DERIVATIVE-AWARE UNCERTAINTY QUANTIFICATION FOR EQUIVARIANT MACHINE-LEARNED INTERATOMIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty quantification (UQ) for machine-learned interatomic potentials (MLIPs) is essential considering potential system distribution shift and data scarcity. It faces unique challenges as MLIPs have to respect physics principles: energies must respect symmetry, forces are derivatives of energy, and posterior samples should ideally remain physically valid potentials. We propose *derivative-aware subnetwork Laplace*, a *post-hoc* Bayesian UQ framework for equivariant MLIPs. Starting from a pre-trained deterministic model, we select to Bayesianize a compact force-informed subnetwork by learning a Gaussian posterior which incorporates both energy and force Jacobians. The resulting global posterior induces coherent sampled energy surfaces and force fields, while preserving invariance, equivariance, and conservative structure when applied within an admissible equivariant parameterization. We evaluate DASL on controlled and molecular benchmarks under distribution shift, showing that derivative-aware posterior geometry improves force uncertainty without retraining or ensemble.

1 INTRODUCTION

Machine-learned interatomic potentials (MLIPs) have emerged as a powerful surrogate for quantum-mechanical calculations, enabling atomistic simulations at substantially reduced computational cost (Behler & Parrinello, 2007; Bartók et al., 2010; Deringer et al., 2019; Unke et al., 2021). Modern equivariant neural potentials can achieve high accuracy while respecting the geometric symmetries of molecular systems (Schütt et al., 2021; Batatia et al., 2022; Musaelian et al., 2023; Batzner et al., 2022). Nevertheless, an MLIP remains an approximation learned from finite data and can become unreliable when queried on configurations that are poorly represented during training (Vandermause et al., 2020; Best et al., 2024). This is particularly important when predicted forces drive the system dynamic simulation and errors may accumulate as trajectories move away from familiar configurations. Reliable deployment of MLIPs therefore requires not only accurate point predictions, but also uncertainty estimates that indicate when the learned potential should not be trusted (Podryabinkin & Shapeev, 2017; Schran et al., 2020).

Uncertainty quantification for MLIPs (Lakshminarayanan et al., 2017; Wen & Tadmor, 2020; Hu et al., 2022; Tan et al., 2023; Thaler et al., 2023; Xie et al., 2023; Williams et al., 2024; Willow et al., 2025; Coscia et al., 2025; Ho et al., 2026) has additional structure that is absent from generic regression problems. An MLIP represents a scalar potential-energy surface, while forces are obtained as its spatial derivatives. At the same time, the energy is invariant to the relevant physical symmetries and the force field transforms equivariantly. Moreover, the energy is invariant to the relevant physical symmetries, while the force field transforms equivariantly. A suitable uncertainty model should preserve these properties rather than treating energy and force uncertainties as unrelated quantities. We therefore view MLIP uncertainty as uncertainty over physically admissible potential-energy surfaces: symmetry-related configurations should have consistently transformed predictive distributions, and energy and force uncertainty should arise from a shared stochastic energy function through the derivative relation above. From a function-space perspective, these requirements directly constrain the covariance structure of the uncertain potential.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

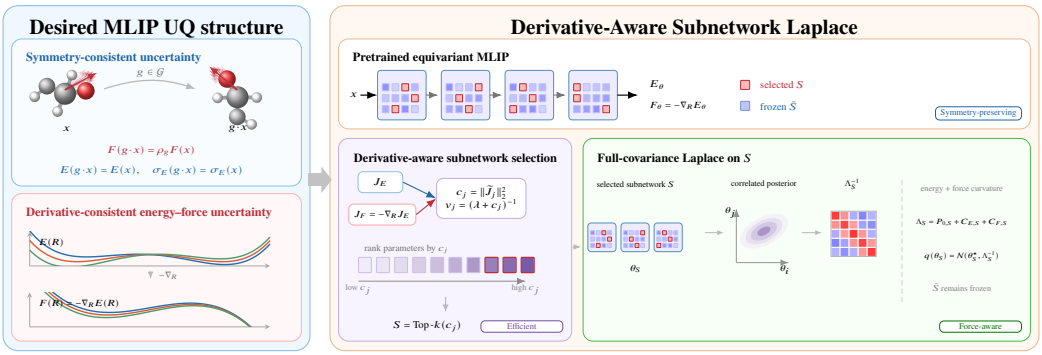

Figure 1: **Conceptual illustration of DASL.** Physically structured uncertainty for MLIPs should preserve molecular symmetries and couple energy and force uncertainty through the conservative relation $F = -\nabla_R E$. DASL realizes these requirements *post hoc* on a pretrained equivariant MLIP by selecting a compact set of derivative-relevant parameters and constructing a full-covariance Laplace posterior over the selected subnetwork.

The second practical challenge is computational overhead by UQ. Many MLIPs are pretrained (Zhu et al., 2023), and obtaining uncertainty estimates should ideally not require training multiple additional potentials or performing Bayesian inference over all network parameters. This consideration is increasingly important for large equivariant architectures and pretrained foundation potentials. We therefore seek a *post-hoc* UQ method that can be attached to an existing MLIP while leaving its deterministic predictor unchanged and preserving the physical structure encoded by the original architecture.

We introduce Derivative-Aware Subnetwork Laplace (DASL), a *post-hoc* Bayesian approximation for pretrained equivariant MLIPs. DASL places a Gaussian posterior over only a compact subset of model parameters while fixing the remainder at their pretrained values, making full-covariance local inference substantially more tractable than Bayesianizing the entire network. The stochastic subnetwork is selected using energy–force sensitivities, and its Laplace curvature incorporates both energy and force Jacobians. Under local linearization, the resulting parameter posterior induces a finite-rank Gaussian process in function space. This view connects the practical DASL construction back to the desired MLIP uncertainty structure: the induced covariance inherits the symmetry of the equivariant model and couples energy and force uncertainty through derivatives of the same stochastic potential.

Our contributions are:

- Structured function-space view of MLIP uncertainty: We formulate two requirements for physics-principled MLIP UQ: the predictive covariance should respect the symmetries of the underlying potential, and energy and force uncertainty should be derivative-consistent components of a shared stochastic energy function.
- Derivative-aware *post-hoc* Bayesian inference: We introduce DASL, which equips a pretrained equivariant MLIP with a compact Gaussian subnetwork posterior without retraining the underlying model. We derive a derivative-aware subnetwork-selection criterion and construct the local Laplace posterior using both energy and force sensitivities.
- Empirical MLIP UQ evaluation: Through controlled and molecular benchmarks, we study how derivative information, subnetwork selection, and posterior size affect uncertainty quality, and compare DASL with alternative post-hoc and retraining-based UQ approaches.

2 Function-Space Structure for MLIP Uncertainty

Uncertainty quantification for machine-learned interatomic potentials should respect the physical structure of the predicted potential-energy surface. In particular, not only must MLIPs adhere to physical domain symmetries, but their quantified uncertainties must also satisfy equivariance constraints and maintain exact differential force–energy consistency. We first study MLIP uncertainty in function space using a Gaussian-process perspective to characterize how these physical constraints shape the covariance between predictions, soliciting two key requirements: uncertainty should trans-

form consistently across symmetry-related configurations, and energy and force uncertainties should arise from a shared latent stochastic energy function through derivative covariance structure.

2.1 Problem Setup and UQ in Function-Space

Let $x = (R, Z)$ denote an atomic configuration, where R collects the atomic coordinates and Z the atomic species. An MLIP parameterized by θ predicts a scalar potential energy $E_\theta(x)$, with the corresponding force field determined by $F_\theta(x) = -\nabla_R E_\theta(x)$. A trained model $\theta^\star$ provides a single estimate of the potential-energy surface and its induced force field, but does not by itself quantify uncertainty in these predictions.

For uncertainty quantification, the object of interest is a distribution over possible energy functions. We represent an uncertain potential by a stochastic energy function $E(x)$ using a Gaussian-process,

$$E \sim \mathcal{GP}(m_E, k_E)\,,$$

with mean $m_E(x) = \mathbb{E}[E(x)]$ and covariance kernel $k_E(x, x') = \mathrm{Cov}[E(x), E(x')]$. The kernel $k_E(x, x')$ describes how uncertainty is correlated across configurations. For molecular potentials, this covariance cannot be arbitrary: it should inherit the symmetry of the underlying physical model and remain consistent with the relation $F = -\nabla_R E$.

2.2 Symmetry-Consistent Posterior Covariance

Let $g \in \mathcal{G}$ denote a physical symmetry acting on an atomic configuration and define the joint energy–force output

$$\Phi(x) = \begin{bmatrix} E(x) \\ F(x) \end{bmatrix}.$$

For an invariant energy and equivariant force field, $\Phi(g \cdot x) = T_g \Phi(x)$, where T_g acts trivially on the scalar energy and according to the corresponding representation on the forces. A physically consistent uncertainty model should inherit the same structure. In particular, its covariance $K_\Phi(x, x') = \mathrm{Cov}[\Phi(x), \Phi(x')]$ should satisfy

$$K_\Phi(g \cdot x, g \cdot x') = T_g K_\Phi(x, x') T_g^\top.$$

Thus, symmetry constrains not only the predictive mean but also the uncertainty: symmetry-related configurations should have consistently transformed predictive distributions rather than unrelated uncertainty estimates.

2.3 Derivative-Consistent Energy–Force Posterior Covariance

Energy and force uncertainty should likewise arise from a shared stochastic potential rather than from separate uncertainty models. Defining $F(x) = -\nabla_R E(x)$, and assuming that k_E is sufficiently differentiable, the joint energy–force covariance is completely determined by the energy kernel:

$$K_\Phi(x, x') = \begin{bmatrix} k_E(x, x') & -\nabla_{R'} k_E(x, x') \\ -\nabla_R k_E(x, x') & \nabla_R \nabla_{R'}^\top k_E(x, x') \end{bmatrix}.$$

Hence, energies and forces are different observations of the same latent stochastic energy function. Energy observations constrain values of this function, while force observations additionally constrain its local derivatives and can therefore provide information that is complementary to energy values alone.

These considerations solicit two requirements for physically structured MLIP UQ: the function-space covariance should respect molecular symmetries, and energy and force uncertainty should be coupled through derivatives of a common stochastic potential. In Section 3, we introduce Derivative-Aware Subnetwork Laplace (DASL) as a computationally tractable *post-hoc* realization of these principles for pretrained equivariant neural potentials.

3 DERIVATIVE-AWARE SUBNETWORK LAPLACE

Section 2 described two function-space requirements for physically structured MLIP uncertainty: the predictive covariance should inherit the symmetries of the molecular potential, and energy and force uncertainty should be coupled through derivatives of a shared stochastic energy function. We now introduce our *post-hoc* Bayesian approximation solution, DASL, which realizes these principles for a pretrained equivariant MLIP. DASL retains the original equivariant architecture, places uncertainty only over a compact subset of parameters, and constructs the local posterior using both energy and force information.

3.1 POST-HOC BAYESIANIZATION OF A PRETRAINED EQUIVARIANT MLIP

Let $\theta^\star$ denote the pretrained MAP parameters of an equivariant MLIP. Rather than retraining the model or introducing an ensemble, DASL constructs a local posterior around this solution while preserving the deterministic backbone. We partition the parameters as $\theta = (\theta_S, \theta_{\bar{S}})$, where S denotes the stochastic subnetwork and $\bar{S}$ its complement. The parameters outside the subnetwork are fixed at their pretrained values, $\theta_{\bar{S}} = \theta^\star_{\bar{S}}$,while uncertainty is introduced only over θ_S.

DASL approximates the subnetwork posterior by a Gaussian centered at the pretrained solution,

$$q(\theta_S) = \mathcal{N}\left(\theta^\star_S, \Sigma_S\right).$$

The remaining questions are how to choose a computationally efficient subnetwork S and how to construct Σ_S so that derivative information from forces is retained.

3.2 EFFICIENT SUBNETWORK SELECTION

A full-covariance posterior approximation over all D parameters is generally impractical, whereas restricting the posterior to k selected parameters reduces covariance storage and factorization from $O(D^2)$ and $O(D^3)$ to $O(k^2)$ and $O(k^3)$, respectively. The remaining question is which parameters should form S. Let

$$\Phi_\theta(x) = \begin{bmatrix} E_\theta(x) \\ F_\theta(x) \end{bmatrix}, \qquad \Sigma_{\text{obs}} = \operatorname{diag}\left(\sigma_E^2, \sigma_F^2 I\right),$$

and let $\widetilde{J}$ denote the stacked, noise-whitened Jacobian of Φ_θ on the training configurations, evaluated at $\theta^\star$. For parameter θ_j, define

$$c_j = \|\widetilde{J}_j\|_2^2, \qquad v_j = (\lambda + c_j)^{-1},$$

where $\widetilde{J}_j$ is the j-th Jacobian column and λ is an isotropic prior precision. The quantity c_j measures the sensitivity of the joint energy–force outputs to parameter j, while v_j is its marginal variance under a diagonal Laplace approximation.

The original subnetwork Laplace (Daxberger et al., 2021b) criterion prioritizes parameters with large v_j, i.e. directions that are most uncertain in weight space. DASL instead selects the k parameters with the smallest v_j, equivalently the largest derivative-aware curvatures c_j. The following result motivates this reversed criterion in function space.

Proposition 1. *Minimum-variance selection maximizes diagonal output-space variance.*

Under the diagonal Laplace approximation, let

$$K_S = \widetilde{J}_S \operatorname{diag}(v_j)_{j\in S} \widetilde{J}_S^\top$$

denote the covariance of the whitened linearized outputs induced by a subnetwork S. For any fixed budget $|S| = k$,

$$\arg\max_{|S|=k} \operatorname{tr}(K_S) = \arg\min_{|S|=k} \sum_{j\in S} v_j.$$

Therefore, the k parameters with the smallest marginal posterior variances maximize the total retained output-space variance under the diagonal approximation.

This result reverses the usual maximum-variance criterion used in subnetwork Laplace (Daxberger et al., 2021b). A small weight-space variance can arise because the outputs are highly sensitive to that parameter direction; the relevant contribution to function-space uncertainty is $c_j v_j$, rather than v_j alone. DASL therefore uses the minimum-variance criterion to select S, before constructing a full-covariance posterior over the selected parameters.

3.3 Derivative-Aware Laplace Approximation

After selecting S, we construct a full-covariance Laplace approximation around $\theta^\star$. Using the generalized Gauss–Newton approximation for the joint energy–force likelihood, the posterior precision decomposes as $\Lambda_S = P_{0,S} + C_{E,S} + C_{F,S}$, where

$$C_{E,S} = \frac{1}{\sigma_E^2} \sum_{i \in \mathcal{D}_{\text{train}}} J_{E,S}(x_i)^\top J_{E,S}(x_i), \qquad C_{F,S} = \frac{1}{\sigma_F^2} \sum_{i \in \mathcal{D}_{\text{train}}} J_{F,S}(x_i)^\top J_{F,S}(x_i).$$

Here, $J_{E,S}(x) = \frac{\partial E_\theta(x)}{\partial \theta_S}$, $J_{F,S}(x) = \frac{\partial F_\theta(x)}{\partial \theta_S}$. The derivative-aware structure follows directly from $F_\theta(x) = -\nabla_R E_\theta(x)$, which implies $J_{F,S}(x) = -\nabla_R J_{E,S}(x)$. Hence, $C_{E,S}$ captures parameter sensitivity of the energy values, whereas $C_{F,S}$ additionally captures sensitivity of the local gradients of the potential-energy surface. The resulting posterior is $q(\theta_S) = \mathcal{N}\left(\theta_S^\star, \Lambda_S^{-1}\right)$.

Force information can also be viewed as additional posterior contraction. Defining

$$\Lambda_{E,S} = P_{0,S} + C_{E,S}, \qquad \Lambda_{EF,S} = \Lambda_{E,S} + C_{F,S},$$

we have $C_{F,S} \succeq 0$, and therefore $\Lambda_{EF,S}^{-1} \preceq \Lambda_{E,S}^{-1}$. Thus, force observations provide additional information in directions that are visible through derivatives of the learned energy surface.

3.4 Induced Function-Space Posterior

To connect DASL to the function-space principles of Section 2, define the energy tangent features of the selected parameters as $\phi_S(x) = J_{E,S}(x)$. Linearizing the energy around the pretrained solution (Immer et al., 2021) gives $E_\theta(x) \approx E_{\theta^\star}(x) + \phi_S(x)(\theta_S - \theta_S^\star)$. Together with $\theta_S - \theta_S^\star \sim \mathcal{N}(0, \Lambda_S^{-1})$, this induces the finite-rank kernel

$$k_E^S(x, x') = \phi_S(x) \Lambda_S^{-1} \phi_S(x')^\top.$$

Thus, selecting a subnetwork is equivalently selecting the tangent-feature subspace used to represent function-space uncertainty. Under the diagonal approximation used for subnetwork selection,

$$k_{E,\text{diag}}^S(x, x') = \sum_{j \in S} v_j \, \phi_j(x) \phi_j(x'),$$

each selected parameter contributes one component to the finite-rank tangent kernel. The DASL criterion selects the components with the largest derivative-aware output-space contribution $c_j v_j = \frac{c_j}{\lambda + c_j}$, thereby choosing a compact tangent kernel intended to retain the uncertainty most relevant to the joint energy–force predictions.

Importantly, this kernel inherits the physical structure established in Section 2. Because the underlying energy model is invariant/equivariant for all admissible parameter values, its tangent features inherit the corresponding symmetry structure, and hence the induced covariance satisfies the symmetry-consistency condition of Section 2.2. Moreover,

$$J_{F,S}(x) = -\nabla_R J_{E,S}(x)$$

implies that the force tangent features are derivatives of the energy tangent features. Consequently, the energy–force covariance blocks are obtained by differentiating k_E^S, recovering the derivative-consistent structure of Section 2.3. DASL can therefore be viewed as selecting and Bayesianizing a compact, physically structured tangent-feature space: subnetwork selection determines the finite-rank kernel, while derivative-aware Laplace determines its posterior. Formal derivations of the finite-rank kernel construction, its tangent-feature interpretation, and the inherited symmetry and derivative-consistency properties are provided in Appendix B.

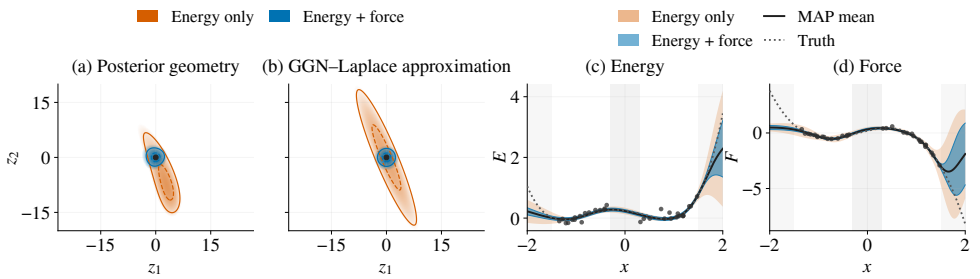

Figure 2: **Effect of derivative information on the Laplace posterior.** (a) (b) Energy-only curvature leaves a weakly constrained parameter direction, whereas incorporating force observations contracts the posterior along directions visible through the force Jacobian. (c) (d) The two approximations share the same MAP predictor but induce different uncertainty in function space.

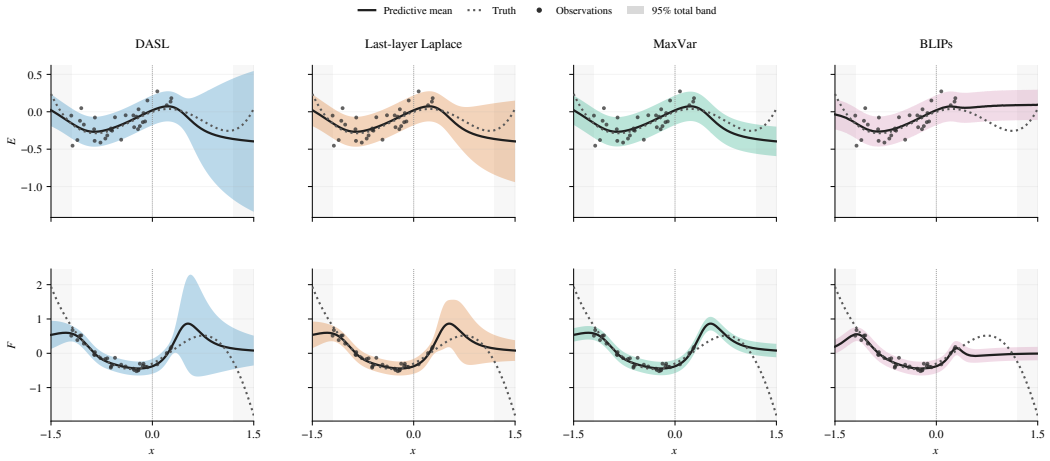

Figure 3: **Comparison of predictive uncertainty on the toy problem.** Energy predictions (top row) and force predictions (bottom row) are shown for DASL, last-layer Laplace (Watson et al., 2021), MaxVar (Daxberger et al., 2021b), and BLIPs (Coscia et al., 2025). Solid curves denote the predictive mean, dotted curves the ground truth, and points the observations; shaded regions indicate the 95% total predictive interval.

4 Experiments

4.1 Controlled Experiments

We first use one-dimensional controlled potentials to illustrate the two mechanisms underlying DASL: derivative-aware posterior construction and subnetwork selection. The scalar energy $E(x)$ induces the force $F(x) = -\partial E(x)/\partial x$, allowing us to visualize both parameter-space uncertainty and the corresponding uncertainty in energy and force predictions.

Effect of force observation. We first compare Laplace approximations constructed from energy observations alone and from both energy and force observations. Figure 2 shows the resulting posterior geometry and predictive uncertainty. Adding force information substantially contracts the posterior along directions that remain weakly constrained by energy values alone; the corresponding GGN–Laplace approximation exhibits the same qualitative contraction.

This change in parameter-space geometry propagates directly to function space. Although the two approximations share the same MAP predictor, their uncertainty differs, particularly for the force prediction, where derivative information more strongly constrains the local shape of the potential. These observations are consistent with the analysis in Section 3: the force curvature C_F contributes additional information through $J_F = -\nabla_R J_E$, thereby reducing uncertainty in directions that are visible through derivatives of the energy surface.

Effect of the subnetwork choice. We next examine how the choice of stochastic parameterization shapes the induced function-space uncertainty. Figure 3 compares DASL, last-layer Laplace (Watson et al., 2021), MaxVar subnetwork Laplace (Daxberger et al., 2021b), and BLIPs (Coscia et al.,

Table 1: **Force uncertainty on the Ammonia OOD test set.** All subnetwork Laplace variants use the same frozen predictive mean.

Method	K	Force NLL ↓	Force CRPS ↓	Spearman ↑
BLIP (Coscia et al., 2025)	–	5.731	38.195	0.882
Deep Ensemble (Lakshminarayanan et al., 2017)	–	5.720	37.119	0.748
Last-layer Laplace	65	5.707	36.753	0.895
MaxVar (Daxberger et al., 2021b)	64	5.708	36.758	0.817
Random	64	5.708	36.758	0.863
DASL	64	5.707	36.753	**0.905**

2025) on the same toy problem. DASL and last-layer Laplace exhibit substantially increasing uncertainty as predictions move away from the observed configurations, with DASL showing the strongest widening in the sparse-data region. In contrast, MaxVar retains the least extrapolative uncertainty, despite selecting parameters with the largest marginal variance in weight space. Both MaxVar and BLIPs also show relatively little distinction between observation-dense and observation-sparse regions in this controlled setting. These results illustrate that the choice of stochastic parameter subspace directly shapes the induced finite-rank kernel and can strongly affect whether uncertainty reflects local data support. The narrowing near $x = 1.5$ reflects a limitation of the local Laplace approximation: it predicts little variation in force around the fitted model, despite a large extrapolation error. We provide a more detailed discussion of the observed BLIPs failure mode in Appendix E.

4.2 Molecular Benchmarks

We next evaluate whether the behavior observed in the controlled example carries over to equivariant molecular potentials. In all experiments, the pretrained MLIP is kept fixed and DASL is fitted *post hoc*. We compare DASL with alternative subnetwork Laplace constructions, including maximum-variance (Daxberger et al., 2021b) and last-layer Laplace (Watson et al., 2021), together with BLIP (Coscia et al., 2025), GMM, deep ensembles (Lakshminarayanan et al., 2017), and other baselines.

Ammonia. Table 1 evaluates force uncertainty on the ammonia inversion benchmark (Tan et al., 2023; Coscia et al., 2025). Among the subnetwork Laplace methods, DASL achieves the strongest uncertainty–error alignment, with a Spearman correlation of 0.905, compared with 0.895 for last-layer Laplace, 0.817 for MaxVar, and 0.863 for random selection. DASL also matches the best Laplace NLL and CRPS (5.707 and 36.753, respectively). The improvement over MaxVar is consistent with the function-space motivation of Section 3: selecting parameters with large marginal weight variance does not necessarily retain the directions most relevant to uncertainty in the energy–force outputs. BLIP also provides strong ranking performance (0.882), while the deep ensemble attains better CRPS than BLIP but substantially weaker rank correlation (0.748). On this controlled molecular benchmark, DASL provides both competitive probabilistic scores and the strongest identification of high-error configurations.

The comparison between DASL and the other subnetwork variants is particularly informative because their deterministic predictions are identical. The improvement in error ranking therefore comes from how the stochastic subspace is chosen rather than from a change in predictive accuracy. In particular, the gap between DASL and MaxVar is consistent with the function-space motivation of Section 4.1: parameter directions with large weight-space variance need not be the directions that contribute most strongly to uncertainty in energy and force predictions.

3BPA under distribution shift. Table 2 evaluates force uncertainty on 3BPA benchmark (Kovács et al., 2021) at $300\,\mathrm{K}$ and $600\,\mathrm{K}$. The different Laplace subnetworks obtain nearly identical NLL and CRPS, with their main differences appearing in uncertainty–error ranking. At $300\,\mathrm{K}$, DASL achieves a Spearman correlation of 0.515, slightly below last-layer Laplace (0.598) and MaxVar (0.586), but above random selection (0.442). At $600\,\mathrm{K}$, this gap narrows: DASL reaches 0.622, compared with 0.642 for last-layer Laplace and 0.637 for MaxVar, while again improving over random selection (0.552). Thus, the proposed MinVar criterion does not uniformly dominate alternative subnetwork choices on 3BPA. Nevertheless, DASL compares favorably with BLIP across both distribution shifts. At $300\,\mathrm{K}$, DASL improves the force NLL from 2.569 to 2.213, CRPS from 1.707 to 1.093, and Spearman correlation from 0.345 to 0.515. The same pattern holds at $600\,\mathrm{K}$, where DASL obtains 2.870 NLL, 1.967 CRPS, and 0.622 Spearman correlation, compared with 3.089, 3.009, and 0.572 for BLIP. The deep ensemble remains strongest overall on this benchmark, but

Table 2: **Force uncertainty on 3BPA at $300\,\mathrm{K}$ and $600\,\mathrm{K}$.** The Laplace variants share the same frozen PaiNN predictive mean.

Test set	Method	Force NLL ↓	Force CRPS ↓	Spearman ↑
300 K	BLIP (Coscia et al., 2025)	2.569	1.707	0.345
	Deep Ensemble (Lakshminarayanan et al., 2017)	1.717	0.790	0.631
	Last-layer Laplace	2.213	1.093	0.598
	MaxVar (Daxberger et al., 2021b)	2.213	1.092	0.586
	Random	2.212	1.092	0.442
	DASL	2.213	1.093	0.515
600 K	BLIP (Coscia et al., 2025)	3.089	3.009	0.572
	Deep Ensemble (Lakshminarayanan et al., 2017)	2.779	1.796	0.666
	Last-layer Laplace	2.864	1.966	0.642
	MaxVar (Daxberger et al., 2021b)	2.869	1.967	0.637
	Random	2.886	1.969	0.552
	DASL	2.870	1.967	0.622

Table 3: **Force uncertainty on MACE-OFF23-small across five molecule-disjoint splits (mean ± sample standard deviation).** The Laplace variants preserve the pretrained predictive mean. Adapted baselines update their mean and therefore are not directly controlled comparisons of the uncertainty approximation.

Method	K	Force NLL ↓	Force CRPS ↓	ρ_F ↑
Frozen pretrained predictor				
Last-layer Laplace	112	-0.779 ± 0.496	0.0428 ± 0.0025	0.464 ± 0.093
Random selection	512	-0.960 ± 0.217	0.0427 ± 0.0024	0.573 ± 0.054
MaxVar	512	-1.058 ± 0.040	0.0434 ± 0.0023	0.601 ± 0.067
DASL	512	$\mathbf{-1.065 \pm 0.030}$	0.0433 ± 0.0023	$\mathbf{0.612 \pm 0.046}$
Adapted predictive model				
BLIP (Coscia et al., 2025)	–	-0.961 ± 0.259	$\mathbf{0.0400 \pm 0.0049}$	0.439 ± 0.039
Readout ensemble	–	-0.570 ± 0.835	0.0428 ± 0.0025	0.486 ± 0.076
Fine-tuned deep ensemble	–	-0.611 ± 0.760	0.0429 ± 0.0026	0.471 ± 0.041
BBB (Blundell et al., 2015)	–	-0.941 ± 0.227	$\mathbf{0.0400 \pm 0.0050}$	0.536 ± 0.045

requires training multiple predictive models. DASL instead operates on the frozen PaiNN (Schütt et al., 2021) predictor, providing competitive uncertainty estimates without retraining the underlying MLIP.

4.3 Post-hoc UQ for a Pretrained Molecular Potential

Setup. We next evaluate the intended post-hoc use case of DASL on the released MACE-OFF23-small molecular potential Kovács et al. (2023). Across five molecule-disjoint splits, the pretrained MACE checkpoint is kept fixed for all Laplace variants, so differences in uncertainty quality arise from the stochastic subspace and posterior approximation rather than changes in the predictive mean. DASL, MaxVar, and random selection use a matched budget of $K = 512$ stochastic parameters, while last-layer Laplace places a posterior over the 112 terminal readout weights. We additionally compare with BLIP, readout and fine-tuned ensembles, and Bayes by Backprop (BBB), which adapt the predictive model and therefore provide complementary rather than directly controlled comparisons. Hyperparameters are selected using validation data; split construction, implementation details, and budget ablations are provided in Appendix D.4.

Fixed-pretrained comparison. Table 3 shows that DASL performs strongest among methods that preserve the released MACE predictor. At the matched $K = 512$ budget, DASL obtains the lowest mean force NLL, -1.065 ± 0.030, and the highest force uncertainty–error correlation, $\rho_F = 0.612 \pm 0.046$. MaxVar is close, with -1.058 ± 0.040 NLL and $\rho_F = 0.601 \pm 0.067$, while random selection reaches -0.960 ± 0.217 and 0.573 ± 0.054, respectively. The advantage of DASL over MaxVar is modest, but both informed selection rules improve over random selection without changing the pretrained predictive mean. Joint-readout Laplace yields weaker ranking performance ($\rho_F = 0.464 \pm 0.093$), although it uses a smaller stochastic parameter budget.

Comparison with other adapted UQ baselines. DASL remains particularly strong even when compared with methods that are allowed to adapt the predictive model. It achieves the highest force uncertainty–error correlation, $\rho_F = 0.612$, outperforming BLIP (0.439), the readout ensemble (0.486), the fine-tuned deep ensemble (0.471), and BBB (0.536). DASL also obtains the lowest mean force NLL (-1.065) among all methods in Table 3. Notably, these gains are achieved while preserving the released MACE predictive mean and Bayesianizing only 512 selected parameters,

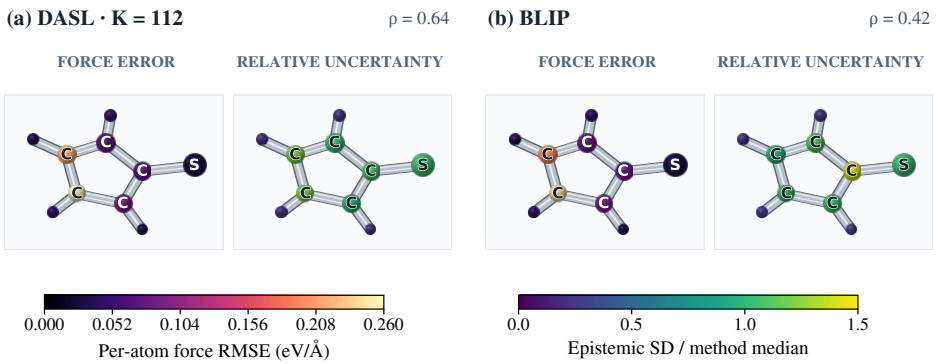

Figure 4: **Per-atom force error and epistemic uncertainty for an illustrative held-out MACE-OFF23 molecule.** Each method shows frozen-model per-atom force RMSE against DFT (left) and epistemic force standard deviation (right). Uncertainty is normalized within each method to emphasize spatial ranking. DASL, MaxVar, and last-layer Laplace preserve the pretrained MACE mean; BLIP uses an adapted predictive mean.

whereas the competing adapted methods modify or retrain parts of the predictive model. BLIP and BBB obtain lower CRPS, indicating that DASL does not dominate every probabilistic metric; nevertheless, its combination of strong error ranking, favorable NLL, and a strictly post-hoc construction demonstrates the practical advantage of selecting a compact, uncertainty-relevant subspace of a pretrained molecular potential. Overall, the MACE experiment demonstrates the intended use case of DASL: informative force uncertainty can be attached to a large pretrained MLIP without changing its original predictions or retraining the model. The improvement over both random and MaxVar selection further indicates that the choice of stochastic subspace matters, while the comparison with adapted baselines shows that competitive—and in uncertainty ranking, stronger—UQ can be obtained using only a compact post-hoc posterior.

Visualization. Figure 4 provides a configuration-level view of this behavior for an illustrative held-out 10-atom C_5H_4S molecule. The two ring carbons farthest from sulfur have the largest frozen-MACE force errors (0.254 and $0.217\,\mathrm{eV/Å}$), and DASL assigns its two largest uncertainties to the same atoms. BLIP exhibits a different spatial uncertainty pattern using its adapted predictive mean. For this molecule, DASL yields a stronger within-configuration uncertainty–error correlation ($\rho = 0.64$) than BLIP ($\rho = 0.42$).

5 Summary and Conclusion

We introduced *Derivative-Aware Subnetwork Laplace* (DASL), a post-hoc Bayesian uncertainty method for pretrained machine-learned interatomic potentials. Motivated by a function-space view of MLIP uncertainty, DASL is designed to preserve the symmetries of the underlying molecular potential and the derivative relation between energy and forces. It achieves this by retaining the pretrained equivariant model, Bayesianizing only a compact subset of parameters, and constructing the local Laplace posterior from both energy and force sensitivities. We further show that the subnetwork-selection problem admits a function-space interpretation: under a diagonal Laplace approximation, selecting parameters with the smallest marginal weight variance maximizes the retained joint energy–force output variance at a fixed budget. The resulting linearized posterior induces a finite-rank tangent kernel that inherits the symmetry and derivative structure of the original potential.

Across controlled and molecular benchmarks, DASL provides useful uncertainty estimates while leaving the pretrained predictive model unchanged. The controlled experiments isolate the effects of derivative-aware curvature and subnetwork selection, while the molecular experiments demonstrate strong uncertainty–error alignment on Ammonia and MACE-OFF23-small and competitive performance under the more challenging 3BPA distribution shifts. In particular, the MACE experiment demonstrates the intended use case of DASL: structured Bayesian uncertainty can be attached to a large pretrained molecular potential by Bayesianizing only a small parameter subset, without retraining the underlying model. Overall, DASL provides a computationally tractable route to uncertainty over physically structured potential-energy surfaces, combining symmetry preservation, derivative-aware inference, and efficient post-hoc Bayesianization.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, mathematical checking, and coding support. This included improving the clarity and organization of the manuscript, checking intermediate derivations and mathematical expressions, and assisting with code development and debugging. Generative AI was not used to generate experimental data or fabricate numerical results, figures, or citations.

All AI-assisted text, mathematical content, and code were reviewed and verified by the authors. Mathematical arguments were checked for correctness, code was tested against the intended computations, and all reported experimental results were validated by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13): 136403, 2010.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

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.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

IR Best, TJ Sullivan, and James R Kermode. Uncertainty quantification in atomistic simulations of silicon using interatomic potentials. *The journal of chemical physics*, 161(6), 2024.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Dario Coscia, Pim de Haan, and Max Welling. Blips: Bayesian learned interatomic potentials. *arXiv preprint arXiv:2508.14022*, 2025.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021a.

Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pp. 2510–2521. PMLR, 2021b.

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.

Cheuk Hin Ho, Christoph Ortner, and Yangshuai Wang. Flexible uncertainty calibration for machine-learned interatomic potentials. *npj Computational Materials*, 2026.

Yuge Hu, Joseph Musielewicz, Zachary W Ulissi, and Andrew J Medford. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Machine Learning: Science and Technology*, 3(4):045028, 2022.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pp. 703–711. PMLR, 2021.

Dávid Péter Kovács, Cas van der Oord, Jiri Kucera, Alice EA Allen, Daniel J Cole, Christoph Ortner, and Gábor Csányi. Linear atomic cluster expansion force fields for organic molecules: beyond rmse. *Journal of chemical theory and computation*, 17(12):7696–7711, 2021.

Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. Mace-off23: Transferable machine learning force fields for organic molecules, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

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. *Nature Communications*, 14(1):579, 2023.

Evgeny V Podryabinkin and Alexander V Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International conference on learning representations*, 2018.

Christoph Schran, Krystof Brezina, and Ondrej Marsalek. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics*, 153 (10), 2020.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Aik Rui Tan, Shingo Urata, Samuel Goldman, Johannes CB Dietschreit, and Rafael Gómez-Bombarelli. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Computational Materials*, 9(1):225, 2023.

Stephan Thaler, Gregor Doehner, and Julija Zavadlav. Scalable bayesian uncertainty quantification for neural network potentials: promise and pitfalls. *Journal of Chemical Theory and Computation*, 19(14):4520–4532, 2023.

Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muller. Machine learning force fields. *Chemical reviews*, 121(16):10142–10186, 2021.

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 Computational Materials*, 6(1):20, 2020.

Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. Latent derivative bayesian last layer networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1198–1206. PMLR, 2021.

Mingjian Wen and Ellad B Tadmor. Uncertainty quantification in molecular simulations with dropout neural network potentials. npj computational materials, 6 (1), august 2020. *URL https://doi. org/10.1038/s41524-020-00390-8*, 2020.

Logan Williams, Khachik Sargsyan, Andrew Rohskopf, and Habib N Najm. Active learning for snap interatomic potentials via bayesian predictive uncertainty. *Computational Materials Science*, 242: 113074, 2024.

Soohaeng Yoo Willow, Tae Hyeon Park, Gi Beom Sim, Sung Wook Moon, Seung Kyu Min, D ChangMo Yang, Hyun Woo Kim, Juho Lee, and Chang Woo Myung. Bayesian e (3)-equivariant interatomic potential with iterative restratification of many-body message passing. *arXiv preprint arXiv:2510.03046*, 2025.

Yu Xie, Jonathan Vandermause, Senja Ramakers, Nakib H Protik, Anders Johansson, and Boris Kozinsky. Uncertainty-aware molecular dynamics from bayesian active learning for phase transformations and thermal transport in sic. *npj Computational Materials*, 9(1):36, 2023.

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

A Related works

A.1 Uncertainty quantification for MLIPs.

Uncertainty quantification for machine-learned interatomic potentials has been studied through several complementary approaches. Deep ensembles remain strong empirical baselines for MLIP uncertainty, but require training, storing, and evaluating multiple full potentials, which is costly for large equivariant or pretrained models (Lakshminarayanan et al., 2017; Tan et al., 2023; Willow et al., 2025). Single-model predictive methods, including mean–variance estimation, evidential regression, mixture-density models, dropout, and variational dropout, reduce this cost and can provide useful marginal uncertainty estimates (Wen & Tadmor, 2020; Tan et al., 2023; Coscia et al., 2025). However, these methods typically learn predictive uncertainty or inject stochasticity at inference time, rather than constructing a global posterior over physically admissible potential-energy surfaces. Calibration-based approaches such as conformal prediction can improve empirical coverage, but similarly do not define posterior samples of invariant energies and equivariant force fields (Hu et al., 2022; Ho et al., 2026). More Bayesian approaches, including SG-MCMC, sparse Gaussian-process force fields, and Bayesian linear potentials, provide coherent posterior uncertainty, but can be computationally expensive or tied to non-neural model classes (Thaler et al., 2023; Xie et al., 2023; Williams et al., 2024).

A.2 Laplace Approximation and Subnetwork Inference

Laplace approximation provides a practical post-hoc approach to Bayesian neural-network inference by approximating the posterior around a trained MAP solution with a Gaussian distribution (Ritter et al., 2018; Daxberger et al., 2021a). Modern variants use structured curvature approximations to make posterior construction tractable for large neural networks. GGN-based and locally linearized formulations additionally provide a convenient function-space interpretation of the resulting approximation (Immer et al., 2021). For networks with millions of parameters, however, maintaining a full-covariance posterior over all weights remains prohibitively expensive. Subnetwork inference addresses this issue by placing a richer posterior over only a selected subset of parameters while fixing the remaining weights at their MAP values (Daxberger et al., 2021b). The original subnetwork criterion prioritizes parameters with large marginal posterior variance, motivated by preserving uncertainty in weight space. Last-layer Laplace provides a simpler alternative by restricting Bayesian inference to the final prediction layer.

DASL builds on the idea of subnetwork Laplace but changes both the information used to construct the posterior and the criterion used to choose the stochastic subspace. Energy and force sensitivities jointly determine the local curvature, making the approximation derivative-aware. Moreover, rather than selecting parameters according to large weight-space variance, DASL selects a tangent-feature subspace according to its contribution to joint energy–force function-space uncertainty. This provides a compact post-hoc posterior that retains the symmetry and derivative structure of the pretrained equivariant potential.

B Properties of the Induced DASL Kernel

This section provides the formal analysis for the function-space interpretation of DASL in Section 3.4. We show that the linearized subnetwork posterior induces a valid finite-rank kernel, that

subnetwork selection restricts the corresponding tangent-feature space, and that the resulting kernel inherits the symmetry and derivative structure of the underlying MLIP.

B.1 Properties of the Induced DASL Kernel

Let

$$\delta\theta_S = \theta_S - \theta_S^\star \sim \mathcal{N}(0, \Lambda_S^{-1}),$$

and define the selected energy tangent features

$$\phi_S(x) = J_{E,S}(x) = \left.\frac{\partial E_\theta(x)}{\partial \theta_S}\right|_{\theta=\theta^\star}.$$

Under the first-order approximation,

$$E_\theta(x) \approx E_{\theta^\star}(x) + \phi_S(x)\delta\theta_S.$$

Since $\delta\theta_S$ is Gaussian, any finite collection of linearized energy predictions is jointly Gaussian. The induced covariance is

$$\begin{aligned} k_E^S(x, x') &= \mathrm{Cov}\left[\phi_S(x)\delta\theta_S, \phi_S(x')\delta\theta_S\right] \\ &= \phi_S(x)\Lambda_S^{-1}\phi_S(x')^\top. \end{aligned}$$

To verify that k_E^S is a valid positive-semidefinite kernel, consider arbitrary configurations $\{x_i\}_{i=1}^n$ and coefficients $\{a_i\}_{i=1}^n$. Then

$$\begin{aligned} \sum_{i,j=1}^n a_i a_j k_E^S(x_i, x_j) &= \left(\sum_i a_i \phi_S(x_i)\right) \Lambda_S^{-1} \left(\sum_j a_j \phi_S(x_j)\right)^\top \\ &\geq 0, \end{aligned}$$

because $\Lambda_S^{-1} \succeq 0$. Hence, the linearized DASL posterior defines a valid Gaussian process with kernel k_E^S. Moreover, the feature map $\phi_S(x)$ has dimension $|S|$. Therefore, for any finite collection of configurations,

$$\mathrm{rank}(K_E^S) \leq |S|,$$

so the induced GP is finite rank.

B.2 Subnetwork Selection as Tangent-Feature Selection

Let the full energy tangent feature vector be

$$\phi(x) = \left[\phi_1(x), \ldots, \phi_D(x)\right], \qquad \phi_j(x) = \left.\frac{\partial E_\theta(x)}{\partial \theta_j}\right|_{\theta=\theta^\star}.$$

Selecting a subnetwork $S \subset \{1, \ldots, D\}$ restricts this feature representation to

$$\phi_S(x) = [\phi_j(x)]_{j\in S}.$$

Thus, subnetwork selection determines the tangent-feature subspace on which the finite-rank kernel is constructed. Under the diagonal approximation used for subnetwork selection,

$$\Lambda_S^{-1} \approx \mathrm{diag}(v_j)_{j\in S},$$

and the corresponding kernel becomes

$$k_{E,\mathrm{diag}}^S(x, x') = \sum_{j\in S} v_j \phi_j(x)\phi_j(x').$$

Hence, each selected parameter contributes one rank-one tangent-kernel component. For the noise-whitened joint energy–force Jacobian $\widetilde{J}$, the derivative-aware curvature of parameter j is

$$c_j = \|\widetilde{J}_j\|_2^2,$$

and its diagonal Laplace variance under isotropic prior precision λ is

$$v_j = (\lambda + c_j)^{-1}.$$

The contribution of parameter j to the total whitened output-space variance is therefore

$$c_j v_j = \frac{c_j}{\lambda + c_j}.$$

As shown in the minimum-variance proposition, selecting the smallest v_j, equivalently the largest c_j, maximizes the retained output-space variance under this diagonal approximation. In the function-space view, DASL therefore selects tangent-kernel components according to their derivative-aware contribution to joint energy–force uncertainty. For the final DASL posterior, Λ_S^{-1} is full covariance, and the induced kernel can be expanded as

$$k_E^S(x, x') = \sum_{j,\ell \in S} [\Lambda_S^{-1}]_{j\ell} \phi_j(x) \phi_\ell(x').$$

Thus, the subnetwork determines the tangent-feature subspace, while the derivative-aware Laplace posterior determines the metric coupling these features within that subspace.

B.3 Inheritance of Symmetry

Assume that the underlying energy model is invariant for every admissible parameter value,

$$E_\theta(g \cdot x) = E_\theta(x), \qquad \forall\, \theta,$$

for every symmetry transformation $g \in \mathcal{G}$. Differentiating this identity with respect to θ_S gives

$$J_{E,S}(g \cdot x) = J_{E,S}(x).$$

Therefore,

$$\begin{aligned} k_E^S(g \cdot x, g \cdot x') &= J_{E,S}(g \cdot x) \Lambda_S^{-1} J_{E,S}(g \cdot x')^\top \\ &= J_{E,S}(x) \Lambda_S^{-1} J_{E,S}(x')^\top \\ &= k_E^S(x, x'). \end{aligned}$$

Hence, the induced energy kernel is invariant. More generally, let

$$\Phi_\theta(x) = \begin{bmatrix} E_\theta(x) \\ F_\theta(x) \end{bmatrix}$$

satisfy

$$\Phi_\theta(g \cdot x) = T_g \Phi_\theta(x)$$

for every admissible θ. Differentiation with respect to the selected parameters yields

$$J_{\Phi,S}(g \cdot x) = T_g J_{\Phi,S}(x).$$

The corresponding DASL covariance

$$K_\Phi^S(x, x') = J_{\Phi,S}(x) \Lambda_S^{-1} J_{\Phi,S}(x')^\top$$

therefore obeys

$$K_\Phi^S(g \cdot x, g \cdot x') = T_g K_\Phi^S(x, x') T_g^\top.$$

Thus, the finite-rank posterior induced by DASL inherits the symmetry structure of the underlying equivariant MLIP.

B.4 Derivative Consistency of the Induced Kernel

Because the force field is obtained from the energy,

$$F_\theta(x) = -\nabla_R E_\theta(x),$$

the corresponding tangent features satisfy

$$J_{F,S}(x) = -\nabla_R J_{E,S}(x),$$

assuming differentiation with respect to coordinates and parameters can be interchanged. The energy–force covariance induced by DASL is

$$\begin{aligned} K^S_{EF}(x, x') &= J_{E,S}(x)\Lambda_S^{-1} J_{F,S}(x')^\top \\ &= -J_{E,S}(x)\Lambda_S^{-1} \left(\nabla_{R'} J_{E,S}(x')\right)^\top \\ &= -\nabla_{R'} k^S_E(x, x'). \end{aligned}$$

Similarly,

$$K^S_{FE}(x, x') = -\nabla_R k^S_E(x, x'),$$

and

$$K^S_{FF}(x, x') = \nabla_R \nabla^\top_{R'} k^S_E(x, x').$$

Hence, the energy and force covariances induced by DASL are generated from the same latent energy kernel. The finite-rank subnetwork approximation therefore preserves the derivative-consistent function-space structure described in Section 2.3.

C DASL Algorithm

Algorithm 1 summarizes the complete DASL procedure. Starting from a pretrained MLIP, DASL first uses derivative-aware energy–force sensitivities to select a compact stochastic subnetwork, and then constructs a full-covariance Laplace posterior over the selected parameters while keeping the remaining network fixed.

D Supplementary Information on Experiments

D.1 Additional Details for the Controlled Experiments

D.1.1 Effect of Force Observations

Potential and observations. We consider the one-dimensional potential

$$E(x) = \frac{(x^2-1)^2}{4} + 0.2x(x^2-1), \qquad F(x) = -\frac{\mathrm{d}E(x)}{\mathrm{d}x}. \tag{1}$$

The learned model is a 1–16–16–1 neural network with Tanh activations and 321 scalar parameters. The model is not given the polynomial form in Equation equation 1; forces are obtained by differentiating its predicted energy.

The principal design contains 32 configurations, with 16 sampled from $[-1.5, -0.3]$ and 16 from $[0.3, 1.5]$. Energy and force labels have independent Gaussian noise with known standard deviations $\sigma_E = \sigma_F = 0.1$. We consider two observation conditions: 32 energy labels, and the same 32 energy labels supplemented by 32 force labels.

MAP fitting and Laplace approximation. The energy-only and energy–force training objectives are

$$\mathcal{L}(\theta) = \frac{1}{2\sigma_E^2} \sum_i \left(E_\theta(x_i) - y_i^E\right)^2 + \frac{\delta_F}{2\sigma_F^2} \sum_i \left(F_\theta(x_i) - y_i^F\right)^2 + \frac{\alpha}{2}\|\theta\|_2^2, \tag{2}$$

where $\delta_F = 1$ when force observations are included and zero otherwise. The prior precision is fixed at $\alpha = 1$. Optimization uses 1,000 Adam steps with learning rate 3×10^{-3}, followed by Levenberg–Marquardt least squares with analytic Jacobians.

DASL selects $K = 33$ scalar coordinates and constructs the dense restricted GGN precision

$$\Lambda_B = \alpha I + \sigma_E^{-2} J_{E,B}^\mathsf{T} J_{E,B} + \delta_F \sigma_F^{-2} J_{F,B}^\mathsf{T} J_{F,B}, \tag{3}$$

where B denotes the selected subnetwork. Predictions use the linearized covariance Λ_B^{-1}.

Algorithm 1 Derivative-Aware Subnetwork Laplace (DASL)

Require: Pretrained MLIP parameters $\theta^\star$; training set $\mathcal{D}_{\text{train}}$; subnetwork budget k; prior precision λ; energy and force noise scales σ_E, σ_F.
Ensure: Selected subnetwork S and Gaussian posterior $q(\theta_S) = \mathcal{N}(\theta_S^\star, \Lambda_S^{-1})$.
1: Freeze the pretrained model at $\theta^\star$.
2: **for** each trainable parameter θ_j **do**
3: Compute energy sensitivity

$$J_{E,j}(x_i) = \left.\frac{\partial E_\theta(x_i)}{\partial \theta_j}\right|_{\theta=\theta^\star}.$$

4: Compute force sensitivity

$$J_{F,j}(x_i) = \left.\frac{\partial F_\theta(x_i)}{\partial \theta_j}\right|_{\theta=\theta^\star} = -\nabla_R J_{E,j}(x_i).$$

5: Compute derivative-aware diagonal curvature

$$c_j = \sum_{i\in\mathcal{D}_{\text{train}}} \left[\frac{1}{\sigma_E^2} J_{E,j}(x_i)^\top J_{E,j}(x_i) + \frac{1}{\sigma_F^2} J_{F,j}(x_i)^\top J_{F,j}(x_i)\right].$$

6: Compute diagonal posterior variance

$$v_j = (\lambda + c_j)^{-1}.$$

7: **end for**
8: Select the k parameters with the smallest v_j:

$$S = \arg\min_{|S|=k} \sum_{j\in S} v_j.$$

9: Construct the selected energy and force Jacobians $J_{E,S}$ and $J_{F,S}$.
10: Form the full-covariance posterior precision

$$\Lambda_S = P_{0,S} + \frac{1}{\sigma_E^2} \sum_{i\in\mathcal{D}_{\text{train}}} J_{E,S}(x_i)^\top J_{E,S}(x_i) + \frac{1}{\sigma_F^2} \sum_{i\in\mathcal{D}_{\text{train}}} J_{F,S}(x_i)^\top J_{F,S}(x_i).$$

11: Define the DASL posterior

$$q(\theta_S) = \mathcal{N}\left(\theta_S^\star, \Lambda_S^{-1}\right).$$

12: **return** S, $q(\theta_S)$.

D.1.2 Observation Density and Subnetwork Choice

Experimental design. Figure 3 evaluates predictive uncertainty when the observations are distributed unevenly across the input domain. The training set contains 32 configurations in $[-1.2, 1.2]$: 28 observations lie in the left half, and four lie in the right half. The right half is therefore the sparsely observed region. The potential, network architecture, training support, and known energy and force noise standard deviations $\sigma_E = \sigma_F = 0.1$ are held fixed. The experiment is repeated over five paired data and initialization seeds; the figure shows the preselected seed zero.

DASL, last-layer Laplace, and MaxVar use $K = 33$ stochastic parameters and share the same deterministic mean within each seed. After selecting their coordinates, all three methods construct a dense restricted joint energy–force covariance. MaxVar in this experiment is the non-SWAG rule that selects parameters with the largest diagonal posterior-variance proxies,

$$\left(\alpha + \operatorname{diag} C_{EF}\right)^{-1}, \tag{4}$$

where C_{EF} is the diagonal joint energy–force curvature and the prior precision is $\alpha = 1$. This method is distinct from the SWAG-based MaxVar baseline used in the molecular experiments.

The method denoted by BLIPs is the supplied amortized variational toy adaptation. It initializes its predictive mean from the paired deterministic checkpoint and subsequently updates both the mean

network and a 1,285-parameter variance network for 5,000 additional steps. Predictive moments are estimated using 1,000 Monte Carlo draws. Consequently, the three Laplace methods form a fixed-mean comparison, whereas BLIPs has a separately optimized predictive mean.

D.2 Additional Details for Ammonia OOD Benchmark

Data and evaluation sets. The released ammonia training file contains 78 configurations. We use the first 64 configurations for model fitting and the remaining 14 for checkpoint validation. The official test set contains 129 configurations spanning the ammonia inversion pathway. Because the fitting data are concentrated near equilibrium, the test set represents a controlled geometric and energetic distribution shift. Energies and forces are reported in kcal/mol and kcal/mol/Å, respectively.

Deterministic model. The predictive model is a PaiNN network (Schütt et al., 2021) with three interaction blocks, hidden width 128, a 5 Å cutoff, and 16 radial basis functions. It is trained for 500 epochs with batch size 64 using

$$\mathcal{L}_{\text{train}} = 0.1|E_\theta - E| + \text{mean}|F_\theta - F|. \tag{5}$$

Forces are obtained as the negative derivative of the predicted energy. Checkpoint selection uses the separate 14-configuration validation set.

All Laplace approximations are constructed around the same trained checkpoint and leave its parameters unchanged. They therefore share the same predictive mean and point accuracy. Differences among these methods arise only from the selected stochastic subspace and its posterior covariance.

Laplace methods and parameter budgets. Last-layer Laplace places a dense posterior over 65 parameters. DASL, MaxVar, and random selection use $K = 64$, approximately matching the last-layer parameter count. Every selected subnetwork uses a dense restricted joint energy–force GGN covariance.

DASL selects parameters using their joint energy–force sensitivity. MaxVar use SWAG (Maddox et al., 2019) to estimate parameter variances: marginal parameter variances are estimated from 40 SGD snapshots of a separate model copy, using learning rate 10^{-3}, momentum 0.9, batch size 16, and no weight decay. Random-selection results average five independently sampled subnetworks at each K.

Alternative uncertainty methods. BLIP (Coscia et al., 2025) uses the pinned upstream local-reparameterization and KL components, with predictive moments estimated from 100 stochastic predictions. Unlike the Laplace methods, BLIP adapts part of the predictive model and therefore has a separately fitted mean. The deep ensemble contains five independently trained PaiNN members.

The Gaussian mixture model fits a two-component full-covariance mixture to atom-level latent representations. Its negative latent log density is used as an uncertainty-ranking score. Because this quantity is not a predictive force variance, we do not assign GMM an NLL or CRPS value.

Hyperparaters. The NLL and CRPS values in Table 1 use the same predictive noise floor for every method:

$$\sigma_E = 15 \text{ kcal/mol}, \qquad \sigma_F = 70 \text{ kcal/mol/Å}. \tag{6}$$

The total predictive variance is

$$v_t^{\text{pred}} = v_t^{\text{epi}} + v_t^{\text{native}} + \sigma_t^2. \tag{7}$$

D.2.1 Effect of subnetwork size

Table 4 reports the DASL budget ablation using the shared noise scale from Equation equation 6. Increasing K improves force NLL and CRPS, while the uncertainty-error correlation changes comparatively little and is not monotonic at the intermediate budgets.

The small change in ρ_F indicates that $K = 64$ already captures much of the information required to rank high-error structures. Increasing the subnetwork size mainly changes the scale and covariance of the predictive distribution.

Table 4: **DASL budget sensitivity on the official ammonia test set.** Coverage is for nominal 90% predictive intervals. Force CRPS is in kcal/mol/Å.

K	Force NLL ↓	Force CRPS ↓	Coverage	ρ_F ↑
64	5.707	36.753	0.891	0.905
128	5.705	36.744	0.891	0.901
256	5.702	36.731	0.892	0.905
512	5.696	36.703	0.892	0.914

D.3 Additional Details For 3BPA Results

Dataset and evaluation splits. We train the interatomic potential on 100 configurations sampled at 300 K and use an additional 50 configurations for checkpoint selection and uncertainty-hyperparameter tuning. The model is evaluated on four test sets containing 1,669 configurations at 300 K and 2,138 configurations at 600 K.

Model and training details. The deterministic potential is a PaiNN model with hidden dimension 128, three interaction blocks, a 5 Å cutoff, and 16 radial basis functions. It is trained for 1,000 epochs with batch size 16 using the joint energy–force objective

$$\mathcal{L} = 0.1|E - E^\star| + \frac{1}{3N}\sum_{i=1}^{N}\sum_{d=1}^{3}|F_{i,d} - F^\star_{i,d}|. \tag{8}$$

Energies are reported in $\mathrm{kcal\,mol^{-1}}$ and forces in $\mathrm{kcal\,mol^{-1}\,Å^{-1}}$.

All Laplace methods use the same frozen trained network and therefore share the same predictive mean. They differ only in the parameters retained in the posterior approximation. Last-layer Laplace contains 65 parameters. The primary DASL, MaxVar, and random-subnetwork comparison uses $K = 64$, which approximately matches this parameter count. We additionally evaluate DASL with $K \in \{64, 128, 256, 512\}$.

Hyperparameter selection. Observation-noise scales are selected by fivefold geometry-grouped cross-validation. The configurations are assigned to 20 groups using average-linkage clustering of their within-molecule pair-distance representations. Complete groups are held out during each fold, including during checkpoint selection. The candidate energy and force noise scales are

$$\sigma_E, \sigma_F \in \{0.1, 0.3, 1, 3, 10\}.$$

The selection objective is the equally weighted average of the energy and force marginal negative log likelihoods. This procedure selects $\sigma_E = 1$ and $\sigma_F = 3$ for all Laplace variants. Prior precision is optimized conditionally for each posterior approximation using the frozen trained checkpoint.

Evaluation metrics. Forces are evaluated using negative log likelihood (NLL), continuous ranked probability score (CRPS), and Spearman rank correlation. For a configuration x, the uncertainty score and force-error score are

$$u_F(x) = \frac{1}{3N_x}\sum_{i=1}^{N_x}\sum_{d=1}^{3}\mathrm{Var}_{\mathrm{epi}}[F_{i,d}(x)]\,, \tag{9}$$

and

$$e_F(x) = \frac{1}{3N_x}\sum_{i=1}^{N_x}\sum_{d=1}^{3}\left(F_{i,d}(x) - F^\star_{i,d}(x)\right)^2. \tag{10}$$

We report Spearman's correlation $\rho = \mathrm{corr}_{\mathrm{rank}}(u_F, e_F)$. Unlike NLL and CRPS, this ranking metric depends only on epistemic variance and is unchanged by the addition of a constant observation-noise floor.

Table 5: Effect of the DASL subnetwork size on 3BPA. Lower NLL and CRPS are better; higher Spearman correlation ρ is better.

Test set	K	Force NLL ↓	Force CRPS ↓	ρ ↑
300 K	64	**2.213**	**1.093**	0.515
	128	2.214	1.094	0.538
	256	2.216	1.095	0.536
	512	2.220	1.098	**0.544**
600 K	64	2.870	1.967	0.622
	128	2.850	1.965	**0.636**
	256	2.823	1.961	0.627
	512	**2.779**	**1.955**	0.630

D.3.1 Effect of the DASL subnetwork size.

Table 5 reports the DASL ablation over the number of selected parameters. Increasing K has little effect on the in-distribution 300 K results, but improves force NLL and CRPS under temperature shifts. Uncertainty ranking also improves from 0.622 to 0.630. These results indicate that retaining additional sensitivity-selected directions becomes more useful as the test distribution moves farther from the training distribution.

On the dihedral split, increasing the subnetwork size does not produce the same improvement. Predictive scores remain nearly unchanged, and the rank correlation is low for every value of K. Thus, a larger DASL subnetwork is most beneficial for the temperature shifts represented in these experiments, but does not by itself resolve uncertainty ranking under the distinct conformational shift.

D.4 Post-hoc Uncertainty and Fine-tuning on MACE-OFF23-small

Dataset and pretrained model. We evaluate uncertainty estimation for the released MACE-OFF23-small interatomic potential Kovács et al. (2023). The checkpoint contains 694,320 floating-point parameters. Reference energies and forces are taken from the PubChem portion of the released MACE-OFF23 evaluation archive.

We restrict the data to neutral, connected molecules containing at most 24 atoms. Canonical molecular identities are constructed after removing atom mappings and stereochemical annotations, and one conformer is sampled for each selected molecular group. Within each split, molecular groups are disjoint across the adaptation, validation, and test partitions.

We construct five data splits, each containing 64 adaptation structures, 32 validation structures, and 128 test structures. All five splits use the same released MACE checkpoint. The splits may overlap with one another; therefore, the reported standard deviations quantify variation across post-hoc molecular data splits rather than variation across independently pretrained foundation models. We did not independently audit possible overlap between these molecules and the original MACE pretraining set.

Comparison protocol. The experiment contains two categories of uncertainty estimators. The Laplace methods leave the pretrained MACE parameters unchanged and attach a Gaussian posterior to selected coordinates. In contrast, BLIP (Coscia et al., 2025), Bayes by Backprop (Blundell et al., 2015), and the ensemble (Lakshminarayanan et al., 2017) baselines update some or all of the pretrained parameters. All results report the mean and sample standard deviation across the five molecular splits. Energy quantities are measured in eV and force quantities in eV/Å.

Laplace approximations. All Laplace variants use the frozen pretrained predictive mean and a selected-coordinate generalized Gauss–Newton covariance constructed from the joint energy–force likelihood. Conditional prior precision is fitted on the adaptation partition using frozen-center empirical Bayes.

DASL, MaxVar, and random selection are evaluated with

$$K \in \{64, 128, 256, 512\}.$$

We additionally use $K = 112$ for the matched comparison to the last-layer laplace in the main text. Random-selection results are averaged over three independently sampled subnetworks within each split.

The MaxVar baseline ranks marginal parameter variances estimated from 40 SGD snapshots of a separate model copy. Snapshot collection uses learning rate 10^{-3}, momentum 0.9, batch size 16, no weight decay, and the loss

$$0.1|E - E^\star| + \text{mean}|F - F^\star|.$$

Only the parameter ranking is transferred from this optimization copy; all Laplace predictions use the original pretrained checkpoint. This is the SWAG-based MaxVar baseline used in the molecular experiments.

We define joint-readout Laplace as a posterior over the terminal linear maps of both additive MACE interaction readouts. It contains 96 weights from the first readout and 16 weights from the second readout, for 112 parameters in total. A separate final-readout control contains only the latter 16 weights. Neither posterior includes all 1,648 parameters in the complete MACE readout modules.

BLIP adaptation. Because the original BLIP implementation does not directly support all irreducible-representation-indexed MACE operations, we construct a MACE-compatible BLIP adaptation using the upstream local-reparameterization layer and KL implementation Coscia et al. (2025). Eight scalar radial linear layers are converted while all equivariant tensor products and readout operations retain their original form.

A shared posterior network with two 64-unit SiLU hidden layers maps squared edge distances to eight variance multipliers. The posterior network contains 4,808 parameters. Its 66,560 radial mean weights are fine-tuned, while all non-radial parameters and buffers remain fixed. The resulting method is a radial MACE adaptation of BLIP rather than a previously published native BLIP–MACE architecture.

The BLIP model is trained for 50 epochs with batch size 8 using AdamW without weight decay. The learning rates are 10^{-5} for the radial mean weights and 10^{-4} for the posterior network. The training objective is

$$\mathcal{L}_{\text{BLIP}} = 0.1\ \text{Huber}(E, E^\star) + \text{mean}\,\text{Huber}(F, F^\star) + 10^{-3}\mathcal{L}_{\text{KL}}, \tag{11}$$

where the Huber threshold is 0.01. The prior probability is 0.01, and the posterior probabilities are constrained to $[10^{-5}, 0.1]$. We use the final training epoch and do not perform test-based checkpoint selection. Predictive moments are estimated using 100 Monte Carlo samples.

A deterministic radial fine-tuning control updates the same 66,560 radial mean parameters using the same data, optimizer, and data loss, but without the variational posterior.

Ensemble baselines. The readout ensemble contains five members. Each member is initialized from the released checkpoint, trained on an independent bootstrap sample of the 64 adaptation structures, and updates all 1,648 readout-module parameters. The remainder of the pretrained MACE model is frozen.

The fine-tuned deep ensemble also contains five bootstrap-trained members, but updates all 694,320 parameters in each member. Thus, it is an ensemble obtained by fine-tuning a common pretrained checkpoint rather than an ensemble of independently pretrained MACE models. Epistemic variance for both ensemble methods is the unbiased sample variance across the five members.

Bayes by Backprop. The full-network Bayes-by-Backprop baseline places a mean-field Gaussian posterior over all 694,320 MACE parameters, giving 1,388,640 trainable mean and scale values. The Gaussian prior is centered at the released checkpoint. For each parameter tensor, the prior standard deviation is 0.01 times the larger of the checkpoint root-mean-square magnitude and 0.01.

Training uses 50 epochs and a Gaussian energy–force likelihood with fixed training standard deviations of 0.03 eV and 0.1 eV/Å, respectively. Prediction uses 100 globally coherent weight samples. Forces are computed as derivatives of each sampled energy. This baseline is our MACE adaptation of Bayes by Backprop rather than an unchanged upstream implementation.

Table 6: MACE-OFF23-small results across five molecular data splits, reported as mean ± sample standard deviation. Energy and force quantities are in eV and eV/Å, respectively. NLL and CRPS include validation-selected observation noise. Laplace methods preserve the checkpoint mean, whereas the adapted baselines update it.

Method	K	E MAE ↓	F MAE ↓	E NLL ↓	F NLL ↓	E CRPS ↓	F CRPS ↓	ρ_F ↑
Frozen MACE + noise	0	0.0394 ± 0.0053	0.0475 ± 0.0035	-0.299 ± 1.784	-0.569 ± 0.837	0.0340 ± 0.0059	0.0428 ± 0.0025	—
Joint-readout Laplace	112	0.0394 ± 0.0053	0.0475 ± 0.0035	-0.911 ± 0.842	-0.779 ± 0.496	0.0305 ± 0.0049	0.0428 ± 0.0025	0.464 ± 0.093
Random	512	0.0394 ± 0.0053	0.0475 ± 0.0035	-0.939 ± 0.897	-0.960 ± 0.217	0.0304 ± 0.0049	0.0427 ± 0.0024	0.573 ± 0.054
MaxVar (SWAG)	512	0.0394 ± 0.0053	0.0475 ± 0.0035	-1.482 ± 0.112	-1.058 ± 0.040	0.0322 ± 0.0046	0.0434 ± 0.0023	0.601 ± 0.067
DASL	512	0.0394 ± 0.0053	0.0475 ± 0.0035	-1.513 ± 0.070	-1.065 ± 0.030	0.0321 ± 0.0047	0.0433 ± 0.0023	0.612 ± 0.046
Deterministic radial fine-tuning	—	0.0410 ± 0.0058	0.0476 ± 0.0035	-0.201 ± 1.930	-0.569 ± 0.836	0.0348 ± 0.0057	0.0428 ± 0.0025	—
BLIP	—	0.0402 ± 0.0056	0.0477 ± 0.0035	-1.073 ± 0.762	-0.961 ± 0.259	0.0307 ± 0.0052	0.0400 ± 0.0049	0.439 ± 0.039
Readout ensemble	—	0.0409 ± 0.0052	0.0475 ± 0.0035	1.215 ± 3.808	-0.570 ± 0.836	0.0340 ± 0.0040	0.0428 ± 0.0025	0.486 ± 0.076
Fine-tuned deep ensemble	—	0.0424 ± 0.0052	0.0478 ± 0.0036	0.344 ± 1.938	-0.611 ± 0.760	0.0327 ± 0.0049	0.0429 ± 0.0026	0.471 ± 0.041
Full-network BBB	—	0.0550 ± 0.0023	0.0476 ± 0.0035	-0.183 ± 0.032	-0.941 ± 0.227	0.0904 ± 0.0040	0.0400 ± 0.0050	0.536 ± 0.045

Observation noise and evaluation. For every fitted method and split, the reported observation-noise scales are chosen using only the validation partition. We search over

$$\sigma_E \in \{0.01, 0.03, 0.1\}\ \text{eV}, \qquad \sigma_F \in \{0.03, 0.1, 0.3\}\ \text{eV/Å}.$$

The selected pair minimizes the equally weighted energy and force marginal NLL. The test labels do not enter posterior fitting or hyperparameter selection.

Predictive NLL and CRPS use total variance,

$$\mathrm{Var}(y \mid x, \mathcal{D}) = \mathrm{Var}_{\mathrm{epi}}(y \mid x, \mathcal{D}) + \sigma^2_{\mathrm{obs}}. \tag{12}$$

For BLIP and BBB, the Monte Carlo predictive distributions are summarized using moment-matched Gaussian marginals.

Force MAE pools all Cartesian components. Force NLL and CRPS are first averaged across the force components of each molecule and subsequently averaged across molecules. The force ranking metric is

$$\rho_F = \mathrm{corr}_{\mathrm{rank}}\left(\frac{1}{3N_x}\sum_{i,d}\mathrm{Var}_{\mathrm{epi}}[F_{i,d}(x)], \frac{1}{3N_x}\sum_{i,d}(F_{i,d}(x) - F^{\star}_{i,d}(x))^2\right). \tag{13}$$

Because ρ_F uses epistemic variance only, it is unaffected by the addition of a constant observation-noise floor.

Complete comparison. Table 6 reports the complete comparison. For the subnetwork methods, we show $K = 512$ because it provides the strongest predictive NLL in the budget ablation. Joint-readout Laplace retains its natural parameter count of 112.

The Laplace methods preserve the pretrained predictor and consequently have identical energy and force MAE. At $K = 512$, DASL achieves the lowest mean energy and force NLL among the evaluated methods and the highest mean force uncertainty–error correlation. Its results are close to MaxVar, and the five-split experiment does not establish a statistically significant difference between these two selection strategies. Both informed selectors produce a higher mean correlation than random selection.

BLIP obtains a lower force CRPS than the reported Laplace variants, whereas its force uncertainty–error correlation is lower. BBB obtains a similarly low force CRPS and a moderate ranking correlation, but its energy MAE and energy CRPS are substantially worse. Neither the readout ensemble nor the fine-tuned deep ensemble improves consistently over the frozen MACE predictor in this limited-data setting.

These comparisons demonstrate that no single metric gives a uniform ordering. DASL performs strongly in predictive likelihood and uncertainty ranking while leaving the pretrained mean unchanged. BLIP and BBB produce lower force CRPS but require stochastic fine-tuning and do not provide the same ranking performance.

Subnetwork-size ablation. Table 7 compares DASL, MaxVar, and random selection over increasing subnetwork budgets. The predictive mean is identical for every row.

Table 7: MACE subnetwork-budget ablation across five molecular data splits. Values are mean $\pm$ sample standard deviation. Predictive scores include validation-selected observation noise, while ρ_F uses epistemic variance only.

Method	K	E NLL ↓	F NLL ↓	E CRPS ↓	F CRPS ↓	ρ_F ↑
Random	64	0.811 ± 3.221	-0.659 ± 0.689	0.0330 ± 0.0045	0.0428 ± 0.0025	0.509 ± 0.057
MaxVar (SWAG)	64	-0.187 ± 1.895	-0.804 ± 0.442	0.0304 ± 0.0049	0.0428 ± 0.0025	0.596 ± 0.061
DASL	64	-0.638 ± 1.272	-0.728 ± 0.581	0.0304 ± 0.0050	0.0428 ± 0.0025	0.611 ± 0.044
Random	128	0.389 ± 2.531	-0.750 ± 0.529	0.0307 ± 0.0050	0.0428 ± 0.0025	0.533 ± 0.069
MaxVar (SWAG)	128	-0.956 ± 0.870	-0.915 ± 0.283	0.0304 ± 0.0049	0.0427 ± 0.0025	0.600 ± 0.071
DASL	128	-1.146 ± 0.636	-0.907 ± 0.272	0.0303 ± 0.0050	0.0428 ± 0.0024	0.614 ± 0.042
Random	256	-0.315 ± 1.589	-0.863 ± 0.348	0.0305 ± 0.0049	0.0427 ± 0.0025	0.550 ± 0.064
MaxVar (SWAG)	256	-1.333 ± 0.378	-1.027 ± 0.112	0.0307 ± 0.0047	0.0428 ± 0.0024	0.606 ± 0.061
DASL	256	-1.477 ± 0.192	-1.031 ± 0.098	0.0306 ± 0.0047	0.0428 ± 0.0024	0.618 ± 0.041
Random	512	-0.939 ± 0.897	-0.960 ± 0.217	0.0304 ± 0.0049	0.0427 ± 0.0024	0.573 ± 0.054
MaxVar (SWAG)	512	-1.482 ± 0.112	-1.058 ± 0.040	0.0322 ± 0.0046	0.0434 ± 0.0023	0.601 ± 0.067
DASL	512	-1.513 ± 0.070	-1.065 ± 0.030	0.0321 ± 0.0047	0.0433 ± 0.0023	0.612 ± 0.046

Table 8: Equal-parameter MACE readout controls across five molecular splits. Final-readout Laplace contains 16 parameters; joint-readout Laplace contains 112 parameters.

Method	K	E NLL	F NLL	E CRPS	F CRPS	ρ_F
Final-readout Laplace	16	0.694 ± 3.062	-0.600 ± 0.790	0.0307 ± 0.0050	0.0428 ± 0.0025	0.610 ± 0.036
DASL	16	0.907 ± 3.360	-0.589 ± 0.804	0.0307 ± 0.0050	0.0428 ± 0.0025	0.629 ± 0.057
Joint-readout Laplace	112	-0.911 ± 0.842	-0.779 ± 0.496	0.0305 ± 0.0049	0.0428 ± 0.0025	0.464 ± 0.093
DASL	112	-1.053 ± 0.758	-0.884 ± 0.306	0.0303 ± 0.0050	0.0428 ± 0.0024	0.608 ± 0.045

Increasing the subnetwork budget improves the mean NLL of DASL and MaxVar. For DASL, energy NLL improves from -0.638 at $K = 64$ to -1.513 at $K = 512$, while force NLL improves from -0.728 to -1.065. However, the force CRPS does not improve monotonically and slightly increases at the largest budget. The force rank correlation also remains relatively stable, between 0.611 and 0.618. Thus, the likelihood improvement with increasing K does not correspond to a comparable improvement in every uncertainty metric.

Across all four budgets, DASL and MaxVar produce higher mean force rank correlation than random selection. DASL has the highest mean correlation at each budget, although its difference from MaxVar is small relative to the variation across splits. The results support sensitivity- or variance-based selection over random selection, but do not establish uniform superiority of DASL over MaxVar.

Matched readout controls. Table 8 compares DASL against the two MACE readout definitions at identical parameter counts.

At $K = 16$, DASL has a slightly higher mean force rank correlation than final-readout Laplace, while the readout posterior has slightly better NLL. At $K = 112$, DASL improves both energy and force NLL and produces a higher mean force rank correlation than joint-readout Laplace. The comparison suggests that the location of the selected parameters matters in addition to the number of uncertain parameters.

Additional per-atom force-error and uncertainty maps. Figures 5 and 6 provide two further held-out molecular examples. The Laplace methods use the same frozen MACE predictive mean and $K = 112$ uncertain weights; BLIP uses an adapted mean. Uncertainty is normalized by each method's median within a molecule, so these maps compare where methods place uncertainty, not its absolute scale. The within-molecule correlations reported in the figures are distinct from the across-molecule ρ_F reported in the tables.

In Figure 5, a carbon and an oxygen have the largest force errors (0.100 and 0.099 eV/Å). DASL ranks these same atoms as its two most uncertain, whereas the other methods place their uncertainty maximum on a nitrogen or a lower-error carbon. The per-atom uncertainty–error correlation is 0.43 for DASL, compared with 0.19 for MaxVar, -0.08 for last-layer Laplace, and 0.16 for BLIP.

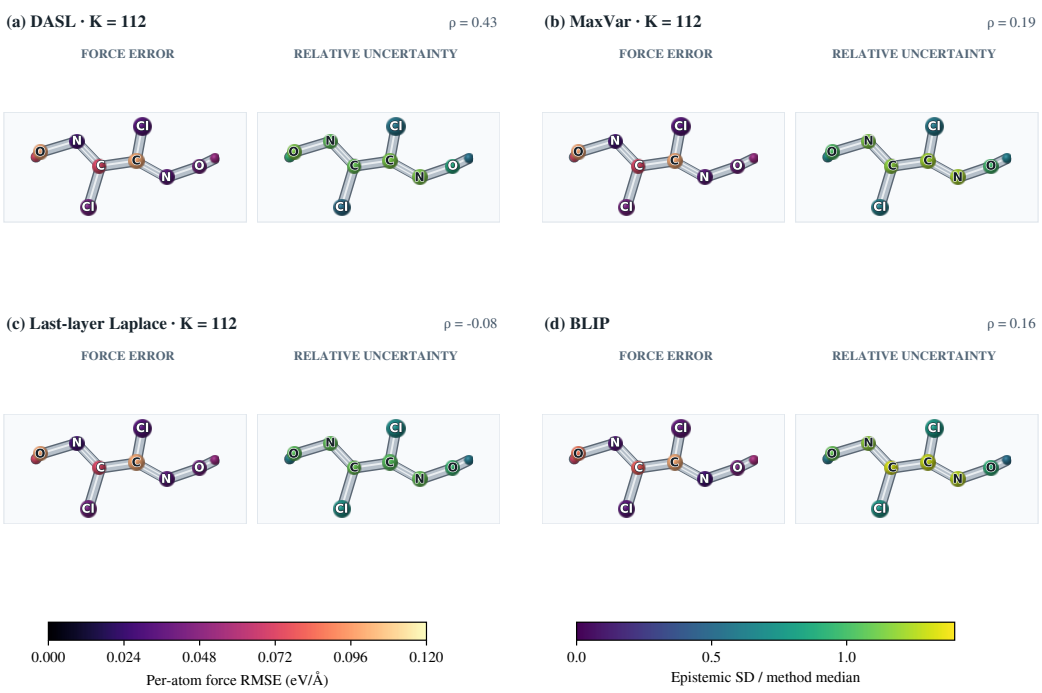

Figure 5: Per-atom force error and uncertainty for a second test molecule, `ON=C(Cl)C(Cl)=NO`, from the seed-0 MACE-OFF23 split. Each card pairs force error against DFT (left) with relative epistemic force uncertainty (right): (a) DASL, (b) MaxVar, (c) last-layer Laplace, and (d) BLIP.

In Figure 6, all methods assign their highest uncertainty to phosphorus, which also has the largest force error. DASL additionally ranks the bonded carbon second, matching its frozen-MACE error rank. MaxVar emphasizes the fluorines, while last-layer Laplace gives greater relative uncertainty to a chlorine. DASL has the highest per-atom correlation (0.68, compared with 0.32, 0.36, and 0.32), although no method perfectly orders the atoms. Both molecules were selected post hoc for illustration; the five-split test-set results, rather than these examples, support comparative conclusions.

E Discussion of BLIP's Weak Sensitivity to Input Data Density

In our controlled experiment, we observe only a small distinction in BLIP's predictive uncertainty between densely and sparsely observed regions. We provide a structural interpretation of this behavior based on its variational formulation.

E.1 BLIP uses an input-conditioned variational distribution

BLIP introduces stochastic weights into the message and update functions of an MPNN. Its generative construction places independent priors over the message and update weights and assumes further factorization across layers. Importantly, however, the approximate posterior over these weights is conditioned on the input graph. In particular, BLIP writes

$$q_\phi\left(\theta_l^M, \theta_l^U \mid h^0, a\right) = q_\phi\left(\theta_l^M \mid h^0, a\right) q_\phi\left(\theta_l^U \mid h^0\right),$$

with Gaussian variational factors whose variances are controlled by adaptive coefficients α_l and β_l. These coefficients are functions of the initial node and edge features and are produced by a shared inference network. At inference time, the coefficients are computed for the query structure and then reused across Monte Carlo samples.

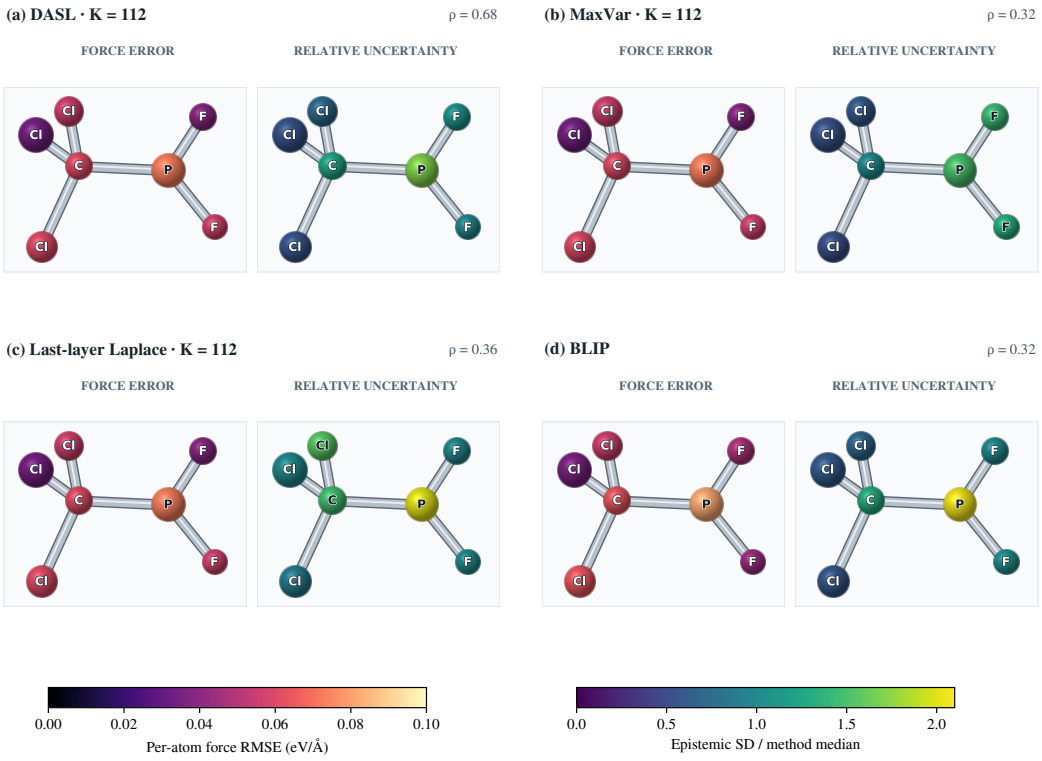

Figure 6: Per-atom force error and uncertainty for a seven-atom MACE-OFF23 test molecule, `FP(F)C(Cl)(Cl)Cl`. Each method card shows force error against DFT (left) beside relative epistemic force uncertainty (right): (a) DASL, (b) MaxVar, (c) last-layer Laplace, and (d) BLIP.

Implicit local-latent structure in BLIP. A conventional BNN places a prior over a single global weight vector θ, so that all observations are conditionally independent given the same latent weights,

$$p(\mathcal{D}, \theta) = p(\theta) \prod_{i=1}^{N} p(y_i \mid x_i, \theta).$$

Although the prior may factorize across weight coordinates, it does not factorize across inputs: the same θ generates predictions for every configuration, and all observations update one common posterior $p(\theta \mid \mathcal{D})$. DASL preserves this global structure on the selected parameter subspace. BLIP instead learns an input-conditioned variational distribution $q_\phi(\theta \mid x)$, regularized toward an input-independent Gaussian prior. A natural probabilistic interpretation is therefore to associate each input with a local stochastic weight variable θ_i. Since BLIP does not specify a correlated joint prior over $\{\theta_i\}$ across configurations, the corresponding local-latent model takes the factorized form

$$p(\theta_{1:N}) = \prod_{i=1}^{N} p(\theta_i), \qquad p(y_{1:N} \mid x_{1:N}, \theta_{1:N}) = \prod_{i=1}^{N} p(y_i \mid x_i, \theta_i).$$

The resulting posterior also factorizes across inputs. Thus, observations at one configuration do not directly update the latent stochastic weights associated with another configuration. The amortization network shares parameters across inputs and can learn similar uncertainty for similar structures, but this sharing occurs through the deterministic map $x \mapsto q_\phi(\theta \mid x)$, rather than through posterior covariance of a shared latent weight vector.

In a conventional BNN or DASL, all observations are coupled through a shared latent parameter vector θ. The posterior has the form

$$p(\theta \mid \mathcal{D}) \propto p(\theta) \prod_{i=1}^{N} p(y_i \mid x_i, \theta),$$

so every observation contributes information about the same latent variable. As additional observations accumulate, they progressively constrain the shared posterior, and this contraction propagates to predictions at other configurations through their dependence on the same θ. Under the local-latent interpretation of BLIP, the situation is different. Associate each input x_i with a stochastic parameter variable θ_i, and suppose the prior and observation model factorize as

$$p(\theta_{1:N}) = \prod_{i=1}^{N} p(\theta_i),$$

and

$$p(y_{1:N} \mid x_{1:N}, \theta_{1:N}) = \prod_{i=1}^{N} p(y_i \mid x_i, \theta_i).$$

The resulting posterior also factorizes,

$$p(\theta_{1:N} \mid \mathcal{D}) = \prod_{i=1}^{N} p(\theta_i \mid x_i, y_i).$$

Now consider a query configuration x_* with associated latent θ_*. If the prior contains no dependence between θ_* and the training latents $\theta_{1:N}$, then

$$p(\theta_*, \theta_{1:N}) = p(\theta_*) \prod_{i=1}^{N} p(\theta_i).$$

Since the training observations depend only on $\theta_{1:N}$,

$$p(\theta_* \mid \mathcal{D}, x_*) = p(\theta_* \mid x_*).$$

Thus, observations at other configurations do not directly update the latent variable associated with the query. In particular, increasing the number of nearby observations does not, by itself, reduce the

posterior uncertainty of θ_*. The missing ingredient is a statistical dependency that allows information from the observed latents to propagate to the query latent. This contrasts directly with a shared Bayesian posterior. In DASL,

$$\Lambda_S = P_{0,S} + \sum_{i=1}^{N} J_{\Phi,S}(x_i)^\top \Sigma_{\text{obs}}^{-1} J_{\Phi,S}(x_i),$$

so every observation contributes to the precision of the same stochastic parameter vector θ_S. When multiple observations constrain similar directions, their contributions accumulate explicitly, producing posterior contraction. The induced covariance

$$K_S(x, x') = J_{\Phi,S}(x) \Lambda_S^{-1} J_{\Phi,S}(x')^\top$$

then propagates this information across configurations. For BLIP, the shared amortization network can still learn to assign different variational variances to different inputs. Therefore, BLIP may empirically learn lower uncertainty in well-supported regions. However, this behavior is learned by the deterministic mapping

$$x \mapsto q_\phi(\theta_x \mid x),$$

rather than arising from posterior conditioning through a correlated latent model. Consequently, contraction with increasing local observation density is not structurally guaranteed: without cross-input latent dependence, there is no Bayesian information-propagation path through which additional nearby observations directly reduce uncertainty at the query. This provides a natural interpretation of our controlled experiment. We observe only a weak distinction in BLIP's uncertainty between densely and sparsely sampled regions. Under the independent local-latent interpretation above, this behavior is consistent with the absence of an explicit mechanism for propagating posterior information from neighboring observations to the query latent.