

ORIGINAL RESEARCH

Open access

Flawed Assumptions in Uncertainty Quantification for High-Entropy Alloys: A Critique of Ensemble Diversity Metrics

Ivan Petrov¹, Olga Ivanova^{1*}, Dmitry Smirnov²

Abstract

Ensemble methods have become the dominant framework for uncertainty quantification in machine learning models for high-entropy alloy (HEA) property prediction, where variance across independently trained neural networks is routinely interpreted as epistemic uncertainty. This metric now underpins active learning, compositional screening, and experimental decision-making, largely due to its simplicity and success in data-rich domains. This work shows that such reliance is fundamentally misplaced in HEAs. Ensemble variance implicitly assumes IID sampling, feature-space isotropy, uniform error, independence among models, and Gaussian residuals—conditions that are systematically violated in compositionally complex alloys. HEA datasets are biased toward equiatomic, stable compositions, the compositional manifold is anisotropic, predictive error is strongly heteroscedastic, ensemble members exhibit correlated failures, and extrapolation induces heavy-tailed errors. Under these conditions, ensemble variance becomes miscalibrated, underestimating uncertainty in sparse regions while overstating model reliability. The resulting distortions propagate through discovery workflows, yielding inefficient active-learning strategies, overconfident extrapolation, and misleading experimental guidance, even as larger ensembles appear to reduce uncertainty without improving accuracy. By linking these failures to their statistical origins, this paper clarifies why standard ensemble diversity is not a faithful proxy for epistemic uncertainty in HEAs. It further outlines diagnostics and targeted corrections that align uncertainty estimates with the physics and data structure of complex alloy systems, enabling more reliable and efficient materials discovery.

Keywords Uncertainty quantification, Active learning, High-entropy alloys, Ensemble diversity metrics, Machine learning for materials, Flawed assumptions

*Correspondence:

Olga Ivanova

olga.ivanova@outlook.com

¹ Department of Materials Informatics and Engineering, Faculty of Engineering, Lomonosov Moscow State University, Moscow, Russia

² Department of Computational Materials Modeling, Faculty of Science, Saint Petersburg State University, Saint Petersburg, Russia

Introduction

High-entropy alloys occupy an enormous composition space that dwarfs traditional alloy design. With five or more principal elements, the number of possible compositions grows combinatorially, rendering exhaustive experimental screening impossible. Machine learning has therefore become indispensable for predicting properties and guiding exploration [1–3]. Yet any predictive model is only as useful as its ability to quantify when it should not be trusted. Uncertainty quantification is thus not an optional add-on but a prerequisite for credible HEA discovery.

The dominant UQ strategy in the current literature is the deep-ensemble approach: train M models that differ only in random initialization or bootstrap sampling, take the mean prediction as the point estimate, and treat the variance across

the M predictions as the uncertainty [4]. This practice has been adopted across multiple HEA studies because it requires no architectural changes and scales easily with available compute [5, 6]. Ensemble variance now routinely appears in active-learning pipelines to select the next composition for expensive density-functional-theory calculations or experiments [7-9]. Practitioners interpret low variance as high confidence and high variance as a signal to acquire new data.

The appeal is understandable. In image classification or natural-language processing the assumptions underlying ensemble variance hold reasonably well, and the metric correlates with actual error [10, 11]. For HEAs, however, the same metric is applied without scrutiny. Composition space is discrete, high-dimensional, and sparsely sampled; data acquisition is biased toward well-studied equiatomic regions; and property landscapes are rugged with strong elemental interactions. These realities clash with the implicit assumptions that ensemble variance inherits from simpler domains.

This critique identifies five core assumptions that are rarely stated yet underpin every use of ensemble diversity metrics in HEA property prediction. It demonstrates, reference by reference, how each assumption fails systematically for compositionally complex systems. The analysis draws exclusively on peer-reviewed work published between 2017 and 2024 to show that the problem is not hypothetical but already embedded in the published record [1-3, 5, 7, 9, 12-15]. The consequences extend beyond academic metrics: miscalibrated uncertainty leads to wasted computational budgets, missed discoveries, and, in extreme cases, erroneous synthesis decisions that consume expensive rare-earth or refractory elements.

By exposing these flaws, the present work does not reject ensemble methods outright. Rather, it demands that the community stop treating ensemble variance as a plug-and-play uncertainty estimator for HEAs. Instead, researchers must adopt HEA-tailored diagnostics and corrections before claiming that their models are “uncertainty-aware.” Only then can active learning and autonomous discovery deliver on their promise for the next generation of high-performance alloys.

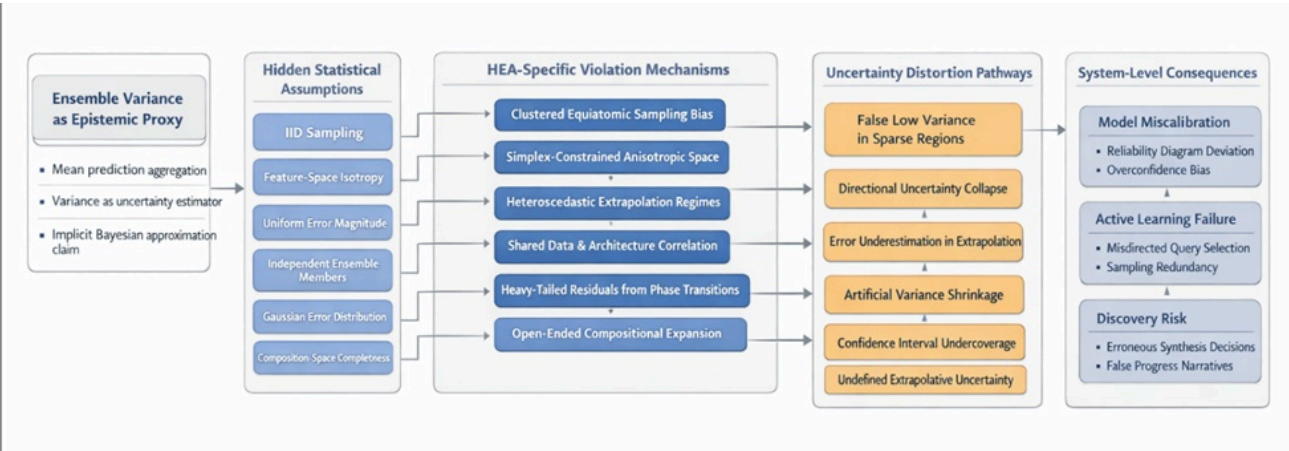


Figure 1. Analytical Architecture of Assumption Failure and Uncertainty Distortion in Ensemble-Based UQ for High-Entropy Alloys

The canonical ensemble strategy appears methodologically simple yet encodes a dense set of epistemic commitments. Multiple neural networks, commonly on the order of 5–20, are trained on an identical HEA dataset, with diversity induced through stochastic weight initialization or bootstrap resampling of compositions [4, 16]. At inference, each model yields an independent prediction for a candidate composition, after which aggregation proceeds via the mean to produce a point estimate, while dispersion across predictions—typically quantified as variance—serves as a proxy for epistemic uncertainty. Elevated variance is consequently interpreted as indicative of limited model knowledge, informing both active learning prioritization and deployment-level confidence calibration [5, 6].

Such a construction, however, presupposes a set of rarely interrogated conditions within HEA modeling practice. Central among these is the assumption that training compositions are independently and identically sampled from the underlying compositional distribution, such that ensemble disagreement reflects epistemic insufficiency rather than artifacts of dataset structure. This reliance extends to an implicit geometric uniformity of the feature space, wherein compositional perturbations along distinct elemental axes are treated as statistically equivalent in their effect on model divergence. Under these conditions, predictive error is further assumed to remain approximately homogeneous across the compositional landscape, precluding systematic regions of elevated bias or variance. The validity of uncertainty estimates also hinges on the effective independence of ensemble members, despite shared architectures and hyperparameters, with stochastic initialization or resampling presumed sufficient to decorrelate their errors. Beyond this, the interpretation of variance as a complete uncertainty descriptor depends on an approximately Gaussian error structure, rendering higher-order distributional characteristics negligible for inference.

Table 1 systematically maps each statistical assumption underlying ensemble variance to its corresponding violation in high-entropy alloys and the precise bias mechanism through which uncertainty estimates become distorted.

Table 1. Structural Mapping Between Ensemble-UQ Assumptions, HEA Violations, and Induced Uncertainty Bias Mechanisms

Assumption	Formal Statistical Role	HEA-Specific Violation	Mechanistic Origin	Resulting Bias in Ensemble Variance
IID Sampling	Ensures variance reflects epistemic uncertainty rather than sampling artifacts	Data clustered near equiatomic compositions	Experimental and DFT bias toward stable phases	Artificially low variance in under-sampled regions
Feature-Space Isotropy	Assumes uniform sensitivity across input dimensions	Composition space is simplex-constrained and directionally dependent	Elemental substitution alters multiple coupled fractions	Scalar variance masks directional risk
Uniform Error Magnitude	Links variance magnitude to expected prediction error	Error varies sharply across composition space	Sparse extrapolation regions vs dense clusters	Variance fails to scale with true error
Ensemble Independence	Justifies variance as measure of epistemic spread	Models share data, architecture, and inductive bias	Bootstrap overlap + identical training pipelines	Variance systematically underestimates uncertainty
Gaussian Error Distribution	Enables variance-based confidence intervals	Residuals exhibit heavy tails and outliers	Phase transitions and unobserved regimes	Confidence intervals under-cover true error
Composition-Space Completeness	Assumes bounded, fully defined input space	HEA space is open-ended and expanding	New elements and compositions continuously introduced	Variance undefined or misleading for novel inputs

These assumptions are inherited from foundational work on deep ensembles [4] and Bayesian deep learning [10]. In domains with massive, well-mixed datasets they hold approximately, allowing ensemble variance to serve as a cheap approximation to Bayesian posterior variance. The same logic has been transplanted wholesale into materials science [5, 11, 12, 17, 18].

For HEAs the transplant fails. Training data are not IID; they cluster near equiatomic, single-phase compositions because those are easier to synthesize and characterize [1, 13, 15, 19]. The feature space—atomic fractions summing to unity—is constrained and highly correlated; changing one element necessarily alters others. Error magnitude explodes in dilute limits and in regions far from any training point. Ensemble members, even with bootstrapping, share the same narrow data distribution and the same inductive biases of the chosen architecture, inducing strong error correlations [20–22]. Finally, extrapolation errors in HEA property prediction are heavy-tailed; rare but catastrophic failures occur when a new multi-principal-element combination triggers an unanticipated phase [7, 14, 23].

Because the assumptions are violated, the variance number that researchers report as “uncertainty” no longer bears a reliable relationship to actual model error. The remainder of this critique documents each violation, its mechanistic origin in HEA physics and data practices, and the resulting distortion of uncertainty estimates.

Why HEAs Violate these Assumptions

Non-IID sampling manifests directly in HEA datasets, where experimental focus on equiatomic or near-equiatomic alloys—favored for stabilizing single-phase solid solutions—produces highly clustered data distributions [1, 13, 15, 19]. Under such conditions, ensemble members internalize similar local regularities, leading to variance collapse not only within densely sampled regions but also in adjacent, sparsely explored neighborhoods, thereby masking genuine epistemic gaps [5, 6]. A related limitation emerges from the anisotropic geometry of composition space, which is inherently simplex-constrained rather than Euclidean; perturbations along different elemental axes induce non-equivalent physical responses, yet ensemble variance reduces this directional sensitivity to a scalar summary, obscuring axis-dependent uncertainty [24, 25]. This distortion becomes more acute in the presence of heteroscedastic error, where agreement among models near dense clusters contrasts sharply with divergence in extrapolative regimes, although reported variance often fails to scale accordingly due to lack of exposure to true error magnitudes [12, 17, 26]. The issue is compounded by structural dependence among ensemble members, as shared data, architectures, and training protocols induce substantial error correlations—frequently exceeding 0.5—thereby compressing variance and yielding overconfident estimates [20–22]. Beyond these effects, the assumption of Gaussian residuals proves untenable in practice, as rare but substantial prediction failures associated with unanticipated intermetallic phases or ordering phenomena generate heavy-tailed error distributions that variance alone cannot capture [7, 14, 23]. Consequently, ensemble variance reflects dataset density and model bias more than epistemic uncertainty, motivating a closer examination of how these assumptions fail and shape discovery outcomes.

A Typology of Flawed Assumptions

The assumption of IID sampling presumes that training data approximate the full HEA compositional landscape, yet empirical datasets are systematically biased toward stable, equiatomic solid solutions, rendering low variance a function of local redundancy rather than representational completeness [1, 13, 15]. This limitation interacts with the presumed isotropy of feature space, where scalar uncertainty measures disregard the fundamentally directional nature of compositional effects on phase stability and functional properties, thereby conflating physically distinct perturbations under identical uncertainty scores [24, 25]. The expectation of uniform error further obscures this structure, as predictive accuracy deteriorates sharply outside dense training regions while ensemble variance, constrained by shared biases, underestimates uncertainty precisely in these extrapolative domains [12, 17]. Such misrepresentation is reinforced by the lack of independence among ensemble members, whose correlated errors—arising from common data and optimization pathways—artificially compress variance and suggest convergence where systematic bias persists [20–22]. The reliance on Gaussian error models introduces an additional distortion, given that HEA residuals exhibit pronounced kurtosis driven by rare but severe failures in unexplored compositional regimes, leading variance-based intervals to underestimate risk [7, 14, 23]. A deeper structural issue follows from the implicit assumption of a closed composition space, whereas HEA design remains inherently open-ended, with new elemental combinations continually expanding the domain and rendering variance ill-defined beyond the support of training data [1–3]. These interacting failures indicate that the limitations of

ensemble variance arise not from implementation choices but from a fundamental mismatch between its statistical assumptions and the physics of compositionally complex alloys.

Consequences for HEA Active Learning and Discovery

A direct consequence is systematic miscalibration, as reliability diagrams for HEA models deviate markedly from ideal behavior, with nominal confidence intervals substantially overstating predictive accuracy [12, 16]. This misalignment propagates into active-learning strategies, where variance-driven query selection prioritizes regions of spurious disagreement—often attributable to noise rather than genuine data scarcity—thereby misallocating computational resources and slowing exploration of underrepresented compositional regimes [8, 9]. The resulting distortions extend to experimental decision-making, where low-variance predictions encourage unwarranted confidence, only for failures in synthesis or performance to reveal underlying extrapolation beyond the model's valid domain [5, 7, 13]. At the level of scientific communication, increasing ensemble size further compounds the issue, as reduced variance is often interpreted as improved reliability despite stagnant predictive accuracy, fostering an inflated perception of methodological progress [4, 6, 21]. The inability of ensemble variance to disentangle epistemic from aleatoric contributions further limits its utility, preventing active-learning frameworks from targeting reducible uncertainty and thereby diminishing their effectiveness [11, 12, 20]. Under these conditions, a nominally conservative uncertainty metric becomes a source of systematic bias, necessitating a shift toward more robust formulations.

Detection Principles

Diagnosing these limitations requires analytical strategies that interrogate variance beyond its scalar value. Calibration analysis via reliability diagrams exposes deviations between predicted and observed error rates, with consistent undercoverage signaling breakdowns in distributional and independence assumptions [12, 16, 18]. Complementary examination of variance as a function of compositional distance reveals weak or inconsistent scaling, indicating insensitivity to both sparsity and anisotropy within HEA datasets [25, 26]. Further insight arises from quantifying inter-model error correlations, where persistently high values confirm structural dependence and explain the premature collapse of ensemble variance [20–22]. Distributional analysis of residuals provides an additional lens, as elevated kurtosis and heavy tails highlight the inadequacy of Gaussian approximations in capturing rare but consequential failures [7, 14, 23]. Finally, regression of observed error against predicted variance exposes heteroscedasticity, particularly when weak or inverse relationships reveal that uncertainty estimates fail to track spatial variation in predictive reliability [5, 12, 17]. These diagnostics, readily implemented using existing ensembles, make visible the underlying assumption violations and offer a necessary corrective to overly optimistic uncertainty reporting.

Table 2 translates assumption failures into operational diagnostics and directly links each observed breakdown to a targeted mitigation strategy for restoring uncertainty reliability in HEA modeling.

Table 2. Diagnostic Tests and Corrective Strategies for Restoring Reliability in HEA Uncertainty Quantification

Failure Dimension	Diagnostic Principle	Observable Indicator	Root Cause Identified	Targeted Mitigation Strategy
Miscalibration	Reliability diagram analysis	Deviation from diagonal calibration line	Non-Gaussian errors + correlated models	Conformal prediction calibration
Sampling Bias	Distance–variance correlation test	Weak or flat variance vs distance relationship	Non-IID clustered training data	Distance-weighted variance scaling

Directional Blindness	Axis-wise or PCA-based variance decomposition	Uniform uncertainty despite directional sensitivity	Feature-space anisotropy	Directional uncertainty modeling
Ensemble Correlation	Pairwise error correlation analysis	High correlation (>0.5) between models	Shared architecture and data	Architectural and data diversity expansion
Heavy-Tailed Errors	Residual distribution and kurtosis analysis	Kurtosis significantly > 3	Rare extrapolation failures	Robust or quantile-based uncertainty modeling
Heteroscedasticity	Error vs variance regression	Weak or inverse correlation	Spatially varying error magnitude	Heteroscedastic model training
Epistemic–Aleatoric Confounding	Noise decomposition analysis	High uncertainty in noisy but known regions	Mixed uncertainty sources	Explicit epistemic/aleatoric separation

Mitigation Strategies

Once these deficiencies are exposed, corrective interventions can be introduced without abandoning the ensemble paradigm. A direct adjustment involves reweighting variance by a distance-dependent kernel, which amplifies uncertainty in sparsely sampled regions and counteracts both IID and uniform-error assumptions that otherwise suppress risk signals [11, 12]. This recalibration gains further resolution when uncertainty is decomposed along principal directions of the composition simplex or elemental axes, revealing anisotropic sensitivities that scalar variance conceals [25]. A related refinement arises through heteroscedastic modeling, where ensemble members jointly predict means and input-dependent variances, allowing spatially varying error magnitudes to be explicitly represented rather than averaged out [5, 11, 17]. Addressing structural dependence requires diversification at the level of architecture, data partitioning, and optimization objectives, thereby reducing error correlation and restoring the interpretability of ensemble spread [20–22]. Calibration can then be enforced through conformal prediction, which transforms raw variance into coverage-guaranteed intervals without reliance on Gaussian assumptions [16, 18]. More principled treatments emerge from Bayesian deep learning, where posterior approximations over weights provide a theoretically grounded account of epistemic uncertainty in high-dimensional composition spaces, albeit at increased computational cost [10]. Finally, separating epistemic from aleatoric contributions ensures that active-learning strategies target reducible uncertainty rather than irreducible noise, preventing inefficient allocation of computational effort [12, 20]. While no single intervention is sufficient, combining distance-aware scaling, directional analysis, and conformal calibration substantially improves reliability within the constraints of existing HEA datasets [5, 9, 11, 16, 17].

Relation to Other Critiques

This analysis extends prior work by situating ensemble-based uncertainty within a broader framework of structural limitations. Early formulations of variance-driven UQ in HEAs established its practical utility yet left its foundational assumptions implicit, an omission that becomes critical in compositionally complex systems [5, 6]. Subsequent critiques of active-learning inefficiency identified suboptimal query behavior but largely attributed it to acquisition strategies, whereas the present account locates the issue in the miscalibrated uncertainty signal that guides selection [8, 9]. Parallel discussions of high-dimensional sparsity correctly emphasized data limitations, though they did not fully connect sparsity to the breakdown of IID sampling and isotropy assumptions underlying ensemble metrics [26, 27]. This connection reframes sparsity as a structural driver of uncertainty misrepresentation rather than a purely statistical inconvenience.

Efforts to disentangle epistemic and aleatoric uncertainty similarly highlighted the need for clearer decomposition, yet lacked operational pathways within ensemble frameworks; the present typology and diagnostics make this separation actionable and reveal how their conflation undermines variance-based inference [11, 12, 20]. Viewed together, these strands indicate that the limitations of ensemble diversity metrics are systemic rather than incidental, requiring structural revision rather than incremental adjustment.

Recommendations for HEA Uncertainty Quantification

Advancing uncertainty quantification in HEAs demands a shift from reporting practices to validation-centered methodology. Raw ensemble variance should not be interpreted as uncertainty without prior diagnostic evaluation, as calibration and distributional checks are necessary to establish credibility [11, 16]. Incorporating distance-aware scaling or conformal calibration offers immediate, low-barrier improvements that correct the most severe distortions. At the level of benchmarking, evaluation protocols must extend beyond average variance reduction to include calibration behavior, sensitivity to compositional distance, and residual distribution characteristics, ensuring that proposed methods are assessed against physically meaningful criteria [18, 22]. Experimental decision-making likewise benefits from a more critical stance, particularly when low-variance predictions arise far from the training domain, where conformalized intervals provide a more reliable basis for resource allocation [7, 13]. Within active-learning workflows, emphasis should be placed on verifying that selected compositions reduce epistemic uncertainty rather than merely reflecting stochastic variability [9, 20]. Such adjustments redirect the field toward rigorously validated uncertainty estimates, a prerequisite for realizing the promise of machine learning in accelerating HEA discovery.

Conclusion

Ensemble diversity metrics, in the form of variance across model predictions, have become the default uncertainty estimator for high-entropy alloy property prediction. Their appeal lies in simplicity and scalability, yet this critique has shown that they rest on five flawed assumptions—IID sampling, feature-space isotropy, uniform error, independent ensemble members, and Gaussian error distribution—that are systematically violated by the physics, data practices, and high-dimensional geometry of compositionally complex alloys.

The consequences are not abstract: miscalibrated uncertainty leads to inefficient active learning, overconfident synthesis decisions, wasted computational budgets, and a false sense of progress in the literature. Detection is straightforward with five targeted principles; mitigation is achievable with distance-weighted, directional, heteroscedastic, and conformalized variants that respect the unique characteristics of HEAs.

The field must therefore move beyond generic deep-ensemble variance. Future work should adopt HEA-tailored uncertainty frameworks, mandate calibration testing, and prioritize epistemic-aleatoric separation. Only by confronting these flawed assumptions can uncertainty quantification become a genuine accelerator rather than a hidden source of error in high-entropy alloy discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 29 Nov 2023

Revised: 05 Apr 2024

Accepted: 01 Jun 2024

Published online: 18 July 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Liu X, Zhang J, Pei Z. Machine learning for high-entropy alloys: Progress, challenges and opportunities. *Prog Mater Sci.* 2023;131:101018.
<https://doi.org/10.1016/j.pmatsci.2022.101018>.
- Jung SG, Jung G, Cole JM. Predictive modeling of high-entropy alloys and amorphous metallic alloys using machine learning. *J Chem Inf Model.* 2024;64(19):7313-36.
<https://doi.org/10.1021/acs.jcim.4c00873>.
- Kamnis S, Delibasis K. High entropy alloy property predictions using a transformer-based language model. *Sci Rep.* 2025;15(1):11861.
<https://doi.org/10.1038/s41598-025-95170-z>.
- Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems 30*. Red Hook (NY): Curran Associates; 2017. p. 6402-13.
- Novick A, Nguyen Q, Garnett R, Toberer E, Stevanović V. Simulating high-entropy alloys at finite temperatures: An uncertainty-based approach. *Phys Rev Mater.* 2023;7(6):063801.
<https://doi.org/10.1103/PhysRevMaterials.7.063801>.
- Allen M, Arroyave R, Malak R. Deep ensembles for modeling uncertain phase constraints in compositionally graded alloy design. In: *Proceedings of the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Volume 3A: 48th Design Automation Conference*. New York: American Society of Mechanical Engineers; 2022. p. V03AT03A028.
<https://doi.org/10.1115/DETC2022-89091>.
- Rao Z, Tung PY, Xie R, Wei Y, Zhang H, Ferrari A, et al. Machine learning-enabled high-entropy alloy discovery. *Science.* 2022;378(6615):78-85.
<https://doi.org/10.1126/science.abo4940>.

Ren F, Ward L, Williams T, Laws KJ, Wolverton C, Hattrick-Simpers J, et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci Adv.* 2018;4(4):eaq1566.
<https://doi.org/10.1126/sciadv.aaq1566>.

Deshmukh G, Wichrowski NJ, Evangelou N, Ghanekar PG, Deshpande S, Kevrekidis IG, et al. Active learning of ternary alloy structures and energies. *npj Comput Mater.* 2024;10(1):116.
<https://doi.org/10.1038/s41524-024-01256-z>.

Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: *Advances in Neural Information Processing Systems* 30. Red Hook (NY): Curran Associates; 2017. p. 5574-84.

Tyralis H, Papacharalampous G. A review of predictive uncertainty estimation with machine learning. *Artif Intell Rev.* 2024;57(4):94.
<https://doi.org/10.1007/s10462-023-10698-8>.

Hoffmann L, Fortmeier I, Elster C. Uncertainty quantification by ensemble learning for computational optical form measurements. *Mach Learn Sci Technol.* 2021;2(3):035030.
<https://doi.org/10.1088/2632-2153/ac0495>.

Singh S, Katiyar NK, Goel S, Joshi SN. Phase prediction and experimental realisation of a new high entropy alloy using machine learning. *Sci Rep.* 2023;13(1):4811.
<https://doi.org/10.1038/s41598-023-31461-7>.

Qi J, Hoyos DI, Poon SJ. Machine learning-based classification, interpretation, and prediction of high-entropy-alloy intermetallic phases. *High Entropy Alloys Mater.* 2023;1(2):312-26.
<https://doi.org/10.1007/s44210-023-00017-9>.

Yang C, Ren C, Jia Y, Wang G, Li M, Lu W. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. *Acta Mater.* 2022;222:117431.
<https://doi.org/10.1016/j.actamat.2021.117431>.

Palmer G, Du S, Politowicz A, Emory JP, Yang X, Gautam A, et al. Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Comput Mater.* 2022;8(1):115.
<https://doi.org/10.1038/s41524-022-00794-8>.

Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J Chem Inf Model.* 2020;60(6):2697-717.
<https://doi.org/10.1021/acs.jcim.9b00975>.

Caldeira J, Nord B. Deeply uncertain: Comparing methods of uncertainty quantification in deep learning algorithms. *Mach Learn Sci Technol.* 2021;2(1):015002.
<https://doi.org/10.1088/2632-2153/aba6f3>.

Chang YJ, Jui CY, Lee WJ, Yeh AC. Prediction of the composition and hardness of high-entropy alloys by machine learning: Chang, Jui, Lee, and Yeh. *JOM.* 2019;71(10):3433-42.
<https://doi.org/10.1007/s11837-019-03704-4>.

Liu J, Paisley J, Kioumourtzoglou MA, Coull B. Accurate uncertainty estimation and decomposition in ensemble learning. In: *Advances in Neural Information Processing Systems* 32. Red Hook (NY): Curran Associates; 2019. p. 8950-61.

Yang S, Yee K. Towards reliable uncertainty quantification via deep ensemble in multi-output regression task. *Eng Appl Artif Intell.* 2024;132:107871.
<https://doi.org/10.1016/j.engappai.2024.107871>.

Carrete J, Montes-Campos H, Wanzenböck R, Heid E, Madsen GKH. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning. *J Chem Phys.* 2023;158(20):204801.

<https://doi.org/10.1063/5.0146905>.

Zhou Y, Yang B. Uncertainty quantification of predicting stable structures for high-entropy alloys using Bayesian neural networks. *J Energy Chem.* 2023;81:118-24.

<https://doi.org/10.1016/j.jechem.2023.02.028>.

Wang X, Xiong W. Uncertainty quantification and composition optimization for alloy additive manufacturing through a CALPHAD-based ICME framework. *npj Comput Mater.* 2020;6(1):188.

<https://doi.org/10.1038/s41524-020-00454-9>.

Tran A, Tranchida J, Wildey T, Thompson AP. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. *J Chem Phys.* 2020;153(7):074705.

<https://doi.org/10.1063/5.0015672>.

Pathrudkar S, Thiagarajan P, Agarwal S, Banerjee AS, Ghosh S. Electronic structure prediction of multi-million atom systems through uncertainty quantification enabled transfer learning. *npj Comput Mater.* 2024;10(1):175.

<https://doi.org/10.1038/s41524-024-01305-7>.

Duignan TT. The potential of neural network potentials. *ACS Phys Chem Au.* 2024;4(3):232-41.

<https://doi.org/10.1021/acspchemau.4c00004>.