Coarse-Graining and Forecasting Atomic Material Simulations with Descriptors

Thomas Swinburne

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
(Dated: November 14, 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.

Materials evolve via complex, non-intuitive 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. Whilst 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 non-equilibrium 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 post-mortem 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. Whilst 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 $\bar{\mathbf{D}} \in \mathbb{R}^{\sim 100}$, Eq. 1. The main results are that 1) 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. 2) 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, 60–62] 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 environ-

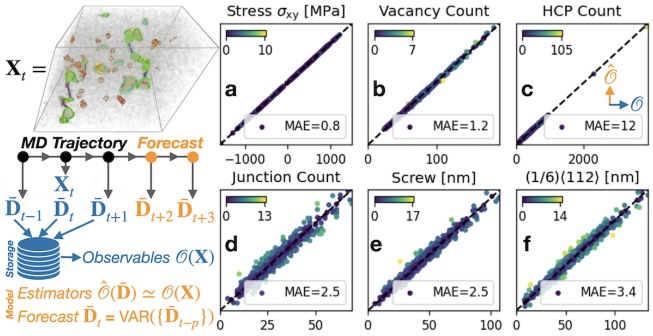

FIG. 1. Coarse graining of dislocation networks in Al under cyclic shear, detailed in the SM[59]. Left: Global descriptor vectors $\bar{\mathbf{D}}$ (1) are stored every 1-10ps and positions $\mathbf{X}$ every 100-500ps. $\{\mathbf{X}, \bar{\mathbf{D}}\}$ data is used to train estimators $\hat{\mathcal{O}}(\bar{\mathbf{D}})$ 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. Non-scalar a) σ_{xy}, estimated with $\bar{\mathbf{D}} \oplus \bar{\mathbf{V}}$[59]. See also $g(r)$ in Fig. 2.

ment of an atom i as input and returns a permutation-invariant scalar, vector or tensor depending on the regression target (e.g. energies, forces)[63, 64]). Descriptors which approximate a many-body atomic basis[48–52] have found use in *linear* estimators $\hat{\mathcal{O}} \simeq \mathbf{\Theta}_{\mathcal{O}} \cdot \mathbf{D} + \Theta^0_{\mathcal{O}}$ of some target observable $\mathcal{O}(\mathbf{X})$, where $\mathbf{\Theta}_{\mathcal{O}} \in \mathbb{R}^D$ and $\Theta^0_{\mathcal{O}} \in \mathbb{R}$ are parameters. For $\mathcal{O} = E$, the atomic potential energy, these can reach state of the art accuracy[61, 64, 65], 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, 65] are used, detailed in the supplementary material (SM)[59], summing over all atoms to give the global descriptors

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

where $\otimes$ is the outer (dyadic) product[66]. Figure (1) shows linear estimators

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

applied to dislocation networks in aluminum[59], 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 $\hat{a}_l(\mathbf{D})$ 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[67, 68], all predictions were stable under widely varying test/train ratios and truncation of training data range. Matrix-valued observables such as the stress $\mathcal{O}(\mathbf{X}) = \boldsymbol{\sigma} \in \mathbb{R}^{3\times 3}$ can be estimated by building equivariant estimators with $\bar{\mathbf{V}}$; the simplest ($l = 0$[64]) example is simply $\hat{\mathcal{O}}(\bar{\mathbf{D}}) = \mathbf{\Theta}_{\mathcal{O}} \cdot \bar{\mathbf{V}} \in \mathbb{R}^{3\times 3}$. Examples for the non-scalar shear stress σ_{xy} are shown in figure (1) and the SM[59]. However, in the following only $\bar{\mathbf{D}}$ is used for forecasting, targeting the scalar pressure $\mathrm{Tr}(\boldsymbol{\sigma})$, as model parameters are scalars and $\bar{\mathbf{D}}$ has a metric distance[48].

Whilst (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 voxelised signal and the constraints required to preserve dislocation topology in forecasts. 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 dimension 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[59]. 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 figure 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 $\bar{\mathbf{X}}_n = \tau_D^{-1} \int_0^{\bar{\tau}} \mathbf{X}(t_n + t)\mathrm{d}t$ over a period $\bar{\tau} \simeq 20 - 50$fs to reduce noise from thermal fluctuations[69], 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 maximise 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, 70]

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

For $P > 1$ a Wold transformation[71] $\mathbf{Z}_n = 1 \oplus \bar{\mathbf{D}}_n \cdots \oplus \bar{\mathbf{D}}_{n-p} \in \mathbb{R}^{1+P\tilde{D}}$ casts (3) as a Markovian Ornstien-Uhlenbeck equation[72] $\mathbf{Z}_{n+1} = \mathbf{T}\mathbf{Z}_n + \mathbf{W}_n$. The maximum likelihood estimator of $\mathbf{T}$ is simply the least squares solution, with $\mathbf{S}$ determined from the

residual covariance[70]. To minimize generalization error a bagging[73–75] approach was developed, applying Bayesian ridge regression[76] 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/LSTM) neural networks[77, 78] or neural differential equations[79] which require significant resources, regularisation/correction schemes[47], and limited in practice to data dimension $\tilde{D} < 10$[79, 80]. A Chapman-Komologorov test[6] for the transfer matrix $\mathbf{T}$ is provided in the SM[59], but in practice the light computational demand also permits a convergence test of model architecture by increasing P[59].

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 non-stationary 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, 65, 81]; here, the Mahalanobis outlier distance[53] is used for the unimodal descriptor distribution[56, 59]. With training data mean $\boldsymbol{\mu}_{\rm tr}$ and covariance $\boldsymbol{\Sigma}_{\rm tr}$ estimated via a shinkage estimator[82], the squared Mahalanobis distance reads

$$\mathcal{M}(\bar{\mathbf{D}}) = \left[\bar{\mathbf{D}} - \boldsymbol{\mu}_{\rm tr}\right] \boldsymbol{\Sigma}_{\rm tr}^{-1} \left[\mathbf{D} - \boldsymbol{\mu}_{\rm tr}\right] / \tilde{D}. \tag{4}$$

Importantly, (4) is independent of the VAR(P) forecast model (3); points drawn from a low density region of $\rho_{\rm tr}$ 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[59], with $\langle\mathcal{M}\rangle$ constant. However, model parameters cannot be assumed static, with a time dependence bounded from below by $1/\tau_M = 1/(M\tau_{\rm tr})$[83], where M is the ensemble size and $\tau_{\rm tr}$ training duration. This drift can be estimated by propagating epistemic uncertainty in the steady state to an uncertainty $\sigma^2_{\mathcal{M}}$ in $\mathcal{M}(\bar{\mathbf{D}})$, which should be accumulated[59], leading to an additional linear growth in (4) of

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

where $\mathcal{M}_\sigma(t) = \sigma^2_{\mathcal{M}} t/\delta\tau$ 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 $\eta = \tau_{\rm pred}/(M\tau_{\rm tr})$, where $\mathcal{M}(\tau_{\rm pred}) \equiv \mathcal{M}_0(M\tau_{\rm tr}) = 2$, 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, 84], but for large sizes and high temperatures the landscape of energy minimia is vast and insufficiently metastable for current acceleration methods[40]. The current application to M =60 4000-atom EAM-Pt[85] nanoparticles at 900K is thus out-of-scope for existing methods.

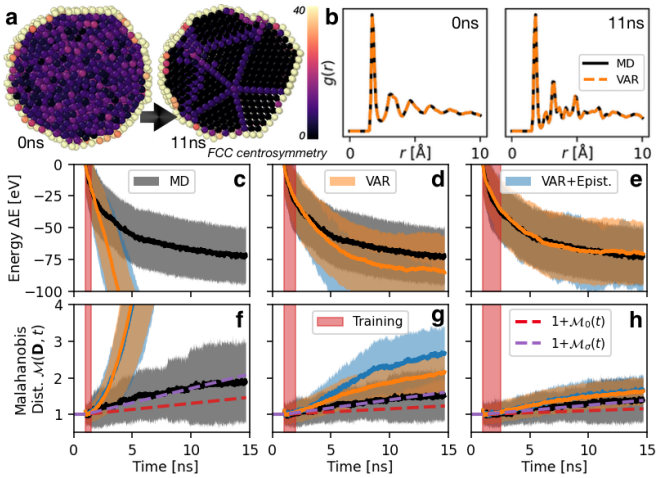

FIG. 2. Annealing of Pt nanoparticles. a) Representative structure at 0ns and 11ns. b) The average RDF $g(r)$ and the corresponding descriptor estimation[59]$g(r;\bar{\mathbf{D}})$. c)-h) Ensemble data with $M = 60$, $\tau_{\rm tr} = 0.5, 1.0, 1.5$ns, training starting at t =1ns (left-right, red shade). Mean is solid line, with standard deviation as bands. Black: MD data. Orange: VAR forecasts from 1ns. Blue: VAR forecast with epistemic errors. c)-e) Potential energy change from 1ns 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 $\mathcal{M}_\sigma$ (purple dash) are also shown.

The initial structure was formed by quenching from the liquid state and annealing for 100ps to give a highly disordered but predominantly fcc structure ($c_{\rm FCC} \simeq 0.5$). Descriptor trajectories were extracted every 1.5 ps, with a full structural analysis undertaken every 100ps, though the dataset was sparsified by taking $\delta\tau$ = 15ps and removing intermediate snapshots. Autoregressive models (3) were constructed with $P = 1 - 3$ and $\tau_{\rm tr}$=0.5, 1.0 or 1.5ns, 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 $c_{\rm FCC}$. MD data used $\mathcal{M}(\mathbf{D})$, eq. (4), which closely follows the theoretical lower bound $\mathcal{M}_\sigma(t)$. Whilst forecasts improve with training data, crucially, the magnitude of $\mathcal{M}(\bar{\mathbf{D}}, t)$ can independently confirm their reliability.

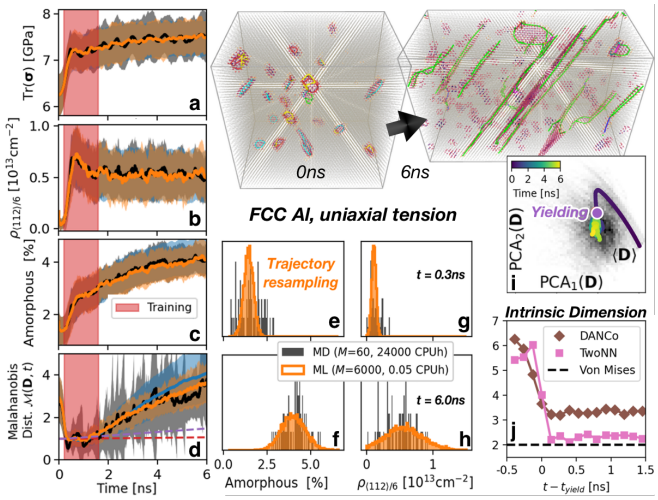

FIG. 3. Yielding of Al under uniaxial tension. Colors follow figure 2. Forecasts are from 0.2ns, with $\tau_{\text{tr}} = 1.2$ns and $P = 5$. a) Pressure, b) $\langle 112\rangle/6$ dislocation density and c) amorphous content. d) Mahalanobis distance. e)-h) Trajectory resampling from 0.3ns (pre-yield), with 100× larger ensemble. i) PCA analysis of the ensemble mean $\langle\bar{\mathbf{D}}\rangle$, clearly showing a localization on yield. Individual trajectories shown as histogram in grayscale. j) ID of the descriptor manifold, estimated via TwoNN[86] and DANCo[87]. Both show a collapse on yielding, but remain above the Von Mises lower bound.

Yielding of Al under uniaxial tension Dislocations carry plastic deformation, forming dense networks under irradiation[20] or extended loading[88]; understanding network evolution remains a grand challenge of physics and engineering[17, 89, 90]. Atomic simulations continue to discover mechanisms even in model systems[90], in part due to the difficulty in analysing atomic data[36]. An ensemble of $M = 60$ dense dislocation networks were formed in an EAM model of Al[91] 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[59] shows application to cyclic shear loading. Whilst typical in MD[17, 90], 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 the deformation of real microstructures. The results are summarized in figure 3a)-d), using the linear estimators (2). Increasing τ_{tr} decreased error and uncertainty, with optimal results found using $P = 5$[59]. Training only on pre-yield 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 pre-transition 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 figure 3e)-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 0.3-0.31ns. The forecasted ensemble captures multiple important structural evolutions that, whilst 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[92], the expected sharp peak in HCP content at yield, accompanied by a growth, peak then steady state in the number of dislocation junctions (see SM[59]). 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[93]. The global descriptor signal $\bar{\bar{\mathbf{D}}}$ can classify yielding with minimal training data, see [59]; figure 3i) shows how the ensemble average $\langle\bar{\bar{\mathbf{D}}}\rangle$ clearly localizes post yield.

Models for yielding invoke the concept of a yield surface in 6D stress space[94], which for metallic systems is typically the Von Mises yield surface, isosurfaces of the J_2 invariant with intrinsic dimension (ID)[58, 87] 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[86, 87, 95], which typically underestimate[87], were applied to the full $\bar{\mathbf{D}} \oplus \bar{\mathbf{V}}$ dataset. Figure 3j) shows both ID estimates collapse from 5-7 to 2-3 on yield, consistent with the Von Mises lower bound of 2. Whilst 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 equivarant descriptor signal to capture long-range correlations.

ACKNOWLEDGMENTS

I thank M-C Marinica, L Truskinovsky for stimulating discussion and anonymous referees for their careful reading of the manuscript. Support from ANR grant 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@cnrs.fr

[1] D. J. Wales, *Energy Landscapes*, edited by C. U. Press (Cambridge, 2003).
[2] D. Perez, B. P. Uberuaga, Y. Shim, J. G. Amar, and A. F. Voter, Annual Reports in computational chemistry **5**, 79 (2009).
[3] S. Plimpton, Journal Computational Physics **117**, 1 (1995).
[4] D. Perez, E. D. Cubuk, A. Waterland, E. Kaxiras, and A. F. Voter, Journal of chemical theory and computation **12**, 18 (2015).
[5] A. Laio and M. Parrinello, Proceedings of the National Academy of Sciences **99**, 12562 (2002).
[6] A. Mardt, L. Pasquali, H. Wu, and F. Noé, Nature communications **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.*, Modelling and Simulation in Materials Science and Engineering **28**, 043001 (2020).
[8] L. Bonati, G. Piccini, and M. Parrinello, Proceedings of the National Academy of Sciences **118**, e2113533118 (2021).
[9] T. D. Swinburne and D. Perez, Modelling and Simulation in Materials Science and Engineering **30**, 034004 (2022).
[10] J. D. Honeycutt and H. C. Andersen, Journal of Physical Chemistry **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, Proceedings of the National Academy of Sciences **112**, E5769 (2015).
[15] A. Stukowski, V. V. Bulatov, and A. Arsenlis, Modelling and Simulation in Materials Science and Engineering **20**, 085007 (2012).
[16] G. Wu, H. Song, and D. Lin, Computational Materials Science **144**, 322 (2018).
[17] L. A. Zepeda-Ruiz, A. Stukowski, T. Oppelstrup, and V. V. Bulatov, Nature **550**, 492 (2017).
[18] W. Setyawan, G. Nandipati, K. J. Roche, H. L. Heinisch, B. D. Wirth, and R. J. Kurtz, Journal of Nuclear Materials **462**, 329 (2015).
[19] N. V. Priezjev, Journal of Non-Crystalline 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.*, Physical Review Letters **125**, 225503 (2020).
[21] T. Swinburne and D. Perez, npj Computational Materials **6**, 190 (2020).
[22] A. Garmon, V. Ramakrishnaiah, and D. Perez, Parallel Computing , 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.*, Advanced Energy Materials **12**, 2102638 (2022).
[24] L.-F. Zhu, J. Janssen, S. Ishibashi, F. Körmann, B. Grabowski, and J. Neugebauer, Computational Materials Science **187**, 110065 (2021).
[25] J. Andrews, O. Gkountouna, and E. Blaisten-Barojas, Chemical Science **13**, 7021 (2022).
[26] C. Le Bris, T. Lelievre, M. Luskin, and D. Perez, Monte Carlo Methods and Applications **18**, 119 (2012).
[27] A. F. Voter, Physical Review B **57**, R13985 (1998).
[28] G. Henkelman and H. Jónsson, The Journal of chemical physics **111**, 7010 (1999).
[29] M. So and A. Voter, The Journal of Chemical Physics **112**, 9599 (2000).
[30] A. F. Voter, Physical Review Letters **78**, 3908 (1997).
[31] G. Henkelman, B. P. Uberuaga, and H. Jonsson, The Journal of Chemical Physics **113**, 9901 (2000).
[32] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, The Journal of chemical physics **128**, 144120 (2008).
[33] T. Lelièvre, G. Stoltz, and M. Rousset, *Free energy computations: A mathematical perspective* (World Scientific, 2010).
[34] T. D. Swinburne and M.-C. Marinica, Phys. Rev. Lett. **120**, 135503 (2018).
[35] J. Rogal, The European Physical Journal 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. , (2022).
[37] G. Bussi and A. Laio, Nature Reviews Physics **2**, 200 (2020).
[38] F. Noé and F. Nuske, Multiscale Modeling & Simulation **11**, 635 (2013).
[39] R. Huang, L.-T. Lo, Y. Wen, A. Voter, and D. Perez, The Journal of chemical physics **147**, 152717 (2017).
[40] R. Huang, Y. Wen, A. Voter, and D. Perez, Physical Review Materials **2**, 126002 (2018).
[41] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, and J. C. Grossman, Nature communications **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 Computational Materials **5**, 125 (2019).
[44] L. Klein, A. Y. Foong, T. E. Fjelde, B. Mlodozeniec, M. Brockschmidt, S. Nowozin, F. Noé, and R. Tomioka, arXiv preprint arXiv:2302.01170 (2023).
[45] C. Han, P. Zhang, D. Bluestein, G. Cong, and Y. Deng, Journal of Computational Physics **427**, 110053 (2021).
[46] B. Leimkuhler, D. T. Margul, and M. E. Tuckerman, Molecular Physics **111**, 3579 (2013).

[47] X. Fu, T. Xie, N. J. Rebello, B. D. Olsen, and T. Jaakkola, arXiv preprint arXiv:2204.10348 (2022).

[48] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, J. Comp. Phys. **285**, 316 (2015).

[49] A. Shapeev, Multiscale Model. Sim. **14**, 1153 (2016).

[50] A. M. Goryaeva, J. Dérès, C. Lapointe, P. Grigorev, T. D. Swinburne, J. R. Kermode, L. Ventelon, J. Baima, and M.-C. Marinica, Phys. Rev. Materials **5**, 103803 (2021).

[51] A. E. Allen, G. Dusson, C. Ortner, and G. Csányi, Machine Learning: Science and Technology **2**, 025017 (2021).

[52] Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, *et al.*, npj Computational Materials **7**, 1 (2021).

[53] P. C. Mahalanobis (National Institute of Science of India, 1936).

[54] E. V. Podryabinkin and A. V. Shapeev, Computational Materials Science **140**, 171 (2017).

[55] N. Bernstein, G. Csányi, and V. L. Deringer, npj Computational Materials **5**, 1 (2019).

[56] A. M. Goryaeva, C. Lapointe, C. Dai, J. Dérès, J.-B. Maillet, and M.-C. Marinica, Nat. Commun. **11**, 4691 (2020).

[57] H. Lütkepohl, *New introduction to multiple time series analysis* (Springer Science & Business Media, 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] See Supplemental Material [url] for a convergence tests and extended figures, which includes Refs [96, 97],.

[60] V. L. Deringer, M. A. Caro, and G. Csányi, Advanced Materials **31**, 1902765 (2019).

[61] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).

[62] I. Batatia, D. P. Kovács, G. N. Simm, C. Ortner, and G. Csányi, arXiv preprint arXiv:2206.07697 (2022).

[63] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Science advances **3**, e1603015 (2017).

[64] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, Nature communications **13**, 1 (2022).

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

[66] In practice, the tensor of symmetric matricies $\bar{\mathbf{V}}_s \equiv \bar{\mathbf{V}} + \bar{\mathbf{V}}^\top$ is calculated in current implementations[3, 50].

[67] C. Lapointe, T. D. Swinburne, L. Thiry, S. Mallat, L. Proville, C. S. Becquart, and M.-C. Marinica, Physical Review Materials **4**, 063802 (2020).

[68] C. Lapointe, T. D. Swinburne, L. Proville, C. S. Becquart, N. Mousseau, and M.-C. Marinica, Under Review (2022).

[69] T. D. Swinburne, S. L. Dudarev, and A. P. Sutton, Physical Review Letters **113**, 215501 (2014).

[70] S. Karlsson, Handbook of economic forecasting **2**, 791 (2013).

[71] H. O. Wold, The Annals of Mathematical Statistics **19**, 558 (1948).

[72] W. Coffey and Y. P. Kalmykov, *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*, Vol. 27 (World Scientific, 2012).

[73] L. Breiman, Machine learning **24**, 123 (1996).

[74] B. Lakshminarayanan, A. Pritzel, and C. Blundell, Advances in neural information processing systems **30** (2017).

[75] F. Petropoulos, R. J. Hyndman, and C. Bergmeir, European Journal of Operational Research **268**, 545 (2018).

[76] 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, Journal of Machine Learning Research **12**, 2825 (2011).

[77] Y. Yu, X. Si, C. Hu, and J. Zhang, Neural computation **31**, 1235 (2019).

[78] P. R. Vlachas, J. Zavadlav, M. Praprotnik, and P. Koumoutsakos, Journal of Chemical Theory and Computation **18**, 538 (2021).

[79] P. Kidger, J. Morrill, J. Foster, and T. Lyons, Advances in Neural Information Processing Systems **33**, 6696 (2020).

[80] F. Liu, M. Cai, L. Wang, and Y. Lu, IEEE Access **7**, 26102 (2019).

[81] C. Zeni, A. Anelli, A. Glielmo, and K. Rossi, Physical Review B **105**, 165141 (2022).

[82] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, IEEE transactions on signal processing **58**, 5016 (2010).

[83] T. D. Swinburne and D. Perez, Phys. Rev. Materials **2**, 053802 (2018).

[84] F. Baletto, R. Ferrando, A. Fortunelli, F. Montalenti, and C. Mottet, The Journal of chemical physics **116**, 3856 (2002).

[85] C. Liu, J. Cohen, J. Adams, and A. Voter, Surface science **253**, 334 (1991).

[86] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, Scientific reports **7**, 12140 (2017).

[87] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, Pattern recognition **47**, 2569 (2014).

[88] S. Lavenstein, Y. Gu, D. Madisetti, and J. A. El-Awady, Science **370**, eabb2690 (2020).

[89] R. Baggio, E. Arbib, P. Biscari, S. Conti, L. Truskinovsky, G. Zanzotto, and O. Salman, Physical Review Letters **123**, 205501 (2019).

[90] N. Bertin, W. Cai, S. Aubry, A. Arsenlis, and V. V. Bulatov, arXiv preprint arXiv:2210.14343 (2022).

[91] B. Onat and S. Durukanoğlu, Journal of Physics: Condensed Matter **26**, 035404 (2013).

[92] J. Marian, W. Cai, and V. V. Bulatov, Nature materials **3**, 158 (2004).

[93] R. B. Sills, N. Bertin, A. Aghaei, and W. Cai, Physical review letters **121**, 085501 (2018).

[94] J. P. Hirth and J. Lothe, *Theory Of Dislocations* (Malabar, FL Krieger, 1991).

[95] J. Bac, E. M. Mirkes, A. N. Gorban, I. Tyukin, and A. Zinovyev, Entropy **23**, 1368 (2021).

[96] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, Journal of Nonlinear Science **25**, 1307 (2015).

[97] C. E. Rasmussen, *Gaussian Processes in Machine Learning* (Springer, Berlin, Heidelberg, 2004).