Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 277

Leveraging domain knowledge for optimal initialization in Bayesian materials optimization

Trevor Hastings, *[ab] James Paramore, [b] Brady Butler [c] and Raymundo Arróyave [bde]

Bayesian optimization (BO) has emerged as an effective strategy to accelerate the discovery of new materials by efficiently exploring complex and high-dimensional design spaces. However, the success of BO methods greatly depends on how well the optimization campaign is initialized—the selection of initial data points from which the optimization starts. In this study, we focus on improving these initial datasets by incorporating materials science expertise into the selection process. We identify common challenges and sources of uncertainty when choosing these starting points and propose practical guidelines for using expert-defined criteria to create more informative initial datasets. By evaluating these methods through simulations and real-world alloy design problems, we demonstrate that using domain-informed criteria leads to initial datasets that are more diverse and representative. This enhanced starting point significantly improves the efficiency and effectiveness of subsequent optimization efforts. We also introduce clear metrics for assessing the quality and diversity of initial datasets, providing a straightforward way to compare different initialization strategies. Our approach offers a robust and widely applicable framework to enhance Bayesian optimization across various materials discovery scenarios.

Received 14th August 2025
Accepted 18th November 2025

DOI: 10.1039/d5dd00361j

rsc.li/digitaldiscovery

1 Introduction

New demands for novel and transformative materials—driven by rapidly evolving applications in energy, sustainability, electronics, aerospace, and other emerging technologies—have significantly accelerated the need for more efficient materials discovery and development processes, prompting a greater adoption of machine learning (ML) and adaptive design methods.[1–4] These methods are commonly evaluated using metrics such as hypervolume improvement (a measure that quantifies the expansion of the solution space covered by optimal points in multiobjective optimization), cumulative regret (the accumulated difference between the results of selected solutions and the theoretical optimal results across iterations) or training loss (a measure of predictive accuracy over iterative evaluations).[5–7] However, such evaluations often remain limited to simulated or narrowly scoped scenarios, rarely extending to practical considerations such as the initialization of the campaign—the initial set of queries of the design space—and its significant impact on the performance of the optimization scheme. Few studies systematically address the methodological and logistical challenges inherent in designing and deploying closed-loop materials discovery campaigns in real world.[8–10]

Among adaptive discovery methods, Bayesian optimization (BO)[11] is especially suited to material discovery, where experimental and computational evaluations are expensive and time consuming.[11–14] BO utilizes a probabilistic model, typically a Gaussian process (GP),[15] to iteratively select promising points by balancing exploration (sampling uncertain regions) and exploitation (sampling regions predicted to have high performance). Although Bayesian inference has a long-established history,[16] its systematic integration into practical experimental workflows, particularly concerning initialization strategies, remains relatively underexplored.[17,18] BO has demonstrated broad success in applications ranging from single-objective optimization (where a single metric or property is optimized) to complex multi-objective design (optimizing multiple properties simultaneously) due to its ability to treat the relationship between input variables and outcomes as a black-box. Furthermore, batch implementations of BO allow parallel evaluations, aligning closely with practical constraints and timelines in real-world discovery campaigns.[17,19]

[a]*Bush Combat Development Complex, Texas A&M University System, 3479 TAMU, College Station, TX, 77843-3479, USA. E-mail: trevorhastings@tamu.edu*

[b]*Department of Materials Science and Engineering, Texas A&M University, 3003 TAMU, College Station, TX 77843-3003, USA*

[c]*DEVCOM Army Research Laboratory, 3003 TAMU, College Station, TX 77843-3003, USA*

[d]*J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University, 3003 TAMU, College Station, TX 77843-3003, USA*

[e]*Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas A&M University, 3003 TAMU, College Station, TX 77843-3003, USA*

Despite the growing adoption of BO in materials discovery, published efforts frequently overlook critical issues associated with selecting the initial queries (*i.e.* the initial dataset) used to jump start the campaign. In the context of genetic algorithm (GA)-based optimization, Maaranen *et al.*[20] examined how different strategies for generating initial populations influence optimization performance, particularly in early stopping scenarios. While their work focused on evolutionary methods, the underlying observation—that initialization can have a dominant effect when only a limited number of iterations are possible—is highly relevant to experimental materials discovery, where datasets are generally sparse and queries are expensive. These conditions are not unique to GA-based methods; they apply equally to BO and other adaptive/iterative optimization strategies commonly used in materials research. Despite this, many experimental studies begin with randomly or heuristically selected initial datasets, often without explicit discussion of their rationale or potential impact on optimization outcomes.[5] A related effort has explored the initialization role alongside human interventions in an autonomous workflow for piezoresponsive materials;[21] this work relied on latent space representations while the present work focuses on domain knowledge and physically interpretable material pro perties. It is worth noting a multi-decade gap here, and that the premiere summary background works on setting up BO problems, such as,[12] do not have sections dedicated to initialization.

Moreover, researchers typically assemble optimization workflows from modular algorithmic components—such as machine learning regressors (*e.g.*, Gaussian processes), kernel functions (*e.g.*, squared exponential), and acquisition functions (*e.g.*, expected hypervolume improvement)—available in widely used machine learning libraries (*e.g.*, PyTorch, scikit-learn).[22–24] However, the availability of these modular tools does not guarantee effective integration into experimental workflows, particularly with respect to initial dataset selection. Without assessments explicitly designed to evaluate the quality of initial datasets—metrics independent of inaccessible ground truths—researchers lack essential guidance for selecting initial points that optimize subsequent performance.

In this paper, we directly address these critical challenges by systematically examining the impact of initial dataset selection on optimization outcomes. We propose a structured, domain-informed methodology for selecting initial design points, incorporating materials science principles such as subsystem complexity (complexity defined by the number of constituent elements or phases) and configurational entropy (a measure of disorder or diversity in atomic arrangements). Using these domain-informed criteria, we demonstrate an improved representativeness and diversity of the initial data sets, significantly improving subsequent optimization efficiency. We use practical metrics to evaluate the quality of initial datasets without requiring knowledge of optimal solutions. We validate this approach using representative design spaces involving both computational and experimental alloy datasets. This work provides robust and practical methodological guidance for materials researchers, significantly improving the reproducibility and effectiveness materials discovery efforts that rely on these techniques.

2 Methods

2.1 Optimization

The Bayesian optimization (BO) approach used in this work is based on the BIRDSHOT framework,[19,25] a recent algorithm developed for efficient, multi-objective discovery in high-dimensional design spaces. While BIRDSHOT has been previously applied in experimental materials campaigns, here we employ it as a general-purpose BO engine to study the influence of initialization on the efficiency of batch BO campaigns.

BO is well-suited for optimizing expensive, black-box objective functions, which are common in scientific and engineering applications where each evaluation is costly. The method constructs a surrogate model of the objective function—here, a Gaussian Process Regressor (GPR)—that provides both a mean prediction and an uncertainty estimate. This uncertainty quantification enables BO to balance exploitation (sampling candidates expected to perform well) with exploration (sampling candidates where the model is uncertain), typically *via* an acquisition function. The GP surrogate provides an uncertainty-aware model over the current design space as a guide for exploration. In the multi-objective setting, we employ Expected Hypervolume Improvement (EHVI), which assigns a scalar utility to each candidate based on its contribution to the expansion of the current Pareto front.

To support parallel experimentation and accelerate convergence, BIRDSHOT implements *batch Bayesian optimization*. In each iteration, an ensemble of GPR models is constructed with randomized hyperparameters, allowing for diverse candidate proposals. From these proposals, a representative batch is selected using *k*-medoids clustering, ensuring both coverage and diversity in the queried designs. This ensemble-based selection strategy reduces sensitivity to hyperparameter tuning and guards against premature convergence to local optima—factors that are particularly important when the BO campaign begins with limited data. As a result, all batch data retrains the ensemble of GPs (no static partitioning with validation or test sets are used).

In this study, we leverage BIRDSHOT's batch optimization capabilities to investigate how different initialization strategies affect the quality and efficiency of BO performance. We analyze the impact of initialization on model accuracy, exploration coverage, and Pareto front convergence, providing guidance for the deployment of BO in resource-constrained discovery workflows. Complete details of the formation of GP surrogate models, acquisition strategy, and batch selection procedure are provided in *Section 5: SI*.

2.2 Initial sampling and features

The datasets used in this study were generated on high-performance computing infrastructure *via* CALPHAD-based simulations performed using Thermo-Calc. Two major compositional design spaces were explored.

The first dataset corresponds to a face-centered cubic (FCC) high-entropy alloy system in the composition space $Al_{x_1}V_{x_2}Cr_{x_3}Fe_{x_4}Co_{x_5}Ni_{x_6}$, where x_i ranges from 0 to 0.95 in

increments of 0.05. This results in approximately 50 000 unique compositions prior to any down-selection. Property outputs in this data set include CALPHAD-predicted room-temperature density and room-temperature heat capacity. The second dataset corresponds to a body-centered cubic (BCC) high-entropy alloy system in the space $Ti_{x_1}V_{x_2}Nb_{x_3}Mo_{x_4}Hf_{x_5}Ta_{x_6}W_{x_7}$, where x_i ranges from 0 to 0.975 in increments of 0.025 prior to filtering. Property outputs include ROM-calculated Shannon configurational entropy and CALPHAD-calculated room temperature thermal conductivity. In addition, the BCC dataset contains experimental measurements of specific hardness and specific modulus for 48 selected compositions.

The material datasets used in this work consist of data derived from thermodynamic and kinetic property predictions obtained *via* CALPHAD-based software. The Al–V–Cr–Fe–Co–Ni system is primarily composed of elements that form face-centered cubic (FCC) alloys, and equilibrium properties including solidification range (K), room-temperature density (g cm^{-3}), heat capacity (J mol^{-1} K^{-1}), and thermal conductivity (W m^{-1} K^{-1}) were calculated from Thermo-Calc. The Ti–V–Nb–Mo–Hf–Ta–W system contains body-centered cubic (BCC) forming alloys, with further methodology described in ref. 25.

Both of these datasets correspond to experimental BO works in the field of alloy design, where single phase solid solutions enable a simple correspondence between simulation and laboratory validation. CALPHAD-derived material property output spaces is physically interpretable and provides a consistency for benchmarking black box problems. Unlike simulated environments that get to draw conclusions from repeat trials, batch physical experimentation typically only has time and funding for one attempt. The net effects of an initialization strategy become veritably relevant.

These datasets were down-selected to create initial datasets for this work. The initial data sets were created using two distinct sampling strategies: uniform random sampling and *k*-medoids clustering. Each strategy was repeated 1000 times to ensure statistical robustness. For computational datasets, batches of 20 samples were used in each initialization, while experimental batches were limited to 8 samples due to availability constraints. These initialization methods and their corresponding performance are further discussed in Section 3*: Results and discussion*.

To evaluate the diversity and representativeness of initial batches, three geometric features were computed for each selected subset: the bounding box area, the centroid deviation, and the average convex hull size. The bounding box area is defined as the smallest axis-aligned rectangular region enclosing the selected points in the multi-objective output space. The centroid deviation is computed as the average Euclidean distance between each point and the centroid of the selected set in the objective space, serving as a compactness metric. Convex hulls were computed by identifying the local neighborhood of each candidate (as if it were a medoid) and measuring the size of the convex region that encloses those neighbors. The average hull size over all points in the batch serves as a proxy for disparity and structural dispersion. These metrics were computed using standard numerical libraries: bounding box and centroid distances were calculated using NumPy, while convex hulls were computed using the scipy.spatial module. Medoid clustering was implemented using the *k* medoids function from the scikit-learn-extra package.

2.3 Visualization

To visualize high-dimensional compositional spaces in two dimensions, this work employs affine projections based on barycentric coordinate systems. Each data point in the composition space—represented as a normalized vector $p = [x_1, x_2, \ldots, x_n]$ such that $x_i = 1$—is projected into two-dimensional Cartesian space *via* a linear transformation. The resulting 2D layout forms a regular polygon, where the number of vertices equals the number of elements (or bins) in the system. These projections provide an interpretable and structure-preserving means of visualizing high-entropy alloy compositions and are based on the methodology described in ref. 26.

To construct the full projection space at a given resolution, we first generate a set of all possible barycentric coordinates by discretizing the composition simplex. Each barycentric vector is then projected into R^2 using a matrix of vertex coordinates that defines a regular polygon in two dimensions. Experimental, random, or off-grid points are projected in the same manner, allowing consistent visualization of arbitrary subsets relative to the entire compositional domain.

Visual annotations, including figure labels and overlays, were created using Inkscape, an open-source vector graphics editor.[27] All visuals in this study were rendered at high resolution to maintain geometric precision and readability in both digital and print formats.

3 Results and discussion

3.1 Variance within optimizations

We first define a multi-objective optimization problem over the FCC Al–V–Cr–Fe–Co–Ni alloy composition space, sampled at 5 at% intervals. This space serves as a testbed for materials-related black-box optimization, with property outputs modeled *via* CALPHAD and including solidification range (K), room-temperature density (g cm^{-3}), heat capacity (J mol^{-1} K^{-1}), and thermal conductivity (W m^{-1} K^{-1}) (see Section 2*: Methods*). To reduce computational overhead and emulate realistic experimental limitations, the original space of approximately 50 000 compositions was down-selected to 2000 representative points using the *k*-medoids clustering algorithm.[28]

This sub-sampling procedure preserves the diversity of the original design space while enabling tractable evaluation during Bayesian optimization. Unlike centroid-based methods such as *k*-means, *k*-medoids selects actual data points as cluster centers, making it well suited for generating representative and physically valid subsets in high-dimensional, multi-objective alloy design spaces. A visualization of the resulting optimization domain is provided in Fig. 1.

An optimization policy can be simulated over the reduced design space by initializing with a small subset of data and iteratively selecting new points for evaluation, treating the CALPHAD-

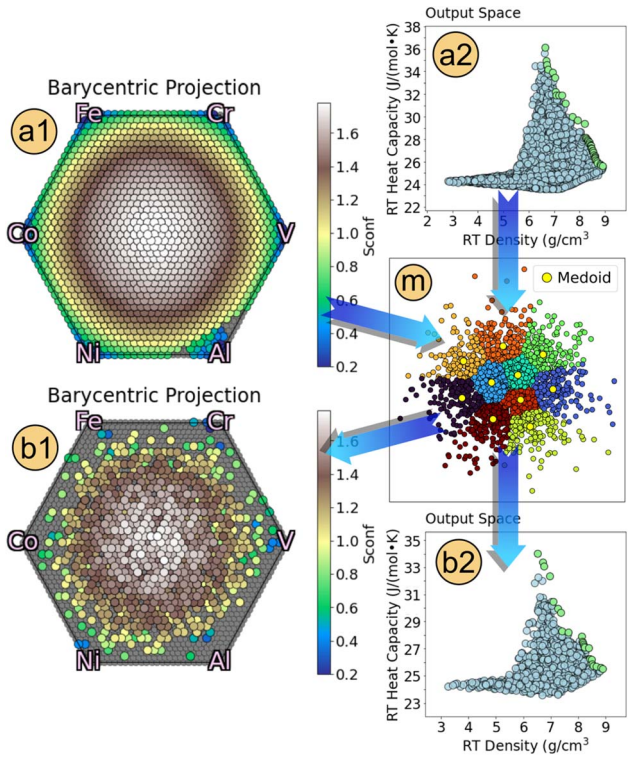

Fig. 1 A high dimensional space is reduced using a clustering algorithm in order to provide a representative dataset that retains the topological qualities of the original. (a1) The input dimensions (atomic fractions of 6 elements) projected onto a 2D plane with barycentric coordinates, sorted by configurational entropy for approximately 50 000 data points. (a2) Two example output dimensions of the same dataset. (m) An example schematic of a medoid function applied to arbitrary 2D data. (b1) The input dimensions after clustering to 2000 data points. (b2) The output dimensions after clustering. Even though only 4% of the dataset is being shown, it retains a similar topology to the original.

modeled properties as if they were experimentally measured. To implement this, we employ the BIRDSHOT method,[19] as described in Section 2. From this ensemble, candidate points are proposed based on expected hypervolume improvement (EHVI). Due to overlap in GP predictions, the resulting candidate pool is typically redundant and is reduced to a set of unique suggestions. These are then clustered using the *k*-medoids algorithm to select a final batch of diverse, high-utility points without imposing an explicit ranking (see Section 2*: Methods* and ref. 25). This simulation framework enables efficient, uncertainty-aware optimization under realistic data constraints.

As a demonstration, we examine the impact of two distinct initialization strategies on optimization performance. The first strategy employs a uniform random selection of initial points, while the second uses a subset deliberately chosen to span the output space through diversity-aware sampling. In both cases, the same BIRDSHOT policy is subsequently applied for 10 iterations, using a realistic batch size of 20 candidates per iteration. The optimization aims at joint maximization of two objectives, with all other conditions kept constant. The resulting optimization trajectories for both initialization schemes are presented in Fig. 2.

3.2 Initialization effects on optimization behavior

Based on the results shown in Fig. 2, several key observations can be made regarding the influence of initialization on optimization dynamics:

3.2.1 Initial conditions strongly influence perceived success. The composition of the starting dataset can significantly affect the observed optimization trajectory and, by extension, the interpretation of success. When high-performing points are present in the initial dataset, subsequent gains may appear minimal, potentially misleading performance-based metrics to suggest that optimization is ineffective. In contrast, when seed points under-sample high-performing regions, even modest improvements may seem disproportionately large. In both cases, common progress metrics are confounded by initialization effects, which complicates fair cross-comparisons of optimization performance.

3.2.2 Random initialization limits interpretability. Initializing the campaign with randomly selected points leaves the experimenter with little contextual knowledge of their location in the design space, making it difficult to understand or trust subsequent improvements.

(a) This problem is especially pronounced in sequential (non-batch) optimization, where early decisions are made with sparse and potentially unrepresentative data.

(b) Alternative strategies such as grid-based sampling, input-space clustering, or point-inhibition methods (*e.g.*, nearest-neighbor maximization) attempt to improve upon random

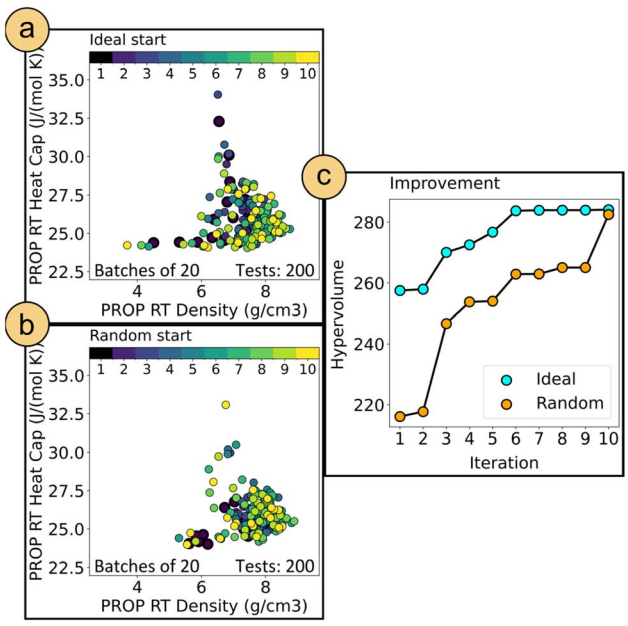

Fig. 2 Batch Bayesian optimization (BBO) simulation in the FCC alloy design space, targeting the maximization of two CALPHAD-predicted properties. (a) Optimization trajectory initialized with a dataset based directly on the output data as an example of broad coverage of the output space. (b) Trajectory initialized with a randomly selected dataset. (c) Hypervolume progression as a function of batch iteration for both strategies. Although the same optimization policy was applied in both cases, the quality of the initial dataset resulted in a performance difference equivalent to approximately 100 additional evaluations.

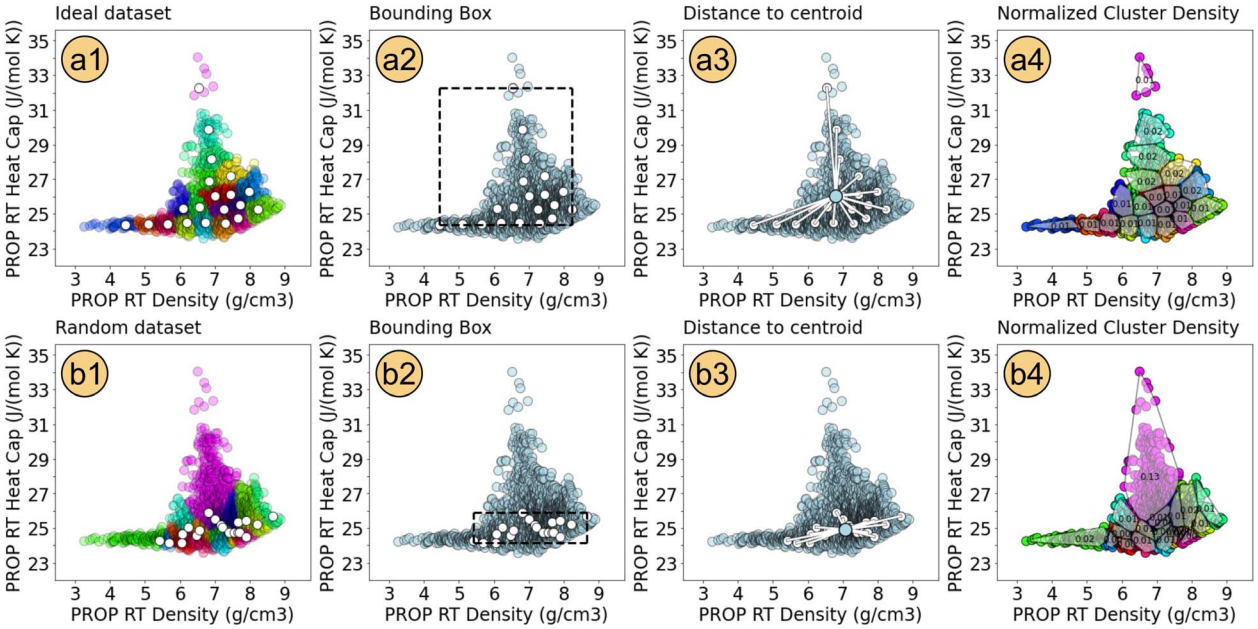

Fig. 3 Comparison of two initial datasets in the objective space. Panel (a1) shows an ideal initialization obtained *via* k-medoids clustering on the output dimensions, yielding broad and representative coverage. Panel (b1) shows a worst-case initialization generated *via* uniform random sampling, which tends to oversample high-density regions and overlook sparsely populated, high-information areas. Three quantitative metrics are used to characterize dataset quality: bounding box volume (a2 and b2), which reflects the overall spread of the dataset; mean distance to centroid (a3 and b3), which quantifies dispersion; and convex hull area (a4 and b4), which measures the extent of coverage in the output space.

initialization by enforcing geometric separation in the input space. However, these approaches do not necessarily ensure adequate coverage of the objective space, particularly when the input–output mapping is non-linear or many-to-one. As a result, they may still produce initial datasets that poorly represent the diversity of achievable outcomes.

3.2.3 Exploration–exploitation trade-offs are topology-dependent. An optimization algorithm's ability to balance exploration and exploitation critically depends on the extent to which the initial dataset captures the underlying structure of the objective landscape. When the initial points cover only a small portion of the design space, the optimization behavior may become indistinguishable from a random search, as nearly every new point offers an apparent improvement. In such cases, evaluating the strategic behavior of the optimizer becomes difficult or even meaningless.

These findings collectively point out the importance of principled initialization in black-box optimization. Without a systematic and transparent methodology for selecting initial data, claims about optimization efficiency, convergence behavior, or exploration capability risk being irreproducible or misleading. Initialization should therefore be treated as a critical component of optimization design, not a procedural afterthought.

3.3 Encoding additional dimensions

A clear operational definition of best- and worst-case initialization scenarios provides a meaningful basis for comparing the quality of different starting datasets. Worst-case initializations are readily characterized as randomly selected subsets that offer minimal coverage or variability in the objective space. Such data sets often fail to expose key regions of the phase space, forcing any downstream optimization policy to first engage in inefficient exploratory sampling before converging on informative solutions.

In contrast, a best-case initialization would consist of a subset that reflects the global structure of the objective space as faithfully as possible, given the available budget. This can be approximated by performing k-medoids clustering directly in the output (objective) space, thereby identifying a representative set of initial points that span the range of achievable responses. By capturing key topological features of the output distribution from the outset, such an initialization reduces the burden on the optimizer to discover underrepresented regions through search alone.

Fig. 3 illustrates the difference between high-quality and low-quality initializations by comparing two candidate datasets in the objective space. Panel (a1) displays a set of 20 medoids selected directly from the output dimensions using k-medoids clustering, along with the associated data clusters. By construction, this initialization provides broad and representative coverage of the output space, capturing both dense and sparse regions.

Panel (b1), on the other hand, shows a random sample of 20 data points, with each remaining point assigned to its nearest selected point to form a "pseudo-medoids" clustering for visual comparison. Because random sampling is biased toward high-density regions, the selected points in this case do not adequately cover low-density, high-information regions of the space. This issue is especially apparent in the present example,

where the majority of the data is concentrated at low heat capacity values, leading the random sample to overlook critical areas in the distribution.

To quantify these differences, panels (a2)–(a4) and (b2)–(b4) present three metrics used to evaluate the geometric quality of the initialization set in the output space: (1) bounding box volume, which measures the overall spread of selected points; (2) mean distance to centroid, which captures dispersion and sensitivity to outliers; and (3) convex hull area (or volume in higher-dimensional spaces), which reflects the extent of topological coverage across the dataset. Initialization sets that are well suited for optimization should provide a rich information sample of both input and output spaces. Desirable initializations are those that can incorporate topological outliers and avoid excessive clustering within locally dense regions.

Each of the proposed metrics is designed to reward broader coverage and penalize redundant, tightly grouped selections that fail to span the full range of objective values. The *bounding box volume* reflects the overall axis-aligned extent of the selected points and serves as a quick diagnostic for whether the initialization spans the full dynamic range of each objective. Although it may overestimate true coverage in non-convex or sparse regions, it remains effective for identifying collapsed or overly concentrated subsets. The *mean distance to the centroid* measures internal dispersion and is sensitive to structural variability; it increases not only with spread but also with the presence of outliers, offering insight into how uniformly the dataset samples around its central tendency. The *convex hull volume* provides a more topology-aware measure of coverage by enclosing the minimal convex region that contains all selected points. This makes it especially valuable for distinguishing between broad but disconnected spreads and genuinely representative samples that trace the boundary of the achievable objective space.

It is important to emphasize that the clustering shown in Fig. 1(m) illustrates just one of many valid ways to partition the output space into representative regions. There is no unique solution to this decomposition, even when accounting for local density or distributional features. For example, the placement of medoids can be rotated or perturbed around the center of the space without meaningfully affecting their representativeness. In more complex datasets, small variations in the boundaries of the clusters – particularly those driven by nearest-neighbor assignments – can lead to multiple equally plausible clustering configurations, each reflecting a different but defensible interpretation of the structure of the space. Despite this inherent variability in how the output space may be partitioned, the metrics introduced here exhibit consistent and interpretable behavior across thousands of subsets constructed both randomly and deliberately. As shown later in Fig. 5, these diagnostics remain robust under non-unique sampling conditions, reliably capturing meaningful differences in dataset structure regardless of the specific clustering configuration.

A common strategy for selecting an initial dataset—once purely random sampling has been ruled out—is to construct a grid that spans the input space as uniformly as possible. For example, in the context of a binary phase diagram, one might sample 11 evenly spaced compositions between 0% and 100% of each element in 10% increments. Although this approach is straightforward and effective in low-dimensional symmetric systems, it becomes increasingly impractical in higher dimensions, particularly when physical constraints restrict the set of

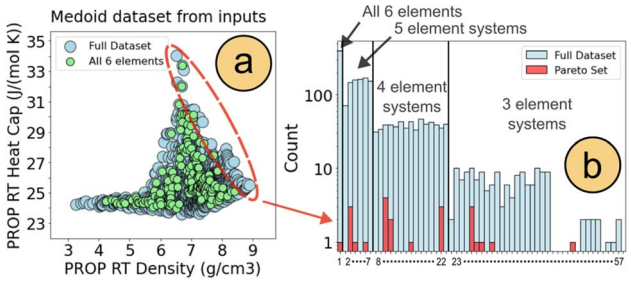

Fig. 4 Distributions in a six-dimensional compositional input space and their relationship to Pareto-optimal outputs. (a) Projection of the output space showing heat capacity *versus* density, with points corresponding to the most populous elemental sub-system (all six elements present) highlighted. (b) Histogram of subsystem populations, with the number of Pareto-optimal points from each subsystem overlaid in red. The discrepancy between input-space population and contribution to the Pareto front illustrates that highly populated compositional regions do not necessarily yield the most optimal candidates. This decoupling underscores the risk of using input-space density as a proxy for importance in optimization strategies.

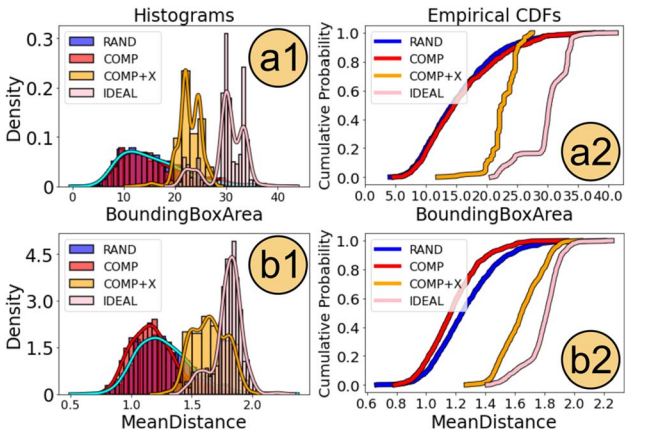

Fig. 5 Histograms and empirical cumulative distribution functions (CDFs) comparing two geometric diversity metrics—bounding box area and mean distance to centroid—across 1000 initialization datasets generated using four distinct sampling strategies: **RAND** (random sampling in input space), **COMP** (composition-only *k*-medoids), **COMP+X** (composition medoids augmented with the subsystem complexity dimension), and **IDEAL** (medoids clustered directly in the output space using known property values). The **IDEAL** case serves as a theoretical benchmark, requiring prior knowledge of target properties and thus representing the upper bound for initialization quality. (a1 and b1) Histograms of bounding box areas and centroid distances for each strategy. (a2 and b2) Empirical CDFs of the same metrics. Incorporating the subsystem complexity feature in the **COMP+X** strategy substantially improves output-space coverage, shifting both distributions closer to the **IDEAL** baseline. These results demonstrate that chemically informed descriptors enhance the representativeness of initial datasets, providing broader property diversity and stronger foundations for subsequent optimization and discovery.

feasible compositions. In such cases, algorithms such as *k*-medoids offer a tractable and interpretable alternative by identifying representative subsets that are well distributed across the valid input space.

For example, in the six-element Al–V–Cr–Fe–Co–Ni design space, selecting 20 representative compositions from a candidate pool of 2000 enables efficient initialization of a high-dimensional optimization campaign. Although one of the six compositional variables is linearly dependent due to the constraint that atomic fractions must sum to one, all dimensions are treated equally when computing pairwise distances. Assigning a consistent length scale to each dimension ensures a well-defined geometry in the input space, even when some features are mathematically redundant. As will be shown later, such derived or constrained features can still contribute a meaningful structure to optimization algorithms, particularly when used in kernel-based models or distance-aware sampling strategies.

This strategy is grounded in a widely held assumption in materials science: small changes in input variables generally lead to small, continuous changes in material properties. As a result, sampling the input space at roughly uniform intervals is often viewed as a reasonable way to ensure early-stage information gain. However, this assumption breaks down in many realistic scenarios and uniform input-space sampling becomes problematic for two main reasons:

(1) It assumes that equal spacing in the input space corresponds to comparably informative samples in output space. This is rarely valid in high-dimensional or nonlinear systems, where local gradients in material properties can vary dramatically and unpredictably. In such settings, equidistant sampling may result in the oversampling of trivial regions and the undersampling of highly informative ones.

(2) Empirically, uniform input-space sampling often produces output distributions that are statistically similar to those resulting from random sampling, particularly when the input–output mapping is non-linear or the output space is unevenly populated. This diminishes any practical advantage of regular grids in many real-world design tasks.

To illustrate point (1), consider a binary phase diagram such as the Al–Cu system.[29] These diagrams often contain structurally simple regions alongside composition intervals with complex multiphase behavior. Suppose one were tasked with reconstructing the Al–Cu phase diagram from 21 cooling curves. A naive strategy might place samples every 5% across the full composition range, but this would result in wasted effort in simple regions and inadequate resolution where the behavior is most intricate, such as between 50–85% Cu. A more effective strategy would concentrate the samples in the region of highest complexity, as visualized in Section 5*: SI*. Although such complexity is not known *a priori* in new systems, domain knowledge, prior experience, and physically grounded heuristics can guide the design of more informative initializations.

Point (2) is demonstrated graphically and may be interpreted as a population-based bias. Input-space distance metrics—such as those used in Euclidean kernels or nearest-neighbor algorithms—implicitly treat all displacements of equal magnitude as equally informative. For example, a 5% increase in one element paired with a 5% decrease in another yields the same distance, regardless of where this perturbation occurs in composition space. However, this symmetry assumption breaks down in constrained or irregular spaces. As shown in Fig. 1, even after down-selection, the 2000 alloy candidates span compositional regions with vastly different densities. Algorithms that rely on geometric distance in input space will tend to oversample high-density regions, implicitly assuming that Pareto-optimal solutions are similarly distributed.

This assumption parallels what is known in the philosophy of science as the *Presumptuous Philosopher* problem:[30–32] reasoning from population priors without direct measurement or justification. Although uniform assumptions may be harmless for provably irrelevant variables, such as the parity of candidate indices—they are epistemically unjustified when applied to output-relevant features. Supposing that optimal solutions mirror the population distribution of candidates introduces a subtle but significant selection bias, which can mislead both experimental design and algorithmic search strategies.

This hypothesis can be empirically tested using the previously described dataset of 2000 alloy compositions, shown in Fig. 4. Fig. 4 (left) illustrates that the most populous elemental subsystem—comprising all six elements—closely mirrors the overall topology of the complete dataset. As a result, restricting sampling to this subsystem does not improve initial diversity relative to random sampling from the full composition space. Fig. 4 (right) presents population distributions for each elemental subsystem. Notably, the Pareto front—defined by optimizing two representative CALPHAD-modeled properties and highlighted in red—shows a marked mismatch between input population density and the location of optimal alloys. Several Pareto-optimal candidates originate from sparsely populated subsystems; in this example, 7 out of 23 Pareto-optimal points fall within low-population groups. This decoupling between subsystem frequency and Pareto quality persists regardless of the specific property pair selected for optimization.

As demonstrated earlier in our toy dataset example, input-space population is a poor predictor of where optimal properties reside. Topologically, such divergence is inevitable: a six-dimensional compositional manifold, even if uniformly sampled, becomes distorted when projected into two- or three-dimensional property space *via* nonlinear chemistry–property mappings. Consequently, regions densely populated in the input space need not correspond to peaks in property performance. However, when optimization algorithms consistently surface candidates from densely sampled regions, it becomes easy to conflate frequency with merit. This tendency risks introducing a form of selection bias, where genuinely superior candidates from underrepresented subsystems are systematically overlooked.

To mitigate this issue, we propose augmenting the input feature space with chemically informed descriptors, specifically a 'subsystem complexity' feature that encodes the unique combination of constituent elements for each alloy. Including

this dimension in the initialization process, *via* *k*-medoids clustering, ensures that medoids are selected not only based on geometric proximity but also on chemical diversity, independent of population density. This approach effectively breaks the symmetry imposed by conventional distance-based methods, which treat equal compositional shifts as equally meaningful. In reality, small changes in minor constituents can have outsized impacts: for example, increasing an element from 30% to 35% may have marginal effects, while reducing another from 5% to 0% can fundamentally alter the behavior of the material—such as fully eliminating Al in stainless steel, which precludes the formation of protective passivation layers of alumina.[33] Since the number of possible elemental combinations far exceeds the number of initialization points, this strategy inherently favors subsystem diversity and helps counteract the population-driven sampling bias. A visual example of this approach, using *k*-medoids clustering over an augmented feature space and visualized *via* affine projections (as in Fig. 1), is provided in Section 5*: SI*.

We demonstrate this strategy using the previously described CCA toy dataset. A discrete 'complexity' feature was added to the input space, ranging from 1 to 57 and corresponding to all unique elemental combinations of 2 to 6 elements (excluding single-element systems) from a six-element pool, as shown in Fig. 4 (right). Clustering was initialized using the *k*-medoids++ method over the compositional dimensions, with or without the complexity feature, depending on the strategy. A total of 1000 initialization datasets were generated under four distinct sampling approaches: **RAND** (random sampling in the input space), **COMP** (only composition medoids), **COMP+X** (composition medoids with added complexity dimension) and **IDEAL** (medoids clustered directly in the output space, serving as a theoretical benchmark for optimal property coverage).

For each initialization set, we computed the boundary box area and the mean distance to the centroid in the objective space, visualizing the distributions of these metrics using histograms and empirical cumulative distribution functions (CDFs), as shown in Fig. 5. These results clearly indicate that **COMP** and **RAND** yield nearly indistinguishable distributions, both of which deviate substantially from the **IDEAL** benchmark. In contrast, the **COMP+X** strategy shifts both metric distributions toward the **IDEAL** regime, demonstrating significantly improved output-space diversity.

Further inspection of Fig. 5 reveals that datasets generated by **RAND** or **COMP** generally exhibit smaller bounding boxes and shorter centroid distances, indicating limited exploration of property space. As shown earlier in Fig. 3, these strategies also tend to exclude alloys with higher heat capacities, which are relatively rare in the dataset. By incorporating the complexity feature, **COMP+X** systematically improves the diversity of selected initial points in the output space, providing a stronger foundation for subsequent optimization. In this specific system—although the pattern is likely generalizable—our results suggest that *greater chemical complexity correlates with greater property diversity*, supporting the use of chemically informed descriptors in initialization design.

3.4 Application to a refractory alloy dataset

We applied this approach to the BCC refractory alloy system Ti–V–Nb–Mo–Hf–Ta–W, selecting a subset of this seven-dimensional composition space based on CALPHAD-modeled properties. Two target outputs were considered: the Shannon configurational entropy (k_B), which explicitly quantifies the alloy complexity, and the room-temperature thermal conductivity, κ_{RT} (W m^{-1} K^{-1}), a key property for high-temperature structural applications.[34] Fig. 6(a) visualizes the input space, including the full design space (gray), a production-constrained region (red), a CALPHAD-feasible region based on stable phase equilibria (orange), and the final optimization subset, colored by configurational entropy.

This dataset and methodology are based on the framework established by Paramore *et al.*, who demonstrated the utility of Bayesian optimization for alloy design in high-dimensional refractory systems.[25] Despite the fact that the input space forms a compact and fully connected manifold, the output space—as shown in Fig. 6(b)—exhibits significant structural discontinuities. In particular, alloys with high thermal conductivity (κ_{RT} > 30 W m^{-1} K^{-1}) are sparsely represented, forming isolated regions in objective space. This

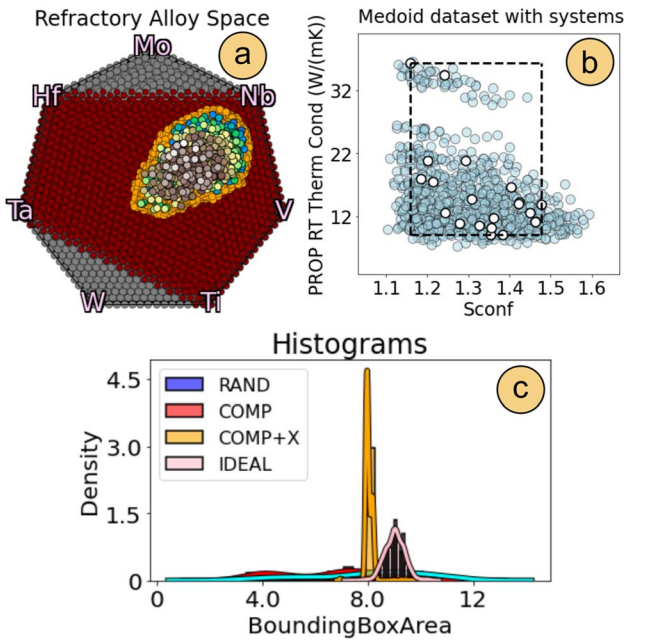

Fig. 6 (a) Barycentric projection of the optimized Ti–V–Nb–Mo–Hf–Ta–W alloy dataset, constrained to CALPHAD-feasible compositions and colored by Shannon configurational entropy. (b) Output space visualization showing thermal conductivity *versus* configurational entropy, with example medoids selected using the COMP+X strategy (composition plus subsystem complexity) highlighted, along with their axis-aligned bounding box. (c) Histogram comparing bounding box areas across 1000 initialization subsets generated using four strategies: RAND (random input sampling), COMP (composition-only medoids), COMP+X (composition medoids augmented with subsystem complexity), and IDEAL (output-space medoids as a theoretical upper bound). Results show that including subsystem complexity (COMP+X) systematically increases output-space spread, yielding consistently larger bounding boxes and improved diversity in the property space.

undersampling of extreme property alloys can hinder downstream tasks such as optimization and discovery, as further illustrated in Fig. 2.

After applying thermodynamic and production feasibility constraints, 24 unique elemental subsystems remained in the BCC refractory alloy design space. One subsystem in particular—the Hf-free system—contained a disproportionately large number of candidates satisfying all constraints, while several others had fewer than ten viable compositions. As in the previous case, composition-based sampling alone would likely overlook these sparsely populated regions, assigning undue weight to the more populous subsystem. Augmenting the design space with a subsystem 'complexity' dimension once again offers a principled solution to this sampling bias.

To evaluate this approach, we generated 1000 initial datasets using the established sampling strategies: **RAND** (random input sampling), **COMP** (composition-only medoids), **COMP+X** (composition medoids augmented with subsystem complexity), and **IDEAL** (output-space medoids serving as a theoretical upper bound). As shown in Fig. 6(c), the **COMP+X** strategy yields initializations with output-space diversity approaching that of the **IDEAL** case. In contrast, **RAND** and **COMP** consistently fail to produce initialization sets with sufficient coverage of the property space. An extended version of this figure similar to Fig. 5 is provided in Section 5*: SI*.

Although earlier analyses demonstrated the utility of encoding subsystem complexity in computational datasets derived from CALPHAD modeling or ideal-mixing assumptions, it is essential to validate this approach using experimental data. Unlike modeled properties, experimental measurements often exhibit additional complexity and variability, which poses challenges to standard sampling strategies. To investigate this, we analyzed a dataset from Paramore *et al.*, comprising 48 experimentally synthesized alloys in the Ti–V–Nb–Mo–Hf–Ta–W system, each tested for specific hardness and specific modulus as part of a batch optimization study. Given the limited size of the dataset, we generated 1000 subsets of eight alloys each to evaluate the robustness of different sampling methods.

Fig. 7 (left) shows the distribution of specific hardness and specific modulus in the output space, along with one representative subset. The subset is visualized *via* its medoid and connections to the other selected alloys. Because the two mechanical properties are strongly correlated, the bounding box area provides limited discriminative power for quantifying diversity. Instead, we focus on mean distance to centroid as a more robust indicator of output-space coverage. Fig. 7 (right) shows a histogram of this metric across 1000 subsets generated *via* three sampling strategies: **RAND** (random selection), **COMP** (medoids selected from compositional inputs), and **COMP+X** (medoids selected from inputs augmented with subsystem complexity). As this is an experimental dataset, no **IDEAL** sampling strategy is available for benchmarking.

Despite the modest dataset size, **RAND** frequently yielded subsets with low centroid distances and limited property diversity. **COMP** provided only marginal improvement over random sampling, whereas **COMP+X** consistently produced more diverse subsets, better spanning the full extent of the

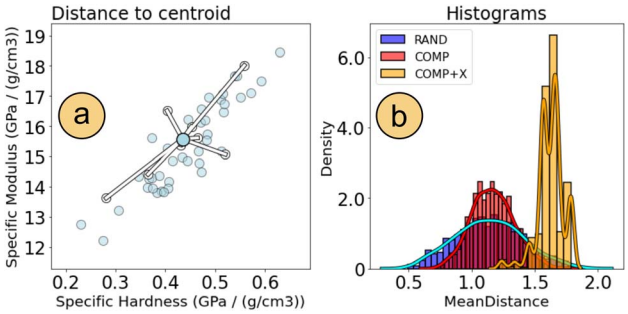

Fig. 7 (a) A set of 48 experimentally tested refractory alloys evaluated for specific hardness and specific modulus. One example subset of 8 alloys is shown, along with their distances to the centroid in output space. (b) Histogram of centroid distances for 1000 subsets generated using three sampling strategies: RAND (random selection), COMP (medoids selected from input compositions), and COMP+X (medoids selected from inputs augmented with elemental combination information). Even when using experimental property values, incorporating elemental diversity in the sampling strategy (COMP+X) results in initial datasets that are more diverse in output space.

measured property space. These results reinforce the finding that augmenting the input space with domain knowledge—such as subsystem complexity—can lead to more representative and effective initializations, even in data-limited experimental regimes.

3.5 Generalizing feature augmentation strategies

The preceding discussion focused on the use of alloy complexity—quantified by the number of constituent elements—as an augmenting dimension for improving diversity in initialization. Although subsystem complexity served as a natural and chemically interpretable feature in this context, the underlying approach is more general. In principle, any domain-knowledge-informed attribute of the materials space can be incorporated as an additional dimension to guide sampling. Although we specifically employed a binary encoding of elemental presence, this feature could be replaced or complemented by other expert-curated properties, such as electronic structure descriptors, processing constraints, or thermodynamic heuristics. The central idea is to enrich the input representation with a physically or chemically meaningful structure, enabling more effective and extensible initialization strategies across a wide range of optimization and discovery workflows.

To illustrate this concept, we examined a subset of the Al–V–Cr–Fe–Co–Ni alloy composition space, explicitly excluding vanadium to focus on alloys containing the remaining five elements. In this example, we incorporate Shannon's configurational entropy, defined as $S_{\mathrm{conf}} = -k_{\mathrm{B}}\sum_{i=1} x_i \ln(x_i)$—as an additional dimension in the clustering process. Since configurational entropy is a convex and nonlinear function of composition, uniform sampling across the input space yields a non-uniform, typically skewed distribution in entropy space. In the context of alloy discovery, one may wish to investigate systems of differing structural or thermo-dynamic complexity as quantified by configurational entropy.

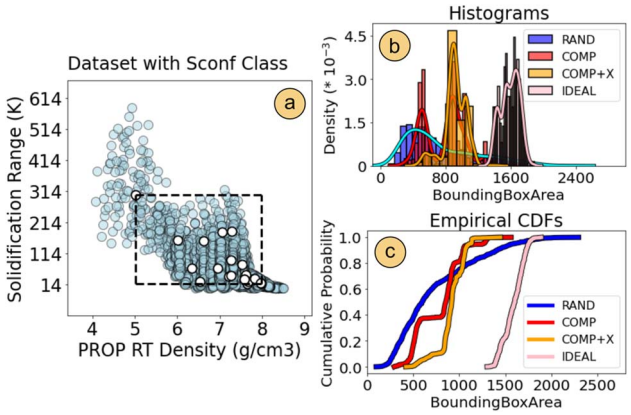

Fig. 8 Example analysis of a subset of the Al–V–Cr–Fe–Co–Ni composition space excluding vanadium, restricted to compositions containing nonzero amounts of the remaining five elements. (a) Output space visualization showing solidification range *versus* room-temperature density, with an example subset selected using an augmented feature space that includes a configurational entropy class. (b) Histogram and (c) empirical CDF of bounding box areas for 1000 such subsets. Even when using a coarse classification such as "high entropy" *versus* "low entropy," incorporating this additional dimension leads to more consistent initialization sets with broader coverage of material properties, improving the diversity of starting conditions for optimization.

To ensure representative sampling across this dimension, we introduced a binary entropy class (low *vs.* high entropy) as an auxiliary input feature.

The FCC alloy space, shown in Fig. 1, was filtered to its vanadium-free subset, yielding 3842 CALPHAD-feasible compositions. The predicted values for the solidification range and the room temperature density are used as representative outputs in Fig. 8 (left). We selected the Shannon entropy of an equiatomic quaternary alloy ($\sim$1.386 k_B) as the threshold to classify alloys as low or high entropy. This binary classification was appended to the input space and used in *k*-medoids clustering. As shown in Fig. 8 (right), augmenting the input representation with entropy class consistently improved property-space coverage across 1000 initialization datasets, compared to clustering on composition alone. This result highlights how even a coarse and physically meaningful classification—when used as an additional feature—can enhance the diversity of initializing populations and improve the quality of early-stage exploration.

3.6 Impact of initialization on optimization outcomes

Given appropriate domain knowledge to augment the feature space, such information can be readily integrated into the Bayesian optimization (BO) framework. Following the methodology outlined in Section 2*: Methods*, we simulated two initialization strategies: one based on clustering over compositional input features (**COMP**), and another using clustering augmented with a descriptor that encodes the identity of nonzero elemental constituents (**COMP+X**). The **COMP** strategy reflects conventional BO practice, where diversity is assumed to arise from compositional variation alone.

To evaluate performance, 50 independent optimization campaigns were conducted for each strategy using the FCC dataset introduced in Fig. 1 and discussed in Section 3.1*: Variance within optimizations*. Each campaign aimed to simultaneously maximize room-temperature density and heat capacity in the Al–V–Cr–Fe–Co–Ni system. Campaigns proceeded over 10 iterations with a batch size of 20, including the initial seed. Each iteration used an ensemble of 1000 Gaussian process regressors (GPRs), each with randomized length scale hyperparameters and acquisition based on expected hypervolume improvement (EHVI), with the final batch selected *via k*-medoids clustering.[35]

Prior to simulation, we anticipated that **COMP+X** would produce higher initial hypervolumes on average, consistent with earlier observations. While hypervolume (HV) was not a direct clustering target, the larger average bounding box area observed in **COMP+X** initializations inherently implies broader coverage of the property space. The simulations thus evaluate whether this initialization advantage consistently improves the optimization performance across surrogate model realizations.

The results are shown in Fig. 9. For each strategy, the mean hypervolume achieved per iteration in all 50 optimization runs is plotted: **COMP+X** in cyan and **COMP** in green. Shaded bands denote the 10th to 90th percentile range. The influence of initialization is evident from the first iteration: 96% of **COMP+X** seed hypervolumes exceed the mean seed HV of the **COMP** runs and 60% surpass the **COMP** 90th percentile. These results indicate that augmenting the input space with subsystem

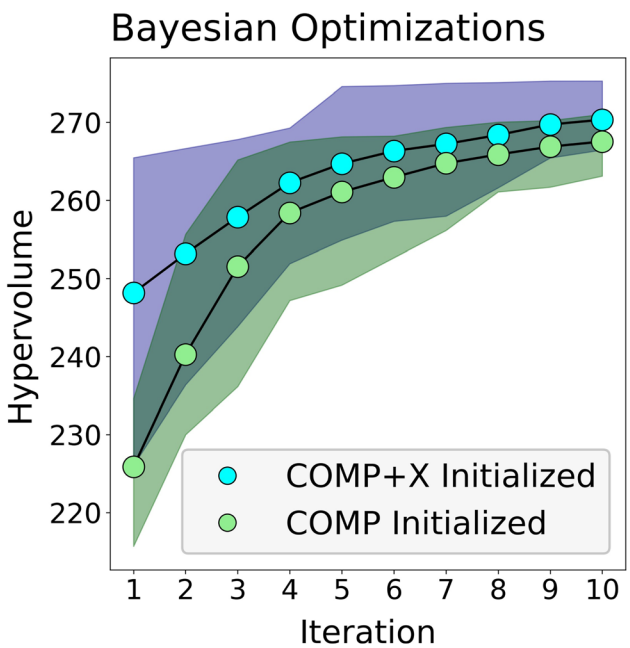

Fig. 9 Comparison of two Bayesian optimization campaigns targeting FCC alloy properties, differing only in their initialization strategy. The COMP strategy specifies diversity based solely on the compositional input dimensions, while the COMP+X strategy incorporates an additional input dimension representing the specific combination of elements present. Both optimization strategies were repeated 50 times using identical surrogate models and acquisition settings.

complexity yields significantly more diverse and higher-performing initial datasets. This early advantage persists throughout the optimization process. The **COMP+X** strategy maintains a consistent lead in hypervolume across iterations, reflecting a durable performance advantage from improved initialization. Although both strategies may con-verge to similar Pareto fronts in the asymptotic limit, the superior starting conditions provided by **COMP+X** facilitate more effective early stage exploration, a central objective of batch BO frameworks operating under resource constraints. The graph in Fig. 9 follows a standard format commonly used in comparative optimization studies, such as those evaluating different acquisition functions or kernel choices. Error bars are intentionally omitted, as they misrepresent statistical uncertainty in this context. In optimization problems, each iteration is inherently dependent on the previous one, and metrics such as hypervolume are strictly non-decreasing by definition. Moreover, hypervolume is bounded above by the value corresponding to full discovery of the Pareto front. Consequently, standard deviations or standard errors—which assume independent and identically distributed (i.i.d.) samples—are not meaningful when overlaid on such trajectories. To complement the analysis of average hypervolume trajectories, we further examine the results of individual optimization runs. This provides insight into the variability and robustness of each strategy and goes beyond conventional summaries by highlighting differences in the consistency and efficacy of the policies across repeated simulations.

The optimization scheme described in Section 2*: Methods* contains multiple sources of variation. Length-scale hyperparameters, selected *via* Latin hypercube sampling, are intentionally randomized in the BBO framework to generate diverse candidate sets across surrogate models without imposing a particular prior. Likewise, as discussed extensively in this work, the initial dataset is selected *via* a policy that incorporates random seeding. By allowing both the initial dataset and the Gaussian process hyperparameters to vary, we account for the inherent stochasticity of BO campaigns. For a sufficiently large number of runs, this variability ensures statistical robustness when comparing competing initialization strategies.

To systematically compare two policies, we proceed as follows:

(1) For each optimization under policy #1, record the value of the performance metric—hypervolume (HV)—at every iteration.

(2) For each optimization under policy #2, record the same metric at the same iteration indices.

(3) Compare the HV of the individual policy #1 optimization to each of the policy #2 runs at each iteration, and compute the fraction of cases where it dominates.

(4) Repeat this procedure for all policy #1 optimizations, aggregating results into a mean dominance percentage with percentile bands.

(5) Reverse the comparison: evaluate each policy #2 optimization against all policy #1 runs using the same procedure.

If policy #1 consistently yields better performance than policy #2, then steps (1)–(4) should produce high dominance percentages, while step (5) should produce noticeably lower ones.

Previous analyses already established this dominance at the first iteration. As shown in Fig. 5, the **COMP+X** initialization strategy reliably yields higher HV in the seed dataset compared to **COMP**. Although previous results focused on the bounding box area and the mean distance to the centroid, these geometric metrics strongly correlate with initial hypervolume. What remains to be determined is the persistence of this advantage across iterations and its cumulative effect on overall optimization performance. These results are summarized in Fig. 10.

Each set of policies' means are shown as scatter points on top of middle %ile bands from 25% to 75%. As expected, the first iteration of Fig. 10 matches the prior histograms. In fact, the **COMP+X** policy is so consistent in this regard that its mean HV value lies outside the inner 50%ile for the first iteration; out of the 50 datasets from the **COMP+X** policy that dominate the **COMP** policy, the vast majority of values are at or near 100%. Continuing onward, only occasionally does variance of an individual simulation cause a **COMP+X** policy to perform worse than standard initialization practices. Conversely, the **COMP** policies are often near 0% domination of the **COMP+X** policies; their non-zero values arise largely from occasional particular iterations that have a large jump in HV (or where they have gotten "lucky"). Fig. 10 puts into context the likelihood of success (given a set of algorithmic changes), which is difficult to parse with "HV *vs.* iteration" visuals alone (*i.e.* Fig. 9).

The findings presented here reinforce and extend the conclusions of Maaranen *et al.*,[20] underscoring the critical

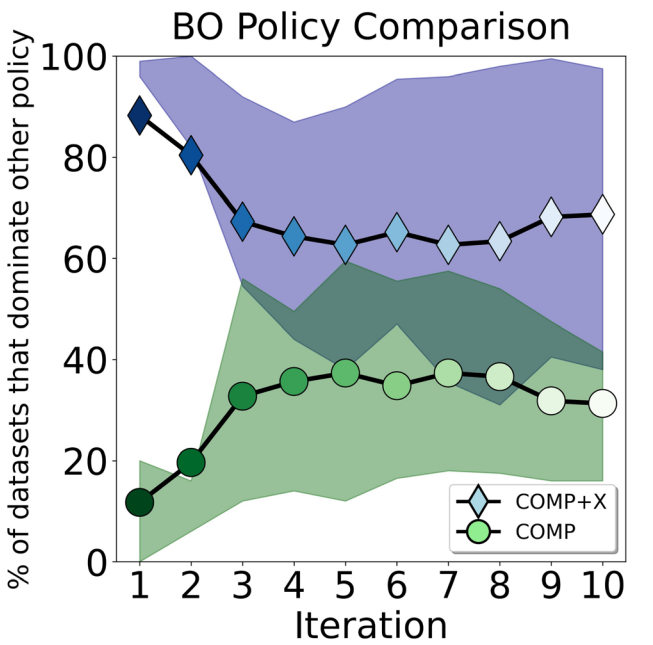

Fig. 10 Comparison of two BBO policies (**COMP+X** and **COMP**) targeting FCC alloy properties, differing only in their initialization strategy. Each optimization's hypervolume (HV) from one policy is compared to the HVs of all optimizations from the opposing policy, across all iterations and simulation runs. Mean values are indicated by scatter points, and the central 25–75% percentile bands are shown in the background. This comparison accounts for both sources of randomness: variation in the initial dataset and stochasticity in GP length scale hyperparameters.

impact of initial dataset selection on optimization performance. Our results show that incorporating expert-defined descriptors into the sampling process yields markedly improved initialization strategies. These informed approaches mitigate the biases inherent in uniform or geometry-based input sampling, leading to more representative coverage of the design space and more efficient exploration of complex materials landscapes. The principal conclusions are summarized as follows:

- Random sampling is inadequate for initializing optimization campaigns and frequently degrades downstream performance, regardless of the method applied.
- Sampling strategies that define diversity solely in terms of geometric input space (*e.g.*, compositional distance) offer little to no improvement over random sampling in terms of output-space coverage.
- Input spaces can and should be augmented with derived or domain-informed features. This work shows that incorporating classification-like descriptors (*e.g.*, subsystem identity or configurational entropy class) leads to more representative and effective initial datasets.

Importantly, these insights are broadly applicable and largely independent of specific algorithmic choices, including surrogate model type, kernel function, acquisition strategy, or batch selection policy. More generally, embedding domain knowledge into input representations provides a flexible and effective means of improving materials optimization workflows, particularly in systems characterized by complex or sparsely sampled design spaces. Future work may further integrate such descriptors into the optimization loop itself—for example, by informing GP kernel length scales or shaping acquisition behavior during batch selection.

4 Conclusions

This study demonstrates that the breadth of material properties first observed is coupled with the strategy for initial dataset selection. Such initializations play an underappreciated role in determining the efficiency of materials discovery efforts, such as those driven by Bayesian optimization (BO). Although BO is widely regarded as a powerful tool for exploring expensive high-dimensional design spaces, its performance is strongly conditioned on the quality and representativeness of the initial data set. We show that conventional initialization strategies—such as random sampling or geometry-based clustering in compositional space—often fail to produce a diverse or informative initial survey, limiting the optimizer's ability to explore the design space effectively.

While this work leverages a BO framework, the novelty primarily lies in the initialization strategy and integration of chemically meaningful descriptors (such as subsystem complexity and configurational entropy). Across a variety of simulated and experimental alloy systems, we find that augmenting the input space with these expert-derived dimensions leads to significantly more diverse initial datasets and improved optimization trajectories. Importantly, these gains are observed without changing the underlying BO engine, acquisition strategy, or surrogate model. Future enhancements of these strategies may utilize the properties of select alloys, such as predicted microstructural elements for processing conditions.

Author contributions

Trevor Hastings: conceptualization, data curation, formal analysis, methodology, writing – original draft. James Paramore: supervision, Brady Butler: supervision, Raymundo Arróyave: supervision, funding acquisition, writing – review & editing.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

The code and datasets used in this project are archived with Zenodo *via* **https://doi.org/10.5281/zenodo.17635456**. The corresponding GitHub repository is available at **https://github.com/trevorhastings/Initialization**.

Supplementary information (SI): the Bayesian methods, a data clustering representation, additional refractory data, and a diagram of sampling strategies. See DOI: **https://doi.org/10.1039/d5dd00361j**.

Acknowledgements

This work was a tertiary investigation under experimental campaigns funded by the United States Army Futures Command University Technology Development Division (UTDD) *via* contract number W911NF-22-F-0032. Cooperative Agreement Number W911NF-22-2-0106 with the Army Research Laboratory is also acknowledged for supporting aspects of this work.

References

1 C. Yang, *et al.*, A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness, *Acta Mater.*, 2022, **222**, 117431.
2 Y. Liu, J. Wang, B. Xiao and J. Shu, Accelerated development of hard high-entropy alloys with data-driven high-throughput experiments, *J. Mater. Inf.*, 2022, **2**, 3.
3 C. Wen, *et al.*, Machine learning assisted design of high entropy alloys with desired property, *Acta Mater.*, 2019, **170**, 109–117.
4 Z. Y. Rao, H. Springer, D. Ponge and Z. M. Li, Combinatorial development of multicomponent Invar alloys via rapid alloy prototyping, *Materialia*, 2022, **21**, 101326.
5 O. Mamun, M. Bause and B. S. M. Ebna Hai, Accelerated development of multi-component alloys in discrete design space using Bayesian multi-objective optimisation, *Mach. Learn.: Sci. Technol.*, 2025, **6**, 015001, DOI: **10.1088/2632-2153/ada47d**.
6 S. Hu, H. Wang, Z. Dai, B. K. H. Low and S. H. Ng, Adjusted Expected Improvement for Cumulative Regret Minimization

in Noisy Bayesian Optimization, *arXiv*, 2024, preprint, arXiv:2205.04901 [cs], DOI: 10.48550/arXiv.2205.04901.
7 K. Swersky, J. Snoek and R. P. Adams Freeze-Thaw Bayesian Optimization, *arXiv*, 2014, preprint, arXiv:1406.3896 [stat], DOI: 10.48550/arXiv.1406.3896.
8 Z. Del Rosario, M. Rupp, Y. Kim, E. Antono and J. Ling, Assessing the frontier: Active learning, model accuracy, and multi-objective candidate discovery and optimization, *J. Chem. Phys.*, 2020, **153**, 024112.
9 K. M. Jablonka, G. M. Jothiappan, S. F. Wang, B. Smit and B. Yoo, Bias free multiobjective active learning for materials design and discovery, *Nat. Commun.*, 2021, **12**, 2312.
10 A. M. Gopakumar, P. V. Balachandran, D. Z. Xue, J. E. Gubernatis and T. Lookman, Multi-objective Optimization for Materials Discovery via Adaptive Design, *Sci. Rep.*, 2018, **8**, 3738.
11 P. I. Frazier, A tutorial on Bayesian optimization, *arXiv*, 2018, preprint, arXiv:1807.02811, DOI: 10.48550/arXiv.1807.02811.
12 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, Taking the Human Out of the Loop: A Review of Bayesian Optimization, *Proc. IEEE*, 2016, **104**, 148–175.
13 R. Arróyave, *et al.*, A perspective on Bayesian methods applied to materials discovery and design, *MRS Commun.*, 2022, **12**, 1037–1049, DOI: 10.1557/s43579-022-00288-0.
14 D. Khatamsaz, *et al.*, Bayesian optimization with active learning of design constraints using an entropy-based approach, *npj Comput. Mater.*, 2023, **9**, 49.
15 C. K. Williams and C. E. Rasmussen *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006, vol. 3.
16 J. Mockus On Bayesian methods for seeking the extremum in *Optimization Techniques IFIP Technical Conference Novosibirsk*, ed. G. Goos, et al., Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 1975, vol. 27, pp. 400–404.
17 R. Couperthwaite, *et al.*, Materials Design Through Batch Bayesian Optimization with Multisource Information Fusion, *JOM*, 2020, 1–13.
18 C. Acemi, *et al.*, Multi-objective, multi-constraint high-throughput design, synthesis, and characterization of tungsten-containing refractory multi-principal element alloys, *Acta Mater.*, 2024, **281**, 120379.
19 T. Hastings, *et al.*, Accelerated Multi-Objective Alloy Discovery through Efficient Bayesian Methods: Application to the FCC High Entropy Alloy Space, *Acta Mater.*, 2025, 121173.
20 H. Maaranen, K. Miettinen and A. Penttinen, On initial populations of a genetic algorithm for continuous optimization problems, *J. Global Optim.*, 2007, **37**, 405–436, DOI: 10.1007/s10898-006-9056-6.
21 B. N. Slautin, Y. Liu, H. Funakubo and S. V. Kalinin, Unraveling the impact of initial choices and in-loop interventions on learning dynamics in autonomous scanning probe microscopy, *J. Appl. Phys.*, 2024, **135**, 154901, DOI: 10.1063/5.0198316.
22 A. Paszke, *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library, *arXiv*, 2019, preprint, arXiv:1912.01703 [cs], DOI: 10.48550/arXiv.1912.01703.
23 F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python, *arXiv*, 2018, preprint, arXiv:1201.0490 [cs], DOI: 10.48550/arXiv.1201.0490.
24 T. Cinquin, *et al.*, What Actually Matters for Materials Discovery: Pitfalls and Recommendations in Bayesian Optimization, in *AI for Accelerated Materials Design-ICLR 2025*, 2025.
25 J. Paramore, *et al.*, Two-Shot Optimization of Compositionally Complex Refractory Alloys, in *SSRN Scholarly Paper*, Rochester, NY, 2024, https://papers.ssrn.com/abstract=5000547.
26 B. Vela, T. Hastings and R. Arróyave, Visualizing High Entropy Alloy Spaces: Methods and Best Practices, *arXiv*, 2024, preprint, arXiv:2408.07681 [cond-mat, physics:physics], DOI: 10.48550/arXiv.2408.07681.
27 B. Harrington, *Inkscape: Open Source Scalable Vector Graphics Editor*, 2004, http://www.inkscape.org/.
28 H.-S. Park and C.-H. Jun, A simple and fast algorithm for K-medoids clustering, *Expert Syst. Appl.*, 2009, **36**, 3336–3341.
29 O. Zobac, A. Kroupa, A. Zemanova and K. W. Richter, Experimental description of the Al-Cu binary phase diagram, *Metall. Mater. Trans. A*, 2019, **50**, 3805–3815.
30 N. Bostrom, *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, Routledge, New York, 2013.
31 J. B. Hartle and M. Srednicki, Are We Typical?, *Phys. Rev. D*, 2007, **75**, 123523.
32 S. M. Carroll, Why Boltzmann Brains Are Bad, *arXiv*, 2017, preprint, arXiv:1702.00850 [hep-th], DOI: 10.48550/arXiv.1702.00850.
33 M. P. Brady, *et al.*, The development of alumina-forming austenitic stainless steels for high-temperature structural use, *JOM*, 2008, **60**, 12–18.
34 B. Vela, *et al.*, High-throughput exploration of the WMoV-TaNbAl refractory multi-principal-element alloys under multiple-property constraints, *Acta Mater.*, 2023, **248**, 118784.
35 G. Zhao, R. Arroyave and X. Qian, *Fast Exact Computation of Expected HyperVolume Improvement*, 2018.