Physics-Informed Gaussian Processes for Hardness Prediction in Refractory High Entropy Alloys

Anonymous Authors[1]

Abstract

Refractory High Entropy Alloys have emerged as a compelling alternative to traditional metals, offering resilience in extreme environments that conventional alloys cannot withstand. With at least four, and often up to six principal elements, the resulting compositional space spans trillions of candidates, far too large to navigate through experiment alone given the cost and time involved. Finding novel alloys which exhibit the required mechanical properties such as hardness necessitates highly predictive models to guide any tractable search. While existing surrogate models to cheaply impute hardness leverage elemental features, they are often architected in a way where the prior physical inductive bias is not explicitly tunable. More importantly, they rely solely on composition-derived descriptors such as valence electron concentration and mixing entropy, missing features that may only be acquired through experimentation such as local inhomogeneity found in non-equilibrium microstructures present in all real materials. In this paper, we demonstrate that integrating a scalable and physically informed differentiable Gaussian process improves predictive performance over existing black-box models. We achieve this by embedding the Maresca-Curtin solid-solution-strengthening model as a prior mean and replacing its fixed atomic volumes with element-resolved effective volumes learned from data. The resulting model achieves a mean absolute error of 35.6 HV on the public Borg benchmark, outperforming the strongest black-box baselines by 1.3x. Furthermore, we convert X-ray diffraction (XRD) spectra and microscopy/EDS-derived elemental segregation partition coefficients into microstructural descriptors, showing that these experimentally derived features further improve performance to 7.5 HV MAE on the as-cast Experimental dataset because dendritic segregation cannot be captured by bulk composition alone. Our central contribution is a non-linear correction scheme that reveals systematic, non-affine residuals in elemental volumes which no standard reparameterization could reproduce. We demonstrate that this scheme successfully recovers the physical regimes identifying whether an alloy is screw or edge dominated. Together, these results establish a multiscale, physics-informed surrogate for hardness prediction in realistic experimental settings where data are costly, heterogeneous, and experimentally constrained.

1. Introduction

High-entropy alloys (HEAs) are multi-principal-element systems with near-equiatomic compositions whose combination of strength, ductility, and thermal stability has driven sustained interest (Miracle and Senkov, 2017; Yang et al., 2025). Refractory HEAs (RHEAs) based on Group IV–VI transition metals are particularly attractive for high-temperature structural applications in aerospace and nuclear environments (Senkov et al., 2018), but the vast combinatorial design space makes exhaustive experimental exploration impractical and motivates predictive surrogate models that can prioritise candidate compositions before synthesis, particularly in experimental settings where data are generated through iterative, lab-driven workflows (Butler et al., 2018; Schmidt et al., 2019).

Existing surrogates fall into two camps with complementary weaknesses. Purely data-driven regressors trained on heuristic compositional descriptors such as atomic radius, valence electron concentration, and mixing entropy achieve reasonable accuracy within their training distribution (Rao et al., 2022; Gao et al., 2023; Vela et al., 2023), but they operate independently of the physical mechanisms governing deformation, which limits both interpretability and out-of-distribution generalization. More broadly, physics-informed and theory-guided machine learning methods

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

seek to address this limitation by incorporating mechanistic structure into data-driven models (Karniadakis et al., 2021; Willard et al., 2022). Physics-based models grounded in solid-solution strengthening theory, most notably the Maresca–Curtin framework (Maresca and Curtin, 2020), are mechanistically interpretable but are limited by the rule-of-mixtures (ROM) approximation for atomic volumes and elastic constants. The ROM approximation becomes inaccurate in multi-principal-element environments where local chemical fluctuations dominate the misfit parameter that drives strengthening (Baruffi et al., 2022), so the dominant source of error in the analytical prior is concentrated in the elemental volumes that the prior consumes rather than in the strengthening equations themselves.

In this work we bridge the two camps. We embed the Maresca–Curtin model as a differentiable prior mean function of a Gaussian process (GP) and replace its fixed ROM-based atomic volumes with element-resolved effective volumes that are learned from data and re-aggregated through the original physics. This formulation corrects the dominant error in the analytical prior while preserving interpretability at the level of individual elements. Unlike prior work that applies corrections in output space, our method learns structured corrections in physically meaningful latent variables while preserving the governing equations. This formulation connects atomic-scale misfit physics, mesoscale microstructural heterogeneity, and macroscopic mechanical response within a single predictive model, aligning naturally with autonomous experimental workflows where data are generated and consumed iteratively (Kusne et al., 2020; Abolhasani and Kumacheva, 2023). This perspective leads to four contributions evaluated below: (i) the resulting physics-informed GP outperforms both black-box baselines and prior GP-based approaches on the public Borg benchmark (Borg et al., 2020; Gao et al., 2023; Vela et al., 2023); (ii) adding microstructural descriptors derived from XRD peak counts and energy dispersive spectroscopy (EDS) measured partition coefficients further improves prediction on a proprietary cast-RHEA dataset (the Experimental dataset), where dendritic segregation is not captured by bulk composition; (iii) the learned correction is small in magnitude (under 5% in volume) yet systematically nonlinear in composition, with per-element misfit-volume surfaces that are $R^2 > 0.999$ to a best-fit plane but have spatially structured, non-affine residuals that no ROM or affine reparameterization could reproduce; and (iv) the model's edge yield-strength prediction tracks measured hardness roughly three times more cleanly in alloys with Hume–Rothery misfit $\delta \geq 0.035$ than in alloys below the threshold, recovering the screw/edge regime identified by Baruffi et al. (2022) despite the model being trained only on hardness.

2. Method

We define a hybrid framework where the analytical Maresca–Curtin model serves as the GP prior mean and a learned correction modifies the misfit input to that model. This section details the prior, the GP, the learnable effective-volume formulation, the microstructural features used on the Experimental dataset, and training. A summary of the notation is provided in Appendix Table 3.

2.1. Maresca–Curtin Prior

For a composition with atomic fractions $\{c_i\}$ and experimental Body-Centered Cubic (BCC) lattice constants $\{a_i\}$, the equilibrium atomic volume and per-element misfit volume are

$$V_{\mathrm{eq}} = \sum_i c_i \frac{a_i^3}{2}, \quad \Delta V_i = \frac{a_i^3}{2} - V_{\mathrm{eq}}. \tag{1}$$

The reduced misfit parameter $\sigma = \sum_i c_i (\Delta V_i)^2$ feeds the Maresca–Curtin equations for the zero-temperature yield stress $\sigma_{y,0}$ and activation energy ΔE_b, which depend on the shear modulus G, Poisson's ratio ν, and Burgers vector b (Maresca and Curtin, 2020). We convert the thermally-softened yield stress $\sigma_y(T)$ to Vickers hardness through the Tabor relation:

$$\mathrm{HV} = \frac{3 \cdot \sigma_y(T)}{9.81}. \tag{2}$$

2.2. Physics-Informed Gaussian Process

We embed the analytical hardness prediction as the prior mean function $m(\mathbf{x})$ of a GP:

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}') + \delta_{ij}\sigma_n^2\big). \tag{3}$$

The kernel k is a Matérn-5/2 with Automatic Relevance Determination (ARD), letting the GP learn a non-parametric correction to the physics prior weighted by local data density.

2.3. Learnable Effective-Volume Misfit

The dominant source of error in the analytical prior is the ROM approximation that produces ΔV_i from elemental BCC lattice constants. We introduce a hierarchy of corrections, all bounded and centred on the ROM baseline.

Shared-σ scaling. The simplest correction applies a composition-dependent scalar to the aggregate ROM misfit:

$$\sigma_{\mathrm{adj}} = \sigma_{\mathrm{rom}} \cdot \exp\big(\beta \tanh g_\theta(\mathbf{c})\big), \tag{4}$$

where g_θ is a single-layer neural network and β bounds the correction magnitude.

Effective-volume reconstruction. For element-level interpretability, we instead learn an effective volume V_i^{eff} for each element:

$$V_i^{\text{eff}} = V_i^{\text{base}} \cdot \exp\big(\beta \tanh h_{\theta,i}(\mathbf{c})\big). \tag{5}$$

The aggregate misfit σ is then reconstructed by re-applying Eq. (1) to the learned V_i^{eff}, ensuring that data-driven adjustments stay physically consistent with the Maresca–Curtin aggregation.

2.4. Microstructural Features

The base feature set for both datasets is eight compositional descriptors (e.g., VEC, δ, $\Delta\chi$) standardized to zero mean and unit variance. This design follows the broader motivation of microstructure-aware materials modeling, where explicit microstructural descriptors are used to bridge characterization data and process–structure–property relationships (DeCost et al., 2017; Peng et al., 2025; Latypov et al., 2019). For the Experimental dataset only, we extend these with per-sample microstructural descriptors that we generate through a dedicated characterization pipeline. Microstructural characterization is performed at the per-sample level so that the model receives physically meaningful inputs rather than bulk compositional proxies. For every synthesized sample, an XRD spectrum is collected and passed through an automated peak-detection workflow to count the number of significant peaks. This count serves as a proxy for the number of minor intermetallic (IM) phases present, given that the major phase is consistently BCC. Rather than reducing XRD data to categorical labels such as BCC+IM, which collapse complex microstructural diversity into discrete classes, peak count provides a more descriptive representation that implicitly encodes phase symmetry where a single BCC phase yields few peaks, additional secondary phases such as disordered BCC or FCC contribute modestly, while highly ordered phases such as silicides and sigma phases yield significantly more peaks. To capture the true microstructural state of each as-cast sample more rigorously, we apply automated segmentation workflows to backscattered secondary electron (BSE) micrographs and EDS elemental maps acquired in a scanning electron microscopy (SEM). This pipeline yields a comprehensive set of per-phase descriptors, including the number of distinct phases, phase morphology quantified through feature size, and elemental partition coefficients k_i as a direct measure of chemical microsegregation, together with a derived core phase fraction f.

To our knowledge this represents one of the first systematic attempts to incorporate true per-sample microstructural descriptors, rather than bulk compositional proxies, as features in machine-learning prediction of mechanical performance for cast refractory alloys, complementing broader work on microstructure-aware materials informatics (Peng et al., 2025; Latypov et al., 2019). A practical advantage of this feature set is its natural extensibility: applying identical characterization to annealed variants of the same compositions would isolate the contribution of microstructural state from chemistry, since annealing normalizes the as-cast segregation while preserving bulk composition. The same descriptors are also not limited to post-hoc experimental measurement; quantities such as the number of equilibrium and non-equilibrium phases, solidification reaction types, and elemental partition coefficients are accessible a priori through established thermodynamic and kinetic simulation frameworks including Scheil solidification and diffusion-based models (Senkov et al., 2018), which opens the prospect of fully simulation-driven feature generation for alloy screening prior to synthesis.

2.5. Optimization

Training uses staged maximum marginal likelihood. In the **warm-up stage**, the physics-prior parameters are frozen and only the GP kernel hyperparameters are optimized. In the **joint stage**, the physics prefactors (α, M, f_L) and the effective-volume neural weights are unfrozen and trained jointly. The bounded-exponential parameterization in Eqs. (4), (5) keeps learned volumes within $e^{\pm\beta}$ of the experimental BCC values (typically $< 10\%$), preventing the correction head from overwhelming the physics baseline.

3. Datasets

Borg. The Borg dataset (Borg et al., 2020) is a literature-curated compilation of Vickers hardness measurements for multi-principal-element alloys. We retain entries with complete hardness, compositional features, and formulae over our supported element set ({Al, Si, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zr, Nb, Mo, Hf, Ta, W}). Missing test temperatures are taken to be room temperature (298.15 K). The elemental lattice constants and elastic stiffness constants used in the prior are listed in Appendix Table 4.

Experimental Dataset. While the Borg benchmark provides a standardized, composition-only evaluation, the Experimental dataset serves as a complementary setting that reflects realistic experimental conditions where additional microstructural information is available but cannot be inferred from composition alone. The Experimental dataset comprises Vickers micro-hardness measurements (Qness Q10A+) of cast refractory HEAs synthesized within an internally developed experimental workflow for rapid alloy exploration. Compositions are determined by SEM-EDS and XRF, XRD peak counts are recorded, and all measurements are at room temperature. Partition coefficients and core phase fractions are obtained from EDS analysis of the as-cast microstructure, providing the inputs for the mi-

Table 1. Five-fold cross-validation on the Borg dataset ($n = 99$). MAE and RMSE are in HV units. The physics-informed GP cuts MAE by $4.4\times$ relative to the analytical Curtin prior and beats the strongest black-box baseline by $1.3\times$. Adding the learnable effective-volume misfit gives the best overall fit.

Model	MAE (HV)	RMSE (HV)
Curtin prior (analytical)	162.5	180.4
Gao RF (Gao et al., 2023)	47.8	68.7
Gao SVR (Gao et al., 2023)	58.6	85.1
PI-GP (analytic mean)	37.5	53.3
PI-GP + shared-σ	37.6	53.1
PI-GP + effective volume	**35.6**	**50.2**

crostructural features in Section 2.4.

4. Experiments

We evaluate four claims in turn: (i) the physics-informed GP improves over baselines on the public Borg benchmark; (ii) adding microstructural features further improves prediction on as-cast Experimental alloys; (iii) the learned misfit-volume correction is small but systematically nonlinear; (iv) the model's edge yield-strength prediction internalizes the screw/edge regime identified by Baruffi et al. (2022). All results are evaluated using 5-fold cross-validation with no overlap of samples or duplicate compositions between training and validation folds.

4.1. Physics-Informed GP on the Borg Benchmark

We evaluate the physics-informed GP (PI-GP) on the Borg dataset against three families of baselines: the analytical Curtin prior (Maresca and Curtin, 2020), the black-box Random Forest and SVR regressors with the engineered descriptors of Gao et al. (2023), and the residual-GP approach of Vela et al. (2023). Table 1 reports five-fold cross-validated MAE and RMSE, and Figure 1 shows the corresponding parity behavior. The bare Curtin prior achieves an MAE of 162.5 HV, which is consistent with the literature observation that ROM-based misfit underpredicts strengthening in non-dilute alloys (Baruffi et al., 2022). Replacing the analytical prediction with the analytic-mean PI-GP cuts MAE to 37.5 HV, a $4.3\times$ improvement that already exceeds the strongest black-box baseline (Gao RF at 47.8 HV) by $1.3\times$ and reflects the value of an informative GP prior in the small-n regime (Williams and Rasmussen, 2006; Gardner et al., 2018). Within the physics-informed family, the shared-σ scaling of Eq. 4 is essentially neutral relative to the analytic-mean PI-GP, indicating that a single alloy-level rescaling of the misfit cannot displace the residual error in the prior. The element-resolved effective-volume correction of Eq. 5 gives a further consistent reduction in MAE and RMSE, reaching 35.6 HV and 50.2 HV respectively, which we read

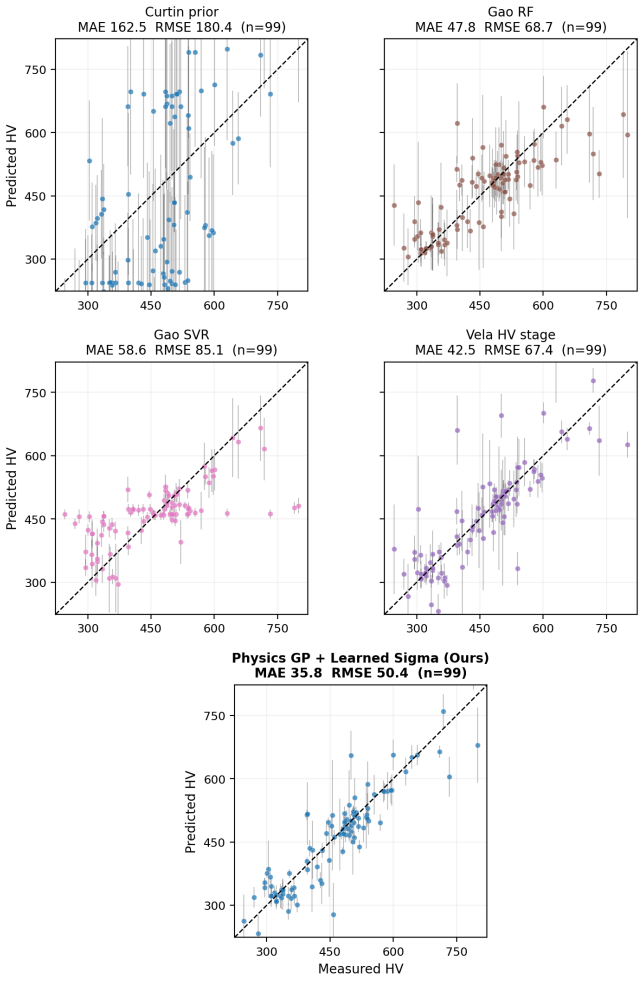

Figure 1. Parity plots on the Borg dataset for the physics-based model family. Vertical whiskers denote the GP predictive standard deviation. Moving from the bare Curtin prior to the physics-informed GP improves the fit, and adding the learnable effective-volume σ correction further tightens parity while preserving the mechanistic structure of the prior.

as evidence that the residual error in the analytical prior is element-specific and is not absorbed by a global misfit miscalibration.

4.2. Microstructural Features on the Experimental Dataset

We now evaluate whether the gains observed on the public Borg benchmark extend to a more realistic experimental setting where microstructural descriptors are available.

The Experimental dataset is substantially larger than Borg and consists of as-cast samples for which we have both XRD peak counts and EDS-derived partition coefficients, allowing a sharper question than on Borg: does experimental information that bulk composition cannot express provide

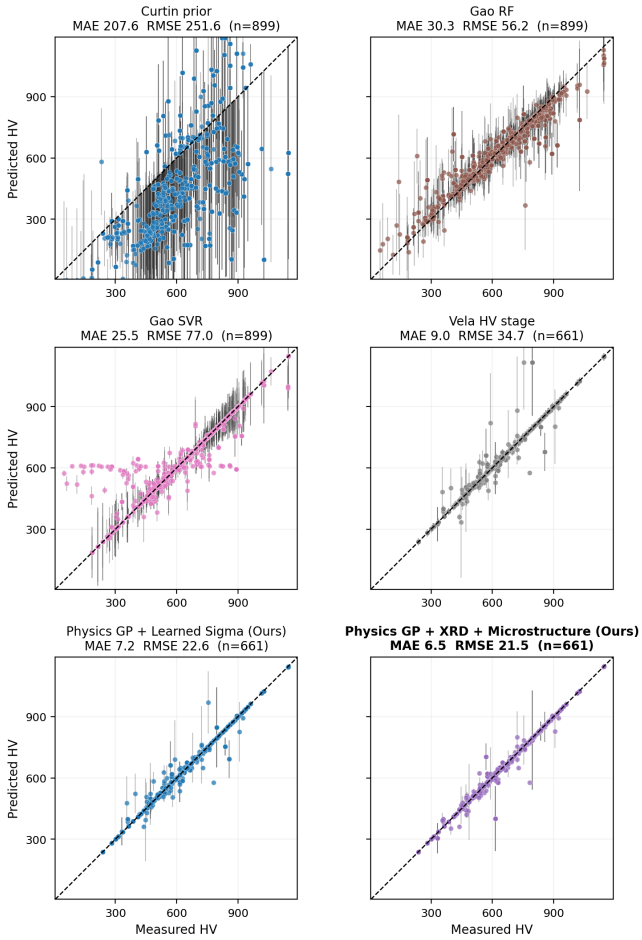

Figure 2. Five-fold cross-validation error summary on the Experimental dataset. Bars give mean MAE/RMSE across folds; whiskers give fold-to-fold standard deviation. Augmenting the physics-informed GP with experimental descriptors (XRD peak count, microstructural features) yields further improvement on top of the composition-only effective-volume model.

additional lift on top of the composition-only PI-GP? Figure 2 reports five-fold cross-validation across four feature ablations and confirms that it does. The composition-only PI-GP cuts MAE by roughly an order of magnitude relative to the bare Curtin prior (from 216.3 HV to 9.1 HV) and beats the strongest black-box baseline (Gao SVR at 11.8 HV) by roughly $1.3\times$, which shows that the composition-only PI-GP already captures the dominant compositional signal before any experimental features are introduced. Adding the single XRD-peak-count descriptor reduces RMSE by approximately 20%, since this descriptor is a coarse proxy for phase multiplicity (single-phase BCC versus multi-phase) and helps the kernel separate alloys that share bulk composition but diverge in their solidification path (Senkov et al., 2018). Adding the 17 microstructure features (core and shell compositional descriptors reconstructed from partition coefficients k_i, together with the core phase fraction f) reduces MAE further, and combining XRD peak count with the microstructural descriptors gives the best fit overall (7.5 HV MAE, 28.3 HV RMSE). The two feature families are complementary because they encode different aspects of as-cast morphology. XRD peak count tells the kernel how many phases are present; the partition coefficients k_i tell it where the chemistry sits within each phase, since the local chemistry of dendritic core and inter-dendritic shell can differ markedly from the bulk SEM-EDS composition that the physics prior consumes. The same descriptors yielded no measurable improvement on Borg because Borg aggregates literature samples without consistent microstructural characterization, and the lift therefore tracks the availability of the underlying experiment rather than the modelling choice in the kernel. A complementary observation is that the black-box baselines are weaker on Experimental than on Borg in relative terms: once the Curtin prior absorbs the bulk compositional trend, the kernel only has small residuals left to model, and tree-based regressors and SVR cannot reach this fine-grained regime. The same observation underpins the residual-GP design of Vela et al. (2023), which is one of the stronger prior baselines for hardness regression on RHEAs.

4.3. Nonlinearity of the Learned Correction

A natural next question is whether the learned correction is doing something physically meaningful or simply retuning the elemental BCC volumes that enter the prior. The bounded parameterization in Eq. (5) keeps the per-element ratio $V_i^{\mathrm{eff}}/V_i^{\mathrm{base}}$ within $e^{\pm\beta}$ of unity, and empirically the median deviation is only a few percent, so a correction this small could in principle be reproduced by a simpler reparameterization that retunes the elemental BCC lattice constants a_i themselves (a global affine reparameterization in V_i). Figure 3 tests this hypothesis on four representative ternary subsystems by plotting the ROM ΔV surface, the learned ΔV surface, their difference, and the residual of the learned surface to its best-fit plane. The learned ΔV surfaces are highly planar with $R^2 > 0.999$ to their best-fit planes in every ternary, so the bulk shape of the correction is consistent with a global retuning of the underlying lattice constants and is therefore not by itself a meaningful contribution beyond what Maresca and Curtin (2020) could have captured by choosing different a_i values.

The residual to the best-fit plane carries a normalized RMSE of order 10^{-3}. Although small, this residual is not random: it exhibits coherent sign across contiguous regions of each ternary, indicating a nonlinear-in-composition contribution to V_i^{eff} that no global reparameterization of the elemental lattice constants a_i could reproduce. Section 4.4 shows that this structure propagates to the dual-mechanism head, where it underlies the model's separation of edge- and screw-dominated alloys.

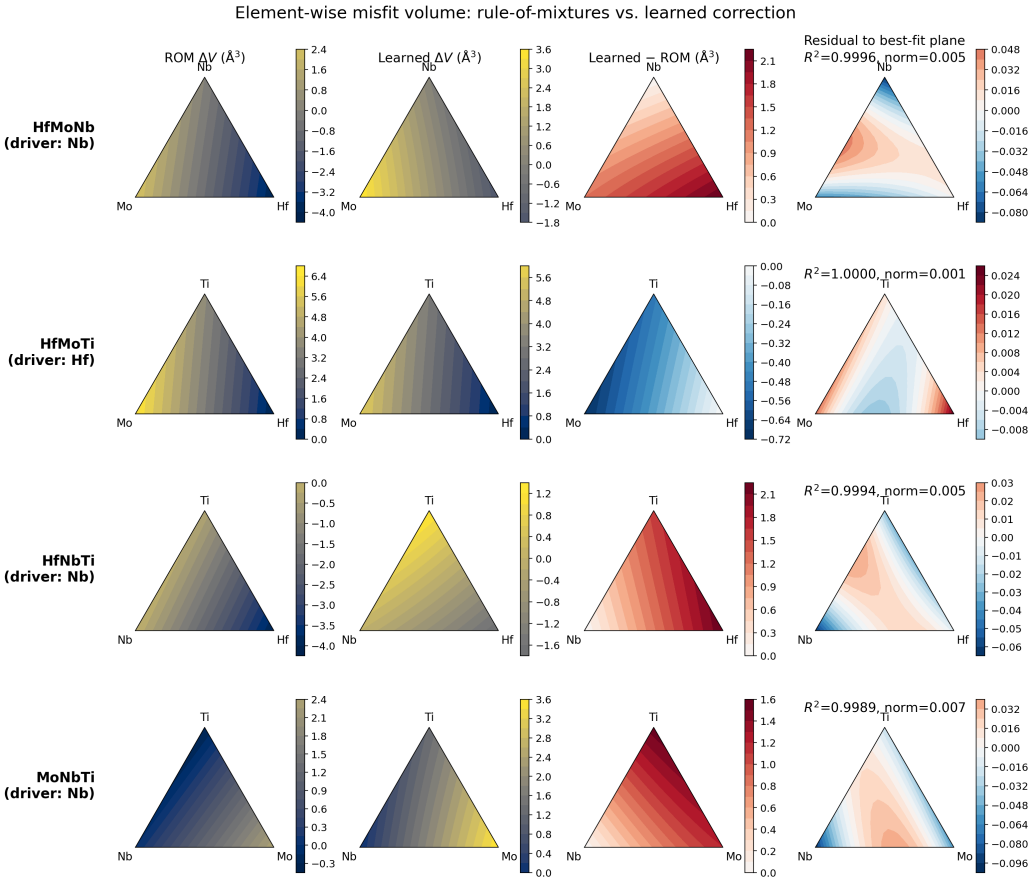

Figure 3. Element-wise misfit-volume surfaces for the best Experimental composition-only model. Each row picks the most nonlinear element for that ternary subsystem and shows, from left to right, the rule-of-mixtures ΔV, the learned ΔV, their difference, and the residual of the learned surface to its best-fit plane (with R^2 and a normalized RMSE). Plane R^2 exceeds 0.999 in all four ternaries, but the residuals are spatially structured rather than random, which demonstrates that the learned correction is small in magnitude yet genuinely nonlinear in composition and goes beyond what any rule-of-mixtures or affine reparameterization could reproduce.

Table 2. Linear fit $H = m\sigma_{y,\mathrm{th}} + c$ between measured Vickers hardness (GPa) and the model's predicted edge yield strength on the Experimental dataset, with one point per alloy (mean H across indents). The misfit threshold $\delta = 0.035$ separating edge- and screw-dominated alloys is from Baruffi et al. (2022). The edge-dominated subset yields a fit roughly three times tighter in R^2 than the screw-dominated subset.

Subset	n	slope	intercept (GPa)	R^2
All Exp. alloys	242	0.93	3.09	0.40
$\delta \geq 0.035$ (edge)	153	0.83	3.70	0.34
$\delta < 0.035$ (screw)	89	0.58	3.67	0.11

4.4. The Model Internalizes the Screw/Edge Regime

Baruffi et al. (2022) report that BCC HEAs above $\delta = 0.035$ are edge-dominated and that on those alloys the Maresca–Curtin edge yield strength correlates with measured hardness, while alloys below the threshold are screw-dominated and show no such correlation. On their 63 wrought RHEAs, applying the screen lifts the linear H-vs-$\sigma_{y,\mathrm{th}}$ fit from $R^2 = 0.47$ across the full set to $R^2 = 0.71$ on the edge-dominated subset. We can run the same analysis on the Experimental dataset because our dual-mechanism head produces a room-temperature edge yield-strength prediction even though the model is trained only on hardness. Table 2 reports the fit on each δ bucket with one point per alloy ($n = 153$ for the edge-dominated subset and $n = 89$ for the screw-dominated one).

The qualitative split observed on wrought RHEAs is preserved on the as-cast Experimental alloys: $R^2 = 0.34$ on the edge-dominated bucket against $R^2 = 0.11$ on the screw-dominated bucket. This is consistent with the underlying physics, since misfit volumes are the dominant driver of strengthening only in the edge-controlled regime (Maresca and Curtin, 2020), while screw-controlled strengthening involves kink-pair mechanisms whose parameters are not exposed by the edge prediction. Absolute R^2 values are lower than those reported by Baruffi et al. (2022) because

the Experimental samples are as-cast: dendritic segregation introduces scatter into the H-vs-$\sigma_{y,\mathrm{th}}$ relationship, the same scatter that motivated the microstructural features in Section 4.2. The methodological conclusion is that the model recovers the screw/edge separation from hardness data alone, with no supervision on yield strength or mechanism labels. For alloy design this is the discrimination one needs: the same surrogate that predicts hardness within a few HV also indicates when the edge theory of Maresca and Curtin (2020) is the appropriate analytical tool.

5. Related Work

Gao et al. (2023) apply Random Forest and SVR with engineered compositional descriptors on the Borg dataset and provide the black-box baselines used in Section 4.1. Vela et al. (2023) take a Bayesian approach to yield-strength prediction in refractory BCC HEAs via a two-stage hierarchical GP that first imputes hardness and then predicts yield strength using the Maresca–Curtin prior with a residual sklearn GP; their grouped two-fold cross-validation accounts for non-independence within alloy series. We extend this line by (i) implementing the Maresca–Curtin model as a fully differentiable prior mean inside the GP rather than as a fixed offset, and (ii) learning structured corrections to the misfit-volume input rather than to the model output.

6. Discussion and Conclusion

Embedding solid-solution strengthening physics into the GP mean function provides an inductive bias that improves prediction accuracy and generalization while preserving element-level interpretability. The emergence of structured, non-affine corrections in effective atomic volumes suggests that local chemical environments induce systematic deviations from rule-of-mixtures assumptions, providing a data-driven refinement to classical solid-solution strengthening theory. The hierarchy of misfit corrections trades flexibility for regularization: the analytic baseline is the most constrained, the shared-σ scaling adds a single alloy-level degree of freedom, and the effective-volume formulation introduces N_{el} element-resolved corrections. This suggests that effective atomic volumes in multi-principal-element alloys cannot be fully described by composition-independent parameters, but instead exhibit systematic dependence on local chemical environments.

The effective-volume model achieves the best overall performance on Borg (35.6 HV MAE) and on Experimental when combined with the microstructural feature set (7.5 HV MAE), identifying the ROM lattice constants as a primary bottleneck of the Curtin prior (Maresca and Curtin, 2020). The bounded-exponential parameterization in Eq. 5 is essential to this result: without bounding, the correction head overfits training noise and degrades out-of-fold predictions, and we expect this regularization to be similarly important whenever a learned correction is layered onto a physically motivated prior. Microstructural descriptors give a second axis of improvement on Experimental that is not available on Borg because Borg lacks consistent microstructural characterization, and this dependence on the availability of the underlying experiment, rather than on modelling choices, is itself useful when planning closed-loop discovery campaigns where each measurement is expensive. The screw/edge analysis (Table 2) further suggests that the learned representation internalizes physically meaningful regime structure even without direct supervision on yield strength, mirroring the screening criterion of Baruffi et al. (2022). Two natural extensions follow: direct supervision on yield-strength data to sharpen the dual-mechanism head where such data are available (Vela et al., 2023), and using the model as an inexpensive surrogate within active learning over the RHEA composition space (Lookman et al., 2019), with simulation-driven microstructural features (Section 2.4) extending the surrogate beyond samples that have been physically characterized. In this setting, the model serves as a natural predictive component within an autonomous or semi-autonomous experimental loop, where candidate compositions can be proposed, synthesized, and evaluated iteratively (Kusne et al., 2020).

7. Limitations

The Experimental dataset used in this work is proprietary and cannot be released publicly due to experimental and commercial constraints. To ensure reproducibility and transparency, we (i) evaluate all models on the fully public Borg benchmark, (ii) report cross-validation protocols in detail, and (iii) isolate the effect of each feature group through ablation studies. We additionally report physically interpretable intermediate quantities (e.g., effective volumes, misfit parameters) that can be independently validated against known theory. We will release model code and trained weights upon publication.

One limitation of this work is the heterogeneity and reproducibility of experimental data across sources. The Experimental dataset includes richer microstructural features, while the Borg dataset (Borg et al., 2020) is primarily compositional, making it challenging to combine the two without introducing confounding effects from differences in measurement protocols and instrumentation. This raises the possibility that some observed performance gains may be influenced by dataset-specific biases. Additionally, the relatively limited size of the Experimental dataset makes it unclear whether the proposed approach will scale or generalize to substantially larger datasets or different data distributions. Finally, variation in experimental measurement techniques

for hardness and microstructure introduces further uncertainty, and it remains an open question which modalities are most informative for predictive modeling.

References

Abolhasani, M., Kumacheva, E., 2023. The rise of self-driving labs in chemical and materials sciences. Nature Synthesis 2, 483–492.

Baruffi, C., Maresca, F., Curtin, W., 2022. Screw vs. edge dislocation strengthening in body-centered-cubic high entropy alloys and implications for guided alloy design. Mrs Communications 12, 1111–1118.

Borg, C.K., Frey, C., Moh, J., Pollock, T.M., Gorsse, S., Miracle, D.B., Senkov, O.N., Meredig, B., Saal, J.E., 2020. Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. Scientific Data 7, 430.

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Machine learning for molecular and materials science. Nature 559, 547–555.

DeCost, B.L., Francis, T., Holm, E.A., 2017. Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures. Acta Materialia 133, 30–40.

Gao, Z., Zhao, F., Gao, S., Xia, T., 2023. Machine learning prediction of hardness in solid solution high entropy alloys. Materials Today Communications 37, 107102.

Gardner, J., Pleiss, G., Weinberger, K.Q., Bindel, D., Wilson, A.G., 2018. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. Advances in neural information processing systems 31.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 3, 422–440.

Kusne, A.G., Yu, H., Wu, C., Zhang, H., Hattrick-Simpers, J., DeCost, B., Sarker, S., Oses, C., Toher, C., Curtarolo, S., et al., 2020. On-the-fly closed-loop materials discovery via bayesian active learning. Nature communications 11, 5966.

Latypov, M.I., Khan, A., Lang, C.A., Kvilekval, K., Polonsky, A.T., Echlin, M.P., Beyerlein, I.J., Manjunath, B., Pollock, T.M., 2019. Bisque for 3d materials science in the cloud: microstructure–property linkages. Integrating Materials and Manufacturing Innovation 8, 52–65.

Lookman, T., Balachandran, P.V., Xue, D., Yuan, R., 2019. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Computational Materials 5, 21.

Maresca, F., Curtin, W.A., 2020. Mechanistic origin of high strength in refractory bcc high entropy alloys up to 1900k. Acta Materialia 182, 235–249.

Miracle, D.B., Senkov, O.N., 2017. A critical review of high entropy alloys and related concepts. Acta materialia 122, 448–511.

Peng, X.L., Fathidoost, M., Lin, B., Yang, Y., Xu, B.X., 2025. What can machine learning help with microstructure-informed materials modeling and design? x. peng et al. MRS Bulletin 50, 61–79.

Rao, Z., Tung, P.Y., Xie, R., Wei, Y., Zhang, H., Ferrari, A., Klaver, T., Körmann, F., Sukumar, P.T., Kwiatkowski da Silva, A., et al., 2022. Machine learning–enabled high-entropy alloy discovery. Science 378, 78–85.

Schmidt, J., Marques, M.R., Botti, S., Marques, M.A., 2019. Recent advances and applications of machine learning in solid-state materials science. npj computational materials 5, 83.

Senkov, O.N., Miracle, D.B., Chaput, K.J., Couzinie, J.P., 2018. Development and exploration of refractory high entropy alloys—a review. Journal of materials research 33, 3092–3128.

Vela, B., Khatamsaz, D., Acemi, C., Karaman, I., Arróyave, R., 2023. Data-augmented modeling for yield strength of refractory high entropy alloys: A bayesian approach. Acta Materialia 261, 119351.

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys 55, 1–37.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian processes for machine learning. volume 2. MIT press Cambridge, MA.

Yang, Y.F., Hu, F., Xia, T., Li, R.H., Bai, J.Y., Zhu, J.Q., Xu, J.Y., Zhang, G.F., 2025. High entropy alloys: A review of preparation techniques, properties and industry applications. Journal of Alloys and Compounds 1010, 177691.

Table 3. Principal symbols and abbreviations.

Symbol	Description
c_i	Atomic fraction of element i
a_i	Experimental BCC lattice constant of element i
V_{eq}	Equilibrium atomic volume (ROM)
ΔV_i	Misfit volume of element i
σ	Reduced misfit parameter
C_{11}, C_{12}, C_{44}	Elastic stiffness constants
K	Bulk modulus
G	Shear modulus
ν	Poisson's ratio
b	Burgers vector magnitude
α	Numerical prefactor (0.04)
f_L	Line-tension parameter ($1/12$)
M	Taylor factor (3)
ΔE_b	Activation energy for dislocation glide
k_B	Boltzmann constant
c, q	Thermal softening exponents
V_i^{eff}	Learnable effective volume of element i
β	Log-bound on volume/sigma corrections
g_θ, h_θ	Neural network correction functions
k_i	Partition coefficient of element i
f	Core phase fraction
RHEA	Refractory High-Entropy Alloy
HV	Vickers Hardness
GP	Gaussian Process
ROM	Rule of Mixtures
ARD	Automatic Relevance Determination
RF	Random Forest
SVR	Support Vector Regression
VEC	Valence Electron Concentration

Table 4. Experimental BCC lattice constants and elastic stiffness constants used in the Curtin prior. Lattice constants are in Å; elastic constants are in GPa.

Element	a (Å)	C_{11}	C_{12}	C_{44}
Cr	2.884	350	68	101
Fe	2.867	226	140	116
Mo	3.147	463	169	109
W	3.165	522	204	161
V	3.030	227	116	47
Nb	3.300	246	134	29
Ta	3.301	267	161	87
Al	3.240	107	61	28
Co	2.820	307	165	75
Ni	2.880	246	147	124
Cu	2.890	168	121	75
Ti	3.320	162	92	47
Zr	3.570	144	72	33
Hf	3.530	176	77	51
Mn	2.990	246	138	110
Si	2.830	167	65	80
Re	3.110	591	361	162