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

Delia McGrath [1] **Michael J. Pavel** [1] **Rohan Yuri Sanspeur** [1]

Abstract

Refractory High Entropy Alloys (RHEAs) offer considerable structural promise in extreme environments, but their vast compositional design space and high synthesis costs drive the need for accurate predictive models to accelerate materials screening. While hardness serves as a proxy for structural performance, existing predictive models lack physical inductive biases and fail to leverage experimentally accessible microstructural data. Here, we develop a physics-informed Gaussian process framework for RHEA hardness prediction that integrates mechanistic solid-solution strengthening theory with experimentally derived descriptors. By embedding the Maresca–Curtin model as the prior mean of a differentiable Gaussian process, we replace rigid rule-of-mixtures atomic volumes with element-resolved effective volumes learned directly from data. On the public Borg benchmark, this framework achieves a mean absolute error of 35.8 HV, outperforming the strongest black-box baselines by 1.3×. Furthermore, by incorporating experimental elemental partition coefficients from energy-dispersive spectroscopy together with phase-sensitive descriptors from X-ray diffraction, we reduce prediction error to 5.9 HV mean absolute error on our proprietary dataset, HADEX. Beyond predictive accuracy, the learned correction scheme reveals systematic nonlinear deviations from standard rule-of-mixtures assumptions, and its learned corrections align with the screw-to-edge strengthening transition identified in prior work. Together, these results establish a multiscale, physics-informed surrogate model ready for experimentally constrained alloy-design workflows.

[1]Radical AI. Correspondence to: Delia McGrath <dmcgrath@radical-ai.com>, Rohan Yuri Sanspeur <ysanspeur@radical-ai.com>.

Accepted at ICML 2026 AI for Science Workshop.

1. Introduction

High Entropy Alloys are a group of multi-principal element systems with near-equiatomic composition that are under sustained research due to their unique combination of thermophysical properties including strength, ductility, thermal stability, and corrosion resistance (Miracle and Senkov, 2017; Yang et al., 2025). Refractory HEAs (RHEAs), a subset of HEAs based on Group IV–VI transition metals with increased melting temperatures, are particularly attractive for extreme environment applications in aerospace and nuclear power generation (Senkov et al., 2018). Realizing this potential, however, requires identifying which compositions within the vast RHEA design space deliver optimal property combinations for a given application. As the alloys contain at least four, and often up to six principal elements, the potential compositional space spans trillions of candidates. This combinatorial explosion renders purely experimental exploration intractable given the prohibitive costs and time requirements of synthesis and analysis. Consequently, there is an urgent need for accurate predictive surrogate models that can provide reasonable estimates for material performance and ultimately prioritize candidate compositions before fabrication (Butler et al., 2018; Schmidt et al., 2019).

The emergence of self-driving laboratories (SDLs), which autonomously execute closed-loop synthesis and characterization cycles, has only heightened this need. These surrogate models must now be flexible enough to be updateable and keep pace with data arriving continuously rather than in the batched, human-paced campaigns of conventional research. Machine learning (ML) models have emerged as promising tools for this purpose, particularly in experimental settings where data are generated through iterative, lab-driven workflows (Tom et al., 2024; Hysmith et al., 2024).

Existing surrogate ML models can be divided into two distinct groups, each with complementary weaknesses. Purely data-driven regressors trained on heuristic compositional descriptors, such as atomic radius, valence electron concentration (VEC), and mixing entropy, which achieve reasonable accuracy within their training distribution (Rao et al., 2022; Gao et al., 2023; Vela et al., 2023). However, these models operate independently of the physical mechanisms governing deformation, which limits both their interpretabil-

ity and out-of-distribution generalization. On the other hand, physics-based models grounded in solid-solution strengthening (SSS) theory, most notably the Maresca–Curtin framework (Maresca and Curtin, 2020), are mechanistically interpretable but have primarily been evaluated under 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). Motivated by this observation, we hypothesize that the dominant source of error in the analytical prior lies in the elemental volumes supplied to the model rather than in the strengthening equations themselves.

In this work, we leverage both these two frameworks. 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 learned from data. Because these volumes are re-aggregated through the original physical equations, our framework functions as a physics-informed GP (PI-GP). Rather than superficially applying corrections in the output space, this formulation learns structured corrections within physically meaningful latent variables while fully preserving the governing strengthening equations, thereby maintaining interpretability at the level of individual elements.

Ultimately, this approach connects atomic-scale misfit physics, mesoscale microstructural heterogeneity, and macroscopic mechanical response within a single predictive framework that aligns naturally with autonomous experimental workflows (Kusne et al., 2020; Abolhasani and Kumacheva, 2023). Together, our work makes the following contributions:

(i) We demonstrate that the PI-GP framework outperforms existing GP-based and black-box baselines on the public Borg benchmark (Borg et al., 2020; Gao et al., 2023; Vela et al., 2023).

(ii) We show that integrating microstructural descriptors such as phase presence derived from X-ray diffraction (XRD) and elemental partition coefficients from energy-dispersive spectroscopy (EDS) with the PI-GP model yields state-of-the-art predictive accuracy on our proprietary Hardness Alloy Dataset from EXperiments (HADEX).

(iii) We show that the learned effective-volume correction remains physically bounded (typically under 5%) yet exhibits systematic nonlinear structures in composition space, revealing non-affine residual behaviors that cannot be captured through standard rule-of-mixtures or simple affine reparameterizations.

(iv) We show that the model's learned corrections carry signal aligned with the screw-to-edge strengthening transition identified by Baruffi et al. (2022), despite training only on macroscopic hardness.

2. Method

We introduce a PI-GP framework where metallurgical mechanics serve as a structured prior mean, while data-driven latent layers adjust unconstrained inputs to relax rigid physical assumptions. Rather than learning a post hoc correction in the output space, our model embeds learnable corrections directly within the physically meaningful latent variables governing SSS.

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 defined under a standard ROM baseline as:

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

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:

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

The prior mean $m(\mathbf{c}) \equiv \mathrm{HV}_{\mathrm{prior}}(\mathbf{c})$ evaluates purely on the compositional coordinates. The kernel $k_{\boldsymbol{\theta}}$ is a Matérn-5/2 with Automatic Relevance Determination (ARD), letting the GP learn a parallel, non-parametric spatial correction to the physics weighted by local data density in the full descriptor space $\mathbf{x}$.

2.3. Learnable Effective-Volume Misfit

We attribute the dominant source of error in the analytical prior to the ROM approximation that produces ΔV_i from elemental BCC lattice constants. We introduce a hierarchy of learnable latent modifications, all bounded and centered on

the ROM baseline, which alter the misfit parameter before it enters the prior mean.

Shared-σ scaling. The simplest correction applies a composition-dependent scalar to the ROM computed misfit parameter, denoted σ_{ROM}, yielding

$$\sigma_{\mathrm{scaled}} = \sigma_{\mathrm{ROM}} \exp\big(\beta \tanh g_\theta(\mathbf{c})\big). \tag{4}$$

where g_θ is a single-layer neural network and β bounds the maximum allowable adjustment magnitude.

Effective-volume reconstruction. For element-level interpretability, we instead map the composition to an element-resolved effective volume V_i^{eff} for each element:

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

where $V_i^{\mathrm{base}} = a_i^3/2$, and $h_{\theta,i}$ represents the coordinate outputs of a parameter-shared neural network head. The aggregate misfit parameter is then reconstructed by re-applying the structural rules of Eq. (1) to these learned effective volumes:

$$V_{\mathrm{eq}}^{\mathrm{eff}} = \sum_i c_i V_i^{\mathrm{eff}}(\mathbf{c}), \quad \Delta V_i^{\mathrm{eff}} = V_i^{\mathrm{eff}}(\mathbf{c}) - V_{\mathrm{eq}}^{\mathrm{eff}}. \tag{6}$$

The resulting misfit parameter is then

$$\sigma = \sum_i c_i (\Delta V_i^{\mathrm{eff}})^2. \tag{7}$$

This reconstructed parameter replaces the ROM-computed misfit parameter in the downstream Maresca–Curtin evaluation. This ensures that the data-driven corrections remain perfectly constrained by the structural equations of dislocation interaction.

2.4. Microstructural Features

The GP operates on a base feature set of eight compositional descriptors $\phi(\mathbf{c}) \in \mathbb{R}^8$ computed from a composition $\mathbf{c} = \{c_i\}$: mean atomic radius $\bar{r}$, radius variance, radius mismatch $\delta_r = \sqrt{\sum_i c_i (1 - r_i/\bar{r})^2}$ (identical to δ_{gate} in Section 2.6), composition-weighted bulk and shear moduli, Poisson mismatch δ_ν, mean VEC, and mean melting temperature. All eight are standardized to zero mean and unit variance across the training fold.

For HADEX we additionally construct a 17-dimensional microstructural feature vector that exposes the as-cast dendritic structure to the GP kernel. The physics prior is unchanged — the Maresca–Curtin model has no slot for per-phase composition, so the microstructural descriptors enter only through the GP kernel, where the ARD-Matérn 5/2 covariance learns per-dimension length scales jointly with the base descriptors. This keeps the mechanistic strengthening prediction tied to the bulk composition while letting the data-driven correction respond to local chemical heterogeneity. The microstructural feature vector is built from three measured quantities per sample: bulk composition $\mathbf{c}$ from SEM-EDS, elemental partition coefficients $\{k_i\}$ from EDS analysis of dendritic core versus inter-dendritic shell regions, and the core phase fraction f. We adopt the standard EPMA core-to-bulk convention in which k_i scales the bulk fraction toward the solidified core, and reconstruct core and shell compositions by

$$\tilde{c}_i^{\mathrm{core}} = k_i c_i, \qquad c_i^{\mathrm{core}} = \frac{\tilde{c}_i^{\mathrm{core}}}{\sum_j \tilde{c}_j^{\mathrm{core}}}, \tag{8}$$

$$\tilde{c}_i^{\mathrm{shell}} = \max\left(0,\ \frac{c_i - f\, c_i^{\mathrm{core}}}{1 - f}\right), \quad c_i^{\mathrm{shell}} = \frac{\tilde{c}_i^{\mathrm{shell}}}{\sum_j \tilde{c}_j^{\mathrm{shell}}}. \tag{9}$$

Eq. 8 normalizes the core composition to sum to one, since $\sum_j \tilde{c}_j^{\mathrm{core}}$ is generally not unity, and Eq. 9 applies a lever-rule mass balance with a non-negativity clamp on shell fractions before renormalization. Both steps relax exact conservation of bulk composition in exchange for descriptors that remain physically interpretable (valid atomic fractions on each phase) when the measured k_i and f are slightly inconsistent, as is typical for EDS-derived partition data. Elements absent from the partition table default to $k_i = 1$ (no segregation). The eight compositional descriptors ϕ are then evaluated on each reconstructed phase, yielding $\phi(\mathbf{c}^{\,\mathrm{core}}) \in \mathbb{R}^8$ and $\phi(\mathbf{c}^{\,\mathrm{shell}}) \in \mathbb{R}^8$, and f is retained as a seventeenth scalar feature. The XRD peak count, when used, enters as a separate scalar feature alongside the 17 microstructural descriptors and is standardized in the same fold-wise manner as the rest of the feature vector. The ablations in Section 4.2 switch each of these feature groups on and off independently.

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.

2.6. Dual-Branch Strengthening and Mechanism Gate

To represent edge- versus screw-dominated strengthening, we evaluate the Curtin expression twice using the same learned misfit parameter σ and the same elastic intermediates, varying only the line-tension parameter: an edge branch at $\alpha_e = 1/12$ and a screw branch at $\alpha_s = 0.5$. Each branch yields a temperature-dependent strength (σ_y^e, σ_y^s) mapped to hardness (HV^e, HV^s) by the same Tabor relation (Eq. 2).

A branch is selected by a Baruffi-style atomic-size mismatch $\delta_{\mathrm{gate}} = \left[\sum_i c_i(1 - r_i/\bar{r})^2\right]^{1/2}$, with $\bar{r} = \sum_i c_i r_i$, converted to an edge-applicability score $p_e = \mathrm{sigmoid}\big(k(\delta_{\mathrm{gate}} - \delta_0)\big)$ with fixed $\delta_0 = 0.035$ and $k = 250$. This score is used in two distinct modes. In *diagnostic* mode, p_e serves only as a mechanism label ($\hat{m} = \mathrm{edge}$ if $p_e \geq 0.5$, else screw) and does not enter the hardness prediction, which remains the edge branch; this is the mode evaluated against literature mechanisms in Section 4.4. In *gated* mode, a learned gate strength $g \in (0, 1)$ blends the branches as $w_e = 1 - g(1 - p_e)$, $w_s = g(1 - p_e)$, giving the convex hardness prior $\mathrm{HV}_{\mathrm{prior}} = w_e \mathrm{HV}^e + w_s \mathrm{HV}^s$. The two modes share the same δ_{gate} but serve different purposes: diagnostic mode gives an interpretable classifier independent of the regression fit, while gated mode lets the prior interpolate smoothly between mechanisms.

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.

HADEX. The HADEX (Hardness Alloy Dataset from EXperiments) dataset comprises Vickers micro-hardness measurements (Qness Q10A+) of as-cast RHEAs synthesized within an internally developed experimental workflow for rapid alloy exploration. All measurements are at room temperature; bulk compositions are determined by SEM-EDS and XRF. While Borg provides a standardized, composition-only evaluation, HADEX is a complementary setting in which per-sample microstructural information is available but cannot be inferred from composition alone (DeCost et al., 2017; Peng et al., 2025; Latypov et al., 2019).

Each sample is additionally characterized by XRD and SEM. An XRD spectrum is collected and passed through an automated peak-detection workflow, yielding a peak count n_{XRD} that serves as a continuous proxy for phase multiplicity (the major phase is consistently BCC across HADEX). Automated segmentation of backscattered secondary electron micrographs and EDS elemental maps yields per-phase quantities including elemental partition coefficients $\{k_i\}$ between dendritic core and inter-dendritic shell, and the core phase fraction f. Together, $\{\mathbf{c}, k_i, f, n_{\mathrm{XRD}}\}$ are the inputs to the microstructural feature vector of Section 2.4.

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 $4.3\times$ 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
Vela HV stage (Vela et al., 2023)	42.5	67.4
PI-GP (analytic mean)	37.5	53.3
PI-GP + shared-σ	37.6	53.1
PI-GP + effective volume	**35.8**	**50.4**

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 HADEX alloys; (iii) the learned misfit-volume correction is small but systematically nonlinear; (iv) the model's learned corrections align with 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-

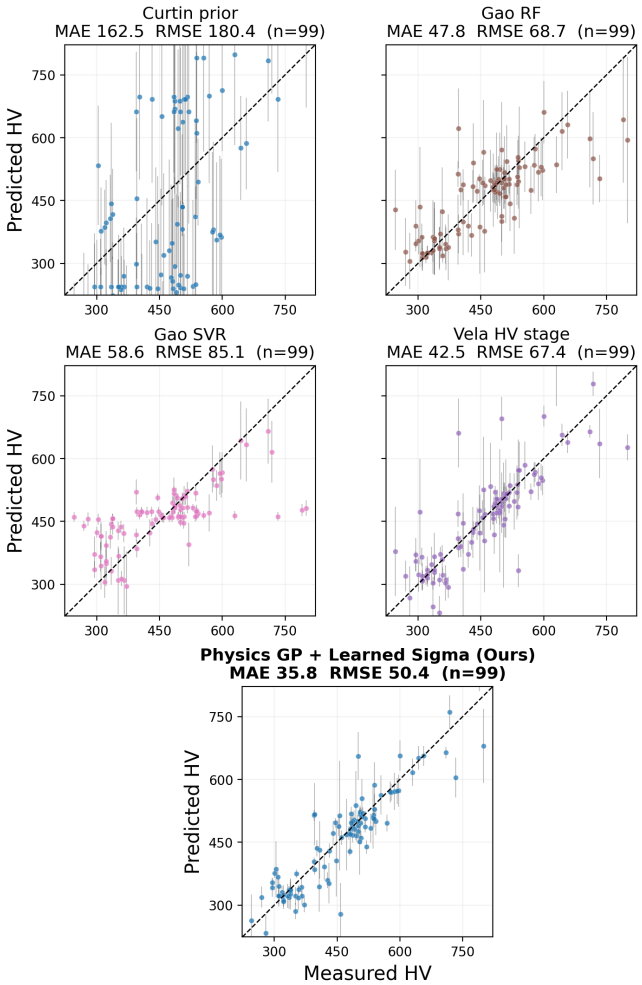

Figure 1. Parity plots on the Borg dataset for the physics-based model family. Vertical whiskers denote the GP predictive standard deviation. Moving from the bare Curtin prior to the physics-informed GP improves the fit, and adding the learnable effective-volume σ correction further tightens parity while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

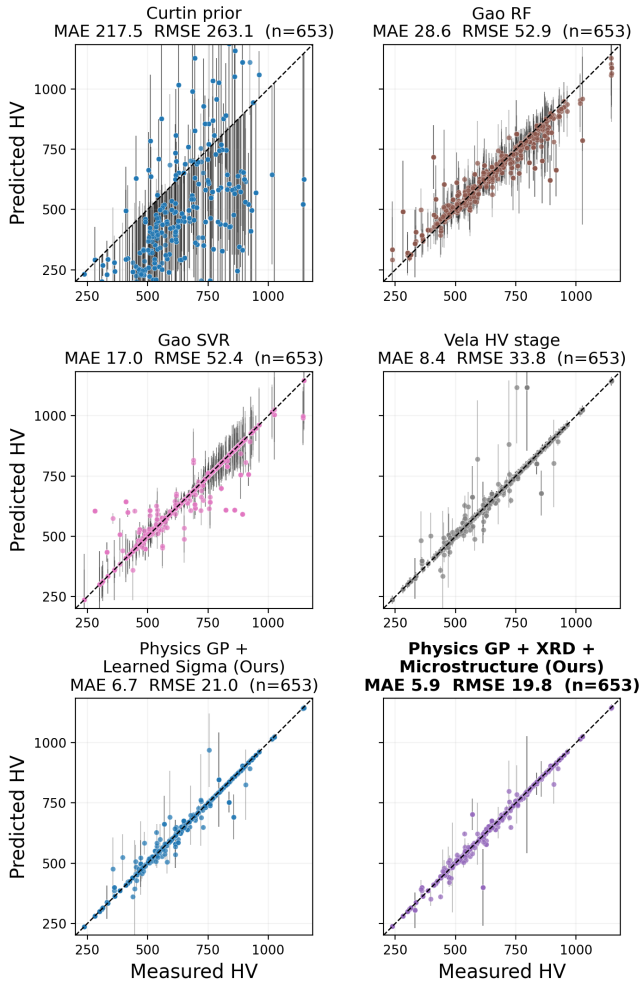

Figure 2. Five-fold cross-validation error summary on the HADEX dataset. Bars give mean MAE/RMSE across folds; whiskers give fold-to-fold standard deviation. Augmenting the physics-informed GP with experimental descriptors (XRD peak count, microstructural features) yields further improvement on top of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

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 35.8 HV and 50.4 HV respectively, which we read 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 HADEX 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 HADEX 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 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 217.5 HV to 6.7 HV) and beats the strongest black-box baseline (Gao SVR at 17.0 HV) by roughly $2.5\times$, 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 (5.9 HV MAE, 19.8 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 HADEX 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 question is whether the learned correction is physically meaningful or merely re-tunes the elemental BCC volumes entering 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, with a median deviation of only a few percent, so a correction this small could in principle be absorbed by a global affine reparameterization of the elemental lattice constants a_i. To further evaluate this, we decompose the learned ΔV surface into this affine component plus a non-affine residual, and show that the residual is both structured and functionally load-bearing.

Figure 3 plots, for four representative ternary subsystems, the ROM ΔV surface, the learned ΔV surface, their difference, and the residual of the learned surface to its best-fit plane. The affine component dominates the bulk shape, which is consistent with global retuning of the underlying lattice constants, and thus by itself reproducible within the analytical framework of Maresca and Curtin (2020). Despite the small residual with a normalized RMSE on the order 10^{-3}, further analysis with Moran's I demonstrates it to be spatially coherent. Moran's I ranges between 0.815 and 0.833 (mean 0.828) across the four surfaces, against a spatial-null expectation of -0.004. The residual therefore encodes a nonlinear-in-composition contribution to V_i^{eff} that no global reparameterization of the elemental a_i could reproduce.

4.4. The Model Internalizes the Screw/Edge Regime

BCC refractory high-entropy alloys undergo a physical transition between edge-dislocation line-tension and screw-dislocation kink-pair regimes, dictated by local atomic size misfit (Baruffi et al., 2022). Because our framework parameterizes both dislocation modes within its diagnostic head (Section 2.6), we ask whether the learned correction internalizes this transition despite receiving no mechanism labels or yield-strength supervision.

We take the misfit screen of Baruffi et al. (2022) as an independent reference. Across the fourteen alloys of Table 2, the fixed diagnostic-mode screen of Section 2.6 agrees with the literature mechanism on 10 of the 12 non-mixed labels. This screen is geometric and does not depend on our learned head; it supplies the reference labels, not the discovery.

The learned non-affine residual carries information about the underlying deformation mechanism. Using the model's edge-yield branch to separate alloys along the Baruffi misfit axis, constraining the correction to its affine component weakens that separation, most strongly in the high-misfit regime (R^2 from 0.45 to 0.37 for $\delta_{\mathrm{gate}} \geq 0.035$) where edge behavior dominates. The correction therefore contributes most where the screw-to-edge transition occurs, consistent with the residual capturing mechanism-relevant structure rather than simply fitting hardness, although the modest correlations and small alloy count ($n = 12$) make this corroborating rather than definitive.

The emergence of this screw/edge boundary from macroscopic hardness alone—without direct supervision on yield strength or mechanism labels—suggests that the learned effective-volume representation captures physically meaningful structure. For closed-loop alloy discovery workflows, this diagnostic behavior is directly useful: the same surrogate that predicts hardness to within a few HV can also indicate when edge-centered analytical mechanics are likely to remain valid.

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 hier-

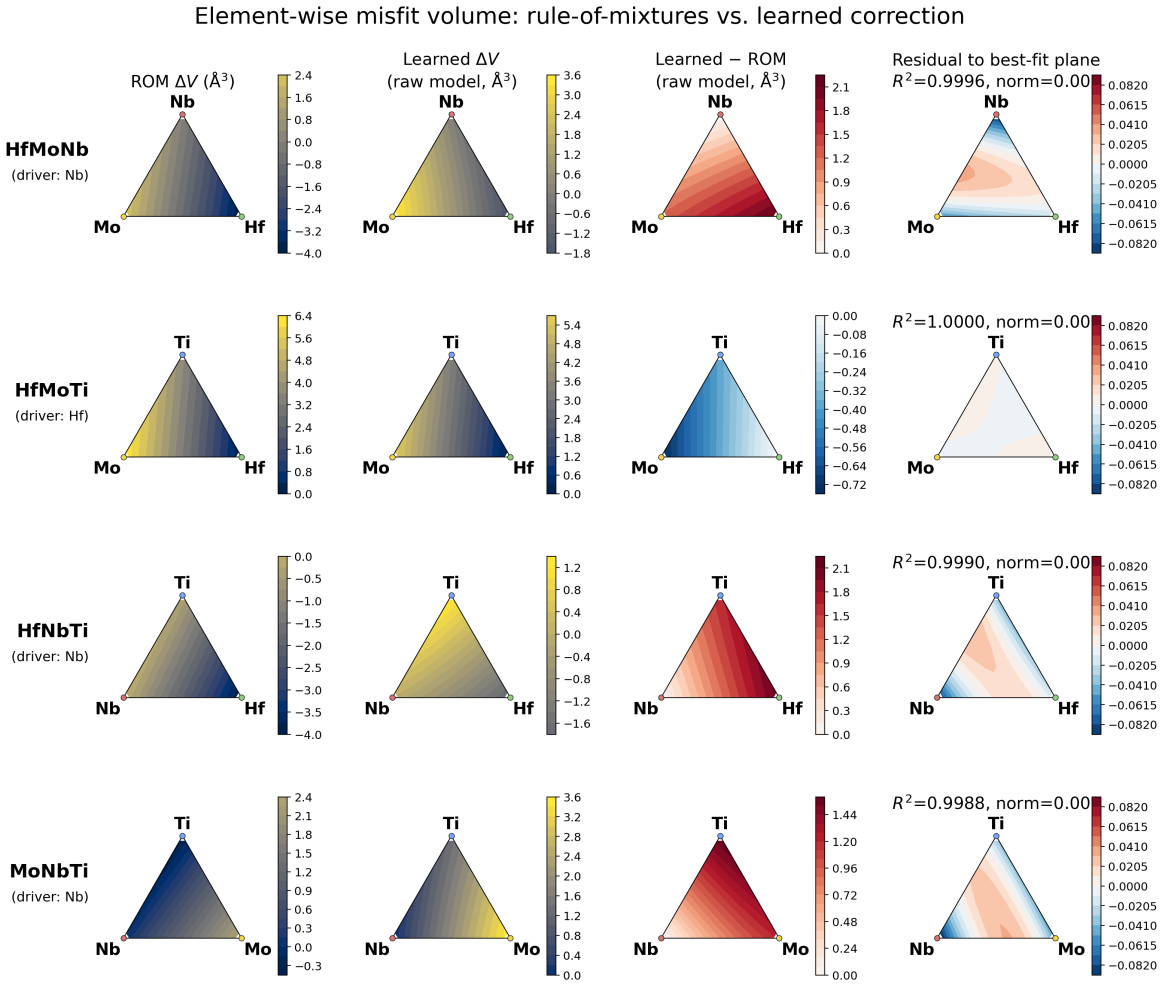

Figure 3. Element-wise misfit-volume surfaces for the best HADEX composition-only model. Each row picks the most nonlinear element for that ternary subsystem and shows, from left to right, the rule-of-mixtures ΔV, the learned ΔV, their difference, and the residual of the learned surface to its best-fit plane (with R^2 and a normalized RMSE). The residuals contain Moran's I from 0.81-0.83 against a null of -0.0044.

archical 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 SSS 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 SSS 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 (35.8 HV MAE) and on HADEX when combined with the microstructural feature set (5.9 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

Table 2. Baruffi mechanism reference for fourteen literature alloys. The misfit screen of Baruffi et al. (2022) (δ_{gate}, edge probability p_e; $p_s = 100 - p_e$) agrees with the literature mechanism on 10 of 12 non-mixed labels and supplies the reference classifications used below. The two mixed-label alloys have no single ground-truth mechanism and are excluded from scoring. The final column gives the model's gated edge weight w_e ($w_s = 100 - w_e$), used for hardness blending, not classification.

Composition	Baruffi	δ_{gate}	p_e %	Diagnostic	Result	w_e %
Mixed label — excluded from scoring						
W50 Ta50	mixed	0.0246	6.9	screw	—	70.9
W60 Ta40	mixed	0.0242	6.3	screw	—	70.7
Non-mixed label — scored						
W80 Ta20	edge	0.0199	2.3	screw	mismatch	69.4
Ta42 Nb40 V18	edge	0.0321	32.4	screw	mismatch	78.9
Ta90 Hf1 W9	screw	0.0166	1.0	screw	match	69.1
Nb71.7 Hf22.4 W5.9	edge	0.0396	75.7	edge	match	92.4
Nb68.4 Hf20 W6.4 Ta2.6 Zr2.6	edge	0.0402	78.7	edge	match	93.3
Nb65.6 Hf20 W9.2 Ta2.6 Zr2.6	edge	0.0416	83.8	edge	match	94.9
Nb90.7 Hf5.4 Ti2.6 Zr0.74 Ta0.26 W0.26	screw	0.0216	3.4	screw	match	69.8
Nb50 Ti50	screw	0.0034	0.0	screw	match	68.8
Nb39 Ti61	screw	0.0033	0.0	screw	match	68.8
Nb70 Ti18.1 Mo11.9	screw	0.0163	0.9	screw	match	69.0
Nb67.8 Ti18.1 Mo14.1	screw	0.0175	1.2	screw	match	69.1
Nb34.7 Ti32.5 Mo32.8	screw	0.0246	6.8	screw	match	70.9
Screen-vs-literature agreement (non-mixed labels)						10/12

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 HADEX 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 HADEX 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.

One limitation of this work is the heterogeneity and reproducibility of experimental data across sources. The HADEX 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 HADEX 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.

Hysmith, H., Foadian, E., Padhy, S.P., Kalinin, S.V., Moore, R.G., Ovchinnikova, O.S., Ahmadi, M., 2024. The future of self-driving laboratories: from human in the loop interactive ai to gamification. Digital Discovery 3, 621–636.

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.

Tom, G., Schmid, S.P., Baird, S.G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E.M., Skreta, M., et al., 2024. Self-driving laboratories for chemistry and materials science. Chemical Reviews 124, 9633–9732.

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.

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)
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
$g_\theta,\ h_\theta$	Neural network correction functions
$\alpha_e,\ \alpha_s$	Edge / screw line-tension parameters
δ_{gate}	Baruffi atomic-size mismatch
p_e	Edge-applicability score
g	Learned gate strength
$w_e,\ w_s$	Edge / screw blend weights
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