PAPER • OPEN ACCESS

Parameter uncertainties for imperfect surrogate models in the low-noise regime

View the article online for updates and enhancements.

CrossMark

PAPER

Parameter uncertainties for imperfect surrogate models in the low-noise regime

Thomas D Swinburne[1,*] and Danny Perez[2]

[1] Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France
[2] Theoretical Division T-1, Los Alamos National Laboratory, Los Alamos, NM, United States of America
* Author to whom any correspondence should be addressed.

E-mail: thomas.swinburne@cnrs.fr and danny_perez@lanl.gov

Abstract

Bayesian regression determines model parameters by minimizing the expected loss, an upper bound to the true generalization error. However, this loss ignores model form error, or misspecification, meaning parameter uncertainties are significantly underestimated and vanish in the large data limit. As misspecification is the main source of uncertainty for surrogate models of low-noise calculations, such as those arising in atomistic simulation, predictive uncertainties are systematically underestimated. We analyze the true generalization error of misspecified, near-deterministic surrogate models, a regime of broad relevance in science and engineering. We show that posterior parameter distributions must cover every training point to avoid a divergence in the generalization error and design a compatible ansatz which incurs minimal overhead for linear models. The approach is demonstrated on model problems before application to thousand-dimensional datasets in atomistic machine learning. Our efficient misspecification-aware scheme gives accurate prediction and bounding of test errors in terms of parameter uncertainties, allowing this important source of uncertainty to be incorporated in multi-scale computational workflows.

Surrogate models are widely used across science and engineering to efficiently approximate the output of computationally intensive simulation engines [1–7]. Domain expertise is leveraged to design optimal features and model architecture before parameters are inferred from training data generated by the engines. In a broad range of applications, the inference problem shares three key characteristics:

(a) Simulation engines are near *deterministic*: outputs have vanishing aleatoric uncertainty, e.g. the atomic energy $V(\mathbf{X})$ from some quantum chemistry calculation is a near-deterministic function of atomic positions $\mathbf{X}$, with very weak uncertainty due to e.g. convergence of the electronic minimization scheme.

(b) Surrogate models are *misspecified*: no single choice of the P parameters can match all N observations, meaning parameters are intrinsically uncertain.

(c) Surrogate models are *underparametrized*: large quantities of training data are available, such that standard epistemic uncertainty estimates from Bayesian inference vanish, typically corresponding to the regime $P/N \ll 1$.

Misspecification affects both finite capacity models and deep learning approaches with finite training resources [8, 9]. A misspecification-aware learning scheme should minimize the cross entropy between predicted and observed data distributions, known as the *generalization error*, or under certain technical definitions, the population risk (see [10, 11]). However, the generalization error is typically not tractable for learning, both due to numerical instabilities and the lack of theoretical bounds for convergence. Bayesian learning schemes instead minimize an upper bound, the *expected loss* (log likelihood), which ignores misspecification but admits a robust learning scheme. Epistemic uncertainties are thus severely

underestimated, vanishing in the underparametrized limit, with broad implications for surrogate model selection, uncertainty quantification and error propagation, as noted by multiple groups [8, 9, 12–15].

In this paper, we analyze the generalization error of near-deterministic, misspecified and under-parametrized surrogate models, i.e. those typically used to approximate simulation engines in science and engineering. Our central application is to the problem of fitting interatomic potentials for atomic simulations [5, 16–18], where the training data is generated by quantum chemistry calculations. However, our analysis applies to the fitting of any surrogate model to a near-deterministic ground truth. Posterior distributions strictly become unbounded as the aleatoric uncertainty vanishes, meaning the low-noise/near-deterministic limit is only taken as a theoretical device to derive conditions that any posterior distribution must obey to avoid a logarithmic divergence in the generalization error of that distribution. We then design an ansatz that respect this condition and admits a robust learning scheme at finite N, with standard treatment of epistemic (N-dependent) errors, reducing to a constrained loss minimization problem for large N/P. Our final misspecification-aware parameter posterior has finite uncertainty even as $N/P \rightarrow \infty$, whilst the generalization error of the posterior from Bayesian inference diverges. The approach thus captures misspecification uncertainties missed by standard Bayesian regression schemes.

Our implementation for linear models is a robust and to our knowledge unique means to estimate misspecification uncertainties, and thus is the optimal available choice when fitting high-dimensional linear models to low-noise data.

Our main contributions are:

(a) We define *pointwise optimal parameter sets* (POPSs) for each training point, within which model predictions are exact. Parameter distributions must have mass in every POPS to avoid a divergent generalization error.

(b) We use the ensemble of loss minimizers from each POPS to design two ansatz posteriors, a reweighted $\mathcal{O}(N)$ ensemble and a bounding $\mathcal{O}(P)$ hypercube, both of which respect POPS occupancy and give a finite generalization error, lower than the posterior from Bayesian inference.

(c) For linear models, our ansatz can be efficiently evaluated via rank-one updates to a leverage-weighted loss minimizer. This gives an efficient scheme to incorporate misspecification into linear regression, which we show gives minimal computational overhead (approximately a factor of two in training) over Bayesian inference.

(d) A `pip`-installable Python implementation of our misspecification-aware POPS regression algorithm, following the Scikit-learn `linear_model` API [19], is provided at https://github.com/tomswinburne/POPS-Regression.git.

(e) We test the approach on synthetic datasets, showing how our scheme provides robust bounds on test errors. This performance is maintained under application to thousand-dimensional datasets for fitting machine learning interatomic potentials [5, 17] and informatics models [20, 21], untreatable with existing misspecification-aware methods.

The remainder of this paper is organized as follows. In section 1, we review Bayesian regression and PAC-Bayes bounds, highlighting the limitations of standard approaches for misspecified models. Section 2 presents our analysis of the generalization error for near-deterministic surrogate models, introducing the concept of POPSs. In section 3, we develop our misspecification-aware ansatz, providing efficient implementations for linear models. Section 4 demonstrates the effectiveness of our approach on both synthetic datasets and high-dimensional problems in atomic machine learning. Section 5 discusses our method in the context of existing uncertainty quantification and propagation schemes for atomistic simulation. Finally, section 6 summarizes our findings and highlights future directions for research in misspecification-aware regression.

1. Bayesian fitting of surrogate models

This section summarizes known results from Bayesian analysis and PAC-Bayes inequalities [10], focusing on the limit of interest for this paper: fitting a misspecified surrogate model in the case that the ground truth is (near-)deterministic, i.e. known to have vanishing aleatoric error. We provide a glossary of all terms in appendix A

1.1. Deterministic simulation engines
A simulation engine $\mathcal{E} : \mathcal{X} \rightarrow \mathcal{Y}$ takes input $\mathbf{X} \in \mathcal{X}$ and returns $\mathbf{Y} \in \mathcal{Y}$. We denote the dimension of $\mathcal{Y}$ and $\mathcal{X}$ by Y and X respectively. We consider the regime where the aleatoric uncertainty of $\mathcal{E}$ is weak, such that the

output $\mathbf{Y}$ follows a Gaussian distribution

$$\rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X}) = \mathcal{N}(\mathbf{Y}|\mathcal{E}(\mathbf{X}), \boldsymbol{\Sigma}_{\mathcal{Y}}(\mathbf{X})), \tag{1}$$

which is fully defined by the output mean $\langle \mathbf{Y}|\mathbf{X}\rangle \equiv \mathcal{E}(\mathbf{X})$ and covariance $\langle \mathbf{Y}\mathbf{Y}^{\top}|\mathbf{X}\rangle - \mathcal{E}(\mathbf{X})\mathcal{E}^{\top}(\mathbf{X}) \equiv \boldsymbol{\Sigma}_{\mathcal{Y}}(\mathbf{X})$, where $\|\boldsymbol{\Sigma}_{\mathcal{Y}}(\mathbf{X})\| = \mathcal{O}(\epsilon^{2Y})$, and ϵ is assumed small. As the aleatoric error is vanishingly weak, we assume homoskedatiscity $\boldsymbol{\Sigma}_{\mathcal{Y}}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathcal{Y}}$ for simplicity. We will later take the deterministic, underparametrised limit $\epsilon \to 0$, $N/P \to \infty$ assuming that $\epsilon = \mathcal{O}(\sqrt{P/N})$. With weights $\sum_{i=1}^{N} w_i = 1$, the training data $\mathcal{D}_N = \{\mathbf{Y}_i, \mathbf{X}_i\}_{i=1}^{N}$ form an input distribution in $\mathbf{X}$ of

$$\rho_N(\mathbf{X}) \equiv \sum_i w_i \delta(\mathbf{X} - \mathbf{X}_i), \ \rho_{\mathcal{D}}(\mathbf{X}) \equiv \lim_{N \to \infty} \rho_N(\mathbf{X}). \tag{2}$$

Learning requires that we can bound expectations over independent identically distributed (i.i.d.) samples $\mathbf{X}, \mathbf{Y} \in \mathcal{D}$, where $\mathcal{D}$ is the set of all possible input-output pairs, corresponding to the limit of infinite training data,

$$\langle f(\mathbf{Y}, \mathbf{X})\rangle_{\mathcal{D}} \equiv \int_{\mathcal{D}} f(\mathbf{Y}, \mathbf{X}) \rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X}) \rho_{\mathcal{D}}(\mathbf{X}) \, d\mathbf{X} d\mathbf{Y}. \tag{3}$$

For timeseries or other signal data the i.i.d. property is only satisfied for decorrelated segments $\mathbf{X} = \mathbf{x}_{\mathbf{t}=1}^{T}$, meaning $\boldsymbol{\Sigma}_{\mathcal{Y}}$ should capture signal correlations as in generalized least squares regression [10, 12, 22].

1.2. Deterministic surrogate models

A surrogate model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ for $\mathcal{E}$ is defined by P parameters $\boldsymbol{\Theta} \in \mathcal{P}$. Under a parameter distribution $\pi(\boldsymbol{\Theta})$, the predicted distribution has the same form as (1), reading

$$\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}) = \mathcal{N}(\mathbf{Y}|\mathcal{M}(\mathbf{X}, \boldsymbol{\Theta}), \boldsymbol{\Sigma}_{\mathcal{Y}}), \tag{4}$$

$$\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}) = \int_{\mathcal{P}} \rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}) \pi(\boldsymbol{\Theta}) \, d\boldsymbol{\Theta}. \tag{5}$$

Unlike typical Bayesian schemes [23, 24], we do not treat the covariance term $\boldsymbol{\Sigma}_{\mathcal{Y}}$ as fitting parameter but as an intrinsic property of $\mathcal{E}$. This has implications for model selection criteria, discussed in appendix B.

1.3. The generalization error and expected loss

The generalization error $\mathcal{G}[\pi]$ is simply the cross entropy of the predicted distribution $\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X})\rho_N(\mathbf{X})$ to the observed distribution $\rho(\mathbf{Y}|\mathbf{X})\rho_N(\mathbf{X})$; omitting the constant term $\mathcal{G}_{\mathcal{E}} \equiv \langle \ln|\rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X})|\rangle_{\mathcal{D}}$ we have

$$\mathcal{G}[\pi] = -\left\langle \ln\left|\int_{\mathcal{P}} \rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}) \pi(\boldsymbol{\Theta}) \, d\boldsymbol{\Theta}\right|\right\rangle_{\mathcal{D}}. \tag{6}$$

In principle any learning scheme should aim to minimize the functional $\mathcal{G}[\pi]$, but in practice learning schemes target a tractable upper bound, the *expected loss* $\mathcal{L}[\pi] \geqslant \mathcal{G}[\pi]$, often known as the negative log likelihood. Using the Jensen inequality $-\langle \ln x\rangle \leqslant -\ln\langle x\rangle$, as used to derive Gibbs' famous inequality for free energy estimation [25], we find that

$$\mathcal{G}[\pi] \leqslant \mathcal{L}[\pi] = -\int_{\mathcal{P}} \langle \ln|\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta})|\rangle_{\mathcal{D}} \pi(\boldsymbol{\Theta}) \, d\boldsymbol{\Theta}. \tag{7}$$

$\mathcal{L}[\pi]$ is clearly minimized by a delta function $\delta(\boldsymbol{\Theta} - \boldsymbol{\Theta}_{\mathcal{L}}^{*})$, where $\boldsymbol{\Theta}_{\mathcal{L}}^{*} \in \arg\min_{\mathcal{P}} \langle -\ln|\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta})|\rangle_{\mathcal{D}}$. As the delta function has formally zero width, the loss minimizing distribution thus has vanishing parameter uncertainty. The next section shows how this emerges from the inference posterior at finite N/P.

1.4. PAC-Bayes estimation and connection to inference

PAC-Bayes analysis [10] provides concentration inequalities [26] that bound expectations over $\mathcal{D}$ by empirical expectations over $\mathcal{D}_N$ with some probability. Whilst originally designed for classification problems, Germain and coworkers [22] established an important connection between Bayesian inference and probability approximately correct (PAC-Bayes) analysis for regression problems with unbounded losses, which bounds true expectations by empirical likelihood expectations through concentration inequalities. The PAC-Bayes framework provides a robust learning scheme without test-train splitting, and gives a rationalization for Bayesian inference as a form of regularization. In particular, minimizing the PAC-Bayes bound for the expected loss gives the familiar relation between posterior, prior and likelihood from Bayesian

inference. This subsection adapts the results of Germain *et al* [22] and Masegosa *et al* [12] in the near-deterministic limit of interest here. The central result of [22] is that the expected loss $\mathcal{L}[\pi]$ admits the PAC-Bayes upper bound, holding with probability $1 - \xi$, of

$$\mathcal{L}[\pi] \leqslant \int_{\mathcal{P}} \hat{L}_N(\boldsymbol{\Theta}) + \frac{1}{N}\ln\left|\frac{\pi}{\pi_0}\right|\mathrm{d}\pi(\boldsymbol{\Theta}) - \frac{\ln|\xi|}{N} + C_0, \tag{8}$$

where $\hat{L}_N(\boldsymbol{\Theta}) = -\langle \ln|\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X},\boldsymbol{\Theta})\rangle\rangle_N$ is the empirical loss over $\rho_N(\mathbf{X})$, $\pi_0(\boldsymbol{\Theta})$ is some *prior* and C_0 is a constant. The value of the constant C_0 requires certain assumptions on the distribution of the error between the empirical and true expected loss. Following [22], we assume that this error distribution is unbounded but sub-normal, i.e. all error fluctuations are bounded by a normal distribution of variance s^2, giving $C_0[\pi_0] < s^2/2$. We refer the reader to [22] and subsequent works [12, 27] for further discussion. The minimizer of the upper bound (8) is the posterior from Bayesian inference [22], reading, with a parameter λ to enforce normalization

$$\pi_{\mathcal{L},N}(\boldsymbol{\Theta}) = \pi_0(\boldsymbol{\Theta})\exp\left(\lambda - N\hat{L}_N(\boldsymbol{\Theta})\right). \tag{9}$$

As $N/P \to \infty$ the posterior is strongly peaked around a loss minimizer $\boldsymbol{\Theta}_{\mathcal{L}}^* \in \arg\min_{\boldsymbol{\Theta}\in\mathcal{P}} L(\boldsymbol{\Theta})$. Defining $\boldsymbol{\Sigma}_0^{-1}$ as the curvature of $-\ln\pi_0(\boldsymbol{\Theta})$ around $\boldsymbol{\Theta}_{\mathcal{L}}^*$, we use Laplace's method (appendix C) as $N/P \to \infty$ to write

$$\pi_{\mathcal{L}}(\boldsymbol{\Theta}) = \mathcal{N}\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}_{\mathcal{L}}^*, [\boldsymbol{\Sigma}_0 + N\mathcal{I}_{\mathcal{L}}^*]^{-1}\right), \tag{10}$$

where $\mathcal{I}_{\mathcal{L}}^* \equiv \lim_{N/P\to\infty} \boldsymbol{\nabla}_{\boldsymbol{\Theta}}\boldsymbol{\nabla}_{\boldsymbol{\Theta}}^{\top}\hat{L}_N(\boldsymbol{\Theta}_{\mathcal{L}}^*)$ is the Fisher information matrix. If $\mathcal{I}_{\mathcal{L}}^*$ is full rank the prior has vanishing influence and the expected loss is minimized by the sharp distribution $\pi_{\mathcal{L}}(\boldsymbol{\Theta}) \to \delta(\boldsymbol{\Theta} - \boldsymbol{\Theta}_{\mathcal{L}}^*)$, giving vanishing parameter uncertainty. Otherwise, the inference prior may still have finite width but, by construction, these will not have influence on model predictions. The generalization error of the loss minimizing distribution clearly diverges as $1/\epsilon^2$:

$$\lim_{N/P\to\infty} \mathcal{G}[\pi_{\mathcal{L}}] = \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathcal{L}}^*\boldsymbol{\Sigma}_{\mathcal{Y}}^{-1}\right) = \mathcal{O}\left(1/\epsilon^2\right). \tag{11}$$

$\boldsymbol{\Sigma}_{\mathcal{L}}^* = \langle[\mathcal{E}(\mathbf{X}) - \mathcal{M}(\mathbf{X},\boldsymbol{\Theta}_{\mathcal{L}}^*)][\mathcal{E}(\mathbf{X}) - \mathcal{M}(\mathbf{X},\boldsymbol{\Theta}_{\mathcal{L}}^*)]^{\top}\rangle$ is the error covariance around $\boldsymbol{\Theta}_{\mathcal{L}}^*$ and we use $\mathrm{Tr}(\boldsymbol{\Sigma}_{\mathcal{Y}}) = \mathcal{O}(\epsilon^2)$. Under misspecification it is thus clear that minimizing this upper bound is sub-optimal [12] as the generalization error diverges as $1/\epsilon^2$. In the next section we analyze the generalization error as $\epsilon \to 0$ to derive a condition for any $\pi(\boldsymbol{\Theta})$ that ensures a finite generalization error.

1.5. Related work on misspecification-aware Bayesian regression

In recent years, multiple groups have used the PAC-Bayes framework to derive tighter bounds for the generalization error than that provided by the expected loss [8, 9, 11, 12, 14, 22], to build a misspecified-aware regression scheme. However, these efforts focus on probabilistic regression settings, often with neural network models far from the underparametrized limit, meaning aleatoric, epistemic and misspecification errors must be considered jointly.

Masegosa [12] derived second order PAC-Bayes bounds for the generalization error minimized through a variational or ensemble ansatz. Morningstar *et al* [11] developed estimators to assess disagreement between the generalization error and expected loss, deriving specialized PAC bounds for theoretical guarantees. Lahlou *et al* [8] considered misspecification of neural networks, jointly training a minimum loss surrogate and an independent predictor of misspecification error. Lofti *et al* [14] considered misspecification in the context of deep model selection, deriving conditional likelihoods better aligned with the generalization error. As various definitions for the distinction between epistemic and misspecification uncertainties exist [13], we have defined (section 1.4) misspecification as non-aleatoric uncertainties which survive in the underparametrized limit, where the minimizer of PAC-Bayes bounds for the expected loss, which coincides with the posterior from Bayesian inference, predicts vanishing parameter uncertainties. Our epistemic uncertainties thus coincide with standard estimates from loss minimization.

2. Analysis of the generalization error in the near-deterministic limit

We consider the generalization error (6) in the near-deterministic limit, $\mathcal{G}_0[\pi] \equiv \lim_{\epsilon\to 0} \mathcal{G}[\pi]$. The integral over $\mathbf{Y}$ in (6) concentrates at $\mathbf{Y} = \mathcal{E}(\mathbf{X})$; using $\langle\ldots\rangle_{\mathcal{X}} \equiv \int\ldots\rho_{\mathcal{D}}(\mathbf{X})\mathrm{d}\mathbf{X}$ yields, using Laplace's method

(appendix C)

$$\mathcal{G}_0\left[\pi\right] = -\left\langle \ln\left|\rho_{\mathcal{M}}\left(\mathcal{E}\left(\mathbf{X}\right)|\mathbf{X}\right)\right|\right\rangle_{\mathcal{X}} = -\left\langle \ln\left|\int_{\mathcal{P}} \rho_{\mathcal{M}}\left(\mathcal{E}\left(\mathbf{X}\right)|\mathbf{X},\mathbf{\Theta}\right)\pi\left(\mathbf{\Theta}\right)d\mathbf{\Theta}\right|\right\rangle_{\mathcal{X}}. \tag{12}$$

Whilst $\mathbf{Y} = \mathcal{E}(\mathbf{X})$ is a unique maximum, to ensure $\min_{\mathbf{X}\in\mathcal{X}}\rho_{\mathcal{M}}(\mathcal{E}(\mathbf{X})|\mathbf{X}) > 0$ and thus avoid divergence in $\mathcal{G}_0[\pi]$ as $\epsilon \to 0$ we require $\pi(\mathbf{\Theta})$ has mass in every *pointwise optimal parameter set* (POPS)

$$\mathcal{P}_{\mathcal{M}}\left(\mathbf{X}\right) \equiv \left\{\mathbf{\Theta}\,|\,\mathcal{M}\left(\mathbf{X},\mathbf{\Theta}\right) = \mathcal{E}\left(\mathbf{X}\right)\right\}. \tag{13}$$

Any model with a constant term will ensure $\mathcal{P}_{\mathcal{M}}(\mathbf{X}) \neq \emptyset$, i.e. that the POPS is not empty, but under misspecification it clear that $\cap_{\mathbf{X}\in\mathcal{X}}\mathcal{P}_{\mathcal{M}}(\mathbf{X}) = \emptyset$, i.e. that no one parameter choice can be a member of all POPS. To express the final generalization error, we introduce a pointwise mass function

$$\mathrm{m}_{\mathcal{E}}\left(\mathbf{\Theta},\mathbf{X}\right) \equiv \int_{\mathcal{P}} \delta\left(\mathbf{\Theta}-\mathbf{\Theta}'\right)\rho_{\mathcal{M}}\left(\mathcal{E}\left(\mathbf{X}\right)|\mathbf{X},\mathbf{\Theta}'\right)d\mathbf{\Theta}', \tag{14}$$

which gives our first main result, a generalization error for deterministic surrogate models of

$$\mathcal{G}_0\left[\pi\right] \equiv -\left\langle \ln \int_{\mathcal{P}} \mathrm{m}_{\mathcal{E}}\left(\mathbf{\Theta},\mathbf{X}\right)\pi\left(\mathbf{\Theta}\right)d\mathbf{\Theta}\right\rangle_{\mathcal{X}}. \tag{15}$$

To avoid $1/\epsilon^2$ divergence in $\mathcal{G}_0[\pi]$ a valid $\pi(\mathbf{\Theta})$ must satisfy the 'POPS covering' constraint as $\epsilon \to 0$ of

$$\min_{\mathbf{X}\in\mathcal{X}}\max_{\mathbf{\Theta}\in\mathcal{P}_{\mathcal{M}}(\mathbf{X})}\pi\left(\mathbf{\Theta}\right) > \lim_{\epsilon\to 0}\exp\left(-c/\epsilon^2\right), \tag{16}$$

for any $c > 0$. Equation (16) is our second main result, a demonstration that any candidate minimizer of the generalization error must have probability mass in every POPS $\mathcal{P}_{\mathcal{M}}(\mathbf{X})$, whilst the minimizing distribution concentrates on regions where multiple POPS intersect. As shown above, the minimum loss solution $\pi_{\mathcal{L}}(\mathbf{\Theta})$ does not satisfy this requirement and thus $\mathcal{G}[\pi_{\mathcal{L}}] = \mathcal{O}(1/\epsilon^2)$. Importantly, when (16) is satisfied, the predictive distribution $\rho_{\mathcal{M}}(\mathbf{Y},\mathbf{X})$ covers every training data point, i.e. model predictions will envelope observations as $N/P \to \infty$. We now investigate two candidate distributions which satisfy (16), an $\mathcal{O}(N)$ ensemble ansatz and $\mathcal{O}(P)$ hypercube ansatz, finding approximate variational minima that allow for efficient deployment on high-dimensional datasets.

3. POPS-constrained ensemble and hypercube ansatz

The covering constraint (16) requires $\pi(\mathbf{\Theta})$ to have mass in the POPS of each training point. This condition must be satisfied whilst concentrating mass in as small a region in $\mathcal{P}$ as possible to minimize $\mathcal{G}_0$. To achieve this, we first select a set of pointwise fits $\mathbf{\Theta}_{\mathbf{X}}^* \in \mathcal{T}_E^*$ that are each POPS-constrained loss minimizers

$$\mathbf{\Theta}_{\mathbf{X}}^* \in \mathcal{T}_E^*, \quad \mathbf{\Theta}_{\mathbf{X}}^* \equiv \arg\min_{\mathbf{\Theta}_x\in\mathcal{P}_{\mathcal{M}}(\mathbf{X})}L\left(\mathbf{\Theta}_{\mathbf{X}}\right). \tag{17}$$

These pointwise fits will be closely clustered around the minimum loss solution $\mathbf{\Theta}_{\mathcal{L}}^*$, as measured by the Fisher metric $\mathcal{I}_{\mathcal{L}}^*$. In appendix D we explore a weighted ensemble ansatz

$$\pi_E^*\left(\mathbf{\Theta}\right) = \left\langle \mathrm{w}\left(\mathbf{X}\right)\delta\left(\mathbf{\Theta}-\mathbf{\Theta}_{\mathbf{X}}^*\right)\right\rangle, \tag{18}$$

where $\mathrm{w}(\mathbf{X}) > 0$. In particular, we show $\mathcal{T}_E^*$ naturally emerges when choosing pointwise fits $\mathbf{\Theta}_{\mathbf{X}} \in \mathcal{P}_{\mathcal{M}}(\mathbf{X})$ to minimize the ansatz expected loss $\mathcal{L}[\pi_E] = \langle \mathrm{w}(\mathbf{X})L(\mathbf{\Theta}_X)\rangle$.

The ensemble ansatz produces highly informative max/min bounds for any test point, providing an envelope for worst case errors, as we show in figure 1. However, even when employing the variational optimal reweighting $\mathrm{w}^*(\mathbf{X})$ (see appendix D.1), the ensemble ansatz systematically underestimates moments of the test error distribution, as shown in figure 2. In addition, both resampling and storage are $\mathcal{O}(N)$, which can become problematic for the $N/P \to \infty$ limit of interest. As a result, whilst further investigation of ensemble schemes and alternatives to the variational optimal reweighting $\mathrm{w}^*(\mathbf{X})$ (appendix D.1), we now look for an alternative ansatz beyond the ensemble approach.

For practical applications, the parameter distribution should ideally require at most $\mathcal{O}(P^2)$ effort for storage and $\mathcal{O}(P)$ for resampling. We have found this can be achieved by finding the minimal hypercube $\mathcal{H}(\mathcal{T}_E^*) \in \mathcal{P}$ which encompasses all members of the loss minimizing POPS-ensemble $\mathbf{\Theta}_{\mathbf{X}}^* \in \mathcal{T}_E^*$. As $\mathcal{H}(\mathcal{T}_E^*)$

intersects all POPS by construction, Equation (16) is satisfied by the hypercube ansatz uniform over $\mathcal{H}(\mathcal{T}_E^*)$, i.e.

$$\pi_{\mathcal{H}}^*(\boldsymbol{\Theta}) = V_{\mathcal{H}}^{-1}\delta\left(\boldsymbol{\Theta} \in \mathcal{H}\left(\mathcal{T}_E^*\right)\right), \tag{19}$$

where $V_{\mathcal{H}} = \int_{\mathcal{P}} \delta(\boldsymbol{\Theta} \in \mathcal{H}(\mathcal{T}_E^*)) \mathrm{d}\boldsymbol{\Theta}$. Crucially, resampling $\pi_{\mathcal{H}}(\boldsymbol{\Theta})$ requires only $\mathcal{O}(P)$ effort whilst giving a conservative estimate and bounding of test errors.

To determine $\mathcal{H}(\mathcal{T}_E^*)$ we use singular value decomposition on the POPS-ensemble $\mathcal{T}_E^*$ to obtain the right eigenmatrix $\mathbf{V} \in \mathbb{R}^{R \times P}$, $R \leqslant P$, $\mathbf{VV}^\top = \mathbb{I}_R$, requiring $\mathcal{O}(P^3)$ effort for evaluation. Uniformly sampling the full max/min hypercube of the projected ensemble $\tilde{\mathcal{T}}_E^* = \{\mathbf{V}\boldsymbol{\Theta}|\boldsymbol{\Theta} \in \mathcal{T}_E^*\}$ returns samples $\tilde{\boldsymbol{\Theta}}^\top \mathbf{V}$ of $\pi_{\mathcal{H}}^*$ for $\mathcal{O}(R)$ effort and $\mathcal{O}(RP)$ storage. We find this efficient hypercube posterior to give excellent bounding and prediction of test errors for $P \in [3, 1600]$, as shown in figures 2 and 3.

3.1. POPS-constrained loss minimization for linear models

Both the ensemble and hypercube ansatz require POPS-constrained loss minimizing fits $\boldsymbol{\Theta}_{\mathbf{X}}^* \in \mathcal{P}_{\mathcal{M}}(\mathbf{X})$. Practical application therefore requires an efficient means to perform this constrained minimization. Linear models use feature functions $\mathcal{F} : \mathcal{X} \to \mathcal{P} \times \mathcal{Y}$ to map inputs $\mathbf{X} \in \mathcal{X}$ to features $\mathbf{F}(\mathbf{X}) \in \mathcal{P} \times \mathcal{Y}$, giving $\mathcal{M}(\mathbf{X}, \boldsymbol{\Theta}) = \boldsymbol{\Theta} \cdot \mathbf{F}(\mathbf{X})$ with parameters $\boldsymbol{\Theta} \in \mathcal{P}$. Minimizing the PAC-Bayes bound (8) with a Gaussian prior $\pi_0(\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{\Theta}|\mathbf{0}, \boldsymbol{\Sigma}_0)$ gives a global loss minimizer $\boldsymbol{\Theta}_{\mathcal{L}}^*$ of

$$\mathbf{A} = \left[\boldsymbol{\Sigma}_{\mathcal{Y}}\boldsymbol{\Sigma}_0^{-1}/N + \langle \mathbf{F}(\mathbf{X})[\mathbf{F}(\mathbf{X})]^\top\rangle\right]_{\mathcal{X}}^{-1}, \quad \boldsymbol{\Theta}_{\mathcal{L}}^* = \mathbf{A}\langle \mathbf{F}(\mathbf{X})\mathcal{E}(\mathbf{X})\rangle_{\mathcal{X}}, \tag{20}$$

The epistemic covariance for small N/P can be incorporated in a variational PAC-Bayes scheme (appendix D.1) that will be developed further in future work. However, in the following, we concentrate on misspecification uncertainties in the limit $N/P \to \infty$, where the influence of $\boldsymbol{\Sigma}_0$ is negligible. For linear models the POPS-constraint is simply $\boldsymbol{\Theta}_{\mathbf{X}} \cdot \mathbf{F}(\mathbf{X}) \equiv \mathcal{E}(\mathbf{X})$. The loss minimizing pointwise fits $\boldsymbol{\Theta}_{\mathbf{X}}^*$ can be evaluated analytically via the efficient rank-one updates

$$\boldsymbol{\Theta}_{\mathbf{X}}^* = \boldsymbol{\Theta}_{\mathcal{L}}^* + \frac{\mathcal{E}(\mathbf{X}) - \boldsymbol{\Theta}_{\mathcal{L}}^* \cdot \mathbf{F}(\mathbf{X})}{\mathrm{h}(\mathbf{X})}\mathbf{AF}(\mathbf{X}), \tag{21}$$

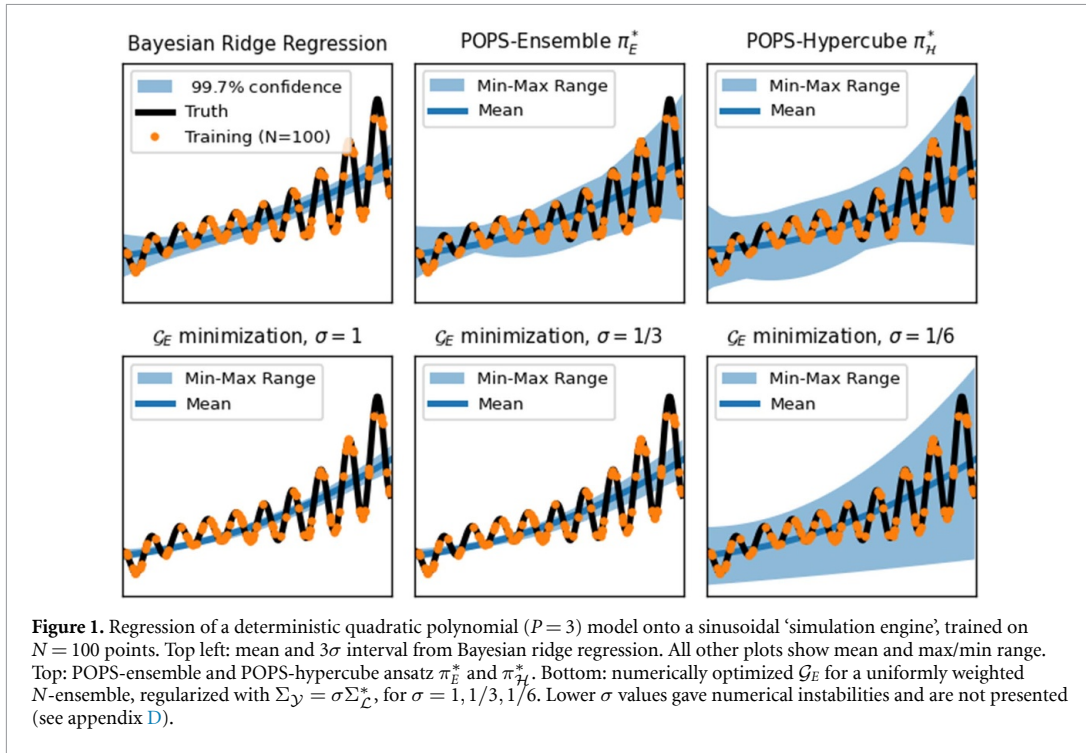where $\mathrm{h}(\mathbf{X}) = [\mathbf{F}(\mathbf{X})]^\top \mathbf{AF}(\mathbf{X})$ is the *leverage*, an outlier measure closely related to the distance of [28]. The minimum loss solution is the leverage-weighted centroid of pointwise optima, $\boldsymbol{\Theta}_{\mathcal{L}}^* = \langle \mathrm{h}(\mathbf{X})\boldsymbol{\Theta}_{\mathbf{X}}^*\rangle_{\mathcal{X}}$. Equation (21) is our third main result, a scheme to incorporate misspecification uncertainty into least-squares linear regression for minimal overhead, requiring only $\mathcal{O}(N)$ inner products over feature vectors, which can then be used to construct the ensemble or hypercube ansatz given above. Details of the implementation are provided in appendix E. In appendix F we compare the performance of our `POPSRegression` implementation against `BayesianRidge` from `sklearn`, finding an minimal overhead of around $2\times$ for fitting and essentially identical performance for prediction, even when returning max/min bounds in addition to standard deviations.

4. Numerical Experiments

In this section we apply our POPS approach to a variety of synthetic and real datasets, comparing our POPS-ensemble and POPS-hypercube ansatz to Bayesian inference and the minimization of the PAC-Bayes generalization error $\mathcal{G}_E$. We also discuss our approach in the context of existing methods for uncertainty quantification in atomistic machine learning and discuss how these parameter uncertainties can be propagated up scales to predict material properties with uncertainty [29].

4.1. Comparison to Bayesian ridge regression and minimization of $\mathcal{G}_E$

We applied Bayesian ridge regression and our POPS-ansatz to regress a quadratic polynomial against a sinusoidal function, the constant, linear and quadratic features giving $P = 3$ model parameters. Appendix G shows the same procedure for a model with $P = 6$ independent features. For these simple models it is possible to approximately minimize the generalization error $\mathcal{G}_E$ of some discrete model ensemble, providing we regularize with a finite aleatoric uncertainty $\boldsymbol{\Sigma}_{\mathcal{Y}} = \sigma\boldsymbol{\Sigma}_{\mathcal{L}}^*$. With increasing σ we have $\mathcal{G} \to \mathcal{L}$ and thus expect the ensemble to concentrate on $\boldsymbol{\Theta}_{\mathcal{L}}^*$. In the small σ regime of interest minimizing $\mathcal{G}_E$ becomes increasingly unstable, meaning that the final minimization is highly sensitive to initial conditions and frequently diverges, leading to conditioning issues. We found the same small σ instability when applying an ensemble approach from [12] which targets a second-order PAC-Bayes loss bound, designed for misspecified probabilistic models where σ can be large. These instabilities show existing misspecification-aware methods require

Figure 1. Regression of a deterministic quadratic polynomial ($P = 3$) model onto a sinusoidal 'simulation engine', trained on $N = 100$ points. Top left: mean and 3σ interval from Bayesian ridge regression. All other plots show mean and max/min range. Top: POPS-ensemble and POPS-hypercube ansatz π_E^* and $\pi_{\mathcal{H}}^*$. Bottom: numerically optimized $\mathcal{G}_E$ for a uniformly weighted N-ensemble, regularized with $\Sigma_{\mathcal{Y}} = \sigma \Sigma_{\mathcal{L}}^*$, for $\sigma = 1, 1/3, 1/6$. Lower σ values gave numerical instabilities and are not presented (see appendix D).
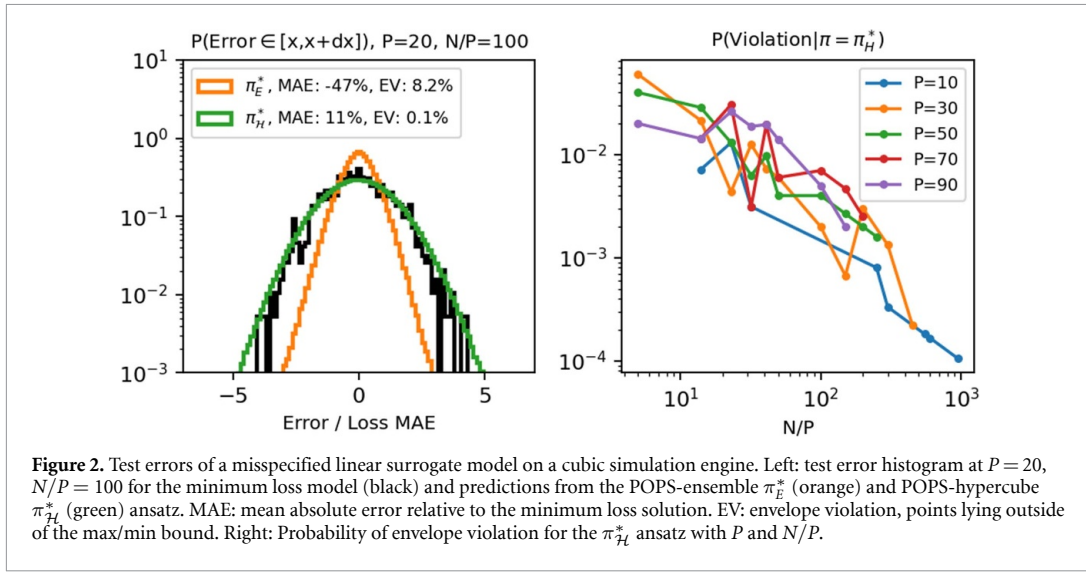
careful hyperparameter tuning and low model dimension to avoid vanishing gradients or other numerical issues. Our POPS-ansatz, in contrast, can easily sample $P = \mathcal{O}(1000)$ or higher dimensional data.

Figure 1 shows results for $\sigma = 1, 1/3, 1/6$, at the lower limit for which minimization converges. As expected from section 1.4, Bayesian ridge regression significantly underestimates test errors even at the 99.7% confidence level (three standard deviations) which as discussed above will worsen with increasing N/P. The max/min range of the numerical ensemble shows similar behavior for $\sigma = 1, 1/3$ whilst at $\sigma = 1/6$ test errors are bounded but significantly overestimated, which worsens as σ decreases (whilst stable). Importantly, in all cases our POPS ansatz π_E^* and $\pi_{\mathcal{H}}^*$ give excellent bounding of test errors and are stable in the near-deterministic limit $\sigma \to 0$. As expected, the hypercube ansatz $\pi_{\mathcal{H}}^*$ gave slightly more conservative bounds. In appendix G we provide a similar test of directly minimizing the generalization error for higher dimensional datasets, with $P = 6$ independent features. As for the polynomial case, we find that strong regularization is required for minimization to be stable, and in this limit the solution closely follows that of loss minimization. At weaker regularization, the direct minimization becomes systematically closer to the prediction of the POPS ansatz $\pi_{\mathcal{H}}^*$. These albeit simple results give strong evidence that our ansatz is indeed a stable and highly efficient means to find approximate minimizers of $\mathcal{G}$ that would otherwise be intractable.

4.2. Test errors in medium-dimensional linear regression

We now consider higher/medium dimensional problems, $P = 10 - 100$, where numerical minimization of $\mathcal{G}_E$ is not possible. The simulation engines were constrained to be misspecified from the linear surrogate models by combining features in a cubic polynomial. We also investigated the influence of adding a simple Gaussian noise term to simulates the presence of variability orthogonal to the chosen feature set. In all cases, independent test and training data were generated, typically with a 10:90 test:train split. However, all conclusions remain robust to varying the precise nature of data generation. In appendix G we provide examples showing this performance remains when varying the form of simulation engine to quadratic and sinusoidal functions or randomly sampling the linear feature coefficients from some predetermined subset of models.

As shown, the $\mathcal{O}(P)$ hypercube ansatz $\pi_{\mathcal{H}}^*$ provides an excellent prediction and bounding, of the test error distribution whilst the $\mathcal{O}(N)$ ensemble π_E^* underestimates moments of the error distribution and the max/min envelope bound. Importantly, the envelope violation rate of the hypercube ansatz $\pi_{\mathcal{H}}^*$ drops with N/P, as shown in figure 2 for a range of P and N/P. The simultaneous prediction and *bounding* of test errors is highly valuable for surrogate model deployment, allowing not only the prediction of expected errors but a robust assessment of worst case error scenarios.

Figure 2. Test errors of a misspecified linear surrogate model on a cubic simulation engine. Left: test error histogram at $P = 20$, $N/P = 100$ for the minimum loss model (black) and predictions from the POPS-ensemble π_E^* (orange) and POPS-hypercube $\pi_{\mathcal{H}}^*$ (green) ansatz. MAE: mean absolute error relative to the minimum loss solution. EV: envelope violation, points lying outside of the max/min bound. Right: Probability of envelope violation for the $\pi_{\mathcal{H}}^*$ ansatz with P and N/P.
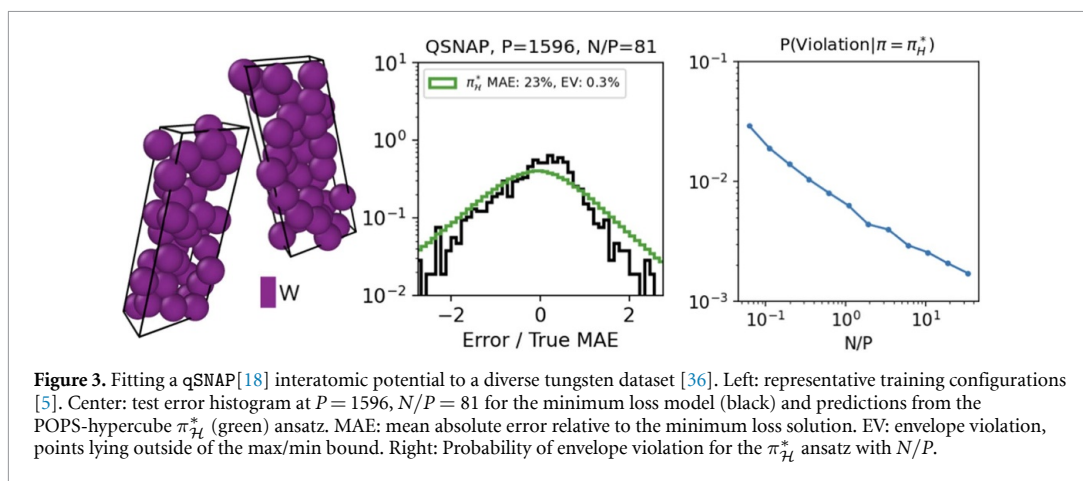
4.3. Application to machine learning interatomic potentials

We now demonstrate the performance of the method on challenging high-dimensional $P = \mathcal{O}(100 - 1000)$ regression tasks from atomic machine learning. Atomic-scale simulations of materials were traditionally broadly classified into two types:

(a) Empirical simulations with physically-motivated parametric models of interatomic interactions. These allow for fast simulations at a qualitative level of accuracy.

(b) First-principle quantum simulations obtained from approximate solutions of the full Schrodinger equation. These incur a high computational cost and exhibit poor scalability with system size.

The use of machine learning techniques in atomic simulation has blurred the lines between these two limits by promising near-quantum accuracy at a much more favorable computational cost and scaling [2]. The total energy of the system V is typically factored into a sum of per-atom contributions V_i, which are themselves parameterized in term of a feature vector that describes the local atomic environment $\mathbf{F}_i(\{\mathbf{x}_j | r_{ij} < r_c\})$, where r_c is a interaction range. Training a machine learning model proceeds by minimizing a squared error loss between reference energies and gradients. While certain approaches are systematically improvable, the computational cost associated with repeated evaluations can limit the complexity of the models that are selected for applications. Further, quantum calculations can be tightly converged, so that their intrinsic error (w.r.t. a fully converged calculation) can be made small compared to typical errors incurred by computationally-efficient models. This regime corresponds to the underparametrized deterministic setting considered here. Estimating the errors incurred by such machine learning models is highly sought after to assess the robustness of conclusions drawn from the simulations [30–33]. In the following, we explore the application of our POPS approach for a linear interatomic potential model $V_i = \boldsymbol{\Theta} \cdot \mathbf{F}_i$, with forces obtained by taking spatial gradients. Whilst we choose a particlar form of linear interatomic potential model, the presented approach is general to the wide range of available implementations [16, 18, 34, 35].

Our first application is to pure tungsten, a key material for nuclear fusion applications [16]. Atomic training data was generated using a recently introduced information-theoretic approach [5, 36] that aims to maximize an estimated feature entropy for a given set of feature functions, giving an an extremely (ideally, maximally) diverse training set. 10 000 atomistic configurations containing between 2 and 32 tungsten atoms were generated and characterized using Density Functional Theory [37], yielding 167 922 training data points in total (10 000 energies and 157 922 gradient components). Per-atom energies are expressed as a linear combination of $P = 1596$ so-called bispectrum components within the 'Quadratic' Spectral Neighborhood Analysis Potential (SNAP) formalism [18], known as qSNAP, which can (in spite of what the name suggests) be expressed in linear model form $V_i = \boldsymbol{\Theta} \cdot \mathbf{F}_i$ by defining the $\mathbf{F}_i$ as products of standard SNAP features. Numerical experiments were carried out by randomly sub-selecting N training points (either energies or gradient components), and by generating a discrete ensemble of N pointwise-optimal models according to equation (21), up to a maximal value of $N = 129,853$.

Figure 3. Fitting a `qSNAP`[18] interatomic potential to a diverse tungsten dataset [36]. Left: representative training configurations [5]. Center: test error histogram at $P = 1596$, $N/P = 81$ for the minimum loss model (black) and predictions from the POPS-hypercube $\pi_{\mathcal{H}}^*$ (green) ansatz. MAE: mean absolute error relative to the minimum loss solution. EV: envelope violation, points lying outside of the max/min bound. Right: Probability of envelope violation for the $\pi_{\mathcal{H}}^*$ ansatz with N/P.

The properties of the ensemble are then evaluated on a complementary set of 10 000 hold-out testing points. Error analysis is carried out in the context of the POPS-hypercube ansatz. Even on the full dataset, the analysis requires only on the order of 20 minutes on 1 CPU core. It is therefore extremely lightweight in addition to being easily parallelizable, enabling its use for extremely large datasets. The results show that the envelope violation on test data varies between 2% and 0.2% for $N/P = 0.06$ and 33, respectively, in qualitative agreement with the results shown in figure 2. This demonstrates that worst-case model errors are extremely well captured by the POPS-hypercube ensemble. As reported in the left panel of figure 3, these excellent bounds do not trivially result from the resampled error distribution being overly pessimistic, as the distribution of errors induced by a uniform parameter resampling within the POPS-hypercube provides an excellent predictor of the actual error distribution with respect to the minimum-loss model. The MAE estimated from the resampled models exceeds the measured value by only 23%, while providing a very good description of the overall error distribution.
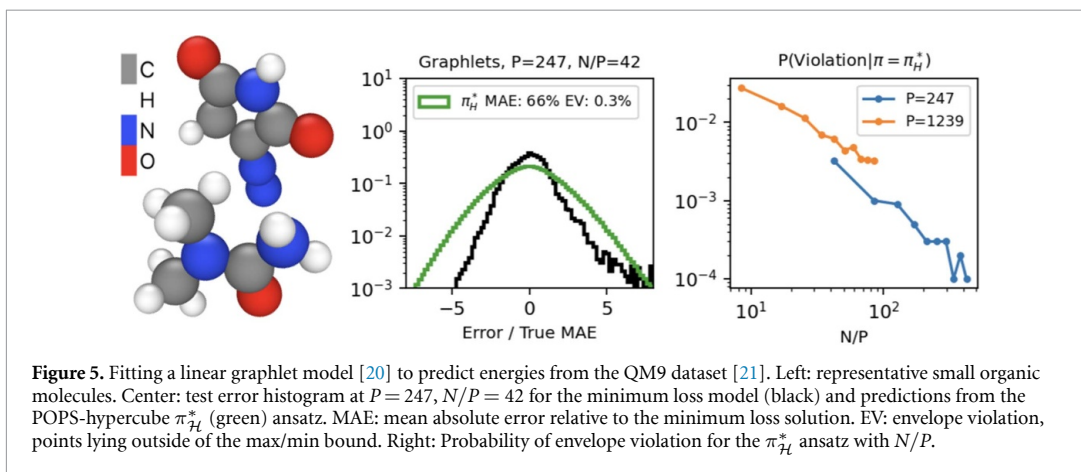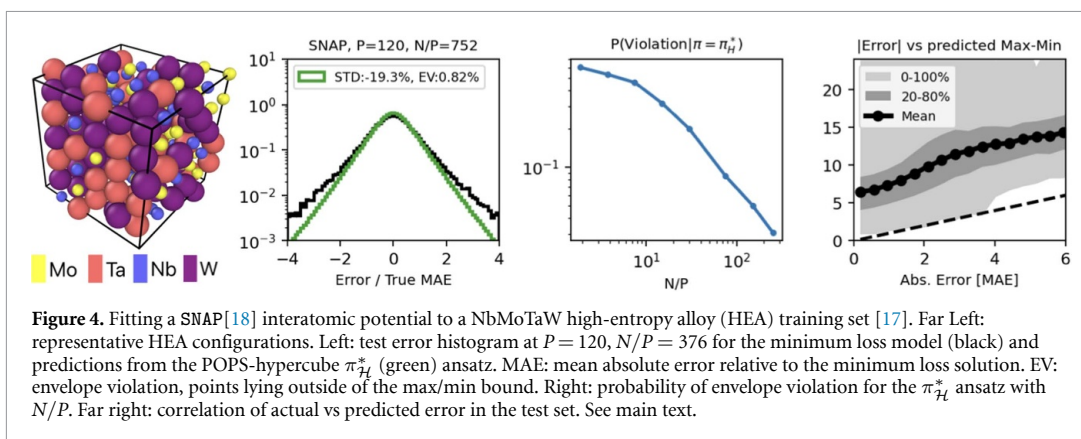
Our second test case fits a simple `SNAP` potential to *ab initio* data for the NbMoTaW high-entropy alloy from Li *et al* [17]. In this case, we use a simple form of multi-component descriptor that is linear in the number of species, and additionally choose a small set of descriptor functions, giving a total of $P = 120$ adjustable parameters to fit $N = 43,250$ force observations of this chemically complex system, with a further $N = 121248$ force observations held out for testing. As can be seen in figure 4, the POPS-hypercube shows excellent performance in bounding prediction errors and in estimating the test error distribution, albeit with a slight underestimation in the tails of the error distribution.

To investigate the ability of our POPS approach to capture these large-error predictions, we also plot a common metric for uncertainty quantification schemes, the correlation between predicted and observed errors. Whilst such a correspondence is not expected to be exact, as the observed error is a stochastic quantity, there is clear benfit in the ability to asses, with confidence, whether a given prediction is expected to have small or large error. In the present setting, an appropriate comparison is the absolute error of the loss minimizing model against the predicted by the hypercube posterior prediction. A well perfoming model should a) bound the absolute error from above, i.e. not predict low error when the actual error is high and b) have a max-min range which increases for points with large error. As shown in the right of figure 4, we see these properties are indeed observed in the HEA example. An in-depth study of our POPS apporach in active learning schemes will be the subject of future work.

4.4. Application to molecular cheminformatics models

As a final application, we apply the POPS scheme to a recently introduced linear graphlet model [20] for predicting molecular properties. Whilst superficially similar to the interatomic potential application, the graphlet features are quite distinct as they are designed to produce explanatory models that detail the contribution of different local bond environment to the total atomic energy rather than providing a smoothly differentiable interatomic potential. The graphlet features, which are essentially discrete counts of the different possible subgraphs of the molecular graph, have a much higher sparsity than the atomic descriptors employed in the previous examples.

Here, we consider the formation energies of around 50 000 small organic molecules from the well-known QM9 dataset of small organic molecules [21, 38]. Restricting feature extraction to graphlets of maximum

Figure 4. Fitting a `SNAP`[18] interatomic potential to a NbMoTaW high-entropy alloy (HEA) training set [17]. Far Left: representative HEA configurations. Left: test error histogram at $P = 120$, $N/P = 376$ for the minimum loss model (black) and predictions from the POPS-hypercube $\pi^*_{\mathcal{H}}$ (green) ansatz. MAE: mean absolute error relative to the minimum loss solution. EV: envelope violation, points lying outside of the max/min bound. Right: probability of envelope violation for the $\pi^*_{\mathcal{H}}$ ansatz with N/P. Far right: correlation of actual vs predicted error in the test set. See main text.

Figure 5. Fitting a linear graphlet model [20] to predict energies from the QM9 dataset [21]. Left: representative small organic molecules. Center: test error histogram at $P = 247$, $N/P = 42$ for the minimum loss model (black) and predictions from the POPS-hypercube $\pi^*_{\mathcal{H}}$ (green) ansatz. MAE: mean absolute error relative to the minimum loss solution. EV: envelope violation, points lying outside of the max/min bound. Right: Probability of envelope violation for the $\pi^*_{\mathcal{H}}$ ansatz with N/P.

order 3 or 4 gave a total of $P = 247$ or $P = 1239$ features, respectively. Once again, the formation energy model is linear in the graphlet features. As shown in figure 5, the POPS-hypercube again provides an excellent prediction and bounding of test errors, with the probability of envelope violation decreasing with increasing N/P. A detailed study of the use of our POPS approach for linear models in a wide range of applications will be the subject of a separate publication.

5. Comparison against existing approaches for uncertainty quantification and propagation in atomic simulation

As discussed in the previous section, uncertainty quantification schemes for machine learning interatomic potentials are actively sought after to assess the robustness of conclusions drawn from atomic simulations [30–33, 39–41]. Existing approaches for uncertainty quantification can be broadly categorized into two types, those that aim to directly estimate the error distribution of the surrogate model, and those (such as the POPS approach) that aim to capture model errors through the determination of parameter uncertainties.

Whilst direct error estimation is central to active learning workflows [35, 42, 43], these estimates can not typically be used to propagate uncertainty through the entire simulation engine to predict uncertainty on quantities of interest (e.g. formation energies, elastic constants, or relaxed structures). This is because each quantity of interest is an (implicit) function of the parameters of the chosen interatomic potential, and will in general vary with the parameters in a non-trivial manner due to the strong correlations in the simulation routine. For example, computing defect formation energies require comparing two very similar, and thus highly correlated simulations. A direct estimate of energy error for each configuration will in general significantly overestimate the error in the formation energy, as the correlation between the two structures is not accounted for.

As a result, schemes which estimate uncertainty in model parameters are desirable for uncertainty propagation. The propagation can be achieved if we can predict how simulations are affected by variation in

parameters, either through i) implicit differentiation methods [29] ii) reweighting of existing simulations using e.g. thermodynamic perturbation theory [15], or iii) direct resampling simulations by drawing samples from the parameter posterior distribution [39]. A detailed study combining the first two of these approaches (implicit differentiation and thermodynamic perturbation) with the POPS procedure will be the subject of a forthcoming work.

The POPS regression procedure presented above is a simple and efficient approach which returns uncertainties on the parameters of misspecified models. To the best of our knowledge, existing schemes for uncertainty quantification use as their objective function some form of loss, and thus cannot account for misspecification in parameter uncertainties of the model. In this sense, the POPS scheme is unique in its ability to robustly capture the parameter uncertainty of misspecified linear models and provide max-min bounds for predictions.

The most directly-comparable scheme to POPS is ensembling, which typically train an ensemble of models on random subsets of the training data, a process known as bagging [44]. Various forms of ensemble methods are a popular approach to quantify uncertainty in neural network models [45] and have been successfully applied to uncertainty quantification for neural network interatomic potentials [15, 40, 46].

It is known that bagging and ensemble methods more generally are most effective for high-capacity models in presence of appreciable aleatoric uncertainty [44]. In the low-noise regime, i.e. with weak aleatoric uncertainty, ensemble uncertainty estimates can be expected to significantly underestimate the true uncertainty, even when using high-capacity models such as neural networks. For low-capacity models, such as the underparametrized linear models considered here, bagging methods are not expected to produce useful uncertainty estimates. In most settings, loss-based uncertainty estimates require significant calibration [15, 40, 46], or the use of closely-related methods such as conformal prediction [39], which similarly require some form of calibration. In addition, all loss-based schemes should give vanishing parameter uncertainty in the large data limit, as discussed in section 1. Whilst bagging schemes for neural networks do provide finite parameter uncertainty, this is predominantly due to the multi-modal loss landscape, and has unclear relationship to the true uncertainty and generalization error of the model. Indeed, even the POPS-ensemble ansatz was found to underestimate moments of the error distribution, motivating the hypercube ansatz. An interesting direction for future work is to investigate how the POPS coverage criterion, Equation (16), can be used to systematically construct improved ensemble methods in the low-noise regime.

We end this section with a comment on the wider validity of the POPS method for atomic simulation. Clearly, the requirement to be in the underparametrised limit $N \gg P$, can in practice lead to a large amount of training data when P is large. In addition, this condition implicitly assumes that the model is not applied in a strongly extrapolative regime, which is formally true for any ensemble method [40]. Whilst we have not found this to be an issue in practical applications, this theoretical requirement is also common to ensemble methods [40]. In practical tests we do not find this to be problematic as the POPS scheme will typically give an overestimate of model uncertainty in regions of low training data density, which is a robust means to quantify extrapolation in high dimension [47]. Nevertheless, we recommend using one of the many extrapolation grades / outlier measures for high-dimensional linear models [34, 35, 47, 48] to ensure that the POPS conditions are satisfied.

6. Conclusion

In spite of its ubiquity in practical applications, misspecification is often ignored in Bayesian regression approaches based on expected loss minimization, which can lead to erroneous conclusions regarding parameter uncertainty quantification, surrogate model selection, and error propagation. In this work, we show that the important near-deterministic underparameterized regime is amenable to approximation by two formally simple ansatz that exploits the concentration of the cross-entropy minimizing ensemble onto pointwise optimal parameter sets (POPSs). For linear regression problems, this ensemble can be obtained extremely efficiently using rank-one perturbations of the expected loss solution, from which our optimally weighted POPS-ensemble or POPS-bounding hypercube ansatz can be efficiently obtained. Our POPS-constrained ansatz are shown to produce excellent parameter distributions which give accurate prediction of test error distributions and highly informative worst-case bounds on model predictions, in the important misspecified, near-deterministic regime where existing methods fail.

Importantly, POPS encode model error as misspecification uncertainty on model parameters, which is essential for the propagation of uncertainty to simulation results [15, 29]. The application of POPS parameter uncertainties to multi-scale uncertainty propagation will be the subject of forthcoming work. In

addition, a clear extension of the POPS approach is to non-linear models, for interatomic potentials [34, 46] or more widely in scientific machine learning. As the loss function is non-convex, the central challenge will be to efficiently generate an optimal set of POPS-constrained loss minimizers, which can be achieved analytically in the linear case.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/tomswinburne/POPS-Regression.git. Data will be available from 2024 August 20.

Acknowledgment

Appendix A. Glossary of terms

Table 1 contains a glossary of the key algebraic symbols used in the paper.

Table 1. Glossary of key algebraic symbols used in the paper.

Symbol	Description
$\mathcal{E}$	Simulation engine
$\mathcal{M}$	Surrogate model
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
$\mathbf{X}$	Input vector
$\mathbf{Y}$	Output vector
X	Input vector dimension
Y	Output vector dimension
$\mathbf{\Theta}$	Model parameters
$\mathcal{P}$	Parameter space
N	Number of training points
P	Number of model parameters
$\rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X})$	Output distribution of simulation engine
$\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta})$	Output distribution of surrogate model for given parameter choice
$\pi(\mathbf{\Theta})$	Posterior parameter distribution
$\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X})$	Model output distribution $\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta})$ averaged over $\pi(\mathbf{\Theta})$
$\mathcal{G}[\pi]$	Generalization error of $\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X})$ to $\rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X})$ under $\pi(\mathbf{\Theta})$
$\mathcal{L}[\pi]$	Expected loss of $\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X})$ to $\rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X})$ under $\pi(\mathbf{\Theta})$
$\hat{L}_N(\mathbf{\Theta})$	Empirical loss of $\rho_{\mathcal{M}}(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta})$ to $\rho_{\mathcal{E}}(\mathbf{Y}|\mathbf{X})$ with N training points
$\mathcal{I}_{\mathcal{L}}^*$	$X \times X$ Fisher information matrix from empirical loss
$\mathbf{\Sigma}_{\mathcal{Y}}$	$Y \times Y$ Aleatoric covariance matrix (assumed vanishing, determinant $\mathcal{O}(\epsilon^{2Y})$)
ϵ	Dimensionless measure of aleatoric uncertainty (assumed small)
$\mathbf{\Sigma}_{\mathcal{L}}^*$	$Y \times Y$ Error covariance matrix from loss minimization
$\pi_0(\mathbf{\Theta})$	Prior parameter distribution
$\pi_{\mathcal{L},N}(\mathbf{\Theta})$	Posterior parameter distribution from Bayesian inference
$\pi_E^*(\mathbf{\Theta})$	Posterior parameter distribution from POPS-constrained ensemble ansatz
$\pi_H^*(\mathbf{\Theta})$	Posterior parameter distribution from POPS-constrained hypercube ansatz

Appendix B. Model selection criteria for deterministic models

The Bayesian information criterion (BIC), used to discriminate between surrogate models, is derived by approximating (twice) the negative log evidence as $N/P \to \infty$, assuming a slowly varying prior $\pi_0(\boldsymbol{\Theta}) = \mathcal{O}(1)$. We have

$$\text{BIC} = -2\ln\left|\int \exp\left(-N\hat{L}(\boldsymbol{\Theta})\right)\pi_0(\boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{\Theta}\right| \to N\text{Tr}\left(\boldsymbol{\Sigma}_{\mathcal{L}}^*\boldsymbol{\Sigma}_{\mathcal{Y}}^{-1}\right) + \ln\left(\|\boldsymbol{\Sigma}_{\mathcal{Y}}\|^N N^P\|\boldsymbol{\mathcal{I}}_{\mathcal{L}}^*\|\right) + \mathcal{O}(1). \quad (\text{B1})$$

We note that the value of the PAC-Bayes loss bound (8) satisfies $\text{BIC} = 2N\hat{\mathcal{L}}_N[\pi_{\mathcal{L},N}] - 2C_0$, with C_0 model independent, as discussed in [22]. In standard derivations, $\boldsymbol{\Sigma}_{\mathcal{Y}}$ is set to the variational maximum likelihood solution $\boldsymbol{\Sigma}_{\mathcal{Y}} = \boldsymbol{\Sigma}_{\mathcal{L}}^*$, yielding the familar expression $\text{BIC} = P\ln N + N\ln\|\boldsymbol{\Sigma}_{\mathcal{L}}^*\| + \ln\|\boldsymbol{\mathcal{I}}_{\mathcal{L}}^*\|$, where model independent terms are neglected (principally, terms containing only N and the aleatoric covariance $\boldsymbol{\Sigma}_{\mathcal{Y}}$). However, for deterministic regression problems, $\boldsymbol{\Sigma}_{\mathcal{Y}}$ is fixed. Neglecting model-independent terms, this gives an alternative BIC for deterministic models of

$$BIC = N\text{Tr}\left(\boldsymbol{\Sigma}_{\mathcal{L}}^*\boldsymbol{\Sigma}_{\mathcal{Y}}^{-1}\right) + P\ln N + \ln\|\boldsymbol{\mathcal{I}}_{\mathcal{L}}^*\|. \quad (\text{B2})$$

In physical terms, this corresponds to an estimation of a partition function based on an harmonic approximation of the Hamiltonian around its minimum.

Appendix C. 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. The method applies to a function $f(x)$ which is twice differentiable and has a unique maximum in some closed interval $[a,b]$ which may be the entire real line. This encompasses all applications in this paper; the extension to multidimensional functions is straightforward. The method approximates the integral using a second-order Taylor expansion of $f(x)$ around the maximum of the function, which reduces the problem to a Gaussian integral. Laplace's method then reads

$$\lim_{N\to\infty}\int_a^b \exp[Nf(x)]\,\mathrm{d}x = \sqrt{\frac{2\pi}{N}}\sqrt{\frac{\mathrm{d}^2 f(x^*)}{\mathrm{d}x^2}}^{-1}\exp[Nf(x^*)], \quad x^* = \arg\max_{x\in[a,b]}f(x). \quad (\text{C1})$$

For further information we refer the reader to e.g. [49].

Appendix D. POPS-constrained ensemble ansatz

Selecting some set of pointwise fits and $\mathcal{T}_E = \{\boldsymbol{\Theta}_{\mathbf{X}} \in \mathcal{P}_{\mathcal{M}}(\mathbf{X})\}_{\mathcal{X}}$ and weights $\mathcal{W}_E = \{\mathrm{w}_E(\mathbf{X}) > 0\}_{\mathcal{X}}$ such that $\langle \mathrm{w}_E(\mathbf{X})\rangle_{\mathcal{X}} = 1$, the covering constraint is satisfied by

$$\pi_E(\boldsymbol{\Theta}|\mathcal{T}_E, \mathcal{W}_E) \equiv \langle \mathrm{w}_E(\mathbf{X})\mathcal{N}\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}_{\mathbf{X}}, [\boldsymbol{\Sigma}_0 + N\boldsymbol{\mathcal{I}}_{\mathcal{L}}^*]^{-1}\right)\rangle_{\mathcal{X}}, \quad (\text{D1})$$

For any choice of weights $\mathcal{W}_E$, an approximate optimal set $\boldsymbol{\Theta}_{\mathbf{X}}^* \in \mathcal{T}_E^*$ can be determined in a variational setting at finite N, using (D1) in the PAC-Bayes loss bound (8) for the loss $\mathcal{L}[\pi_E] = \langle \mathrm{w}_E(\mathbf{X})L(\boldsymbol{\Theta}_{\mathbf{X}})\rangle_{\mathcal{X}}$. As $P/N \to 0$, we show in the sub-section below that this reduces to the set of POPS-constrained loss minimizers

$$\boldsymbol{\Theta}_{\mathbf{X}}^* \in \mathcal{T}_E^*, \quad \boldsymbol{\Theta}_{\mathbf{X}}^* \equiv \arg\min_{\boldsymbol{\Theta}_x \in \mathcal{P}_{\mathcal{M}}(\mathbf{X})}L(\boldsymbol{\Theta}_{\mathbf{X}}) \quad (\text{D2})$$

The POPS-constrained loss-minimization ensures the fits will be as close as possible under the Fisher metric $\boldsymbol{\mathcal{I}}_{\mathcal{L}}^*$, such that $\boldsymbol{\Theta}_{\mathbf{X}}^* \to \boldsymbol{\Theta}_{\mathcal{L}}^*$ for specified models. Under misspecification, the generalization error of (D1) using $\mathcal{T}_E^*$ has the $P/N \to 0$ limit

$$\mathcal{G}_0[\pi_E] = -\left\langle \ln\langle \mathrm{w}_E(\mathbf{X}')\,\mathrm{m}_{\mathcal{E}}(\boldsymbol{\Theta}_{\mathbf{X}'}^*, \mathbf{X})\rangle_{\mathbf{X}' \in \mathcal{X}}\right\rangle_{\mathbf{X} \in \mathcal{X}}, \quad (\text{D3})$$

where $\mathrm{m}_{\mathcal{E}}(\boldsymbol{\Theta}, \mathbf{X})$ is the mass function introduced in (14). In general we expect a uniform weighting $\mathrm{w}_E(\mathbf{X}) = 1$ to be sub-optimal for generalization as the loss-minimizing fits $\boldsymbol{\Theta}_{\mathbf{X}}^*$ will be too close to $\boldsymbol{\Theta}_{\mathcal{L}}^*$.

To optimize $\mathrm{w}_E(\mathbf{X})$ we define a density $\rho^*_{\mathbf{X}}(\mathbf{X}') = \mathrm{m}_{\mathcal{E}}(\boldsymbol{\Theta}^*_{\mathbf{X}'}, \mathbf{X}) / \langle \mathrm{m}_{\mathcal{E}}(\boldsymbol{\Theta}^*_{\mathbf{X}'}, \mathbf{X}) \rangle_{\mathbf{X}' \in \mathcal{X}}$ of pointwise fits $\boldsymbol{\Theta}^*_{\mathbf{X}'}$ satisfying $|\mathcal{M}(\mathbf{X}, \boldsymbol{\Theta}^*_{\mathbf{X}'}) - \mathcal{E}(\mathbf{X})| = \mathcal{O}(\epsilon)$, then use Jensen's inequality to write the 'loss-like' bound

$$
\begin{aligned}
\mathcal{G}_0\left[\pi_E\right] &= \mathcal{G}_U - \left\langle \ln \left\langle \mathrm{w}_E(\mathbf{X}') \rho^*_{\mathbf{X}}(\mathbf{X}') \right\rangle_{\mathbf{X}' \in \mathcal{X}} \right\rangle_{\mathbf{X} \in \mathcal{X}} \\
&\leqslant \mathcal{G}_U - \left\langle \ln \left| \mathrm{w}_E(\mathbf{X}') \right| \left\langle \rho^*_{\mathbf{X}}(\mathbf{X}') \right\rangle_{\mathbf{X} \in \mathcal{X}} \right\rangle_{\mathbf{X}' \in \mathcal{X}},
\end{aligned}
\tag{D4}
$$

where $\mathcal{G}_U$ is the generalization error (D3) with uniform weighting $\mathrm{w}_E(\mathbf{X}) = 1$. Under the constraint $\langle \mathrm{w}_E(\mathbf{X}) \rangle_{\mathcal{X}} = 1$ it is simple to show this upper bound is minimized by

$$
\mathrm{w}^*_E(\mathbf{X}) \in \mathcal{W}^*_E, \quad \mathrm{w}^*_E(\mathbf{X}) = \lambda \left\langle \rho^*_{\mathbf{X}'}(\mathbf{X}) \right\rangle_{\mathbf{X}' \in \mathcal{X}}.
\tag{D5}
$$

where λ ensures $\langle \mathrm{w}^*_E(\mathbf{X}) \rangle_{\mathcal{X}} = 1$. Evaluation requires at most $\mathcal{O}(N^2)$ effort but only $\mathcal{O}(N)$ for storage and resampling. The final ensemble ansatz parameter posterior then reads

$$
\pi^*_E(\boldsymbol{\Theta}) \equiv \left\langle \mathrm{w}^*_E(\mathbf{X}) \delta(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*_{\mathbf{X}}) \right\rangle_{\mathcal{X}}.
\tag{D6}
$$

Equation (D6) is a variationally optimized ensemble ansatz for misspecified regression problems. The worst case of extreme misspecification, corresponding to non-intersecting POPS $\mathcal{P}_{\mathcal{M}}(\mathbf{X}') \cap \mathcal{P}_{\mathcal{M}}(\mathbf{X}) = \emptyset$, gives $\rho^*_{\mathbf{X}'}(\mathbf{X}) = \delta_{\mathbf{X}\mathbf{X}'}$ and thus uniform weights $\mathrm{w}^*_E(\mathbf{X}) = 1$. Interestingly, it is simple to show that optimal weights are also uniform in the large ϵ limit, as $\rho^*_{\mathbf{X}'}(\mathbf{X})$ is then uniform. For real systems we find $\mathrm{w}^*_E(\mathbf{X})$ is highly non-uniform and smallest for points near to $\boldsymbol{\Theta}^*_{\mathcal{L}}$, as expected. However, while this approaches improves the description of the parameter uncertainty compared to the uniform weight variant described in the main text, numerical tests reveal that moments of the predicted ensemble errors are still underestimated, as the 'loss like' bounds still bring all points too close to the minimum loss value. This observation (as well as the reduced evaluation cost) motivates the hypercube approach described in the text. However, the max/min bounds of the ensemble, which are insensitive to our choice of $\mathrm{w}(\mathbf{X})$, remain robust, as shown in figure 1.

D.1. Variational optimization
For finite N/P, we have access to training data $\mathcal{D}_N = \{\mathbf{Y}_i, \mathbf{X}_i\}_{i=1}^N$ to give averages over $\langle \ldots \rangle_N$, with some set of pointwise fits $\mathcal{T}_E = \{\boldsymbol{\Theta}_i\}_{i=1}^N$, $\boldsymbol{\Theta}_i \in \mathcal{P}_{\mathcal{M}}(\mathbf{X}_i)$. Our ensemble ansatz then reads

$$
\pi_E(\boldsymbol{\Theta}) = \left\langle \mathrm{w}_{E,i} \mathcal{N}\left(\boldsymbol{\Theta} | \boldsymbol{\Theta}_i, \sigma^2 \mathbb{I}\right) \right\rangle_N = \sum_{\mathbf{X}_i \in \mathcal{D}_N} \mathrm{w}_{E,i} \mathrm{w}_i \mathcal{N}\left(\boldsymbol{\Theta} | \boldsymbol{\Theta}_i, \sigma^2 \mathbb{I}\right),
\tag{D7}
$$

where $\mathrm{w}_{E,i}$ are the discrete ensemble weights and w_i are the training data weights.

The empirical loss $\hat{\mathcal{L}}_N[\pi_E]$, which contributes All π-dependent terms in the PAC-Bayes upper bound (8) then read

$$
\hat{\mathcal{L}}_N[\pi_E] = \int \hat{L}_N(\boldsymbol{\Theta}) \pi_E(\boldsymbol{\Theta}), \quad KL(\pi_E|\pi_0) = \int \ln \left| \frac{\pi_E(\boldsymbol{\Theta})}{\pi_0(\boldsymbol{\Theta})} \right| \pi_E(\boldsymbol{\Theta}) \, \mathrm{d}\boldsymbol{\Theta}.
\tag{D8}
$$

To proceed, we define a POPS-constrained variation $\delta_i \boldsymbol{\Theta}_i$ such that $\boldsymbol{\Theta}_i + \delta_i \boldsymbol{\Theta}_i \in \mathcal{P}_{\mathcal{M}}(\mathbf{X}_i)$, which gives a variation in π_E of

$$
\delta_i \pi_E(\boldsymbol{\Theta}) \equiv \frac{\mathrm{w}_{E,i} \mathrm{w}_i}{\sigma^2} \delta \boldsymbol{\Theta}_i^\top [\boldsymbol{\Theta} - \boldsymbol{\Theta}_i] \mathcal{N}\left(\boldsymbol{\Theta} | \boldsymbol{\Theta}_i, \sigma^2 \mathbb{I}\right).
\tag{D9}
$$

The upper bound (8) thus has a POPS-constrained variation

$$
\begin{aligned}
\delta_i\left(\hat{\mathcal{L}}_N[\pi_E] + \frac{1}{N} KL(\pi_E|\pi_0)\right) &= \frac{\mathrm{w}_{E,i}\mathrm{w}_i}{\sigma^2} \int \delta\boldsymbol{\Theta}_i^\top [\boldsymbol{\Theta} - \boldsymbol{\Theta}_i]\left(\hat{L}_N(\boldsymbol{\Theta}) + \frac{1}{N}\ln\left|\frac{\pi_E(\boldsymbol{\Theta})}{\pi_0(\boldsymbol{\Theta})}\right|\right) \mathcal{N}\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}_i,\sigma^2\mathbb{I}\right)\mathrm{d}\boldsymbol{\Theta}, \\
&= \mathrm{w}_{E_i}\mathrm{w}_i\delta\boldsymbol{\Theta}_i^\top \int \nabla_{\boldsymbol{\Theta}}\left(\hat{L}_N(\boldsymbol{\Theta}) + \frac{1}{N}\ln\left|\frac{\pi_E(\boldsymbol{\Theta})}{\pi_0(\boldsymbol{\Theta})}\right|\right)\mathcal{N}\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}_i,\sigma^2\mathbb{I}\right)\mathrm{d}\boldsymbol{\Theta}, \quad \text{(D10)}
\end{aligned}
$$

where we use integration by parts. Aside from the POPS constrained variation, this is a standard variational result which in principle allows for variational minimization of $\pi_E(\boldsymbol{\Theta})$ at finite N for both linear and non-linear models.

In the limit $\sigma \to 0$ we make the approximation that, for $\boldsymbol{\Theta} - \boldsymbol{\Theta}_i = \delta\boldsymbol{\Theta}$, $\pi_E(\boldsymbol{\Theta}_i + \delta\boldsymbol{\Theta}) \simeq (1/N)\mathcal{N}(\delta\boldsymbol{\Theta}|\mathbf{0}, \sigma^2\mathbb{I})$. With a Gaussian prior of zero mean and covariance $\boldsymbol{\Sigma}_0$ we have

$$
\int \nabla_{\boldsymbol{\Theta}} \ln\left|\frac{\pi_E(\boldsymbol{\Theta})}{\pi_0(\boldsymbol{\Theta})}\right| \mathcal{N}\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}_i,\sigma^2\mathbb{I}\right)\mathrm{d}\boldsymbol{\Theta} = \int \left(\frac{\boldsymbol{\Theta}-\boldsymbol{\Theta}_i}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Theta}\right)\mathcal{N}\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}_i,\sigma^2\mathbb{I}\right)\mathrm{d}\boldsymbol{\Theta} \to \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Theta}_i, \quad \sigma \to 0.
\tag{D11}
$$

In the present work we are primarily concerned with the underparametrized, or large data limit $P/N \to 0$, where the influence of the prior distribution is negligible, as demonstrated in the main text. In this regime, for small σ the variation of the upper bound (8) has a limiting form of

$$\delta_i \hat{\mathcal{L}}_N [\pi_E] = \mathrm{w}_{E_i} \mathrm{w}_i \delta_i \boldsymbol{\Theta}_i^\top \left[\nabla_{\boldsymbol{\Theta}} \hat{\mathcal{L}}_N (\boldsymbol{\Theta}_i) + \frac{1}{N} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Theta}_i \right] + \mathcal{O}\left(\sigma^2\right). \tag{D12}$$

This result shows the POPS-constrained variation of the PAC-Bayes loss upper bound converges to the gradient of the squared loss evaluated at a POPS-constrained parameter choice, to within a positive multiplicative constant. As a result, in the underparametrized limit, the variational optimal choice of of POPS-constrained parameter $\boldsymbol{\Theta}_{\mathbf{X}}$ will be given by

$$\boldsymbol{\Theta}_i^* \equiv \arg \min_{\boldsymbol{\Theta} \in \mathcal{P}_{\mathcal{M}}(\mathbf{X}_i)} L(\boldsymbol{\Theta}), \tag{D13}$$

in agreement with the result given in the main text.

Appendix E. Implementation of the ansatz

Algorithm 1. Find POPS-Constrained loss minimizers.

Require $\mathbf{X} \in \mathbb{R}^{N \times P}, \mathbf{y} \in \mathbb{R}^N, \boldsymbol{\Sigma}_0 \in \mathbb{R}^{P \times P}$ (prior covariance)

1: $\mathbf{C} \leftarrow \mathbf{X}^\top \mathbf{X} + \frac{\boldsymbol{\Sigma}_0}{N}$ ▷Compute regularized feature covariance matrix
2: $\mathbf{w} \leftarrow \mathrm{solve}(\mathbf{C}, \mathbf{X}^\top \mathbf{y})$ ▷Solve $\mathbf{Cw} = \mathbf{X}^\top \mathbf{y}$ for least squares solution
3: $\mathbf{e} \leftarrow \mathbf{y} - \mathbf{Xw}$ ▷Calculate residuals
4: $\mathbf{A} \leftarrow \mathrm{solve}(\mathbf{C}, \mathbf{X}^\top)$ ▷Solve $\mathbf{CA} = \mathbf{X}^\top$ for influence matrix
5: $\mathbf{h} \leftarrow \mathrm{diag}(\mathbf{XA})$ ▷Compute vector of leverage scores
6: $\mathbf{T} \leftarrow \mathbf{A} \odot (\mathbf{e} \oslash \mathbf{h})$ ▷Element-wise product and division for POPS-constrained corrections to $\mathbf{w}$
7: **Return** $\mathbf{w}, \mathbf{T}$ ▷Return the loss minimizer and POPS-constrained corrections

---

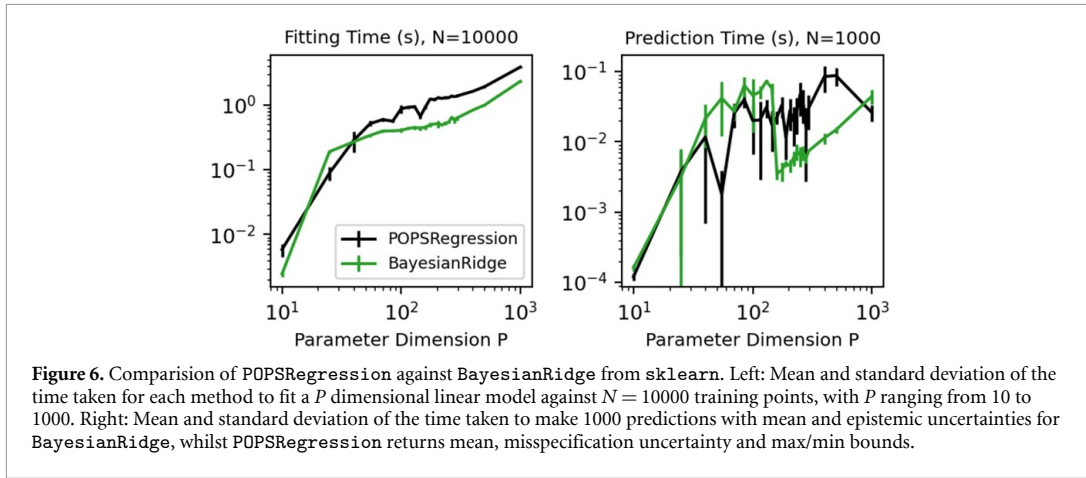**Algorithm 2.** Generate POPS-constrained hypercube samples.

Require $\mathbf{T} \in \mathbb{R}^{N \times P}$ (POPS-constrained corrections from Algorithm 1)

1: $\mathbf{U}, \mathbf{S}, \mathbf{V}^\top \leftarrow \mathrm{SVD}(\mathbf{T})$ ▷Perform SVD on T
2: $R \leftarrow \mathrm{rank}(\mathbf{S})$ ▷Determine rank of S
3: $\mathbf{V}_R \leftarrow \mathbf{V}[:, :R]$ ▷Keep only the first R right singular vectors
4: $\tilde{\mathbf{T}} \leftarrow \mathbf{T}\mathbf{V}_R$ ▷Project T onto the singular vectors
5: $\mathbf{l} \leftarrow \min(\tilde{\mathbf{T}}, \mathrm{axis} = 0)$ ▷Find min bounds of projections
6: $\mathbf{u} \leftarrow \max(\tilde{\mathbf{T}}, \mathrm{axis} = 0)$ ▷Find max bounds of projections
7: **for** $i = 1$ to M **do**   ▷Generate M samples
8:     $\mathbf{x} \leftarrow \mathrm{UniformRandom}(\mathbf{l}, \mathbf{u})$ ▷Sample uniformly within bounds
9:     $\mathbf{T}_{\mathrm{sample}, i} \leftarrow \mathbf{x}\mathbf{V}_R^\top$ ▷Project back to original basis
10: **end for**
11: **Return** $\mathbf{T}_{\mathrm{sample}}$ ▷Return M hypercube samples

---

A slightly modified implementation of this POPS-constrained linear regression algorithm, wrapping the Scikit-learn `BayesianRidge` class [19], is available on GitHub at https://github.com/tomswinburne/POPS-Regression.git. It can be installed (requires Scikit-learn) via pip: `pip install POPSRegression`.

Appendix F. Performance comparison against Bayesian Ridge

To demonstrate the performance of `POPSRegression` against `BayesianRidge` from `sklearn`, we fit against the medium-dimensional cubic 'simulation engine' described in the main text, with P ranging from 10 to 1000 and $N = 10000$. Figure 6 shows `POPSRegression` incurs a minimal ($2\times$) overhead over `BayesianRidge`, whilst existing misspecification aware schemes do not converge in the low-noise regime.
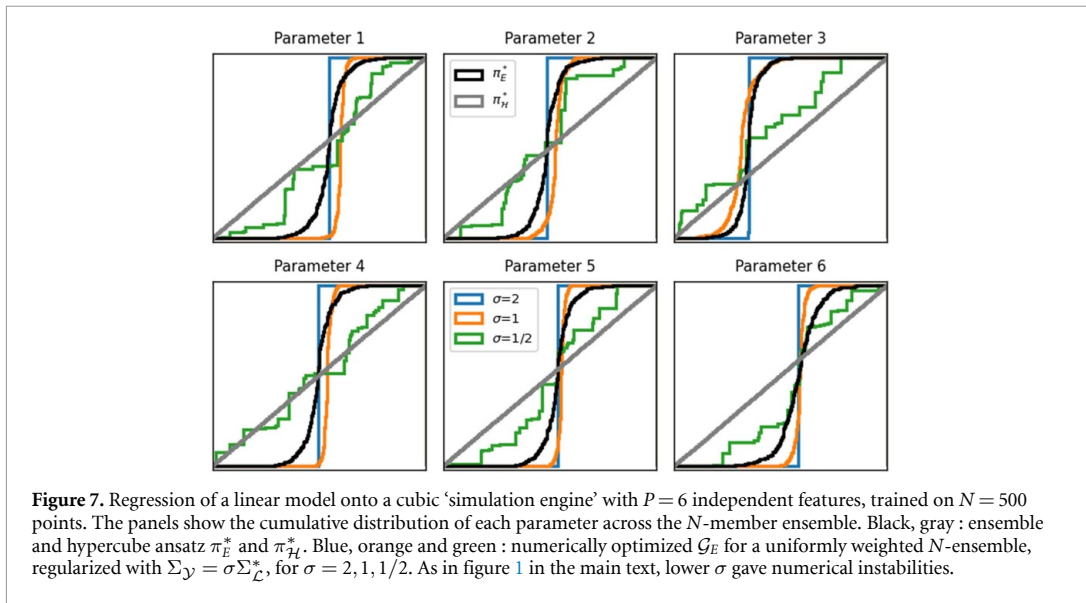
Figure 6. Comparision of `POPSRegression` against `BayesianRidge` from `sklearn`. Left: Mean and standard deviation of the time taken for each method to fit a P dimensional linear model against $N = 10000$ training points, with P ranging from 10 to 1000. Right: Mean and standard deviation of the time taken to make 1000 predictions with mean and epistemic uncertainties for `BayesianRidge`, whilst `POPSRegression` returns mean, misspecification uncertainty and max/min bounds.

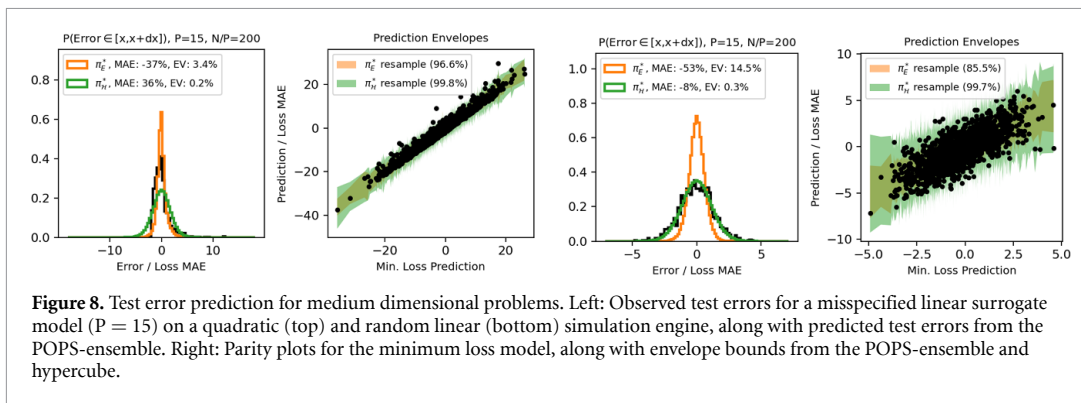## Appendix G. Additional plots of test error prediction

G.1. Low dimensional models

We applied the same procedure as for the polynomical example to a model system with $P = 6$ independent features to simulate more realistic datasets. As before, the numerical minimization routine could only treat a small range of σ values. Figure 7 shows the resulting parameter distributions. At large $\sigma = 2$, the minimizer quickly concentrates all ensemble members on the minimum loss solution; at $\sigma = 1$ the ensemble distribution is notably sharper than the POPS-ensemble π_E^*, whilst for low sigma $\sigma = 1/2$ the minimization only has marginal stability, but becomes closer to the POPS-bounding hypercube $\pi_{\mathcal{H}}^*$. In addition, a minimizer initialized with the POPS-ensemble did not change appreciably. These albeit simple results give strong evidence that our ansatz is indeed a stable and highly efficient means to find approximate minimizers of $\mathcal{G}$ that would otherwise be intractable.

G.2. Medium dimensional models

The plots below are identical in presentation to the top row of figures in the main text, but for the quadratic and random linear simulation engines described in the main text.

Figure 7. Regression of a linear model onto a cubic 'simulation engine' with $P = 6$ independent features, trained on $N = 500$ points. The panels show the cumulative distribution of each parameter across the N-member ensemble. Black, gray : ensemble and hypercube ansatz π_E^* and $\pi_{\mathcal{H}}^*$. Blue, orange and green : numerically optimized $\mathcal{G}_E$ for a uniformly weighted N-ensemble, regularized with $\Sigma_{\mathcal{Y}} = \sigma \Sigma_{\mathcal{L}}^*$, for $\sigma = 2, 1, 1/2$. As in figure 1 in the main text, lower σ gave numerical instabilities.

Figure 8. Test error prediction for medium dimensional problems. Left: Observed test errors for a misspecified linear surrogate model (P = 15) on a quadratic (top) and random linear (bottom) simulation engine, along with predicted test errors from the POPS-ensemble. Right: Parity plots for the minimum loss model, along with envelope bounds from the POPS-ensemble and hypercube.

ORCID iDs

Thomas D Swinburne ● https://orcid.org/0000-0002-3255-4257
Danny Perez ● https://orcid.org/0000-0003-3028-5249

References

[1] Alizadeh R, Allen J K and Mistree F 2020 *Res. Eng. Des.* **31** 275
[2] Deringer V L, Caro M A and Csányi G 2019 *Adv. Mater.* **31** 1902765
[3] Lapointe C, Swinburne T D, Thiry L, Mallat S, Proville L, Becquart C S and Marinica M-C 2020 *Phys. Rev. Mater.* **4** 063802
[4] Nyshadham C, Rupp M, Bekker B, Shapeev A V, Mueller T, Rosenbrock C W, Csányi G, Wingate D W and Hart G L 2019 *npj Comput. Mater.* **5** 51
[5] Montes de Oca Zapiain D, Wood M A, Lubbers N, Pereyra C Z, Thompson A P and Perez D 2022 *npj Comput. Mater.* **8** 189
[6] Bonatti C, Berisha B and Mohr D 2022 *Int. J. Plast.* **158** 103430
[7] Kudela J and Matousek R 2022 *Soft Comput.* **26** 13709
[8] Lahlou S, Jain M, Nekoei H, Butoi V I, Bertin P, Rector-Brooks J, Korablyov M and Bengio Y 2021 (arXiv:2102.08501)
[9] Psaros A F, Meng X, Zou Z, Guo L and Karniadakis G E 2023 *J. Comput. Phys.* **477** 111902
[10] Alquier P 2021 (arXiv:2110.11216)
[11] Morningstar W R, Alemi A and Dillon J V 2022 *International Conference on Artificial Intelligence and Statistics* (PMLR) pp 8270–98
[12] Masegosa A 2020 *Advances in Neural Information Processing Systems* vol **33** p 5479
[13] Kato Y, Tax D M and Loog M 2022 (arXiv:2210.16938)
[14] Lotfi S, Izmailov P, Benton G, Goldblum M and Wilson A G 2022 *International Conference on Machine Learning* (PMLR) pp 14223–47
[15] Imbalzano G, Zhuang Y, Kapil V, Rossi K, Engel E A, Grasselli F and Ceriotti M 2021 *J. Chem. Phys.* **154** 074102
[16] Goryaeva A M, Dérès J, Lapointe C, Grigorev P, Swinburne T D, Kermode J R, Ventelon L, Baima J and Marinica M-C 2021 *Phys. Rev. Mater.* **5** 103803
[17] Li X-G, Chen C, Zheng H, Zuo Y and Ong S P 2020 *npj Comput. Mater.* **6** 70
[18] Wood M A and Thompson A P 2018 *J. Chem. Phys.* **148** 241721
[19] Pedregosa F *et al* 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825
[20] Tynes M, Taylor M G, Janssen J, Burrill D J, Perez D, Yang P and Lubbers N 2024 *Digit. Discovery* **3** 1980
[21] Ruddigkeit L, van Deursen R, Blum L C and Reymond J-L 2012 *J. Chem. Inf. Model.* **52** 2864
[22] Germain P, Bach F, Lacoste A and Lacoste-Julien S 2016 *Advances in Neural Information Processing Systems* vol **29**
[23] Bhat H S and Kumar N 2010 *School of Natural Sciences* (University of California) vol 99 p 58
[24] von Toussaint U 2011 *Rev. Mod. Phys.* **83** 943
[25] Gibbs J W 1902 *Elementary Principles in Statistical Mechanics* (Charles Scribner's Sons)
[26] Hoeffding W 1994 *The Collected Works of Wassily Hoeffding* 409
[27] Shalaeva V, Esfahani A F, Germain P and Petreczky M 2020 *Proc. of the AAAI Conf. on Artificial Intelligence* vol 34 pp 5660–7
[28] Mahalanobis P C 1936 *Procedings of the National Institute of Science of India* (National Institute of Science of India)
[29] Maliyov I, Grigorev P and Swinburne T D 2024 (arXiv:2407.02414)
[30] Wen M and Tadmor E B 2020 *npj Comput. Mater.* **6** 124
[31] Li Y, Xiao W and Wang P 2018 *Asme International Mechanical Engineering Congress and Exposition* vol 52170 (American Society of Mechanical Engineers) p V012T11A030
[32] Gabriel J J, Paulson N H, Duong T C, Tavazza F, Becker C A, Chaudhuri S and Stan M 2021 *JOM* **73** 149
[33] Wan S, Sinclair R C and Coveney P V 2021 *Phil. Trans. R. Soc.* A **379** 20200082
[34] Lysogorskiy Y *et al* 2021 *npj Comput. Mater.* **7** 1
[35] Podryabinkin E V and Shapeev A V 2017 *Comput. Mater. Sci.* **140** 171
[36] Karabin M and Perez D 2020 An entropy-maximization approach to automated training set generation for interatomic potentials *J. Chem. Phys.* **153** 094110
[37] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[38] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 *Sci. Data* **1** 140022
[39] Best I R, Sullivan T J and Kermode J R 2024 *J. Chem. Phys.* **161** 064112
[40] Kellner M and Ceriotti M 2024 Uncertainty quantification by direct propagation of shallow ensembles *Mach. Learn.: Sci. Technol.* **5** 035006
[41] Thaler S, Doehner G and Zavadlav J 2023 *J. Chem. Theory Comput.* **19** 4520
[42] Li Z, Kermode J R and De Vita A 2015 *Phys. Rev. Lett.* **114** 096405

[43] Vandermause J, Torrisi S B, Batzner S, Xie Y, Sun L, Kolpak A M and Kozinsky B 2020 *npj Comput. Mater.* **6** 20

[44] Breiman L 1996 *Mach. Learn.* **24** 123

[45] Abe T, Buchanan E K, Pleiss G, Zemel R and Cunningham J P 2022 *Advances in Neural Information Processing Systems* vol **35** p 33646

[46] Vita J A, Samanta A, Zhou F and Lordi V 2024 (arXiv:2402.00853)

[47] Zeni C, Anelli A, Glielmo A and Rossi K 2022 *Phys. Rev.* B **105** 165141

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

[49] Wong R 2001 *Asymptotic Approximations of Integrals* (SIAM)