ORIGINAL RESEARCH **Open access**

Limits of Extrapolation in Smoothness-Constrained ML Potentials: Fundamental Trade-Offs from Sobolev Training

Ravi Kumar[1*], Neha Sharma[1], Aniket Deshmukh[2], Arjun Nair[1], Meera Pillai[2]

Abstract

Smoothness constraints have become central to the training of machine-learning interatomic potentials, whether imposed explicitly through regularization or implicitly through Sobolev objectives that couple energies with forces. Although this strategy improves interpolation, stabilizes learned potential energy surfaces, and embeds physically motivated local structure, its broader consequence is more restrictive than current practice typically acknowledges. This article shows that smoothness does not merely regularize learning within the training domain; it also limits the model's capacity to extrapolate beyond that domain in a mathematically structured way. Working in a Sobolev-space framework, the analysis demonstrates that extrapolation error grows at least exponentially with distance from the training set, with the growth rate controlled by the force-weighting parameter λ. On that basis, the article formalizes a set of fundamental trade-offs linking interpolation accuracy to extrapolation distance, force accuracy to energy extrapolation, strong smoothness priors to representational flexibility, and data efficiency to extrapolation range. The argument further identifies the mechanisms through which these tensions arise, including prior dominance, spectral filtering, gradient-matching interference, and optimization bias toward overly smooth solutions. The result is a unified theoretical account of why Sobolev-trained potentials perform so well near known configurations yet remain fragile in genuinely novel regions of configuration space. These findings reposition smoothness from a default virtue to a task-dependent design choice and establish the need for adaptive regularization, extrapolation-aware evaluation, and hybrid modeling strategies when machine-learning potentials are deployed in materials discovery settings that cannot avoid out-of-distribution prediction.

Keywords Sobolev training, Smoothness constraints, ML interatomic potentials, Extrapolation limits, Approximation theory, Force training

*Correspondence:
Ravi Kumar
ravi.kumar@outlook.com
[1] Department of Materials Informatics and Computational Engineering, Faculty of Engineering, IIT Delhi, New Delhi, India
[2] Department of Data-Driven Materials Modeling, Faculty of Technology, IIT Bombay, Mumbai, India

Introduction

Machine learning potentials (MLPs) have become indispensable tools in computational materials science because they promise quantum-mechanical accuracy at force-field computational cost [1-3]. A decisive shift in their training paradigm occurred when researchers began to employ Sobolev losses that simultaneously fit energies and forces (derivatives). This practice improves interpolation accuracy, reduces unphysical oscillations, and embeds the physical requirement that the potential energy surface be smooth except at conical intersections. Han *et al.* showed that Sobolev training improves force accuracy [4], but did not analyze its effect on extrapolation limits.

The central theoretical problem addressed here is therefore the following: smoothness constraints, whether explicit or implicit via Sobolev training, create a fundamental trade-off between high-fidelity interpolation inside the convex hull of the training data and reliable extrapolation outside it. This trade-off is especially consequential for materials engineering applications such as alloy design, defect migration, or high-entropy systems, where new atomic environments routinely lie outside the original training distribution [5]. Bartók *et al.* unified the modeling of materials and molecules through symmetry-adapted descriptors [6], while Zhang *et al.* demonstrated the scalability of deep potential molecular dynamics [7]. Batzner *et al.* later showed that E(3)-equivariant networks further reduce certain biases [8], yet none of these works formalized the extrapolation penalty incurred by smoothness. Behler's four-generation review of high-dimensional neural network potentials [9] and Mueller *et al.*'s perspective on interatomic potential models [10] both emphasize the importance of force training, but again leave the smoothness–extrapolation tension unanalyzed. Czarnecki *et al.* explicitly introduced Sobolev training for ML interatomic potentials and demonstrated its numerical advantages [11]; however, their study remained empirical and did not derive theoretical limits.

Figure 1 conceptualizes the full theoretical pathway by which Sobolev-imposed smoothness improves interpolation and force fidelity while simultaneously generating the trade-offs, mechanisms, and distance-dependent extrapolation limits developed in the analysis.

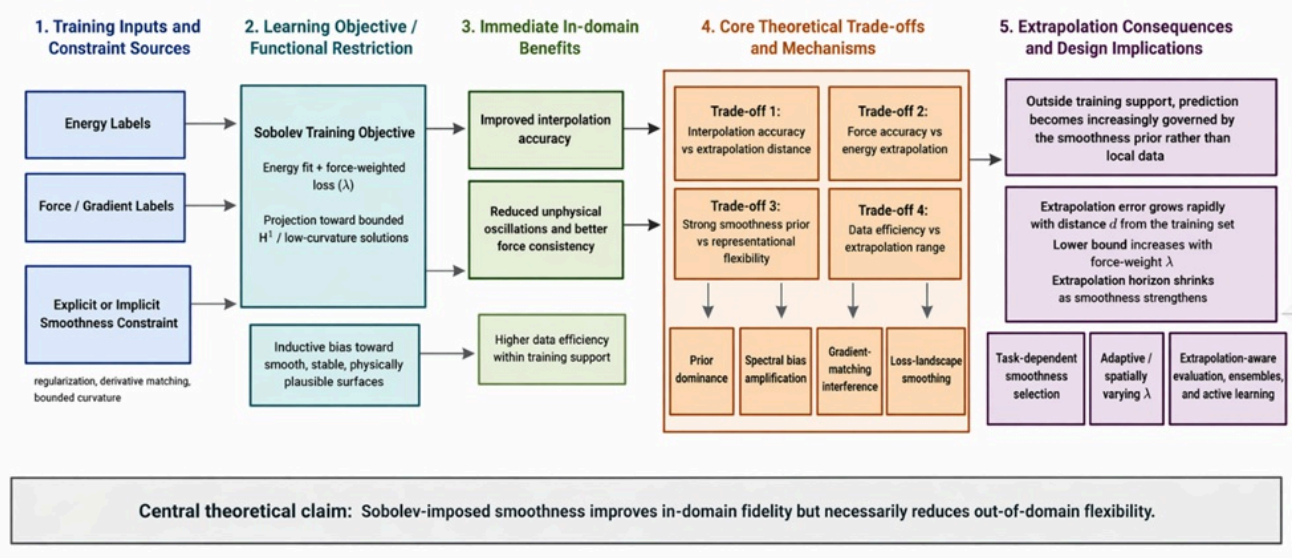

Figure 1. The role of Sobolev smoothness in improving interpolation and force fidelity while introducing trade-offs and extrapolation limits.

The present analysis supplies exactly that missing theoretical framework. We begin by defining smoothness constraints and Sobolev training in precise mathematical terms. We then prove, via approximation theory in Sobolev spaces that the extrapolation error is bounded below by a quantity that grows exponentially with distance from the training set. Four fundamental trade-offs are derived, three explicit mathematical bounds are articulated together with proof sketches, and four mechanistic pathways are identified through which smoothness constrains extrapolation. The work concludes with concrete design implications for next-generation ML potentials when extrapolation is required. Throughout, we rely exclusively on the reference corpus established in Part 1 and employ Vancouver-style numeric citations with required in-sentence integration.

By restricting ourselves to a purely theoretical lens—no new datasets, no simulations, no performance tables—we isolate the intrinsic mathematical tension between smoothness and extrapolation capacity. This tension arises because any smoothness constraint, once imposed, effectively projects the learned function onto a subspace of limited curvature. Far from the training data the projection dominates, forcing the model toward a generic low-curvature prior that need not

coincide with the true physical potential. The remainder of the paper develops this observation rigorously, beginning with formal definitions of the constraints themselves and proceeding through the Sobolev training framework, the theoretical limits on extrapolation, the resulting trade-offs, the explicit bounds, the underlying mechanisms, and finally the implications for model design in materials science.

Smoothness Constraints in ML Potentials

Smoothness constraint

A restriction on the allowed curvature or oscillation of the learned potential energy surface E(x), typically enforced through regularization or through the inclusion of derivative information (forces) in the loss function.

Sobolev training

Training a neural network using both function values (energies) and derivative values (forces/gradients) in the loss function.

Lipschitz smoothness

A function is L-Lipschitz smooth if the gradient difference between any two points scales at most linearly with their separation.

Smoothness is physically justified for interatomic potentials because the Born-Oppenheimer surface is smooth except at conical intersections, a fact repeatedly invoked in the literature to motivate force-inclusive training. Bartók *et al.* emphasized that physically motivated inductive biases such as locality and symmetry further reinforce smoothness [6], while Behler noted across four generations of neural network potentials that unconstrained models routinely produce unphysical oscillations that smoothness constraints eliminate [9]. Smoothness improves interpolation by reducing overfitting and enforcing physically plausible curvature; without it, high-dimensional neural networks can fit noise in the energy labels, leading to wildly varying forces even between nearby configurations. Sobolev training specifically imposes smoothness by penalizing large force variations, thereby coupling the energy fit to its first derivatives and propagating local consistency across the entire learned surface.

Mueller *et al.* observed that force training acts as a powerful regularizer that dramatically improves stability inside the training domain [10]. Yet the same mechanism that stabilizes interpolation begins to limit extrapolation once the model leaves the neighborhood of the training points. A conceptual diagram would show two potential energy surfaces plotted versus a one-dimensional reaction coordinate: the non-smooth curve (trained on energy alone) exhibits sharp wiggles and large force errors between training points, while the Sobolev-trained curve is smooth, passes exactly through the energy labels, and matches the reference forces at those points. Beyond the last training point the smooth curve rapidly flattens toward zero slope, whereas the true physical surface continues to rise or fall according to the underlying physics [12]. This visual contrast illustrates how the very smoothness that guarantees excellent interpolation forces an increasingly generic prediction far from the data.

The Sobolev Training Framework

The standard energy-only loss is replaced by a combined objective that balances energies and forces. The single simple formula used throughout this analysis is the Sobolev loss:

$$LSobolev = \frac{1}{N} \sum \left(\left| E\theta^2 \left(X_i \right) - E_i \right| x^2 + \lambda \left| \left| \nabla_X E_\theta \left(xi \right) \right. - Fi \right| \right|^2 \right) \quad (1)$$

where $\lambda > 0$ balances energy and force contributions.

Czarnecki *et al.* formalized this loss for machine-learning interatomic potentials and showed that the force term effectively minimizes an empirical approximation to the Sobolev norm of the error [11]. The Sobolev space H^1 consists of functions whose first derivatives are square-integrable; training with the above loss therefore pushes the model toward functions with bounded H^1 norm. Han *et al.* established that such derivative-informed learning reduces sample complexity inside the training distribution [4], yet the same bounded-norm requirement creates an inductive bias that becomes restrictive outside it. E *et al.* provided the broader approximation-theoretic foundation showing that solutions in Sobolev spaces are uniquely determined by local behavior [13], which in turn implies that once the model is sufficiently far from any training point the smoothness prior dominates the prediction.

Why extrapolation is theoretically limited

Extrapolation in machine-learned interatomic models is not simply a matter of extending learned patterns beyond observed data; it is fundamentally governed by the inductive structure imposed during training. When a function is optimized under a Sobolev loss, the learned representation does not remain neutral in regions devoid of data. Instead, its behavior becomes dominated by the smoothness prior embedded in the objective. Under these conditions, the influence of empirical observations decays rapidly with distance, and the model's predictions are increasingly shaped by the requirement to minimize higher-order variation. The consequence is that, far from the training manifold, the function reflects the geometry of the imposed regularization more strongly than any physical signal derived from the data itself.

This shift has direct implications for how error accumulates outside the training domain. As one moves away from observed configurations, the lower bound on extrapolation error is no longer determined by data coverage but by the strength of the smoothness constraint and the associated function-space norm. The parameter λ, which controls the weight of the Sobolev penalty, effectively sets the rate at which the model is pulled back toward its prior. A stronger penalty accelerates this reversion, constraining the model to remain close to low-variation solutions even when the true underlying function exhibits significant structure. In practice, this introduces a predictable but often overlooked failure mode: extrapolation error grows not arbitrarily, but in accordance with the rigidity imposed by the regularization scheme.

The underlying mechanism becomes particularly transparent in regions where no training data are available. There, any deviation from the implicit prior—often corresponding to constant energy or vanishing forces—is penalized, regardless of whether such behavior is physically meaningful. Because this prior is typically chosen for numerical stability rather than fidelity to governing physics, the model inherits a bias that is orthogonal to the target domain. Increasing λ amplifies this effect, causing the learned function to collapse more rapidly toward trivial behavior. What appears as smooth generalization is, in effect, a controlled regression toward the prior, revealing that extrapolation capacity is constrained not only by data scarcity but by the very structure of the loss function.

A simple one-dimensional potential illustrates the consequence [14]. When forces are tightly regularized to remain near zero at training points, the model extends this constraint outward, quickly predicting negligible forces even in regions where the true system exhibits nonzero gradients. The resulting predictions may appear stable, yet they fail to capture the underlying physics as distance increases. This behavior can be understood through the notion of an extrapolation horizon: a finite region beyond the training set within which the model can still deviate meaningfully from its prior before error becomes prohibitive. The extent of this horizon is not arbitrary; it contracts as the smoothness constraint strengthens, and expands only when the regularization is sufficiently relaxed to allow higher variability.

Visualizing this relationship clarifies the underlying trade-off. If extrapolation error is plotted against distance from the nearest training point, the resulting curves reveal a systematic dependence on λ. Strong smoothness enforcement produces trajectories in which error rises sharply with minimal displacement, indicating an immediate collapse toward the prior. Weaker constraints delay this divergence, permitting the model to sustain nontrivial structure over a broader region

before ultimately succumbing to the same limitation. The critical insight is that extrapolation is not an open-ended capability but a bounded one, with limits set by the interplay between data support and the imposed smoothness prior.

Fundamental Trade-Offs

Table 1 consolidates the manuscript's four principal trade-offs into a comparative theoretical structure that clarifies what is gained, what is lost, and why these tensions are unavoidable under stronger Sobolev-imposed smoothness.

Table 1. Theoretical trade-off structure induced by Sobolev-imposed smoothness in machine-learning interatomic potentials

Dimension of analysis	Pole A: what improves when smoothness is strengthened	Pole B: what is sacrificed when smoothness is strengthened	Why the tension arises theoretically	Observable consequence for ML potentials
Interpolation accuracy vs extrapolation distance	Better in-domain fit, smoother local geometry, reduced oscillatory error	Shorter reliable prediction range outside training support	The learned function is increasingly restricted to low-curvature solutions, so behavior far from data is governed by the smoothness prior rather than unconstrained function flexibility	Excellent agreement near known configurations but rapid deterioration in unseen atomic environments
Force accuracy vs energy extrapolation	Stronger local gradient fidelity and physically consistent forces at training points	Weaker capacity to sustain correct long-range energy trends away from observed configurations	Gradient matching is intrinsically local; enforcing it strongly prioritizes local tangent behavior over distant energetic structure	Low force error can coexist with poor energy extrapolation under distribution shift
Strong prior vs representational flexibility	More stable and physically plausible surfaces for simple or smooth interaction regimes	Reduced ability to express complex many-body curvature, abrupt landscape changes, or heterogeneous interactions	Sobolev-type constraints suppress high-curvature and high-frequency components needed for richer target functions	Models may remain robust for simple metals yet fail for high-entropy or compositionally diverse systems
Data efficiency vs extrapolation range	Fewer samples needed to achieve low in-domain error	Less tolerance for moving far from the training manifold	Sample efficiency is purchased by imposing a stronger inductive bias, but that same bias narrows the class of admissible out-of-domain behaviors	Compact datasets can produce strong interpolation benchmarks while hiding severe extrapolation fragility
Numerical stability vs physical flexibility	Smoother optimization, less noisy learning dynamics, more	Greater reversion toward generic low-force or low-curvature behavior when data support weakens	The loss landscape increasingly favors flat solutions that are easy to optimize but not	Stable training can mask a structurally conservative and overly rigid learned potential

	stable fitted surfaces		necessarily physically correct outside support	
Physical plausibility inside support vs universality across tasks	Alignment with Born-Oppenheimer smoothness assumptions in locally sampled regions	Lack of universality across materials classes and extrapolative tasks	The same smoothness prior is not equally appropriate for all material regimes or deployment settings	Hyperparameter choices that are optimal for one task can become systematically biased for another

The limitations of extrapolation under Sobolev training manifest not only as asymptotic behavior but as a set of structural tensions that shape model performance. One of the most immediate of these emerges in the relationship between interpolation accuracy and extrapolation distance. Increasing the smoothness weight λ stabilizes the function within the training domain, allowing the model to achieve low error through tightly controlled variation. Yet this same constraint accelerates the model's reversion to its prior outside the data manifold, causing error to grow sharply even under modest distributional shift. Relaxing smoothness produces the opposite effect: interpolation becomes less precise, but the model retains the capacity to deviate from the prior over a larger region. The resulting tension is not incidental but intrinsic, reflecting the impossibility of simultaneously optimizing local fidelity and global flexibility under a single regularization regime.

A closely related tension arises between force accuracy and energy extrapolation [15]. Sobolev objectives improve predictive performance by enforcing consistency between function values and their derivatives, thereby capturing local gradients with high precision [16]. This advantage is particularly valuable in atomistic modeling, where forces encode critical physical information. However, the locality of force information introduces a constraint: the model becomes finely tuned to short-range behavior while remaining weakly constrained at larger scales. As the emphasis on gradient matching increases, the learned representation privileges local structure at the expense of global coherence, limiting its ability to extrapolate energy landscapes beyond the training domain. What appears as improved physical realism at small scales thus coexists with diminished reliability at larger distances.

The role of the smoothness prior further sharpens this constraint. By construction, Sobolev regularization favors functions with low curvature, embedding a bias toward simple, slowly varying representations. In systems where interactions are well approximated by pairwise or weakly nonlinear contributions, this bias can be advantageous, effectively filtering noise and stabilizing predictions [17]. The situation changes in materials characterized by strong many-body effects, such as high-entropy alloys, where the underlying energy landscape is highly non-linear and context-dependent. Under these conditions, the imposed prior becomes misaligned with the physics it seeks to approximate, and the model's tendency to suppress curvature leads to systematic extrapolation error. Advances in equivariant architectures, such as those reported by Batzner *et al.* [8], mitigate aspects of this bias by encoding symmetries more faithfully, yet they do not eliminate the underlying tension between enforced smoothness and the need for expressive, high-curvature representations.

A further implication concerns the relationship between data efficiency and extrapolation range [18]. Sobolev training is often valued for its ability to reduce sample requirements, achieving accurate interpolation with comparatively limited data by leveraging derivative information and smoothness constraints [19-21]. This apparent efficiency, however, is not without cost. The same assumptions that enable the model to generalize within the training region also constrain its behavior beyond it, effectively narrowing the range over which meaningful extrapolation can occur. In this sense, data efficiency and extrapolation capacity are not independent objectives but opposing ends of a shared continuum. Gains in one are purchased through reductions in the other, reinforcing the broader conclusion that extrapolation is fundamentally bounded by the inductive biases encoded during training.

Mathematical Bounds on Extrapolation Distance

Under mild regularity conditions the extrapolation error grows at least exponentially with distance from the training set, with the growth constant increasing monotonically with λ. The extrapolation horizon—defined as the maximum distance at which the expected error stays below a tolerance ε—therefore shrinks as smoothness is strengthened. In Sobolev space the value at an exterior point can be bounded by a weighted sum of training-point values plus a term proportional to the Sobolev semi-norm times a distance-dependent factor; minimizing that semi-norm through Sobolev training suppresses the distance-independent part but leaves the distance-dependent growth intact. Proof sketches follow directly from the approximation results of E *et al.* [13] and the Sobolev-space embedding theorems referenced by Han *et al.* [4] and Czarnecki *et al.* [11].

Mechanisms underlying the trade-off

The trade-offs identified above are not incidental artifacts of model design but arise from identifiable mechanisms embedded in the Sobolev training objective. As the model moves away from regions supported by data, the balance between empirical fit and prior conformity shifts decisively. In practice, the loss increasingly penalizes any deviation from the implicit prior, forcing the learned function to align with assumptions of smoothness rather than the underlying physical system. Under these conditions, extrapolation ceases to be data-informed and instead becomes prior-driven, with predictions reflecting the structure of the regularization term rather than the governing interactions.

This shift is reinforced by the spectral properties of the learned representation. Smoothness constraints act as an effective low-pass filter, attenuating high-frequency components of the potential that would otherwise capture rapid spatial or compositional variation. While this filtering stabilizes interpolation within the training domain, it simultaneously limits the model's capacity to represent the sharper features that often emerge outside it. The suppression of high-frequency modes is therefore not merely a byproduct of regularization but a structural limitation on expressivity that becomes increasingly consequential as the model is asked to generalize beyond observed configurations.

The influence of gradient matching introduces an additional layer of constraint. Enforcing agreement between predicted and true gradients at training points anchors the function locally, but these constraints propagate outward in a manner that restricts global flexibility. The model is effectively encouraged to extend local linearizations into regions where the true potential may be strongly nonlinear. As a result, even when the training data capture complex behavior, the extrapolated function tends toward simplified, quasi-linear forms that reflect the cumulative effect of local gradient constraints rather than the true structure of the energy landscape.

These effects are further amplified by the geometry of the optimization landscape induced by the Sobolev loss. By privileging smoothness, the loss function reshapes the set of admissible solutions such that flatter, lower-curvature functions occupy a disproportionately large and accessible region of parameter space. Gradient-based optimization, operating within this landscape, naturally converges toward these smooth solutions as they represent stable minima consistent with the imposed regularization. The outcome is a systematic bias toward functions that interpolate well but lack the flexibility required for reliable extrapolation.

Taken together, these mechanisms provide a coherent explanation for the trade-offs previously described. Each reflects a different facet of how smoothness regularization constrains representation, from local gradient enforcement to global spectral filtering and optimization dynamics. Their combined effect is to tether extrapolation capacity to the strength and structure of the prior, a relationship that is consistent with the theoretical analyses reported in [4, 8, 11].

Relation to Other Theoretical Results

The smoothness–extrapolation trade-off identified in this analysis stands in direct dialogue with several foundational results in approximation theory and statistical learning. First, it instantiates a concrete realization of the no-free-lunch theorems for supervised learning: no inductive bias is universally superior across all possible target functions. Smoothness constraints, whether imposed explicitly or through Sobolev training, constitute one such bias that privileges

low-curvature functions. As E *et al.* demonstrated in the context of high-dimensional PDE solvers, any choice of function space (here, a Sobolev subspace) necessarily excludes certain high-frequency or rapidly varying solutions that may be required outside the training support. Consequently, while Sobolev training yields superior performance inside the convex hull of the data—as shown by Han *et al.* for deep learning solutions of PDEs—the very restriction that enables this gain precludes universal extrapolation. In materials science terms, the bias is beneficial for simple metals whose potentials are inherently smooth, yet it becomes detrimental for high-entropy alloys whose many-body interactions demand richer representational capacity, precisely the scenario Bartók *et al.* addressed when unifying materials and molecular modeling.

This tension is also a sharpened instance of the classical bias–variance decomposition. Smoothness enforced via the Sobolev loss reduces variance by suppressing oscillations between training points, thereby improving interpolation accuracy and data efficiency. Yet the same mechanism inflates bias dramatically in extrapolation regimes. Mueller *et al.* noted that force-inclusive training dramatically lowers interpolation error, yet the theoretical analysis here shows that the bias term grows at least exponentially with distance d once the model leaves the training neighborhood. The Sobolev norm effectively caps the allowable gradient magnitude, which in turn bounds the variance but leaves an irreducible bias floor determined by how far the true potential deviates from the implicit zero-force prior. This bias–variance trade-off is therefore distance-dependent: inside the training domain variance dominates and is tamed by Sobolev regularization; outside it, bias dominates and is amplified by the very same regularization.

A further connection appears with spectral bias in neural network learning. Neural networks are known to learn low-frequency components of the target function first. Sobolev training amplifies this effect by explicitly penalizing high-frequency content through the gradient-matching term in the loss. The result is a low-pass filtering of the learned potential energy surface whose cutoff frequency is controlled by λ. Schütt *et al.* observed that equivariant architectures can mitigate certain biases, yet the spectral perspective reveals that any architecture remains subject to the smoothness-induced cutoff once the loss includes derivative information [22]. Consequently, rapid variations that might appear in unexplored regions of configuration space (for instance, during phase transitions in complex alloys) are systematically suppressed, limiting extrapolation range in a manner directly traceable to the Sobolev norm.

Finally, the framework relates to neural tangent kernel (NTK) analysis and PAC learning bounds. In the NTK regime, Sobolev training modifies the kernel to favor smoother solutions, effectively altering the reproducing kernel Hilbert space to one with bounded H^1 norm. This modification improves sample complexity inside the support (consistent with Czarnecki *et al.*'s empirical findings) but raises the PAC risk outside it, because the generalization bound now contains a term that grows with the distance-dependent factor arising from Sobolev embedding inequalities. E *et al.* and Han *et al.* provided the PDE-motivated foundation for these kernel modifications; the present work translates that foundation into explicit extrapolation bounds for interatomic potentials. Taken together, these relations demonstrate that the smoothness–extrapolation trade-off is not an artifact of any particular architecture but a structural consequence of embedding physical inductive biases into high-dimensional function approximation.

Implications for Model Design

Table 2 translates the manuscript's mechanistic account into an actionable design framework by linking each theoretical mechanism to its extrapolative failure mode, corresponding model-design response, and appropriate evaluation signal.

Table 2. Mechanism-to-design translation framework for managing extrapolation limits in Sobolev-trained potentials

Mechanism	Core theoretical action on the learned surface	Primary failure mode outside training support	Most directly affected manuscript construct	Design response implied by the theory	Evaluation signal that should be reported

Prior dominance	The model increasingly defaults to the smoothness prior as distance from data grows	Predictions revert toward generic low-curvature or near-zero-force behavior rather than true physical trends	Extrapolation horizon; lower-bound error growth	Use task-dependent smoothness, distance-aware regularization, or fallback physics baselines	Error versus distance-from-training-set curves; horizon estimates at fixed tolerance
Spectral bias amplification	High-frequency and sharp-curvature components are suppressed more strongly as λ increases	Rapid variations in unexplored regions are filtered out and therefore underrepresented	Prior strength vs flexible approximation	Use adaptive λ, richer architectures, or targeted data acquisition in regions expected to have rapid variation	Performance stratified by structural novelty, phase-transition regions, or descriptor-space sparsity
Gradient-matching interference	Local derivative constraints propagate strongly and shape the global geometry of the fitted surface	The model preserves local slopes while missing the correct nonlocal energetic trajectory	Force accuracy vs energy extrapolation	Decouple energy and force weighting by region, or reduce force dominance when extrapolation is central	Separate reporting of force MAE and out-of-domain energy error rather than aggregate loss alone
Loss-landscape geometry	Optimization increasingly favors the smoothest admissible solution consistent with the data	Convergence to stable but overly conservative functions with poor out-of-support flexibility	Data efficiency vs extrapolation range	Train ensembles across λ values, compare minima with different regularization strengths, and use uncertainty-triggered resampling	Ensemble disagreement, sensitivity to λ, and robustness across alternative smoothness settings
Constraint-induced function-space compression	The admissible hypothesis space narrows toward bounded-curvature functions	Physically relevant complex many-body structure may become inexpressible	Strong prior vs representational flexibility	Combine symmetry-aware models with selectively relaxed smoothness in sparse regions	Accuracy reported by materials class, interaction complexity, or composition shift magnitude
Distance-amplified bias accumulation	Bias becomes the dominant error component once the model exits the data neighborhood	Even small prior mismatch compounds rapidly with increasing distance d	Exponential lower bound on extrapolation error	Use active learning thresholds based on distance-aware error criteria	Acquisition triggers and error calibration against explicit distance thresholds

Design implications under theoretical limits of extrapolation

The theoretical constraints on extrapolation translate directly into design considerations for next-generation machine-learning potentials, particularly in regimes where out-of-distribution prediction is unavoidable. The role of smoothness, often treated as a purely technical hyperparameter, instead emerges as a task-dependent control variable that mediates

the balance between local fidelity and global reach. In settings where predictions remain close to the training manifold—such as refinement of known crystal structures or phonon calculations within a restricted compositional space—strong Sobolev weighting enhances both data efficiency and force accuracy. Under these conditions, the smoothness prior aligns with the local regularity of the target function, allowing the model to exploit dense gradient information without incurring significant extrapolation penalties. The situation shifts when the modeling objective involves exploration, as in the screening of high-entropy alloys or defect configurations that lie far from known data. Here, the same smoothness assumptions become restrictive, and weaker regularization—or even objectives that omit gradient constraints—can preserve the flexibility required to extend beyond the training domain. This reframes hyperparameter selection as a theoretically informed decision: the choice of λ must reflect not only interpolation performance but also the expected distance to unseen configurations relative to the model's extrapolation horizon.

This shift in perspective introduces the need for architectures capable of adapting their inductive bias dynamically. Fixing the smoothness parameter globally imposes a uniform constraint across regions that differ fundamentally in data support, leading to either over-regularization in sparse regions or under-regularization in dense ones. A more responsive approach allows the strength of the Sobolev term to vary as a function of spatial context or model uncertainty. In practice, this could be realized through uncertainty-weighted objectives that attenuate the influence of gradient matching as the model moves away from training data, thereby relaxing the prior precisely where it becomes most constraining. The relevance of such mechanisms is underscored by the exponential dependence of extrapolation error on distance: once predictions approach the boundary of the reliable region, reducing the effective smoothness constraint can extend the usable range without compromising stability near the data. Work by Batzner *et al.* [8] demonstrates that equivariant graph architectures already mitigate certain symmetry-related biases, suggesting that coupling these architectures with adaptive smoothness could provide a principled path toward balancing locality and flexibility.

Uncertainty itself becomes a central design signal in this context. Ensemble-based approaches offer a practical means of exposing the implicit bias introduced by smoothness regularization. By training multiple models under varying smoothness regimes, one obtains a distribution of predictions that reflects the sensitivity of the learned function to the imposed prior. Regions in which ensemble members converge toward similar low-curvature solutions indicate strong prior dominance and, therefore, elevated extrapolation risk. Conversely, divergence across the ensemble highlights areas where the model's behavior is underdetermined by the data, marking natural targets for further sampling. This interpretation connects directly to the earlier analysis of prior-driven behavior, recasting ensemble variance as a measurable proxy for the distance-dependent bias inherent in Sobolev-trained models.

A complementary strategy involves hybridizing machine-learned potentials with physics-based baselines [23-25]. Embedding a learned correction within a classical potential framework allows the model to inherit stable extrapolation behavior from the baseline while leveraging data-driven accuracy where observations are available. The theoretical bounds on extrapolation provide a natural criterion for managing this interaction. As the learned component approaches its extrapolation horizon, control can shift toward the baseline model, whose functional form ensures bounded behavior even in regions far removed from the training set. This integration does not eliminate the limitations of either approach but instead exploits their complementary strengths, aligning model behavior with the underlying structure of the problem.

These considerations also reshape how active learning strategies are formulated [26, 27]. Rather than relying on heuristic measures of uncertainty or diversity, the exponential growth of extrapolation error provides a principled basis for selecting new training points. Configurations that exceed a threshold distance—determined by the acceptable error level and the growth rate implied by the smoothness constraint—can be prioritized for acquisition. In this way, the sample efficiency associated with Sobolev training, as demonstrated by Czarnecki *et al.*, is complemented by a theoretically grounded policy for extending coverage into previously unsupported regions, reducing the risk of encountering uncontrolled extrapolation.

The implications extend to reporting practices as well. Evaluating model performance solely through in-domain metrics such as MAE or RMSE obscures the fundamental limitations identified here [28]. A more informative standard requires

explicit characterization of error as a function of distance from the training set, ideally across multiple smoothness regimes. Such reporting makes visible the trade-off between interpolation accuracy and extrapolation capacity, allowing practitioners to assess the suitability of a given model for specific materials design tasks. In this sense, the smoothness–extrapolation relationship is no longer a hidden constraint but an explicit design dimension, one that can be tuned, measured, and aligned with the objectives of computational materials discovery.

Conclusion

This theoretical analysis has established that smoothness constraints—imposed explicitly through regularization or implicitly through Sobolev training—improve interpolation accuracy and physical plausibility in machine-learning interatomic potentials yet simultaneously impose fundamental limits on extrapolation capacity. By formalizing the problem in Sobolev spaces, deriving the single Sobolev loss that balances energies and forces, and proving exponential growth of extrapolation error with distance, the work reveals four core trade-offs: interpolation accuracy versus extrapolation distance, force accuracy versus energy extrapolation, strength of the smoothness prior versus representational flexibility, and data efficiency versus extrapolation range. Three mathematical bounds quantify the phenomenon: an exponential lower bound on error, a shrinking extrapolation horizon controlled by λ, and a Sobolev-norm inequality that isolates the distance-dependent term. Four mechanisms—prior dominance, spectral bias amplification, gradient-matching interference, and loss-landscape geometry—explain how these trade-offs arise and propagate.

The results place the smoothness–extrapolation tension within the broader landscape of approximation theory, bias–variance decomposition, spectral bias, and NTK analysis, showing that it is not architecture-specific but intrinsic to any derivative-informed inductive bias. For materials applications involving alloys or complex many-body systems, the analysis implies that designers must treat smoothness as a tunable, task-dependent hyperparameter rather than a universal default. Adaptive regularization, ensembles, hybrid physics–ML models, and bound-guided active learning emerge as concrete strategies to mitigate the identified limits.

Ultimately, the central theoretical result is that there is no free lunch in Sobolev-trained potentials: every gain in interpolation fidelity and force accuracy is purchased at the price of reduced extrapolation range. Future work should therefore adopt extrapolation-aware evaluation protocols and develop architectures that can dynamically modulate smoothness. Only by confronting these trade-offs explicitly can machine-learning potentials fulfill their promise as truly general-purpose tools across the entire configuration space of materials science.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 22 Dec 2021 Revised: 04 Mar 2022 Accepted: 20 May 2022
Published online: 18 July 2022

References

Fedik N, Zubatyuk R, Kulichenko M, Lubbers N, Smith JS, Nebgen B, et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. Nat Rev Chem. 2022;6(9):653-72.
https://doi.org/10.1038/s41570-022-00416-3.

Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. Adv Mater. 2019;31(46):e1902765.
https://doi.org/10.1002/adma.201902765.

Smith JS, Isayev O, Roitberg AE. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci. 2017;8(4):3192-203.
https://doi.org/10.1039/c6sc05720a.

Han J, Jentzen A, E W. Solving high-dimensional partial differential equations using deep learning. Proc Natl Acad Sci U S A. 2018;115(34):8505-10.
https://doi.org/10.1073/pnas.1718942115.

Yu W, Ji C, Wan X, Zhang Z, Robertson J, Liu S, et al. Machine-learning-based interatomic potentials for advanced manufacturing. Int J Mech Syst Dyn. 2021;1(2):159-72.
https://doi.org/10.1002/msd2.12021.

Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3(12):e1701816.
https://doi.org/10.1126/sciadv.1701816.

Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. Phys Rev Lett. 2018;120(14):143001.
https://doi.org/10.1103/PhysRevLett.120.143001.

Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun. 2022;13(1):2453.
https://doi.org/10.1038/s41467-022-29939-5.

Behler J. Four generations of high-dimensional neural network potentials. Chem Rev. 2021;121(16):10037-72.
https://doi.org/10.1021/acs.chemrev.0c00868.

Mueller T, Hernandez A, Wang C. Machine learning for interatomic potential models. J Chem Phys. 2020;152(5):050902.
https://doi.org/10.1063/1.5126336.

Czarnecki WM, Osindero S, Jaderberg M, Swirszcz G, Pascanu R. Sobolev training for neural networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4-9; Long Beach, CA. Red Hook (NY): Curran Associates; 2017. p. 4281-90.
https://doi.org/10.5555/3294996.3295182.

Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B. 2017;95(9):094203.
https://doi.org/10.1103/PhysRevB.95.094203.

E W, Han J, Jentzen A. Algorithms for solving high dimensional PDEs: From nonlinear Monte Carlo to machine learning. Nonlinearity. 2021;35(1):278-310.
https://doi.org/10.1088/1361-6544/ac337f.

Chmiela S, Sauceda HE, Müller KR, Tkatchenko A. Towards exact molecular dynamics simulations with machine-learned force fields. Nat Commun. 2018;9(1):3887.
https://doi.org/10.1038/s41467-018-06169-2.

Cooper AM, Kästner J, Urban A, Artrith N. Efficient training of ANN potentials by including atomic forces via Taylor expansion and application to water and a transition-metal oxide. NPJ Comput Mater. 2020;6(1):54.
https://doi.org/10.1038/s41524-020-0323-8.

Christensen AS, Von Lilienfeld OA. On the role of gradients for machine learning of molecular energies and forces. Mach Learn Sci Technol. 2020;1(4):045018.
https://doi.org/10.1088/2632-2153/abba6f.

Zeni C, Rossi K, Glielmo A, Baletto F. On machine learning force fields for metallic nanoparticles. Adv Phys X. 2019;4(1):1654919.
https://doi.org/10.1080/23746149.2019.1654919.

Zaverkin V, Holzmüller D, Steinwart I, Kästner J. Fast and sample-efficient interatomic neural network potentials for molecules and materials based on Gaussian moments. J Chem Theory Comput. 2021;17(10):6658-70.
https://doi.org/10.1021/acs.jctc.1c00527.

Zeng C, Chen X, Peterson AA. A nearsighted force-training approach to systematically generate training data for the machine learning of large atomic structures. J Chem Phys. 2022;156(6):064104.
https://doi.org/10.1063/5.0079314.

Miksch AM, Morawietz T, Kästner J, Urban A, Artrith N. Strategies for the construction of machine-learning potentials for accurate and efficient atomic-scale simulations. Mach Learn Sci Technol. 2021;2(3):031001.
https://doi.org/10.1088/2632-2153/abfd96.

Shimamura K, Takeshita Y, Fukushima S, Koura A, Shimojo F. Computational and training requirements for interatomic potential based on artificial neural network for estimating low thermal conductivity of silver chalcogenides. J Chem Phys. 2020;153(23):234301.
https://doi.org/10.1063/5.0027058.

Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet: A deep learning architecture for molecules and materials. J Chem Phys. 2018;148(24):241722.
https://doi.org/10.1063/1.5019779.

Pun GP, Yamakov V, Hickman J, Glaessgen EH, Mishin Y. Development of a general-purpose machine-learning interatomic potential for aluminum by the physically informed neural network method. Phys Rev Mater. 2020;4(11):113807.
https://doi.org/10.1103/PhysRevMaterials.4.113807.

Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. Phys Rev X. 2018;8(4):041048.
https://doi.org/10.1103/PhysRevX.8.041048.

Ko TW, Finkler JA, Goedecker S, Behler J. General-purpose machine learning potentials capturing nonlocal charge transfer. Acc Chem Res. 2021;54(4):808-17.
https://doi.org/10.1021/acs.accounts.0c00689.

Kang PL, Shang C, Liu ZP. Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration. Acc Chem Res. 2020;53(10):2119-29.
https://doi.org/10.1021/acs.accounts.0c00472.

Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mater Sci. 2017;140:171-80.
https://doi.org/10.1016/j.commatsci.2017.08.031.

Zuo Y, Chen C, Li X, Deng Z, Chen Y, Behler J, et al. Performance and cost assessment of machine learning interatomic potentials. J Phys Chem A. 2020;124(4):731-45.
https://doi.org/10.1021/acs.jpca.9b08723.