disordered phases have presented persistent challenges for simulations, requiring large system sizes, long simulation times, and transferable atomic interaction models (that are valid for all relevant structural and bonding environments). ML-driven interatomic potentials are an emerging and powerful approach to address these challenges[19–21], with pressure-induced transitions between *crystalline* phases of silicon having been among the very first applications of these methods[22], and crystal nucleation in the liquid among more recent ones[23]. We have previously carried out pilot studies of disordered silicon based on molecular-dynamics (MD) simulations with a quantum-accurate Gaussian approximation potential (GAP) ML model[24,25], using system sizes between 512 and 4,096 atoms, and considering only the ambient-pressure regime at that

time[26,27]. In the present work, we now use much more extensive GAP-MD simulations of a system containing 100,000 silicon atoms to resolve the atomistic mechanisms of the various structural transitions – including those at very high pressures and densities, which had been incompletely understood (Extended Data Fig. 1–2). Comprising several million individual timesteps at this system size, such simulations would previously have only been possible with empirically parameterised force fields of (necessarily) limited accuracy and transferability[28,29]. However, we demonstrate that such a simple force field is unable to reproduce the pressure-induced changes in silicon, which are observed experimentally and found in the present study. ML potentials are currently gaining immense popularity[19–21], although their use for larger system sizes than in the present work has largely focused on technical capability demonstrations[30] or on transition pathways between crystalline phases[31].

Vitrification of silicon

The first mechanism to be studied here in atomistic detail is the liquid–amorphous transition. Cooling liquid silicon at a properly chosen rate yields a glassy *a*-Si network with a structure compatible with experimental observations, as we have established for small GAP model structures[26,27]. We now carried out a quench simulation for our 100,000-atom system, reducing the temperature at a rate of 10^{11} K s^{-1} in the relevant interval (Fig. 1a). The large system size and (relatively) slow cooling allow us to pinpoint the transition from a supercooled high-density liquid (sc-HDL) to a low-density amorphous (LDA) phase, as the volume increased by about 10% between 1,195 and 1,175 K (Fig. 1a). While our system at 1,500 K appeared to be fully disordered (Fig. 1b), we observed an onset of spatial heterogeneity ("patchiness") during cooling, shown at 1,195 K, just before the transition set in. At this stage, regions with high coordination numbers (red in Fig. 1b) coexisted with others that were much closer to fourfold, "diamond-like" coordination (white), and spatial fluctuations occurred on the length scale of a few nanometres. Upon further cooling (1,195 → 1,175 K), we then observed a rapid transition to a largely fourfold

coordinated, glassy network, concomitant with a sudden drop in the atomic mobility (as monitored by the mean-square displacement, MSD; Fig. 1a). Beside the coordination numbers, the overall, short- to medium-range structural similarity to crystalline silicon also increased sharply during the transition: we measure that using the Smooth Overlap of Atomic Positions (SOAP) kernel[32], which defines a quantitative value for the structural similarity to diamond-type silicon, ranging between zero and one for each atom (Fig. 1c)[27,33]; the same kernel was used to fit the GAP. We finally link the evolution of the spatial (and purely structural) heterogeneity with that of local energetic stability: the predicted atomic energy, ε_{ML}, derived from the GAP regression model, can serve as an indicator for the stability of individual atomic environments in liquid and amorphous silicon[27]. Those regions that are low-coordinated (white in Fig. 1b) and similar to diamond-type silicon (light green in Fig. 1c) also have low – that is, favourable – ML atomic energies (blue in Fig. 1d), and vice versa. Remarkably, the distribution of ε_{ML} and its evolution during the sc-HDL → LDA transition (between 1,195 and 1,175 K) can be deconvoluted into contributions from four-, five-, and sixfold coordinated environments (Fig. 1e). This approach complements the colour-coded plots in Fig. 1d by giving insight into the entire system – collecting local information for 21 simulation snapshots, or 2.1 million distinct atomic environments.

Structural transitions under pressure

The second mechanism, and perhaps the most intriguing question in the context of the present work, concerns the structural transformations of *a*-Si under high pressure. Diamond-anvil cell (DAC) experiments have indicated an amorphous–amorphous transition upon compressing *a*-Si to several gigapascals, evidenced by the sudden disappearance of high-frequency Raman fingerprints and by a concomitant sharp increase of the electrical conductivity (a semiconductor–metal transition), both indicative of a major change in atomistic structure[10–12]. Increasing the pressure even further, to about 14 GPa, was seen to induce crystallisation of the simple hexagonal (*sh*) phase of silicon (thereby demarcating the existence limit of dense disordered phases)[13,14],

although the experimental results may depend on the nature, origin, and purity of the sample[15,34], and the mere appearance of Bragg peaks in X-ray diffraction (XRD) does not fully explain the mechanism of crystallisation. Furthermore, while experiments made it possible to identify the transition in the first place, they can provide relatively little insight into the atomistic structure of the amorphous high-density phase(s). Over the years, computer simulations have led to predictions of various high-pressure structures, including those with $N = 5$ predominantly[11,12,17] and those with much higher coordination numbers[16], presumably depending on the computational method used. No previous atomistic simulation has been able to reproduce the pressure-induced crystallisation of amorphous silicon, to our best knowledge. Motivated by these outstanding questions, we carried out ML-driven simulations of the 100,000-atom *a*-Si system under isothermal compression. Hydrostatic pressure was applied at a constant rate of 0.1 GPa ps^{-1}, while the temperature was held at 500 K: high enough to overcome local energy barriers, but below the melting line.

The evolution of our *a*-Si system with increasing pressure is visualised in Fig. 2a–e, which reveals multiple interesting phenomena. Up to 11 GPa, most atoms remained in fourfold-coordinated (LDA-like) environments. However, regions of higher coordination emerged (magenta in Fig. 2a), consistent with the notion of a "high-density amorphous" (HDA) phase. A striking result is the *coexistence* of LDA- and HDA-like regions at the same temperature and pressure; that is, the simulations indicate the presence of polyamorphism over a range of several GPa, rather than an abrupt transition to an almost completely fivefold-coordinated single HDA phase. Being able to capture this phenomenon at all requires system sizes beyond the nanometric length scale. We note that McMillan *et al.* explicitly mention the presence of both polyamorphs on decompression, inferred from Raman data at the time[11], and that Moras *et al.* described the simulation of a gradual transition between LDA- and HDA-like *a*-Si under hydrostatic pressure, as well as the disappearance of this effect under shear[35].

Upon further compression, beginning at around 12 GPa, much higher coordinated ($N \geq 7$) regions suddenly emerged in our simulation (orange in Fig. 2b), again exhibiting spatial heterogeneity on a scale of several nanometres. These highly-coordinated regions rapidly coalesced into a dense form that is distinct from both LDA and HDA (Fig. 2c). We refer to this phase as "very-high-density amorphous" (VHDA), in line with conventions in the field[16–18]. The rapid structural collapse during VHDA formation reduced the volume from around 18 to around 14 Å^3 per atom (Fig. 2f). Vibrational densities of states, which are consistent with experimental evidence from Raman measurements and corroborate the disappearance of the high-frequency modes as a consequence of the structural transition, are presented in Extended Data Fig. 3.

Importantly, this VHDA phase was transient in our pressurization simulations, and crystalline regions rapidly nucleated (Fig. 2d), in agreement with experiments: DAC XRD measurements showed sharp diffraction peaks, consistent with an *sh* phase ("Si-V"; Ref. 36), beginning to appear upon compression of an amorphous sample to around 14 GPa[13]. The remarkable finding of the present work is not just the formation of *sh* at high pressure (that, alone, has been deduced from free-energy estimations[37] and observed by XRD[13,14]), but the observation of a multistep crystallisation process which proceeds through an entirely distinct VHDA precursor – at variance with the assumption in previous work of direct HDA → crystalline transitions[13,14,37]. Having reached 20 GPa (a few tens of picoseconds after the crystallisation had first set in), our system had fully transformed into a polycrystalline ("pc") phase exhibiting hexagonally packed layers, stacked to form an *sh* structure (Fig. 2e). Disordered regions between the grains remained, as expected for poly- and nano-crystalline materials (Fig. 2g). The small number of crystallites observed in our simulation (Fig. 2e) suggests a nucleation-controlled mechanism with a critical nucleus size of at least several atoms. It is challenging to quantify the critical nucleus size, due to the highly disordered nature of the preceding VHDA phase, but we may refer to an earlier, DFT-based thermodynamic estimate of a critical-nucleus diameter of ≈ 0.7 nm at 14 GPa[13], much smaller than our simulation system size of > 10 nm. We note that an early DFT simulation[16] on

a 216-atom *a*-Si model predicted an abrupt collapse of the tetrahedral network near 16 GPa (which we may now interpret as VHDA formation), though the tiny cell and short simulations revealed nothing about the stability of the structure, and did not show crystallisation[16]. The pressure-induced crystallisation of amorphous solids appears to be a rather infrequent occurrence: two such instances include $Ge_2Sb_2Te_5$ (Ref. 38) and $Ce_{75}Al_{25}$ (Ref. 39), but neither seem to involve (transient) VHDA-like phases.

To test the robustness of our observation, we developed a separate ML potential fit to results of the strongly constrained and appropriately normed (SCAN) functional[40], which also predicts VHDA formation and crystallisation (Fig. 2h and Extended Data Fig. 4), as does a random phase approximation (RPA) correction to GAP-18 (Fig. 2h and Extended Data Fig. 5). We note that the potentials in these tests nucleated β-tin-like (rather than *sh*) crystallites, presumably because of a slight shift in the delicate balance between both high-pressure forms. We finally performed a negative control, using an empirically fitted force field[41] that has been widely used to study disordered silicon, which showed neither the VHDA formation nor the subsequent crystallisation (Fig. 2h and Extended Data Fig. 6).

To further substantiate the series of transformations observed in Fig. 2a–e, we computed excess enthalpies, ΔH_{ac}, compared to the respective most stable crystalline form of silicon at the same pressure. Their evolution is consistent with the subsequent transformations proposed here. At 0 GPa, we obtained ΔH_{ac} = +0.15 eV at.$^{-1}$ for LDA, and this value did not change notably upon initial compression. At 13 GPa, the VHDA phase is slightly favoured (+0.13 eV at.$^{-1}$) over the LDA / HDA polyamorph (+0.15 eV at.$^{-1}$). Compared to all these non-crystalline phases, the pc-*sh* structure that ultimately formed is much more stable, being only +0.02 eV at.$^{-1}$ higher in enthalpy than the single-crystalline *sh* phase at 20 GPa (Extended Data Fig. 7). The driving force for crystallisation can further be demonstrated by using, once more, the stability of individual atoms as determined by the ML regression model. To include effects of pressure, we define a machine-learned *enthalpy* per atom, $h_{ML}(i) = \varepsilon_{ML}(i) + pV/N$, which we reference to the enthalpy

of the respective most stable crystalline phase ("per-atom excess enthalpy", in analogy to the above-mentioned ΔH_{ac} for macroscopic systems). Figure 2i shows the results by colour-coding. In the VHDA phase, the atomic-scale structural disorder is reflected in a seemingly random distribution of more stable (*blue*) and less stable (*yellow*) atomic environments. In contrast, the emerging *sh* crystallites at 15 GPa provide spatial regions of stability. At 20 GPa, the excess enthalpy in the grains is close to that of the crystalline phase, and the grain boundaries "light up" as expected (Fig. 2i). These results emphasise the usefulness of quantum-accurate ML-driven simulations, not only for amorphous but also for polycrystalline materials[30], in which the precise atomistic structure of grain boundaries is a largely unresolved question.

Electronic fingerprints from machine learning

Among the experimental indicators for the amorphous–amorphous transition in silicon is a sudden increase in the electrical conductivity[11]. We studied the electronic structure of our 100,000-atom systems using two approaches, details of which are given in the Methods section. On the one hand, we carried out tight-binding (TB) computations to obtain the electronic DOS directly. On the other hand, we used a recently introduced ML approach[42] to develop a regression model for the DOS in disordered silicon, requiring only atomic coordinates as input. The new parameterisation is "learned" from hybrid-DFT data for representative structural models of all relevant polyamorphs, including VHDA, as well as the pertinent crystalline phases. With this model in hand, we are able to make hybrid-DFT-quality predictions for the electronic DOS within minutes, whereas the direct computation at this high level would have been restricted to system sizes of a few hundred atoms at most. The value at the Fermi level, DOS(E_F), is a primary signature of electrical conductivity[43], and its dramatic increase during compression (Fig. 3a–c) indicates metallisation in the transient VHDA phase, qualitatively consistent with the rapid conductivity increase between 10–12 GPa observed in DAC experiments[11]. At 13 GPa, where the VHDA formation was complete in our simulation, the pseudogap was entirely filled in (marked

by an arrow in Fig. 3c). The prediction of this distinct electronic feature might be tested by ultrafast spectroscopy techniques, which have been previously applied to the liquid–liquid phase transition in silicon[44] and can access timescales which indeed correspond to those in our simulations. ML models for the DOS, as shown in Fig. 3, might play a key role in this regard, by giving access to experimentally relevant system sizes (unlike DFT). Another implication of the onset of metallicity is a possible link to superconductivity, in analogy to what has been observed for the metallic high-pressure form of the heavier congener, amorphous germanium[45], and indeed for crystalline *sh* silicon (with a critical temperature of about 8 K at 14.8 GPa)[46]. This question, however, requires further experimental study.

Finally, by combining the structural information (from SOAP similarity, as used in Fig. 2g) and the machine-learned electronic fingerprints, we may construct structure–property maps for atomic environments using kernel principal covariates regression[47]. This approach yields 2D slices that map out the atomic environments, arranged so as to reflect structural diversity and also the relationship between structure and metallicity, for which the locally-averaged ML DOS(E_F) is used as a proxy[42]. We then arranged the slices in 3D to study their evolution through the transitions, with pressure as the third coordinate (Fig. 3e). We observed a rather uniform distribution of data points in LDA silicon at 0 GPa, reflecting the coexistence of locally-ordered semiconducting environments, and highly-defective environments that contribute to the DOS in the gap. The distribution gradually shifted and broadened towards environments with higher local DOS(E_F) as polyamorphic HDA regions developed up to 11 GPa. The structural collapse at 12 GPa led to a new maximum in the map: this indicates a transition between two distinct phases, also seen in Fig. 2b. The VHDA phase was localised in a very different region of the map than the LDA / HDA environments, consistent with the marked increase in coordination numbers (Fig. 2c) and local DOS(E_F) contributions. For the *sh* crystallites (at 20 GPa), the data points remained in an overall similar region of the map but became more sharply focused compared to VHDA silicon, and shifted slightly to a region of lower DOS(E_F), indicative of the formation of

a small pseudogap (also seen in Fig. 3d). We expect that such maps, both in 2D and in 3D, will become useful tools for studying structural and electronic transitions in diverse phases of matter.

Conclusions and outlook

Our simulations have described and explained the full range of phase transitions in disordered silicon, up to the experimentally established limit (namely, crystallisation), consistent with experimental observations. Beyond this one specific material, however, they demonstrate the ability of atomistic ML methods to enable scientific discovery. Giving access to quantum-accurate predictions of structure, stability, and properties, these methods can reveal as-yet unknown phenomena: including structural and electronic fingerprints of individual atoms but also polyamorphism, polycrystallinity, and other forms of nanoscale heterogeneity. Simulations of disordered materials have thereby taken a qualitative step forward: from simple structural models to realistic, predictive, and fully atomistic descriptions of material systems under experimentally challenging conditions.

1. Elliott, S. R. Medium-range structural order in covalent amorphous solids. *Nature* **354**, 445–452 (1991).

2. Sheng, H. W., Luo, W. K., Alamgir, F. M., Bai, J. M. & Ma, E. Atomic packing and short-to-medium-range order in metallic glasses. *Nature* **439**, 419–425 (2006).

3. Xie, R. *et al.* Hyperuniformity in amorphous silicon based on the measurement of the infinite-wavelength limit of the structure factor. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13250–13254 (2013).

4. Keen, D. A. & Goodwin, A. L. The crystallography of correlated disorder. *Nature* **521**, 303–309 (2015).

5. Hedler, A., Klaumünzer, S. L. & Wesch, W. Amorphous silicon exhibits a glass transition. *Nat. Mater.* **3**, 804–809 (2004).

6. Wilding, M. C., Wilson, M. & McMillan, P. F. Structural studies and polymorphism in amorphous solids and liquids at high pressure. *Chem. Soc. Rev.* **35**, 964 (2006).

7. Sheng, H. W. *et al.* Polyamorphism in a metallic glass. *Nat. Mater.* **6**, 192–197 (2007).

8. Debenedetti, P. G., Sciortino, F. & Zerze, G. H. Second critical point in two realistic models of water. *Science* **369**, 289–292 (2020).

9. Cheng, B., Mazzola, G., Pickard, C. J. & Ceriotti, M. Evidence for supercritical behaviour of high-pressure liquid hydrogen. *Nature* **585**, 217–220 (2020).

10. Deb, S. K., Wilding, M., Somayazulu, M. & McMillan, P. F. Pressure-induced amorphization and an amorphous–amorphous transition in densified porous silicon. *Nature* **414**, 528–530 (2001).

11. McMillan, P. F., Wilson, M., Daisenberger, D. & Machon, D. A density-driven phase transition between semiconducting and metallic polyamorphs of silicon. *Nat. Mater.* **4**, 680–684 (2005).

12. Daisenberger, D. *et al.* Polyamorphic amorphous silicon at high pressure: Raman and spatially resolved X-ray scattering and molecular dynamics studies. *J. Phys. Chem. B* **115**, 14246–14255 (2011).

13. Pandey, K. K., Garg, N., Shanavas, K. V., Sharma, S. M. & Sikka, S. K. Pressure induced crystallization in amorphous silicon. *J. Appl. Phys.* **109**, 113511 (2011).

14. Garg, N., Pandey, K. K., Shanavas, K. V., Betty, C. A. & Sharma, S. M. Memory effect in low-density amorphous silicon under pressure. *Phys. Rev. B* **83**, 115202 (2011).

15. Haberl, B., Guthrie, M., Sprouster, D. J., Williams, J. S. & Bradby, J. E. New insight into pressure-induced phase transitions of amorphous silicon: the role of impurities. *J. Appl. Crystallogr.* **46**, 758–768 (2013).

16. Durandurdu, M. & Drabold, D. A. Ab initio simulation of first-order amorphous-to-amorphous phase transition of silicon. *Phys. Rev. B* **64**, 014101 (2001).

17. Morishita, T. High density amorphous form and polyamorphic transformations of silicon. *Phys. Rev. Lett.* **93**, 055503 (2004).

18. Daisenberger, D. *et al.* High-pressure x-ray scattering and computer simulation studies of density-induced polyamorphism in silicon. *Phys. Rev. B* **75**, 224118 (2007).

19. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56**, 12828–12840 (2017).

20. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

21. Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, 1902765 (2019).

22. Behler, J., Martoňák, R., Donadio, D. & Parrinello, M. Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. *Phys. Rev. Lett.* **100**, 185501 (2008).

23. Bonati, L. & Parrinello, M. Silicon liquid structure and crystal nucleation from *ab initio* deep metadynamics. *Phys. Rev. Lett.* **121**, 265701 (2018).

24. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).

25. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).

26. Deringer, V. L. *et al.* Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *J. Phys. Chem. Lett.* **9**, 2879–2885 (2018).

27. Bernstein, N. *et al.* Quantifying chemical structure and machine-learned atomic energies in amorphous and liquid silicon. *Angew. Chem. Int. Ed.* **58**, 7057–7061 (2019).

28. Hejna, M., Steinhardt, P. J. & Torquato, S. Nearly hyperuniform network models of amorphous silicon. *Phys. Rev. B* **87**, 245204 (2013).

29. Dahal, D., Atta-Fynn, R., Elliott, S. R. & Biswas, P. Hyperuniformity and static structure factor of amorphous silicon in the infinite-wavelength limit. *J. Phys.: Conf. Ser.* **1252**, 012003 (2019).

30. Jia, W. *et al.* Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. *arXiv preprint* arXiv:2005.00223 [physics.comp-ph] (2020).

31. Khaliullin, R. Z., Eshet, H., Kühne, T. D., Behler, J. & Parrinello, M. Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nat. Mater.* **10**, 693–697 (2011).

32. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).

33. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).

34. Imai, M., Mitamura, T., Yaoita, K. & Tsuji, K. Pressure-induced phase transition of crystalline and amorphous silicon and germanium at low temperatures. *High Pres. Res.* **15**, 167–189 (1996).

35. Moras, G. *et al.* Shear melting of silicon and diamond and the disappearance of the poly-amorphic transition under shear. *Phys. Rev. Mater.* **2**, 083601 (2018).

36. Hu, J. Z. & Spain, I. L. Phases of silicon at high pressure. *Solid State Commun.* **51**, 263–266 (1984).

37. Shanavas, K. V., Pandey, K. K., Garg, N. & Sharma, S. M. Computer simulations of crystallization kinetics in amorphous silicon under pressure. *J. Appl. Phys.* **111**, 063509 (2012).

38. Xu, M. *et al.* Pressure-induced crystallization of amorphous $Ge_2Sb_2Te_5$. *J. Appl. Phys.* **108**, 083519 (2010).

39. Wu, M., Tse, J. S., Wang, S. Y., Wang, C. Z. & Jiang, J. Z. Origin of pressure-induced crystallization of $Ce_{75}Al_{25}$ metallic glass. *Nat. Commun.* **6**, 6493 (2015).

40. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402 (2015).

41. Stillinger, F. H. & Weber, T. A. Computer simulation of local order in condensed phases of silicon. *Phys. Rev. B* **31**, 5262–5271 (1985).

42. Ben Mahmoud, C., Anelli, A., Csányi, G. & Ceriotti, M. Learning the electronic density of states in condensed matter. Submitted to *Phys. Rev. B*; preprint available at arXiv:2006.11803 [cond-mat.mtrl-sci] (2020).

43. Mott, N. F. & Davis, E. A. *Electronic processes in non-crystalline materials*. (Oxford University Press, 2012).

44. Beye, M., Sorgenfrei, F., Schlotter, W. F., Wurth, W. & Fohlisch, A. The liquid-liquid phase transition in silicon revealed by snapshots of valence electrons. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16772–16776 (2010).

45. Barkalov, O. I. *et al.* Pressure-induced transformations and superconductivity of amorphous germanium. *Phys. Rev. B* **82**, 020507 (2010).

46. Mignot, J. M., Chouteau, G. & Martinez, G. High pressure superconductivity of silicon. *Physica B+C* **135B**, 235–238 (1985).

47. Helfrecht, B., Cersonsky, R. K., Fraux, G. & Ceriotti, M. Structure-property maps with Kernel Principal Covariates Regression. *Mach. Learn.: Sci. Technol.* (2020) doi:10.1088/2632-2153/aba9ef.

Figure legends

Figure 1: Vitrification of supercooled liquid silicon. (**a**) Evolution of the temperature, *T*, the cell volume, *V*, and the change in atomic mean-square displacement, ΔMSD (obtained by subtracting a moving average) in the relevant region of the ML-driven simulation trajectory from supercooled high-density liquid (sc-HDL) to low-density amorphous (LDA) states. (**b**) Structural snapshots during the quench, taken at the beginning (*top*), just before (*middle*), and just after the structural transition (*bottom*). Simulation cells are shown in plan view, offering the same perspective in all panels. Atoms are drawn as opaque spheres, and so the slice thickness is a few Å at most. Coordination numbers, *N* (spatial cut-off = 3.1 Å), are indicated by colour coding. (**c**) Same for the SOAP-kernel similarity to ideal diamond-type crystalline ("**dia**") Si. (**d**) Same for the ML atomic energy, ε_{ML} (referenced to **dia**-Si). (**e**) Evolution of ε_{ML} shown as kernel-density estimates ("smoothed histograms"), similar to Ref. 27, evaluated here for a 100,000-atom system at 1 K temperature increments between 1,195 and 1,175 K, and shown separately according to coordination numbers, *N*. The arrows indicate the direction of evolution of the curves with decreasing temperature, *i.e.* during the quench from the liquid to the amorphous state.

Figure 2: Amorphous silicon at high and very high pressure. (**a**–**e**) Structural snapshots during isothermal compression at 500 K using the GAP-18 model, showing the coexistence of LDA-like ($N = 4$) and HDA-like ($N > 4$) regions up to 11 GPa, the collapse into a transient VHDA phase ($N \gg 4$) at 12–13 GPa, and finally the formation of simple-hexagonal (*sh*) crystallites. Colour coding indicates coordination numbers, *N* (spatial cut-off = 2.85 Å). (**f**) Volume versus pressure during this simulation. The transition pressure, as well as the onset of crystallisation (indicated by dashed lines), are consistent with experimental reports within a few GPa[13]; see text. The inset shows the evolution of coordination numbers, *N*, during the structural transitions. (**g**) SOAP kernel similarity to simple hexagonal (*sh*) silicon. This analysis shows the system at 13 GPa to be fully disordered on the atomic scale and homogeneous on the nanometric scale. In contrast, *sh*-like crystallites have begun to form at 15 GPa, leading to nm-scale inhomogeneity. (**h**) As panel (f), now comparing three simulations with different interatomic potential models but otherwise similar parameters. A new ML potential using the SCAN functional (*red line*) as well as an RPA-corrected difference model (*yellow line*) both reproduce the structural collapse, VHDA formation, and eventual crystallisation; the established empirical Stillinger–Weber (SW) potential (*grey dashed line*) does not predict either of these effects (see also Extended Data Fig.

4–6). (**i**) ML-based prediction of atomic contributions to the enthalpy (defined here as ε_{ML} + pV/N), indicating the local stabilisation of the *sh*-like regions.

Figure 3: Electronic fingerprints of structural transitions. (**a**–**d**) Electronic densities of states (DOS) at various stages of the compression run (*cf.* Fig. 2a–e). Black lines indicate the result of an ML model for hybrid DFT data using the HSE06 functional; grey shading indicates the associated uncertainty quantification (Methods). Blue dashed lines show the result of direct tight-binding (TB) computations for the 100,000-atom systems. Note that the TB basis set is minimal (one s and three p valence orbitals per atom), and therefore states above the Fermi level, E_F, are less well represented because of incompleteness effects. A red arrow marks the filling-in of the pseudogap upon VHDA formation, as discussed in the text. In all plots, E_F is set as the energy zero. (**e**) Evolution of the atomic environments during our compression simulation, visualised using kernel principal covariates regression (KPCovR)[47]. The axes (components) provide the 2D projection of the SOAP kernel[32] features that gives the best balance between discriminating the structural diversity of the environments, and linearly predicting the locally-averaged ML DOS(E_F). The latter quantity, as a fingerprint of electronic structure and metallisation, is used to colour-code the points associated with individual atomic environments. Contour lines indicate the distribution of atomic environments in the KPCovR space and emphasise the structural and electronic transition upon VHDA formation.

Methods

ML-driven modelling of dense disordered silicon. Our primary results are based on a recently introduced general-purpose GAP for silicon[25], henceforth referred to as "GAP-18". Details of the SOAP-GAP approach for fitting ML-based interatomic potential models are given in Refs. 24 and 32. We furthermore refer the reader to previous smaller-scale studies of *a*-Si[25–27] and amorphous carbon[48,49] using this methodology, and to an overview article in Ref. 21.

The fact that an unmodified potential can discover a range of new configurations for which it has not been "trained" (Fig. 2b–e) is a substantial demonstration of transferability: the silicon GAP-18 model has only included ambient-pressure liquid and amorphous configurations but none at high pressures (although we note that it does include the single-crystalline *sh* phase in its construction). The validation of this potential for ambient-pressure *a*-Si has been reported before[26] and included comparison with three key experimental observables: calorimetric excess enthalpies, ^{29}Si solid-state NMR shifts, and the structure factor, $S(q)$[26]. In fact, with the 100,000-atom system in hand, we repeated the calculation of $S(q)$ for completeness, and we obtained practically quantitative agreement with high-resolution experimental data, including the height of the first sharp diffraction peak (FSDP); see Extended Data Fig. 8a. In addition, as a supplement to Fig. 1e which showed the evolution of ML atomic energies during vitrification, we analysed the $S(q)$ of our system along the same, decisive part of the simulation trajectory – which allows us to study the evolution of the FSDP during cooling (Extended Data Fig. 8b).

Validation through a separate ML potential model. A separate database of disordered silicon structures was constructed for the fitting of a new ML model. To explore a wide range of pressures, we chose the unit-cell volume as a simple parameter, which we varied from 20 Å^3 atom^{-1} (almost corresponding to ambient-pressure *a*-Si) down to 11 Å^3 atom^{-1} (extreme compression). We performed GAP-driven constant-volume melt–quench simulations using a Langevin thermostat, as implemented in quippy (`https://github.com/libAtoms/QUIP`); the

protocol was similar to that in Ref. 48. To these structures, we added the small-cell configurations for crystalline diamond-type, β-tin-type, and *sh* silicon from the GAP-18 database. Single-point energies, forces, and virial stresses for all configurations were then evaluated using the SCAN functional[40] and the projector augmented-wave (PAW) method[50] as implemented in the Vienna Ab Initio Simulation Package (VASP)[51,52]. The new potential therefore uses a different DFT functional, treatment of core electrons, and electronic-structure code than GAP-18, to rule out possible artefacts of any part of the ML input data generation. The SCAN-based potential used the same fitting architecture as before, namely a baseline for exchange repulsion at short distances and a Smooth Overlap of Atomic Positions (SOAP)[32] descriptor and kernel, the latter using a cut-off radius of 5.0 Å and a fit using 9,000 representative points. It was found to be required to increase the smoothness of the SOAP kernel to σ_{at} = 0.75 Å, which was previously a key step in the development of the GAP-driven random structure searching (RSS) approach and can help to make potentials more flexible in highly disordered regions of configuration space and in the presence of limited reference data (see below for more details)[53]. The unique identifier of the newly fitted potential parameter files is `GAP_2020_8_8_60_14_23_0_14`.

A Δ-GAP model for beyond-DFT corrections. We also developed another proof-of-concept ML model at the post-DFT level to rule out the possibility that VHDA formation is an artefact of the approximate DFT functional itself. For this, we use the random-phase approximation (RPA), which is an emerging approach for solids[54–56]. Instead of fitting the full RPA PES (which would be an extremely expensive task), we create a machine-learned difference model to be added to an existing baseline – the latter here being constituted by the general-purpose GAP-18 model. The idea behind such a difference fit is sketched in Extended Data Fig. 5a and has been discussed, in the GAP framework, in Ref. 21. As reference points for sampling the potential-energy surfaces at two levels simultaneously, from which the difference ("Δ") model is then constructed (Extended Data Fig. 5a), we use an ensemble of small structures generated using GAP-RSS. This approach, initially introduced in Ref. 53 and extended into a full "self-guided"

fitting framework for ML potentials in Ref. 57, makes it possible to generate potentials with low computational effort. In essence, an initial ensemble of random atomic configurations is created in analogy to the Ab Initio Random Structure Searching (AIRSS) approach[58,59], and in fact using the `buildcell` algorithm of that implementation (including the use of a hard-sphere constraint and space-group symmetry operations to narrow down the search space). An initial GAP model is then fitted to those data, and used to drive structure searches, which iteratively explore a given PES and serve as input for the next round of fitting – extending the reference database up to a specified size and gradually increasing the quality of the evolving GAP[53,57]. Here, we used 900 (110) structures from a large GAP-RSS structural database developed in Ref. 57 to generate RPA-computed training (testing) data, respectively. Each structure contained between 6 and 16 atoms in the unit cell (giving 9,498 atomic environments in the training set in total). Illustrative examples of such GAP-RSS structures are shown in Extended Data Fig. 5b: they include highly disordered atomic environments, allowing us to generate robust potentials in an efficient way[53,57].

The RPA reference computations used the implementation in VASP 5.4.4, a Γ-centred **k**-point mesh with spacing (`KSPACING`) of 0.5655, a plane-wave cut-off of 250 eV, and the VASP rev. 5.4 `Si_GW` PAW potential. The PBE functional[60] was used for the initial wavefunction calculations and also serves as the reference for the difference model; note that the baseline is therefore slightly different from the ground truth in GAP-18, *viz.* PW91. The third step, computing the virtual states, used no long-range Hartree–Fock contribution (`LOPTICS = .FALSE.`), as recommended by the VASP documentation for metallic systems, such as the highly disordered structures considered here. The final RPA correlation energy was evaluated with a grid order (`NOMEGA`) of 16. We fitted the energy difference between RPA and DFT (PBE) using a SOAP-GAP model with 800 representative points, convergence parameters of $\{n_{\max}, l_{\max}\} = \{16, 6\}$, a smoothing of the neighbour density, σ_{at}, of 0.2 Å, and a kernel exponent of $\zeta = 4$. The radial cut-off of the SOAP descriptor was set to 6.0 Å, slightly larger than that used in GAP-18 (5 Å), and

it was combined with radial scaling (`radial_decay=-0.5`). The scaling pre-factor for the energy model was δ = 0.03 eV per atom (corresponding to the approximate distribution of the difference terms to be "learned"; Extended Data Fig. 5c), and the regularisation of the GAP fit was 0.003 eV per atom, the latter corresponding to an "expected error" for the input data. The unique identifier of the potential parameters for the RPA–DFT difference model is `GAP_2020_6_11_0_19_39_52_705`; these need to be combined with the GAP-18 baseline (unique identifier `GAP_2017_6_17_60_4_3_56_165`).

We emphasise that neither this potential nor the SCAN variant discussed above are intended to be a full substitute for the general-purpose model described in Ref. 25. Instead, they are created here to demonstrate the robustness of the presented findings, most importantly, the formation of VHDA, which had not been observed with established empirical interatomic potentials (Extended Data Fig. 6). The development of a full RPA-quality general-purpose ML potential for silicon is envisioned for the future.

Molecular-dynamics simulations. MD simulations for the 100,000-atom systems were carried out using LAMMPS[61], with a Nosé–Hoover thermostat controlling temperature and a barostat controlling hydrostatic pressure[62–64]. The ambient-pressure quench follows the protocol established in our preceding pilot studies, and similarly uses the GAP-18 model: liquid Si at ambient pressure was quenched from 1,500 to 1,250 K at a rate of 10^{13} K s^{-1}, then to 1,050 K at 10^{11} K s^{-1}, and finally to 500 K at 10^{13} K s^{-1}. The change in MSD shown in Fig. 1a, ΔMSD, was evaluated by subtracting a moving average reaching back 10 fs. Pressurisation runs were performed independently for the liquid at temperatures following the melting line (Extended Data Fig. 2a) and for the *a*-Si structure at 500 K, compressing to 20 GPa over 200 ps. The time step in all simulations was 1 fs. For the enthalpy analysis (Extended Data Fig. 7), relevant systems were cooled using 1,000 MD steps and subsequently fully relaxed using a conjugate–gradient algorithm. Enthalpies are referenced to those of the respective most stable crystalline phase, the latter being derived from computing $E(V)$ curves, taking the pressure as a third of the trace of the

stress tensor, and performing a piecewise linear interpolation of the resulting $H(p)$ data for the relevant pressure interval. Vibrational densities of states (VDOS; Extended Data Fig. 3) were obtained for selected, fully optimised structures, which were thermalised at 300 K and the appropriate pressure for 5 ps; the thermostat and barostat were then removed, and constant-energy (NVE) dynamics were carried out for another 1 ps (1,000 time steps). During the NVE simulation, the averaged velocity–velocity autocorrelation function (VACF) was computed at every timestep, as implemented in LAMMPS[61]. The VDOS were then obtained using a Fourier transformation of the VACF, using in parts the `dump2vdos` code [S. Bringuier; `http://www.u.arizona.edu/~stefanb/Codes/dump2VDOS.py`].

Tight-binding computations. TB electronic DOS were obtained using the methods of Ref. 65. A linear-scaling, maximum-entropy method[65] was combined with the tight-binding Hamiltonian of Kwon *et al*.[66], previously used in studies of Urbach tails in *a*-Si.[67] A relatively realistic tight-binding scheme using four orbitals (one s and three p) per site (Ref. 65) was employed to compute the Hamiltonian matrices for snapshots from 0 to 20 GPa, and also for large supercells of the diamond-type and simple hexagonal crystal phases of silicon. The electronic densities of states were computed with 70 Tchebychev polynomial moments extracted from sparse Hamiltonian matrices of dimension 400,000. For each snapshot, the 400,000 × 400,000 matrix was converted into a sparse format. A conservative initial guess, somewhat broader than the exact support of the spectrum, was made; the sparse Hamiltonian was then scaled and shifted onto the range (–1,1). An approximate "impartial vector" reproducing the first three exact moments was obtained,[65] and Tchebychev polynomial moments were extracted from the matrix (which are, in turn, Tchebychev moments of the DOS function of the matrix). The preceding matrix operations were order-N (N being the dimension of the matrix), since they required only matrix-on-vector operations[68] (no matrix multiplications). To obtain an approximate DOS, we solved the resulting Hausdorff moment problem. The Principle of Maximum Entropy[69] was used to solve the moment

problem, both because of its underlying fundamental rationale, and its rapid pointwise convergence[70,71] compared to methods such as the Kernel Polynomial Method[72]. For large numbers of moments, numerical convergence is sensitive to the guessed spectral support, and this is iteratively tuned to the exact support as the number of moments increases. The convergence of the DOS was examined and 70 moments were found to be more than sufficient to obtain accurate pointwise estimates for the DOS across the full spectral range for all of our snapshots. For reference, and to showcase the system sizes accessible to our method, we also include the DOS of the diamond-type structure (computed for a cubic 2,097,172-atom cell, 34.7584 nm on a side), using 170 moments, in Extended Data Fig. 9. This result may be compared to analogous computations in large fullerenes and graphene[73].

Machine-learning model for the electronic DOS. We obtained the hybrid-DFT-quality global DOS, represented in Fig. 3a–d, using the methods of Ref. 42. We use SOAP features with radial scaling[74] and sparsified Gaussian processes (GPs) to build an ML model for the total DOS of a given atomistic structure, by decomposing the latter into a sum of local contributions (LDOS) centred on every atomic environment in the system. We represent the DOS as a target of the ML models by its cumulative distribution function (CDF). This approach yielded systematically lower prediction errors than models utilising the DOS curve directly,[42] because it is sensitive to shifts in peak positions. Once the prediction is obtained, we derive the obtained CDF to obtain the ML DOS curves.

Using this approach, a new parameterisation was developed for the present work that is based on hybrid-DFT data. The SOAP cut-off radius was 6.0 Å; the smoothness parameter was set to σ_{at} = 0.5 Å. The radial scaling parameters correspond to a cut-off function,

$$f_{\mathrm{cut}}(r) = \frac{1}{1 + (r/r_0)^m},$$

where we set the rate parameter r_0 to 3.0 Å and the exponent m to 5. We selected 3000 atomic environments by Farthest Point Sampling (FPS)[75] to be the representative environments for the

sparsified GPs. As a kernel, we used the square of the scalar product between the normalised feature vectors, as discussed in Ref. 42. The training data consisted of 658 structures from Ref. 25, supplemented by 100 small *a*-Si snapshots (64 atoms per cell) at 0 GPa, used in Ref. 42, and 30 small dense disordered silicon structural models (64 atoms per cell) that were drawn from the new reference dataset used to fit the SCAN model, over a range of pressures between 11 and 20 GPa. The latter part serves to properly represent the high-density phases and their electronic DOS. Electronic structure calculations to extract the DOS for labelling the input data were performed using the FHI-aims package[76], with the `intermediate` convergence settings. The HSE06 hybrid functional[77,78], which is known to usually provide reliable estimates of the band structure of systems with small band gaps[79], was used to determine the self-consistent Kohn–Sham eigenvalues, which were then used to compute the reference DOS. The **k**-point spacing was 0.01 $Å^{-1}$.

Uncertainty quantification for the ML DOS model. Instead of using the variance estimator of GPs, we built a committee of 8 models, each containing a subset of 394 structures randomly selected from the training set. This approach has been shown to be more computationally efficient and ensures a proper error propagation[80]. The average prediction of the DOS from the committee of models was taken as the final prediction and their variance as the uncertainty estimate. Since the models of the committee are correlated, we rescaled the variance around the mean, determining the calibration coefficient with a likelihood maximization criterion. The value of the uncertainty estimate at each given energy increment is shown by shading in Fig. 3a–d.

Local DOS and kernel principal covariates regression (KPCovR). We discuss here briefly the definition of the locally averaged DOS which is used in constructing the plots in Fig. 3e; more details may be found in the technical work in Ref. 42. In an additive atom-centred learning framework as we use to predict the DOS, the model for an entire structure is constructed as a simple combination of the predictions for individual (*i*-th) atomic environments, *viz.*

$$\mathrm{DOS}(E) = \sum_{i \in A} \mathrm{LDOS}(E, A_i)\,.$$

Individual predictions do not have to be physically meaningful (*e.g.*, it is entirely possible to predict a locally negative LDOS), but reflect the way the ML model combines atom-centred information to reproduce the total DOS: there might exist a scenario in which the best overall model can be achieved by having two nearby atoms having very different density of states, because one of the two distorted environments always occurs in combination with its neighbour. In this scenario, only the sum of the two LDOS would be physically relevant. Following this reasoning, we use a locally-averaged ("LA") value of the ML DOS prediction[42], *viz.*

$$\mathrm{LADOS}(E, A_i) = \sum_{j \in A} \frac{f_{cut}(r_{ij})\,\mathrm{LDOS}(E, A_j)}{\sum_{k \in A}\, f_{cut}(r_{jk})}\,,$$

where f_{cut} is the same cut-off function used to define the atom-centred representations. In other words, we average the ML predictions of the LDOS over a length scale comparable to that used to define the environments, which eliminates the strong fluctuations of the direct LDOS predictions and leads to a more easily interpretable value. These LADOS values are used, together with the same kernel used to regress the DOS, to build a map of the environments in the large structures (represented in Fig. 3e), that reflects both structural diversity (dissimilarity) and the correlations between structure and the LADOS. To this end, we use the recently-introduced kernel principal covariates regression (KPCovR) method, described in Ref. 47, that can be seen as a modified kernel principal-component analysis in which one uses a modified kernel with a scaling parameter, α,

$$\widetilde{\mathbf{K}} = \alpha\mathbf{K} + (1-\alpha)\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^{T}\,,$$

combining the structural information encoded in $\mathbf{K}$ with the target properties (more precisely, their best GP estimate), $\widehat{\mathbf{Y}}$. Here, we take $\widehat{\mathbf{Y}}$ to contain the LADOS restricted to the [–4, 4] eV

energy interval to highlight the correlation between the local environments and their corresponding (LA)DOS in the vicinity of E_F. The two principal components used to draw the maps in Fig. 3e were determined by training the KPCovR model on 164,000 environments, selected by FPS, from 41 structures at pressures ranging from 0 to 20 GPa. All remaining atomic environments were then projected on these two coordinates and used for further analysis. In the plots of Fig. 3e, the axis for component 1 has been inverted to ease visualisation; note that the numerical axis values are immaterial to the interpretation and are therefore not shown. Original data underlying these plots are provided (see Data availability statement below).

Data availability

Original data supporting this work, including coordinates for all reported structural models, are openly available in the Zenodo repository (https://doi.org/10.5281/zenodo.4174139).

Code availability

The QUIP code, which provides the interface for carrying out GAP-driven simulations with LAMMPS, is publicly available at https://github.com/libAtoms/QUIP; additional information may be found there. The GAP code is available freely for non-commercial research at http://www.libatoms.org/gap/gap_download.html.

48. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203 (2017).

49. Caro, M. A., Csányi, G., Laurila, T. & Deringer, V. L. Machine learning driven simulated deposition of carbon films: From low-density to diamondlike amorphous carbon. *Phys. Rev. B* **102**, 174201 (2020).

50. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

51. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).

52. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).

53. Deringer, V. L., Pickard, C. J. & Csányi, G. Data-driven learning of total and local energies in elemental boron. *Phys. Rev. Lett.* **120**, 156001 (2018).

54. Harl, J. & Kresse, G. Accurate bulk properties from approximate many-body techniques. *Phys. Rev. Lett.* **103**, 056401 (2009).

55. Harl, J., Schimka, L. & Kresse, G. Assessing the quality of the random phase approximation for lattice constants and atomization energies of solids. *Phys. Rev. B* **81**, 115126 (2010).

56. Schimka, L. *et al.* Accurate surface and adsorption energies from many-body perturbation theory. *Nat. Mater.* **9**, 741–744 (2010).

57. Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Comput. Mater.* **5**, 99 (2019).

58. Pickard, C. J. & Needs, R. J. High-pressure phases of silane. *Phys. Rev. Lett.* **97**, 045504 (2006).

59. Pickard, C. J. & Needs, R. J. *Ab initio* random structure searching. *J. Phys.: Condens. Matter* **23**, 053201 (2011).

60. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

61. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).

62. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

63. Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **101**, 4177–4189 (1994).

64. Shinoda, W., Shiga, M. & Mikami, M. Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. *Phys. Rev. B* **69**, 134103 (2004).

65. Drabold, D. A. & Sankey, O. F. Maximum entropy approach for linear scaling in the electronic structure problem. *Phys. Rev. Lett.* **70**, 3631–3634 (1993).

66. Kwon, I., Biswas, R., Wang, C. Z., Ho, K. M. & Soukoulis, C. M. Transferable tight-binding models for silicon. *Phys. Rev. B* **49**, 7242–7250 (1994).

67. Drabold, D. A., Li, Y., Cai, B. & Zhang, M. Urbach tails of amorphous silicon. *Phys. Rev. B* **83**, 045201 (2011).

68. Skilling, J. in *Maximum entropy and Bayesian methods* (ed. Skilling, J.) (Kluwer, 1989).

69. Jaynes, E. T. *Probability theory: the logic of science*. (Cambridge University Press, 2003).

70. Mead, L. R. & Papanicolaou, N. Maximum entropy in the problem of moments. *J. Math. Phys.* **25**, 2404–2417 (1984).

71. Bandyopadhyay, K., Bhattacharya, A. K., Biswas, P. & Drabold, D. A. Maximum entropy and the problem of moments: A stable algorithm. *Phys. Rev. E* **71**, 057701 (2005).

72. Weiße, A., Wellein, G., Alvermann, A. & Fehske, H. The kernel polynomial method. *Rev. Mod. Phys.* **78**, 275–306 (2006).

73. Drabold, D. A., Ordejón, P., Dong, J. & Martin, R. M. Spectral properties of large fullerenes: From cluster to crystal. *Sol. State Commun.* **96**, 833–838 (1995).

74. Willatt, M. J., Musil, F. & Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **20**, 29661–29668 (2018).

75. Imbalzano, G. *et al.* Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **148**, 241730 (2018).

76. Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).

77. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).

78. Krukau, A. V., Vydrov, O. A., Izmaylov, A. F. & Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **125**, 224106 (2006).

79. Borlido, P. *et al.* Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **15**, 5069–5079 (2019).

80. Musil, F., Willatt, M. J., Langovoy, M. A. & Ceriotti, M. Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.* **15**, 906–915 (2019).

81. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2010).

82. Bundy, F. P. Phase diagrams of silicon and germanium to 200 kbar, 1000°C. *J. Chem. Phys.* **41**, 3809–3814 (1964).

83. Funamori, N. & Tsuji, K. Pressure-induced structural change of liquid silicon. *Phys. Rev. Lett.* **88**, 255508 (2002).

84. Dharma-wardana, M. W. C., Klug, D. D. & Remsing, R. C. Liquid-liquid phase transitions in silicon. *Phys. Rev. Lett.* **125**, 075702 (2020).

85. Desgranges, C. & Delhommelle, J. Unraveling liquid polymorphism in silicon driven out-of-equilibrium. *J. Chem. Phys.* **153**, 054502 (2020).

86. Needs, R. J. & Martin, R. M. Transition from β-tin to simple hexagonal silicon under pressure. *Phys. Rev. B* **30**, 5390–5392 (1984).

87. Laaziri, K. *et al.* High-energy x-ray diffraction study of pure amorphous silicon. *Phys. Rev. B* **60**, 13520–13533 (1999).

Acknowledgements

V.L.D. acknowledges a Leverhulme Early Career Fellowship and support from the Isaac Newton Trust. Parts of the simulations reported here were carried out during his previous affiliation with the University of Cambridge (until August 2019). N.B. acknowledges support from the Office of Naval Research through the U.S. Naval Research Laboratory's core basic research program, and computer time through the U.S. DOD HPCMPO at the AFRL DSRC. D.A.D. acknowledges support from the U.S. NSF under award DMR 1506836. M.C. and C.B. acknowledge support by the Swiss National Science Foundation (Project No. 200021-182057), and by the NCCR MARVEL, funded by the Swiss National Science Foundation. This work used the ARCHER UK National Supercomputing Service *via* a Resource Allocation Panel award (project e599) and the UKCP consortium (EPSRC grant EP/P022596/1). All structural drawings were created using OVITO[81]. We thank A. P. Bartók for technical help.

Author contributions

V.L.D., G.C., and S.R.E. initiated the project. V.L.D. and N.B. performed the ambient-pressure simulations; V.L.D. performed the high-pressure simulations; V.L.D., N.B., and G.C. carried out further validation with other ML potentials. D.A.D. performed the tight-binding electronic-structure computations. C.B. and M.C. performed the electronic DOS ML predictions and the KPCovR analysis. V.L.D., M.W., D.A.D., and S.R.E. analysed the data and developed the main conclusions regarding high-pressure phases. All authors contributed to discussions. V.L.D. drafted the paper, and all authors contributed to its final version.

Competing interests

G.C. is listed as an inventor on a patent filed by Cambridge Enterprise Ltd. related to SOAP and GAP (US patent 8843509, filed on 5 June 2009 and published on 23 September 2014). The other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to V.L.D.

Reprints and permissions information is available at http://www.nature.com/reprints.

Extended Data Figure legends

Extended Data Figure 1: Machine-learning-driven modelling beyond the nanometric length scale. The fully relaxed amorphous silicon (*a*-Si) structure with 100,000 atoms is shown. The smaller boxes on the left show the size of a 512-atom system from a recent study[26], marking the limit of current DFT methods for simulations over several nanoseconds, and that of a 4,096-atom system in our pilot GAP-MD studies[26]. All boxes are drawn to scale.

Extended Data Figure 2: Compression of liquid silicon. (**a**) Overview of the transition pathways investigated in the present work. The quench (vitrification) and compression runs are discussed in the main text. We focus here on additional data that we have collected for validation: *viz.* the description of the high-temperature liquid. We melted a structure at 1,800 K, above the melting point of diamond-type silicon, and then compressed it by simultaneously adapting the thermostat and barostat settings so as to trace the estimated phase coexistence lines given by Bundy[82], in analogy to Ref. 83. The temperature was reduced by 41 K GPa^{-1} to follow the diamond melting line, up to the estimated triple point at 15 GPa, after which the slope was inverted and followed the metallic silicon melting line (+14 K GPa^{-1})[82]. The compression rate was 0.5 GPa ps^{-1}. (**b**) Structure factors of liquid silicon during this compression run. Computed values from our simulations (*red*) are overlaid on experimental reference data by Funamori and Tsuji[83] (*black*) for which the estimated temperatures are at about 50 K above the melting line[83], closely mirrored by our computations. In the original experimental work[83], arrows indicate the location of the maxima (labelled there as Q_1 and Q_2) and a shoulder in the first peak (Q_h), the latter being gradually diminished at higher pressure, all of which are correctly described by our simulations. In the context of liquid–liquid transitions, we mention in passing very recent density-functional[84] and empirical force-field based studies[85]; such research questions may become worthwhile targets for future GAP-driven studies as well. Reprinted figure with permission from [Funamori, N. & Tsuji, K. *Phys. Rev. Lett.* **88**, 255508 (2002)]. Copyright 2002 by the American Physical Society.

Extended Data Figure 3: Vibrational densities of states (VDOS). We obtained these by Fourier transformation of the velocity–velocity autocorrelation function, as described in the Methods section. Two characteristic features associated with the amorphous–amorphous transition under high pressure, observed in previous Raman spectroscopy experiments[10–12], are reproduced by these simulations. First, the peak at large wavenumbers persists throughout the LDA / HDA coexistence but then disappears entirely. Second, the VHDA formation is associated with the

formation of another peak at intermediate wavenumbers. It is noted that this feature appears in both the simulated VHDA and the polycrystalline *sh* system. The predictions here might be tested, in the future, by combined *in situ* X-ray diffraction and Raman spectroscopy, because the former technique will easily be able to distinguish VHDA silicon from crystalline phases.

Extended Data Figure 4: Reproducibility of VHDA formation with a separate ML potential. (**a**) Snapshots from a compression simulation using the same starting structure and protocol as for the main result (Fig. 2a–e), but now using a newly fitted GAP ML model based on SCAN meta-GGA input data (Methods section). This simulation confirms the structural collapse at high pressure, seen in the third panel, and the subsequent crystallisation. The SCAN-level ML potential initially nucleated β-Sn-like crystallites ($N = 6$; red colour on the atoms), which is explained in the following. (**b**) Energy–volume curves for relevant crystalline allotropes of silicon, computed using the GAP-18 model (based on PW91 data; *top*) and the new SCAN-based model (*bottom*). In both cases, the sequence of diamond-type → β-Sn-type → simple hexagonal (*sh*) with increasing pressure (decreasing cell volume) is correctly reproduced, consistent with early DFT studies[86]. With SCAN, the β-Sn-like phase is favoured over a wider range of pressures; the cross-over between the two $E(V)$ curves is indicated by arrows in both panels. Note that the absolute energies for both allotropes are very similar, leading to a delicate balance between both. (**c**) Oblique view of the simulation cell from the SCAN simulation after having reached 20 GPa. Initially, β-tin-like crystallites had formed ($N = 6$; red); then, an *sh* grain emerged ($N = 8$; orange). Note that the absolute pressure values at which the subsequent transitions occur are slightly different between the GAP-18 (Fig. 2a–e) and SCAN (Extended Data Fig. 4a) simulations, but the VHDA phases and subsequent formation of polycrystalline phases are clearly observed in both. The same is not the case with an established, empirically fitted interatomic potential, as shown in Extended Data Fig. 6b below.

Extended Data Figure 5: Beyond-DFT modelling with a Δ-GAP ML fit. (**a**) Schematic illustration of the approach, as discussed in the Methods section. The key ideas are that: (*i*) the random-phase approximation (RPA) potential-energy surface (PES) can only be sampled at selected points, because of the computational cost, and that: (*ii*) the difference Δ(RPA–DFT), indicated by red shading, varies more smoothly than the full PES and is therefore more easily amenable to an ML fit. (**b**) Example structural snapshots from a GAP-RSS search (Ref. 57). We use such very small simulation cells to represent large structural diversity in ML-potential fitting where computational cost is at a premium. (**c**) Quality-of-fit for the difference model, shown in the form of a scatter plot for the training data (*blue*) and a separate test set (*green*) of the ML prediction

(vertical axis) against the "ground truth" to be learned (horizontal axis). The distribution of the target values, σ, is given as inset, alongside the root mean square error (RMSE) measures for training and testing set. (**d**) Snapshots from a compression simulation using the same starting structure and protocol as for the GAP-18 (Fig. 2a–e) and SCAN (Extended Data Fig. 4) results, but now using the GAP-18 + Δ-GAP(RPA–DFT) difference ML potential. The collapse into VHDA is clearly reproduced, as is the subsequent nucleation of crystallites; the result at 20 GPa is a poly-crystalline β-Sn-like phase (*cf.* Extended Data Fig. 4a).

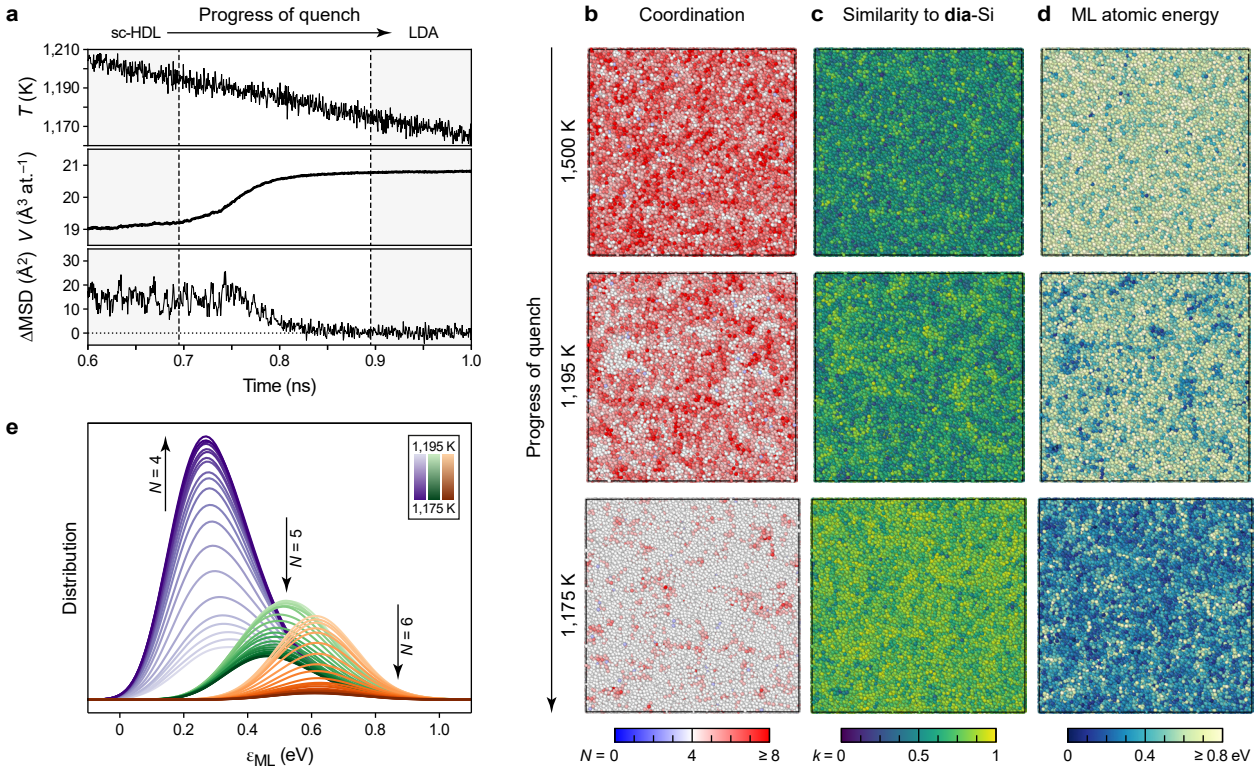
Extended Data Figure 6: Describing VHDA formation and crystallisation requires quantum-accurate simulations. In panel (**a**), we show the results of our ML-driven simulation, with the collapse to VHDA between 12 and 13 GPa, and the crystallisation between 15 and 20 GPa. In panel (**b**), we show results for the same simulation setup but now using the empirical Stillinger–Weber (SW) potential[41], which had been the state of the art for 100,000-atom simulations of silicon so far. Here, neither VHDA formation nor the subsequent crystallisation are observed.

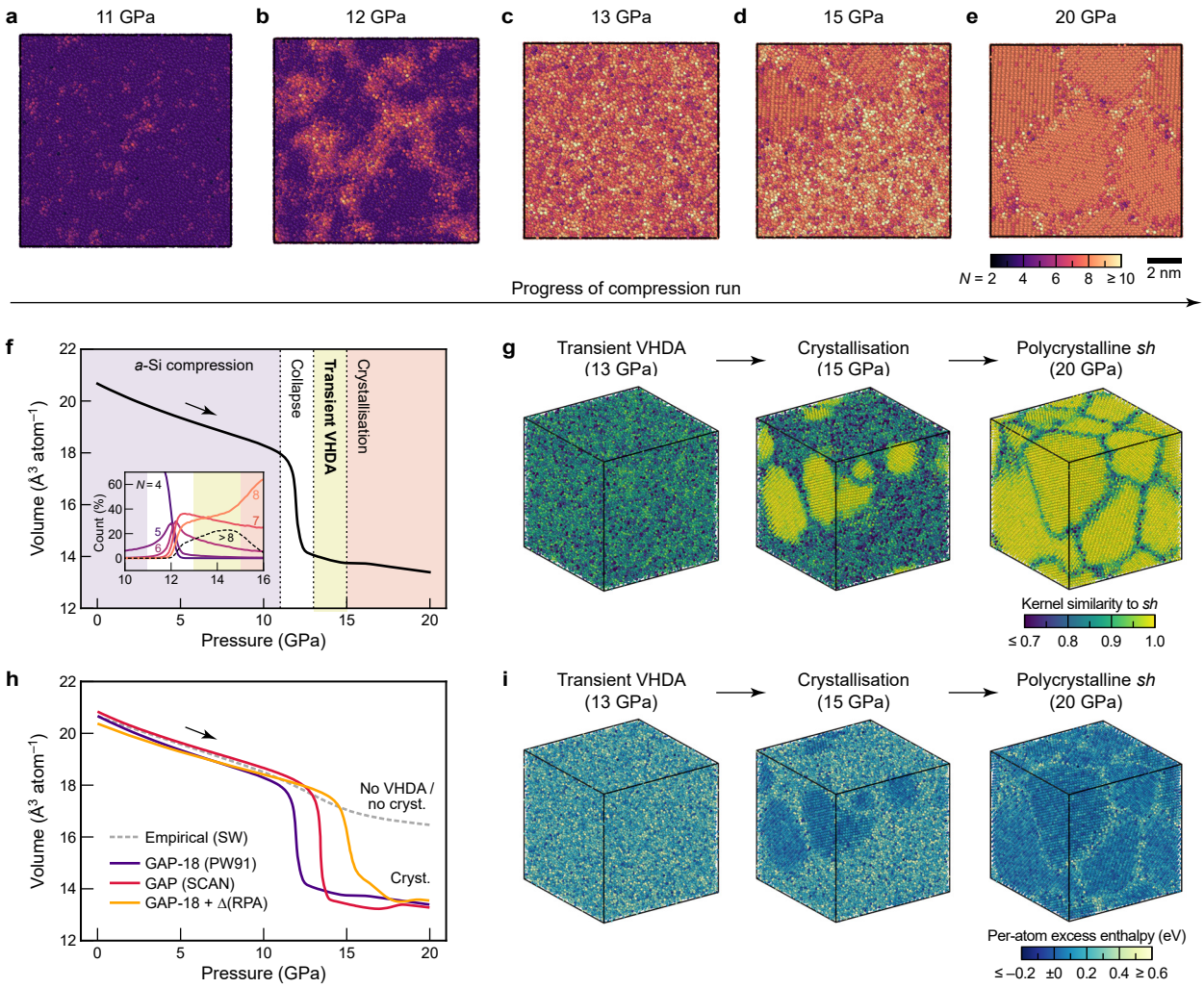
Extended Data Figure 7: The enthalpy landscape of metastable disordered forms of silicon. (**a**) Computed enthalpy of 100,000-atom systems, given relative to the respective most stable crystalline form at any given pressure (diamond-type → β-Sn-type → sh); see Methods section. The red line shows the result for snapshots along the 500 K compression trajectory. Squares indicate results for snapshots which have been frozen in by rapid MD quenching (over 1 ps) and subsequently relaxed with a conjugate-gradient algorithm, all at the given external pressure. The shaded area is a guide to the eye and corresponds to the enthalpy difference between the 500 K and fully relaxed *a*-Si structures at 0 GPa. Relevant structures are shown: note the near-perfect ordering of layers in the polycrystalline ("pc") sample. (**b**) Enthalpy changes associated with the structural changes during compression. Copies of the 10 GPa structure (LDA / HDA polyamorph) were relaxed with increased external pressure (open symbols); this direct relaxation freezes the structure and does not allow it to transform to VHDA. A direct comparison between two competing phases at 13 GPa is therefore possible (labelled as ΔH_1) and indicates the preference for VHDA formation. The enthalpy is lowered much further upon crystallisation (ΔH_2). A dashed vertical line emphasises the change in the crystalline reference, from diamond-type (**dia**) to β-Sn-type silicon. (**c**) Relaxation of copies of the pc-*sh* structure with decreased external pressure, mirroring decompression of a sample in experiment. The relative enthalpic stability over a relatively wide pressure range is qualitatively consistent with the observation of a hysteresis

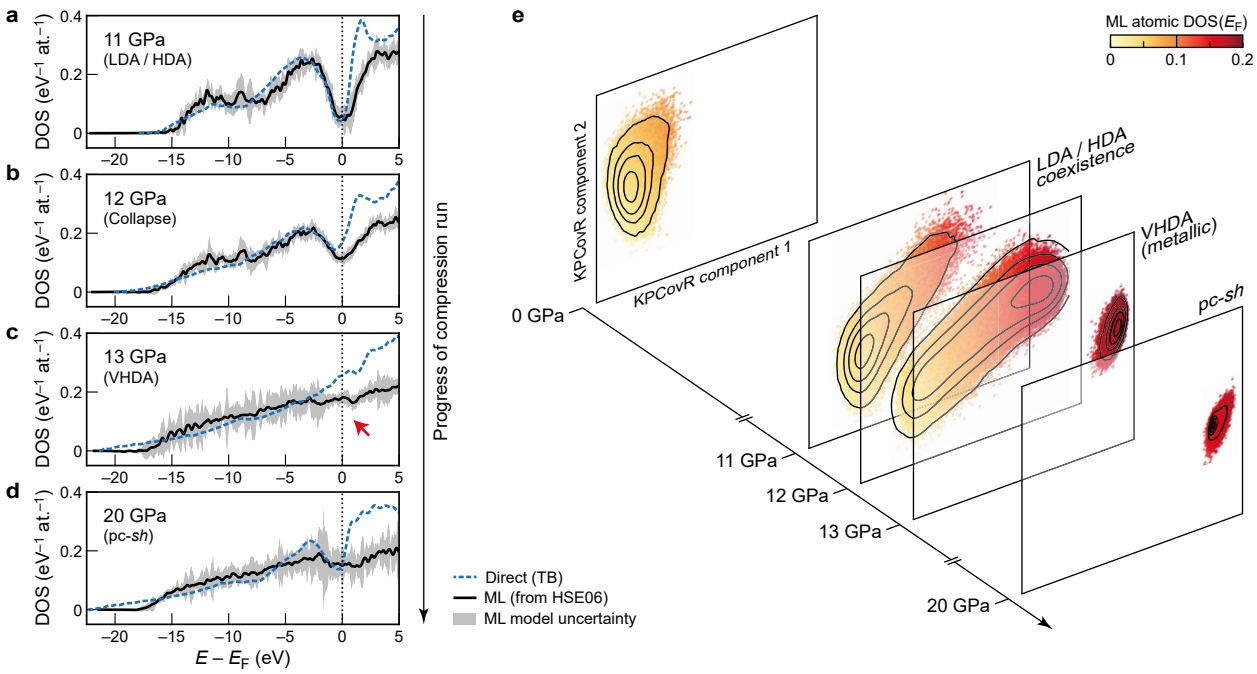
upon decompression: *e.g.*, in Ref. 10, the LDA phase was fully recovered only after decompression to about 4 GPa.

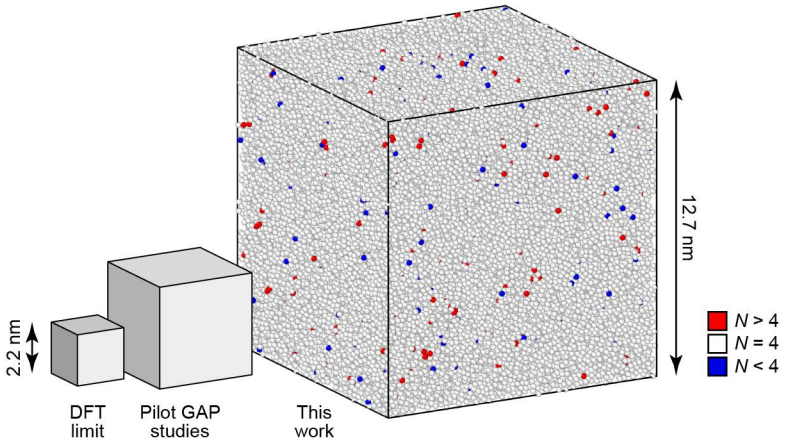
Extended Data Figure 8: Computed structure factors. (**a**) The static structure factor, $S(q)$, as a probe for medium-range structural order, has been evaluated for the fully relaxed amorphous system. The computed result, including the height of the first sharp diffraction peak (FSDP), is in excellent agreement with experimental data taken from Ref. 87. The inset shows a radial distribution function, $g(r)$, for the same structure, indicating long-range correlations beyond the first nanometre, which our ML-driven simulations can access. A dashed line at ≈ 11 Å illustrates the limit of DFT modelling (half the cell length of the smallest system sketched in Extended Data Fig. 1). (**b**) Computed structure factors during quenching. The plot on the left-hand side shows the evolution of simulated structure factors through the relevant part of the liquid-quenching trajectory in the vicinity of the glass transition, plotted in 1 K temperature increments. The emergence of the FSDP (between 1.5 and 2.0 $Å^{-1}$), as well as the structuring of the third peak (between 5 and 6 $Å^{-1}$), are clearly visible. On the right-hand side, a detailed view is provided of the evolution of the FSDP with decreasing temperature, using the same colour scale as on the left-hand side.

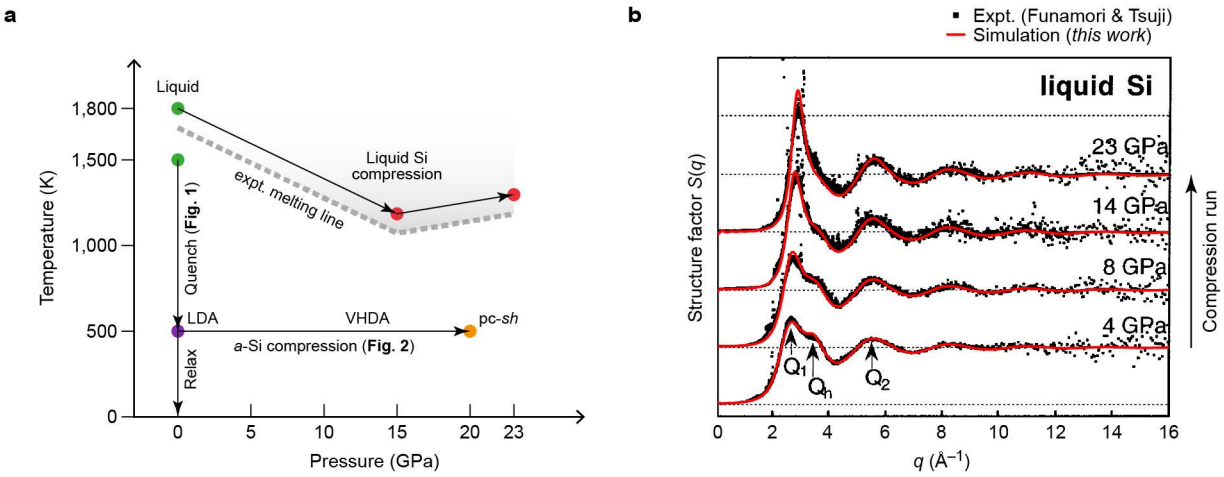
Extended Data Figure 9: Tight-binding DOS for an ultra-large system. Supplementing the TB electronic-structure computations in Fig. 3a–d, this figure shows the electronic DOS computed with the same approach but for a diamond-type crystalline silicon supercell at atmospheric pressure, containing > 2 million atoms (see details in the Methods section). The energy scale is set by Ref. 66.

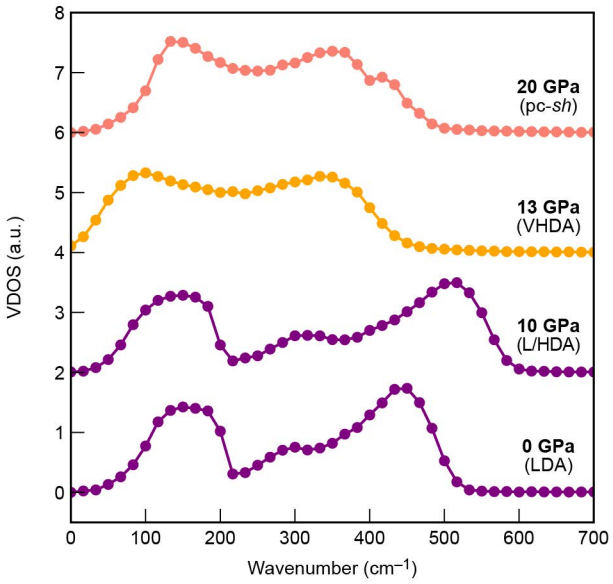

a
Progress of quench
sc-HDL
LDA
T (K)
1,210
1,190
1,170
V (Å³ at.⁻¹)
21
20
19
ΔMSD (Å²)
30
20
10
0
0.6
0.7
0.8
0.9
1.0
Time (ns)
b
Coordination
c
Similarity to dia-Si
d
ML atomic energy
Progress of quench
1,500 K
1,195 K
1,175 K
N = 0
4
≥ 8
k = 0
0.5
1
0
0.4
≥ 0.8 eV
e
Distribution
N = 4
N = 5
N = 6
1,195 K
1,175 K
0
0.2
0.4
0.6
0.8
1.0
εML (eV)

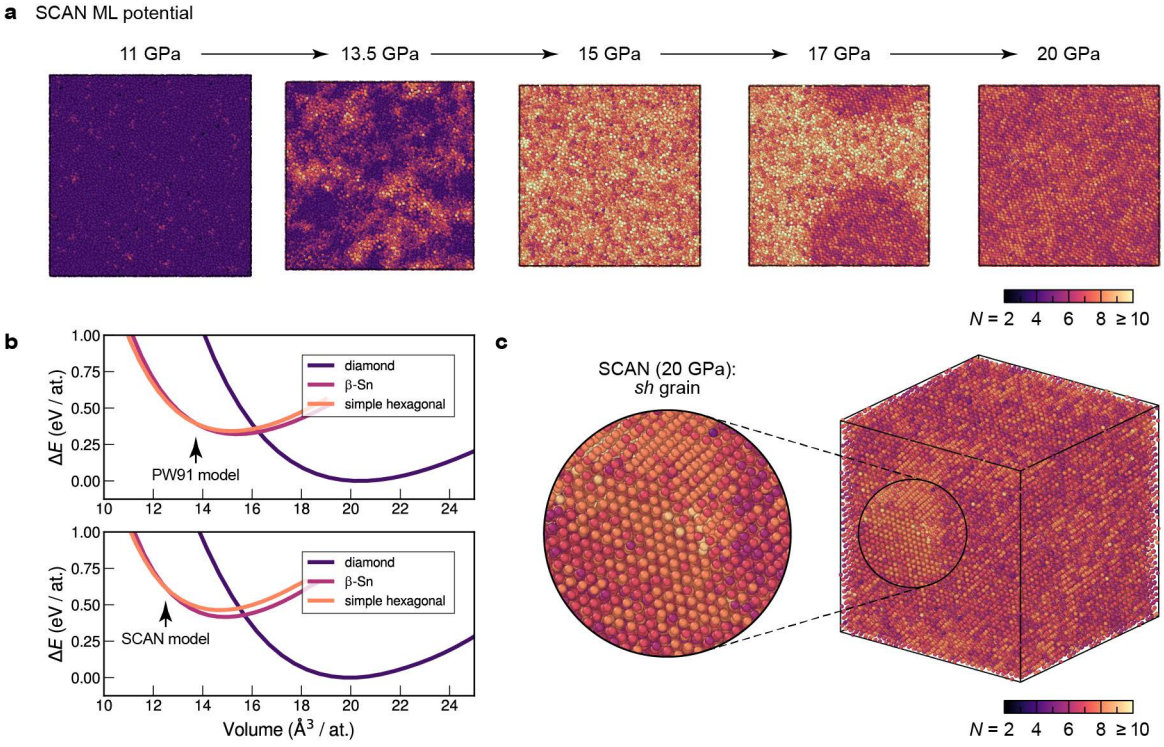

a
11 GPa
b
12 GPa
c
13 GPa
d
15 GPa
e
20 GPa
N = 2 4 6 8 ≥ 10
2 nm
Progress of compression run
f
a-Si compression
Collapse
Transient VHDA
Crystallisation
Volume (Å3 atom−1)
Pressure (GPa)
Count (%)
N = 4
5
6
7
8
>8
g
Transient VHDA (13 GPa)
Crystallisation (15 GPa)
Polycrystalline sh (20 GPa)
Kernel similarity to sh
≤ 0.7 0.8 0.9 1.0
h
No VHDA / no cryst.
Cryst.
Empirical (SW)
GAP-18 (PW91)
GAP (SCAN)
GAP-18 + Δ(RPA)
Volume (Å3 atom−1)
Pressure (GPa)
i
Transient VHDA (13 GPa)
Crystallisation (15 GPa)
Polycrystalline sh (20 GPa)
Per-atom excess enthalpy (eV)
≤ –0.2 ±0 0.2 0.4 ≥ 0.6

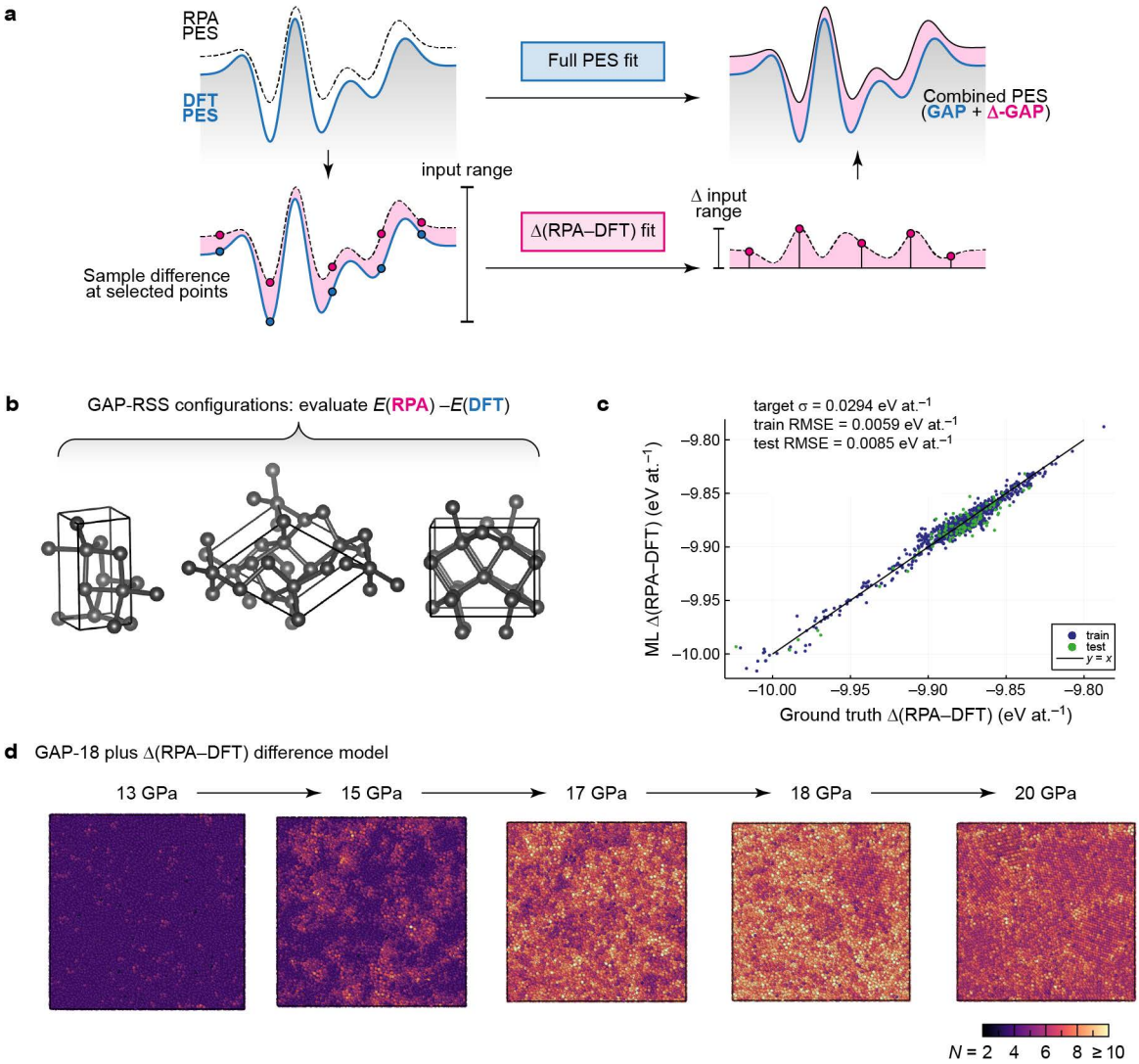

a
11 GPa
(LDA / HDA)
b
12 GPa
(Collapse)
c
13 GPa
(VHDA)
d
20 GPa
(pc-sh)
DOS (eV⁻¹ at.⁻¹)
0.4
0.2
0
−20
−15
−10
−5
0
5
E − E_F (eV)
Progress of compression run
Direct (TB)
ML (from HSE06)
ML model uncertainty
e
ML atomic DOS(E_F)
0
0.1
0.2
KPCovR component 2
KPCovR component 1
0 GPa
11 GPa
12 GPa
13 GPa
20 GPa
LDA / HDA
coexistence
VHDA
(metallic)
pc-sh

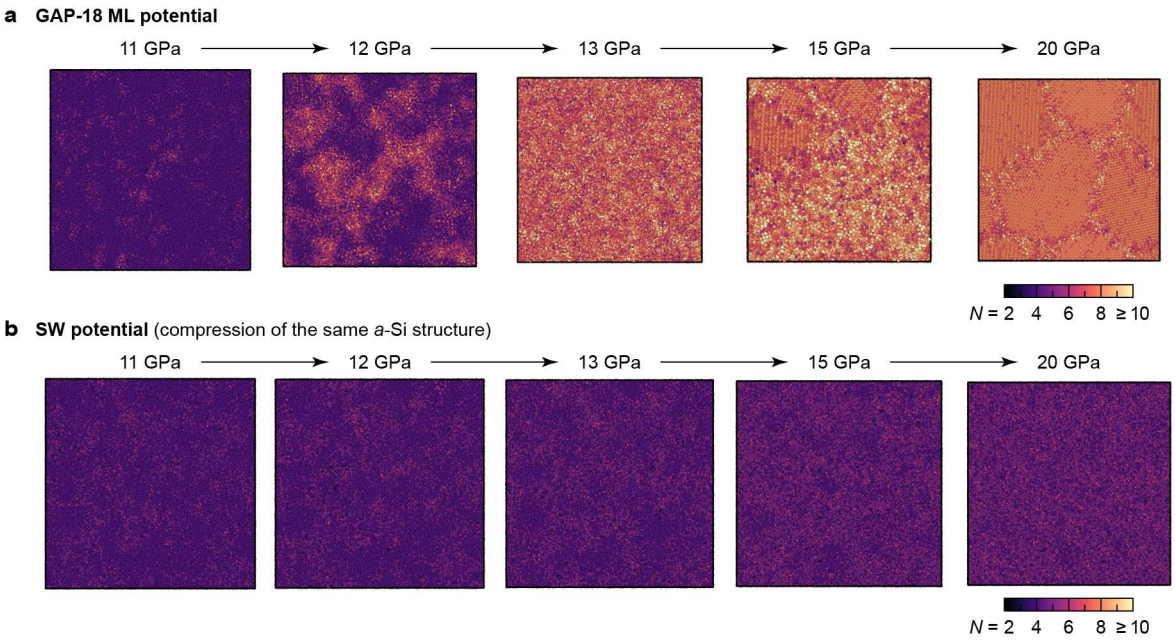

2.2 nm
12.7 nm
DFT limit
Pilot GAP studies
This work
N > 4
N = 4
N < 4

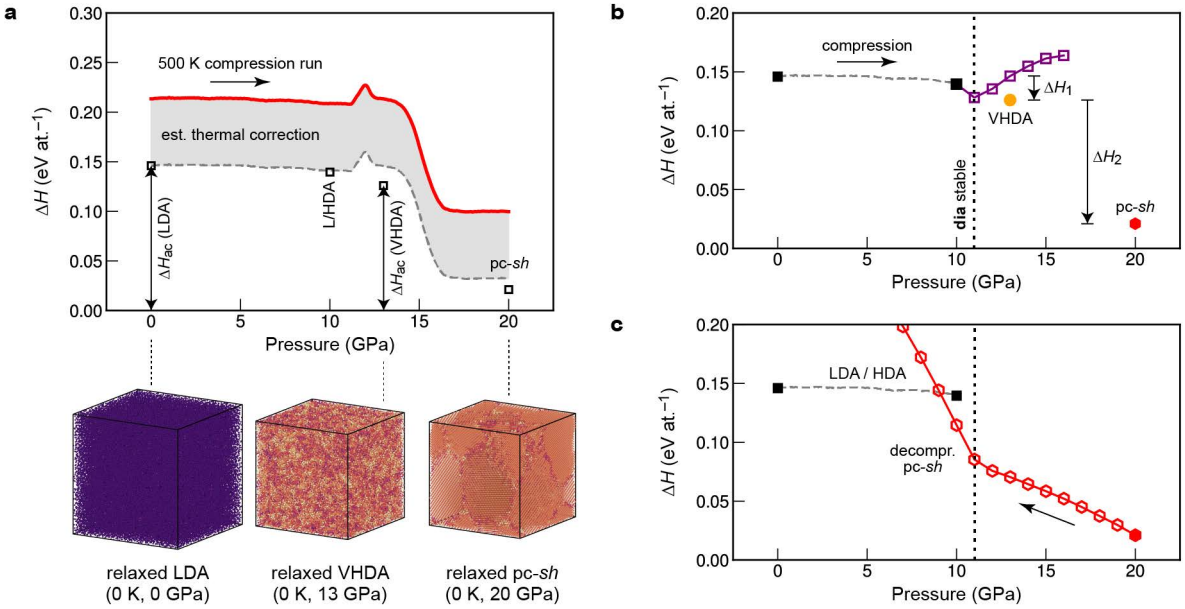

a
Liquid
1,800
1,500
1,000
500
0
Liquid Si
compression
expt. melting line
Quench (Fig. 1)
LDA
VHDA
pc-sh
a-Si compression (Fig. 2)
Relax
0
5
10
15
20
23
Pressure (GPa)
Temperature (K)
b
Expt. (Funamori & Tsuji)
Simulation (this work)
liquid Si
23 GPa
14 GPa
8 GPa
4 GPa
Q1
Qn
Q2
Compression run
Structure factor S(q)
0
2
4
6
8
10
12
14
16
q (Å−1)

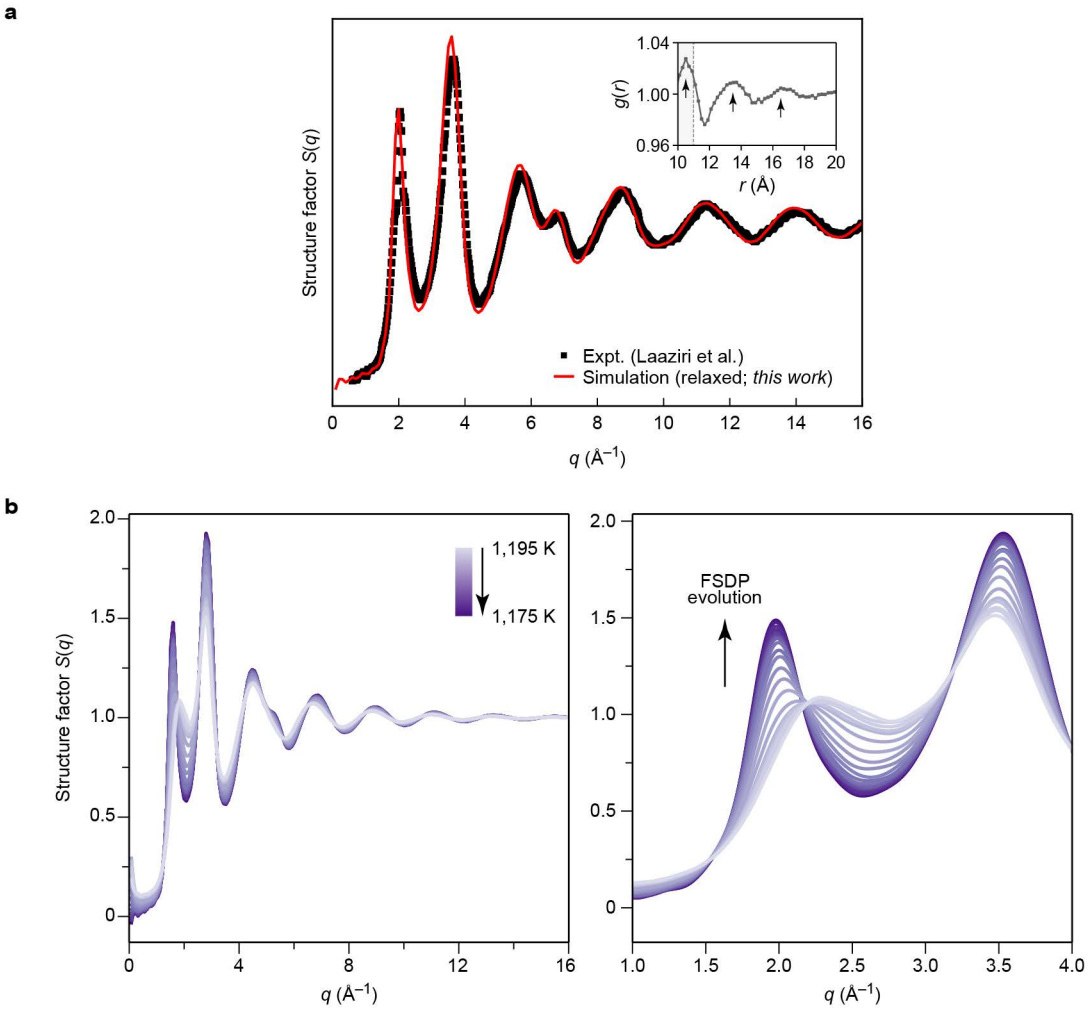

VDOS (a.u.)
Wavenumber (cm^{-1})
20 GPa
(pc-sh)
13 GPa
(VHDA)
10 GPa
(L/HDA)
0 GPa
(LDA)
0
1
2
3
4
5
6
7
8
0
100
200
300
400
500
600
700

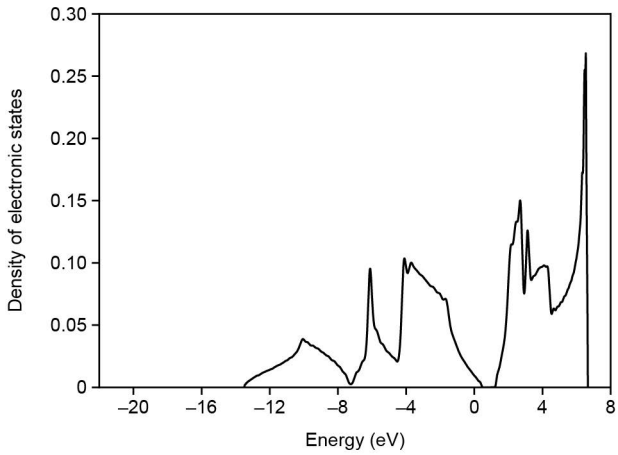

a SCAN ML potential
11 GPa
13.5 GPa
15 GPa
17 GPa
20 GPa
N = 2 4 6 8 ≥ 10
b
ΔE (eV / at.)
diamond
β-Sn
simple hexagonal
PW91 model
SCAN model
Volume (Å3 / at.)
c
SCAN (20 GPa):
sh grain
N = 2 4 6 8 ≥ 10


a
RPA PES
DFT PES
Full PES fit
Combined PES (GAP + Δ-GAP)
input range
Δ input range
Δ(RPA–DFT) fit
Sample difference at selected points
b
GAP-RSS configurations: evaluate E(RPA) –E(DFT)
c
target σ = 0.0294 eV at.−1
train RMSE = 0.0059 eV at.−1
test RMSE = 0.0085 eV at.−1
ML Δ(RPA–DFT) (eV at.−1)
Ground truth Δ(RPA–DFT) (eV at.−1)
train
test
y = x
d
GAP-18 plus Δ(RPA–DFT) difference model
13 GPa
15 GPa
17 GPa
18 GPa
20 GPa
N = 2 4 6 8 ≥ 10


a GAP-18 ML potential
11 GPa
12 GPa
13 GPa
15 GPa
20 GPa
N = 2 4 6 8 ≥ 10
b SW potential (compression of the same a-Si structure)
11 GPa
12 GPa
13 GPa
15 GPa
20 GPa
N = 2 4 6 8 ≥ 10


a
0.30
0.25
0.20
0.15
0.10
0.05
0.00
ΔH (eV at.⁻¹)
500 K compression run
est. thermal correction
ΔH_ac (LDA)
L/HDA
ΔH_ac (VHDA)
pc-sh
0
5
10
15
20
Pressure (GPa)
relaxed LDA
(0 K, 0 GPa)
relaxed VHDA
(0 K, 13 GPa)
relaxed pc-sh
(0 K, 20 GPa)
b
compression
ΔH1
VHDA
ΔH2
dia stable
pc-sh
c
LDA / HDA
decompr.
pc-sh


a
Structure factor S(q)
q (Å−1)
g(r)
r (Å)
Expt. (Laaziri et al.)
Simulation (relaxed; this work)
b
Structure factor S(q)
1,195 K
1,175 K
q (Å−1)
FSDP
evolution
q (Å−1)


0.30
0.25
0.20
0.15
0.10
0.05
0
Density of electronic states
–20
–16
–12
–8
–4
0
4
8
Energy (eV)