

Coarse-Graining and Forecasting Atomic Material Simulations with Descriptors

Thomas D. Swinburne^{*}

Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France



(Received 10 March 2023; revised 21 July 2023; accepted 13 November 2023; published 8 December 2023)

Atomic simulations of materials require significant resources to generate, store, and analyze. Here, descriptor functions are proposed as a general, metric latent space for atomic structures, ideal for use in large-scale simulations. Descriptors can regress a broad range of properties, including character-dependent dislocation densities, stress states, or radial distribution functions. A vector autoregressive model can generate trajectories over yield points, resample from new initial conditions and forecast trajectory futures. A forecast confidence, essential for practical application, is derived by propagating forecasts through the Mahalanobis outlier distance, providing a powerful tool to assess coarse-grained models. Application to nanoparticles and yielding of nanoscale dislocation networks confirms low uncertainty forecasts are accurate and resampling allows for the propagation of smooth property distributions. Yielding is associated with a collapse in the intrinsic dimension of the descriptor manifold, which is discussed in relation to the yield surface.

DOI: 10.1103/PhysRevLett.131.236101

Introduction.—Materials evolve via complex, nonintuitive atomic mechanisms spanning a wide range of time and length scales [1,2]. Atomic simulations (MD) with empirical force fields offer exceptional insight, but although spatial decomposition schemes give excellent (weak) parallel scaling with system size [3], serial time integration limits trajectory duration, irrespective of available processors [4]. The ubiquity yet high cost of MD means development of predictive techniques to coarse grain (CG) in space or time is an active research area [2,5–9]. Resolving material defects requires large system sizes, necessitating efficient and scalable CG techniques. While many structural analysis tools exist [10–15], none provide generic *compression* of atomic data with a clear metric for similarity or diversity, nor is it clear *a priori* how to select CG properties, leading to massive storage requirements at scale [16,17]. A further challenge is that simulations of materials are typically nonequilibrium and exhibit, in part due to timescale limitations, partially disordered structures with a dense kinetic spectrum and an unknown steady state, often with external driving [18–20]. To harness modern parallel computers there is thus a recognized need to *resample* sparse simulation data and to *forecast* simulation futures, both for physical insight and to maximize the information yield of additional computational effort [4,21–25].

However, the complexities of material deformation limit the applicability of current CG and acceleration schemes, which require identification of a clear timescale separation [26] to allow parallel time accumulation [2,4,9,21,26–29] or the design of low rank (typically 1–4) collective variables (CV) which can be used to bias dynamics [5,8,30–34]. Despite many recent advances [8,35,36]

general CVs for extended defects remain elusive [34,36], instead requiring specialized simulation setups with only a few active mechanisms such as nucleation [8] or the migration of isolated defects [34,35]. Exploring *unseen* regions of configuration space is known to be uncontrolled as low rank CVs may not remain descriptive [37]. These issues extend to the powerful postmortem analysis tools [6,38–42], which learn collective variables that obey a discrete state Markov model in order to identify kinetically important configurations with implied transition timescales. While all-atom [43–46] or coarse-beaded [47] generative models may provide a route for accelerated time stepping, they are currently only competitive to direct time integration for fairly small equilibrium systems with a static or slowly varying bonding topology, and so cannot be applied to large-scale simulations of material deformation where a highly transient, heterogeneous atomic connectivity is fundamental.

In this Letter, atomic descriptor functions [48–52] are proposed as an efficient, general, and uncertainty-aware coarse-graining approach, mapping atomic positions $\mathbf{X} \in \mathbb{R}^{N \times 3}$ to a global vector $\mathbf{D} \in \mathbb{R}^{\sim 100}$, Eq. (1). The main results are that (i) descriptors can classify and regress a remarkable range of structural properties (see figures) and permit a data-driven model extrapolation measure [53], transferring advances in active learning [54–56] to atomic CG. This generality means CG targets need not be specified *a priori*, giving huge compression in storage and efficiencies in analysis at scale. (ii) Descriptor trajectories can be efficiently resampled and forecasted via a vector autoregressive (VAR) model [57], with, crucially, a robust forecast *uncertainty* derived from the descriptor outlier measure (5). This allows rapid assessment of when

forecasts can be trusted or when additional training is needed, essential for practical usage but typically missing in existing schemes. The approach is applied to analyze and forecast systems essentially untreatable with existing methods, the annealing of large nanoparticles, and yielding of nanoscale dislocation networks under cyclic shear and uniaxial tension [17]. The intrinsic dimension [58] of the descriptor manifold is shown to collapse on yielding, which is discussed in relation to the yield surface.

Descriptor coarse graining.—Descriptors [56,59–61] map atomic coordinates $\mathbf{X} \in \mathbb{R}^{N \times 3}$ to $\mathbf{D}(\mathbf{X}) \in \mathbb{R}^{N \times D}$, where each element $[\mathbf{D}(\mathbf{X})]_{ij}$ takes the local atomic environment of an atom i as input and returns a permutation-invariant scalar, vector, or tensor depending on the regression target (e.g., energies, forces) [62,63]. Descriptors which approximate a many-body atomic basis [48–52] have found use in linear estimators $\hat{\mathcal{O}} \simeq \Theta_{\mathcal{O}} \cdot \mathbf{D} + \Theta_{\mathcal{O}}^0$ of some target observable $\mathcal{O}(\mathbf{X})$, where $\Theta_{\mathcal{O}} \in \mathbb{R}^D$ and $\Theta_{\mathcal{O}}^0 \in \mathbb{R}$ are parameters. For $\mathcal{O} = E$, the atomic potential energy, these can reach state of the art accuracy [60,63,64], often with lower computational cost and simpler fitting [48–52]. The first main result of this Letter is that linear estimators can capture essentially any structural property which could be of relevance to a coarse-grained model. The widely used [3] bSO(4) descriptors [3,48,50,64] are used, detailed in the Supplemental Material (SM) [65], summing over all atoms to give the global descriptors

$$\bar{\mathbf{D}} = \sum_i \mathbf{D}_i \in \mathbb{R}^D, \quad \bar{\mathbf{V}} = \sum_i \mathbf{X}_i \otimes \nabla_{\mathbf{X}} \mathbf{D}_i \in \mathbb{R}^{D \times 3 \times 3}, \quad (1)$$

where \otimes is the outer (dyadic) product [68]. Figure 1 shows linear estimators

$$\hat{\mathcal{O}}(\bar{\mathbf{D}}) = \Theta_{\mathcal{O}} \cdot \bar{\mathbf{D}} + \Theta_{\mathcal{O}}^0, \quad (2)$$

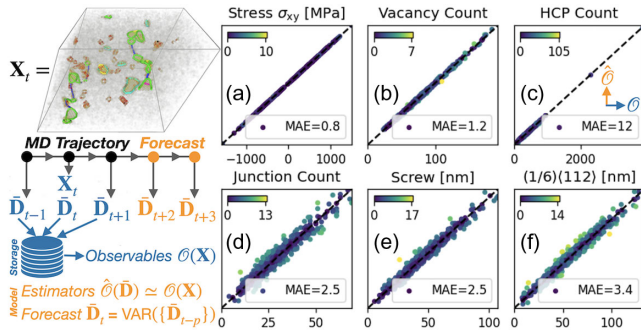


FIG. 1. Coarse graining of dislocation networks in Al under cyclic shear, detailed in the SM [65]. Left: global descriptor vectors $\bar{\mathbf{D}}$ (1) are stored every 1–10 ps and positions \mathbf{X} every 100–500 ps. $\{\mathbf{X}, \bar{\mathbf{D}}\}$ data is used to train estimators $\hat{\mathcal{O}}(\bar{\mathbf{D}})$ of observables $\mathcal{O}(\mathbf{X})$ and a VAR forecaster (3). (a)–(f) \mathcal{O} vs $\hat{\mathcal{O}}$ from over 20 targets, including (d) dislocation junctions and total length of (e) screw or (f) $\langle 112 \rangle / 6$ dislocations. Nonscalar (a) σ_{xy} , estimated with $\bar{\mathbf{D}} \oplus \bar{\mathbf{V}}$ [65]. See also $g(r)$ in Fig. 2.

applied to dislocation networks in aluminum [65], accurately capturing a broad range of properties including dislocation junction densities, character-dependent line densities, and crystal structure content. Similar results were found for the nanoparticle ensemble and a range of dislocated solids in fcc and bcc materials. Dislocation properties were extracted with OVITO-DXA [15] which has some intrinsic noise due to the discretization parameters. It is also possible to capture the radial distribution function (RDF) $g(r)$ by estimating coefficients $\hat{a}_l(\bar{\mathbf{D}})$ of a basis expansion $g(r) \equiv \sum_l a_l u_l(r)$, as shown in Fig. 2. As found in previous work targeting vibrational entropies [69,70], all predictions were stable under widely varying test and train ratios and truncation of training data range. Matrix-valued observables such as the stress $\mathcal{O}(\mathbf{X}) = \sigma \in \mathbb{R}^{3 \times 3}$ can be estimated by building equivariant estimators with $\bar{\mathbf{V}}$; the simplest ($l = 0$ [63]) example is simply $\hat{\mathcal{O}}(\bar{\mathbf{D}}) = \Theta_{\mathcal{O}} \cdot \bar{\mathbf{V}} \in \mathbb{R}^{3 \times 3}$. Examples for the nonscalar shear stress σ_{xy} are shown in Fig. 1 and the SM [65]. However, in the following only $\bar{\mathbf{D}}$ is used for forecasting, targeting the scalar pressure $\text{Tr}(\sigma)$, as model parameters are scalars and $\bar{\mathbf{D}}$ has a metric distance [48].

While (2) is trained on the global descriptor signal (1), spatial dependence will be required as the simulation volume increases, achieved by averaging over atoms in some voxel discretization. Future work will investigate this voxelized signal and the constraints required to preserve dislocation topology in forecasts.

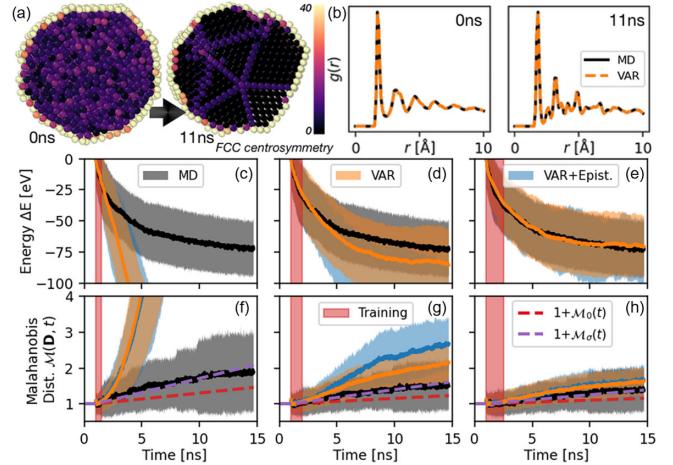


FIG. 2. Annealing of Pt nanoparticles. (a) Representative structure at 0 ns and 11 ns. (b) The average RDF $g(r)$ and the corresponding descriptor estimation [65] $g(r; \bar{\mathbf{D}})$. (c)–(h) Ensemble data with $M = 60$, $\tau_{\text{tr}} = 0.5, 1.0, 1.5$ ns, training starting at $t = 1$ ns (left to right, red shade). Mean is the solid line, with standard deviation as bands. Black: MD data. Orange: VAR forecasts from 1 ns. Blue: VAR forecast with epistemic errors. (c)–(e) Potential energy change from 1 ns mean. (f)–(h) Mahalanobis distances, MD: $\mathcal{M}(\mathbf{D}_t)$, Eq. (4); forecasts: $\mathcal{M}(t)$, Eq. (5). The theoretical lower bounds \mathcal{M}_0 (red dash) and \mathcal{M}_σ (purple dash) are also shown.

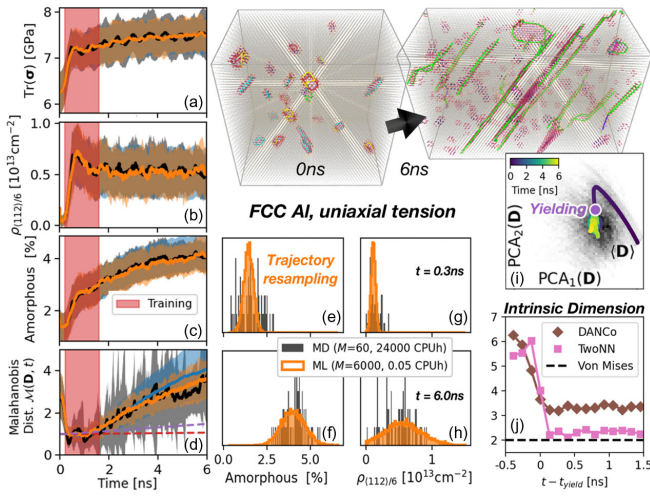


FIG. 3. Yielding of Al under uniaxial tension. Colors follow Fig. 2. Forecasts are from 0.2 ns, with $\tau_{\text{tr}} = 1.2$ ns and $P = 5$. (a) Pressure, (b) $\langle 112 \rangle / 6$ dislocation density, and (c) amorphous content. (d) Mahalanobis distance. (e)–(h) Trajectory resampling from 0.3 ns (preyield), with $100\times$ larger ensemble. (i) PCA analysis of the ensemble mean ($\bar{\mathbf{D}}$), clearly showing a localization on yield. Individual trajectories shown as histogram in grayscale. (j) ID of the descriptor manifold, estimated via TwoNN [71] and DANCo [72]. Both show a collapse on yielding but remain above the Von Mises lower bound.

The accuracy and scope of (2) has particular relevance for massively parallel workflows, as only $\bar{\mathbf{D}}, \bar{\mathbf{V}}$ need to be stored to later extract almost any global observable of interest *a posteriori* after training on a small database of stored positions, offering massive data compression.

Unimodality and generation of descriptor data.—As the descriptors have a metric distance, similar atomic structures will be close in descriptor space. In addition, their distribution in sufficiently high dimensions can be expected to be unimodal, routinely invoked in active learning schemes [54,55] and more recently in the analysis of defect structures [56]. Evidence for nanoparticle and dislocation ensembles is provided in the SM [65]. It is then simple to *generate* plausible descriptor vectors by fitting and sampling a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to the descriptor dataset. An example of this is shown below in Fig. 3, where the observed descriptor initial conditions are densely interpolated, allowing the evolution of observable distributions to be monitored.

Resampling and forecasting of descriptor trajectories.—At regular intervals $t_n = n\delta\tau$, $\delta\tau \simeq 10$ ps, a “snapshot” is taken by time averaging $\bar{\mathbf{X}}_n = \tau_D^{-1} \int_0^{\bar{\tau}} \mathbf{X}(t_n + t) dt$ over a period $\bar{\tau} \simeq 20\text{--}50$ fs to reduce noise from thermal fluctuations [73], then calculating descriptor vectors $\bar{\mathbf{D}}_n = \bar{\mathbf{D}}(\bar{\mathbf{X}}_n)$. A small database of positions $\bar{\mathbf{X}}_n$ is built by recording 1%–5% of snapshots, though positions could be selected adaptively to maximize training diversity. An ensemble of M simulations thus produces M discrete time trajectories $\{\bar{\mathbf{D}}_n\}$,

which are used to train a P -state vector autoregressive VAR(P) model [57,74]

$$\bar{\mathbf{D}}_{n+1} = \sum_{p=0}^{P-1} \mathbf{T}_p \bar{\mathbf{D}}_{n-p} + \mathbf{c} + \mathbf{w}_n, \quad \langle \mathbf{w}_n^T \mathbf{w}_m \rangle = \mathbf{S} \delta_{nm}. \quad (3)$$

For $P > 1$ a Wold transformation [75] $\mathbf{Z}_n = \mathbf{1} \oplus \bar{\mathbf{D}}_n \dots \oplus \bar{\mathbf{D}}_{n-p} \in \mathbb{R}^{1+P\bar{D}}$ casts (3) as a Markovian Ornstein-Uhlenbeck equation [76] $\mathbf{Z}_{n+1} = \mathbf{T}\mathbf{Z}_n + \mathbf{W}_n$. The maximum likelihood estimator of \mathbf{T} is simply the least squares solution, with \mathbf{S} determined from the residual covariance [74]. To minimize generalization error, a bagging [77–79] approach was developed, applying Bayesian ridge regression [80] to random overlapping subsets. Results were stable under 10–40 subsets each with 10%–40% coverage, giving epistemic uncertainties $\delta\mathbf{T}, \delta\mathbf{S}$ from the covariance across subsets. Training is robust and requires only a few CPU minutes, a key advantage over (RNN and LSTM) neural networks [81,82] or neural differential equations [83] which require significant resources, regularization, and correction schemes [47], and limited in practice to data dimension $\bar{D} < 10$ [83,84]. A Chapman-Komologorov test [6] for the transfer matrix \mathbf{T} is provided in the SM [65], but in practice the light computational demand also permits a convergence test of model architecture by increasing P [65].

Deriving a forecast uncertainty.—Practical application of (3) requires a robust measure of forecast uncertainty [4,21–25], which should be larger for configurations further from the training data independent of epistemic errors. This is particularly relevant to the nonstationary dynamics of material deformation. As uncertainty to previously unseen macroscopic changes is clearly not quantifiable [21], the following bound is conditional on the simulation *ensemble* remaining unimodal and not undergoing macroscopic changes. Many extrapolation grade estimators have been developed for active learning of energy models [54,56,58,64,85]; here, the Mahalanobis outlier distance [53] is used for the unimodal descriptor distribution [56,65]. With training data mean $\boldsymbol{\mu}_{\text{tr}}$ and covariance $\boldsymbol{\Sigma}_{\text{tr}}$ estimated via a shrinkage estimator [86], the squared Mahalanobis distance reads

$$\mathcal{M}(\bar{\mathbf{D}}) = [\bar{\mathbf{D}} - \boldsymbol{\mu}_{\text{tr}}] \boldsymbol{\Sigma}_{\text{tr}}^{-1} [\bar{\mathbf{D}} - \boldsymbol{\mu}_{\text{tr}}] / \bar{D}. \quad (4)$$

Importantly, (4) is independent of the VAR(P) forecast model (3); points drawn from a low density region of ρ_{tr} will have a large Mahalanobis distance, even if epistemic uncertainties $\delta\mathbf{T}$ are small. At long forecasting times, (3) will reach its high dimensional steady state [65], with $\langle \mathcal{M} \rangle$ constant. However, model parameters cannot be assumed static, with a time dependence bounded from below by $1/\tau_M = 1/(M\tau_{\text{tr}})$ [87], where M is the ensemble size and τ_{tr} training duration. This drift can be estimated by propagating epistemic uncertainty in the steady state to

an uncertainty $\sigma_{\mathcal{M}}^2$ in $\mathcal{M}(\bar{\mathbf{D}})$, which should be accumulated [65], leading to an additional linear growth in (4) of

$$\mathcal{M}(t_n) = \langle \mathcal{M}(\bar{\mathbf{D}}_n) \rangle + \mathcal{M}_\sigma(t_n), \quad \mathcal{M}_\sigma(t_n) \geq \mathcal{M}_0(t_n), \quad (5)$$

where $\mathcal{M}_\sigma(t) = \sigma_{\mathcal{M}}^2 t / \delta\tau$ and $\mathcal{M}_0(t) = t / \tau_M$. Equation (5) is the main theoretical result of this Letter, an uncertainty metric for forecasting via (3). An approximate parallel efficiency is implied by $\eta = \tau_{\text{pred}} / (M\tau_{\text{tr}})$, where $\mathcal{M}(\tau_{\text{pred}}) \equiv \mathcal{M}_0(M\tau_{\text{tr}}) = 2$, giving $\eta = 1$ when $\mathcal{M}(t_n) = \mathcal{M}_0(t)$.

Annealing of Pt Nanoparticles.—Metallic nanoparticles are important functional materials for catalysis; 50–150 atom clusters have been extensively studied in simulations [1,39,40,88], but for large sizes and high temperatures the landscape of energy minima is vast and insufficiently metastable for current acceleration methods [40]. The current application to $M = 60$ 4000-atom EAM-Pt [89] nanoparticles at 900 K is thus out of scope for existing methods.

The initial structure was formed by quenching from the liquid state and annealing for 100 ps to give a highly disordered but predominantly fcc structure ($c_{\text{FCC}} \simeq 0.5$). Descriptor trajectories were extracted every 1.5 ps, with a full structural analysis undertaken every 100 ps, though the dataset was sparsified by taking $\delta\tau = 15$ ps and removing intermediate snapshots. Autoregressive models (3) were constructed with $P = 1$ –3 and $\tau_{\text{tr}} = 0.5, 1.0, \text{ or } 1.5$ ns, with $P = 1$ shown. Generated trajectories had initial conditions from the start of the training data, to both resample then forecast observed trajectories. Figure 2 displays the ensemble simulation data, model predictions, and epistemic errors for the formation energy, the RDF $g(r)$, and the Mahalanobis uncertainty (5). The RDF reflects the significant growth in FCC crystal structure, as can also be directly extracted through estimation of c_{FCC} . MD data used $\mathcal{M}(\mathbf{D})$, Eq. (4), which closely follows the theoretical lower bound $\mathcal{M}_\sigma(t)$. While forecasts improve with training data, crucially, the magnitude of $\mathcal{M}(\bar{\mathbf{D}}, t)$ can independently confirm their reliability.

Yielding of Al under uniaxial tension.—Dislocations carry plastic deformation, forming dense networks under irradiation [20] or extended loading [90]; understanding network evolution remains a grand challenge of physics and engineering [17,91,92]. Atomic simulations continue to discover mechanisms even in model systems [92], in part due to the difficulty in analyzing atomic data [36]. An ensemble of $M = 60$ dense dislocation networks were formed in an embedded atom method (EAM) model of Al [93] by creating simulation boxes of around 1.5×10^5 atoms, orientated to $[10\bar{1}]$, $[111]$, $[1\bar{2}1]$, with populations of interstitial loops with density $\rho_{\text{dis}} \in [10^{11}, 10^{13}] \text{ cm}^{-2}$. Uniaxial tension was applied at a rate $\dot{\epsilon}_{xx} = 1 \times 10^8 \text{ s}^{-1}$ along $[10\bar{1}]$, allowing other supercell dimensions to relax [17]. The SM [65] shows application to cyclic shear loading. While typical in MD [17,92], the small system sizes and large strain rates

suppress correlations from long-range elastic interactions and the role of e.g., thermally activated mechanisms which will clearly influence network evolution and thus yield behavior. Future work will use spatially dependent descriptor signals from voxelization (discussed above) to find trends in how dislocation network evolution depends on system size, dislocation density, and loading conditions, required to connect such atomic simulations to the deformation of real microstructures. The results are summarized in Figs. 3(a)–3(d), using the linear estimators (2). Increasing τ_{tr} decreased error and uncertainty, with optimal results found using $P = 5$ [65]. Training only on preyield structures led to unstable forecasts as yield is characterized by a qualitative change in the descriptor manifold as detailed below. $\mathcal{M}(\bar{\mathbf{D}}, t)$ also diverged at the yield point, clearly indicating that more training data is required. This again demonstrates the utility and critical importance of a forecast uncertainty to assess data-driven predictions. However, accurate forecasting of structural transitions only from pretransition data remains an important topic for future research. Resampling allows for ensembles to be increased by orders of magnitude for negligible CPU effort, giving the smooth distributions shown in Figs. 3(e)–3(h). Initial descriptor states were generated as described above from $\rho_{\text{tr}} \simeq \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, fit from the descriptor ensemble at times 0.3–0.31 ns.

The forecasted ensemble captures multiple important structural evolutions that, while known for this well-studied system [17], confirm the accuracy of the VAR approach. Forecasts correctly predict the growth of amorphous atomic environments due to defect production under continued loading [94], the expected sharp peak in hexagonal close packed (HCP) content at yield, accompanied by a growth peak, then steady state in the number of dislocation junctions (see SM [65]). Distributions can tighten or widen, here indicating the evolution in dislocation character—initial populations of $\langle 100 \rangle / 3$ Hirth dislocation loops decay to a tight distribution close to zero upon loading, accompanied by an emergence of a broad, stable distribution of $\langle 112 \rangle / 6$ dislocation lines that carry the plastic flow [17]. The joint stability of junctions, dislocation density, and stress is consistent with a Kocks-Mecking steady state [95]. The global descriptor signal $\bar{\mathbf{D}}$ can classify yielding with minimal training data, see [65]; Fig. 3(i) shows how the ensemble average $\langle \bar{\mathbf{D}} \rangle$ clearly localizes postyield.

Models for yielding invoke the concept of a yield surface in 6D stress space [96], which for metallic systems is typically the Von Mises yield surface, isosurfaces of the J_2 invariant with intrinsic dimension (ID) [58,72] of 2. Yielding is thus expected to be accompanied by a drop in the ID of the stress trajectory; furthermore, as descriptors can perfectly predict stress, their ID is an upper bound to the stress ID. Two empirical estimators [71,72,97], which typically underestimate [72], were applied to the full $\bar{\mathbf{D}} \oplus \bar{\mathbf{V}}$ dataset. Figure 3(j) shows both ID estimates collapse from 5–7 to 2–3 on yield, consistent with the

Von Mises lower bound of 2. While larger-scale studies are essential, this suggests the existence of a generalized yield manifold, which could allow for data-driven construction of much richer structure-property relationships.

Conclusions.—This Letter has promoted descriptors as a general, uncertainty aware coarse-grained representation of atomic structures ideal for analysis, resampling, and forecasting. The descriptor manifold holds promise for future research on structural transitions such as yielding, alongside the use of forecasting in resource allocation [4,21–25] and extension to a spatially dependent, fully equivariant descriptor signal to capture long-range correlations.

I thank M.-C. Marinica and L. Truskinovsky for stimulating discussion and anonymous referees for their careful reading of the manuscript. Support from ANR Grant No. ANR-19-CE46-0006-1, IDRIS allocations A0090910965, A0120913455, Euratom Grant No. 633053, and the IPAM program *New Mathematics for the Exascale* are gratefully acknowledged.

*thomas.swinburne@cnsr.fr

- [1] D. J. Wales, *Energy Landscapes*, 1st edition (Cambridge University Press, Cambridge, England, 2003).
- [2] D. Perez, B. P. Uberuaga, Y. Shim, J. G. Amar, and A. F. Voter, *Annu. Rep. Comput. Chem.* **5**, 79 (2009).
- [3] S. Plimpton, *J. Comput. Phys.* **117**, 1 (1995).
- [4] D. Perez, E. D. Cubuk, A. Waterland, E. Kaxiras, and A. F. Voter, *J. Chem. Theory Comput.* **12**, 18 (2015).
- [5] A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562 (2002).
- [6] A. Mardt, L. Pasquali, H. Wu, and F. Noé, *Nat. Commun.* **9**, 1 (2018).
- [7] E. Van Der Giessen, P. A. Schultz, N. Bertin, V. V. Bulatov, W. Cai, G. Csányi, S. M. Foiles, M. G. Geers, C. González, M. Hütter *et al.*, *Model. Simul. Mater. Sci. Eng.* **28**, 043001 (2020).
- [8] L. Bonati, G. Piccini, and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2113533118 (2021).
- [9] T. D. Swinburne and D. Perez, *Model. Simul. Mater. Sci. Eng.* **30**, 034004 (2022).
- [10] J. D. Honeycutt and H. C. Andersen, *J. Phys. Chem.* **91**, 4950 (1987).
- [11] C. L. Kelchner, S. J. Plimpton, and J. C. Hamilton, *Phys. Rev. B* **58**, 11085 (1998).
- [12] G. J. Ackland and A. P. Jones, *Phys. Rev. B* **73**, 054104 (2006).
- [13] E. A. Lazar, J. K. Mason, R. D. MacPherson, and D. J. Srolovitz, *Phys. Rev. Lett.* **109**, 095505 (2012).
- [14] E. A. Lazar, J. Han, and D. J. Srolovitz, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5769 (2015).
- [15] A. Stukowski, V. V. Bulatov, and A. Arsenlis, *Model. Simul. Mater. Sci. Eng.* **20**, 085007 (2012).
- [16] G. Wu, H. Song, and D. Lin, *Comput. Mater. Sci.* **144**, 322 (2018).
- [17] L. A. Zepeda-Ruiz, A. Stukowski, T. Opperstrup, and V. V. Bulatov, *Nature (London)* **550**, 492 (2017).
- [18] W. Setyawan, G. Nandipati, K. J. Roche, H. L. Heinisch, B. D. Wirth, and R. J. Kurtz, *J. Nucl. Mater.* **462**, 329 (2015).
- [19] N. V. Priezjev, *J. Non-Cryst. Solids* **479**, 42 (2018).
- [20] D. R. Mason, S. Das, P. M. Derlet, S. L. Dudarev, A. J. London, H. Yu, N. W. Phillips, D. Yang, K. Mizohata, R. Xu *et al.*, *Phys. Rev. Lett.* **125**, 225503 (2020).
- [21] T. Swinburne and D. Perez, *npj Comput. Mater.* **6**, 190 (2020).
- [22] A. Garmon, V. Ramakrishnaiah, and D. Perez, *Parallel Comput.* **112**, 102936 (2022).
- [23] J. Schaarschmidt, J. Yuan, T. Strunk, I. Kondov, S. P. Huber, G. Pizzi, L. Kahle, F. T. Bölle, I. E. Castelli, T. Vegge *et al.*, *Adv. Energy Mater.* **12**, 2102638 (2022).
- [24] L.-F. Zhu, J. Janssen, S. Ishibashi, F. Körmann, B. Grabowski, and J. Neugebauer, *Comput. Mater. Sci.* **187**, 110065 (2021).
- [25] J. Andrews, O. Gkountouna, and E. Blaisten-Barojas, *Chem. Sci.* **13**, 7021 (2022).
- [26] C. Le Bris, T. Lelièvre, M. Luskin, and D. Perez, *Monte Carlo Methods Appl.* **18**, 119 (2012).
- [27] A. F. Voter, *Phys. Rev. B* **57**, R13985 (1998).
- [28] G. Henkelman and H. Jónsson, *J. Chem. Phys.* **111**, 7010 (1999).
- [29] M. So and A. Voter, *J. Chem. Phys.* **112**, 9599 (2000).
- [30] A. F. Voter, *Phys. Rev. Lett.* **78**, 3908 (1997).
- [31] G. Henkelman, B. P. Uberuaga, and H. Jónsson, *J. Chem. Phys.* **113**, 9901 (2000).
- [32] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, *J. Chem. Phys.* **128**, 144120 (2008).
- [33] T. Lelièvre, G. Stoltz, and M. Rousset, *Free Energy Computations: A Mathematical Perspective* (World Scientific, Singapore, 2010).
- [34] T. D. Swinburne and M.-C. Marinica, *Phys. Rev. Lett.* **120**, 135503 (2018).
- [35] J. Rogal, *Eur. Phys. J. B* **94**, 1 (2021).
- [36] J. Baima, A. S. Goryaeva, T. Swinburne, J.-B. Maillet, M. Nastar, and M. C. Marinica, *Phys. Chem. Chem. Phys.* **24**, 23152 (2022).
- [37] G. Bussi and A. Laio, *Nat. Rev. Phys.* **2**, 200 (2020).
- [38] F. Noé and F. Nuske, *Multiscale Model. Simul.* **11**, 635 (2013).
- [39] R. Huang, L.-T. Lo, Y. Wen, A. Voter, and D. Perez, *J. Chem. Phys.* **147**, 152717 (2017).
- [40] R. Huang, Y. Wen, A. F. Voter, and D. Perez, *Phys. Rev. Mater.* **2**, 126002 (2018).
- [41] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, and J. C. Grossman, *Nat. Commun.* **10**, 1 (2019).
- [42] S. Soltani, C. W. Sinclair, and J. Rottler, *Phys. Rev. E* **106**, 025308 (2022).
- [43] W. Wang and R. Gómez-Bombarelli, *npj Comput. Mater.* **5**, 125 (2019).
- [44] L. Klein, A. Y. Foong, T. E. Fjelde, B. Mlodozieniec, M. Brockschmidt, S. Nowozin, F. Noé, and R. Tomioka, arXiv:2302.01170.
- [45] C. Han, P. Zhang, D. Bluestein, G. Cong, and Y. Deng, *J. Comput. Phys.* **427**, 110053 (2021).
- [46] B. Leimkuhler, D. T. Margul, and M. E. Tuckerman, *Mol. Phys.* **111**, 3579 (2013).
- [47] X. Fu, T. Xie, N. J. Rebello, B. D. Olsen, and T. Jaakkola, arXiv:2204.10348.

- [48] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, *J. Comput. Phys.* **285**, 316 (2015).
- [49] A. Shapeev, *Multiscale Model. Simul.* **14**, 1153 (2016).
- [50] A. M. Goryaeva, J. D  r  s, C. Lapointe, P. Grigorev, T. D. Swinburne, J. R. Kermode, L. Ventelon, J. Baima, and M.-C. Marinica, *Phys. Rev. Mater.* **5**, 103803 (2021).
- [51] A. E. Allen, G. Dusson, C. Ortner, and G. Cs  nyi, *Mach. Learn. Sci. Technol.* **2**, 025017 (2021).
- [52] Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Cs  nyi, C. Ortner *et al.*, *npj Comput. Mater.* **7**, 1 (2021).
- [53] P. C. Mahalanobis (National Institute of Science of India, 1936).
- [54] E. V. Podryabinkin and A. V. Shapeev, *Comput. Mater. Sci.* **140**, 171 (2017).
- [55] N. Bernstein, G. Cs  nyi, and V. L. Deringer, *npj Comput. Mater.* **5**, 1 (2019).
- [56] A. M. Goryaeva, C. Lapointe, C. Dai, J. D  r  s, J.-B. Maillet, and M.-C. Marinica, *Nat. Commun.* **11**, 4691 (2020).
- [57] H. L  tkepohl, *New Introduction to Multiple Time Series Analysis* (Springer Science & Business Media, New York, 2005).
- [58] A. Glielmo, I. Macocco, D. Doimo, M. Carli, C. Zeni, R. Wild, M. d'Errico, A. Rodriguez, and A. Laio, *Patterns* **3**, 100589 (2022).
- [59] V. L. Deringer, M. A. Caro, and G. Cs  nyi, *Adv. Mater.* **31**, 1902765 (2019).
- [60] A. P. Bart  k, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Cs  nyi, and M. Ceriotti, *Sci. Adv.* **3**, e1701816 (2017).
- [61] I. Batatia, D. P. Kov  cs, G. N. Simm, C. Ortner, and G. Cs  nyi, *arXiv:2206.07697*.
- [62] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Sch  tt, and K.-R. M  ller, *Sci. Adv.* **3**, e1603015 (2017).
- [63] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, *Nat. Commun.* **13**, 1 (2022).
- [64] A. P. Bart  k, M. C. Payne, R. Kondor, and G. Cs  nyi, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [65] See Supplemental Material, which includes Refs. [66,67], at <http://link.aps.org/supplemental/10.1103/PhysRevLett.131.236101> for convergence tests and extended figures.
- [66] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *J. Nonlinear Sci.* **25**, 1307 (2015).
- [67] C. E. Rasmussen, *Gaussian Processes in Machine Learning* (Springer, Berlin, Heidelberg, 2004).
- [68] In practice, the tensor of symmetric matrices $\bar{\mathbf{V}}_{\xi} \equiv \bar{\mathbf{V}} + \bar{\mathbf{V}}^T$ is calculated in current implementations [3,50].
- [69] C. Lapointe, T. D. Swinburne, L. Thiry, S. Mallat, L. Proville, C. S. Becquart, and M.-C. Marinica, *Phys. Rev. Mater.* **4**, 063802 (2020).
- [70] C. Lapointe, T. D. Swinburne, L. Proville, C. S. Becquart, N. Mousseau, and M.-C. Marinica, *Phys. Rev. Mater.* **6**, 113803 (2022).
- [71] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, *Sci. Rep.* **7**, 12140 (2017).
- [72] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, *Pattern Recognit.* **47**, 2569 (2014).
- [73] T. D. Swinburne, S. L. Dudarev, and A. P. Sutton, *Phys. Rev. Lett.* **113**, 215501 (2014).
- [74] S. Karlsson, *Handbook of Economic Forecasting* (Elsevier, France, 2013), Vol. 2, p. 791.
- [75] H. O. Wold, *Ann. Math. Stat.* **19**, 558 (1948).
- [76] W. Coffey and Y. P. Kalmykov, *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry and Electrical Engineering* (World Scientific, Singapore, 2012), Vol. 27.
- [77] L. Breiman, *Mach. Learn.* **24**, 123 (1996).
- [78] B. Lakshminarayanan, A. Pritzel, and C. Blundell, *Adv. Neural Inf. Process. Syst.* **30**, 6402 (2017).
- [79] F. Petropoulos, R. J. Hyndman, and C. Bergmeir, *Eur. J. Oper. Res.* **268**, 545 (2018).
- [80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [81] Y. Yu, X. Si, C. Hu, and J. Zhang, *Neural Comput.* **31**, 1235 (2019).
- [82] P. R. Vlachas, J. Zavadlav, M. Praprotnik, and P. Koumoutsakos, *J. Chem. Theory Comput.* **18**, 538 (2021).
- [83] P. Kidger, J. Morrill, J. Foster, and T. Lyons, *Adv. Neural Inf. Process. Syst.* **33**, 6696 (2020).
- [84] F. Liu, M. Cai, L. Wang, and Y. Lu, *IEEE Access* **7**, 26102 (2019).
- [85] C. Zeni, A. Anelli, A. Glielmo, and K. Rossi, *Phys. Rev. B* **105**, 165141 (2022).
- [86] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, *IEEE Trans. Signal Process.* **58**, 5016 (2010).
- [87] T. D. Swinburne and D. Perez, *Phys. Rev. Mater.* **2**, 053802 (2018).
- [88] F. Baletto, R. Ferrando, A. Fortunelli, F. Montalenti, and C. Mottet, *J. Chem. Phys.* **116**, 3856 (2002).
- [89] C. Liu, J. Cohen, J. Adams, and A. Voter, *Surf. Sci.* **253**, 334 (1991).
- [90] S. Lavenstein, Y. Gu, D. Madiseti, and J. A. El-Awady, *Science* **370**, eabb2690 (2020).
- [91] R. Baggio, E. Arbib, P. Biscari, S. Conti, L. Truskinovsky, G. Zanzotto, and O. U. Salman, *Phys. Rev. Lett.* **123**, 205501 (2019).
- [92] N. Bertin, W. Cai, S. Aubry, A. Arsenlis, and V. V. Bulatov, *arXiv:2210.14343*.
- [93] B. Onat and S. Durukano  lu, *J. Phys. Condens. Matter* **26**, 035404 (2013).
- [94] J. Marian, W. Cai, and V. V. Bulatov, *Nat. Mater.* **3**, 158 (2004).
- [95] R. B. Sills, N. Bertin, A. Aghaei, and W. Cai, *Phys. Rev. Lett.* **121**, 085501 (2018).
- [96] J. P. Hirth and J. Lothe, *Theory of Dislocations* (Malabar, FL Krieger, 1991).
- [97] J. Bac, E. M. Mirkes, A. N. Gorban, I. Tyukin, and A. Zinovyev, *Entropy* **23**, 1368 (2021).