

Score-based denoising for atomic structure identification

Tim Hsu^{*1}, Babak Sadigh¹, Nicolas Bertin¹, Cheol Woo Park¹, James Chapman², Vasily Bulatov¹, and Fei Zhou^{†1}

¹*Lawrence Livermore National Laboratory, Livermore, CA, United States*

²*Boston University, Boston, MA, United States*

Abstract

We propose an effective method for removing thermal vibrations that complicate the task of analyzing complex dynamics in atomistic simulation of condensed matter. Our method iteratively subtracts thermal noises or perturbations in atomic positions using a denoising score function trained on synthetically noised but otherwise perfect crystal lattices. The resulting denoised structures clearly reveal underlying crystal order while retaining disorder associated with crystal defects. Purely geometric, agnostic to interatomic potentials, and trained without inputs from explicit simulations, our denoiser can be applied to simulation data generated from vastly different interatomic interactions. The denoiser is shown to improve existing classification methods such as common neighbor analysis and polyhedral template matching, reaching perfect classification accuracy on a recent benchmark dataset of thermally perturbed structures up to the melting point. Demonstrated here in a wide variety of atomistic simulation contexts, the denoiser is general, robust, and readily extendable to delineate order from disorder in structurally and chemically complex materials.

1 Introduction

In molecular dynamics (MD) of condensed matter, characterization methods for the simulated atomic configurations aim to unravel meaningful structural features such as crystalline phases and defects. As the simulations are typically carried out at finite temperatures, accurate characterization of structures and defects is complicated by perturbations in atomic positions induced by thermal vibrations. To this end, increasingly sophisticated methods have been proposed over the years for identification of local atomic motifs in simulated configurations [1–9].

Existing characterization methods usually focus on either ordered crystalline phases or crystal defects. For example, the common neighbor analysis (CNA) algorithm [2] identifies simple crystal structures such as the body-centered cubic (BCC), face-centered cubic (FCC), and hexagonal closed-packed (HCP). Other commonly used methods for structure identification include bond order analysis [1, 10], centrosymmetry analysis [11], adaptive template analysis [12], and polyhedral template matching (PTM) [5]. On the other hand, the dislocation extraction algorithm (DXA) [13, 14] identifies dislocation defects within an *a priori* known ordered crystalline environment. All of the mentioned methods rely heavily on domain knowledge, physical intuition, and heuristics (for review, see for example [3]). As such, they are often application- and/or structure-specific, and are

^{*}hsu16@llnl.gov

[†]zhou6@llnl.gov

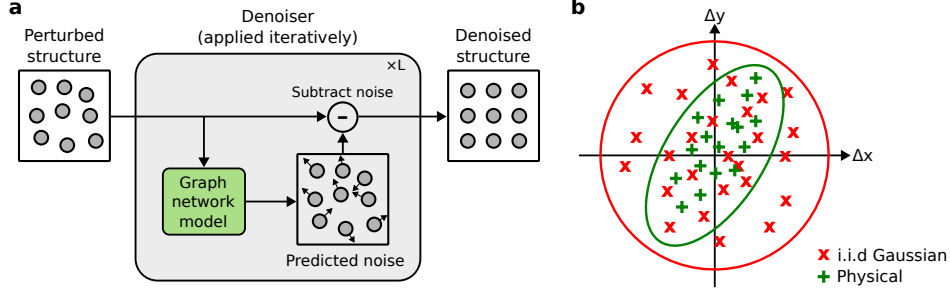


Figure 1: Schematics of (a) the denoising model, in which a graph neural network predicts the noise to be subtracted from an input structure (this operation is applied over multiple iterations); (b) displacements considered, including the artificial i.i.d. Gaussian noise for training (red \times) and physical correlated thermal perturbations expected in MD for inference (green $+$). Contours of the respective uncorrelated and correlated distributions are shown for clarity.

not always easy to generalize beyond their original scope of applicability. More recently, data-driven machine-learning (ML) approaches are being developed for performing ordered phase classification and sometimes defect detection [15–22], often employing existing tools such as Steinhardt order parameters [1] for featurization. While comparatively more straightforward to develop with modern ML pipelines, these emerging methods require considerable amounts of carefully curated training data and are often informed by material-specific physics and domain knowledge which limit transferability of the trained models.

In this work, we take a probabilistic and physics-agnostic approach by supplementing structure characterization tasks with a *denoising* model. Consider a structure classifier $C(\mathbf{x})$ where \mathbf{x} denotes a configuration consisted of atomic coordinates \mathbf{r} and auxiliary information \mathbf{z} regarding cell dimensions and atom types. For a cell of N atoms, $C_i(\mathbf{x}) \in \{1, \dots, N_s\}$ is a classification of atom i ($1 \leq i \leq N$) into one of the N_s candidate structure types. Since $C(\mathbf{x})$ is discrete, there exists a finite region of perturbing displacements $\{\boldsymbol{\epsilon}\}$ within which $C(\mathbf{x} + \boldsymbol{\epsilon}) \equiv C(\mathbf{x})$. Classification is easy for configurations within the $\boldsymbol{\epsilon}$ -neighborhoods of the ideal candidate structures, while those beyond the boundaries pose challenges to C . Rather than directly improving C , we propose to first optimize any configuration \mathbf{x} by bringing it towards the “closest” ideal structures (details in Section 4.3) before classification. This amounts to an improved composite classifier $C \circ D$, where D is a denoising function to remove perturbations with respect to certain ideal reference structures. The probabilistic interpretation is that iterative applications of D , or D^n , gradually increase the similarity of a configuration towards the references to improve classification accuracy. In this way, the problems of classifying ordered crystalline phases and revealing and locating disordered crystal defects can be unified, in analogy with determining order-disorder features from thermal fluctuation. It should be noted that the denoised configurations obtained in our probabilistic approach are mathematical constructs and do not necessarily have direct physical meaning.

To implement such a denoiser, we trained a graph network model, based on the equivariant NequIP architecture [23, 24], to denoise heavily perturbed structures and reveal the underlying order-disorder (Fig. 1a). Given a distribution of pairs of noiseless ideal reference topologies (e.g., BCC, FCC, and HCP) and their noised (with independent and identically distributed or i.i.d. Gaussian random displacements, see Fig. 1b) counterparts, training our denoiser is equivalent to learning a score function [25], which in this work is a gradient field in the atomic coordinate space converging towards points of maximum likelihood that correspond to the ideal reference topologies (more details in Section 4.1). Equipped with this theoretic knowledge, our denoiser can be considered as an itera-

tive scheme that optimizes perturbed structures towards ideal topologies. The score function plays a central role in modern generative models such as the denoising diffusion probabilistic model (DDPM) [26–28] for sampling realistic data from a high dimensional data distribution [29]. Here it is applied for denoising rather than generative applications. Iterative denoising with score functions allows us to approach perfect identification of ordered crystalline phases in several case studies with significant improvement over existing classifiers. It is interesting that similar advantages of iterative use of score functions were observed in DDPM with annealed Langevin dynamics for generative purpose [26–28]. Our findings support the view that iterative models break down challenging problems into smaller, manageable steps.

Compared to other existing denoising methods, e.g., energy minimization (or steepest descent mapping [30]) and vibration-averaging, our method does not have energy-based bias (details in Section 4.1), and only requires the instantaneous snapshot as the input (i.e., only one timeframe of input). In contrast, energy minimization is intrinsically biased due to the use of interatomic potential that may favor certain phases over others, and vibration-averaging requires a tuned averaging window over multiple snapshots where fast processes may be overly smeared. Further, utilizing synthetically noised/perturbed structures as training data, our approach does not rely on physical knowledge other than the ideal reference structures, and is a purely geometric algorithm complementary to existing physics-based techniques. As such, in contrast to data-hungry approaches, our denoiser does not need physics simulation data for training.

Prioritizing single-element systems in this work, the denoising capabilities of our model is demonstrated on several challenging applications, including identification of transient crystal phases during Cu solidification from melt, and characterization of dislocations and point defect debris in BCC Ta undergoing plastic deformation. Importantly, our denoiser does not overzealously denoise the disordered melt into ordered phases. Further, it is shown to help reveal and locate point defects, dislocations, and grain boundaries at high temperatures (approaching melting point) where again the model is observed not to denoise or rearrange crystal defects into ideal lattice motifs. Additional demonstrations on two-element SiO_2 polymorphs are also provided. Besides denoising, the underlying neural network architecture of the denoiser can be extended to classify the denoised atomic environments. At this stage of development we mainly rely instead on existing methods such as CNA, PTM, and DXA to perform the final characterization. With appropriate optimization, we envision that our denoising algorithm would be a robust and highly efficient filter integrated in the workflows of massive MD simulations for the purpose of on-the-fly data compression and post-processing analyses.

2 Results

In our approach, denoising a thermally perturbed configuration \mathbf{x} entails iteratively subtracting the perturbation predicted by a graph network model \mathcal{E}_θ with parameters θ (Algorithm 1). In code implementation, \mathbf{x} is expressed in terms of the atomic coordinates \mathbf{r} and auxiliary information \mathbf{z} regarding cell dimensions and atom types. Our denoiser is an optimization algorithm that modifies input noisy structures towards maximal data likelihood (further explained in Section 4.1, with toy visualization in Supplementary Fig. 1). By including the ideal FCC, HCP, and BCC lattices in the training data, our model attempts to evolve an input perturbed structure towards one of the three ideal lattices depending on which lattice type it most resembles geometrically. Importantly, as demonstrated in the results that follow, our denoiser does not excessively alter the topology of disordered structures, including liquid/melt phase, point defects, dislocations, and grain boundaries far removed from the ideal lattice topologies. This property renders our method safe against overzealous denoising, thus retaining meaningful disordered features in input structures.

Algorithm 1 Denoising process

Require: Denoiser $D(\cdot)$, perturbed structure \mathbf{x} with atomic coordinates \mathbf{r} and auxiliary information \mathbf{z} , and a pre-defined number of iterations k
repeat
 $\mathbf{r} \leftarrow D(\mathbf{x}) := \mathbf{r} - \boldsymbol{\varepsilon}_\theta(\mathbf{r}, \mathbf{z})$
until convergence or k is reached

Trained with purely synthetic data, our denoiser is applied to a wide variety of MD-perturbed systems: (1) BCC, FCC, and HCP Cu simulated above the melting point T_m , as well as liquid/melt Cu perturbed around T_m ; (2) the recently published “DC3” benchmark dataset for crystal structure identification [20] containing BCC, FCC, HCP, cubic diamond, simple cubic, and hexagonal diamond structures perturbed over a wide temperature range from zero Kelvin to above melting; (3) hard-to-detect transient crystal phases momentarily forming during solidification of Cu from melt; (4) FCC, HCP, and BCC Cu containing point defects; (5) BCC Ta containing complex dislocation networks and point defect clusters; (6) BCC Ta containing grain boundaries; and finally (7) SiO₂ polymorphs β -quartz, α -cristobalite, and β -cristobalite. In (1) and (3)–(6), the denoiser is shown to reduce or eliminate thermal perturbation, making it trivial to identify the underlying crystal structures while not destroying meaningful disordered features such as point defects, dislocations, and grain boundaries. In (2), followed by classifiers such as a-CNA and PTM, the denoiser achieves perfect classification accuracies in all systems at T_m . In (7), the generalizability of the denoiser to multi-element complex materials is validated. The results for each case study are detailed below.

2.1 Denoising FCC, HCP, BCC, and liquid/melt Cu

The first demonstration focuses on denoising solid BCC, FCC, and HCP crystals perturbed by thermal vibrations at above the melting point (3400 K), and liquid/melt Cu at around the melting point (3000 K), as shown in Fig. 2a. Before denoising, the popular adaptive CNA (a-CNA) algorithm [3] classifies most of the solid atoms (82, 77, and 77% in BCC, FCC, and HCP, respectively) as disordered, i.e. not belonging to any of the three crystal lattice types. After just one iteration of denoising, the number of misclassified atoms is significantly reduced (to 14, 5, and 3% for BCC, FCC, and HCP, respectively). The following iterations, typically within 5–8 steps, remove the remaining minor perturbations. The denoised solids resemble perfect FCC, HCP, and BCC lattices, thus trivializing subsequent phase classification. The Steinhardt order parameters \bar{q}_4 and \bar{q}_6 [10] computed before and after denoising confirm that virtually all thermal perturbations imparted on the solids are removed (Fig. 2b). Note that after denoising, the Steinhardt quantities for BCC/FCC/HCP appear to be a single point, but are in fact about 1,000 points (each point corresponds to an atom) overlapped together.

Interestingly, denoising the Cu melt phase leaves nearly all atoms to remain disordered, as indicated by a-CNA labeling all such atoms as *other* or *unknown* (Fig. 2a), even though the atomic displacements over the denoising iterations are roughly the same as that for the solid phases (Supplementary Fig. 2). Additionally, the very first peak of the radial pair distribution function (RDF) becomes sharper and splits after denoising, and the peaks at the medium-range distances also become slightly sharper (Fig. 2c). In short, how the denoiser precisely impacts the disordered structure is not yet fully understood (further details in Supplementary Fig. 3), and should be more thoroughly investigated in future work. The fact that the Cu melt phase remains disordered even after denoising can be a useful property of our model not explicitly learned from its training data, which consists of only perfect and randomly distorted but otherwise ordered crystal lattices. In simulations involving solid-liquid coexistence, an example of which will be shown later (Section 2.3), we certainly wish our

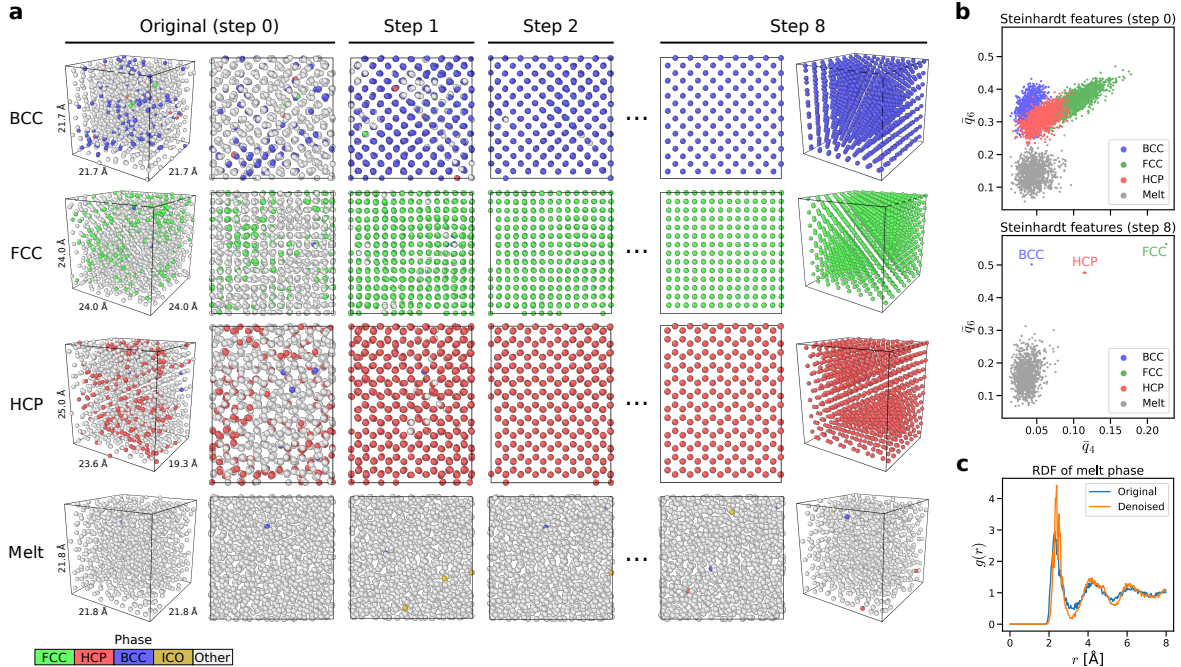


Figure 2: Iterative application of the denoiser to small cells of BCC, FCC, HCP, and liquid/melt Cu. (a) Visualization of the structures along the denoising iterations. (b) Steinhardt features \bar{q}_4 and \bar{q}_6 [10] before and after the denoising. (c) Radial distribution function of the melt phase before and after the denoising. In (a), the structures are shown in orthogonal views along densely packed crystallographic directions, with additional perspective 3D views for steps 0 and 8. The atoms are colored according to a-CNA prediction implemented in OVITO [31]. The solid phases and the melt have been annealed at 3400 K ($1.1T_m$) for 400 ps and 3000 K ($1.0T_m$) for 600 ps, respectively. ICO stands for icosahedral coordination.

model to denoise only the thermally distorted crystal lattices while leaving truly disordered phases disordered.

2.2 Classification accuracy on the DC3 benchmark dataset

Recently Chung et al. [20] proposed a machine-learned crystal classification method, the data-centric crystal classifier (DC3). The authors provided a benchmark dataset that spans a wide variety of chemical composition, crystal structure types, and temperatures ranging from 0 K to melting. Compared to other existing methods [2, 5–9], DC3 was shown to achieve the best accuracy on most systems in the dataset. Since the dataset includes three more crystal types, cubic diamond (*cd*), hexagonal diamond (*hd*), and simple cubic (*sc*) in addition to FCC, BCC, and HCP, we trained a separate denoiser on all six reference structures available in the DC3 dataset. This denoiser accepts systems of at most 2 elements to accommodate the NaCl binary system (hydrogens are absent in the H_2O structures in the benchmark). We summarize in Table 1 the improved classification accuracy at the melting T_m from classifying denoised structures with a-CNA and PTM, as well as the performance of DC3 [20] and other existing methods [2, 5–9] tested in the DC3 paper. The most dramatic improvement can be seen in a-CNA, rising from near-bottom 15–57% accuracy to on par with DC3 after just one single denoising step. With eight denoising iterations, the accuracy of a-CNA and PTM converges to the perfect classification score of 100% on all but Fe snapshots (Supplementary Fig. 6). Confirmed by

Table 1: Accuracy comparison between our classification approach (in bold font), the DC3 classifier [20], and the methods tested in the DC3 work on the DC3 benchmark dataset [20]. The accuracy value measured at the melting point, computed as the fraction of correctly labeled atoms, is shown in percentage. In our denoising approach, the RMSD (root-mean-square-deviation) cutoff for PTM is 0.1. Missing entries imply non-applicable structure types for the chosen classifier.

	Al (fcc)	Fe (bcc)	Ti (hcp)	Si (cd)	H ₂ O (hd)	NaCl (sc)	Ar (fcc)	Li (bcc)	Mg (hcp)	Ge (cd)
a-CNA	50.3	39.9	15.7				57.1	34.4	47.3	
PTM ^a	95.9	84.3	82.8	99.9	100.0	94.6	96.9	83.1	95.7	99.2
Denoiser (1 step) + a-CNA	98.5	96.9	81.7				98.6	96.9	97.1	
Denoiser (1 step) + PTM	100.0	99.5	98.7	100.0	100.0	99.3	100.0	99.6	99.9	100.0
Denoiser (8 steps) + a-CNA	100.0	99.7 ^b	100.0				100.0	100.0	100.0	
Denoiser (8 steps) + PTM	100.0	99.9 ^b	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DC3 ^a	96.9	86.8	89.4	99.0	99.2	95.6	97.5	85.8	97.4	100.0
i-CNA ^a	68.5	56.6	27.9				75.7	51.6	64.9	
AJA ^a	66.9	35.6	42.4				74.0	34.1	67.0	
VoroTop ^a	24.0	61.2	57.5				23.1	57.2	57.4	
Chill+ ^a				99.8	98.8					100.0

^aTaken from the DC3 paper [20].

^bThere is some trace amount of point defects in the BCC Fe snapshots at $T = T_m$ (Supplementary Fig. 6b).

communications with the DC3 authors, some of the Fe snapshots contain Frenkel pairs that formed spontaneously from intense thermal fluctuations at T_m . Therefore our denoising approach indeed reaches perfect classification accuracy for BCC iron by revealing the unsuspected point defects. The difference between the performance from the a-CNA backend classifier and that from PTM on BCC Fe, namely 99.7% and 99.9%, can be attributed to the different ways the classifiers define atoms as defective or unknown.

While Table 1 lists the performances of various methods only at the melting point, Fig. 3a,b,d,e shows, over a wide temperature range, significant improvement in classification accuracy of a-CNA and PTM before and after denoising. Importantly, after denoising, both a-CNA, a common, ubiquitous classifier that is noticeably less performant, and the more updated and sophisticated PTM have essentially the same ideal performance up to the melting point, and close to perfect performance even at temperatures far above T_m . A non-exhaustive inspection of structures above T_m reveals increasing defects, which explain the ostensible dip in accuracy beyond melting. Overall, the results shown so far demonstrate the utility of the denoiser for further improving existing phase classifiers and minimizing a wide range of thermal perturbations.

We performed further tests to assess how much of the improvement could be attributed to the classifier method or to the score-based denoising idea. Using the same NequIP architecture, we trained a separate classifier based on the DC3 dataset (training details in the Methods section). What differentiates the NequIP-based denoiser and classifier is their training labels. The former was trained to predict the applied displacement vectors, which are continuous, unlimited, and information-rich targets. In comparison, the classifier had to match the structure class labels, which are discrete, far less informative, and limited to the training set. As shown in Fig. 3c, the classifier alone is reasonably accurate and retains more than 90% accuracy even above melting, which is very similar to the performance of the DC3 method [20]. Fig. 3f shows that even this high accuracy can be further improved to a “perfect” 100% well above melting by denoising. We conclude from Fig. 3 that score-based denoising does play a decisive role in achieving a very high level of accuracy for classifiers of various sophistication: the classification task was made almost trivial by denoising. Here we make a cautionary note about data-driven ML classifiers trained with labels. The NequIP

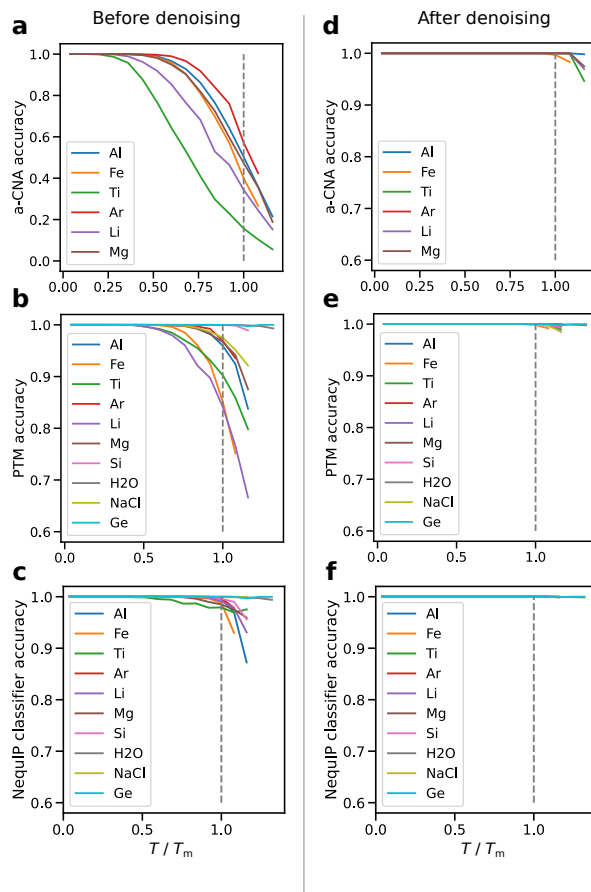


Figure 3: Classification performances of a-CNA, PTM, and an in-house NequIP classifier, on the DC3 benchmark dataset over a wide temperature range, are significantly improved after denoising. **(a, b, c)** a-CNA, PTM, and NequIP classification accuracies from nearly 0K to above melting. **(d, e, f)** Classification accuracies of the same models over the same temperature range, after denoising is applied. The melting point ($T = T_m$) is indicated by the dashed line.

classifier is obviously overconfident by predicting perfectly ordered solids well above melting in Fig. 3f, as point defects increasingly emerge at high temperatures as discussed previously (and shown in Supplementary Fig. 6). Such defects represent rare but incorrect labels that can mislead classifiers to ignore the defects during training. Mislabeling is a major headache for supervised learning methods and hard to systematically detect and correct in large training sets. In contrast, our self-supervised denoiser does not rely on class labels, and was able to reveal, combined with non-data-driven a-CNA or PTM, increasing existence of defects, as seen in Fig. 3d,e.

2.3 Denoising Cu solidification trajectory

Our denoiser is further tested here on an MD trajectory of Cu solidification from melt. Previously studied by Sadigh, et al. [32], solid nuclei appearing in the initial transient stages of Cu solidification are polymorphic, containing BCC, FCC, HCP, and disordered melt phase simultaneously. Characterization of such a complex transient behavior is challenging and presents a useful test case for our method. As shown in Fig. 4a, denoising four transient configurations of the trajectory results in drastic improvement in subsequent phase classification by the a-CNA algorithm. This improvement is

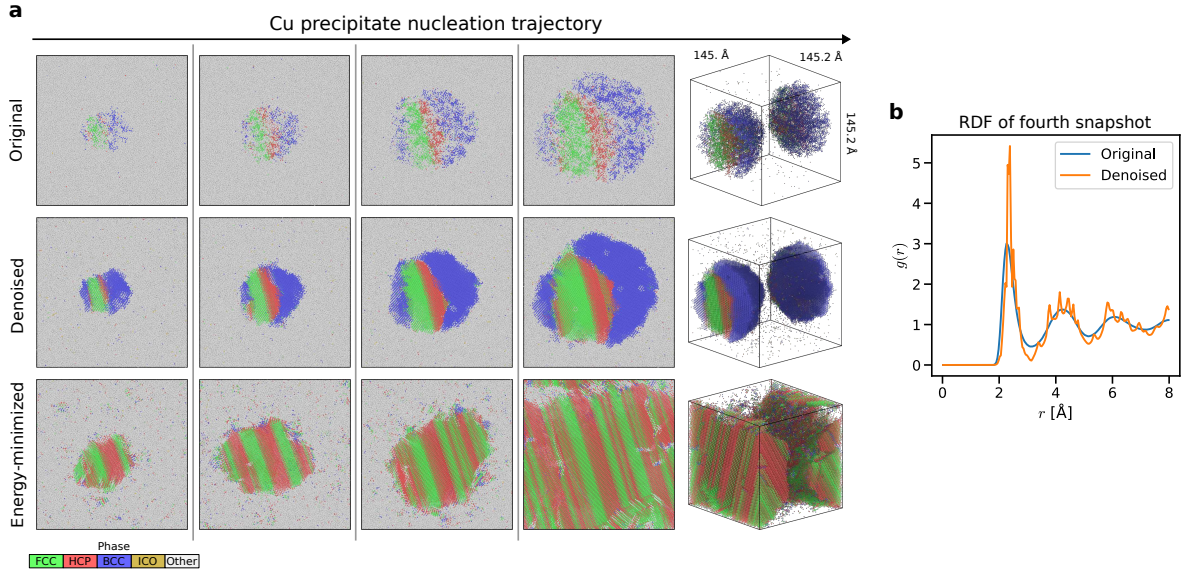


Figure 4: Denoising a dynamic trajectory of Cu solidification (314,926 atoms) significantly improves (a-CNA) phase classification without introducing unreasonably short interatomic distances. **(a)** Four consecutive snapshots along the trajectory, in original, denoised, and energy minimized states, are shown in orthogonal views with an additional perspective view for the fourth snapshot. **(b)** RDF of the fourth snapshot before and after denoising. In the perspective view, the atoms classified as *other* by a-CNA are rendered transparent.

manifested in much denser, correctly classified labels (FCC, HCP, and BCC) on the atoms within the ordered solid nucleus, as well as in considerably sharper boundaries between the phases. Here again the atoms in the disordered melt remain disordered as their a-CNA labels remain largely unchanged. Notably, a few atoms labeled as *other* (i.e., unknown) are observed within the sharply defined crystal phases even after denoising. These “unknown” atomic motifs turn out to be point defects, e.g. vacancies and interstitials, as will be discussed in greater detail in the next section. Also, they appear mostly in the BCC phase likely due to the metastability of the BCC phase in Cu entropically stabilized under the high pressure (70 GPa) of the simulation. Otherwise the same BCC phase of Cu is only marginally metastable as manifested in the appearance of soft modes in its phonon spectrum. At high temperatures close to melting, such soft modes may well result in some of the atoms within the BCC phase to significantly deviate from their ideal lattice positions, resulting in formation of point defects.

It can be informative to compare denoising to steepest descent energy minimization (EM), another common method for filtering out thermal vibrations. In stark contrast to denoising, EM not only reduces the thermal perturbation but also greatly changes the nucleus structure beyond recognition (third row of Fig. 4a). Namely, under EM the solid nucleus grows considerably larger, and the transient BCC phase disappears in favor of the more stable FCC and HCP phases. The striking difference between EM and denoising can be attributed to their contrasting assumptions. The EM mapping clearly favors solid phases of lowest ground state energy and may overzealously nudge atoms towards such phases, exactly as it happens in the considered example. The denoising process, from a complementary and purely geometric perspective, relies on less biased, equal prior probabilities of the reference phases included in training.

To further investigate whether denoising introduces unwanted or unphysical artifacts, in Fig. 4b we plot the RDF of the last snapshot from Fig. 4a. The RDF of the denoised structure generally matches that of the original. Having been trained to reduce thermal perturbations, the denoiser likely

has also learned not to bring atoms to excessively short distances of each other (that would be unphysical). The sharper peaks of the denoised RDF are attributed to the three ordered crystal phases, each of which contributes its own discrete set of sharply defined interatomic distances.

Note that the denoiser does in fact “denoise” some atoms within the melt phase into local environments regarded as crystalline (BCC/FCC/HCP) by the a-CNA classification, as evidenced by the increased number of solid labels in the melt region after denoising (Fig. 4a). The appearance of such “crystalline” atoms reflects that even in fully disordered liquid, statistically a small fraction of the atomic motifs may momentarily resemble a crystal phase. Our denoiser then acts locally and further enhances such resemblance, thus making a-CNA (a strictly local classifier itself) to recognize such atoms as crystalline. Observing that these misclassified atoms are few and far isolated, we accept this minor artifact as a worthwhile trade-off. We emphasize that our current denoiser model has been exclusively trained on randomly displaced solid reference structures and no liquid structure information has been included. Therefore the denoised configurations in the melt phase do not bear direct physical significance. Future probabilistic models informed with liquid physics may be useful but they are beyond the scope of this work, which is about proving the principle that preprocessing thermally excited atomic configurations via iterative denoising can dramatically improve the performance of downstream classifiers, which we illustrate by identifying polymorphic solid phase distributions obtained during rapid phase transformation from the melt, and by locating defect (non-crystalline) atoms in plastically deformed crystals at elevated temperatures.

2.4 Denoising FCC, HCP, and BCC Cu containing point defects

As an example of point defect characterization, the denoiser is applied to FCC, HCP, and BCC Cu each containing an intentionally inserted extra atom followed by annealing in MD at 3400 K ($1.1T_m$) for 400 ps. These structures were denoised into ideal lattices with local regions of disorder unknown to a-CNA (Fig. 5). The Wigner-Seitz defect analysis (OVITO) confirms that these regions of disorder indeed correspond to point defects. Notably, for the FCC crystal, the thermal vibrations were sufficiently intense to spontaneously generate two more point defects, a Frenkel pair of one vacancy and one interstitial.

The example in Fig. 5 demonstrates desirable outcomes of denoising crystal structures containing point defects. Although the denoiser aims to modify local atomic motifs towards ideal topology, it cannot do so on regions of point defects simply due to extra or missing atoms. In such a case, the model appears to not significantly alter the local topologies around the defects while denoising the rest of the bulk into an ideal lattice.

2.5 Denoising BCC Ta containing dislocations

Our model is similarly effective for denoising structures containing lattice dislocations. For a toy example, a hexagon-shaped dislocation loop inserted into BCC Ta was annealed at 2500 K ($0.8T_m$) and subsequently denoised (Fig. 6a). Again, satisfyingly, the denoiser does not significantly alter local atomic configurations near the dislocation loop while cleanly denoising the surrounding crystal bulk.

Note that although the denoiser was trained on ideal and synthetically noised Cu lattices, it is applicable to Ta or any other elemental crystal of FCC, BCC or HCP lattice structure. This transferability is achieved by simply scaling the input structure to match the interatomic distance of the corresponding Cu phase. The output structure would then be re-scaled back to its original dimensions.

As a more realistic and difficult test, the denoiser was applied to help reveal a complex dislocation network in a BCC Ta crystal subjected to plastic deformation at 2000 K or $0.6T_m$ (Fig. 6b). Similar

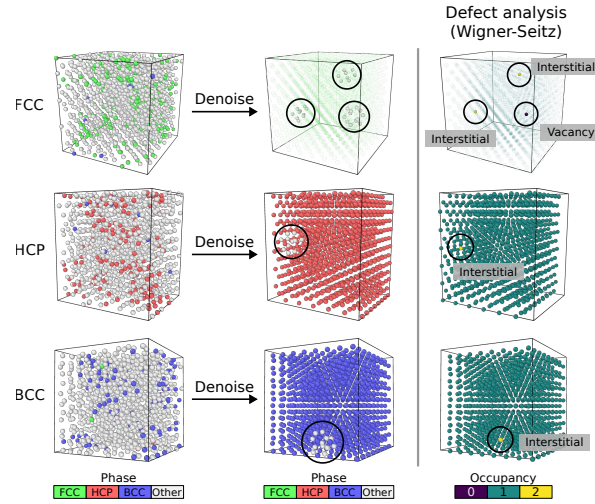


Figure 5: Denoising facilitates point defect identification in thermally perturbed FCC, HCP, and BCC Cu crystals. Point defects completely obscured by thermal vibrations are clearly revealed after denoising (circled in black). When applied to the denoised structures, the Wigner-Seitz defect analysis (OVITO) correctly assigns mass content to each crystal defect: site occupancies 0, 1, and 2 correspond to a vacancy, a regular atom, and an interstitial, respectively. After denoising, the regular atoms in the FCC crystal (top row) are rendered semi-transparent to more clearly reveal the point defects.

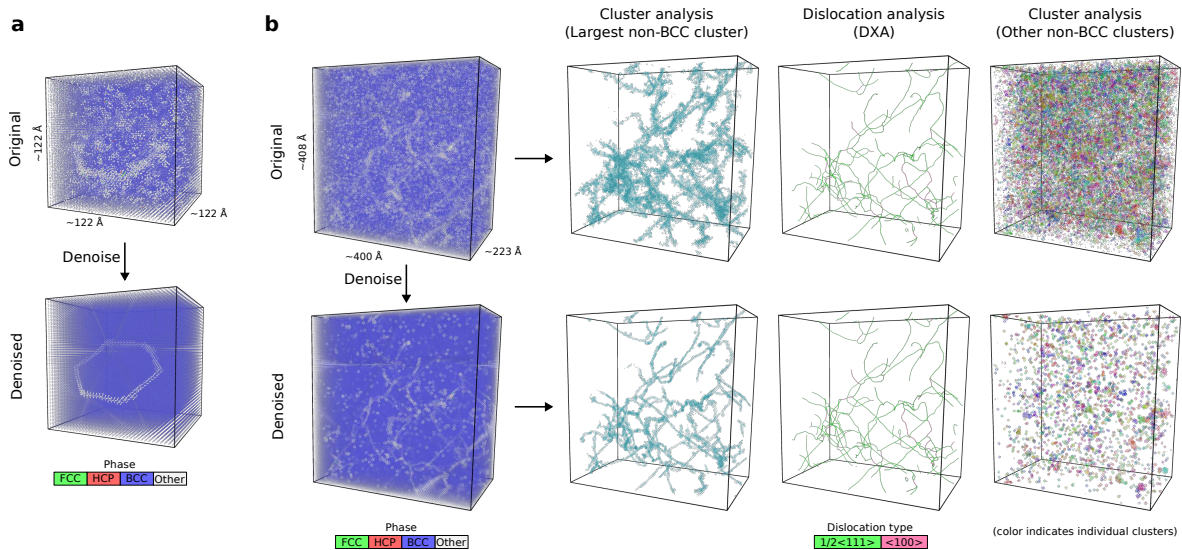


Figure 6: Denoising BCC single crystal Ta containing dislocations (and point defects). **(a)** A single dislocation loop annealed at 2500 K ($0.8T_m$) for 50 ps. **(b)** A snapshot of a relatively large MD simulation (1.97 million atoms) of single crystal Ta undergoing deformation at 2000 K ($0.6T_m$), which results in a network of entangled dislocation lines and a large number of point defect clusters. The DXA method for dislocation analysis (OVITO) confirms the presence of dislocations and delineates the dislocation network topology. The cluster analysis (OVITO) aims to separate the point defect clusters from the dislocations. The atoms are normally colored by a-CNA prediction except for the cluster/dislocation analysis, in which the BCC atoms are rendered transparent.

to the case of the single dislocation loop, the dislocation network (as colored by a-CNA) is more sharply defined after denoising. Subsequent application of the DXA algorithm [14] to the original and the denoised configurations results in nearly identical dislocation networks, which testifies to the exceptional robustness of DXA against thermal perturbation, and confirms that the dislocations are better captured by a-CNA after denoising.

Despite DXA’s already high performance in dislocation characterization, the denoiser still benefits or complements DXA by facilitating the characterization of the point defect clusters that were either left as debris in the wake of dislocation motion or produced by dragging jogs formed at dislocation intersections [33]. Focusing on non-BCC atoms clustered within a cutoff distance of 3.2 Å, a large cluster (corresponding to the dislocation network) and a high concentration of small clusters are observed. Without denoising, the small clusters may simply be manifestation of perturbation based on visual interpretation. However, after denoising, the small clusters resemble and likely capture the point defect debris. This is unlikely a case of the denoiser failing to denoise non-dislocation regions into perfect lattice for two reasons: (1) the presence of the point defect debris is known *a priori*, and (2) the denoiser clearly denoises the non-dislocation region in the toy case of the single dislocation loop, with virtually no point defects left (Fig. 6a).

2.6 Denoising BCC Ta containing grain boundaries

To test how our method performs on crystals containing grain boundaries, the denoiser was applied to a Ta bi-crystal containing two tilt boundaries. Prior to denoising, the bi-crystal was annealed at a high temperature of 2500 K ($0.8T_m$) for 50 ps. As shown in Fig. 7a, denoising does not alter the topology of the defects, and results in two near-perfect BCC crystals separated by two perfectly planar grain boundaries, with trace amount of point defects likely emitted from the boundaries into the bi-crystal interior. A more complex test case is shown in Fig. 7b, where a polycrystal consisted of 12 grains of BCC Ta had been similarly annealed at 2500 K ($0.8T_m$) for 50 ps. Here again, denoising removes thermal perturbation while still revealing a few point defects in the grain interiors.

2.7 Denoising SiO₂ polymorphs

Finally, to validate the generalizability of our approach, we trained a separate denoiser for minimizing thermal perturbation in SiO₂ systems. Similar to the demonstration shown in Fig. 2, the denoiser was applied to MD-perturbed high-temperature silica polymorphs β -quartz, α -cristobalite, and β -cristobalite (Fig. 8). Although we also trained the denoiser with reference α -quartz topology, the inherent similarity between α - and β -quartz complicates the task of generating stable and distinctive MD-perturbed snapshots of α -quartz compared to β -quartz. Therefore α -quartz is not included in this preliminary result. Regardless, the denoiser removes virtually all thermal perturbation in the silica polymorphs. Future work may include extending to larger multi-element systems in the presence of disordered features.

3 Discussion

In developing our method we have taken to a geometric and statistical perspective on delineating order-disorder features in MD simulations of solids, a problem particularly difficult at elevated temperatures approaching the melting point. We regard two seemingly distinct tasks of classifying ordered phases and locating disordered defects as fundamentally the same problem, which is addressed by denoising. Based on a statistical score function, the denoiser presented and tested in this work effectively reduces and minimizes thermal perturbation in ordered solids without impacting isolated

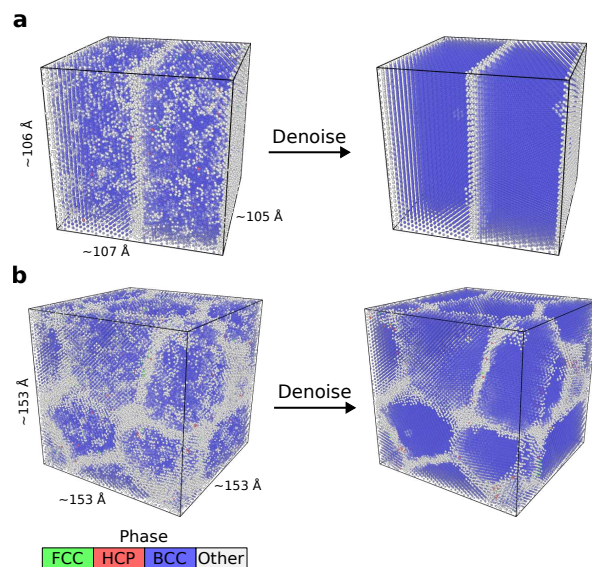


Figure 7: The denoising process helps reveal grain boundaries obscured by thermal perturbation. **(a)** A bi-crystal of BCC Ta (64,000 atoms) containing two planar grain boundaries. **(b)** A polycrystal of BCC Ta (187,921 atoms) containing a network of grain boundaries. Both crystals were annealed in MD simulations at 2500 K ($0.8T_m$) for 50 ps which caused minor coarsening in the polycrystal. After denoising, trace amount of point defects becomes visible in both examples. The atoms are classified by a-CNA. The BCC atoms are rendered slightly semi-transparent.

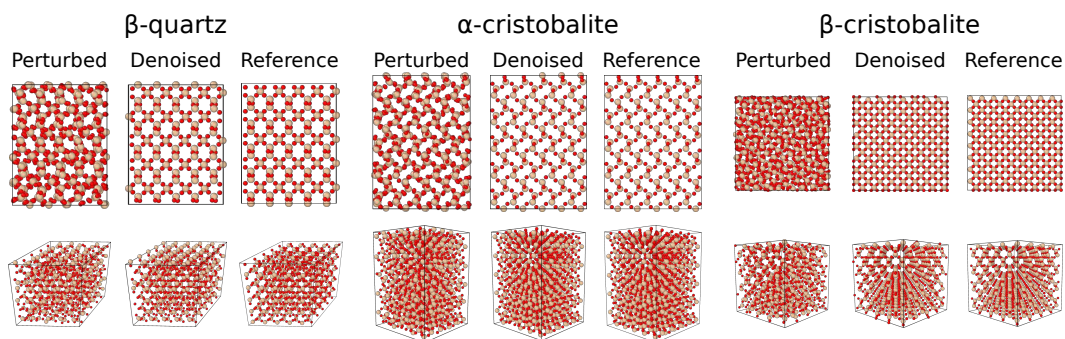


Figure 8: A separately trained denoiser effectively removes thermal perturbations (simulated at 1000 K or roughly $0.5T_m$) in SiO_2 β -quartz, α -cristobalite, and β -cristobalite polymorphs. The denoised structures match the ideal references used during training. 16 iterations were used for denoising. The silicon atoms are colored in beige, and the oxygens in red.

disordered defects and the liquid/melt phase. Combined with various crystal classifiers, it was able to achieve perfect classification accuracy in a public benchmark dataset over existing methods. Further tests against no or single-step denoising as well as supervised classifier based on the same network architecture provided more evidence for the effectiveness of our approach. To support our conclusions, in the preceding sections we applied denoising as a pre-processor to reveal the underlying “anomaly” structures across the entire spectrum of crystal disorder, namely (0D) point defects, (1D) dislocation lines, (2D) grain boundary, and (3D) liquid phases, all distinct from the ordered phases that the denoiser was trained with. For future work, more comprehensive and quantitative effort is necessary to validate the merit of the denoiser as a pre-processor for defect characterization methods. For denoising BCC/FCC/HCP, our model was trained only once and only on the three ordered structures but then shown to successfully but not excessively denoise the much more complex structures used for testing. We relate this useful ability of not overly denoising defective atoms to the inductive bias of our graph network model with its limited number of message-passing steps. Trained entirely on synthetic data, our model is not derived from any deep physical insights about topology and geometry of the reference structures and, as such, does not require careful data curation. Thus, its extension to other ordered structures should be straightforward.

Our method holds unique advantages and disadvantages over two other methods widely used for reducing thermal perturbation: energy minimization (EM) and vibration-averaging (VA). With the right parameters and done over a small number of iterations, EM can lead to similar results as that of the denoiser. However, over many iterations, EM may grow or shrink certain phases due to the intrinsic bias associated with the use of interatomic potential. An example of EM being overzealous and distorting configurations beyond recognition with respect to the original state is in Section 2.3. The purely geometric denoiser, on the other hand, does not require a known or developed interatomic potential. Further, the denoiser is unbiased in the sense that the reference lattices used in training have equal prior probability. Although the denoising graph model is more expensive than VA (which entails simple averaging operations), VA can potentially smear out atomic motion by averaging over a time interval. Such a distortion can be fairly significant since, to average out thermal vibrations, VA requires time averaging intervals of hundreds or even thousands of time steps. Our denoising method, on the other hand, treats every time snapshot separately and does not coarse-grain over time. As is often the case, no one method used for denoising is singularly superior to all other existing methods. We hope that our approach finds its own place among existing and emerging methods for structure and defect classification, and can serve as an accurate and efficient pre-processing filter to facilitate application of more computationally demanding methods of structural analysis such as DXA.

Based on an equivariant graph network model architecture, our denoising model is readily extendable to more complex reference structures and materials by incorporating additional information such as atom types into the graph embedding. In addition to extending the method to chemically complex systems, our ongoing and future efforts may also focus on its computational efficiency and scalability. Finally, beyond atomic structures, the ability of our model to achieve state-of-the-art classification accuracy through iterative denoising score functions suggests the idea may be useful in other disciplines for enhanced accuracy and/or robustness against fluctuations.

4 Methods

4.1 Theoretical justification of the denoising model

Our approach to denoising builds on the theory of statistical learning of score functions [25] that establishes equivalence between denoising and score matching. Consider a probability distribution function $q(\mathbf{x})$ that exists in principle but is analytically intractable due to the high dimensionality of

the data space $\mathbf{x} \in \mathbb{R}^d$ ($d \gg 1$). Focusing on approximating the gradient of the log-probability density $\nabla_{\mathbf{x}} \log q(\mathbf{x})$ —also known as the score function [34]—rather than $q(\mathbf{x})$ itself circumvents the often intractable problem of finding the normalization constant for $q(\mathbf{x})$. Score matching then amounts to finding an approximating model $s_{\theta}(\mathbf{x})$ with parameters θ to match the score function, with the *score matching* loss [34]

$$L_{\text{SM}} = \frac{1}{2} \mathbb{E}_{q(\mathbf{x})} \left[\|s_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x})\|^2 \right]. \quad (1)$$

Nevertheless, the term $\nabla_{\mathbf{x}} \log q(\mathbf{x})$ is still unknown. To address this, consider approximating $q(\mathbf{x})$ by adding isotropic Gaussian noises of variance σ^2 to the clean data samples \mathbf{x} , resulting in noised samples $\mathbf{x}' = \mathbf{x} + \sigma \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and the approximating distribution

$$q_{\sigma}(\mathbf{x}'|\mathbf{x}) = Z e^{-\|\mathbf{x}' - \mathbf{x}\|^2 / 2\sigma^2}, \quad (2)$$

where Z is a normalization constant. This way, instead of the original loss, we minimize the *denoising score matching* loss based on the key insight from Ref. [25] to train with pairs of clean and corrupted data points:

$$L_{\text{DSM}} = \frac{1}{2} \mathbb{E}_{q_{\sigma}(\mathbf{x}'|\mathbf{x})} \left[\|s_{\theta}(\mathbf{x}') - \nabla_{\mathbf{x}'} \log q_{\sigma}(\mathbf{x}'|\mathbf{x})\|^2 \right], \quad (3)$$

where the new score function $\nabla_{\mathbf{x}'} \log q_{\sigma}(\mathbf{x}'|\mathbf{x})$ can be computed via

$$\begin{aligned} \log q_{\sigma}(\mathbf{x}'|\mathbf{x}) &= \log(Z) - \frac{1}{2\sigma^2} \|\mathbf{x}' - \mathbf{x}\|^2, \\ \nabla_{\mathbf{x}'} \log q_{\sigma}(\mathbf{x}'|\mathbf{x}) &= -\frac{1}{\sigma^2} (\mathbf{x}' - \mathbf{x}) = -\frac{\boldsymbol{\epsilon}}{\sigma}, \end{aligned} \quad (4)$$

revealing that the score function points from noisy samples \mathbf{x}' to clean ones \mathbf{x} . This observation also implies that learning the score function is equivalent to training a denoising model. To better see this, note that the denoising score matching loss can now be simplified into

$$L_{\text{DSM}} = \frac{1}{2} \mathbb{E}_{q_{\sigma}(\mathbf{x}'|\mathbf{x})} \left[\left\| s_{\theta}(\mathbf{x}') + \frac{\boldsymbol{\epsilon}}{\sigma} \right\|^2 \right]. \quad (5)$$

After scaling equation (5) by a factor of σ and incorporating a noise prediction model $\boldsymbol{\epsilon}_{\theta}(\mathbf{x}') = -\sigma s_{\theta}(\mathbf{x}')$ that aims to predict the applied noise, then the loss function can be written as

$$L_{\text{DSM}} = \frac{1}{2} \mathbb{E}_{q_{\sigma}(\mathbf{x}'|\mathbf{x})} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta}(\mathbf{x}')\|^2 \right], \quad (6)$$

finally establishing the connection between the score function model $\mathbf{s}_{\theta}(\mathbf{x}')$ and the denoising model $\boldsymbol{\epsilon}_{\theta}(\mathbf{x}')$ by $\boldsymbol{\epsilon}_{\theta}(\mathbf{x}') = -\sigma \mathbf{s}_{\theta}(\mathbf{x}')$. This clarifies the meaning of $\boldsymbol{\epsilon}_{\theta}$ in Algorithm 1: it is the scaled score function defining the noise added to clean data. While the noise amplitude σ is a hyper-parameter, it can be estimated by a fitted/trained denoising model itself from a noisy input. Hereafter, we refer to $\boldsymbol{\epsilon}_{\theta}$ as a score function for brevity. Along the above steps for connecting between score matching and denoising, we have omitted certain details for brevity. For a rigorous formulation, see [25].

The ideal score function is a gradient field in the data space that converges to clean data points used to train the denoiser. For better intuition, a toy score function is visualized in Supplementary Fig. 1, which illustrates that following the score function is the same as denoising. In our context where ideal FCC, HCP and BCC lattices were used for training, a perturbed (noised) input configuration may be denoised into one of the three ideal structures that it resembles the most. At the same time, a highly perturbed input configuration bearing no resemblance to any of the three ideal reference configurations is unlikely to be meaningfully denoised resulting in unknown or divergent

values of predicted noise. The case studies considered in this work all suggest that our denoising model does not significantly alter such disordered structures, including melt, point defects, dislocations, and grain boundaries. This property is instrumental in allowing the denoising model to reveal underlying crystalline order without impacting meaningful disordered features in thermally perturbed configurations.

The score function plays a central role in modern likelihood-based generative models such as the denoising diffusion probabilistic model (DDPM) [26, 27] and score-based generative model [28], which can be unified under the same framework [28]. Among its numerous recent achievements [29], DDPM has been applied to crystal and molecular structure generations [35, 36]. In this work, however, we apply the score matching method for denoising rather than generative applications, and focus on a limited number of reference crystal structures instead of many (thousands or millions) training images/structures.

Now the similarity and distinction between our score-based denoiser and energy minimization are clear from a statistical point of view. Conceptually and heuristically both move atoms to make a structure more “reasonable” or “likely” with regard to a predefined probability. Quantitatively both are driven by a vector field proportional to $\nabla_{\mathbf{x}} \log P(\mathbf{x})$, i.e. a score. Our denoiser is unbiased towards any specific structure type $\mathbf{x}^{(i)}$ since the probability $P \propto \sum_i \mathcal{N}(\mathbf{x}^{(i)}, \sigma^2)$, a kernel density estimation using Gaussian kernel, is unbiased with equal probability at any structure $P(\mathbf{x}^{(i)})$. In contrast, for energy minimization the Boltzmann distribution $P \propto \exp[-U(\mathbf{x})/k_B T]$ always favors low energy structures.

4.2 Model training

Our clean data samples are reference crystal structures of interest represented by the atomic coordinates \mathbf{r} and the auxiliary information \mathbf{z} : $\mathbf{x}^{(i)} \rightarrow (\mathbf{r}^{(i)}, \mathbf{z}^{(i)})$. The noise prediction model $\boldsymbol{\epsilon}_\theta$ was trained with entirely synthetic data (Algorithm 2), which is generated by adding Gaussian noises to the atomic coordinates $\mathbf{r}' = \mathbf{r} + \sigma \boldsymbol{\epsilon}$, with $\sigma \sim \mathcal{U}(0, \sigma_{\max})$ drawn uniformly up to $\sigma_{\max} \approx 13\%$ of the shortest interatomic distance, adhering to Lindemann’s law on mean-squared thermal displacement of solids before melting [37]. Since the denoising method is always applied iteratively to remove thermal fluctuation in structures and each time there is a decreasing amount of fluctuation remaining, we trained the score model with different levels of noise for better generalizability. In spite of the vibrational degrees of freedom being strongly coupled in solids, as schematically depicted in Fig. 1b, it nevertheless suffices for the denoising model to be trained solely on artificial i.i.d. Gaussian random displacements, so-called Einstein vibrational modes. It is reasonable to argue that so long as the chosen displacement amplitude σ_{\max} contains the thermal displacements observed in MD, no matter how correlated they might be, the denoiser should be able to converge to correct reference structures. Our implemented loss function, which scales equation (6) by $2\sigma^2$, is

$$L = \mathbb{E}_{\mathbf{r}, \mathbf{z}, \sigma, \boldsymbol{\epsilon}} [\|\sigma \boldsymbol{\epsilon} - \boldsymbol{\Delta}_\theta(\mathbf{r} + \sigma \boldsymbol{\epsilon}, \mathbf{z})\|^2], \quad (7)$$

where $\boldsymbol{\Delta}_\theta = \sigma \boldsymbol{\epsilon}_\theta$ predicts not only the perturbation direction $\boldsymbol{\epsilon}$ but amplitude σ .

The idea of mixing training data with random noises is not new in either general-purpose or scientific machine learning. Adding a small amount of noise to the training data, known sometimes as the “noise trick”, is a well-established data augmentation or regularization technique in general-purpose machine learning to reduce overfitting and increase model robustness [38]. For example, Zhou et al. [39, 40] adopted a hybrid training data pipeline of MD trajectories and Gaussian noise displacements to fit the potential energy surface of crystalline solids. A similar method was used by Chung et al. [20] to identify ordered solid phases. The noise trick was also adopted to train GNN

Algorithm 2 Denoiser training

Require: Ideal reference dataset \mathbf{D} , initial model parameters θ , maximum noise amplitude σ_{\max} , gradient descent optimizer Optim, and learning rate η

repeat

 Sample $\mathbf{x} \sim \mathbf{D}$

 Sample $\sigma \sim \mathcal{U}(0, \sigma_{\max})$

 Sample $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$

$(\mathbf{r}, \mathbf{z}) \leftarrow \mathbf{x}$

$L \leftarrow \mathbb{E}_{\mathbf{r}, \mathbf{z}, \sigma, \boldsymbol{\varepsilon}} [\|\sigma \boldsymbol{\varepsilon} - \Delta_{\theta}(\mathbf{r} + \sigma \boldsymbol{\varepsilon}, \mathbf{z})\|^2]$

$\theta \leftarrow \text{Optim}(L, \theta, \eta)$

until convergence

surrogate models for physical simulations [41, 42]. This work, however, makes denoising perturbed inputs the centerpiece rather than merely a regularization technique.

The model was trained with randomly drawn $\mathbf{x} \in \{\text{FCC}, \text{BCC}, \text{HCP}\}$, using the AdamW optimizer [43] and a learning rate of 2×10^{-4} , over 20,000 weight updates in minibatches of 32 samples. Each FCC/HCP/BCC cell consists of roughly 1000 atoms. The training was carried out using PyTorch [44] and PyTorch-Geometric [45]. All other training parameters, if unspecified in this work, default to values per PyTorch 1.11.0 and PyTorch Geometric 2.0.4.

4.3 Equivariant graph network model (NequIP)

The development of a score model is entirely similar to that of an interatomic potential, as both are tasked with predicting a vector field (score vs. force) as the gradient of a scalar quantity (logarithmic probability vs. total energy) with respect to atomic coordinates. Before delving into the details of the adopted graph neural network (GNN) model, we provide a generic overview of the GNN architecture. We start with a neighbor or edge list $\{ij\}$ based on a distance cutoff, as well as atomic or node features \mathbf{h}_i (an array of atomic attributes such as partial charge and atomic number). For each edge ij , the effects of j on i , called a message of the edge, is calculated $\mathbf{m}_{ij} = M(\mathbf{h}_i, \mathbf{h}_j, \mathbf{r}_j - \mathbf{r}_i)$. Then the messages are aggregated, typically through summation, into node i : $\mathbf{m}_i = \sum_j \mathbf{m}_{ij}$, and the node features are updated using its current values and the aggregated message $\mathbf{h}'_i = U(\mathbf{h}_i, \mathbf{m}_i)$. This so-called message-passing step is similar to and a generalization of the EAM potential, which models the charge contribution from j to i , aggregates the contributions to i , and describes the aggregated effects with an embedding function. The difference is that GNN adopts machine-learned message function M and update function U and applies multiple message-passing operators iteratively. Afterwards, GNN computes per-node contributions through a decoder $D(\mathbf{h}_i)$.

The denoising model output is a vector (on each atom) that should be equivariant under translation, rotation and mirror operations—the same requirements for force fields or interatomic potentials. We adopted a customized version of the E(3)-equivariant NequIP model [23, 24], which guarantees such equivariance. NequIP is primarily built upon the idea of equivariant tensor product between two inputs of irreducible representations, or irreps, resulting in another irreps as the output. Unlike regular tensor products, the tensor products in NequIP are parametrized by learnable weights and are therefore termed *WeightedTP* in this work. Since the exact mathematical details of the equivariant tensor product can be dense and complex, we refer to the original work for their precise description [24]. We chose to directly predict vector outputs (noising displacements) rather than a scalar output.

The main components of our NequIP variant consist of the initial embedding, the interaction layers, and the final self-interaction layer to produce the noise output (Fig. 9).

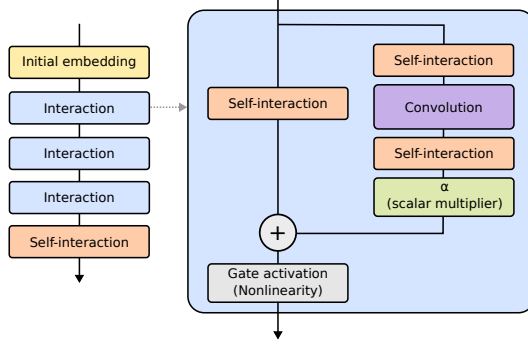


Figure 9: The NequIP model architecture.

In the initial embedding, the structure input is converted to an atomic graph, with \mathbf{h}_i as some attributes for node or atom i ; $\tilde{\mathbf{h}}_i$ as another set of attributes for the same node i ; and \mathbf{e}_{ij} as the vector for the *directed* edge from node i to node j . Transformed from atom type information by a trainable embedding matrix, \mathbf{h}_i initially only holds scalar information ($l = 0$, where l is the tensor rank or the degree of representation) but is typically expanded to hold information of higher tensor ranks ($l = 1, 2, \dots$) in subsequent layers. $\tilde{\mathbf{h}}_i$, also transformed from atom type information by a trainable embedding matrix, holds only scalar information and does not change throughout the model layers. The embedding matrix is not needed for the single-element systems but necessary for the SiO_2 polymorphs. We used Atomic Simulation Environment [46] and PyTorch-Geometric [45] for the conversion to graphs.

Each interaction layer consists of several sub-operations: the self-interaction, the convolution, SkipInit [47], and the gate activation. The self-interaction updates the attributes of each node \mathbf{h}_i via the WeightedTP operation with $\tilde{\mathbf{h}}_i$ and does not aggregate information from neighbor nodes:

$$\mathbf{h}'_i = \text{WeightedTP}(\mathbf{h}_i, \tilde{\mathbf{h}}_i) \quad (8)$$

The convolution updates the attributes of each node \mathbf{h}_i as the sum of the WeightedTP operations between the neighboring nodes \mathbf{h}_j and the spherical harmonics of neighboring edges $Y(\hat{\mathbf{e}}_{ij})$, with the weights learned from the edge distances $\|\mathbf{e}_{ij}\|$ via a multilayer perceptron (MLP):

$$\mathbf{h}'_i = \frac{1}{Z} \sum_{j \in N(i)} \text{WeightedTP}_{\|\mathbf{e}_{ij}\|}(\mathbf{h}_j, Y(\hat{\mathbf{e}}_{ij})) \quad (9)$$

where $\hat{\mathbf{e}}_{ij}$ is the normalized version of \mathbf{e}_{ij} and is therefore a unit vector pointing from node i to node j , $N(i)$ denotes the neighbor nodes of node i , and Z is a normalization constant. The MLP contains one hidden layer. The initial layer of the MLP is the basis function values expanded from the edge distance. For the SkipInit mechanism [47], the scalar multipliers α are learned from yet another WeightedTP operation between \mathbf{h}_i and $\tilde{\mathbf{h}}_i$ (similar to the self-interaction operation). The gate activation applies equivariant nonlinearities [48] to the node attributes.

In the end, the final self-interaction layer serves to transform the node attributes $\mathbf{h}_i^{(L-1)}$ from the second to last layer $L - 1$, which may hold scalar, vectorial, and tensorial features at node i , into a single vector describing the noise output:

$$\mathbf{h}_i^{(L)} = (\boldsymbol{\epsilon}_\theta)_i = \text{WeightedTP}(\mathbf{h}_i^{(L-1)}, \tilde{\mathbf{h}}_i^{(L-1)}), \quad (10)$$

The complexity of the model is largely determined by the specified irreps format for the node and edge attributes. For example, an array of 4 scalars and 8 vectors can be written as $4 \times 0e + 8 \times 1o$,

with the numbers 4 and 8 describing the multiplicities, the numbers 0 and 1 describing the tensor rank, and the letters e (even) and o (odd) describing the parity. Higher multiplicities and tensor ranks can often result in better performance but also larger memory and computational requirements. We intentionally kept the model complexity small in favor of scalability to structures of millions of atoms. Here we provide rough estimates of the runtime numbers of the NequIP-based denoiser. However, we caution that compute speed is not a major focus of this preliminary work, especially since there can be speed limitations with NequIP. We have since focused on optimizing for speed with efforts such as parallelizing the workload across multiple GPUs, but such efforts are perhaps better suited for a future report. Also, the same denoiser formulation can be implemented with other more optimized graph network models. For 1000 atoms or less, the current denoiser model takes sub-second to run on a typical modern CPU. For 1000–100,000 atoms it takes sub-minute. For 1–10 million atoms it can take 1–20 minutes. Again, we expect that these runtimes can be much further optimized. The model settings are listed in Supplementary Table 1.

4.4 NequIP classifier

The NequIP-based classifier was trained on 90% of the Al, Fe, Ti, Si, H₂O, and NaCl snapshots from the DC3 benchmark dataset, with the rest being the validation dataset. Specifically, for each of the Al, Fe, Ti, Si, H₂O, and NaCl systems, there are ten snapshot files per temperature increment, and nine were taken for training, with the remaining one for validation. In other words, this classifier was not trained on any of the Ar, Li, Mg, and Ge structure data from the DC3 dataset. Applying the classifier to Ar, Li, Mg, and Ge amounts to scaling the unit cell dimensions to match that of the corresponding training data with the same structure type. For example, Ge is slightly larger than Si, and therefore is scaled down (roughly 90%) before feeding to the classifier. The model parameters of the NequIP classifier are listed in Supplementary Table 1.

4.5 Molecular dynamics simulations

This section describes the MD simulations used to demonstrate the capabilities of our denoising method. These are simulations of (1) BCC, FCC and HCP Cu structures, both defect-free and with point defects; (2) a solid crystal nucleus growing inside melted Cu; (3) crystal plasticity in single crystal of Ta in the BCC phase; (4) Ta grain boundaries; and (5) SiO₂ in β -quartz and cristobalite polymorphs. All simulations were performed with periodic boundary conditions using LAMMPS [49].

The MD simulations for Cu using the embedded-atom method (EAM) potential by Mishin *et al.* [50] were performed in BCC, FCC and HCP cells containing 1024, 1372, and 1152 atoms respectively. Since the bulk BCC phase is dynamically unstable in Cu with imaginary phonon frequencies at ambient conditions, the calculations were performed at a pressure of 60 GPa, where the BCC phase becomes metastable.

Melting points of the three phases at 60 GPa have been calculated to be 3030 K for BCC, 3066 K for HCP, and 3073 K for FCC [32]. Although FCC remains the thermodynamically stable and thus preferred phase below 3073 K, free energies of three phases are very close under these pressure and temperature conditions. At slightly higher pressures (71.6 GPa and 85 GPa), the phase diagram of the model of Cu contains triple points where two of the three solid phases and the liquid phase coexist [32]. We have taken advantage of these thermodynamic proximity of three crystal phases and the melt to set up an MD simulation of a polymorphic critical solid nucleus simultaneously containing all three solid phases surrounded by melt. The simulation contained 314,928 Cu atoms and was initiated in an isobaric-isoenthalpic (NPH) ensemble at 70 GPa from a small near-equilibrium nucleus with

coexisting FCC and HCP ordered regions containing about 200 FCC atoms and 300 HCP atoms, respectively. Upon switching to an isobaric-isothermal (NPT) ensemble at the same pressure and temperature 2800 K, the solid nucleus grows and partially transforms to the BCC phase.

Interatomic interactions in tantalum were modeled using a well-known EAM potential developed by Li et al. [51]. For simulations of crystal plasticity in Ta described in Section 2.5, the crystals were created by arranging atoms in a BCC lattice within a cubic or an orthorhombic periodic supercell with repeat vectors aligned along the cube axes of the BCC lattice. Dislocations were seeded into the crystals in the form of one or several hexagon-shaped prismatic loops of the vacancy type, following the procedure introduced in [52]. For the configuration in Fig. 6a, a single dislocation loop was introduced at the center of a cube-shaped simulation box made of 101,853 atoms and annealed at temperature 2000 K. The complex network of dislocations shown in Fig. 6b was generated by initially introducing 12 randomly positioned dislocation loops into a ~ 2 million atoms box, annealing the model at 2500 K and zero pressure and then subjecting the crystal to uniaxial compression along the [001] crystallographic axis at a “true” strain rate of 2×10^8 /s for 2 ns while maintaining pressure near zero in an NPH ensemble.

For Ta grain boundaries in Section 2.6, the periodic bi-crystal containing two $\Sigma 5(310)$ symmetric tilt grain boundaries was created by joining two crystal blocks of different lattice orientations obtained by rotating two half-crystals in the opposite directions along the common $\langle 100 \rangle$ tilt axis. The Ta polycrystal was assembled using atomsk [53] from 12 randomly seeded grains. Both the bi-crystal and the polycrystal were annealed at 2500 K and zero pressure.

For silica, the Tersoff potential developed by Munetoh et al. [54] was adopted. Simulations were performed in the NVT ensemble with unit cell parameters taken from the Encyclopedia of Crystallographic Prototypes [55–57], and with temperature ramping from 600 K to 2000 K in steps of 200 K every 10 ps. $5 \times 5 \times 5$, $5 \times 5 \times 5$, and $4 \times 4 \times 4$ supercells were used for the β -quartz (space group 181), α -cristobalite (92), and β -cristobalite (227), respectively.

Computational details for the DC3 dataset can be found in Ref. [20].

Acknowledgements

TH and FZ acknowledge support by the Critical Materials Institute, an Energy Innovation Hub funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, and Advanced Materials and Manufacturing Technologies Office. BS, CP, NB and VB are partially supported by the Laboratory Directed Research and Development (LDRD) program (22-ERD-016) at Lawrence Livermore National Laboratory. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under contract No. DE-AC52-07NA27344. JC was partially supported by the Department of Mechanical Engineering’s startup grant at Boston University. We thank Dr. R. Freitas for providing the DC3 benchmark dataset and Dr. A. Stukowski for useful discussions. Computing support for this work came from the LLNL Institutional Computing Grand Challenge program.

Author Contributions

TH implemented the denoising graph network model and analyzed the denoised MD configurations. BS, NB, and CP generated the MD data to be denoised. FZ supervised the research with inputs from all authors.

Data Availability

All data required to reproduce this work, namely the molecular dynamics snapshots (ideal reference, thermally perturbed, and denoised) can be requested from Tim Hsu and Fei Zhou. Some small, simple structure data for demonstration purposes are included in the code repository of this work (see Code Availability). The DC3 dataset used for benchmarking can be found at <https://github.com/freitas-rodrigo/DC3>.

Code Availability

The source code, the trained denoiser and classifier models, and a demo for denoising simple structures are available at <http://www.github.com/llnl/graphite>.

Competing interests

The authors state that there is no conflict of interest.

References

1. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28**, 784 (1983).
2. Honeycutt, J. D. & Andersen, H. C. Molecular dynamics study of melting and freezing of small Lennard-Jones clusters. *J. Phys. Chem.* **91**, 4950–4963 (1987).
3. Stukowski, A. Structure identification methods for atomistic simulations of crystalline materials. *Model. Simul. Mater. Sci. Eng.* **20**, 045021 (2012).
4. Tanaka, H., Tong, H., Shi, R. & Russo, J. Revealing key structural features hidden in liquids and glasses. *Nat. Rev. Phys.* **1**, 333–348 (2019).
5. Larsen, P. M., Schmidt, S. & Schiøtz, J. Robust structural identification via polyhedral template matching. *Model. Simul. Mater. Sci. Eng.* **24**, 055007 (2016).
6. Larsen, P. M. Revisiting the common neighbour analysis and the centrosymmetry parameter. *Preprint at <https://arxiv.org/abs/2003.08879>* (2020).
7. Ackland, G. & Jones, A. Applications of local crystal structure measures in experiment and simulation. *Phys. Rev. B* **73**, 054104 (2006).
8. Lazar, E. A., Han, J. & Srolovitz, D. J. Topological framework for local structure analysis in condensed matter. *Proc. Natl. Acad. Sci.* **112**, E5769–E5776 (2015).
9. Nguyen, A. H. & Molinero, V. Identification of clathrate hydrates, hexagonal ice, cubic ice, and liquid water in simulations: The CHILL+ algorithm. *J. Phys. Chem. B* **119**, 9369–9376 (2015).
10. Lechner, W. & Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **129**, 114707 (2008).
11. Kelchner, C. L., Plimpton, S. & Hamilton, J. Dislocation nucleation and defect structure during surface indentation. *Phys. Rev. B* **58**, 11085 (1998).
12. Sapozhnikov, F., Ionov, G. & Dremov, V. An adaptive template method for analyzing crystal structures and defects in molecular dynamics simulations of high-rate deformations. *Russ. J. Phys. Chem. B* **2**, 238–245 (2008).

13. Stukowski, A. & Albe, K. Extracting dislocations and non-dislocation crystal defects from atomistic simulation data. *Model. Simul. Mater. Sci. Eng.* **18**, 085001 (2010).
14. Stukowski, A., Bulatov, V. V. & Arsenlis, A. Automated identification and indexing of dislocations in crystal interfaces. *Model. Simul. Mater. Sci. Eng.* **20**, 085007 (2012).
15. Kim, Q., Ko, J.-H., Kim, S. & Jhe, W. GCIceNet: a graph convolutional network for accurate classification of water phases. *Phys. Chem. Chem. Phys.* **22**, 26340–26350 (2020).
16. Swanson, K., Trivedi, S., Lequieu, J., Swanson, K. & Kondor, R. Deep learning for automated classification and characterization of amorphous materials. *Soft Matter* **16**, 435–446 (2020).
17. Doi, H., Takahashi, K. Z. & Aoyagi, T. Mining of effective local order parameters for classifying crystal structures: A machine learning study. *J. Chem. Phys.* **152**, 214501 (2020).
18. Becker, S., Devijver, E., Molinier, R. & Jakse, N. Unsupervised topological learning for identification of atomic structures. *Phys. Rev. E* **105**, 045304 (2022).
19. Leitherer, A., Ziletti, A. & Ghiringhelli, L. M. Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning. *Nat. Commun.* **12**, 6234 (2021).
20. Chung, H. W., Freitas, R., Cheon, G. & Reed, E. J. Data-centric framework for crystal structure identification in atomistic simulations using machine learning. *Phys. Rev. Mater.* **6**, 043801 (2022).
21. Hernandez, V. F., Marques, M. S. & Bordin, J. R. Phase classification using neural networks: application to supercooled, polymorphic core-softened mixtures. *J. Phys. Condens. Matter* **34**, 024002 (2021).
22. Chapman, J., Goldman, N. & Wood, B. C. Efficient and universal characterization of atomic structures through a topological graph order parameter. *npj Comput. Mater.* **8**, 37 (2022).
23. Geiger, M. *et al.* *e3nn/e3nn: 2021-05-04* version 0.2.9. May 2021. <https://doi.org/10.5281/zenodo.4735637>.
24. Batzner, S. *et al.* E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
25. Vincent, P. A connection between score matching and denoising autoencoders. *Neural Comput.* **23**, 1661–1674 (2011).
26. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. *Deep unsupervised learning using nonequilibrium thermodynamics* in *International Conference on Machine Learning* (2015), 2256–2265.
27. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020).
28. Song, Y. *et al.* *Score-Based Generative Modeling through Stochastic Differential Equations* in *International Conference on Learning Representations* (2020).
29. Yang, L. *et al.* Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.* **56**, 1–39 (2023).
30. Stillinger, F. & Stillinger, D. Expanded solid matter: Two-dimensional LJ modeling. *Mech. Mater.* **38**, 958–968 (2006).
31. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).

32. Sadigh, B., Zepeda-Ruiz, L. & Belof, J. L. Metastable–solid phase diagrams derived from polymorphic solidification kinetics. *Proc. Natl. Acad. Sci.* **118**, e2017809118 (2021).
33. Stimac, J. C., Bertin, N., Mason, J. K. & Bulatov, V. V. Energy storage under high-rate compression of single crystal tantalum. *Acta Mater.* **239**, 118253 (2022).
34. Hyvärinen, A. & Dayan, P. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** (2005).
35. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. S. *Crystal Diffusion Variational Autoencoder for Periodic Material Generation* in *International Conference on Learning Representations* (2021).
36. Xu, M. *et al.* *GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation* in *International Conference on Learning Representations* (2021).
37. Lindemann, F. A. The calculation of molecular vibration frequencies. *Phys. Z.* **11**, 609–612 (1910).
38. Vincent, P. *et al.* Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11** (2010).
39. Zhou, F., Nielson, W., Xia, Y. & Ozoliņš, V. Lattice anharmonicity and thermal conductivity from compressive sensing of first-principles calculations. *Phys. Rev. Lett.* **113**, 185501 (2014).
40. Zhou, F., Nielson, W., Xia, Y. & Ozoliņš, V. Compressive sensing lattice dynamics. I. General formalism. *Phys. Rev. B* **100**, 184308 (2019).
41. Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. & Battaglia, P. *Learning Mesh-Based Simulation with Graph Networks* in *International Conference on Learning Representations* (2020).
42. Sanchez-Gonzalez, A. *et al.* *Learning to simulate complex physics with graph networks* in *International conference on machine learning* (2020), 8459–8468.
43. Loshchilov, I. & Hutter, F. *Decoupled Weight Decay Regularization* in *International Conference on Learning Representations* (2018).
44. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* **32**, 8026–8037 (2019).
45. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. Preprint at <https://arxiv.org/abs/1903.02428> (2019).
46. Larsen, A. H. *et al.* The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
47. De, S. & Smith, S. Batch normalization biases residual blocks towards the identity function in deep networks. *Advances in Neural Information Processing Systems* **33**, 19964–19975 (2020).
48. Weiler, M., Geiger, M., Welling, M., Boomsma, W. & Cohen, T. S. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems* **31** (2018).
49. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
50. Mishin, Y., Mehl, M., Papaconstantopoulos, D., Voter, A. & Kress, J. Structural stability and lattice defects in copper: Ab initio, tight-binding, and embedded-atom calculations. *Phys. Rev. B* **63**, 224106 (2001).

51. Li, Y., Siegel, D. J., Adams, J. B. & Liu, X.-Y. Embedded-atom-method tantalum potential developed by the force-matching method. *Phys. Rev. B* **67**, 125101 (2003).
52. Zepeda-Ruiz, L. A., Stukowski, A., Opperstrup, T. & Bulatov, V. V. Probing the limits of metal plasticity with molecular dynamics simulations. *Nature* **550**, 492 (2017).
53. Hirel, P. AtomsK: A tool for manipulating and converting atomic data files. *Comput. Phys. Commun.* **197**, 212–219 (2015).
54. Munetoh, S., Motooka, T., Moriguchi, K. & Shintani, A. Interatomic potential for Si–O systems using Tersoff parameterization. *Comput. Mater. Sci.* **39**, 334–339 (2007).
55. Mehl, M. J. *et al.* The AFLOW library of crystallographic prototypes: part 1. *Comput. Mater. Sci.* **136**, S1–S828 (2017).
56. Hicks, D. *et al.* The AFLOW library of crystallographic prototypes: part 2. *Comput. Mater. Sci.* **161**, S1–S1011 (2019).
57. Hicks, D. *et al.* The AFLOW library of crystallographic prototypes: part 3. *Comput. Mater. Sci.* **199**, 110450 (2021).