*ChemPhysChem*

Chemistry Europe
European Chemical Societies Publishing
ChemPhysChem

RESEARCH ARTICLE

# D2VP: A Visual First Platform for Diagnosing Data Distribution Mismatch Toward Reliable Machine Learning in Materials Science

Xuqiang Shao[1] | Zhaoyan Dong[1] | Piao Ma[2,3] | Limin Li[2,3] | Tongao Yao[4,5] | Yujie Yan[4,5] | Mingzhe Li[4,5] | Zhengyang Gao[4,5] | Weijie Yang[4,5]

[1]Department of Computer Science, North China Electric Power University, Baoding, China | [2]Gusu Laboratory of Materials, Suzhou, China | [3]Suzhou MatSource Technology Co. Ltd., Suzhou, China | [4]Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding, China | [5]Department of Power Engineering, North China Electric Power University, Baoding, China

**Correspondence:** Weijie Yang (yangwj@ncepu.edu.cn)

**Received:** 27 September 2025 | **Revised:** 24 March 2026 | **Accepted:** 26 April 2026

**Keywords:** data distribution mismatch | data–centric I | machine learning reliability | materials informatics | visual analytics

## ABSTRACT

The reliability of machine learning (ML) models in chemistry and materials discovery is often compromised by Data Distribution Mismatch (DDM), a systematic deviation between training data distributions and target chemical spaces manifesting in three dimensions: over-density (redundant sampling of known regions), sparsity (insufficient coverage of the feature space), and extrapolation risk (predictions beyond the training domain). To address this challenge, we propose a data-centric, visual-first methodology that transforms reliability assessment from abstract metrics into evidence-based visual inquiry, operationalized through D2VP (Data Diagnosis & Visualization Platform), an open-source interactive system where "2" denotes the dual capability of diagnosing distribution anomalies and visualizing high-dimensional chemical spaces in a human-in–the-loop environment. Case studies validate the methodology across each DDM dimension. Redundancy-reduced sampling reduces dataset size by over 60% with negligible accuracy loss. Exploration-based sampling guides new experiments; adding just 10% of targeted data reduces mean absolute error by over 80% for models trained on sparse data. The credibility analysis toolkit identifies high-risk extrapolation scenarios, including a case with a training-to-prediction ratio of nearly 1:400. By integrating these capabilities, D2VP transforms abstract reliability metrics into actionable scientific insights, providing a domain-agnostic framework for building trustworthy predictive models across the AI for science landscape.

## 1 | Introduction

Machine learning (ML) has emerged as a transformative paradigm in chemical research, joining experiment, theory, and computation as a foundational pillar of modern science. By enabling the rapid prediction of material properties [1–3], inverse molecular design [4–6], and automated synthesis planning [7–9], these data-driven approaches are fundamentally reshaping the research landscape. Yet, as the field transitions from proof-of-concept to practical application, the focus must shift from merely demonstrating predictive potential to rigorously ensuring the operational reliability of these models in high-stakes scientific discovery.

The reliability of an ML prediction hinges on two interconnected factors: the quality of the model architecture and, more fundamentally, the integrity of its training data. While algorithmic advances continue to dominate the literature [10–15], a growing body of evidence suggests that data quality issues which we term Data Distribution Mismatch (DDM), are often the root cause of unreliable predictions in real-world applications.

However, for materials scientists, systematically improving the quality and integrity of data often presents a more direct path to reliable models than pursuing complex algorithmic development. This viewpoint is the foundation of data-centric AI, a

paradigm that prioritizes systematic data improvement over perpetual model tuning [16, 17]. Supporting this perspective, evidence indicates that for a given prediction, the similarity of the query point to the training distribution is a more critical determinant of accuracy than the specific algorithm employed [18]. Grounded in this data-centric philosophy, this work introduces a systematic methodology to diagnose and mitigate DDM.

We formally define Data Distribution Mismatch (DDM) as a systematic deviation between training data distributions and target chemical spaces [19, 20], manifesting across three distinct dimensions. The first dimension, over-density, refers to redundant sampling of known regions, characteristic of the long–tailed distributions common in materials research [21–24] where human-centric research biases concentrate efforts on materials similar to known high-performers [25, 26]. Studies show that prediction errors for materials in sparse data regions can be over 50% higher than for well-represented regions [27]. The second dimension, sparsity, describes insufficient coverage of the feature space, creating cognitive blind spots where predictive accuracy can plummet from over 90% to below 60% [20], rendering predictions unreliable [28]. The third dimension, extrapolation risk, manifests during application, where models trained on limited experimental data must screen vast virtual libraries, often forcing training-to-prediction ratios exceeding 1:400.

This extrapolation dimension of DDM is formally linked to the out-of-distribution (OOD) generalization problem and the challenge of defining a model's applicability domain (AD) [29–33]. Recent work reveals that performance can degrade significantly when models encounter novel inputs. Increasing training data volume alone does not reliably resolve this problem [28, 34–36]. Collectively, these three manifestations of DDM constitute a primary barrier to achieving robust and interpretable machine learning in scientific research [28, 37].

Despite this critical barrier, a significant methodological gap persists. While platforms such as the Materials Project [38], NOMAD [39], and Materials Cloud [40] have revolutionized access to computational materials data, their primary focus remains on data aggregation and standardization rather than systematic DDM diagnosis and mitigation. Current ML toolchains often provide fragmented ecosystems of disconnected components [28, 41], leading to rigid workflows [42] ill-equipped to address DDM consequences. Furthermore, many existing methods demand deep ML expertise, creating adoption barriers for domain experts [8, 19]. A deficit exists in intuitive, human-in-the-loop systems that empower scientists to directly interact with and steer the data analysis process [43, 44].

In response, we developed a visual-first, data-centric methodology operationalized through **D2VP** (**D**ata **D**iagnosis & **V**isualization **P**latform), an open-source interactive platform designed to fill this methodological void. The D2VP platform provides integrated tools to diagnose and mitigate each DDM dimension within an intuitive visual environment. We validate this methodology through case studies demonstrating: (i) redundancy-reduced sampling to address over-density, reducing training sets by over 60% with negligible accuracy loss; (ii) exploration-based sampling to resolve sparsity, improving model performance by over an order of magnitude with targeted data augmentation; and (iii) credibility analysis to manage extrapolation risk, identifying high-risk predictions in scenarios with training-to-prediction ratios approaching 1:400.

## 2 | Results and Discussion

### 2.1 | Visualizing Feature Space for Enhanced Reliability: The D2VP Approach

The primary contribution of this work is not a novel predictive algorithm. Rather, we present D2VP, an interactive visual platform that redirects attention from algorithmic prediction to researcher empowerment. The platform equips domain experts with tools for direct data interrogation and analytical reasoning. Within this environment, researchers can intuitively diagnose data quality issues and systematically address foundational problems that compromise model reliability. Figure 1 illustrates this paradigm shift: from a conventional "black-box" approach relying on abstract metrics (Figure 1a) to a human-in-the-loop methodology grounded in visual, evidence-based scientific inquiry (Figure 1b).

The D2VP platform operationalizes this paradigm shift through three integrated functional modules (Figure 1, lower panel). The Core Analytics Engine forms the foundation, translating high-dimensional feature spaces into intuitive, low-dimensional projected representations via three user-selectable methods: PCA for global linear structure, t-SNE for local cluster resolution, and UMAP for balanced global-local topology. Built upon this engine, the Training Set Optimization Workflow addresses two complementary problems. It first identifies a representative core subset from over-dense data (redundancy-reduced sampling), then augments sparse regions with targeted new candidates (exploration-based sampling). The Reliability Analysis Workflow operates in parallel, projecting both training and prediction sets into a shared visual space to generate comparative distribution plots that expose extrapolation risk. Together, these modules provide integrated analytical functions to address each DDM dimension: (i) over-density, (ii) sparsity, and (iii) extrapolation risk. The subsequent sections validate each function through targeted case studies.

### 2.2 | D2VP Methodology I: Enhancing Training Efficiency via Redundancy-Reduced Sampling

Developing ML models with large datasets presents a fundamental trade-off: computational efficiency versus predictive accuracy. A complete dataset may maximize accuracy, but at prohibitive training cost; too little data accelerates training but degrades performance. Redundancy-reduced sampling addresses this dilemma by targeting the over-density dimension of DDM. As illustrated in Figure 2a, D2VP provides visual controls that enable researchers to identify a representative core subset that maintains high accuracy while reducing training overhead. We validated this capability across three benchmark datasets spanning 900-65,000 samples, 16-131 feature dimensions, and domains including formation energy, aqueous solubility, and hydrogen absorption.

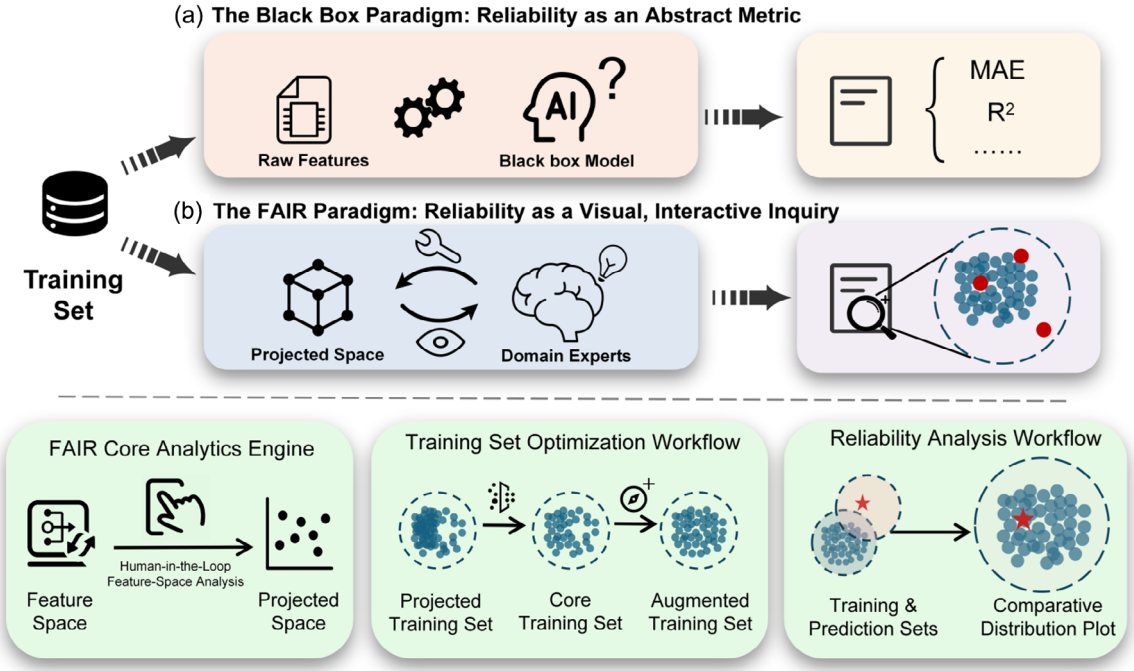


**FIGURE 1** | Opaque automation versus interactive inquiry. Paradigm shift from black-box to visual-first methodology and the D2VP functional architecture. (a) The conventional approach processes raw features through an opaque model, yielding abstract metrics (e.g., MAE, $R^2$) with limited actionable insight. (b) The D2VP methodology projects high-dimensional feature spaces into an intuitive visual representation, enabling human-in-the-loop exploration with transparent, evidence-based insights. The lower panel details the three core functional modules: the Core Analytics Engine translates raw feature spaces into interactive projected representations; the Training Set Optimization Workflow distills a representative core subset and augments it with targeted new data; the Reliability Analysis Workflow generates comparative distribution plots to assess prediction credibility.

In novel materials exploration, high-throughput computations often generate feature spaces congested with redundant candidates (Figure 2b). This over-density creates a computational bottleneck. Our approach addresses this by distilling essential chemical diversity from the noise. Instead of training on exhaustive datasets, the method identifies a structural core. This allows researchers to construct robust models in hours rather than days, accelerating design iteration cycles.

In drug discovery, aqueous solubility prediction acts as a primary filter for millions of candidates (Figure 2c). Experimental assessment of every molecule is infeasible. Here, the sampling strategy functions as a structural sieve. By isolating a representative core subset, the method ensures that limited resources are allocated only to unique structural motifs. This prevents the waste of experimental capacity on chemically redundant compounds.

For hydrogen storage alloys, data scarcity accentuates the risk of overfitting to noise (Figure 2d). In such constrained fields, redundant data points can artificially inflate model confidence without adding informational value. By enforcing a minimally redundant support set, our approach forces the model to learn from distinct physical variations rather than repetitive signal. This results in models that are arguably more generalizable despite being trained on fewer samples.

We applied this sampling function to generate subsets ranging from 10% to 100% of the original data across all three benchmarks. We then evaluated the trade-off between training time and Test Mean Absolute Error (MAE). The quantitative trends are detailed in Figure 2b–d.

Analysis reveals a consistent trend across all datasets. Training time scales linearly with sample size, whereas Test MAE charts a trajectory of diminishing returns. Error rates drop precipitously in the initial sampling phase, plateauing near 40% data usage. Beyond this threshold, additional data yields negligible accuracy gains. Consider formation energy (Figure 2b): tripling data from 40% to 100% reduces MAE by only 0.015 eV/atom. The overlaid scatter plots elucidate this phenomenon: the core topological structure of the feature space is effectively captured even at low sampling densities.

These results validate redundancy-reduced sampling as a robust solution for the "over-density" dimension of DDM. In large-scale contexts (Figure 2b,c) [45, 46], the function accelerates prototyping by eliminating redundant computation. Crucially, this principle extends to data-scarce regimes (Figure 2d) [47], where defining a representative core is essential to avoid overfitting. By systematically resolving the efficiency-accuracy trade-off, the methodology replaces ad-hoc data selection with a rigorous, topology-based standard for training set construction.

### 2.3 | D2VP Methodology II: Accelerating Discovery With Exploration-Based Sampling

Beyond addressing over-density, our methodology integrates a proactive exploration-based sampling strategy to counteract data sparsity. This function shifts the analytical paradigm from retrospective assessment to prospective experimental guidance (Figure 3a).

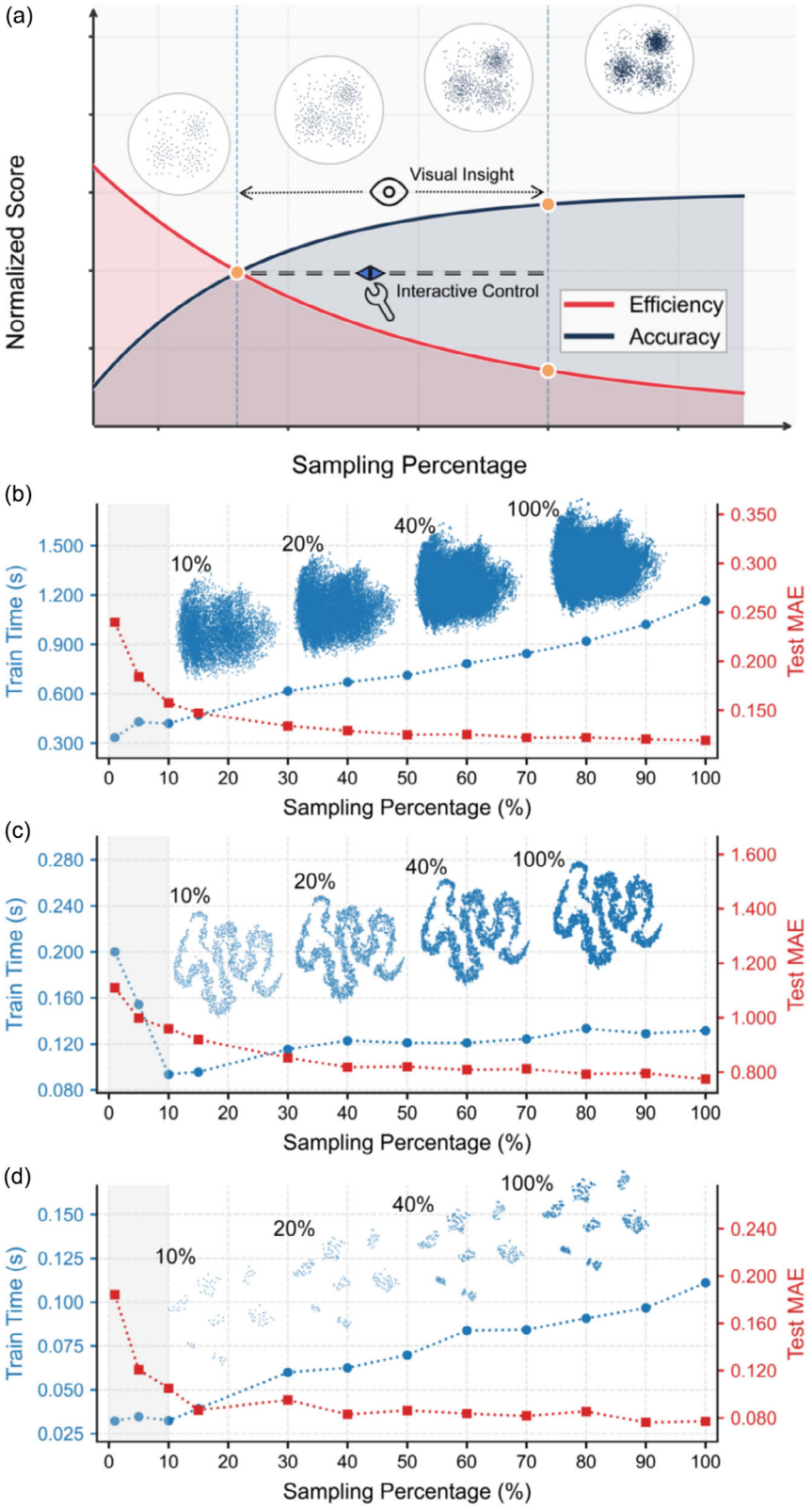


**FIGURE 2** | Quantitative validation of redundancy-reduced sampling. (a) Conceptual trade-off between efficiency and accuracy. D2VP enables users to identify a representative core subset, bypassing the extremes of all-data or minimal-data training. (b–d) Performance benchmarks across three datasets. Subplots compare training time (blue circles) and Test Mean Absolute Error (red squares) against sampling percentage. Scatter plots overlay the spatial distribution of sampled data at 10%, 20%, 40%, and 100%. Datasets: (b) formation energy (~65k samples) [45]; (c) aqueous solubility (~10k samples) [46]; (d) hydrogen absorption (~900 samples) [47].

We validated this capability on two benchmarks. First, using a zero-point energy (ZPE) dataset [47], we addressed the challenge of targeted refinement. Once a baseline model exists, the critical question becomes: which new experiments will most effectively correct model weaknesses? Our approach transforms the model into a prescriptive guide. By pinpointing "unknown unknowns" in sparse regions, the methodology provides actionable recommendations for data acquisition, maximizing the scientific return on investment for high-cost experiments.

Second, we applied the strategy to perovskite stability prediction [48]. Perovskite design involves a vast, complex search space where stability rules remain partially understood. Random exploration in such frontiers is inefficient. Here, the methodology

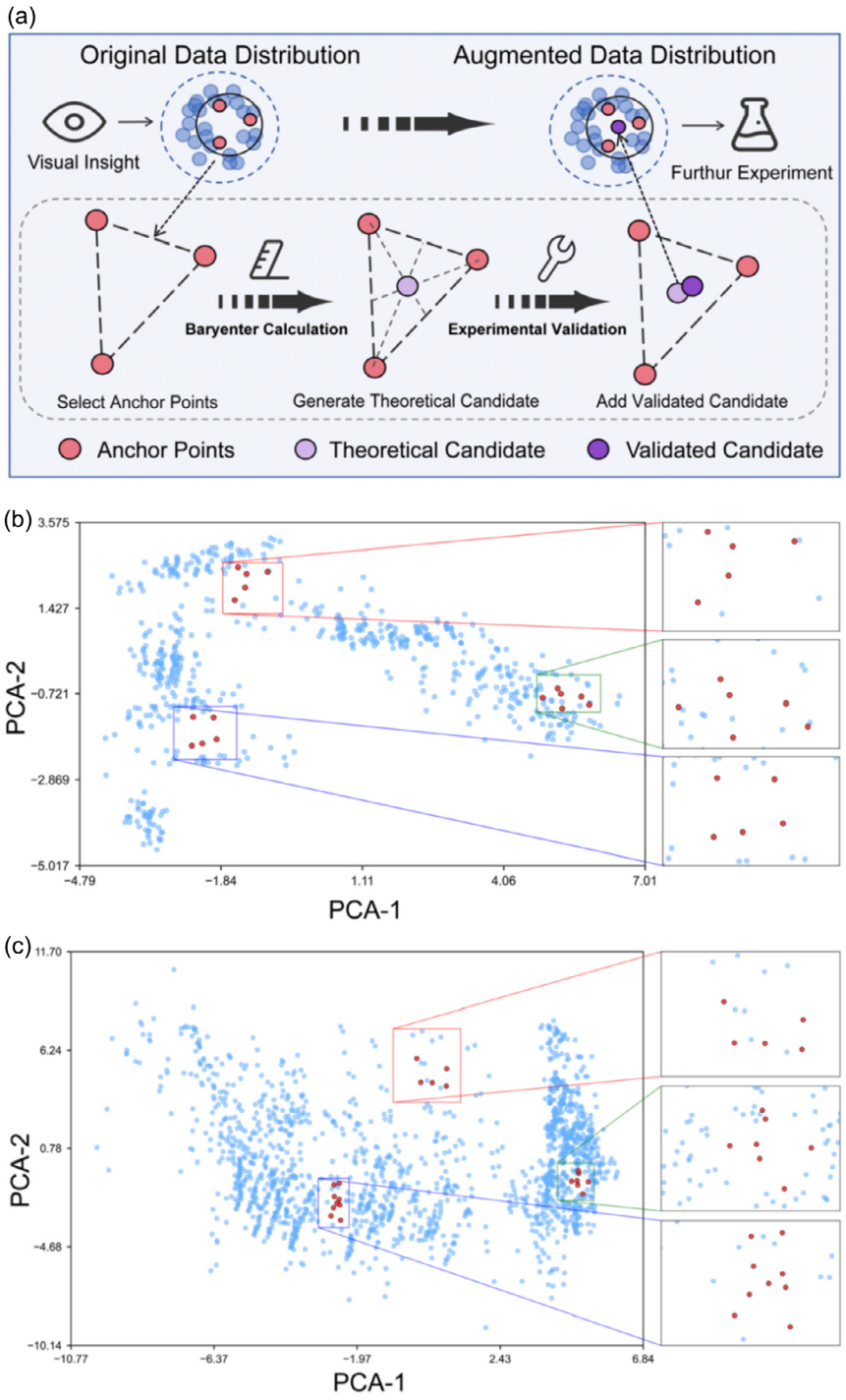


**FIGURE 3** | Validation of exploration-based sampling. (a) Schematic of sampling principle: anchor points in sparse regions define a theoretical candidate (barycenter), guiding experimental validation and augmented data distribution. (b,c) Visual outcomes for ZPE [47] and perovskite [48] datasets. Supplementary points (red) populate sparse "data voids" defined by original distribution (blue). (d,e) Quantitative impact. Box plots (left axis) and Mean MAE (red line, right axis) track performance gains as validated candidates are incrementally added to the training set.

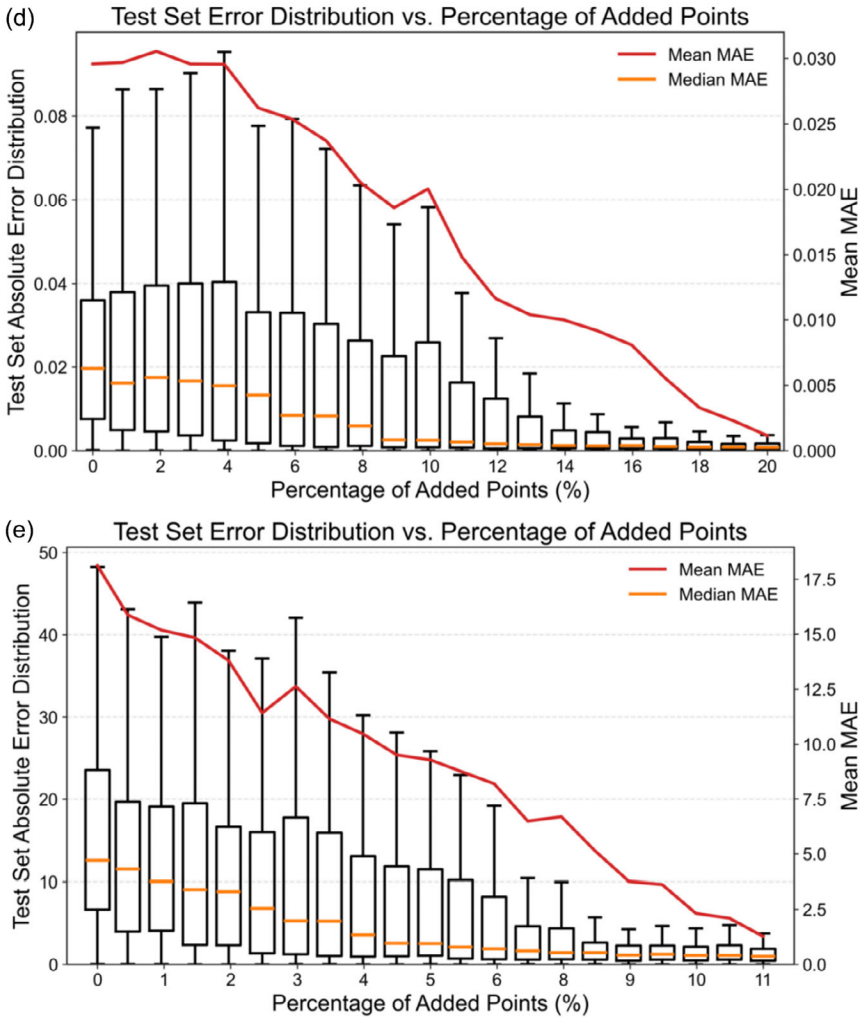


**FIGURE 3** | Continued.

serves as a strategic navigation tool, enabling researchers to systematically probe uncharted territories rather than clinging to known material families. This intelligent search strategy accelerates discovery by replacing serendipity with data-driven precision.

For both cases, we split the original data (9:1 training/test) and applied exploration-based sampling to the training set. We went beyond simulation: candidates proposed by the algorithm were experimentally validated to obtain ground-truth values (red markers in Figure 3b,c). These new points were incrementally added to the training set to quantify their impact on model performance (Figure 3d,e).

This strategy augments datasets by targeting sparse regions (Figure 3a). The workflow transforms 2D visual intuition into high-dimensional precision. Researchers initiate the process by identifying a "data void" in the projected space and selecting three "Anchor Points". The algorithm then retrieves their feature vectors to compute a barycenter, termed the "Theoretical Candidate". This vector provides a specific quantitative coordinate for new experiments. Post-validation, the resulting "Validated Candidate" is assimilated into the dataset. While experimental variance may shift the final projection, the process effectively populates the targeted sparse region.

Figure 3b,c visualize this mechanism in action. Supplementary points (red) do not scatter randomly but cluster within the "voids" of the original distribution (blue). We achieved this targeted placement via two complementary modes. An automated mode uses Delaunay triangulation [49] to identify the largest empty topological regions. Complementing this, a user-driven mode enables experts to manually select anchors based on domain intuition. This combination fuzes algorithmic rigor with expert knowledge to guide experimentation.

Figure 3d,e quantify the impact. In both cases, error rates decline as the model assimilates the recommended data. For the hydrogen absorption dataset (Figure 3d), the method refines an already effective model. Strikingly, for the perovskite benchmark (Figure 3e), adding just 10% of targeted data reduced the mean MAE by over 80%. This contrast underscores the methodology's dual utility: it fine-tunes high-performing models and, perhaps more significantly, stabilizes models trained on sparse data. By resolving sparsity, the approach accelerates knowledge acquisition in novel, data-limited domains.

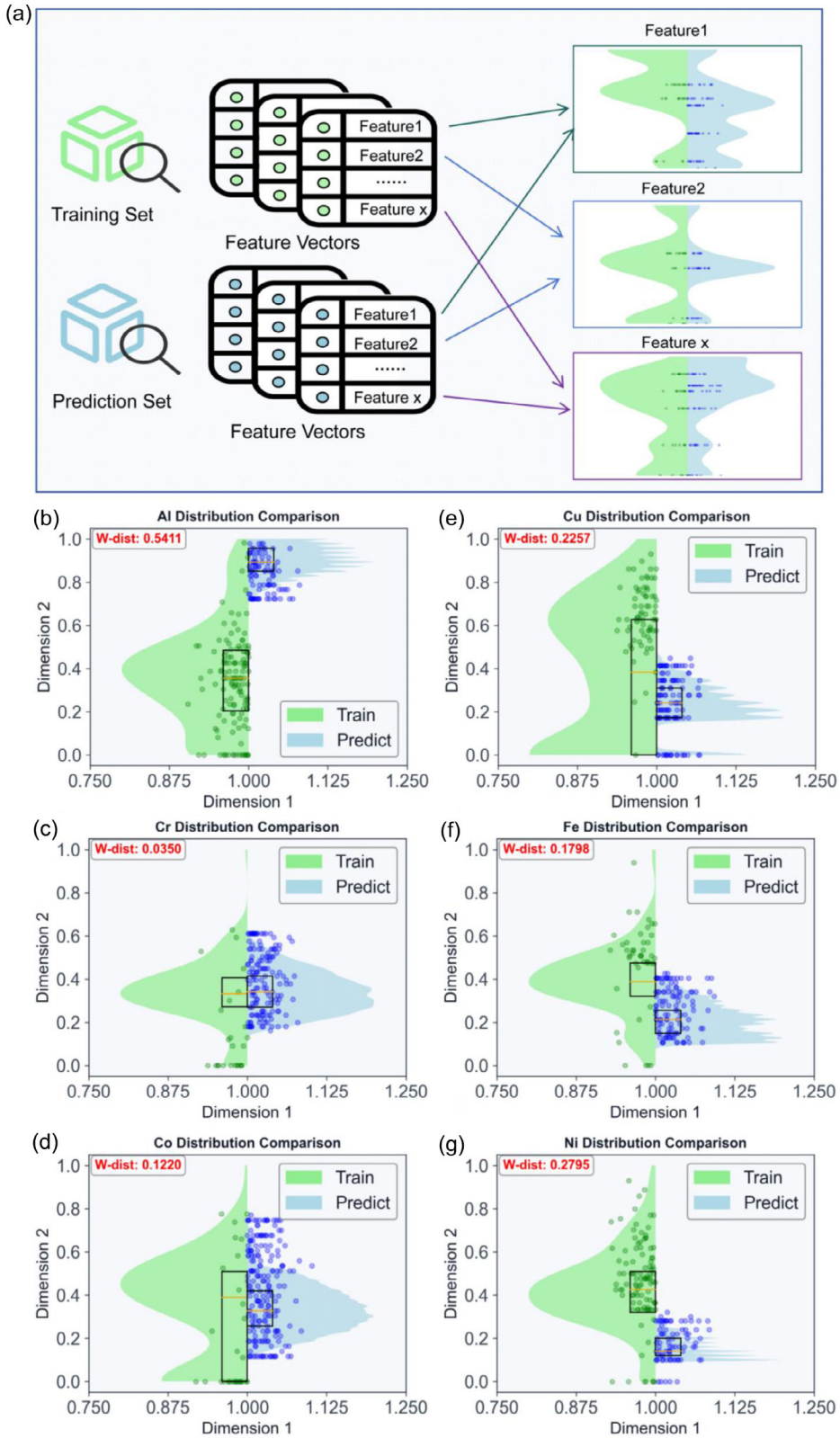


**FIGURE 4** | Feature-level diagnosis of extrapolation risk. (a) Diagnostic workflow schematic: feature vectors from training and prediction sets are compared via split-violin plots. (b–g) Application to HEA hardness prediction [51]. Each subplot compares one compositional feature's distribution between training (green) and prediction (blue) sets. Embedded box plots indicate interquartile range and median. Data are normalized prior to visualization. Wasserstein distance (W-dist) quantifies distributional divergence. The benchmark uses a training-to-prediction ratio of ~1:392, revealing severe mismatches for Al (b) and Ni (g).

### 2.4 | D2VP Methodology III: Statistical Diagnosis of Extrapolation Risk

Before holistic risk assessment, our diagnostic approach enables detailed feature-level scrutiny to evaluate prediction credibility (Figure 4a). This workflow extracts feature vectors from both training and prediction sets, comparing their distributions directly. We couple qualitative visual analysis (split-violin plots) with quantitative metrics (Wasserstein distance [50]) to assess distributional divergence. We applied this method to a high-entropy alloy (HEA) hardness benchmark [51], a domain where extrapolation risks are particularly acute.

HEA exploration often involves training on limited data to screen vast compositional spaces. Predicting exceptional properties is valuable only if the model operates within its applicability domain. Our feature-level diagnosis serves as a mandatory credibility check. It determines whether a candidate's distinctiveness reflects genuine material novelty or merely dangerous feature-space deviation. This analysis provides a scientific basis to distrust unsupported extrapolations, preventing the waste of experimental efforts on statistical artifacts. Figure 4b–g presents the comparative analysis for six key compositional features.

Traditional metrics like MAE yield a single averaged score that obscures prediction-specific risks. Our feature-level diagnosis overcomes this limitation. By coupling split-violin plots with Wasserstein distance, the method exposes how individual features contribute to distributional mismatch.

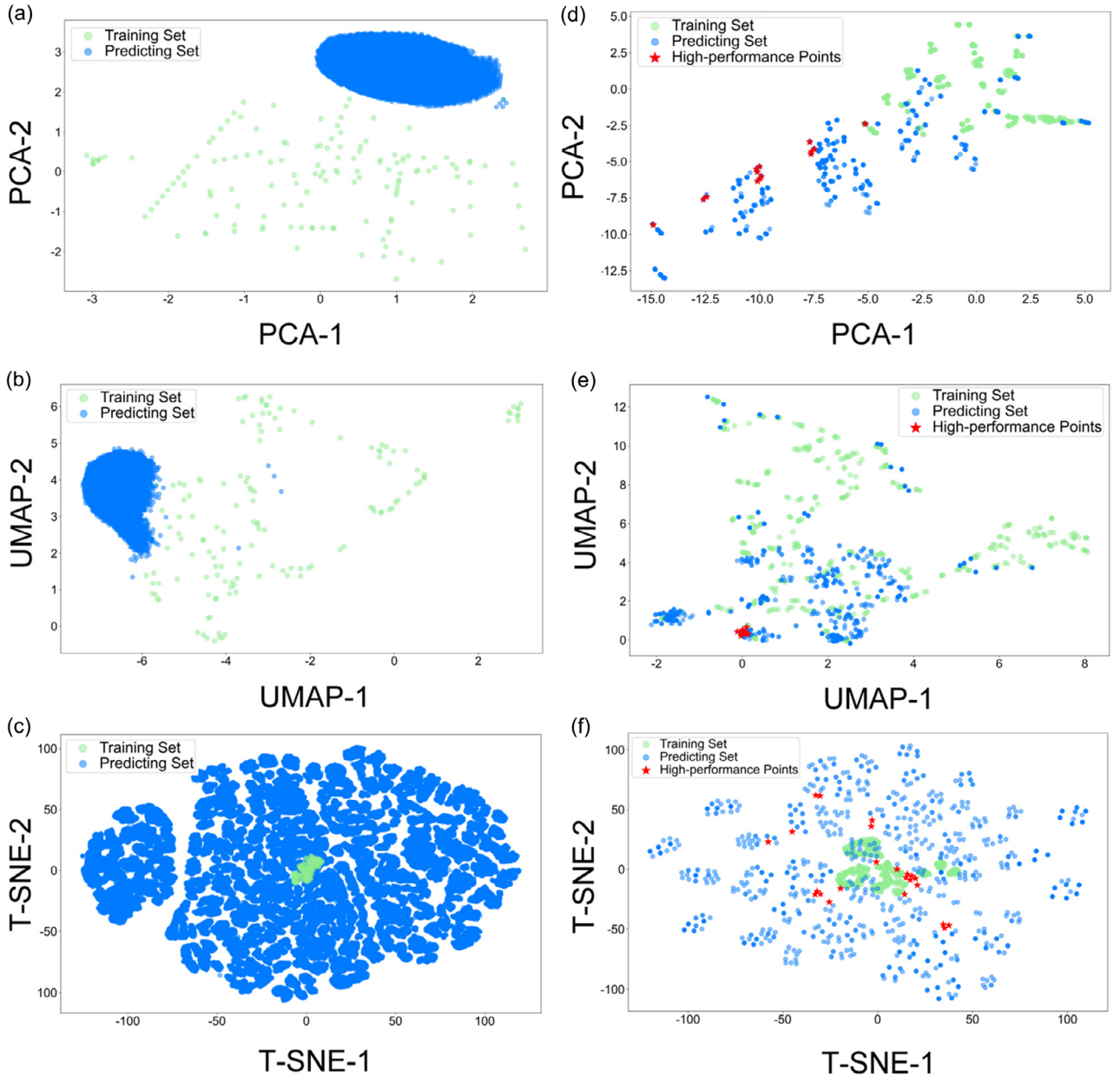


**FIGURE 5** | Holistic assessment of extrapolation risk via global visualization. (a–c) HEA design case [51], representing high-risk extrapolation. (d–f) HER catalyst design case [52], assessing reliability of high-performance predictions. Each subplot projects training (green) and prediction (blue) sets into 2D space using (a,d) PCA, (b,e) UMAP, or (c,f) t-SNE. Red stars mark "high-performance" candidates identified in source literature. The HEA case reveals large-scale spatial separation; the HER case exposes the subtler out-of-domain nature of top candidates despite apparent spatial overlap.

Applying this workflow to the HEA benchmark immediately reveals distributional drift (Figure 4b–g). Al and Ni exhibit the largest Wasserstein distances (0.5411 and 0.2795, respectively), indicating severe feature-space departure. Cu and Fe show similar divergence, confirming the mismatch is systemic rather than isolated. This quantitative evidence flags the model as operating in a high-risk, out-of-distribution regime, providing a decisive insight for reliability assessment.

This analysis validates feature-level diagnosis as an essential preliminary step. Researchers can interrogate model applicability before committing to costly experiments. Yet while this granular view pinpoints problematic features, it does not deliver a single holistic verdict. The subsequent section addresses this gap through global, low-dimensional visualization of the feature space.

## 2.5 | D2VP Methodology III: Holistic Reliability Assessment via Global Visualization

The preceding analysis identified specific sources of extrapolation risk but did not deliver a holistic verdict. To address this gap, we now present our global visual assessment capability. This component confronts the third DDM dimension: over-extrapolation. We applied three dimensionality reduction methods (PCA, UMAP, t-SNE) to generate comparative distribution plots for both HEA [51] and HER [52] benchmarks. Results are shown in Figure 5.

PCA (Figure 5a) captures the most fundamental separation, mapping training and prediction sets to entirely distinct regions. t-SNE (Figure 5c) resolves local structure: the training set collapses into a tight cluster, engulfed by a dispersed galaxy of prediction points. UMAP (Figure 5b) preserves global topology, depicting the prediction set as a dense "comet head" detached from the sparse training "tail". Together, these three views confirm that the model operates in a feature space fundamentally dissimilar to its training domain, spanning global distance, local structure, and overall topology.

The HER catalyst benchmark (Figure 5d–f) reveals a subtler but equally critical extrapolation pattern. At first glance, PCA and UMAP (Figure 5d,e) suggest reliable predictions due to apparent spatial overlap. However, the "High-performance Points" (red stars) cluster at the feature-space periphery, far from the training core. t-SNE (Figure 5f) sharpens this insight: most top candidates form isolated clusters composed exclusively of prediction data. The implication is stark. A seemingly validated prediction may still originate from high-risk extrapolation. This diagnostic capability enables researchers to distinguish robust, knowledge-driven discovery from potentially coincidental success.

Comparing the three dimensionality reduction methods clarifies their diagnostic strengths. PCA (Figure 5a,d) offers a macroscopic view based on global variance, revealing large-scale separation. t-SNE (Figure 5c,f) excels at resolving fine-grained local structure, decomposing data into distinct clusters. UMAP (Figure 5b,e) balances both perspectives, preserving local clustering while maintaining global topology. Including all three is a deliberate design choice. This multi-lens toolkit enables researchers to select the projection best suited to their analytical question.

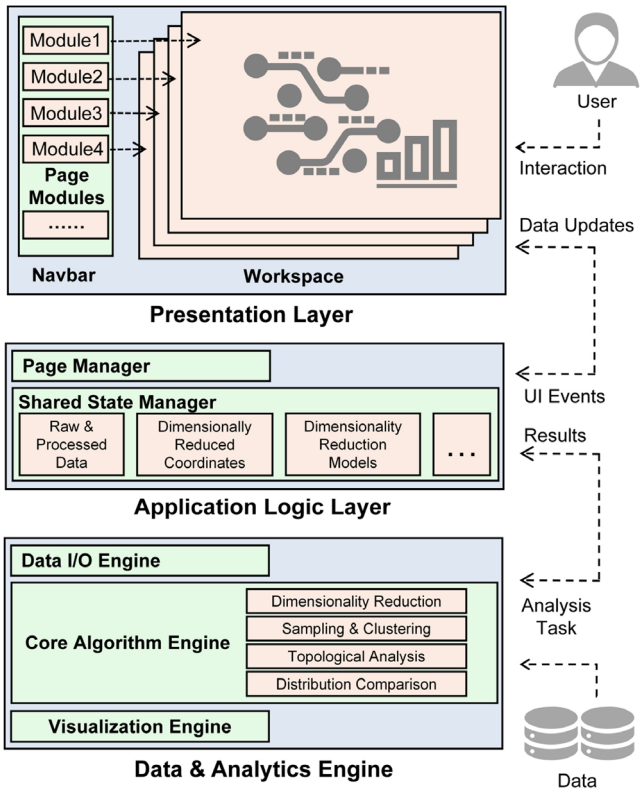


**FIGURE 6** | Software architecture of the D2VP platform. The modular three-layer design decouples user-facing interface modules from backend computation. The Shared State Manager coordinates algorithm and visualization engines, enabling the interactive analysis workflow central to this methodology.

Figure 5 exposes two forms of over-extrapolation central to DDM. The first is overt spatial separation (HEA benchmark). The second is structural extrapolation concealed within apparent overlap (HER benchmark). Our comparative distribution plots render both patterns transparent. This approach transforms reliability assessment from abstract metrics into an intuitive, evidence-based inquiry.

## 2.6 | Platform Architecture

To operationalize the visual-first methodology, we implemented D2VP as a modular, interactive platform (Figure 6). The architecture comprises three layers: (1) a Core Algorithm Engine encapsulating dimensionality reduction, sampling, and distribution comparison routines; (2) a Visualization Engine rendering analytical outputs; and (3) a Shared State Manager coordinating data flow and enabling real-time, human-in-the–loop interaction. This decoupled design separates data processing from user interface logic, facilitating extensibility and maintenance. The platform is released as open-source software (https://github.com/Weijie-Yang/D2VP) to ensure reproducibility and community contribution.

## 3 | Conclusion

Machine learning adoption in scientific discovery is fundamentally challenged by Data Distribution Mismatch (DDM), a condition where dataset biases and structural imbalances compromise

model reliability. Our validation studies confirm this phenomenon is pervasive and a primary driver of ML misuse in scientific applications. This paper introduced a data-centric, visual-first methodology to diagnose and mitigate DDM at its source, operationalized through the D2VP platform.

Validation studies in materials science confirmed the efficacy of this integrated approach. The redundancy-reduced sampling function addresses data over-density, enabling favorable trade-offs between training efficiency and accuracy. The exploration-based sampling function resolves data sparsity by guiding experimental efforts toward underrepresented regions. The credibility analysis toolkit identifies both overt and subtle forms of over-extrapolation, safeguarding the integrity of scientific findings.

These validated functionalities signal a paradigm shift in reliability assessment. Our approach transforms model evaluation from abstract metrics into an interactive, evidence-based process of visual inquiry. The core philosophy merges domain expertise with data-driven visual insights for more confident decision-making. This data-centric paradigm holds potential beyond materials science. As demonstrated by additional validation studies in the Supporting Information, D2VP offers a generalizable framework for enhancing model reliability across the AI for Science landscape.

D2VP is an interactive, standalone platform designed for domain experts without extensive ML expertise. Its modular architecture, which comprises algorithm, visualization, and state-management engines, enables real-time, human-in-the-loop analysis. To encourage adoption and further development, the platform is open-source and freely available at https://github.com/Weijie-Yang/D2VP. A visual walkthrough of each module appears in Figures S1-S5, and complete workflows are demonstrated in Supplementary Videos 1–5.

## 4 | Methods

### 4.1 | Dimensionality Reduction Methods

D2VP provides three complementary dimensionality reduction techniques for visualizing high-dimensional feature spaces [53]. Principal Component Analysis (PCA) [54, 55] captures global variance by projecting data onto orthogonal principal components, revealing large-scale distributional structure. t-Distributed Stochastic Neighbor Embedding (t-SNE) [56] preserves local neighborhoods by minimizing Kullback–Leibler divergence between high- and low-dimensional similarity distributions, resolving fine-grained cluster structure. Uniform Manifold Approximation and Projection (UMAP) [57] balances local and global topology via fuzzy simplicial set optimization, enabling scalable visualization of large datasets. Users can switch between these projections to interrogate data structure from multiple geometric perspectives.

### 4.2 | Redundancy-Reduced Sampling Methods

To address data over-density, D2VP offers three sampling strategies. Cluster-based sampling [58] partitions the feature space via k-means and selects cluster centroids as representative points. Diversity sampling [59] maximizes pairwise distances to retain structurally distinct samples. Random sampling [60] provides an unbiased baseline. Users select a strategy based on dataset characteristics; all three reduce training set size while preserving topological coverage.

### 4.3 | Exploration-Based Sampling Methods

To address data sparsity, D2VP provides a targeted sampling workflow that guides experiments toward underrepresented feature-space regions. The process proceeds through four stages: (1) identify sparse regions in the 2D projection, (2) select three anchor points bounding the target void, (3) compute a theoretical candidate via barycentric interpolation [61, 62] of the anchors' high-dimensional feature vectors, and (4) validate experimentally. Two operational modes are available. An automated mode employs Delaunay triangulation [49] to detect maximal empty triangles. A user-driven mode enables manual anchor selection based on domain intuition. Together, these modes translate visual insight into actionable experimental coordinates.

These newly synthesized data points provided critical sampling guidance. They highlighted specific areas within the feature space where additional experimental samples would be most beneficial. This strategic direction then allows D2VP to minimize redundant experiments and accelerate the discovery process by focusing on unexplored yet potentially valuable areas.

### 4.4 | Wasserstein Distance for Distribution Comparison

In addition to the qualitative visual comparison provided by the split-violin plots, a quantitative metric is required to formally measure the dissimilarity between the feature distributions of the training and prediction sets. For this purpose, the Wasserstein-1 distance, also known as the Earth Mover's Distance (EMD) [50], is employed. Conceptually, the W-dist represents the minimum "cost" or "work" required to transform one distribution into another, akin to moving a pile of earth from one shape to another. A larger distance implies a greater difference between the two distributions.

For any two one-dimensional distributions P and Q, the Wasserstein-1 distance was calculated as a scalar measure of the minimum transport cost required to transform one empirical distribution into the other [63]. Within the D2VP platform, this metric was computed for each individual feature, providing a single value that quantified the dissimilarity between the training and prediction sets for that feature; larger values indicate more significant distributional shift.

**Acknowledgements**

This work was supported by Beijing Natural Science Foundation (no. 2262076), Natural Science Foundation of Hebei (no. E2025502039), Fundamental Research Fund for the Central Universities (no. 2025JC008 and 2025MS131), Suzhou MatSource Technology Co., Ltd. (Suzhou,

China) and Gusu Laboratory of Materials (Suzhou, China), grant number Y2501.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The D2VP platform, including the full source code, a pre-compiled stand-alone executable for Windows, and video tutorials demonstrating the key functionalities, is openly available in our GitHub repository at https://github.com/Weijie-Yang/D2VP. All datasets used and/or analyzed during the current study are also available in this repository or from the original sources cited in the manuscript.

### References

1. T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, (Microsoft Research, Redmond, WA, 2009).

2. A. Agrawal, and A. Choudhary, "Perspective: Materials Informatics and Big Data: Realization of the 'Fourth Paradigm' of Science in Materials Science," *APL Materials* 4 (2016): 053208.

3. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine Learning for Molecular and Materials Science," *Nature* 559 (2018): 547–555.

4. R. Batra, L. Song, and R. Ramprasad, "Emerging Materials Intelligence Ecosystems Propelled by Machine Learning," *Nature Reviews Materials* 6, no. 8 (2021): 655–678.

5. D. Zhang, X. Jia, H. Liu, et al., "Cloud Synthesis: A Global Closed-Loop Feedback Powered by Autonomous AI-Driven Catalyst Design Agent," *AI Agent* 1 (2025): 2.

6. S. M. Moosavi, K. M. Jablonka, and B. Smit, "The Role of Machine Learning in the Understanding and Design of Materials," *Journal of the American Chemical Society* 142, no. 48 (2020): 20273–20287.

7. D. Morgan, and R. Jacobs, "Opportunities and Challenges for Machine Learning in Materials Science," *Annual Review of Materials Research* 50 (2020): 71–103.

8. Y. Lin, and P. Ou, "From Descriptors to Machine Learning Interatomic Potentials: A Review of AI-Accelerated Electrocatalyst Design," *AI Agent* 1 (2025): 3.

9. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (2019): 206–215.

10. M. Kulichenko, B. Nebgen, N. Lubbers, et al., "Data Generation for Machine Learning Interatomic Potentials and Beyond," *Chemical Reviews* 124, no. 24 (2024): 13681–13714.

11. Y. Liu, Z. Yang, X. Zou, et al., "Data Quantity Governance for Machine Learning in Materials Science," *National Science Review* 10, no. 7 (2023): nwad125.

12. S. Kim, J. Xu, W. Shang, Z. Xu, E. Lee, and T. Luo, "A Review on Machine Learning-Guided Design of Energy Materials," *Progress in Energy* 6, no. 4 (2024): 042005.

13. R. Liu, and A. Wallqvist, "Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds," *Journal of Chemical Information and Modeling* 59, no. 1 (2019): 181–189.

14. E. M. Askenazi, E. A. Lazar, and I. Grinberg, "Identification of High-Reliability Regions of Machine Learning Predictions Based on Materials Chemistry," *Journal of Chemical Information and Modeling* 63, no. 23 (2023): 7350–7362.

15. Y. Zhang, and C. Ling, "A Strategy to Apply Machine Learning to Small Datasets in Materials Science," *npj Computational Materials* 4 (2018): 25.

16. C. J. Bartel, "Data-Centric Approach to Improve Machine Learning Models for Inorganic Materials," *Patterns* 2, no. 11 (2021): 100382.

17. D. Zhang, and H. Li, "The Hidden Engine of AI in Electrocatalysis: Databases and Knowledge Graphs at Work," *Molecular Chemistry & Engineering* 1, no. 1 (2025): 100003.

18. R. Liu, H. Wang, K. P. Glover, M. G. Feasel, and A. Wallqvist, "Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure-Activity Relationship Models Based on Deep Neural Networks?," *Journal of Chemical Information and Modeling* 59, no. 1 (2019): 117–126.

19. N. Fujinuma, B. L. DeCost, J. Hattrick-Simpers, and S. E. Lofland, "Reflections on the Future of Machine Learning for Materials Research," arXiv:2112.09764 (2021).

20. A. Davariashtiyani, B. Wang, S. Hajinazar, E. Zurek, and S. Kadkhodaei, "Impact of Data Bias on Machine Learning for Crystal Compound Synthesizability Predictions," *Machine Learning: Science and Technology* 5, no. 4 (2024): 040501.

21. H. Wang, W. Guan, J. Chen, Z. Wang, and D. Zhou, "Towards Heterogeneous Long-Tailed Learning: Benchmarking, Metrics, and Toolbox," *Advances in Neural Information Processing Systems* 37 (2024): 73098–73123.

22. C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning," *Journal of Chemical Information and Modeling* 61, no. 6 (2021): 2623–2640.

23. R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "Data Mining with Imbalanced Class Distributions: Concepts and Methods," in Proceedings of the 4th Indian International Conference on Artificial Intelligence (IICAI 2009) (2009) 359–376.

24. H. Zhang, W. Chen, J. M. Rondinelli, and W. Chen, "ET-AL: Entropy-Targeted Active Learning for Bias Mitigation in Materials Data," *Applied Physics Reviews* 10, no. 2 (2023): 021403.

25. V. Tshitoyan, J. Dagdelen, L. Weston, et al., "Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature," *Nature* 571 (2019): 95–98.

26. X. Jia, A. Lynch, Y. Huang, et al., "Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis," *Nature* 573 (2019): 251–255.

27. M. Kumagai, Y. Ando, A. Tanaka, K. Tsuda, Y. Katsura, and K. Kurosaki, "Effects of Data Bias on Machine-Learning-Based Material Discovery Using Experimental Property Data," *Science and Technology of Advanced Materials: Methods* 2, no. 1 (2022): 302–309.

28. B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y. Han, "Reliable and Explainable Machine-Learning Methods for Accelerated Material Discovery," *npj Computational Materials* 5 (2019): 108.

29. K. Li, A. N. Rubungo, X. Lei, et al., "Probing Out-of-Distribution Generalization in Machine Learning for Materials," *Communications Materials* 6 (2025): 9.

30. D. Horvath, G. Marcou, and A. Varnek, "Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models," *Journal of Chemical Information and Modeling* 49, no. 7 (2009): 1762–1776.

31. S. Dimitrov, G. Dimitrova, T. Pavlov, et al., "A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models," *Journal of Chemical Information and Modeling* 45, no. 4 (2005): 839–849.

32. C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler, "Identifying Domains of Applicability of Machine Learning Models for Materials Science," *Nature Communications* 11 (2020): 4428.

33. L. E. Schultz, Y. Wang, R. Jacobs, and D. Morgan, "A General Approach for Determining Applicability Domain of Machine Learning Models," *npj Computational Materials* 11 (2025): 95.

34. N. Udayangani, H. M. Dolatabadi, S. Erfani, and C. Leckie, "Exploiting Inter-Sample Information for Long-Tailed Out-of-Distribution Detection," in Proceedings of the Winter Conference on Applications of Computer Vision (WACV) (2025) 8535–8544.

35. E. S. Muckley, J. E. Saal, B. Meredig, C. S. Roper, and J. H. Martin, "Interpretable Models for Extrapolation in Scientific Machine Learning," *Digital Discovery* 2 (2023): 1425–1435.

36. M. Yu, Y.-N. Zhou, Q. Wang, and F. Yan, "Extrapolation Validation (EV): A Universal Validation Method for Mitigating Machine Learning Extrapolation Risk," *Digital Discovery* 3, no. 5 (2024): 1058–1067.

37. F. Oviedo, J. Lavista Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and Explainable Machine Learning for Materials Science and Chemistry," *Accounts of Materials Research* 3, no. 6 (2022): 597–607.

38. A. Jain, J. Montoya, S. Dwaraknath, et al., "The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools," in *Handbook of Materials Modeling: Methods: Theory and Modeling*, ed. W. Andreoni, and S. Yip (Springer International Publishing, 2020) 1751–1784.

39. M. Scheidgen, L. Himanen, A. N. Ladines, et al., "NOMAD: A Distributed Web-Based Platform for Managing Materials Science Research Data," *Journal of Open Source Software* 8, no. 90 (2023): 5388.

40. L. Talirz, S. Kumbhar, E. Passaro, et al., "Materials Cloud, a Platform for Open Computational Science," *Scientific Data* 7 (2020): 299.

41. Y. Du, Y. Wang, Y. Huang, et al., "M2Hub: Unlocking the Potential of Machine Learning for Materials Discovery," *Advances in Neural Information Processing Systems* 36 (2023): 77359–77378.

42. O. Antons, and J. C. Arlinghaus, "Data-Driven and Autonomous Manufacturing Control in Cyber-Physical Production Systems," *Computers in Industry* 141 (2022): 103711.

43. X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A Survey of Human-in-the-Loop for Machine Learning," *Future Generation Computer Systems* 135 (2022): 364–381.

44. J. J. Thomas, and K. A. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (IEEE Computer Society, 2005).

45. C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, "A Critical Examination of Compound Stability Predictions from Machine-Learned Formation Energies," *npj Computational Materials* 6 (2020): 97.

46. M. C. Sorkun, A. Khetan, and S. Er, "AqSolDB, a Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds," *Scientific Data* 6 (2019): 143.

47. V. Korostelev, J. Wagner, and K. Klyukin, "Simple Local Environment Descriptors for Accurate Prediction of Hydrogen Absorption and Migration in Metal Alloys," *Journal of Materials Chemistry A* 11, no. 43 (2023): 23576–23588.

48. W. Li, R. Jacobs, and D. Morgan, "Predicting the Thermodynamic Stability of Perovskite Oxides Using Machine Learning Models," *Computational Materials Science* 150 (2018): 454–463.

49. B. Delaunay, "Sur la Sphère Vide," *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk* 7 (1934): 793–800.

50. Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision* 40 (2000): 99–121.

51. C. Wen, Y. Zhang, C. Wang, et al., "Machine Learning Assisted Design of High Entropy Alloys with Desired Property," *Acta Materialia* 170 (2019): 109–117.

52. H. Li, S. Xu, M. Wang, et al., "Computational Design of (100) Alloy Surfaces for the Hydrogen Evolution Reaction," *Journal of Materials Chemistry A* 8, no. 35 (2020): 17987–17997.

53. R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London. Series A* 222 (1922): 309–368.

54. K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine* 2, no. 11 (1901): 559–572.

55. J. Shlens, "A Tutorial on Principal Component Analysis," arXiv:1404.1100 (2014).

56. L. van der Maaten, and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research* 9 (2008): 2579–2605.

57. L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv:1802.03426 (2018).

58. J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1 (1967) 281–297.

59. B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648 (University of Wisconsin-Madison Department of Computer Sciences, Madison, WI, 2009).

60. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, 2009).

61. A. F. Möbius, *Der Barycentrische Calcul: ein neues Hülfsmittel zur analytischen Behandlung der Geometrie, dargestellt und insbesondere auf die Bildung neuer Classen von Aufgaben und die Entwickelung mehrerer Eigenschaften der Kegelschnitte angewendet* (J. A. Barth, 1827).

62. A. Z. Peixinho, B. C. Benato, L. G. Nonato, and A. X. Falcão, "Delaunay Triangulation Data Augmentation Guided by Visual Analytics for Deep Learning," in 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) (IEEE, 2018) 384–391.

63. G. Peyré, and M. Cuturi, "Computational Optimal Transport: With Applications to Data Science," *Foundations and Trends in Machine Learning* 11, no. 5-6 (2019): 355–607.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.