Probabilistic Deep Learning for Simultaneous Multi-Objective Design of High-Entropy Alloys

# Pointers

1. Finding chemistries that have the trade-off with four properties.
2. Challenges related to the dataset and how the approach addresses these challe

# Introduction

What are HEAs used for?

This work is built on the work by Giles et al. [1] on predicting and optimizing the yield strength of high-entropy alloys (HEAs) using machine learning. This work utilizes a range of material descriptors, including atomic composition, lattice distortion, and temperature-dependent properties, to guide the design of HEAs with enhanced mechanical properties. Furthermore, the paper maximizes the yield strength at a specified temperature for a given base high-entropy alloy (HEA) using Differential Evolution as an optimizer. The optimizer iteratively explores different combinations of element fractions within fixed bounds, aiming to find the composition that results in the highest yield strength. The objective function used for optimization is defined as the reciprocal of yield strength, encouraging the algorithm to seek compositions that maximize this property. Linear constraints are applied to ensure that the sum of element fractions equals 1, and the optimization process continues until a composition with the desired yield strength is found or a specified termination criterion is met.

The drawback of this method is that it doesn’t consider the simultaneous optimization of multiple material properties, such as ductility, and hardness. This limitation restricts the applicability of the method in scenarios involving multifunctional material design, where a holistic consideration of multiple properties is essential for achieving optimal material performance. Thus, there is a need for simultaneous optimization of multiple material properties that enables exploring the trade-offs between different properties meeting the requirements of multifunctional material design.

Designing High Entropy Alloys (HEAs) presents a plethora of challenges arising from the high-dimensional chemical complexity of these materials. Multicomponent alloys involve large number of elements across various compositions and processing conditions, leading to an intricate design space. While data used for ML predictions need to be homogenous, sizable, and representative of the entire HEA space, obtaining such data is often hindered by heterogeneity and the need for extensive sampling. Hence, predictive models may struggle to provide accurate insights or recommendations for these unexplored element spaces. Moreover, ML model confidence is influenced by data quality and quantity, with sparse regions in the design space resulting in uncertainties.

## Design under Uncertainty

Methods used in the literature for tackling data uncertainty:

1. Dempster-Shafer Theory
2. Fuzzy Sets
3. Interval Methods
4. Bayesian Methods
5. Information Theory
6. Probability Boxes
7. Sampling-based Monte Carlo method

Bayesian methods excel in managing data uncertainty by integrating prior knowledge with observed data, providing probabilistic estimates that reflect both parameter and predictive uncertainty.

Partially Sampled Space that doesn’t fully represent the full sample space of elemental composition and synthesis conditions.

## Methodology

We have formulated a computational framework for multi-property optimization of material properties using chemical composition and various synthesis parameters as input descriptors. Our framework leverages the use of Bayesian Neural Network BNN with Variational Inference (BNN-VI) and Non-dominated Sorting to tackle the challenges of designing High Entropy Alloys (HEAs). By employing BNNs, we introduce a probabilistic framework that can model the uncertainty inherent in HEA data due to its high-dimensional and heterogeneous nature. In our approach, we start by developing Bayesian Neural Network (BNN) models to predict three key mechanical properties of High Entropy Alloys (HEAs) such as Yield Stress, Ductility and Hardness. Subsequently, we employ these models to predict the properties for new and unexplored alloy compositions, generating synthetic data to expand our understanding of HEA properties across a broader compositional space for the given material space. This synthetic dataset serves as a crucial asset for optimizing material properties. We employ probabilistic regression model to make uncertainty-based property predictions for the synthetically generated alloys. To assess and optimize the performance of these materials holistically, we leverage the concept of non-dominated sorting, which allows us to identify a Pareto front from the properties of the synthetic dataset. This computational pipeline enables us to uncover promising HEAs with optimal combinations of mechanical properties, overcoming the challenge of high-dimensionality, data scarcity and complex and unexplored design spaces. We screen for HEA with optimal chemical composition and synthesis condition that meet multiple design criteria for performance.



Figure :Computational Framework for HEA Simultaneous Property Optimization

# Results

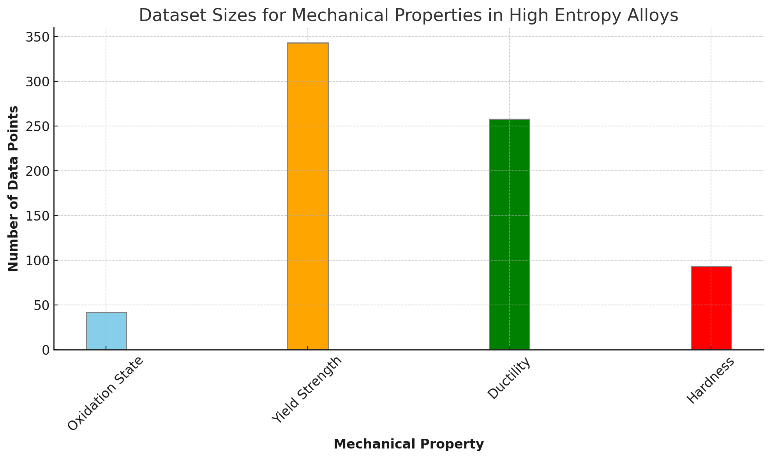
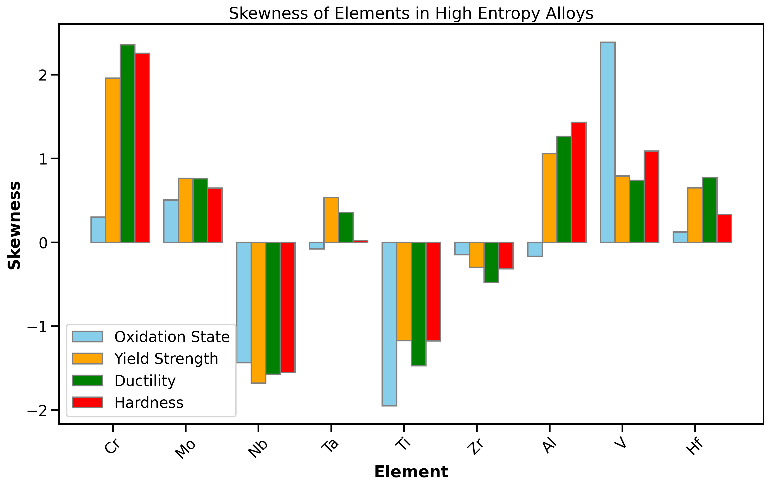
The overall dataset on HEAs contain 103 different elements. However, not all columns are filled. Including such a sparse data can interfere with modeling the property and can produce higher prediction uncertainty. To mitigate this, we have filtered out the elements with less than 50 non-zero compostion fraction, and normalizes the remaining data to ensure that the compositions sum to 1 for each alloy. This data preprocessing is done to focus on the most relevant and consistent set of elements and to ensure that the compositions are on a consistent scale for analysis. After filtering, there are total 9 elements considered namely, Al, Ti, V, Cr, Zr, Nb, Mo, Hf, Ta. These elements, along with their chemical compositions, were integral as input descriptors in our analysis.

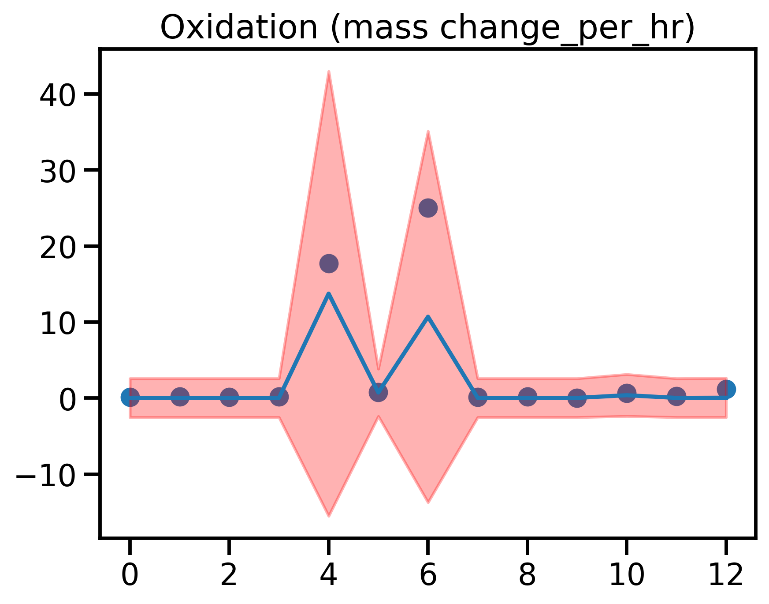
Furthermore, we incorporated seven synthesis conditions as additional input descriptors. These conditions include Test Temperature (Celsius), Strain Rate (s^-1), Arc Melting, HIP Time (hours), Anneal Temp (Celsius), Anneal Time (hours), and Anneal Cooling Rate (degC/min). The inclusion of Phase type as an input descriptor further enriched our dataset. This comprehensive approach, combining elemental composition, synthesis conditions, and phase types, allowed for an extensive exploration of the high-dimensional chemical space, crucial for discovering new alloy compositions with tailored properties.

Our research also extended to a second dataset aimed at predicting the Oxidation State (mass change/hr.). This dataset utilized the same elemental composition space and included synthesis conditions like Initial condition, Temperature, and Phase Type as inputs.

Figure 2 highlights the disparities in dataset sizes and provides a skewness overview, enabling a better understanding of the underlying distribution of mechanical properties in HEAs. The analysis of skewness and dataset size for the mechanical properties of High Entropy Alloys (HEAs) shows varying degrees of asymmetry and data sufficiency. The Oxidation State and Hardness exhibit significant positive skewness, suggesting the presence of high-value outliers and potential non-normative distribution tail characteristics, which, coupled with their relatively smaller dataset sizes of 42 and 93 data points respectively, introduce uncertainty into the dataset. Yield Strength, with a substantial dataset size of 343 and a moderate positive skewness, presents a lower level of uncertainty, indicating a more symmetric and representative distribution of values that could lend to more confident model predictions. Ductility, despite a reasonable dataset size of 258, displays negative skewness, highlighting the need for cautious interpretation of lower-end values. These insights underline the necessity for a predictive modeling approach that can accommodate dataset imperfections and provide probabilistic understandings of prediction uncertainties. Creating a unified approach with diverse skewness in data that optimizes material properties simultaneously is a challenge. Given the uncertainty existing in the training dataset, employing Bayesian Neural Networks with Variational Inference (BNN-VI) is a mathematically feasible approach to mitigate the challenges of modeling mechanical properties of HEAs. By incorporating Variational Inference, BNN-VI allows for the approximation of complex posterior distributions over model parameters.

Figure : Comparative Visualization of Dataset Sizes and Skewness in High Entropy Alloys (HEAs). (a) Bar chart representing the number of training and testing data points available for each mechanical property: Oxidation State (mass change/hr), Yield Strength (MPa), Ductility (%), and Hardness (HV). (b) Grouped bar plot illustrating the skewness of the same mechanical properties across various elements commonly found in HEAs. Each cluster of bars represents an element, while individual bars within the group are color-coded to match the properties in part (a)



A graph of graph with blue and red lines

Description automatically generatedA graph of ductility

Description automatically generatedA graph of blue and pink lines

Description automatically generated

Figure : Visualization of prediction and associated uncertainty made by our proposed Bayesian Network regression model on the testing set of HEAs.The shaded region (red) visualizes a 95% confidence intervals around central predictions in blue line. The blue dots represent the observed values of the material properties.

Separate BNN models were fitted for each mechanical property. Figure 3 illustrates the evaluation and visualization of predictions made by our BNN model, including the prediction uncertainty for each property. We adopted a 95% confidence interval around the mean predictions, represented as a shaded region in our plots, to indicate the likely range of true material properties considering the model uncertainty.

Results:

1. Individual plots
2. Test values, mean, std for BNN
3. Test, observed for regular

Yield Stress:

Oxidation state:

PICP: 1.0, MPIW: 12.701

The Bayesian Neural Network (BNN) showcases a robust predictive capacity, as evidenced by an R2 of 0.7136, indicating it accounts for approximately 71.36% of the variance in the target variable. While the Mean Squared Error (MSE) and Mean Absolute Error (MAE) stand at 19.0224 and 1.6251 respectively, the true strength of the BNN lies in its probabilistic approach. The mean predictions closely align with the observed data, as the BNN mean prediction for an observed value of 17.7 is 13.734468, and for 25.0, it is 10.69452, illustrating the model's effectiveness in capturing the trend in data. The confidence intervals, ranging from [-2.5339284 to 2.578587] for smaller values and expanding to [-15.510445 to 42.97938] for larger observed values, reflect the model's calibrated uncertainty estimates.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Properties | Bayesian Regression | | | Regular Regression | | |
|  | R^2 | MSE | MAE | R^2 | MSE | MAE |
| Yield Stress | 0.7036 | 62307.96 | 192.460 | 0.7053 | 61960.43 | 178.707 |
| Ductility |  |  |  |  |  |  |
| Hardness |  |  |  |  |  |  |
| Oxidation State | 0.7136 | 17.0713 | 1.592 | 0.6808 | 19.0224 | 1.625 |

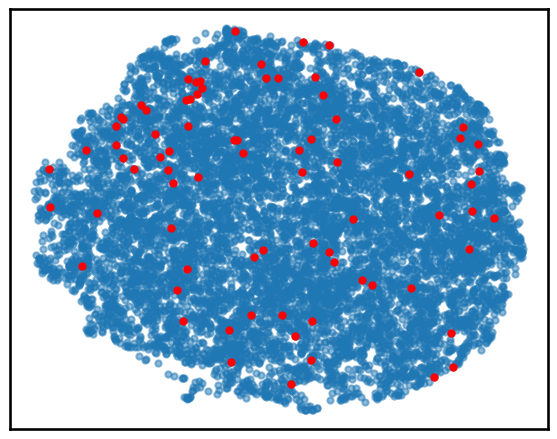


Figure : UMAP projection of chemical composition of synthetically generated alloys for the same elemental space. The red dots are the HEAs that have simultaneously optimized properties.

Figure : Pairwise scatterplot for Ductility, Hardness and Yield Stress for all the HEAs identified on the Pareto front

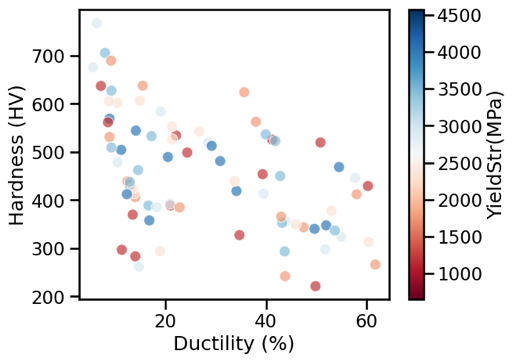
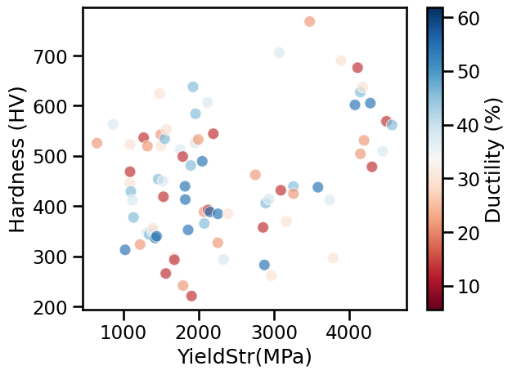
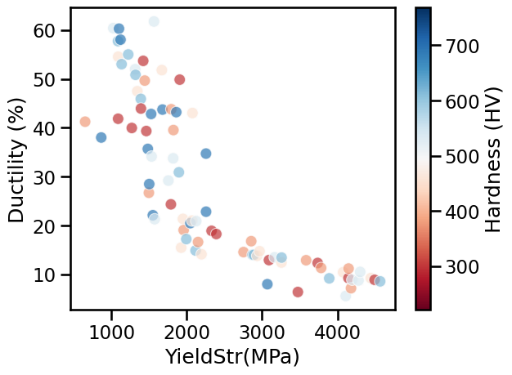
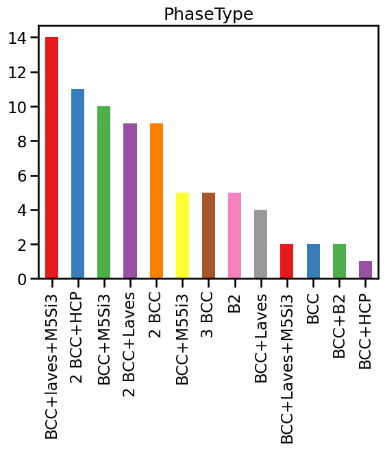
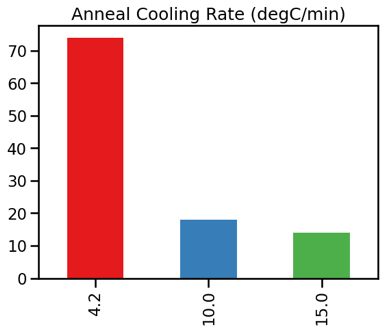
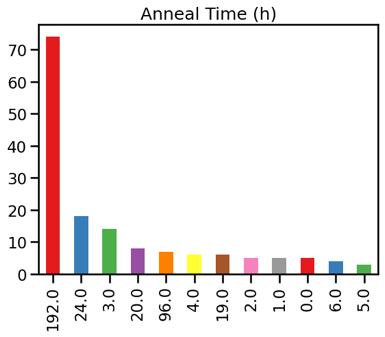
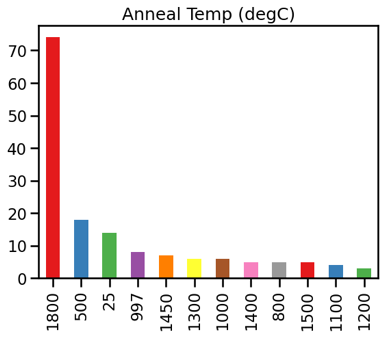
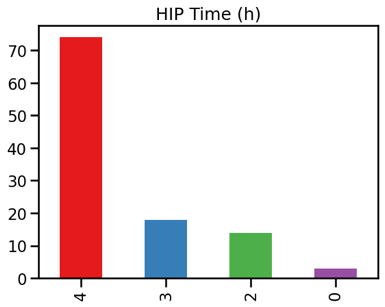
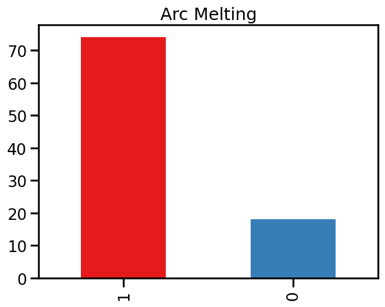
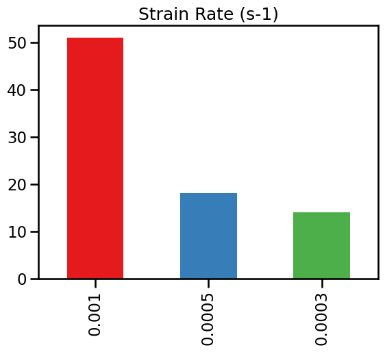
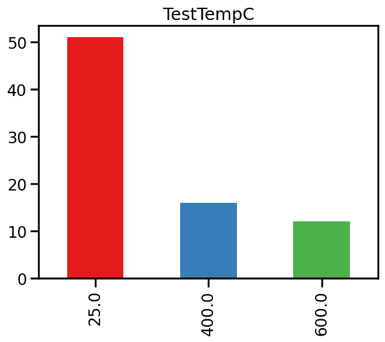


Figure : Value counts for different synthesis conditions for the HEAs that are identified on the Pareto Front



# Method

We utilized two distinct datasets to predict various properties of High Entropy Alloys (HEAs). The first dataset focused on the chemical compositions of the alloys and included a range of synthesis conditions, such as Test Temperature (Celsius), Strain Rate (s^-1), Arc Melting, HIP Time (hours), Anneal Temp (Celsius), Anneal Time (hours), Anneal Cooling Rate (degC/min), and Phase Type. This dataset was instrumental in modeling three key mechanical properties: Ductility (%), Yield Stress (MPa), and Hardness (HV).

The second dataset, while also encompassing the chemical composition of the alloys, differed in its focus on synthesis conditions, specifically targeting Phase Type, Initial condition, and Temperature. This dataset was primarily used to predict the oxidation state, measured as mass change per hour.

Let us consider HEAs to be denoted as , with each HEA as , to , and the target properties Yield Stress as , Hardness as and Ductility as respectively. The input to our machine learning model consists of two types of information for a given HEA, the chemical composition and the synthesis parameters. Let be a vector representing the chemical composition of an alloy, where denotes the concentration (or fraction) of the -th elemental component in the alloy. The composition vector can be represented as where and . The first constraint ensures that the concentration of each element must be a non-negative value less than or equal to 1, indicating a valid and physically meaningful composition. The second constraint ensures that the sum of all elemental concentrations in an alloy must equal 1.

## Bayesian Neural Network

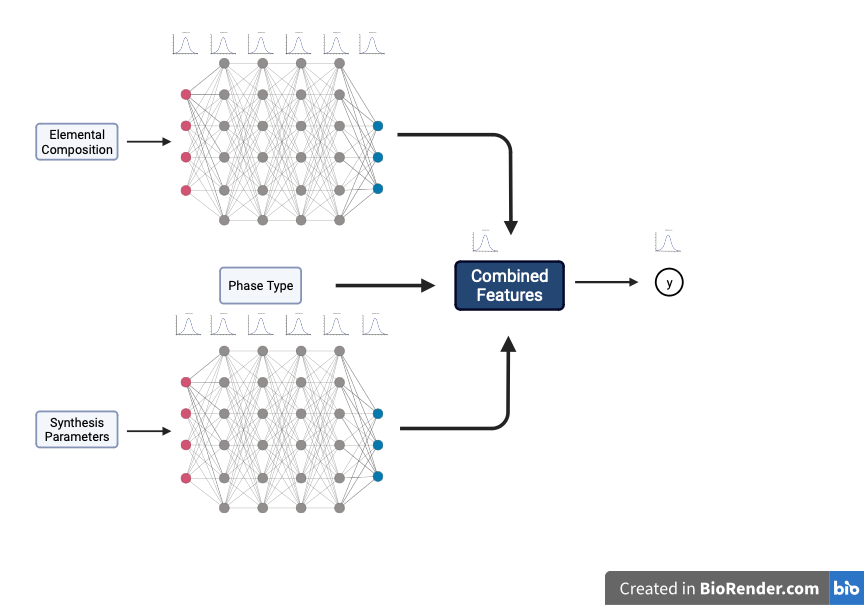


Figure : Architecture of Bayesian Neural Network with probabilistic weights and an output mechanical property.

As noted previously, each dataset is quite small and lacks representation. Additionally, we have different size of dataset available for different property. Thus, we have used four different Bayesian Neural Networks (BNNs) for each property. These models take into account prior knowledge in data uncertainty and iteratively updates itself with the availability of the training data.

We have used a Bayesian Neural Network with Variational Inference (BNN-VI) as a modeling approach to model the mechanical properties, given the chemical composition and the synthesis parameters. A normal regression model assumes a form assuming a deterministic parameter . A Bayesian Neural Network (BNN) assumes a form where is treated as a random variable following a posterior distribution given the data . Following Bayes’ theorem, the posterior distribution is proportional to the prior distribution as . Thus, the posterior predictive distribution for a given mechanical property , is given as,

The Bayesian Neural Network (BNN) architecture used in this study is designed to capture the underlying patterns and uncertainties in the data. It comprises several probabilistic layers, each with a specific mathematical role. The input layer, tailored to the problem's dimensionality, serves as the entry point for data. The model architecture shows separate layers for latent space exploration for the chemical composition and synthesis parameters. Following the input layers, we introduce three hidden probabilistic dense layers, each equipped with 20 neurons and employing the rectified linear unit (ReLU) activation function. These hidden layers facilitate non-linear transformations, enhancing the network's capacity to represent intricate relationships within the data. Importantly, the weights of these neurons are probabilistic,

That is, we assume a Gaussian prior for each weight having a mean and variance . The pivotal component of the architecture is the use of probabilistic dense layer, which generates probabilistic predictions. This layer produces two essential parameters: the mean and the variance of a Gaussian distribution. These parameters characterize the mean and covariance of a target material property. The selection of 20 neurons in each hidden layer and the depth of the architecture is based on empirical experimentation and can be further fine-tuned according to the specific dataset and problem domain. The primary objective is to capture intricate, non-linear data patterns while quantifying predictive uncertainty.

To solve for the posterior distribution, we assume a variational distribution that approximates the true posterior . The variational distribution is a known functional form for whose parameter we wish to estimate by minimizing the Kullback-Leibler divergence between and the true posterior .

## Synthetic Alloy Generation

In this subsection we outline the steps to generate synthetic alloy compositions for the purpose of expanding the scope of material property prediction to unexplored regions of the compositional space and thereby predicting the material properties. To create synthetic alloy compositions, a matrix of random values is generated with shape matching the desired number of synthetic alloys and the number of elemental components in the dataset. These random values are then normalized to ensure that the resulting compositions sum to unity, adhering to the constraint of a valid alloy composition. The composition generation process allows for the exploration of diverse alloy compositions across the high-dimensional chemical space. In addition to alloy composition, we have artificially generated synthesis conditions adhering to the list of conditions in the primary dataset. For each synthetic alloy, the synthesis conditions are randomly sampled from the set of conditions observed in the original dataset. This introduces variability in the synthesis process, reflecting the real-world situations where material properties can vary based on processing parameters.

## Non-dominated Sorting

In this section, we describe the non-dominated sorting for identifying the Pareto front from the list of material properties for synthetically generated alloys. We represent the set , to as the collection of synthetically generated alloys with different elemental composition values and synthesis condition. For each synthetic alloy, we have estimated the probability distribution of the target material property such as Yield Stress ­, Hardness and Ductility .

We have taken the mean of these three distributions namely as our candidate for performing non-dominated sorting. Given the set as the set of synthetic alloys, our goal is to identify the Pareto front, which consists of non-dominated solutions. Non-dominated solutions are those that are not dominated by any other solution in the dataset. Mathematically, an alloy with mean mechanical properties as is said to dominate an alloy with mean mechanical properties as if:

or or

and

or or

# Bibliography

|  |  |
| --- | --- |
| [1] | S. A. Giles, D. Sengupta, S. R. Broderick and K. Rajan, "Machine-learning-based intelligent framework for discovering refractory high-entropy alloys with improved high-temperature yield strength," *npj Computational Materials,* vol. 8, November 2022. |

# Appendix

