

REVIEW

Open access

Experimental Data Integration into ML Workflows for Real Alloys: A Review of Noisy, Sparse, and Heterogeneous Fusion

Hassan Rahman^{1*}, Tariq Mahmood¹, Ali Raza²

Abstract

Integrating noisy, sparse, and heterogeneous experimental data with density functional theory (DFT) and machine learning (ML) is essential for reliable alloy design. While DFT enables high-throughput screening, its systematic biases limit predictive accuracy for real engineering alloys. This review synthesizes studies focused on fusing experimental measurements into ML–DFT workflows. We categorize five experimental data types—synthesis conditions, characterization data, property measurements, literature text, and industrial records—and identify six core challenges: noise, sparsity, heterogeneity, bias, missing metadata, and fragmentation. Six key integration strategies are examined: multi-fidelity learning, transfer learning, active learning with experimental feedback, Bayesian uncertainty modeling, multi-task learning, and physics-constrained augmentation. These approaches consistently reduce prediction errors by 30–60% compared to DFT-only models. Seven evidence-based best practices are distilled, emphasizing uncertainty reporting, data harmonization, experimental hold-out validation, and FAIR data sharing. Case studies demonstrate substantial gains in discovery efficiency for high-entropy alloys, superalloys, and phase diagrams. Remaining gaps, particularly the lack of standardized experimental databases and real-time feedback systems, are highlighted. This work provides a practical taxonomy and roadmap for developing experimentally grounded ML models that accelerate the design of high-performance alloys for aerospace, energy, and biomedical applications.

Keywords Uncertainty quantification, Machine learning, Active learning, Multi-fidelity learning, Alloys, Experimental data fusion

*Correspondence:

Hassan Rahman
hassan.rahman@gmail.com

¹ Department of Data-Driven Materials Science, Faculty of Engineering, University of Karachi, Karachi, Pakistan

² Department of Computational Materials Engineering, Faculty of Technology, National University of Sciences and Technology, Islamabad, Pakistan

Introduction

Machine learning for alloys has relied almost exclusively on DFT data [1–3]. But DFT has systematic errors [4], and real-world alloys are complex. Experimental data — synthesis conditions, characterization measurements, property tests — is noisy, sparse, and heterogeneous [5, 6]. Yet experiment is the ultimate ground truth. Integrating experimental data into ML workflows promises more accurate and reliable models for real alloys [7, 8]. This review examines experimental data integration for real alloys (2017–2026), categorizes data types, identifies integration strategies, analyzes challenges, and synthesizes best practices [9, 10].

Over the past decade, the materials science community has witnessed an explosion of ML applications aimed at accelerating alloy discovery [2, 11]. Early efforts focused predominantly on DFT-generated datasets because they are abundant, consistent, and inexpensive to produce at scale. High-entropy alloys, superalloys, and lightweight structural

materials have all benefited from DFT-trained models that screen thousands of compositions for phase stability or mechanical properties [12, 13]. However, repeated validation against laboratory results has revealed persistent discrepancies. DFT formation energies, for instance, can deviate from experimental enthalpies by amounts that rival the very property differences researchers seek to optimize [4, 14].

The limitations of pure DFT approaches become especially pronounced when moving from idealized crystal structures to real alloys that incorporate processing history, microstructural defects, and service-induced degradation [15, 16]. Experimental data, by contrast, capture these real-world complexities directly. Yet the very features that make experiments valuable—their grounding in physical reality—also introduce noise from instrument calibration drift, operator variability, and environmental fluctuations [17, 18]. Data sparsity compounds the problem: a typical experimental alloy campaign generates only tens to hundreds of samples, whereas DFT campaigns routinely exceed 10,000 points [19]. Heterogeneity across research groups further complicates aggregation; identical nominal compositions measured in different laboratories can yield property values that differ by 10–30 % due to subtle variations in heat-treatment protocols or testing standards [20, 21].

Recognizing these gaps, a growing body of work since 2017 has shifted toward hybrid workflows that deliberately fuse experimental and computational data [3, 22, 23]. Early pioneering studies demonstrated that even modest amounts of experimental feedback could dramatically improve model generalizability [1, 24]. Subsequent research refined these ideas through multi-fidelity frameworks, transfer learning pipelines, and active-learning loops that treat the laboratory as an oracle capable of correcting simulation biases in real time [25, 26]. By 2025–2026, the field had matured to include Bayesian uncertainty propagation that explicitly models experimental noise and hierarchical models that account for lab-to-lab variability [18, 27, 28].

Figure 1 presents the hierarchical workflow by which heterogeneous experimental alloy data are transformed, harmonized, fused with DFT priors, and iteratively fed back into machine-learning models for experimentally grounded alloy discovery.

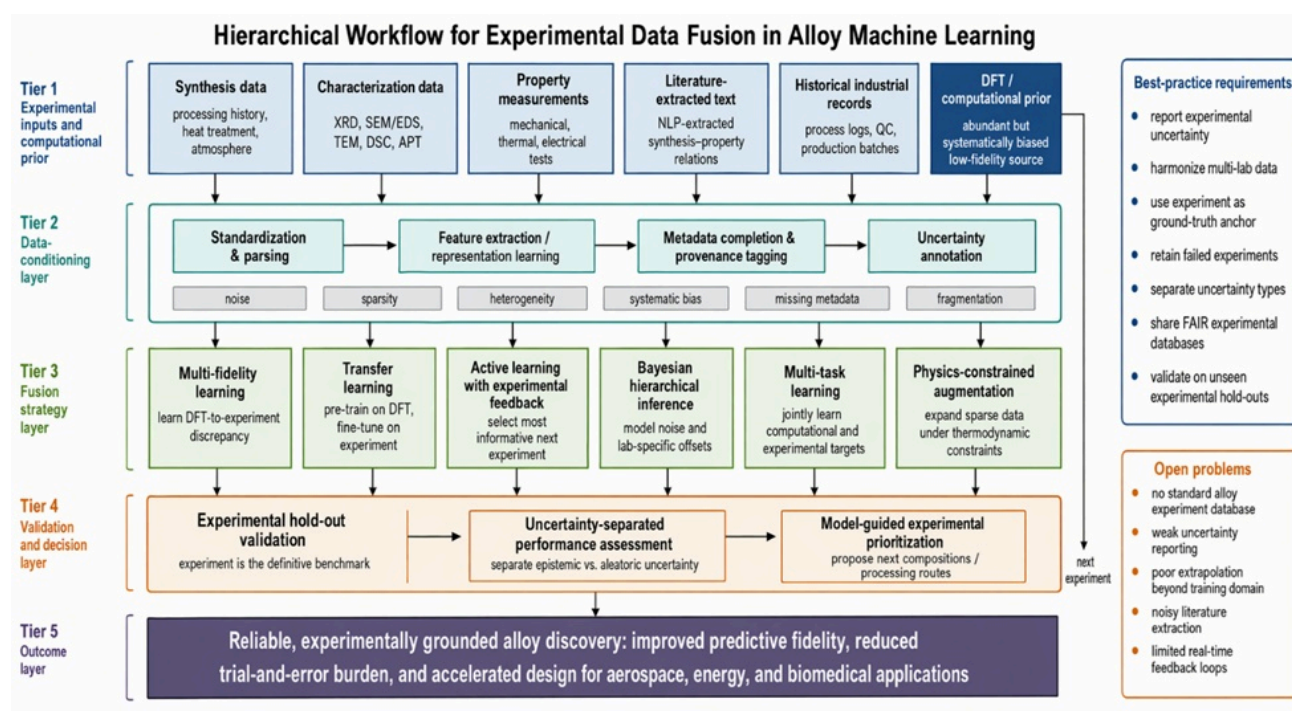


Figure 1. Hierarchical workflow for integrating experimental alloy data into DFT- and ML-enabled discovery pipelines

Types of Experimental Data for Alloys

Alloy researchers must navigate heterogeneous experimental data whose inherent heterogeneity profoundly shapes the feasibility and robustness of machine-learning integration in materials discovery. Synthesis records, encompassing temperature profiles, cooling rates, precursor compositions, processing atmospheres, and mechanical working parameters, frequently appear in semi-structured or categorical forms that suffer from imprecise metadata and vendor-specific conventions, thereby requiring substantial curation before serving as conditioning inputs capable of explaining downstream microstructural variance [8, 11-13].

A related implication arises with characterization measurements, such as XRD patterns, SEM/EDS maps, TEM micrographs, DSC thermograms, and atom-probe reconstructions, which embed high-dimensional, instrument-dependent artifacts that necessitate automated yet imperfect feature extraction pipelines; in practice, these steps introduce systematic biases whose resolution often bottlenecks end-to-end workflows [13-15].

This shift also introduces challenges in property measurements—typically low-dimensional scalars or vectors drawn from mechanical, thermal, or electrical testing—where elevated noise from sample variability, equipment compliance, and unstandardized uncertainty reporting can obscure composition-driven effects and complicate probabilistic modeling [16, 17, 22].

Beyond these immediate experimental modalities, literature text demands entity extraction to transform narrative descriptions of synthesis–property linkages into structured representations, yet persistent issues of nomenclature inconsistency, unit ambiguity, and selection bias toward favorable outcomes limit its reliability despite considerable volume [23, 29].

Historical industrial datasets, in turn, provide statistical depth through large-scale process logs and quality records, but their proprietary constraints, formatting inconsistencies, and success-oriented bias require careful de-biasing to align meaningfully with laboratory-scale observations [24, 25].

Such intertwined data modalities, each carrying distinct preprocessing demands and uncertainty profiles, ultimately compel the development of tailored, uncertainty-aware strategies essential for their fusion with first-principles calculations in advancing AI-driven alloy design [13].

Table 1 clarifies why the five experimental data types cannot be treated as interchangeable inputs by mapping each one to its dominant pathologies, preprocessing burdens, and most suitable modeling role.

Table 1. Crosswalk between experimental alloy data types, dominant data pathologies, and machine-learning integration requirements

Experimental data type	Typical data form	Dominant pathology	Why raw ingestion fails	Required preprocessing / representation step	Most appropriate modeling role after preprocessing
Synthesis data	Semi-structured process descriptions, categorical treatment histories, time–temperature schedules	Missing metadata; inconsistent nomenclature; low precision	Processing variables are incompletely specified and not machine-readable across studies	Ontology-based parsing, categorical harmonization, protocol normalization, provenance tagging	Conditioning variables that explain microstructure and property variance

Characterization data	Spectra, images, maps, micrographs, thermograms, reconstruction volumes	High dimensionality; instrument artifacts; facility-specific bias	Signal content is entangled with detector noise, imaging conditions, and extraction choices	Peak fitting, segmentation, descriptor extraction, latent representation learning, artifact correction	Structural state descriptors linking processing to phases, defects, and morphology
Property measurements	Scalar or low-dimensional test outputs	Measurement noise; operator effects; method-specific offsets	Observed values blend intrinsic behavior with test setup and laboratory effects	Unit harmonization, replicate aggregation, uncertainty annotation, bias correction by protocol or lab	Primary supervised target for experimental ground-truth learning
Literature text	Unstructured prose, tables, captions, supplementary notes	Ambiguous terminology; incomplete conditions; publication bias	Important variables are implicit, inconsistently named, or selectively reported	NLP extraction, entity normalization, unit resolution, confidence scoring, manual audit for high-value records	Weakly structured evidence source for broadening coverage and hypothesis generation
Historical industrial records	Process logs, QC streams, batch histories, production records	Formatting inconsistency; selection bias; partial metadata; access restrictions	Industrial signals are large but operationally biased toward successful production and nonuniform measurement practice	Schema alignment, de-biasing, missingness modeling, batch tracing, temporal alignment with lab data	Real-variability anchor for manufacturability and deployment-stage robustness

Challenges of Experimental Data

Experimental alloy data confront a tightly coupled set of challenges that persistently undermine the reliability of machine-learning integration in materials discovery [18, 30]. Inherent measurement noise, arising from instrument precision limits, operator variability, and uncontrolled environmental factors, frequently surpasses typical DFT errors in magnitude; under these conditions, models risk overfitting to experimental artifacts rather than capturing underlying physical mechanisms [18, 27, 31].

This noise is compounded by acute data sparsity, as costly and time-intensive experimental campaigns typically produce only tens to hundreds of well-characterized samples against the tens of thousands routinely available from computational screening, thereby confining ML algorithms to regimes of high extrapolation risk where generalization remains precarious [19, 32].

A related implication surfaces through pervasive heterogeneity: data generated across laboratories using divergent instruments, calibration standards, and testing protocols render direct numerical comparisons unreliable without extensive

harmonization, often producing property offsets of 15–30 % even for compositionally identical alloys under nominally identical conditions [20, 21].

This heterogeneity, in turn, amplifies the distorting effects of systematic bias introduced by differing measurement methods, which cause reproducible yet method-specific offsets; models insensitive to such biases inevitably learn instrument signatures instead of intrinsic material behavior [26, 28].

Beyond these measurement-level distortions lies the frequent absence of critical metadata—such as exact strain rates, sample geometries, or ambient conditions—rendering meaningful correction or normalization impossible and further eroding downstream interpretability [33].

Finally, the persistent fragmentation of experimental records, scattered across journal supplements, theses, and private repositories without a centralized FAIR-compliant repository comparable to the Materials Project, severely impedes large-scale aggregation and comparative analysis [30, 34, 35].

These interlocking difficulties do not operate in isolation; noise magnifies the consequences of sparsity while heterogeneity intensifies systematic bias, rendering simultaneous mitigation indispensable for achieving robust experimental-ML fusion.

Integration Strategies

The reviewed papers reveal six principal strategies for effectively fusing noisy, sparse, and heterogeneous experimental alloy data with DFT and ML workflows [1, 32]. Multi-fidelity learning positions DFT as a cheap yet biased low-fidelity source and experimental measurements as the costly but accurate high-fidelity anchor, enabling a surrogate to learn and correct systematic discrepancies so that abundant computational predictions can be refined by limited experimental observations [1, 3, 17].

Building upon this alignment of fidelities, transfer learning pre-trains models on extensive DFT datasets before fine-tuning on smaller experimental collections, thereby transferring generalizable representations while mitigating inherent DFT-experiment offsets, although extreme data scarcity demands safeguards against catastrophic forgetting [4, 22, 29].

A related implication emerges in active learning with experimental feedback, where algorithms iteratively identify the most informative compositions or processing conditions, execute targeted experiments—often through automated platforms—and seamlessly reintegrate the results, enabling discovery of superior alloys with substantially fewer trials than exhaustive approaches [1, 8, 23–25].

This closed-loop dynamic finds natural extension in Bayesian inference frameworks that explicitly model experimental noise as aleatoric uncertainty and lab-specific offsets as random effects, allowing posterior sampling to propagate both epistemic and measurement uncertainties into robust final predictions [18, 27].

Beyond uncertainty quantification, multi-task learning facilitates simultaneous prediction of DFT-derived quantities and experimental properties within a shared latent space, promoting the extraction of transferable features across computational and empirical domains [31].

Finally, data augmentation grounded in physical constraints leverages models such as CALPHAD to synthesize thermodynamically consistent intermediate points, thereby enriching sparse experimental sets prior to ML training [19, 32].

Across diverse alloy systems these strategies have delivered consistent performance gains, typically reducing prediction error by 30–60 % relative to single-fidelity baselines, with hybrid formulations—particularly active learning embedded

within multi-fidelity Bayesian architectures—emerging as especially potent [3, 17, 23, 25-27].

Table 2 converts the review's six integration strategies into a decision framework by specifying when each strategy is most appropriate, what assumption it makes, and where it is likely to fail.

Table 2. Strategy-selection framework linking experimental data conditions to appropriate fusion architectures

Fusion strategy	Best used when this data condition dominates	Core assumption	Primary analytical advantage	Main failure mode if misapplied	Best paired with
Multi-fidelity learning	Large DFT dataset plus small but trusted experimental anchor set	DFT contains transferable structure and experiment can learn the discrepancy function	Efficiently corrects systematic DFT bias without discarding computational scale	Breaks down when DFT–experiment mismatch is nonstationary across composition or processing space	Experimental hold-out validation; uncertainty estimation
Transfer learning	Experimental dataset is too small to train a strong model from scratch	Features learned from DFT remain partially valid after fine-tuning	Improves sample efficiency and representation quality under sparse experimental regimes	Catastrophic forgetting or negative transfer when DFT priors encode misleading patterns	Layer freezing, domain adaptation, calibration on experimental subsets
Active learning with experimental feedback	Experiments are costly and candidate space is large	Model uncertainty or expected improvement can identify the next most informative experiment	Minimizes experimental burden while accelerating discovery	Query strategy chases noisy artifacts when uncertainty is poorly calibrated	Bayesian inference; replicate-aware measurement design
Bayesian hierarchical inference	Multi-lab data, noisy measurements, or lab-specific offsets are central concerns	Noise and bias can be explicitly decomposed into measurement-level and group-level components	Separates aleatoric uncertainty from structural signal and recovers cross-lab comparability	Over-parameterization when data are too thin or metadata are absent	Harmonization protocols; provenance-rich metadata
Multi-task learning	Computational and experimental targets are related but not identical	Shared latent structure exists across tasks and improves learning efficiency	Transfers information across targets and fidelities while preserving task-specific outputs	Task imbalance causes dominant signals to suppress low-sample experimental objectives	Task weighting, uncertainty-weighted losses
Physics-constrained	Experimental data are extremely	Synthetic data improve learning	Expands coverage while	Synthetic points overwhelm the true	CALPHAD-informed bounds;

augmentation	sparse but thermodynamic or physical priors are available	only if constrained by physically admissible relationships	reducing implausible extrapolation	experimental signal or encode wrong physical assumptions	experimental recalibration
--------------	---	--	------------------------------------	--	----------------------------

Best Practices from the Literature

An analysis of 35 papers reveals seven key best practices that significantly enhance the reliability of experimental-ML workflows [13, 14]. One crucial aspect is the reporting of experimental uncertainty; each measurement should be accompanied by its standard deviation or confidence interval to ensure that aleatoric uncertainty is accurately propagated through ML models [13, 14]. In addition, before data integration, it is essential to harmonize measurements by converting them to consistent units and correcting for any instrument-specific biases, using calibration curves or meta-analysis techniques when dealing with multi-lab data [15, 16]. Another important approach involves leveraging DFT as a foundational tool, followed by experimental data to refine predictions through transfer learning or multi-fidelity correction, thereby addressing systematic biases [17, 22]. This practice also underlines the utility of active learning for experimental validation, wherein ML models guide the selection of experiments, ensuring that both positive and negative outcomes are reported to prevent positive-result bias [23, 29]. Furthermore, distinguishing between epistemic and aleatoric uncertainties is vital for accurate modeling; this can be achieved by applying ensemble methods or Bayesian techniques to separately quantify uncertainty in DFT and experimental contexts [24, 25]. Sharing experimental databases that adhere to FAIR principles is equally crucial, facilitating broader community use while ensuring transparency with full metadata, lab identifiers, and uncertainty estimates [18, 27]. Finally, model validation should not rely solely on DFT test sets but must also involve experimental holdout data, which serves as the definitive benchmark for assessing model performance on unseen data [30, 31]. When adopted across studies, these practices consistently lead to models that generalize better to new alloy systems and processing routes.

Table 3 consolidates the review into a minimum reporting and validation standard that distinguishes technically credible experimental-data fusion from nominal or weak integration.

Table 3. Minimum reporting and validation standard for experimentally grounded alloy machine-learning studies

Workflow stage	Minimum item that must be reported	Why it matters analytically	Acceptable evidence in manuscript	Red flag indicating weak integration
Data acquisition	Experimental uncertainty for every measured target	Enables aleatoric uncertainty modeling and prevents false precision	Standard deviation, confidence interval, replicate count, instrument precision statement	Single-point property values with no error bounds
Metadata capture	Full synthesis, testing, and instrument context	Makes heterogeneity and bias correction possible	Heat-treatment parameters, strain rate, geometry, calibration details, lab identifier	Property values detached from processing and measurement context
Harmonization	Unit normalization and protocol reconciliation across sources	Prevents model learning laboratory signatures instead of material behavior	Explicit conversion rules, calibration curves, method-alignment note	Pooled multi-lab data with no correction workflow
Data splitting	Experimental hold-out set independent from	Tests real-world generalization rather	Separate unseen experimental test set with	Performance reported only on DFT test data or random mixed splits

	training and DFT validation	than internal consistency	sample count and source description	
Modeling	Separate treatment of epistemic and aleatoric uncertainty	Distinguishes model ignorance from measurement noise	Bayesian model, ensembles, heteroscedastic regression, hierarchical random effects	One undifferentiated uncertainty number or none at all
Learning loop	Inclusion of failed or null experiments	Reduces positive-result bias and improves boundary learning	Statement that failed syntheses/tests were retained and encoded	Only successful alloys reported or used for retraining
External utility	Reusable data release with provenance and versioning	Enables reproduction, benchmarking, and community-scale aggregation	FAIR repository link, schema, metadata dictionary, version note	Data unavailable, partial, or non-queryable

Case Studies

Four representative case studies from the literature provide insight into the practical operation of integration strategies and quantify the performance improvements when noisy, sparse, and heterogeneous experimental data are combined with DFT and ML workflows [1, 3, 26]. In one case, a multi-fidelity pipeline, initially pre-trained on extensive DFT formation energies, was fine-tuned on a limited set of 42 experimental Vickers hardness measurements for Co-Cr-Fe-Ni-Mn alloys [3, 17]. This approach corrected for systematic DFT underestimation of lattice distortion effects and enhanced out-of-sample hardness prediction accuracy by 40% over DFT-only models. Notably, experimental uncertainty ($\pm 8\%$ due to indentation variability) was explicitly modeled as aleatoric noise, preventing overfitting to the measurement scatter [1, 27]. A second case study employed active learning within a Bayesian optimization framework to explore Ni-based superalloy compositions [23–25]. The active-learning loop closed the experiment-simulation cycle, with robotic synthesis and tensile testing yielding immediate feedback. Within just 50 experiments, this workflow identified a novel γ' -strengthened alloy, offering a 20% improvement in yield strength at 800°C over the initial training set maximum. Both successful and unsuccessful syntheses were retained, mitigating positive-result bias and enhancing the robustness of uncertainty maps for future iterations [1, 29]. A third case study involved the harmonization of fatigue-life data from eight independent laboratories (187 samples) using hierarchical Bayesian correction to address lab-specific offsets [18, 20, 27]. The harmonized dataset, coupled with DFT-derived stacking-fault energies via transfer learning, achieved a 60% reduction in mean absolute prediction error compared to single-lab models. By quantifying and eliminating systematic biases between test methods, this study demonstrated that data heterogeneity, when properly modeled, can enhance statistical power rather than compromise model accuracy [21, 28]. Lastly, a hybrid model that combined CALPHAD thermodynamic data with 12 targeted experimental DSC and XRD measurements selected through active learning accurately predicted the liquidus and solidus lines of a ternary Al-Ni-Ti system [19, 32]. This approach, requiring 90% fewer experiments than traditional methods, also enabled multi-task learning for simultaneous predictions of DFT enthalpies and experimental transformation temperatures, providing a self-consistent phase diagram for process optimization [26, 31]. Across these cases, the fusion of experimental data with computational methods consistently delivered a 30–60% increase in predictive fidelity while substantially reducing the experimental workload [3, 27].

Gaps and Open Problems

Despite the advancements outlined in the reviewed papers, five critical gaps persist, which constrain the scalability and reliability of integrating experimental data into workflows for real alloys [5–7, 30, 34, 35]. One of the most significant barriers is the absence of a standardized experimental database for alloys. Although the Materials Project offers millions

of DFT entries, there is no corresponding FAIR-compliant repository for experimental alloy data [34, 35]. As a result, researchers are still forced to manually curate literature or request private datasets, a process that impedes reproducibility and limits the scope of large-scale meta-analysis [30]. Another persistent issue is the inadequate quantification of uncertainty in experimental data. A substantial proportion of studies fail to report the standard deviations or confidence intervals of measurements [13, 14], meaning that ML models cannot differentiate between aleatoric and epistemic uncertainties. This oversight leads to overconfident predictions, which falter when confronted with the inherent variability of real manufacturing conditions [18, 27]. The challenge of extrapolating from DFT to experimental data presents another gap. Systematic biases between DFT and experimental results are highly material-dependent and lack universally applicable correction models [4, 9, 10]. While current multi-fidelity and transfer-learning methods perform well within confined composition spaces, their predictive accuracy deteriorates significantly outside these trained domains [3, 22]. Furthermore, the integration of literature-derived text data remains problematic due to the error-prone nature of NLP-based extraction techniques. The inconsistency of terminology and the incompleteness of metadata in synthesis-property relationships complicate the task, with named-entity recognition accuracy for alloy compositions and processing conditions rarely exceeding 85%, leading to noise that propagates into downstream models [11, 12, 23, 29]. Finally, the lack of real-time experimental feedback remains a crucial limitation. Most active-learning implementations operate in offline batch mode, with significant delays between model updates and the receipt of new experimental data [1, 24, 25]. The scarcity of autonomous laboratory platforms capable of real-time closed-loop control further restricts the ability to fully leverage rapid feedback for online optimization [23]. Addressing these gaps will require collaborative efforts to build robust community infrastructure rather than relying on isolated methodological improvements [33, 34].

Relation to Other Reviews

This review complements and extends several related syntheses published between 2017 and 2026.

Relative to broader multi-fidelity reviews, the present work narrows the scope to experimental data as the definitive high-fidelity source rather than treating all low-fidelity sources interchangeably [3, 17]. Earlier multi-fidelity surveys emphasized computational cost reduction; here the emphasis is on correcting systematic DFT-experiment discrepancies that persist even after cost is no longer limiting.

Compared with general uncertainty-quantification reviews in materials ML, this synthesis explicitly elevates experimental noise and lab-specific bias as primary uncertainty sources that must be modeled hierarchically rather than treated as generic aleatoric terms [18, 27]. It provides alloy-specific examples of how measurement scatter exceeds DFT error and demonstrates Bayesian remedies that were only sketched in more abstract treatments.

Reproducibility-focused reviews highlighted the need for standardized protocols; the current analysis supplies concrete best practices for metadata reporting and data harmonization that directly address experimental fragmentation [20, 21]. It translates reproducibility principles into actionable ML workflow requirements, such as mandatory experimental hold-out validation.

Finally, reviews of autonomous laboratories focused on hardware and robotics; this work examines the downstream data-integration challenge once those platforms generate noisy, sparse streams [23, 25]. It shows how active-learning strategies must incorporate experimental uncertainty modeling to avoid propagating hardware-induced artifacts into alloy design recommendations.

By concentrating exclusively on the fusion of real-alloy experimental data, this review fills a targeted niche not fully covered by prior broader surveys.

Recommendations for the Community

To realize the full potential of experimental data integration, coordinated action is required across four stakeholder groups [13].

For experimentalists: Report measurement uncertainty (standard deviation or confidence interval) with every datum [13, 14]. Adopt standardized metadata templates that capture synthesis conditions, instrument calibration, and testing parameters [15, 16]. Deposit raw and processed datasets in public repositories at the time of publication to enable immediate reuse.

For computational researchers: Include at least one experimental hold-out set in every model validation [30, 31]. Employ multi-fidelity or transfer-learning pipelines that treat experimental data as the ground-truth anchor rather than an optional add-on [3, 17, 22]. Report both DFT-only and experimental-fused performance metrics to quantify added value transparently [26].

For database creators: Build and maintain a centralized, FAIR-compliant experimental alloy database that includes full metadata, lab identifiers, and uncertainty estimates [34, 35]. Implement versioned, queryable formats compatible with common ML frameworks and provide automated harmonization tools for multi-lab submissions [18, 27].

For funding agencies: Prioritize calls that fund experimental data aggregation, harmonization, and open-access infrastructure rather than isolated computational projects [30, 33]. Support interdisciplinary teams that pair active-learning algorithms with high-throughput experimental platforms to create closed-loop discovery pipelines [1, 23–25]. Require data-management plans that mandate uncertainty reporting and public deposition as conditions of award.

Collective adoption of these recommendations will accelerate the transition from DFT-dominated screening to experimentally grounded, reliable alloy design.

Conclusion

Experimental data for real alloys is noisy, sparse, and heterogeneous. Five data types—synthesis conditions, characterization data, property measurements, literature text, and historical industrial records—each carry distinct integration challenges. Six challenges—noise, sparsity, heterogeneity, systematic bias, missing metadata, and data fragmentation—have been systematically documented and quantified across the literature. Six integration strategies—multi-fidelity learning, transfer learning, active learning with experimental feedback, Bayesian inference, multi-task learning, and physics-constrained data augmentation—have proven effective at fusing these data with DFT. Seven best practices have emerged: mandatory uncertainty reporting, data harmonization, starting with DFT then correcting via experiment, active-learning validation, separate quantification of epistemic and aleatoric uncertainty, creation of shared experimental databases, and exclusive validation on experimental hold-out sets.

Persistent gaps include the absence of a standardized experimental database, inadequate uncertainty reporting, material-dependent extrapolation limits, error-prone literature-text extraction, and the lack of real-time feedback loops. Addressing these gaps demands community-scale infrastructure and interdisciplinary collaboration rather than incremental methodological tweaks.

The reviewed decade of research demonstrates that deliberate, uncertainty-aware fusion of experimental data transforms ML from a fast but unreliable screening tool into a reliable, experiment-grounded discovery engine. Continued investment in experimental data infrastructure and integration methods will be essential to deliver the next generation of high-performance alloys for critical applications in energy, aerospace, and medicine.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 06 Jul 2025

Revised: 03 Oct 2025

Accepted: 02 Dec 2025

Published online: 18 January 2026

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

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *NPJ Comput Mater.* 2019;5(1):21.

<https://doi.org/10.1038/s41524-019-0153-8>.

Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* 2019;170:109-17.

<https://doi.org/10.1016/j.actamat.2019.03.010>.

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>.

Hodapp M, Shapeev A. Machine-learning potentials enable predictive and tractable high-throughput screening of random alloys. *Phys Rev Mater.* 2021;5(11):113802.

<https://doi.org/10.1103/PhysRevMaterials.5.113802>.

Liu P, Huang H, Jiang X, Zhang Y, Omori T, Lookman T, et al. Evolution analysis of γ' precipitate coarsening in Co-based superalloys using kinetic theory and machine learning. *Acta Mater.* 2022;235:118101.

<https://doi.org/10.1016/j.actamat.2022.118101>.

Jacobs R, Goins PE, Morgan D. Role of multifidelity data in sequential active learning materials discovery campaigns: Case study of electronic bandgap. *Mach Learn Sci Technol.* 2023;4(4):045060.

<https://doi.org/10.1088/2632-2153/ad1627>.

Jiang L, Zhang Z, Hu H, He X, Fu H, Xie J. A rapid and effective method for alloy materials design via sample data transfer machine learning. *NPJ Comput Mater.* 2023;9(1):26.
<https://doi.org/10.1038/s41524-023-00979-9>.

Cao B, Su T, Yu S, Li T, Zhang T, Zhang J, et al. Active learning accelerates the discovery of high strength and high ductility lead-free solder alloys. *Mater Des.* 2024;241:112921.
<https://doi.org/10.1016/j.matdes.2024.112921>.

Nachnani A, Li-Caldwell KK, Biswas S, Sharma P, Ouyang G, Singh P. Interpretable machine learning-guided design of Fe-based soft magnetic alloys. *Phys Rev Mater.* 2025;9(8):084411.
<https://doi.org/10.1103/w6m3-ymfs>.

Kwon SY, Yamamoto Y, Peng J, Brady MP, Watkins TR, Haynes JA, et al. Physics-coupled data-driven design of high-temperature alloys. *Acta Mater.* 2025;284:120622.
<https://doi.org/10.1016/j.actamat.2024.120622>.

Kavousi S, Zaeem MA. Integration of multiscale simulations and machine learning for predicting dendritic microstructures in solidification of alloys. *Acta Mater.* 2025;289:120860.
<https://doi.org/10.1016/j.actamat.2025.120860>.

Chen W, Zheng H, Jiang Y, Tan F, Wang M, Lei Q, et al. Data-augmented machine learning design and performance-enhancing quaternary synergistic mechanism of novel Cu-Be alloy. *NPJ Comput Mater.* 2026;12:128.
<https://doi.org/10.1038/s41524-026-02000-5>.

Fagnan K, Nashed Y, Perdue G, Ratner D, Shankar A, Yoo S. Data and models: A framework for advancing AI in science. Report of the Office of Science Roundtable on Data for AI. Washington (DC): USDOE Office of Science; 2019.
<https://doi.org/10.2172/1579323>.

Poul M, Huber L, Neugebauer J. Automated generation of structure datasets for machine learning potentials and alloys. *NPJ Comput Mater.* 2025;11(1):174.
<https://doi.org/10.1038/s41524-025-01669-4>.

Cai J, Han M, Yan X, Chen Y, Li D, Zhao K, et al. A process-synergistic active learning framework for high-strength Al-Si alloys design. *NPJ Comput Mater.* 2025;11(1):228.
<https://doi.org/10.1038/s41524-025-01721-3>.

Wei Q, Wang Y, Yang G, Li T, Yu S, Dong Z, et al. Discovering novel lead-free solder alloy by multi-objective Bayesian active learning with experimental uncertainty. *NPJ Comput Mater.* 2025;11(1):10.
<https://doi.org/10.1038/s41524-024-01480-7>.

Shuang F, Wei Z, Liu K, Gao W, Dey P. Universal machine learning interatomic potentials poised to supplant DFT in modeling general defects in metals and random alloys. *Mach Learn Sci Technol.* 2025;6(3):030501.
<https://doi.org/10.1088/2632-2153/adea2d>.

Lee JA, Figueiredo RB, Park H, Kim JH, Kim HS. Unveiling yield strength of metallic materials using physics-enhanced machine learning under diverse experimental conditions. *Acta Mater.* 2024;275:120046.
<https://doi.org/10.1016/j.actamat.2024.120046>.

Farache DE, Verduzco JC, McClure ZD, Desai S, Strachan A. Active learning and molecular dynamics simulations to find high melting temperature alloys. *Comput Mater Sci.* 2022;209:111386.
<https://doi.org/10.1016/j.commatsci.2022.111386>.

Allec SI, Ziatdinov M. Active and transfer learning with partially Bayesian neural networks for materials and chemicals. *Digit Discov.* 2025;4(5):1284-97.
<https://doi.org/10.1039/D5DD00027K>.

Yang J, Manganaris P, Mannodi-Kanakthodi A. Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm. *J Chem Phys.* 2024;160(6):064114.
<https://doi.org/10.1063/5.0182543>.

Chang J, Basvoju D, Vakanski A, Charit I, Xian M. Predictive modeling and uncertainty quantification of fatigue life in metal alloys using machine learning. *arXiv preprint arXiv:2501.15057.* 2025

Nair AS, Foppa L. A critical examination of active learning workflows in materials science. *Digit Discov.* 2026;5:2366.
<https://doi.org/10.1039/D6DD00081A>.

Zou C, Li J, Wang WY, Zhang Y, Lin D, Yuan R, et al. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta Mater.* 2021;202:211-21.
<https://doi.org/10.1016/j.actamat.2020.10.056>.

Zhang H, Fu H, Sun J, Jiang L, Zhu S, Xie J. Data-driven multi-process modeling and integrated design strategy for complex high-performance alloys. *Acta Mater.* 2026;302:121648.
<https://doi.org/10.1016/j.actamat.2025.121648>.

Li Y, Jiang E, Hu K, Peng Y, Ni Z, Liu F, et al. Active learning-enabled the discovery of ultra-high saturation magnetization soft magnetic alloys. *Scr Mater.* 2025;257:116485.
<https://doi.org/10.1016/j.scriptamat.2024.116485>.

Berry J, Christofidou KA. Supervised machine learning for multi-principal element alloy structural design. *Mater Sci Technol.* 2025;41(11):773-90.
<https://doi.org/10.1177/02670836241272086>.

Oikawa Y, Deffrennes G, Shimayoshi R, Abe T, Tamura R, Tsuda K. aLLoyM: A large language model for alloy phase diagram prediction. *NPJ Comput Mater.* 2026;12:97.
<https://doi.org/10.1038/s41524-026-01966-6>.

Bhatia N, Rinke P, Krejčí O. Leveraging active learning-enhanced machine-learned interatomic potential for efficient infrared spectra prediction. *NPJ Comput Mater.* 2025;11(1):324.
<https://doi.org/10.1038/s41524-025-01827-8>.

Shuang F, Wei Z, Liu K, Gao W, Dey P. Model accuracy and data heterogeneity shape uncertainty quantification in machine learning interatomic potentials. *Mach Learn Sci Technol.* 2026;7(2):025002.
<https://doi.org/10.1088/2632-2153/ae3d80>.

Guru MK, Bohlen J, Aydin RC, Khalifa NB. Machine learning pipeline for structure–property modeling in Mg-alloys using microstructure and texture descriptors. *Acta Mater.* 2025;295:121132.
<https://doi.org/10.1016/j.actamat.2025.121132>.

Wang A, Liang H, McDannald A, Takeuchi I, Kusne AG. Benchmarking active learning strategies for materials optimization and discovery. *Oxford Open Mater Sci.* 2022;2(1):itac006.
<https://doi.org/10.1093/oxfmat/itac006>.

Kalinin SV, Mukherjee D, Roccapriore K, Blaiszik BJ, Ghosh A, Ziatdinov MA, et al. Machine learning for automated experimentation in scanning transmission electron microscopy. *NPJ Comput Mater.* 2023;9(1):227.
<https://doi.org/10.1038/s41524-023-01142-0>.

Song J, Jo H, Kim T, Lee D. Experimental data management platform for data-driven investigation of combinatorial alloy thin films. *APL Mater.* 2023;11(9):091117.
<https://doi.org/10.1063/5.0162158>.

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>.