Differentiable, model-agnostic free energy calculation

Thomas D. Swinburne
University of Michigan and CNRS
tswin@umich.edu

Clovis Lapointe, Mihai-Cosmin Marinica
CEA and University Paris-Saclay
{clovis.lapointe,mihai-cosmin.marinica}@cea.fr

Abstract

The vibrational free energy is essential to predict finite temperature material properties. Current methods employ slow, largely sequential sampling with a fixed machine learning interatomic potential (MLIP) to satisfy the tight 1-2meV/atom (1/40-1/20 kcal/mol) convergence requirements. Forward or back propagation of MLIP parameters is not practically possible, meaning estimates cannot be used in objective functions for alignment to reference data or distillation. For the broad class of generalized linear MLIPs we show free energies can be cast as the Legendre transform of a high-dimensional descriptor entropy, accurately estimated via score matching. Our main result is a model-agnostic estimator which returns meV/atom accurate, end-to-end free energies as a function of MLIP parameters. Sampling is efficient and highly parallel, requiring 10x fewer force calls and 100-1000x less walltime than a single thermodynamic integration estimate. Tensor compression allows lightweight storage and inference is instantaneous. In forward propagation, a single estimator predicts a broad ensemble high temperature thermodynamic integration calculations for W. In back-propagation, we fine-tune the $\alpha - \gamma$ transition temperature in an Fe model from 2000K to 1063K, a first demonstration of MLIP alignment against known phase boundaries.

1 Introduction

A realistic material design scheme must account for thermal vibrations, essential to target basic properties such as phase stability, heat capacity, elastic constants or thermal expansion coefficients. In atomic simulation, the computational task is to compute the vibrational (Helmholtz) free energy $\mathcal{F}$ over some set of crystalline phases at a range of temperatures and volumes. Accurate phase prediction requires tightly converged estimates of $\mathcal{F}$, to within 1-2 meV/atom, or 1/40-1/20 kcal/mol. Machine learning interatomic potentials (MLIPs) are becoming a viable replacement to *ab initio* calculation, but remain misspecified[83]. For uncertainty quantification[83, 54, 69, 14, 101], inverse design or top-down learning[43, 88, 17, 70], schemes which allow end-to-end differentiation through MLIP simulations are actively sought. However, estimating $\mathcal{F}$ requires high dimensional integration, one of the most challenging tasks in computational science, the central difficulty in e.g. evidence calculation[93, 16, 61] or density estimation[52]. Current schemes (figure 1, appendix A) perform slow, largely sequential stratified sampling[48] with a *single choice* of best-fit MLIP parameters. Whilst established, such estiamtes are not differentiable and as such finite temperature properties cannot be included in objective functions for alignment against known data or model distillation.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat.

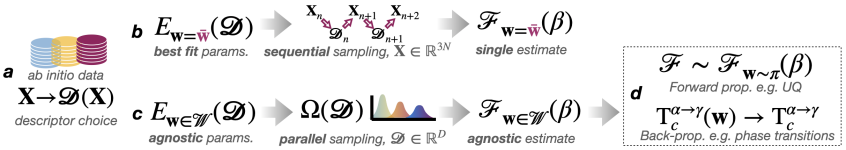

Figure 1: Model agnostic sampling. a) Atomic simulations choose features (descriptors) adapted to some *ab initio* data. b) Current methods sample with a *single* choice of MLIP parameters $\mathbf{w} = \bar{\mathbf{w}}$, potentially using descriptors to bias proposals[87, 2, 97, 72]. c) Our approach (D-DOS) learns an *agnostic* estimator which returns free energies and gradients for any $\mathbf{w} \in \mathcal{W}$. d) D-DOS enables rapid forward/back propagation for UQ or inverse fine-tuning, e.g. targeting phase transitions.

1.1 Main contributions

We provide a model-*agnostic* free energy estimation scheme for atomic simulation, introducing the descriptor density of states (D-DOS) $\Omega(\boldsymbol{\mathcal{D}})$. The associated descriptor entropy $\mathcal{S}(\boldsymbol{\mathcal{D}})$ is a Legendre conjugate to $\mathcal{F}$ and can be accurately estimated by score-matching. We show this allows meV/atom accurate prediction without *a priori* specification of model parameters, demonstrating application in forward propagation for UQ and in back propagation to fine-tune the $\alpha - \gamma$ transition temperature in a model of Fe, to our knowledge a first demonstration of MLIP alignment to phase boundaries.

1.2 Related Work

A range of specialized techniques to estimate $\mathcal{F}$ are well established, all some form of stratified sampling[59] from an analytically tractable reference model[102, 48, 100, 103, 65] (appendix A). Recent studies[100, 65, 21] have shown MLIPs can provide near-*ab initio* accurate free energy predictions, especially when fine-tuned for specific phases[48]. We focus on the popular models[80, 90, 62, 44, 71], including message-passing neural networks[9, 76], where atomic configurations are encoded using (possibly learned) *descriptor* functions ensuring outputs are symmetric under permutation, translation and rotation. Multiple works have noted descriptors are an ideal latent space for generative models of dynamics[84] or thermodynamic samples, using e.g. normalizing flows[87, 2, 97, 72] or variational autoencoders[5] to accelerate convergence. While perturbative approaches allow some forward propagation for UQ[54], no methods to date allow practical back-propagation, as gradient evaluation requires converging an expectation value- this is the computational effort of established methods such as thermodynamic integration. We note that policy gradient algorithms from reinforcement learning (RL) such as REINFORCE[96] have a conceptual similarity with free energy gradients, as both can be express the gradient of a log density as an expectation. However, in the RL setting these gradients are used in stochastic optimization, where one only requires an *unbiased estimation* of the expectation, rather than a converged value, which requires much less computational effort. To our knowledge, the approach we present here is the only method which is able to rapidly evaluate accurate gradients over a range of parameter values simultaneously, allowing for the inclusion of free energies in loss functions for fine-tuning / alignment purposes.

2 Methodology

2.1 Generalized linear machine learning interatomic potentials

With atomic positions $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and species $\mathbf{S} \in \mathbb{Z}^N$ in a periodic supercell $\mathbf{C} \in \mathbb{R}^{3\times 3}$, a general MLIP energy writes $E_{\mathbf{w}}(\mathbf{X}) = \sum_{i=1}^{N} E^1_{\mathbf{w}}(\mathbf{D}_i)$, where the descriptor vector $\mathbf{D}_i(\mathbf{X}, \mathbf{S}) \in \mathbb{R}^d$ depends only on atoms in the vicinity of i (appendix A). and $\mathbf{w}$ is a vector of parameters. In this paper we consider MLIPs of the generalized linear form, with parameters $\mathbf{w} \in \mathbb{R}^D$,

$$E_{\mathbf{w}}(\mathbf{X}) \equiv N\mathbf{w} \cdot \hat{\boldsymbol{\mathcal{D}}}(\mathbf{X}), \quad \hat{\boldsymbol{\mathcal{D}}} \equiv \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\phi}}(\mathbf{D}_i) \in \mathbb{R}^D, \tag{1}$$

where $\hat{\boldsymbol{\phi}}(\mathbf{D}_i) = [\hat{\phi}_1(\mathbf{D}_i), \ldots, \hat{\phi}_D(\mathbf{D}_i)]$ is a D-dimensional featurization of the $\mathbf{D}_i(\mathbf{X}, \mathbf{S})$. Importantly, the vector $\boldsymbol{\mathcal{D}}$ is independent of $\mathbf{w}$ and the dimension D is *intensive*, independent of N. A wide variety of MLIPs can be cast into the general linear form (1). Clearly, these include the broad class of

linear-in-descriptor models, where $\hat{\boldsymbol{\phi}}(\mathbf{D}_i) = \mathbf{D}_i \in \mathbb{R}^d$, including `MTP`[80], `ACFS` [12, 13], `SNAP`[90], `SOAP` [6, 27, 28], `ACE`[31, 32, 62], `MILADY`[44, 35] and `POD`[71] descriptors. Linear descriptor models can reach extremely high (< 2 meV/atom) accuracy to *ab initio* data[21], with robust UQ[83] and often excellent dynamical stability, essential for thermodynamic sampling[102, 48, 100, 103, 65]. Polynomial or kernel featurizations are regularly used to increase flexibility, e.g. `qSNAP`[79], `PiP`[3] `GAP`[7, 30], n-body kernels [41, 42, 91, 99], kernel[100] etc. We can encompass foundational models such as `MACE`[9] by Taylor expanding in parameters[24] or only adjusting a subset of parameters, an approach adopted when fine-tuning recent neural network models[11, 68, 10, 9, 18, 23, 8]. For example, in the `MACE` architecture[10], the input to the final readout layer is taken as $\mathbf{D}_i$, giving the featurization $\hat{\boldsymbol{\phi}}_{\texttt{MACE}}(\mathbf{D}_i) = \mathbf{D}_i \oplus f(\mathbf{D}_i) \in \mathbb{R}^{d+1}$, where $f(\mathbf{D}_i)$ is frozen one layer neural network. Recent work has shown this allows UQ for linear models[83] to be applied[76] to the `MACE-MPA-0` foundation model[9]. Generalized linear models can distill $E_{\mathbf{w}}(\mathbf{D}_i)$ as recently demonstrated for `ACE` on `MACE` [77]. To retain connection with $\mathbf{w}$ one must find features $\hat{\boldsymbol{\psi}}(\mathbf{w})$ such that $\sum_i \|\hat{\boldsymbol{\psi}}(\mathbf{w})^\top \hat{\boldsymbol{\phi}}(\mathbf{D}_i) - E_{\mathbf{w}}(\mathbf{D}_i)\|^2$, an extension which will be reported elsewhere.

2.2 Free energy and the descriptor DOS

Our primary thermodynamic property is the NVT free energy $\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p)$ for vibrations around a crystalline phase $p \in \mathcal{P} = \{\text{BCC, FCC, hcp, } \dots\}$. The Gibbs free energy reads $\mathcal{G}_{\mathbf{w}}(\beta, \boldsymbol{\sigma}, p) \equiv \min_{\mathbf{C}} \mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p) - \mathrm{Tr}(\mathbf{C}\boldsymbol{\sigma})$. Suppressing p and $\mathbf{C}$ for clarity, $\mathcal{F}_{\mathbf{w}}(\beta)$ for linear MLIPs (1) writes

$$\mathcal{F}_{\mathbf{w}}(\beta) = \lim_{N\to\infty} \frac{-1}{N\beta} \ln \left| \lambda_\beta^{-3N} \int e^{-N\beta \mathbf{w}\cdot\boldsymbol{\mathcal{D}}} \Omega(\boldsymbol{\mathcal{D}}) \mathrm{d}\boldsymbol{\mathcal{D}} \right|, \quad \Omega(\boldsymbol{\mathcal{D}}) \equiv \int_{\mathbb{R}^{3N}} \delta(\hat{\boldsymbol{\mathcal{D}}}(\mathbf{X}) - \boldsymbol{\mathcal{D}}) \mathrm{d}\mathbf{X}, \quad (2)$$

where $\lambda_\beta = \hbar\sqrt{2\pi\beta/m}$ is the thermal De Broglie wavelength[39], $m \equiv (\prod_{i=1}^N m_i)^{1/N}$ and $\Omega(\boldsymbol{\mathcal{D}})$ the *descriptor density of states* (D-DOS). Access to $\Omega(\boldsymbol{\mathcal{D}})$ would allow prediction of $\mathcal{F}_{\mathbf{w}}(\beta)$ for *any value* of $\mathbf{w}$ but there are two significant issues: 1) $\Omega(\boldsymbol{\mathcal{D}})$ is ill-conditioned $\int_{\mathbb{R}^D} \Omega(\boldsymbol{\mathcal{D}}) \mathrm{d}\boldsymbol{\mathcal{D}} = V^N$, the key issue in e.g. nested sampling[94, 73] at large N, 2) $D = \mathcal{O}(100 - 1000)$, meaning $\Omega(\boldsymbol{\mathcal{D}})$ cannot be evaluated by quadrature, while Monte Carlo integration cannot give reliable gradients.

2.3 Free energy as a Legendre transform in the thermodynamic limit

We overcome the first ill-conditioning issue in $\Omega(\boldsymbol{\mathcal{D}})$ through a *conditional* D-DOS (CD-DOS)

$$\Omega(\boldsymbol{\mathcal{D}}) \equiv \int_{\mathbb{R}} \Omega(\boldsymbol{\mathcal{D}}|\alpha)\Omega_0(\alpha)\mathrm{d}\alpha, \quad \Omega_0(\alpha) = \int_{\mathbb{R}^{N\times 3}} \delta(\hat{\alpha}(\mathbf{X}) - \alpha)\mathrm{d}\mathbf{X}, \quad (3)$$

where $\hat{\alpha}(\mathbf{X})$ is an *isosurface function* satisfying $\int_{\mathbb{R}} \Omega_0(\alpha)\mathrm{d}\alpha \equiv V^N$, i.e. a folation of configuration space. As detailed in appendix B, we choose $\hat{\alpha}(\mathbf{X}) = \ln|E_0(\mathbf{X})/N|$ or $\hat{\alpha}(\mathbf{X}, \mathbf{P}) = \ln|(\|\mathbf{P}\|^2/2m + E_0(\mathbf{X}))/N|$ such that sampling reduces to generating Gaussian noise or short NVE trajectories, giving $\Omega_0(\alpha)$ analytically or numerically. To avoid quadrature we apply Laplace's method (appendix D.1) in the thermodynamic limit $N \to \infty$ to give the formally exact, integration-free expression

$$\mathcal{F}_{\mathbf{w}}(\beta) \equiv \min_{\alpha, \boldsymbol{\mathcal{D}}} \left(\mathbf{w}\cdot\boldsymbol{\mathcal{D}} - [\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) + \mathcal{S}_0(\alpha)]/\beta \right), \quad (4)$$

as derived in appendix B, where we have defined the intensive descriptor entropies

$$\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) \equiv \lim_{N\to\infty} \ln|\Omega(\boldsymbol{\mathcal{D}}|\alpha)|/N, \quad \mathcal{S}_0(\alpha) \equiv \ln|\Omega_0(\alpha)/V_0^N|, \quad (5)$$

where V_0 ensures $\mathcal{S}_0(\alpha)$ is dimensionless, defined in appendix D. Equation (4) is our central theoretical result, an expression for the free energy where terms estimated via sampling do not require *a priori* specification of parameters. As $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ is normalized one can show $\max_{\boldsymbol{\mathcal{D}}} \mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) = 0$, a crucial point as any estimate $\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) \simeq \nabla\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ can then be integrated to give $\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$.

2.4 Score matching the descriptor entropy

The score matching loss[52] for $\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$ reads, using $\langle \dots \rangle_\alpha$ for $\hat{\alpha}(\mathbf{X}) = \alpha$ averages (see C.1)

$$\mathcal{L}(\boldsymbol{\Theta}|\alpha) \equiv \left\langle \frac{N}{2} \|\boldsymbol{\nabla}\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\|^2 + \boldsymbol{\nabla}^2\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) \right\rangle_\alpha . \quad (6)$$

Table 1: Approximate cost of methods to estimate $\mathcal{F}$ to within the 1-2meV/atom convergence required for phase prediction. $|\Delta\mathcal{F}|$: approx. max deviation from reference, in meV/atom, that can be targeted at 1000 K. Calls: indication of total cost in force calls. Calls/Worker: indicates total wall-time / strong scaling. Only D-DOS is model-agnostic, sampling *once* for all parameter values.

Method	$\lvert\Delta\mathcal{F}\rvert$	**Calls**	**Calls/Worker**	**Agnostic**	**Differentiable**
FEP[39]	10	$\sim 10^6$	$\sim 10^4$	No	No
TI[4, 100]	150	$\sim 10^6$	$\sim 10^5$	No	No
AS[65]	150	$\sim 10^8$	$\sim 10^7$	No	No
D-DOS	**200**	$\sim \mathbf{10^5}$	$\sim \mathbf{10^2}$	**Yes**	**Yes**

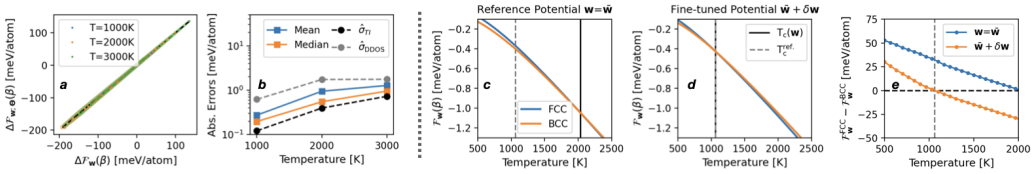

Figure 2: Propagating through D-DOS free energies. a): Accuracy of a *single* D-DOS estimator against an unseen dataset of thermodynamic integration (TI) calculations for an ensemble $\mathbf{w} \in \mathcal{W}$ of BCC W models, up to 3000K. b) Mean (blue) and median (orange) errors across the ensemble are below 1.5meV/atom, comparable to the TI convergence estimate (black), while the D-DOS predicted errors (gray) are robust bounds. c)-e): Aligning Fe model to match $\alpha \to \gamma$ transition. c): An initial Fe model $\mathbf{w} = \bar{\mathbf{w}}$ has FCC (blue) and BCC (orange) free energies that give an $\alpha \to \gamma$ transition at $T_c(\bar{\mathbf{w}}) = 2030$ K. d) Back-propagating through $\mathcal{F}_{\mathbf{w}}^{\mathrm{BCC}} - \mathcal{F}_{\mathbf{w}}^{\mathrm{FCC}}$ in fine tuning loss gives a new set of MLIP parameters where the transition is at the correct value of $T_c(\bar{\mathbf{w}} + \delta\mathbf{w}) = 1063$K. e) Plotting just the FCC-BCC difference shows how small changes in $\mathcal{F}$ give large changes in T_c.

The factor of N emerges from application of integration by parts[52] in the derivation of (6). While $\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$ is intensive, averages over α give rise to terms $\mathcal{O}(N^{-s})$ in $\mathcal{L}(\boldsymbol{\Theta}|\alpha)$, whose minimization formally requires multiscale analysis[74] to solve, (see C.1) but in practice this is not required. We employ a lightweight tensor compression scheme[81] for $\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$, as detailed in appendix C.1, which allows for simple error propagation, as detailed in appendix C.4. For $D = \mathcal{O}(100)$ storage at fixed phase p and supercell $\mathbf{C}$ requires only $3 - 10$ MB storage, allowing broad precomputation. Table 1 compares compute effort for a *single* traditional estimate, showing the highly parallelized D-DOS score matching requires $10\times$ less total effort and $100 - 1000\times$ less wall time. Crucially, D-DOS estimates are uniquely *model agnostic* across a broad parameter range $\mathbf{w} \in \mathcal{W}$ and differentiable, allowing simple inclusion of $\mathcal{F}_{\mathbf{w}}(\beta)$ in objective functions.

3 Results

Our D-DOS estimator is built using a single MLIP $\mathbf{w} = \bar{\mathbf{w}}$ to generate samples on isosurfaces $\hat{\alpha}(\mathbf{X}) = \alpha$ and a score-match a descriptor entropy model $\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$. To test the accuracy of our estimator we generated a broad ensemble of models $\mathbf{w} \in \mathcal{W}$ approximating W, Mo and Fe (appendix F) for which we calculated free energies in BCC, FCC and A15 phases through thermodynamic integration for $\mathrm{T} \in [300, 3000\mathrm{K}]$. In our numerical tests we built MLIPS using the popular BSO(4) descriptor functions, first introduced in the `SNAP` MLIP family[90], giving simply $\boldsymbol{\mathcal{D}}_i = \hat{\boldsymbol{\phi}}(\mathbf{D}_i) = \mathbf{D}_i$. Appllication over broader MLIP families similar to `ACE` or `MACE` will be presented in future work.

3.1 Comparision against thermodynamic integration benchmarks in forward propagation

As a first test in forward propagation, figures 2a) and b) present D-DOS predictions for models $\mathbf{w} \in \mathcal{W}$ approximating BCC W against thermodynamic integration (TI) calculations for each potential. We emphasize that the estimator only uses a harmonic reference model for sampling, with no training on any of the TI calculations. The total sampling budget required only $0.1\times$ cost of a *single* TI calculation, with far superior strong scaling and essentially instantaneous inference. Even

at high homogulous temperatures of 3000K, where the ensemble models show up to 200meV/atom explicit anharmonicity from the reference MLIP $\mathbf{w} = \bar{\mathbf{w}}$ used for sampling, a *single* D-DOS estimator retains the 1-2meV/atom accuracy required for phase prediction. The ability to efficiently precompute and store model-agnostic thermodynamic averages holds many perspectives for error-controlled modelling and allows simulation results to be updated *a posteriori* for e.g. fine-tuning.

3.2 Aligning the $\alpha \rightarrow \gamma$ transition temperature in Fe

Figures 2c),d) and e) illustrate the central result of this short communication, demonstrating how back-propagation allows for the targeting of phase transition temperatures, to our knowledge, a unique ability of the D-DOS procedure. Targets could be calculations, prescribed from higher level simulations to enforce consistency[22] or to experimental data in top-down training[88]. Our demonstration targets the BCC-FCC, or $\alpha \rightarrow \gamma$, transition in Fe. While known to be due to the loss of ferromagnetic ordering[63], in this example of back-propagation we employ non-magnetic models. D-DOS estimators for FCC and BCC phases over a range of atomic volumes allow calculation of NPT free energy difference $\Delta_{\alpha-\gamma}\mathcal{G}_{\mathbf{w};\boldsymbol{\Theta}}(\beta)$. Our regularized loss function reads $\mathcal{L}(\mathbf{w}) = \|\Delta_{\alpha-\gamma}\mathcal{G}_{\mathbf{w};\boldsymbol{\Theta}}(\beta_c)\|^2 + r\mathcal{L}_0(\mathbf{w})$, where $1/(\mathrm{k_B}\beta_c) = 1063K$. We find the subtle changes in potential parameters required to reproduce the desired phase boundary, reducing the $\alpha \rightarrow \gamma$ transition temperature from 2030 K to 1063 K. As the free energy gradient with temperature is only around 0.03 meV/atom/K, the small changes of 30 meV/atom in $\Delta\mathcal{G}$ gives a 1000 K change in $\mathrm{T_c}$.

4 Limitations

In the present form the most significant limitation of our approach is specialisation to generalized linear MLIPS, but as we discuss in 2.1 a wide range of models fall into this class. A general extension to non-linear MLIPs would require estimating the joint density of the total per-atom descriptor vector $\boldsymbol{\mathcal{D}}_N \equiv \bigoplus_i \hat{\boldsymbol{\phi}}(\mathbf{D}_i) \in \mathbb{R}^{N\times D}$, i.e. accounting for local correlations, which will be the subject of future work. In addition, we also have only shown application to solid unary phases, but liquid phases and multi-component systems will be the subject of a forthcoming communication. While this will require additional conditional sampling constraints, this remains a feasible extension of the current framework and will be persued in the near future.

5 Conclusion

This paper presents a new approach to estimate the vibrational free energy of atomic systems, an essential component of any computational material design scheme. Rather than existing methods which return free energy estimates for a specific value of MLIP parameters, we instead return an estimator that can predict free energies over a broad range of model parameters. This is a significant change in approach that not only allows for rapid forward propagation of parameter uncertainties to finite temperature properties and pre-computation of expensive thermodynamic averages, but also uniquely allows for inverse fine-tuning of e.g. phase boundaries through back-propagation, all long-desired capabilities in modern computational materials science workflows.

6 Data availability

Pre-computed samples and a notebook to reproduce a simplified back-propagation result are available at `www.github.com/tomswinburne/DescriptorDOS.git`

7 Funding and acknowledgements

We gratefully acknowledge the hospitality of Institute for Pure and Applied Mathematics at the University of California, Los Angeles (NSF grant DMS-1925919), the Institute for Mathematical and Statistical Innovation, University of Chicago (NSF grant DMS-1929348) and the Institut Pascal at Université Paris-Saclay (ANR grant ANR-11-IDEX-0003-01). TDS gratefully acknowledges support from ANR grant ANR-23-CE46-0006-1, IDRIS allocation A0120913455, Euratom Grant No. 633053 and an Emergence@INP grant from the CNRS. All authors acknowledge the support from GENCI - (CINES/CCRT) computer centre under Grant No. A0170906973.

7.1 Code Availability

An open source, `pip`-installable implementation of the D-DOS code with `LAMMPS`[89] is available at `www.github.com/tomswinburne/DescriptorDOS.git`

References

[1] G. Adjanor, M. Athènes, and F. Calvo. Free energy landscape from path-sampling: application to thestructural transition in LJ38. *The European Physical Journal B - Condensed Matter and Complex Systems*, 53(1):47–60, September 2006.

[2] Rasool Ahmad and Wei Cai. Free energy calculation of crystalline solids using normalizing flows. *Modelling and Simulation in Materials Science and Engineering*, 30(6):065007, 2022.

[3] Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Atomic permutationally invariant polynomials for fitting molecular force fields. *Machine Learning: Science and Technology*, 2(2):025017, February 2021.

[4] Manuel Athènes and Pierre Terrier. Estimating thermodynamic expectations and free energies in expanded ensemble simulations: Systematic variance reduction through conditioning. *J. Chem. Phys.*, 146(19):194101, 05 2017.

[5] Jacopo Baima, Alexandra M Goryaeva, Thomas D Swinburne, Jean-Bernard Maillet, Maylise Nastar, and Mihai-Cosmin Marinica. Capabilities and limits of autoencoders for extracting collective variables in atomistic materials science. *Physical Chemistry Chemical Physics*, 24(38):23152–23163, 2022.

[6] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[7] Albert P. Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, August 2015.

[8] Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor NC Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e (3)-equivariant atom-centred interatomic potentials. *Nature Machine Intelligence*, pages 1–12, 2025.

[9] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.

[10] Ilyes Batatia, Dávid Péter Kovács, Gregor NC Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *arXiv preprint arXiv:2206.07697*, 2022.

[11] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[12] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, February 2011.

[13] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics*, 145(17):170901, November 2016.

[14] I. R. Best, T. J. Sullivan, and J. R. Kermode. Uncertainty quantification in atomistic simulations of silicon using interatomic potentials. *The Journal of Chemical Physics*, 161(6):064112, 08 2024.

[15] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[16] Harish S Bhat and Nitesh Kumar. On the derivation of the bayesian information criterion. *School of Natural Sciences, University of California*, 99:58, 2010.

[17] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-Lopez, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit

differentiation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5230–5242. Curran Associates, Inc., 2022.

[18] Anton Bochkarev, Yury Lysogorskiy, and Ralf Drautz. Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing. *Phys. Rev. X*, 14:021036, Jun 2024.

[19] L. Cao, G. Stoltz, T. Lelièvre, M.-C. Marinica, and M. Athènes. Free energy calculations from adaptive molecular dynamics simulations with adiabatic reweighting. *J. Chem. Phys.*, 140(10):104108, 2014.

[20] Aloïs Castellano, François Bottin, Johann Bouchet, Antoine Levitt, and Gabriel Stoltz. A b initio canonical sampling based on variational inference. *Physical Review B*, 106(16):L161110, 2022.

[21] Aloïs Castellano, Romuald Béjaud, Pauline Richard, Olivier Nadeau, Clément Duval, Grégory Geneste, Gabriel Antonius, Johann Bouchet, Antoine Levitt, Gabriel Stoltz, and François Bottin, 2024.

[22] Long-Qing Chen and Yuhong Zhao. From classical thermodynamics to phase-field method. *Progress in Materials Science*, 124:100868, 2022.

[23] Bingqing Cheng. Cartesian atomic cluster expansion for machine learning interatomic potentials. *npj Computational Materials*, 10(1):157, 2024.

[24] Sanggyu Chong, Filippo Bigi, Federico Grasselli, Philip Loche, Matthias Kellner, and Michele Ceriotti. Prediction rigidities for data-driven chemistry. *Faraday Discussions*, 256:322–344, 2025.

[25] J. Comer, J. C. Gumbart, J. Henin, T. Lelièvre, A. Pohorille, and C. Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B*, 119(3):1129, 2015. PMID: 25247823.

[26] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, July 2006.

[27] James P. Darby, James R. Kermode, and Gábor Csányi. Compressing local atomic neighbourhood descriptors. *npj Computational Materials*, 8(1):1–13, August 2022.

[28] James P Darby, Dávid P Kovács, Ilyes Batatia, Miguel A Caro, Gus LW Hart, Christoph Ortner, and Gábor Csányi. Tensor-reduced atomic density representations. *Physical Review Letters*, 131(2):028001, 2023.

[29] Maurice de Koning, A Antonelli, and Sidney Yip. Optimized free-energy evaluation using a single reversible-scaling simulation. *Physical review letters*, 83(20):3973, 1999.

[30] Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

[31] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B*, 99(1):014104, 2019.

[32] Ralf Drautz. Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys. Rev. B*, 102(2):024104, 2020.

[33] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. *SIAM J. Comput.*, 36(1):132, 2006.

[34] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-Error CUR Matrix Decompositions. *SIAM J. Matrix Anal. A.*, 30(2):844, 2008.

[35] Alexandre Dézaphie, Clovis Lapointe, Alexandra M. Goryaeva, Jérôme Creuze, and Mihai-Cosmin Marinica. Designing hybrid descriptors for improved machine learning models in atomistic materials science simulations. *Computational Materials Science*, 246:113459, January 2025.

[36] Richard S Ellis. Large deviations for a general class of random vectors. *The Annals of Probability*, 12(1):1–12, 1984.

[37] W. Fenchel. On Conjugate Convex Functions. *Canadian Journal of Mathematics*, 1(1):73–77, February 1949.

[38] Rodrigo Freitas, Mark Asta, and Maurice De Koning. Nonequilibrium free-energy calculation of solids using lammps. *Computational Materials Science*, 112:333–341, 2016.

[39] Daan Frenkel and Berend Smit. Understanding molecular simulations: from algorithms to applications. *Academic, San Diego*, 1996.

[40] Jürgen Gärtner. On large deviations from the invariant measure. *Theory of Probability & Its Applications*, 22(1):24–39, 1977.

[41] Aldo Glielmo, Peter Sollich, and Alessandro De Vita. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B*, 95:214302, Jun 2017.

[42] Aldo Glielmo, Claudio Zeni, and Alessandro De Vita. Efficient nonparametric n-body force fields from machine learning. *Phys. Rev. B*, 97(18):184307, May 2018.

[43] Carl P Goodrich, Ella M King, Samuel S Schoenholz, Ekin D Cubuk, and Michael P Brenner. Designing self-assembling kinetics with differentiable statistical physics models. *Proc. Natl. Acad. Sci. U.S.A.*, 118(10):e2024083118, 2021.

[44] Alexandra M. Goryaeva, Julien Dérès, Clovis Lapointe, Petr Grigorev, Thomas D. Swinburne, James R. Kermode, Lisa Ventelon, Jacopo Baima, and Mihai-Cosmin Marinica. Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W. *Phys. Rev. Materials*, 5:103803, Oct 2021.

[45] Alexandra M Goryaeva, Julien Dérès, Clovis Lapointe, Petr Grigorev, Thomas D Swinburne, James R Kermode, Lisa Ventelon, Jacopo Baima, and Mihai-Cosmin Marinica. Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc fe and w. *Physical Review Materials*, 5(10):103803, 2021.

[46] Alexandra M. Goryaeva, Jean Bernard Maillet, and Mihai Cosmin Marinica. Towards better efficiency of interatomic linear machine learning potentials. *Computational Materials Science*, 166:200–209, aug 2019.

[47] Alexander Goscinski, Félix Musil, Sergey Pozdnyakov, Jigyasa Nigam, and Michele Ceriotti. Optimal radial basis for density-based atomic representations. *The Journal of Chemical Physics*, 155(10), September 2021.

[48] Blazej Grabowski, Yuji Ikeda, Prashanth Srinivasan, Fritz Körmann, Christoph Freysoldt, Andrew Ian Duff, Alexander Shapeev, and Jörg Neugebauer. Ab initio vibrational free energies including anharmonicity for multicomponent alloys. *npj Computational Materials*, 5(1):1–6, 2019.

[49] Petr Grigorev, Alexandra M Goryaeva, Mihai-Cosmin Marinica, James R Kermode, and Thomas D Swinburne. Calculation of dislocation binding to helium-vacancy defects in tungsten using hybrid ab initio-machine learning methods. *Acta Mater.*, 247:118734, 2023.

[50] G Henkelman, G Johannesson, and H Jonsson. *Methods for finding saddle points and mimnum energy paths*. Springer, 2000.

[51] G Henkelman, B P Uberuaga, and H Jonsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 113(22):9901–9904, 2000.

[52] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[53] Hikaru Ibayashi, Taufeq Mohammed Razakh, Liqiu Yang, Thomas Linker, Marco Olguin, Shinnosuke Hattori, Ye Luo, Rajiv K. Kalia, Aiichiro Nakano, Ken-ichi Nomura, and Priya Vashishta. Allegro-legato: Scalable, fast, and robust neural-network quantum molecular dynamics via sharpness-aware minimization. In Abhinav Bhatele, Jeff Hammond, Marc Baboulin, and Carola Kruse, editors, *High Performance Computing*, pages 223–239, Cham, 2023. Springer Nature Switzerland.

[54] Giulio Imbalzano, Yongbin Zhuang, Venkat Kapil, Kevin Rossi, Edgar A Engel, Federico Grasselli, and Michele Ceriotti. Uncertainty estimation for molecular dynamics and sampling. *The Journal of Chemical Physics*, 154(7), 2021.

[55] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690–2693, Apr 1997.

[56] John G Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3(5):300–313, 1935.

[57] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, 1996.

[58] Georg Kresse and D Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, 59(3):1758, 1999.

[59] Tony Lelièvre, Gabriel Stoltz, and Mathias Rousset. *Free energy computations: a mathematical perspective*. World Scientific, 2010.

[60] Samuel Livingstone, Michael F Faulkner, and Gareth O Roberts. Kinetic energy choice in hamiltonian/hybrid monte carlo. *Biometrika*, 106(2):303–319, 2019.

[61] Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. In *International Conference on Machine Learning*, pages 14223–14247. PMLR, 2022.

[62] Yury Lysogorskiy, Cas van der Oord, Anton Bochkarev, Sarath Menon, Matteo Rinaldi, Thomas Hammerschmidt, Matous Mrovec, Aidan Thompson, Gábor Csányi, Christoph Ortner, et al. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Computational Materials*, 7(1):1–12, 2021.

[63] Pui-Wai Ma, SL Dudarev, and Jan S Wróbel. Dynamic simulation of structural phase transitions in magnetic iron. *Physical Review B*, 96(9):094418, 2017.

[64] Ivan Maliyov, Petr Grigorev, and Thomas D Swinburne. Exploring parameter dependence of atomic minima with implicit differentiation. *npj Computational Materials*, 11(1):22, 2025.

[65] Sarath Menon, Yury Lysogorskiy, Alexander LM Knoll, Niklas Leimeroth, Marvin Poul, Minaam Qamar, Jan Janssen, Matous Mrovec, Jochen Rohrer, Karsten Albe, et al. From electrons to phase diagrams with machine learning potentials using pyiron based automated workflows. *npj Computational Materials*, 10(1):261, 2024.

[66] M. Methfessel and A. T. Paxton. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B*, 40:3616–3621, Aug 1989.

[67] Hendrik J. Monkhorst and James D. Pack. Special points for brillouin-zone integrations. *Phys. Rev. B*, 13:5188–5192, Jun 1976.

[68] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.*, 14(1):579, 2023.

[69] Félix Musil, Michael J Willatt, Mikhail A Langovoy, and Michele Ceriotti. Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.*, 15(2):906–915, 2019.

[70] Juno Nam and Rafael Gomez-Bombarelli, 2024.

[71] Ngoc-Cuong Nguyen. Fast proper orthogonal descriptors for many-body interatomic potentials. *Phys. Rev. B*, 107:144103, Apr 2023.

[72] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[73] Livia B Pártay, Gábor Csányi, and Noam Bernstein. Nested sampling for materials. *The European Physical Journal B*, 94(8):159, 2021.

[74] Grigorios A Pavliotis and Andrew Stuart. *Multiscale methods: averaging and homogenization*, volume 53. Springer Science & Business Media, 2008.

[75] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[76] Danny Perez, Aparna P. A. Subramanyam, Ivan Maliyov, and Thomas D. Swinburne. Uncertainty quantification for misspecified machine learned interatomic potentials. *npj Computational Materials*, 11(1):263, August 2025.

[77] Mariia Radova, Wojciech G. Stark, Connor S. Allen, Reinhard J. Maurer, and Albert P. Bartók, 2025.

[78] JM Rickman and R LeSar. Free-energy calculations in materials research. *Annu. Rev. Mater. Res.*, 32:195, 2002.

[79] A. Rohskopf, C. Sievers, N. Lubbers, M.a. Cusentino, J. Goff, J. Janssen, M. McCarthy, D. Montes Oca de Zapiain, S. Nikolov, K. Sargsyan, D. Sema, E. Sikorski, L. Williams, A.p. Thompson, and M.a. Wood. Fitsnap: Atomistic machine learning with lammps. *Journal of Open Source Software*, 8(84):5118, 2023.

[80] A. Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Sim.*, 14(3):1153–1173, 2016.

[81] Samantha Sherman and Tamara G Kolda. Estimating higher-order moments using symmetric tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1369–1387, 2020.

[82] T.D. Swinburne, C. Lapointe, and M.-C. Marinica. Supplementary material, 2025.

[83] Thomas Swinburne and Danny Perez. Parameter uncertainties for imperfect surrogate models in the low-noise regime. *Machine Learning: Science and Technology*, 2025.

[84] Thomas D. Swinburne. Coarse-graining and forecasting atomic material simulations with descriptors. *Phys. Rev. Lett.*, 131:236101, Dec 2023.

[85] Thomas D. Swinburne and Mihai-Cosmin Marinica. Unsupervised calculation of free energy barriers in large crystalline systems. *Phys. Rev. Lett.*, 120(13):135503, 2018.

[86] Thomas D Swinburne and Mihai-Cosmin Marinica. `PAFI` code, 2023.

[87] Samuel Tamagnone, Alessandro Laio, and Marylou Gabrié. Coarse-grained molecular dynamics with normalizing flows. *Journal of Chemical Theory and Computation*, 20(18):7796–7805, 2024.

[88] Stephan Thaler, Maximilian Stupp, and Julija Zavadlav. Deep coarse-grained potentials via relative entropy minimization. *J. Chem. Phys.*, 157(24), 2022.

[89] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022.

[90] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comp. Phys.*, 285:316, 2015.

[91] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *Npj Comput. Mater.*, 6(1):1, 2020.

[92] Joshua A. Vita and Daniel Schwalbe-Koda. Data efficiency and extrapolation trends in neural network interatomic potentials. *Machine Learning: Science and Technology*, 4(3):035031, August 2023.

[93] Udo Von Toussaint. Bayesian inference in physics. *Reviews of Modern Physics*, 83(3):943–999, 2011.

[94] Fugao Wang and David P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050, 2001.

[95] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Atom-density representations for machine learning. *The Journal of Chemical Physics*, 150(15):154110, April 2019.

[96] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[97] Peter Wirnsberger, George Papamakarios, Borja Ibarz, Sébastien Racaniere, Andrew J Ballard, Alexander Pritzel, and Charles Blundell. Normalizing flows for atomic solids. *Machine Learning: Science and Technology*, 3(2):025009, 2022.

[98] Roderick Wong. *Asymptotic approximations of integrals*. SIAM, 2001.

[99] Yu Xie, Jonathan Vandermause, Lixin Sun, Andrea Cepellotti, and Boris Kozinsky. Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene. *Npj Comput. Mater.*, 7(1):1, 2021.

[100] Anruo Zhong, Clovis Lapointe, Alexandra M. Goryaeva, Jacopo Baima, Manuel Athènes, and Mihai-Cosmin Marinica. Anharmonic thermo-elasticity of tungsten from accelerated bayesian adaptive biasing force calculations with data-driven force fields. *Phys. Rev. Mater.*, 7:023802, Feb 2023.

[101] Albert Zhu, Simon Batzner, Albert Musaelian, and Boris Kozinsky. Fast uncertainty estimates in deep learning interatomic potentials. *The Journal of Chemical Physics*, 158(16):164111, 04 2023.

[102] Li-Fang Zhu, Blazej Grabowski, and Jörg Neugebauer. Efficient approach to compute melting properties fully from ab initio with application to cu. *Physical Review B*, 96(22):224202, 2017.

[103] Li-Fang Zhu, Fritz Körmann, Qing Chen, Malin Selleby, Jörg Neugebauer, and Blazej Grabowski. Accelerating ab initio melting property calculations with machine learning: application to the high entropy alloy tavcrw. *npj Computational Materials*, 10(1):274, 2024.

[104] R. W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8):1420, 1954.

A Thermodynamic sampling of atomic crystal models

This section reviews standard results from classical statistical mechanics for a system of N atoms with specie $\mathbf{s} = [\mathrm{s}_1, \ldots, \mathrm{s}_N] \in \mathbb{Z}^N$, positions $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N\times 3}$ and momenta $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N] \in \mathbb{R}^{N\times 3}$. Atoms are confined to a periodic supercell $\mathbf{C} \in \mathbb{R}^{3\times 3}$ with volume $V = |\mathbf{C}|$ (the determinant), such that scaled positions lie on the unit torus, i.e. $\mathbf{X}\mathbf{C}^{-1} \in \mathbb{T}^{N\times 3}$. In anticipation of later results where we take the limit $N \to \infty$, we write the total energy $U(\mathbf{X}, \mathbf{P})$ as the sum of a potential and kinetic energy, i.e.

$$U(\mathbf{X}, \mathbf{P}) \equiv E_{\mathbf{w}}(\mathbf{X}) + K(\mathbf{P}), \tag{7}$$

where $K(\mathbf{P}) = \sum_{i=1}^{N} \mathrm{p}_i^2/(2m_i)$ and dependence on $\mathbf{s}$ is contained in the potential energy function $E_{\mathbf{w}}(\mathbf{X})$ by model parameters $\mathbf{w}$, the focus of this paper. To express the supercell in an intensive form we define the supercell per atom $\mathbf{C}$ through $\mathbf{C} = \mathbf{N}\mathbf{C}$, where $\mathbf{N} = \mathrm{Diag}(N_x, N_y, N_z)$, such that $|\mathbf{N}| = N$ and the volume per atom is given by $|\mathbf{C}|$. The canonical (NVT) partition function at $\mathrm{T} = 1/(\mathrm{k_B}\beta)$ then writes

$$Z_{\mathbf{w}}^N(\beta, \mathbf{C}) \equiv \lambda_0(\beta)^{-3N} \int_{\mathbb{R}^{3N}} \exp[-\beta E_{\mathbf{w}}(\mathbf{X})] \mathrm{d}\mathbf{X}, \tag{8}$$

where $\lambda_0(\beta) = \hbar\sqrt{2\pi\beta/m}$ is the thermal De Broglie wavelength[39] and $m^N \equiv \prod_{i=1}^{N} m_i$. The NVT free energy per atom is defined in the thermodynamic limit $N \to \infty$:

$$\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}) \equiv \lim_{N\to\infty} \frac{-1}{\beta N} \ln |Z_{\mathbf{w}}^N(\beta, \mathbf{C})|. \tag{9}$$

In practice, the integral over atomic configuration space in (8) is dominated by contributions from some set of *phases* $\mathcal{P} = \{\text{BCC, FCC, hcp, liquid, } \ldots\}$, such that

$$Z_{\mathbf{w}}^N(\beta, \mathbf{C}) = \sum_{p\in\mathcal{P}} Z_{\mathbf{w}}^N(\beta, \mathbf{C}, p), \tag{10}$$

where each term $Z_{\mathbf{w}}^N(\beta, \mathbf{C}, p)$, is an integral over (disjoint) partitions of configuration space, with corresponding phase free energy $\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p)$, defined as in (9). It is simple to show that as $N \to \infty$ the NVT free energy is dominated by a single phase

$$p_{\mathbf{w}}^*(\beta, \mathbf{C}) = \arg\min_{p\in\mathcal{P}} \mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p), \tag{11}$$

as $\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}) = \min_{p\in\mathcal{P}} \mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p)$. Similarly, the NPT free energy of a phase p is obtained by minimizing $\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p)$ at under some constant external stress $\boldsymbol{\sigma}$ (i.e. isotropic pressure $\boldsymbol{\sigma} = (P/3)\mathbb{I}_3$), giving

$$\mathcal{G}_{\mathbf{w}}(\beta, \boldsymbol{\sigma}, p) \equiv \min_{\mathbf{C}} \mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p) - \mathrm{Tr}(\boldsymbol{\sigma}^\top \mathbf{C}), \tag{12}$$

It is clear that estimation of $\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p)$ for general $\beta, \mathbf{C}$ is sufficient to estimate $\mathcal{G}_{\mathbf{w}}(\beta, \boldsymbol{\sigma}, p)$, giving the stable phase at some temperature and pressure as

$$p_{\mathbf{w}}^*(\beta, \boldsymbol{\sigma}) = \arg\min_{p\in\mathcal{P}} \mathcal{G}_{\mathbf{w}}(\beta, \boldsymbol{\sigma}, p), \tag{13}$$

where the $\mathbf{w}$ subscript emphasizes the dependence of p^* on the parameters of the interatomic potential.

In this paper, we will focus on the set of crystalline phases $\mathcal{P}_s \subset \mathcal{P}$, whose configuration space is defined as the set of (potentially large) vibrations around some lattice structure $\mathbf{X}_p^0$, $p \in \mathcal{P}_s$.

A.1 Thermodynamic integration

Accurate calculation of phase stability requires converging per-atom free energy differences between phases to within a few meV/atom at any given temperature and pressure to allow determination of (13). Accurate determination of phase transitions, where free energy differences are formally zero[102, 48, 100, 103, 65], thus requires tight convergence of any estimator. The stringent accuracy requirement has led to the development of sampling techniques to reduce the number of samples required for convergence [56, 39, 78]. In all cases, the starting point is some atomic energy function

$E_0(\mathbf{X})$, whose corresponding phase free energy $\mathcal{F}_0(\beta, \mathbf{C}, p)$ is known either through tabulation, or analytically if $E_0(\mathbf{X})$ is harmonic[56]. We can thus define $\Delta E_{\mathbf{w}}(\mathbf{X}) = E_{\mathbf{w}}(\mathbf{X}) - E_0(\mathbf{X})$ as the energy difference (per-atom) between the target and reference systems, with a free energy difference $\Delta\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p)$. Thermodynamic integration (TI) is a stratified sampling scheme over $E_\eta(\mathbf{X}) = E_0(\mathbf{X}) + \eta\Delta E_{\mathbf{w}}(\mathbf{X})$ for $\eta \in [0, 1]$. Denoting equilibrium averages by $\langle \dots \rangle_\eta$, we obtain

$$\Delta\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p) = \int_0^1 \langle \Delta E_{\mathbf{w}}(\mathbf{X}) \rangle_\eta \, \mathrm{d}\eta. \tag{14}$$

Sampling efficiency often requires constraint functions or resetting to prevent trajectories escaping the metastable basin of a given phase, as discussed in section F.5. In general, the larger the value of $\Delta\mathcal{F}_{\mathbf{w}}$, the finer the integration scheme over η and the more samples will be required for convergence [59, 25].

A.2 Free energy perturbation

Typically used as a complement to thermodynamic integration, if the difference $N\Delta E_{\mathbf{w}}(\mathbf{X})$ is as small as $10/\beta$, corresponding to at most 10 meV/atom at 1000 K for solid state systems ($N \simeq 100$), we can also use free energy perturbation (FEP) to estimate the free energy difference[59, 48, 20]. Using the definition of the free energy $\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C})$ and $\langle \dots \rangle_\eta$ at $\eta = 0$, it is simple to show that

$$\Delta\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p) = -(1/N\beta) \ln\langle \exp[-N\beta\Delta E_{\mathbf{w}}(\mathbf{X})] \rangle_0.$$

In practice, the logarithmic expectation is expressed as a cumulant expansion[104, 59, 21] for increased numerical stability, writing

$$\Delta\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p) = \langle \Delta E_{\mathbf{w}} \rangle_0 - \frac{N\beta}{2} \langle (\delta\Delta E_{\mathbf{w}})^2 \rangle_0 + \dots, \tag{15}$$

where $\delta\Delta E_{\mathbf{w}} = \Delta E_{\mathbf{w}} - \langle \Delta E_{\mathbf{w}} \rangle_0$, and the expansion continues, in principle, to all orders. While (15) gives an expression for the free energy in terms of samples generated solely with a reference potential, in a practical setting we require free energy differences to be very small to allow for convergence. Equation (15) can be shown to be an upper bound to the estimated free energy difference[20] and as such can be used as a convergence measure for a well-chosen reference potential. In this setting, we typically have $\Delta\mathcal{F}_{\mathbf{w}} < 1$ meV/atom (table 1 in the main text).

A.3 Adiabatic Switching

In addition to the above methods which employ equilibrium averages, the adiabatic switching[1, 29, 38, 65] method estimates free energy differences using the well-known Jarzynski equality [55]. The adiabatic switching equality can be written[38]

$$\Delta\mathcal{F}_{\mathbf{w}}(\beta, \mathbf{C}, p) = \frac{1}{2}[\langle W^{\text{irr}} \rangle_{0\to1} - \langle W^{\text{irr}} \rangle_{1\to0}], \tag{16}$$

where W^{irr} is the irreversible work along a thermodynamic path (in the above η is implied, though it is also possible to use the temperature) and $\langle \dots \rangle_{0\to1}$ indicates an ensemble average of around $10 - 30$ simulations. The key quantity is the so-called 'switching time', i.e. the rate at which the thermodynamic path is traversed. For solid-state free energies one typically progresses along the path in $\mathcal{O}(10)$ increments of $\mathcal{O}(10 - 100)$ ps[65], thus requiring around 10^{7-8} force calls per temperature. In this setting, we can target similar free energy differences to thermodynamic integration, i.e. $\mathcal{O}(100)$ meV/atom at 1000 K. The computational costs of the above methods and the present D-DOS approach is discussed in section F, and summarized in table 1 in the main text.

B Conditional descriptor density of states

Our central strategy to control the V^N divergence of $\Omega(\boldsymbol{\mathcal{D}})$, is to introduce the *conditional* descriptor density of states (CD-DOS)

$$\Omega(\boldsymbol{\mathcal{D}}|\alpha) \equiv \int_{\mathbb{R}^{3N}} \frac{\delta\left(\hat{\alpha}(\mathbf{X}) - \alpha\right)}{\Omega(\alpha)} \delta(\hat{\boldsymbol{\mathcal{D}}}(\mathbf{X}) - \boldsymbol{\mathcal{D}}) \mathrm{d}\mathbf{X}, \tag{17}$$

where $\hat{\alpha}(\mathbf{X})$ is the dimensionless isosurface function

$$\hat{\alpha}(\mathbf{X}) \equiv \ln\left|E_0(\mathbf{X})/(N\mathrm{U}_0)\right|, \tag{18}$$

U_0 is a user-defined energy scale and $E_0(\mathbf{X}) \geq 0$ is some reference potential energy. In section B.2 we detail how equation (18) can be generalized to a momentum-dependent $\hat{\alpha}(\mathbf{X}, \mathbf{P})$. In either case, $E_0(\mathbf{X})$ is chosen such that we can calculate, numerically or analytically, the isosurface volume

$$\Omega(\alpha) \equiv \int_{\mathbb{R}^{3N}} \delta\left(\hat{\alpha}(\mathbf{X}) - \alpha\right) \mathrm{d}\mathbf{X}, \tag{19}$$

which contains the exponential divergence as $\int_{\mathbb{R}} \Omega(\alpha)\mathrm{d}\alpha = V^N$. The crucial advantage of the conditional form is that $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ is normalized by construction, $\int_{\mathbb{R}^D} \Omega(\boldsymbol{\mathcal{D}}|\alpha)\mathrm{d}\boldsymbol{\mathcal{D}} = 1$, showing that $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ is the probability density of $\boldsymbol{\mathcal{D}}$ on the isosurface $\hat{\alpha}(\mathbf{X}) = \alpha$ and allowing us to employ density estimation techniques such as score matching[52]. The full D-DOS $\Omega(\boldsymbol{\mathcal{D}})$ is then recovered through integration against α, $\Omega(\boldsymbol{\mathcal{D}}) = \int_{\mathbb{R}} \Omega(\boldsymbol{\mathcal{D}}|\alpha)\Omega(\alpha)\mathrm{d}\alpha$, emphasizing that our goal is to decompose the high-dimensional configuration space into a foliation of isosurfaces $\hat{\alpha}(\mathbf{X}) = \alpha$ where we expect $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ to be tractable for density estimation. The generalization of $\hat{\alpha}(\mathbf{X})$ to include momentum dependence is discussed in section B.2 and general considerations for designing optimal $\hat{\alpha}(\mathbf{X})$ are discussed in section B.3.

B.1 The isosurface and descriptor entropies

The free energy $\mathcal{F}_{\mathbf{w}}(\beta)$, equation (2), is proportional to the logarithm of the partition function $Z_{\mathbf{w}}(\beta)$, i.e. $\mathcal{F}_{\mathbf{w}}(\beta) = (-1/N\beta)\ln|Z_{\mathbf{w}}(\beta)|$. It is thus natural to define *entropies* of the isosurface volume $\Omega(\alpha)$ and CD-DOS $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$. We first define the intensive isosurface entropy

$$\mathcal{S}_0(\alpha) \equiv \lim_{N\to\infty} \frac{1}{N} \ln|\Omega(\alpha)/V_0^N|. \tag{20}$$

The term V_0 ensures $\mathcal{S}_0(\alpha)$ is dimensionless; with $\hat{\alpha}(\mathbf{X})$ we have $V_0 = \lambda_\beta^3$, while with a momentum-dependent $\hat{\alpha}(\mathbf{X}, \mathbf{P})$, discussed in B.2, we have $V_0 = h^3$. It is clear that $\mathcal{S}_0(\alpha)$ is a measure of the configurational entropy per atom of N independent atoms confined to the isosurface. The CD-DOS $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$, equation (17), has a natural entropy definition, the intensive log density

$$\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) \equiv \lim_{N\to\infty} \frac{1}{N} \ln \Omega(\boldsymbol{\mathcal{D}}|\alpha). \tag{21}$$

The CD-DOS entropy $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ measures the proportion of the isosurface phase space volume that has a global descriptor vector $\boldsymbol{\mathcal{D}}$, meaning descriptor values with larger $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ are more likely to be observed under unbiased isosurface sampling. Furthermore, $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ has two properties which greatly facilitate free energy estimation: $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ is intensive (N-independent) for local descriptor functionsA and as $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ is normalized, application of Laplace's method (see C.1) fixes the maximum of $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$:

$$\lim_{N\to\infty} \frac{1}{N} \ln\left|\int \Omega(\boldsymbol{\mathcal{D}}|\alpha)\mathrm{d}\boldsymbol{\mathcal{D}}\right| = \max_{\boldsymbol{\mathcal{D}}\in\mathbb{R}^D} \mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) = 0. \tag{22}$$

This condition is crucial, allowing us to integrate the score $\nabla_{\boldsymbol{\mathcal{D}}}\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ and produce free energy estimates.

B.2 Forms of the isosurface function

As discussed above, free energy estimation will require access to $\mathcal{S}_0(\alpha)$ and a means to generate samples on the isosurface $\hat{\alpha}(\mathbf{X}) = \alpha$. For harmonic reference potentials $S_0(\alpha)$ is given analytically; the isosurface function writes

$$\hat{\alpha}(\mathbf{X}) \equiv \ln\left|[\mathbf{X} - \mathbf{X}_0]^\top \mathbf{H} [\mathbf{X} - \mathbf{X}_0]/(2N\mathrm{U}_0)\right|, \tag{23}$$

where the Hessian $\mathbf{H}$ has $3N$-3 positive eigenmodes and $\mathbf{X}_0$ is the lattice structure. Sampling $\hat{\alpha}(\mathbf{X}) = \alpha$ reduces to generating random unit vectors in $\mathbb{R}^{3N-3}$, while the isosurface entropy (20) reads

$$\mathcal{S}_0(\alpha) \equiv S_0 + 3\alpha/2, \quad V_0 = \lambda_0^3(\beta). \tag{24}$$

The constant S_0 is given by $S_0 = 3/2 + 3/2 \ln |2\beta \mathrm{U}_0/3| - \beta \mathcal{F}_0(\beta)$, where $\mathcal{F}_0(\beta)$ is the familiar free energy per-atom of harmonic atomic systems (see D). We can generalize (18) to arbitrary $E_0(\mathbf{X})$ with the kinetic energy $K(\mathbf{P}) = \sum_{i=1}^{N} \mathrm{p}_i^2/(2m_i)$ such that

$$\hat{\alpha}(\mathbf{X}, \mathbf{P}) \equiv \ln \left| [K(\mathbf{P}) + E_0(\mathbf{X})]/(N\mathrm{U}_0) \right|. \tag{25}$$

Isosurface sampling then reduces to running microcanonical (NVE) dynamics, in close connection with Hamiltonian Monte Carlo[15, 60]. The NVT free energy $\mathcal{F}_0(\beta)$ of the reference system can be expressed as $\mathcal{F}_0(\beta) = \mathcal{U}_0(\beta) - \mathcal{S}_0(\alpha_\beta)/\beta$, where $\mathcal{U}_0(\beta)$ is the internal energy per atom and $\alpha_\beta \equiv \ln |\mathcal{U}_0(\beta)/\mathrm{U}_0|$ (see D). With (25) the isosurface entropy (20) is then the difference between the reference free and internal energies:

$$\mathcal{S}_0(\alpha) = \beta_\alpha \left[\mathcal{U}_0(\beta_\alpha) - \mathcal{F}_0(\beta_\alpha) \right], \quad V_0(\alpha) = h^3, \tag{26}$$

where β_α is defined through $\mathcal{U}_0(\beta_\alpha) \equiv \mathrm{U}_0 \exp(\alpha)$, which will have a unique solution when $\mathcal{U}_0(\beta_\alpha)$ is monotonic. In practice, $\mathcal{U}_0(\beta)$ and $\mathcal{F}_0(\beta)$ are estimated via thermodynamic sampling (see A) over a range of β, interpolating with $\alpha \equiv \ln |\mathcal{U}_0(\beta)/\mathrm{U}_0|$ to estimate $\mathcal{S}_0(\alpha)$. The final modification is to augment the descriptor vector $\boldsymbol{\mathcal{D}}$, concatenating an intensive kinetic energy $D_K = K(\mathbf{P})/N$ (see D)

$$\boldsymbol{\mathcal{D}} \to \boldsymbol{\mathcal{D}} \oplus D_K, \quad \mathbf{w} \to \mathbf{w} \oplus 1, \tag{27}$$

meaning $\mathbf{w} \cdot \boldsymbol{\mathcal{D}}$ now returns the total energy rather than the potential energy. Sampling schemes can thus use $\hat{\alpha}(\mathbf{X})$ with a harmonic reference potential, where $\mathcal{S}_0(\alpha)$ is given analytically, or $\hat{\alpha}(\mathbf{X}, \mathbf{P})$ with any reference potential, where $\mathcal{S}_0(\alpha)$ determined via thermodynamic sampling. All theoretical results below can use either $\mathcal{S}_0(\alpha)$; use of both are demonstrated for solid phases in section **??**. A forthcoming study will apply the momentum-dependent formalism to liquids and melting transitions.

B.3 Criteria for optimal isosurface functions

Estimation of $\mathcal{F}_{\mathbf{w}}(\beta)$ via (4) relies on our ability to accurately approximate $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ by some score-matched estimator $\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$. Strong curvature of $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ in $\boldsymbol{\mathcal{D}}$ and α is crucial for the statistical efficiency of score matching and applicability of Laplace's method. A poor choice of isosurface function $\hat{\alpha}$ will give weaker curvature, as distributions will vary less between isosurfaces, thus amplifying the consequences of any sampling error. A learnable $\hat{\alpha}_{\boldsymbol{\phi}}(\mathbf{X})$ or $\hat{\alpha}_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{P})$ should tune parameters $\boldsymbol{\phi}$ to maximize curvatures in $\mathcal{S}_{\Theta;\boldsymbol{\phi}}(\boldsymbol{\mathcal{D}}|\alpha)$, a direction we leave for future work.

B.4 Free energy evaluation with Laplace's method

Laplace's method, or steepest descents[98], is a common technique for evaluating the limits of integrals (see D). With the definition of $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$, equation (21), we use Laplace's method to evaluate a conditional free energy $\mathcal{F}_{\mathbf{w}}(\beta|\alpha)$, defined on $\hat{\alpha}(\mathbf{X}) = \alpha$:

$$\begin{aligned} \mathcal{F}_{\mathbf{w}}(\beta|\alpha) &\equiv \lim_{N\to\infty} \frac{-1}{N\beta} \ln \left| \int_{\mathbb{R}^D} e^{-N\beta \mathbf{w}\cdot\boldsymbol{\mathcal{D}}} \Omega(\boldsymbol{\mathcal{D}}|\alpha) \mathrm{d}\boldsymbol{\mathcal{D}} \right|, \\ &= \min_{\boldsymbol{\mathcal{D}}\in\mathbb{R}^D} \left(\mathbf{w}\cdot\boldsymbol{\mathcal{D}} - \mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)/\beta \right). \end{aligned} \tag{28}$$

It is clear that $-\beta\mathcal{F}_{\mathbf{w}}(\beta|\alpha)$ is both the Legendre–Fenchel [37] conjugate of the entropy $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ and has a close connection to the cumulant expansion in free energy perturbation [21] a point we discuss further in section B.6 and A. We thus obtain a final free energy expression, again using Laplace's method

$$\begin{aligned} \mathcal{F}_{\mathbf{w}}(\beta) &\equiv \lim_{N\to\infty} \frac{-1}{N\beta} \ln \int_{\mathbb{R}} e^{N[\mathcal{S}_0(\alpha) - \beta\mathcal{F}_{\mathbf{w}}(\beta|\alpha)]} \mathrm{d}\alpha, \\ &= \min_{\alpha\in\mathbb{R}} \left(\mathcal{F}_{\mathbf{w}}(\beta|\alpha) - \mathcal{S}_0(\alpha)/\beta \right), \\ &= \min_{\alpha,\boldsymbol{\mathcal{D}}} \left(\mathbf{w}\cdot\boldsymbol{\mathcal{D}} - [\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) + \mathcal{S}_0(\alpha)]/\beta \right). \end{aligned} \tag{29}$$

Equation (29) is our main result, an integration-free expression for the NVT free energy for generalized linear MLIPs (1). The minimization over α and $\boldsymbol{\mathcal{D}}$ requires

$$\nabla_{\boldsymbol{\mathcal{D}}}\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) = \beta\mathbf{w}, \quad \partial_\alpha\mathcal{F}_{\mathbf{w}}(\beta|\alpha) = \partial_\alpha\mathcal{S}_0(\alpha) \tag{30}$$

which emphasizes the Legendre duality between $\beta\mathbf{w}$ and $\boldsymbol{\mathcal{D}}$. In A we show use of a harmonic reference energy (23) gives $\mathcal{S}_0(\alpha) = S_0 + 3\alpha/2$, meaning $\partial_\alpha\mathcal{S}_0(\alpha) = 3/2$. We also recover familiar results for harmonic models, setting $\mathbf{w}\cdot\boldsymbol{\mathcal{D}} = E_0(\mathbf{X})$; in this case (30) reduces to the equipartition relation $\beta\langle E_0\rangle = 3/2$.

B.5 Gradients of the free energy

The gradient of $\mathcal{F}_{\mathbf{w}}$ with respect to $\mathbf{w}$ allows the inclusion of finite temperature properties in objective functions for inverse design, a unique feature we explore below. With minimizing values $\alpha^*_{\beta,\mathbf{w}}$, $\boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*}$, the $\mathbf{w}$-gradient is simply

$$\nabla_{\mathbf{w}}\mathcal{F}_{\mathbf{w}}(\beta) = \boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*} \in \mathbb{R}^D. \tag{31}$$

The internal energy $\mathcal{U}_{\mathbf{w}}(\beta)$ is also a simple expression involving the minimizing vector $\boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*}$; with $\hat{\alpha}(\mathbf{X})$, equation (23), we have $\mathcal{U}_{\mathbf{w}}(\beta) = 3/(2\beta) + \mathbf{w}\cdot\boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*}$. With the momentum-dependent isosurface $\hat{\alpha}(\mathbf{X},\mathbf{P})$, equation (25), we have $\mathcal{U}_{\mathbf{w}}(\beta) = \mathbf{w}\cdot\boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*}$. Evaluation of higher order gradients requires implicit derivatives[17, 64], e.g. $\partial_\alpha\boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*} \in \mathbb{R}^D$, $\partial_\beta\alpha^* \in \mathbb{R}$ or $\partial_{\mathbf{w}}[\boldsymbol{\mathcal{D}}^*_{\beta,\mathbf{w},\alpha^*}]^\top \in \mathbb{R}^{D\times D}$. Further exploration of finite temperature properties such as thermal expansion will be the focus of future work.

B.6 Connection to free energy perturbation

The conditional free energy $\mathcal{F}_{\mathbf{w}}(\beta|\alpha)$ can be given by a cumulant expansion, using $\langle\ldots\rangle_\alpha$ for isosurface averages

$$\mathcal{F}_{\mathbf{w}}(\beta|\alpha) = \langle\mathbf{w}\cdot\boldsymbol{\mathcal{D}}\rangle_\alpha + \frac{N\beta}{2}\langle(\mathbf{w}\cdot\delta\boldsymbol{\mathcal{D}})^2\rangle_\alpha + \ldots \tag{32}$$

where $\delta\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{D}} - \langle\boldsymbol{\mathcal{D}}\rangle$. The factor of N to ensures intensivity of the covariance (see E). Free energy perturbation (FEP)[104, 59, 21, 82] also expresses the free energy difference as a cumulant expansion over canonical averages with $E_0(\mathbf{X})$. As discussed in B.2, as $N\to\infty$ canonical sampling at β is equivalent to isosurface sampling at $\alpha = \alpha_\beta$, where the relation between β and α_β depends on the form of the isosurface function $\hat{\alpha}(\mathbf{X})$ or $\hat{\alpha}(\mathbf{X},\mathbf{P})$, e.g. (23) or (25). The FEP estimate is thus equivalent to a D-DOS estimate where we *fix* $\alpha = \alpha_\beta$, instead of minimizing over α as in equation (29). When the free energy difference is very small, i.e. the target is very similar to the reference, $\hat{\alpha} = \alpha_\beta$ may be a good approximate minimization. However, in the general case it is clear the D-DOS estimate can strongly differ from FEP estimates. This is evidenced later in Figures **??** b) and **??**c), where the minimizing α value at constant β varies strongly with $\mathbf{w}$, even at relatively low homologous temperatures (1000 K in W, around 1/4 of the melting temperature), while FEP would predict α to be constant with β.

C Numerical implementation of score-matching sampling

C.1 Derivation of the score matching loss

Our starting point is the definition of the score matching loss, the Fisher divergence[52]

$$L(\boldsymbol{\Theta}|\alpha) \equiv \frac{N}{2}\langle\|\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) - \nabla\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)\|^2\rangle_\alpha - L_0 \tag{33}$$

$$= \frac{N}{2}\langle\|\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\|^2\rangle_\alpha - N\langle\nabla\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)\cdot\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\rangle_\alpha, \tag{34}$$

where L_0 is a $\boldsymbol{\Theta}$-independent constant and we set $L(\boldsymbol{\Theta}|\alpha)$ to be $\mathcal{O}(N)$ by convention. As isosurface averages $\hat{\alpha}(\mathbf{X}) = \alpha$ can clearly be written an integral over the D-DOS $\Omega(\boldsymbol{\mathcal{D}}|\alpha) = \exp[N\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)]$, we use integration by parts to write

$$L(\boldsymbol{\Theta}|\alpha) \equiv \frac{N}{2}\langle\|\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\|^2\rangle_\alpha - N\int\nabla\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)\cdot\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\exp[N\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)]\mathrm{d}\boldsymbol{\mathcal{D}}, \tag{35}$$

$$\equiv \frac{N}{2}\langle\|\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\|^2\rangle_\alpha + \langle\nabla\cdot\nabla\mathcal{S}_{\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)\rangle_\alpha, \tag{36}$$

as given in the main text.

C.2 Analysis in the limit $N \to \infty$

Whilst the descriptor entropy $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ is intensive, as shown above, averages of $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ or gradients over α will in general give rise to terms inversely proportional to N. To correctly infer the solution in the limit $N \to \infty$, we make the multiscale hierarchy[74]

$$L(\boldsymbol{\Theta}|\alpha) = \sum_s N^{1-s}L_s(\boldsymbol{\Theta}|\alpha) \tag{37}$$

To solve this hierarchy we can in principle define a multiscale solution $\boldsymbol{\Theta}^{(S)}$ such that

$$\nabla L(\boldsymbol{\Theta}^{(S)}|\alpha) = \mathbf{0} + \mathcal{O}(N^{-S}). \tag{38}$$

We can find a solution $\boldsymbol{\Theta}^{(S)}$ in a recursive fashion, first minimizing $L_0(\boldsymbol{\Theta}|\alpha)$ to give $\boldsymbol{\Theta}^{(0)}$, then minimizing $L_1(\boldsymbol{\Theta}|\alpha)$ under the constraint $\nabla L_0(\boldsymbol{\Theta}|\alpha) = 0$ to give $\boldsymbol{\Theta}^{(1)}$, and so on. However, we only consider models where $L(\boldsymbol{\Theta}|\alpha)$ is linear in $\boldsymbol{\Theta}$, meaning the loss gradient can be decomposed as

$$\sum_s N^{1-s}\left[\mathbf{A}_s\boldsymbol{\Theta} - \mathbf{b}_s\right] = \mathbf{0}. \tag{39}$$

Respecting the multi-scale hierarchy then equates to ensuring $\boldsymbol{\Theta}^{(S+1)} - \boldsymbol{\Theta}^{(S)}$ is in the null space of all $\mathbf{A}_s$, $s \leq S$. Whilst we investigated solving each term in this hierarchy independently, in practice this had negligable improvement over simply minmizing the score matching loss as the linear solve will naturally ensure the solution respects the multi-scale hierarchy to a within numerical tolerance.

C.3 Low-rank compressed score models

We require a low-rank model to efficiently estimate and store any score model, which should also allow efficient minimization for free energy estimation via (29). We use a common tensor compression approach[81] to produce a low-rank model for estimation of higher order moments.

Using $\langle\dots\rangle_\alpha$ to denote isosurface averages, we first estimate the isosurface mean $\hat{\boldsymbol{\mu}}_\alpha = \langle\boldsymbol{\mathcal{D}}\rangle_\alpha$ and intensive covariance $\hat{\boldsymbol{\Sigma}}_\alpha = N\langle\delta\boldsymbol{\mathcal{D}}\delta\boldsymbol{\mathcal{D}}^\top\rangle_\alpha$, where $\delta\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{D}} - \boldsymbol{\mu}_\alpha$, a symmetric matrix which has D orthonormal eigenvectors $\mathbf{v}_{\alpha,l}$, $l \in [1, D]$. Our low-rank score model uses F scalar functions $\mathbf{f}(x) = [f_1(x), \dots, f_F(x)] \in \mathbb{R}^F$, with derivatives $\partial^n\mathbf{f}(x) = [\partial^n f_1(x), \dots, \partial^n f_F(x)] \in \mathbb{R}^F$. We define the D feature vectors of rank F:

$$\mathbf{f}_l(\boldsymbol{\mathcal{D}}|\alpha) \equiv \mathbf{f}([\boldsymbol{\mathcal{D}} - \boldsymbol{\mu}_\alpha]\cdot\mathbf{v}_{\alpha,l}) \in \mathbb{R}^F, \quad l \in [1, D]. \tag{40}$$

In practice, we use polynomial features of typical order $F = 3 - 7$; we note that quadratic models ($F = 2$) are insufficient to capture the anharmonic behavior of the D-DOS shown below. The conditional entropy model then reads, with $\mathbf{\Theta}_l(\alpha) \in \mathbb{R}^F$,

$$\mathcal{S}_{\mathbf{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) \equiv \Theta_0(\alpha) + \sum_{l=1}^{D} \mathbf{f}_l(\boldsymbol{\mathcal{D}}|\alpha) \cdot \mathbf{\Theta}_l(\alpha). \tag{41}$$

The conditional descriptor score then reads

$$\boldsymbol{\nabla}\mathcal{S}_{\mathbf{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) = \sum_{l} \left(\partial\mathbf{f}_l(\boldsymbol{\mathcal{D}}|\alpha) \cdot \mathbf{\Theta}_l(\alpha)\right) \mathbf{v}_{\alpha,l} \in \mathbb{R}^F, \tag{42}$$

giving a score matching loss that is quadratic in $\mathbf{\Theta}_l(\alpha)$; by the orthonormality of the $\mathbf{v}_{\alpha,l}$, minimization reduces to solving the D linear equations of rank F:

$$N\langle\partial\mathbf{f}_l(\boldsymbol{\mathcal{D}}|\alpha)[\partial\mathbf{f}_l(\boldsymbol{\mathcal{D}}|\alpha)]^\top\rangle_\alpha\mathbf{\Theta}_l(\alpha) = -\langle\partial^2\mathbf{f}_l(\boldsymbol{\mathcal{D}}|\alpha)\rangle_\alpha. \tag{43}$$

Solution of (43) for each α fixes $\mathbf{\Theta}_l(\alpha)$, while the constant $\Theta_0(\alpha) \in \mathbb{R}$ is determined by equation (22), i.e. ensures $\max_{\boldsymbol{\mathcal{D}}} \mathcal{S}_{\mathbf{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) = 0$. For some model $\mathbf{w}$, the conditional free energy (28) then reads

$$\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta|\alpha) \equiv \min_{\boldsymbol{\mathcal{D}}} \left(\mathbf{w} \cdot \boldsymbol{\mathcal{D}} - \mathcal{S}_{\mathbf{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)/\beta\right), \tag{44}$$

which is achieved when $\nabla_{\boldsymbol{\mathcal{D}}}\mathcal{S}_{\mathbf{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha) = \beta\mathbf{w}$. We can then interpolate $\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta|\alpha)$ the sampled range of α values to give a final free energy estimate of

$$\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta) \equiv \min_{\alpha} \left(\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta|\alpha) - \mathcal{S}_0(\alpha)/\beta\right), \tag{45}$$

which is achieved when $\partial_\alpha\mathcal{S}_0(\alpha) = \beta\partial_\alpha\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta|\alpha)$. The final minimizing values of the descriptor vector $\boldsymbol{\mathcal{D}}^*$ allows evaluation of the gradient $\partial_{\mathbf{w}}\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta) = \boldsymbol{\mathcal{D}}^*$, equation (31). Equation (45) is the central result of this paper, a closed-form expression for the vibrational free energy of linear MLIPs (1).

C.4 Error analysis and correction of score matched estimates

To estimate errors on the free energy $\mathcal{F}_{\mathbf{w};\mathbf{\Theta}}(\beta)$, we can use standard error estimates to determine the uncertainty on the isosurface mean $\boldsymbol{\mu}_\alpha$ and covariance eigenvectors $\mathbf{v}_{\alpha,l}$ to produce errors $\delta\mathbf{f}_l(\boldsymbol{\mathcal{D}})$ on feature vectors (40). In addition, epistemic uncertainties on expectations in the score matching loss (43) will give uncertainties $\delta\mathbf{\Theta}_l(\alpha)$ on model coefficients $\mathbf{\Theta}_l(\alpha)$, which can be estimated by either subsampling the simulation data to produce an ensemble of model coefficients or extracting posterior uncertainties from Bayesian regression schemes[93]. Propagating these combined uncertainties provides a robust and efficient estimate of sampling errors, as shown in the numerical experiments.

C.5 D-DOS score matching sampling campaign

As detailed in section C.1, when using the harmonic isosurface function (23), our score matching sampling campaign reduces to sampling descriptor distributions on isosurfaces $\hat{\alpha}(\mathbf{X}) = \alpha$ defined by the Hessian $\mathbf{H}$ of some reference potential $E_0(\mathbf{X})$. For momentum-dependent isosurface functions (25) we instead record samples from an ensemble of short NVE runs, which we explore in section **??**. We tested the harmonic isosurface function (23) using one Hessian $\mathbf{H} = \mathbf{H}_x$ per phase for $x = \mathrm{W}, \mathrm{Mo}, \mathrm{Fe}$. Each Hessian was calculated using the appropriate lattice structure and the reference (loss minimizing) potential parameters $\mathbf{w} = \bar{\mathbf{w}}_x$ described in the previous section. With a given isosurface function $\hat{\alpha}(\mathbf{X})$, we generated $\mathcal{O}(10^3)$ independent samples on $\hat{\alpha}(\mathbf{X}) = \alpha$ for a range of α values at constant volume. It is simple to distribute sampling across multiple processors, as the harmonic isosurface samples are trivially independent (see D). This enables a significant reduction in the wall-clock time for sampling over trajectory-based methods such as thermodynamic integration.

Our open-source implementation uses `LAMMPS`[89] to evaluate `SNAP`[90] descriptors; as a rough guide, with $N = \mathcal{O}(10^2)$ atoms, the sampling campaign used to produce the results below required around $\mathcal{O}(10)$ seconds per α value on $\mathcal{O}(10^2)$ CPU cores. A converged score model built from $\mathcal{O}(10)$ α-values was thus achieved in under 5 minutes at each volume V and Hessian choice $\mathbf{H}$. Use of momentum-dependent isosurfaces $\hat{\alpha}(\mathbf{X}, \mathbf{P})$ requires a single free energy estimate, which

could be either from a separate D-DOS estimation or 'traditional' sampling methods. In addition, to allowing for NVE sample decorrelation gives a factor 10 greater sampling effort, i.e. comparable with the effort for a single fixed model sampling. Computational demands quoted are when using $\hat{\alpha}(\mathbf{X})$; future work will investigate schemes to further accelerate momentum-dependent $\hat{\alpha}(\mathbf{X}, \mathbf{P})$.

The final score model requires minimal storage, being only the $\mathcal{O}(100)$ scalars contained in the vector $\mathbf{\Theta}_\alpha$, equation (41), over a range of α values at constant V, $\mathbf{H}$. It is therefore possible to efficiently store many score models to investigate the influence of the reference model on free energy predictions. For example, figure **??** demonstrates how a D-DOS employing an isosurface function $\hat{\alpha}(\mathbf{X})$ using $\mathbf{H}$ from $\bar{\mathbf{w}}_{\mathrm{Mo}} \in \mathcal{W}_{\mathrm{Mo}}$ can accurately predict free energies from the W ensemble, $\mathbf{w} \in \mathcal{W}_{\mathrm{W}}$.

Table 1 provides a rough guide to the computational cost of existing methods, as reported in recent works[100, 65, 21], alongside the D-DOS sampling scheme detailed above. As can be seen, D-DOS is at least an order of magnitude more efficient than TI and up to two orders of magnitude more efficient than AS, even before considering the massive reduction in wall-clock time due to parallelization. We again emphasize that in addition to the modest computational requirements of D-DOS, sampling is *model-agnostic*, only performed for a given choice of descriptor hyperparameters, system volume and function $\hat{\alpha}(\mathbf{X})$ used for isosurface construction. This is the key innovation of the D-DOS approach, allowing rapid forward propagation for uncertainty quantification and, uniquely, back-propagation for inverse design goals.

D Derivation of Legendre transform expression for the free energy

D.1 Summary of Laplace's method

Laplace's method, also known as the steepest descents method, is a well-known identity allowing the evaluation of an integral of an exponentiated function multiplied by a large number. We provide a brief summary of the method here, and we refer the reader to e.g. [98] for further information. The method applies to a function $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$ which is twice differentiable. We partition the domain $\mathbb{R}^n = \cup_{l=1}^L \mathcal{R}_l$ into regions $\mathcal{R}_l$ each with a single maximum $\mathbf{x}_l^*$, where the negative Hessian matrix $\mathbf{H}_l = -\nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^\top f_{|\mathcal{R}_l} \in \mathbb{R}^{n\times n}$ of $Nf(\mathbf{x})$ has $\mathcal{O}(n)$ positive eigenvalues $\lambda_p \geq 0$, no negative eigenvalues, and all entries of $\mathbf{H}_l$ (and thus all λ_p) are independent of N. In the limit $N \to \infty$ the integral in $\mathcal{R}_l$ is dominated by the maximum $\mathbf{x}_l^*$. Proof of Laplace's method uses Taylor expansions of $f(\mathbf{x})$ around $\mathbf{x}_l^*$ to provide upper and lower bounds, which in the limit $N \to \infty$ both converge to the same Gaussian integral, giving

$$\lim_{N\to\infty} \int_{\mathbb{R}^n} \exp[Nf(\mathbf{x})] \mathrm{d}\mathbf{x} = \sum_{l=1}^{L} \frac{\exp[Nf(\mathbf{x}_l^*)]}{\sqrt{(2\pi N)^n |\mathbf{H}_l|}}. \tag{46}$$

In the case of constant n as $N \to \infty$ it is simple to show that the limiting form of the log integral is

$$\lim_{N\to\infty} \frac{1}{N} \ln \int_{\mathbb{R}^n} \exp[Nf(\mathbf{x})] \mathrm{d}\mathbf{x} = \max_l f(\mathbf{x}_l^*). \tag{47}$$

where we use the fact that

$$\lim_{N\to\infty} \frac{1}{N} \ln \sqrt{(2\pi N)^n |\mathbf{H}_l|} = \lim_{N\to\infty} \frac{1}{2N} \left(n \ln |2\pi N| + \sum_{p=1}^{n} \ln \lambda_p \right) = 0. \tag{48}$$

as $\lim_{N\to\infty} n/N = 0$ and $\lim_{N\to\infty} (1/N) \ln |N| = 0$. When the argument of the function has dimension which scales with N, i.e. $n = rN$, $r > 0$, ($r = 3$ for Hessians), the above simplification does not hold. In general, the integral will depend on higher order gradients to correctly take the limit. Note that the above is distinct from the common use of Laplace's method to approximate the partition function integral $\int_{\mathbb{R}^{3N}} \exp[-\beta E(\mathbf{X})] \mathrm{d}\mathbf{X}$; although the dimension of $\mathbf{X}$ is extensive, in this case Laplace's method is used in the low temperature limit $\beta \to \infty$, rather than $N \to \infty$.

D.2 Isosurface for a harmonic solid

D.2.1 Sampling

For solid systems we use a harmonic reference potential energy $E_0(\mathbf{X})$ and isosurface function $\hat{\alpha}(\mathbf{X})$

$$E_0(\mathbf{X}) \equiv \frac{[\mathbf{X} - \mathbf{X}_0]^\top \mathbf{H} [\mathbf{X} - \mathbf{X}_0]}{2}, \quad \hat{\alpha}(\mathbf{X}) \equiv \ln \left| \frac{\hat{E}_0(\mathbf{X})}{N \mathrm{U}_0} \right|. \tag{49}$$

We assume that $\mathbf{H}$ has $3N' = 3N - 3$ positive eigenvalues $\nu_l > 0$, $l > 3$ with normalized eigenvectors $\mathbf{v}_l$. We can thus define normal mode coordinates $\tilde{\mathrm{X}}_l \equiv \mathbf{v}_l \cdot \mathbf{X} / \sqrt{\nu_l}$, $l > 3$. In addition, we have 3 zero modes $\nu_l = 0$, $l = 1, 2, 3$ with eigenvectors selecting the center of mass $\bar{\mathbf{x}}$ multiplied by $\sqrt{N}$, i.e. $\tilde{\mathrm{X}}_l \equiv \sqrt{N} \bar{\mathrm{x}}_l$, $l = 1, 2, 3$, which meaning we can always ensure normal modes have zero net displacement, i.e. enforce $\mathbf{v}_l \cdot \mathbf{1} = 0$, $l > 3$.

In normal mode coordinates, the energy writes

$$E_0(\mathbf{X}) = \frac{1}{2} \sum_{l=4}^{l=3N} \nu_l \|\mathbf{v}_l \cdot \mathbf{X}\|^2 = \frac{1}{2N} \sum_{l=4}^{l=3N} \tilde{\mathrm{X}}_l^2 = \frac{\tilde{\mathrm{R}}^2}{2}. \tag{50}$$

Sampling the isosurface $\hat{\alpha}(\mathbf{X}) = \alpha$ is clearly equivalent to sampling $E_0(\mathbf{X}) = N \mathrm{U}_0 \exp(\alpha)$, which in normal mode coordinates amounts to sampling the surface of a hypersphere with radius $\tilde{\mathrm{R}} = \sqrt{2N\mathrm{U}_0} \exp(\alpha/2)$.

With a unit vector $\mathbf{u} = [\mathrm{u}_1, \ldots, \mathrm{u}_{3N'}] \in \mathbb{R}^{3N'}$ on the $3N'$ dimensional hypersphere, isosurface samples can then be produced through

$$\mathbf{X}_\alpha[\mathbf{u}] \equiv \mathbf{X}_0 + \sum_{l=1}^{3N'} \sqrt{\frac{2N\mathrm{U}_0 \exp(\alpha)}{\nu_l}} \mathrm{u}_l \mathbf{v}_l, \quad \Rightarrow \quad E_0\left(\mathbf{X}_\alpha[\mathbf{u}]\right) = N\mathrm{U}_0 \exp(\alpha) \sum_{l=1}^{3N'} \mathrm{u}_l^2 = N\mathrm{U}_0 \exp(\alpha). \tag{51}$$

Importantly, the sampling procedure can be trivially parallelized as we can generate independent samples $\{\mathbf{u}\}$ on each parallel worker, providing each worker with a unique seed for pseudo-random number generation.

D.2.2 Isosurface volume and isosurface entropy

For harmonic isosurface functions $\hat{\alpha}(\mathbf{X})$, we can express the isosurface volume (19) in normal mode coordinates using standard expressions for change of variables:

$$\Omega(\alpha) = \int_{\mathbb{R}^{3N}} \delta(\hat{\alpha}(\mathbf{X}) - \alpha) \mathrm{d}\mathbf{X} \tag{52}$$

$$= \frac{V}{\prod_{l=4}^{3N} \sqrt{\nu_l}} \int_{\mathbb{R}^{3N'}} \delta\left(\ln \left| \sum_{l=4}^{3N} \tilde{\mathrm{X}}_l^2/(2N) \right| - \alpha \right) \prod_{l=4}^{3N} \mathrm{d}\tilde{\mathrm{X}}_l. \tag{53}$$

Converting to spherical coordinates we find, using the expression for the surface area of a unit sphere in $3N'$ dimensions as $S_{3N'} = 2\pi^{3N'/2}/\Gamma(3N'/2)$, then again changing variables with $\mathrm{d}\tilde{\mathrm{R}} = \sqrt{N\mathrm{U}_0/2} \exp(\alpha/2) \mathrm{d}\alpha$, we find that

$$\Omega(\alpha) = \frac{2V\pi^{3N'/2}}{\Gamma(3N'/2)} \int_{\mathbb{R}_+} \delta\left(\ln \left| \tilde{\mathrm{R}}^2/(2N) \right| - \alpha \right) \tilde{\mathrm{R}}^{3N'-1} \mathrm{d}\tilde{\mathrm{R}} \tag{54}$$

$$= \frac{V\sqrt{\mathrm{U}_0}^{3N'}}{\prod_{l=4}^{3N} \sqrt{\nu_l}} \frac{\sqrt{2\pi N}^{3N'}}{\Gamma(3N'/2)} \exp(3N\alpha/2). \tag{55}$$

We thus see that the isosurface volume for harmonic solids has the general form

$$\Omega(\alpha) = \Omega_0 \exp(3N\alpha/2), \quad \Omega_0 = \frac{V\sqrt{\mathrm{U}_0}^{3N'}}{\prod_{l=4}^{3N} \sqrt{\nu_l}} \frac{\sqrt{2\pi N}^{3N'}}{\Gamma(3N'/2)}, \tag{56}$$

giving an isosurface entropy

$$\mathcal{S}_0(\alpha) \equiv \lim_{N\to\infty} \frac{1}{N} \ln |\Omega(\alpha)/\lambda_0^{3N}(\beta)| = S_0 + 3\alpha/2, \quad S_0 = \ln |\lambda_0^3(\beta)| + (1/N) \ln |\Omega_0|. \tag{57}$$

While we can simplify the expression for the $\ln \Omega_0$ using Stirling's approximation we shall see this is not required.

D.2.3 Isosurface entropy and connection to harmonic free energy

Using standard Gaussian integrals, the partition function of a harmonic system reads, with $\lambda_0(\beta) = h\sqrt{\beta/(2\pi m)}$,

$$Z_0(\beta) = \frac{1}{\lambda_0^{3N}(\beta)} \int_{\mathbb{R}^{3N}} \exp[-\beta E_0(\mathbf{X})] \mathrm{d}\mathbf{X} = \frac{V}{\lambda_0^{3N}(\beta)} \prod_{l=4}^{3N} \frac{1}{\sqrt{2\pi\beta\nu_l}}, \tag{58}$$

giving a free energy in the limit $N \to \infty$

$$\mathcal{F}_0(\beta) \equiv \lim_{N\to\infty} \frac{-1}{N\beta} \ln |Z_0(\beta)| = \frac{1}{N\beta} \sum_{l=4}^{3N} \ln \left| \beta\hbar\sqrt{\nu_l/m} \right|. \tag{59}$$

We can also write $\mathcal{F}_0(\beta)$ using the isosurface entropy defined in equation (57) and applying Laplace's method, i.e.

$$\beta\mathcal{F}_0(\beta) = \lim_{N\to\infty} \frac{-1}{N} \ln \left| \frac{1}{\lambda_0^{3N}(\beta)} \int_{\mathbb{R}} \Omega(\alpha) \exp(-N\beta \mathrm{U}_0 e^{\alpha}) \mathrm{d}\alpha \right| \tag{60}$$

$$= \lim_{N\to\infty} \frac{-1}{N} \ln \left| e^{NS_0} \int_{\mathbb{R}} \exp(3N\alpha/2 - N\beta \mathrm{U}_0 e^{\alpha}) \mathrm{d}\alpha \right|, \tag{61}$$

$$= \min_{\alpha} \mathrm{U}_0 e^{\alpha} - 3\alpha/2 - S_0, \tag{62}$$

$$= \mathrm{U}_0 e^{\alpha} - 3\alpha/2 - S_0 \Big|_{\alpha = -\ln|2\beta \mathrm{U}_0/3|}, \tag{63}$$

$$= 3/2 - S_0 + 3/2 \ln|2\beta \mathrm{U}_0/3|, \tag{64}$$

$$\Rightarrow S_0 = 3/2 + 3/2 \ln|2\beta \mathrm{U}_0/3| - \beta\mathcal{F}_0(\beta) \tag{65}$$

which allows us to express the constant S_0 purely in terms of the harmonic free energy.

D.3 Momentum-dependent isosurface

Estimating the free energy of e.g. liquid or highly anharmonic phases typically requires more complex reference potential energy models than the harmonic form used above. While we leave a comprehensive numerical study for future work, the following details how the CD-DOS treatment can be generalized to a momentum dependent isosurface

$$\hat{\alpha}(\mathbf{X}, \mathbf{P}) \equiv \ln \left| \frac{\hat{K}(\mathbf{P}) + \hat{E}_0(\mathbf{X})}{N\mathrm{U}_0} \right|. \tag{66}$$

using a kinetic energy function $\hat{K}(\mathbf{P}) = \sum_{i=1}^{3N} \mathrm{p}_i^2/(2m_i)$. Isosurface sampling then corresponds to microcanonical (NVE) dynamics with any reference potential, where the per-atom internal energy satisfies $\mathcal{U} = \mathrm{U}_0 \exp(\alpha)$. Such a generalization has close analogies with Hamiltonian Monte Carlo methods[15], which can use generalized kinetic energies[60]. The isosurface volume of $\hat{\alpha}(\mathbf{X}, \mathbf{P}) = \alpha$ is defined as

$$\Omega(\alpha) \equiv \int_{\mathbb{R}^{3N}\times\mathbb{R}^{3N}} \delta\left(\hat{\alpha}(\mathbf{X}, \mathbf{P}) - \alpha\right) \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{P}. \tag{67}$$

and we evaluate the entropy below. In this case, we treat $D_K = \hat{K}(\mathbf{P})/N$ as an additional intensive descriptor to give an extended conditional descriptor density of states

$$\Omega(\boldsymbol{\mathcal{D}} \oplus D_K|\alpha) \equiv N \int_{\mathbb{R}^{3N}\times\mathbb{R}^{3N}} \frac{\delta(\hat{K}(\mathbf{P}) - D_K)\delta\left(\hat{\alpha}(\mathbf{X}, \mathbf{P}) - \alpha\right)}{\Omega(\alpha)} \times \delta\left(\boldsymbol{\mathcal{D}} - (1/N)\sum_{i=1}^{N} \hat{\boldsymbol{\phi}}(\mathbf{D}_i(\mathbf{X}))\right) \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{P}, \tag{68}$$

By the same manipulations as for the momentum-independent case, this extended conditional descriptor density of states is normalized:

$$\int_{\mathbb{R}^D\times\mathbb{R}_+} \Omega(\boldsymbol{\mathcal{D}} \oplus D_K|\alpha) \mathrm{d}\boldsymbol{\mathcal{D}}\mathrm{d}D_K = 1. \tag{69}$$

However, as samples are not independent, the efficacy will depend on the decorrelation time[59] of microcanonical trajectories. A full study of how such momentum-dependent isosurfaces can be used to estimate the descriptor density of states $\Omega(\boldsymbol{\mathcal{D}})$ and thus the free energy of liquid phases and melting temperatures will be the focus of future work.

D.3.1 Isosurface entropy

The isosurface entropy $\mathcal{S}_0(\alpha)$ cannot be evaluated analytically and instead requires free energy estimation schemes such as thermodynamic integration, discussed in A. To see how this emerges, we

use the definition of the isosurface entropy (20) to write the free energy as

$$\beta\mathcal{F}_0(\beta) = \lim_{N\to\infty} \frac{-1}{N} \ln \left| \frac{1}{h^3 N} \int_{\mathbb{R}} \Omega(\alpha) \exp(-N\beta \exp(\alpha)) d\alpha \right|, \tag{70}$$

$$= \lim_{N\to\infty} \frac{-1}{N} \ln \left| \int_{\mathbb{R}} \exp(N\mathcal{S}_0(\alpha) - N\beta \exp(\alpha)) d\alpha \right|, \tag{71}$$

$$= \min_{\alpha} \beta \mathrm{U}_0 \exp(\alpha) - \mathcal{S}_0(\alpha). \tag{72}$$

It is clear that this minimum is satisfied when $\partial_\alpha \mathcal{S}_0(\alpha) = \beta \mathrm{U}_0 \exp(\alpha)$, and at the minimum $\mathrm{U}_0 \exp(\alpha)$ is clearly the internal energy $\mathcal{U}_0(\beta)$. We can therefore define β_α through the condition $\mathcal{U}_0(\beta_\alpha) \equiv \mathrm{U}_0 \exp(\alpha)$ and thus write

$$\beta_\alpha \mathcal{F}_0(\beta_\alpha) = \beta_\alpha \mathcal{U}_0(\beta_\alpha) - \mathcal{S}_0(\alpha). \tag{73}$$

With a tabulation of the intensive per-atom free energy $\mathcal{F}_0(\beta)$ and total internal energy $\mathcal{U}_0(\beta)$ over a range of temperatures $1/\beta$, the isosurface entropy reads

$$\mathcal{S}_0(\alpha) \equiv \beta[\mathcal{U}_0(\beta_\alpha) - \mathcal{F}_0(\beta_\alpha)], \quad \mathcal{U}_0(\beta_\alpha) \equiv \mathrm{U}_0 \exp(\alpha). \tag{74}$$

The value of β_α is uniquely defined when $\mathcal{U}_0(\beta)$ is monotonic with β.

E Intensivity of the descriptor entropy

This appendix provides a proof that the descriptor entropy $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ is intensive.

By the locality of the descriptor energy $E_{\mathbf{w}} = \sum_i E^1_{\mathbf{w}}(\mathbf{D}_i)$, any two per-atom feature vectors $\boldsymbol{\mathcal{D}}_i, \boldsymbol{\mathcal{D}}_j$ will be independent when the corresponding atoms are spatially separated, i.e. $|\mathbf{r}_{ij}| \to \infty$. As a result, the per-atom feature vector $\boldsymbol{\mathcal{D}}_i$ will have nonzero correlation with only a finite number $N_c \ll N$ of other per-atom feature vectors, indexed by some set $\mathcal{N}_i \subset \{1, \dots, N\}$, which has strong implications for the global vector $N\boldsymbol{\mathcal{D}} = \sum_{i=1}^{N} \boldsymbol{\mathcal{D}}_i$. In particular, it is clear that any cumulant[26] of $N\boldsymbol{\mathcal{D}}$ will be extensive, scaling linearly with N as $N \to \infty$. The first cumulant is the mean $N\langle\boldsymbol{\mathcal{D}}\rangle_\alpha$, where $\langle\boldsymbol{\mathcal{D}}\rangle_\alpha$ is clearly intensive. Defining $\delta\boldsymbol{\mathcal{D}}_i \equiv \boldsymbol{\mathcal{D}}_i - \langle\boldsymbol{\mathcal{D}}\rangle_\alpha$ and thus $\delta\boldsymbol{\mathcal{D}}$, the covariance of $N\boldsymbol{\mathcal{D}}$ writes

$$N^2\langle\delta\boldsymbol{\mathcal{D}} \otimes \delta\boldsymbol{\mathcal{D}}\rangle_\alpha = \sum_{i=1}^{N}\sum_{l\in\mathcal{N}_i}\langle\delta\boldsymbol{\mathcal{D}}_i \otimes \delta\boldsymbol{\mathcal{D}}_l\rangle_\alpha = N\boldsymbol{\Sigma}_\alpha \in \mathbb{R}^{D\times D}, \tag{75}$$

where $\boldsymbol{\Sigma}_\alpha$ is an average over each atom i of the sum of N_c covariance matrices between i and neighbors $l \in \mathcal{N}_i$, which is manifestly intensive. The third order cumulant writes

$$N^3\langle\delta\boldsymbol{\mathcal{D}} \otimes \delta\boldsymbol{\mathcal{D}} \otimes \delta\boldsymbol{\mathcal{D}}\rangle_\alpha = \sum_{i=1}^{N}\sum_{l\in\mathcal{N}_i}\sum_{m\in\mathcal{N}_l}\sum_{m=0}^{N_c}\langle\delta\boldsymbol{\mathcal{D}}_i \otimes \delta\boldsymbol{\mathcal{D}}_l \otimes \delta\boldsymbol{\mathcal{D}}_m\rangle_\alpha = N\boldsymbol{\Xi}_\alpha \in \mathbb{R}^{D\times D\times D}. \tag{76}$$

where $\boldsymbol{\Xi}_\alpha$ is an average over each atom i of the $N_c(N_c+1)/2$ third-order correlations between i and neighbors $l \in \mathcal{N}_i$ and $m \in \mathcal{N}_l$. As before, this is manifestly intensive, and we can continue this procedure to arbitrarily high orders. We can therefore define an intensive cumulant generating function [26] of $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ in the form

$$J(\mathbf{v}|\alpha) \equiv \frac{1}{N}\ln\left|\int_{\mathbb{R}^D} e^{N\mathbf{v}\cdot\boldsymbol{\mathcal{D}}}\Omega(\boldsymbol{\mathcal{D}}|\alpha)\mathrm{d}\boldsymbol{\mathcal{D}}\right| \tag{77}$$

$$= \boldsymbol{\mu}_\alpha \cdot \mathbf{v} + \frac{1}{2}\mathbf{v}^\top\boldsymbol{\Sigma}_\alpha\mathbf{v} + \frac{1}{6}\mathbf{v}^\top\boldsymbol{\Xi}_\alpha : \mathbf{v}\otimes\mathbf{v} + \dots, \tag{78}$$

where $\mathbf{v} \in \mathbb{R}^D$ and $J(\mathbf{v}|\alpha)$ are clearly intensive, meaning that we have the identity

$$\int_{\mathbb{R}^D} \exp\left(N\mathbf{v}\cdot\boldsymbol{\mathcal{D}}\right)\Omega\left(\boldsymbol{\mathcal{D}}|\alpha\right)d\boldsymbol{\mathcal{D}} = \exp\left(NJ(\mathbf{v}|\alpha)\right). \tag{79}$$

As the cumulants of $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ are finite, we know that $\Omega(\boldsymbol{\mathcal{D}}|\alpha)$ has a global maximum at finite $\boldsymbol{\mathcal{D}}$. As a result, we define the descriptor entropy as the (negative) Legendre-Fenchel transform of the cumulant generating function:

$$\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha) \equiv \min_{\mathbf{v}} J(\mathbf{v}|\alpha) - \mathbf{v}\cdot\boldsymbol{\mathcal{D}} \equiv \lim_{N\to\infty}\frac{1}{N}\ln|\Omega(\boldsymbol{\mathcal{D}}|\alpha)|, \tag{80}$$

which is clearly both intensive and convex in $\boldsymbol{\mathcal{D}}$. Equation 80 is closely related to Gärtner-Ellis theorem from large deviation theory[40, 36], which generalizes Crámer's theorem from i.i.d. observations to asymptotically independent observations, and shows the rate function is the Legendre-Fenchel transform of the cumulant generating function. We can thus identify the conditional entropy $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ as the negative rate function. We have thus established that $\mathcal{S}(\boldsymbol{\mathcal{D}}|\alpha)$ is intensive, as it is the sum of two manifestly intensive terms. As discussed in the main text, the conditional free energy $\mathcal{F}_{\mathbf{w}}(\beta|\alpha)$ can be expressed in terms of the cumulant generating function, with $\mathbf{v} = -\beta\mathbf{w}$; however, as we detail, we instead use score matching to estimate higher order moments.

F Numerical implementation of the D-DOS scheme

In this section, we describe in detail how the D-DOS sampling scheme is implemented, and how a broad ensemble of free energy estimates was produced using thermodynamic sampling in order to provide stringent tests of D-DOS free energy estimates. We describe the low-rank linear MLIPs employed (F.2), the production of DFT training data (F.3) the production of reference free energy estimates via thermodynamic integration (14) and details of the D-DOS score matching campaign. We focus on MLIPs that approximate the BCC, A15 and FCC phases of tungsten (W), molybdenum (Mo) and iron (Fe)[45].

F.1 *Ab initio* database design for MLIP training

We have performed an iterative construction of the database. The final aim is to have a potential that satisfies the following requirements: (i) it should reproduce the *ab initio* elastic constants at 0 K; (ii) it must provide a reasonable thermal expansion from 0 K to the melting temperature; and (iii) it should mimic the thermodynamics of BCC and A15 phase from 0 K to the melting temperature.

The DFT calculations were performed using VASP [57]. We have used a PAW pseudopotential [58]: we have used PPs with sp core states and 12 valence electrons in the $4s^2 4p^6 4d^5 5s^1$ states. The cut-off energy for plane-waves is 500 eV. In order to sample reciprocal space, we used Monkhorst-Pack [67] method to build a constant k-points density $\rho_k = 1/(24a_0)^3$ for all the computed configurations, which translates in $6 \times 6 \times 6$ k-points for the 128-atom cell of BCC Mo. Methfessel and Paxton [66] smearing algorithm with $\sigma = 0.3$ eV is used. We have used GGA exchange correlation in PBE [75].

Firstly, we generated a minimal *ab initio* database, DB_1, designed to build the initial version of the potentials. These potentials were then used to generate additional configurations similar to the defects we intend to simulate. The configurations were then computed using DFT without structural relaxation and reintegrated into the more complete database, DB_2. We reiterate the procedure from DB_2 to DB_3. All generated configurations are collected in Table 2. In the following, we detail each component of the database. In the end, the different databases are ruled by the following inclusion relations: $\mathrm{DB}_1 \subset \mathrm{DB}_2 \subset \mathrm{DB}_3$.

The `Cxx` class contains configurations involving iso-volumic deformations, from which the values of the bulk modulus B and the anisotropic elastic constants C_{11}, C_{12}, and C_{44} can be easily extracted. This class provides reliable information for the BCC elastic constants of the MLP. We have used 39 deformations. To minimize numerical round-off errors, the *ab initio* energy calculations are performed in $(4a_0)^3$ cubic supercells (128 atoms). The `epsilon_bulk` class corresponds to random deformation at a constant volume of the cubic cell of 2 atoms of BCC. We impose a deformation $\boldsymbol{\epsilon}_0$ to which we add a random tensor $\delta\epsilon$ defined by $\delta\epsilon_{ij} \sim \varepsilon\, \mathcal{N}(0,1)$. ε is the amplitude of random noise and $\mathcal{N}(0,1)$ is a standardized Gaussian distribution. In the end, we apply the following deformation tensor to the configuration : $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_0 + \frac{1}{2}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top\right)$, We apply uniformly distributed deformations between -5% and 5% with a random parameter $\varepsilon = 0.01$. In the end, we generate 1000 random deformed configurations.

The `noised_` classes are designed to mimic molecular dynamics simulations at a given temperature and avoid the computational expense of *ab initio* molecular dynamics. This is achieved by adding carefully thermal noise to the relaxed 0 K configurations of bulk, mono-, di-, and tri-vacancies.

The class `NEB_` corresponds to standard Nudged Elastic Band [51] pathways computed in DFT for the first nearest-neighbor migration of mono-, di- and tri-vacancies. The convergence criterion is defined as the maximum force being less than 10^{-2} eV/Å . Once the first version of the potentials was fitted from DB_1, the MLP potentials were used to generate finite-temperature pathways from the 0 K trajectories. These configurations are included in the `PAFI_` class. The finite temperature configurations are sampled from the PAFI [85] hyperplanes near the saddle point at a given temperature.

The class `heated_cell` corresponds to NPT molecular dynamics simulations at zero pressure, conducted from 300 K to 5000 K for a simulation cell containing perfect bulk BCC and A15, mono-, di-, and tri-vacancies. The heating ramp is applied at a rate of 5 K/ps. From the molecular dynamics performed with the MLP derived from DB_1, configurations were selected between 3000 K and 5000 K (if the potential was stable, see main text discussion). For DB_3, we randomly chose 388

configurations distributed between 1000 K and 4000 K to help stabilize the BCC-to-liquid as well as A15-to-liquid transitions.

DB_1	Temperature in K						Total
	0	875	1750	2625	3500	MDr	
Cxx	13	0	0	0	0	0	13
ϵ_bulk	1000	0	0	0	0	0	1000
noised_bulk	1	10	10	10	10	0	41
noised_V_1	10	10	10	10	10	0	50
noised_V_2	30	30	30	30	30	0	150
NEB_$V_{1,2,3}$	21	0	0	0	0	0	21
Total DB_1	1075	50	50	50	50	0	1275
DB_2							
DB_1	1075	50	50	50	50	0	1275
PAFI_$V_{1,2,3}$	0	11	11	11	8	0	41
heated_cell						9	9
Total DB_2	1075	61	61	61	58	9	1325
DB_3							
DB_2	1075	61	61	61	58	9	1325
heated_cell						388	388
Total DB_3	1075	61	61	61	58	397	1722

Table 2: An iterative list of atomic configurations for the minimal databases $DB_{1,2,3}$. Cxx denotes the deformations used to obtain accurate elastic constants. The cubic cell of the BCC lattice is subjected to various non-zero strains, ϵ, for the class ϵ_bulk. The noised_ configurations, designed to mimic the MD of bulk, mono- and di-vacancies in various configurations are denoted by V_1 and V_2. NEB and PAFI_ represents sampling from 0 K to finite temperature for the vacancy jump, employing the NEB [51, 50] and PAFI methods [85, 86],respectively. MDr denotes the molecular dynamics trajectories for heating the system from 300 K to 5000 K, which provide the class heated_cell. Further details about all the classes can be found in the text.

F.2 Choice of linear MLIP

We build a linear MLIP using the bispectral BSO(4) descriptor functions, first introduced in the SNAP MLIP family[90]. While a quadratic featurization is often used[44, 49] we employ the original linear model, i.e. $\boldsymbol{\mathcal{D}}_i = \hat{\boldsymbol{\phi}}(\mathbf{D}_i) = \mathbf{D}_i \in \mathbb{R}^d$. For unary systems we have $H = 4$ hyperparameters $\mathbf{h}$, the cutoff radius r_c, the number of bispectrum components D and two additional weights in the representation of the atomic density. We refer the reader to the original publications for further details[90]. To test the transferability of the sampling scheme under different reference models $E_0(\mathbf{X})$, we fix $\mathbf{h}$ to be the same for all potentials, regardless of the specie in training data, using a cutoff radius of $r_c = 4.7$Å and $D = 55$ bispectrum components. While we consider models approximating Mo and W (see section F.3), which have similar equilibrium volumes, we note that the bispectrum descriptor is invariant[90] under a homogeneous rescaling of both the atomic configuration and the cutoff radius, i.e. the CD-DOS is invariant for fixed V/r_c^3.

We expect the numerical results of this section to hold directly if we replace the bispectrum descriptor with other "low-dimensional" ($d = D = \mathcal{O}(100)$) descriptors such as POD[71] or hybrid descriptors in MILADY[46, 35]. For models such as MTP[80] or ACE[62], where $d = \mathcal{O}(10^3)$, the features or score model (or both) will accept some rank reduction, e.g. a linear projection $\boldsymbol{\mathcal{D}}_i(\mathbf{D}_i) = \mathbf{P}\mathbf{D}_i$, where $\mathbf{P} \in \mathbb{R}^{D\times d}$, with $D = \mathcal{O}(100)$. The POD scheme applies rank-reduction to the radial part of the descriptor, following [47]. Many other rank-reduction schemes have been proposed in recent years, including linear embedding[27, 95] or tensor sketching [28].

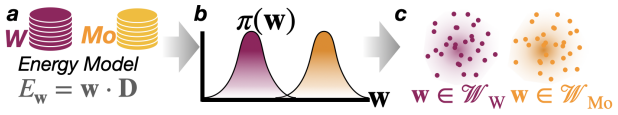

Figure 3: Producing ensemble of potential parameters for testing D-DOS in forward propagation. **a**) A generalized linear MLIP form is chosen to approximate *ab initio* databases, here of Mo and W. **b**) Misspecification-aware Bayesian regression[83] returns a parameter posterior which is broad for simple MLIPs and diverse databases. **c**) The posterior is sampled to produce an ensemble of stable model parameters $\mathbf{w} \in \mathcal{W}_{\mathrm{W}}$ and $\mathbf{w} \in \mathcal{W}_{\mathrm{Mo}}$ used for testing.

F.3 Training data for Fe, W and Mo

The majority of the database configurations for Fe and W are those published in [44]. The W database originates from the defect- and dislocation-oriented database in [44], which was modified and updated with molecular dynamics instances in [100] to improve its suitability for finite-temperature calculations and thermoelasticity of W. Finally, for this study, using the MLIP developed in [100], we prepared multiple samples of W in the A15 phase or liquid within the NPT ensemble, covering temperatures from 100 K to 5000 K. Each system contained 216 atoms. We selected 96 snapshots, which were then recomputed using the same DFT parameterization as in [44, 100]. The Mo database was specifically designed for this study to ensure a well-represented configuration of Mo at high temperatures in the BCC and A15 phases. The detailed components of the database, as well as the ab initio details, are described in the Appendix F.1.

F.4 Ensemble of potential parameters for testing in forward-propagation

From the DFT training databases for $x = \mathrm{W}, \mathrm{Mo}, \mathrm{Fe}$ (F.3), we generate a broad range of parameter values $\mathbf{w} \in \mathcal{W}_x$ for `SNAP` MLIPs (F.2) using a recently introduced [83] Bayesian linear regression scheme. The scheme is designed to produce robust parameter uncertainties for misspecified surrogate models of low-noise calculations, which is precisely the regime encountered when fitting linear MLIPs to DFT data.

Taking training data for $x = \mathrm{W}$ or Mo, the method produces a posterior distribution $\pi(\mathbf{w})$ (Figure 3a-b), with strong guarantees that posterior predictions bound the true DFT result, irrespective of how each training point is weighted. As the `SNAP` form has a relatively small number ($\mathcal{O}(100)$) of adjustable parameters it is strongly misspecified (large model-form error) to the diverse training database and thus the posterior distribution gives a broad range of parameter values.

Each training point was weighted using a procedure described elsewhere [44, 100]. While we also explored randomly varying weights associated with defects and other disordered structures, in all cases we maintained consistently high weights for structures corresponding to small deformations of the cubic unit cell in the BCC, FCC, or A15 phases. This procedure ensures that the resulting potential ensemble yields lattice parameters within a range of 10^{-4} Å and elastic constants that follow a narrow distribution centered around the target DFT average values.

We construct our ensemble $\mathcal{W}_x$, $x = \mathrm{W}, \mathrm{Mo}, \mathrm{Fe}$ by applying CUR sparsification [33, 34] to a large set of posterior samples to extract $\mathcal{O}(100)$ parameter vectors which show sufficient dynamical stability to allow for convergence when performing thermodynamic integration at high temperature (Figure 3c). We also identify a 'reference' value $\bar{\mathbf{w}}_x$, being a stable parameter choice that has the optimal error to training data, i.e. the best overall interatomic potential choice. For each parameter $\mathbf{w}$ we have computed the free energy $\mathcal{F}^L_{\mathbf{w}}(\beta)$, as is detailed in the section F.5.

F.5 Free energies from thermodynamic integration

With a given choice of MLIP parameters $\mathbf{w}$, we employ a recently introduced thermodynamic integration method[19, 100] to calculate the corresponding NVT phase free energies $\mathcal{F}^L_{\mathbf{w}}(\beta)$, equation (9). The thermodynamic scheme first calculates the Hessian matrix $\mathbf{H}_{\mathbf{w}}$ for a given parameter choice, to give a harmonic free energy prediction and to parametrize a 'representative' harmonic reference. Rather than the sequential integration over η as described by equation (14), the employed scheme instead uses a Bayesian reformulation to sample all $\eta \in [0, 1]$ values simultaneously, which

significantly accelerates convergence[19]. In addition, a 'blocking' constraint is used to prevent trajectories escaping the metastable basin of any crystalline phase. We refer the reader to[100] for further details.

Even with these blocking constraints, in many cases phases had poor metastability at high temperatures, in particular the A15 phase, which was rectified by adding more high temperature A15 configurations to training data and restricting the range of potential parameters. These dynamical instabilities reflect general trends observed in long molecular dynamics trajectories, where high-dimensional MLIPs are prone to failure over long time simulations[53, 92].

While there is currently no general solution to the MLIP stability problem, even for the relatively low-dimensional ($D = \mathcal{O}(100)$) descriptors used in this study, it can be mitigated by enriching the training database[100]. In contrast, our score matching procedure only requires stability of the Hessian matrix $\mathbf{H}$ used for the harmonic reference potential $E_0(\mathbf{X})$, a much weaker condition than dynamical stability. The observed accuracy, detailed below, strongly suggests our sampling scheme may be able to predict phase free energies for a much broader range of parameter space than those that can be efficiently sampled via traditional methods. A full exploration of this ability is one of the many future directions we discuss in the conclusions in the main text.

The final sampling campaign to generate reference free energies for comparison against D-DOS estimates required around $\mathcal{O}(10^4)$ CPU hours, or $\mathcal{O}(10^5)$ force calls per model, with blocking analysis[59, 4] applied to estimate the standard error in each free energy estimate. We emphasize that the scheme described in this section represents the state-of-the-art in free energy estimation for MLIPs. Nevertheless, for any given choice of model parameters, free energy estimation requires at least $\mathcal{O}(10^6)$ CPU hours, irrespective of available resources, which significantly complicates uncertainty quantification via forward propagation and completely precludes including finite temperature properties during model training via back-propagation. The model-agnostic D-DOS scheme detailed introduced in this paper provides a first general solution for MLIPs that can be cast into the general linear form used here.

F.6 Systematic error correction for D-DOS free energy estimates

We find the estimated D-DOS errors to be excellent predictors of the observed errors. In addition, both predicted and observed errors are typically very low, around 1-2 meV/atom, rising to 10 meV/atom if the reference model is poorly chosen or the system is particularly anharmonic. These errors were largely corrected via a momentum-dependent isosurface bringing observed and predicted errors back within the stringent 1-2 meV/atom threshold.

However, if tightly converged (<1meV/atom) estimates of the free energy are desired for a given parameter choice $\mathbf{w}$, the close connection between the D-DOS conditional free energy $\mathcal{F}_{\mathbf{w}}(\beta|\alpha)$ and free energy perturbation (FEP), discussed in the main text, offers a systematic correction scheme. Any predicted value of $\mathcal{F}_{\mathbf{w};\boldsymbol{\Theta}}(\beta|\alpha)$ from our score matching estimate can be updated through short isosurface sampling runs, recording the *difference* between observed cumulants of $\mathbf{w} \cdot \boldsymbol{\mathcal{D}}$ and those predicted by $\mathcal{S}_{\mathbf{w};\boldsymbol{\Theta}}(\boldsymbol{\mathcal{D}}|\alpha)$), conducted over a small range of α to account for updated moments changing the minimum solution $\partial_\alpha \mathcal{S}_0(\alpha) = \beta \partial_\alpha \mathcal{F}_w(\beta|\alpha)$. Following established FEP techniques[48, 21] this procedure can then be extended to include *ab initio* data. However, given the accuracy of our D-DOS estimations we focus on exploring the unique abilities of the D-DOS scheme in forward and back parameter propagation, leaving a study of this correction scheme to future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide theoretical derviation and numerical demonstration that we can forward and back-propagate parameter variations through the free energy calculation.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a dedicated limitation section.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result is supported by extensive appendices.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the code required to reproduce all findings and a notebook demonstration the back-propagation application.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code required to reproduce all findings and a notebook demonstration the back-propagation application.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail the score matching model and provide a step-by-step algorithm description in the supplied code.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where applicable (for forward propagation) we include error bars on the predicted and calculated free energies. See figure 1b).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: See table 1 and discussion in results section.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the guidelines and are confident that we are in full conformity.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work focuses on computational aspects of free energy calculation in solids. There are no direct societal impacts in terms of e.g. data privacy or targeting of specific groups.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no such risks in this paper.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We are owners of all code and data provided.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide documented and anonymous code.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No LLM usage in results or writing of paper or code.