Supplementary Information

Attention-Enhanced Variational Learning for Physically Informed Design of Ultra-hard Multicomponent Metallic Glasses

Anurag Bajpai, Dierk Raabe

Max Planck Institute for Sustainable Materials, 40237 Düsseldorf, Germany

Corresponding Authors:

Correspondence to: Anurag Bajpai, Dierk Raabe

a.bajpai@mpie.de (Anurag Bajpai), d.raabe@mpie.de (Dierk Raabe)

S1. Input Dataset Analysis

A dataset of 412 MMG alloy compositions, with applied loads (in Newtons), and corresponding Vickers hardness values was curated using the available literature. Figure SF1(a) illustrates the frequency distribution of constituent elements across the dataset. Elements such as Zr, Cu, and Ni exhibit notably high occurrences, underscoring their common usage as primary constituents in MMGs. Zr-based MMGs are widely investigated due to their superior glass forming ability (GFA) and mechanical properties, while Cu and Ni are frequently added to tailor properties such as ductility, strength, and corrosion resistance. Conversely, rare-earth elements (e.g., Dy, Y, and Er) and refractory elements (*e.g.*, Mo, Nb, Ta) are present at intermediate to lower frequencies, indicating their selective incorporation aimed at specific functional enhancements such as increased thermal stability, improved GFA, or targeted modification of mechanical behaviour. Figure SF1(b) details the hardness (HV) distribution, which prominently displays a multimodal pattern. The presence of distinct peaks around ~500 HV and ~1200 HV suggests the coexistence of multiple classes of MMGs within the dataset, each potentially governed by distinct compositional or structural factors. The bimodal distribution reflects variability not only in constituent elements but also in intrinsic structural features, such as short-range order, free volume content, and cluster packing efficiency, which fundamentally govern mechanical properties in MMGs. Figure SF1(c) presents a scatter plot depicting the relationship between hardness and applied load (N). The observed spread in hardness at similar loading conditions highlights inherent measurement variability, potentially attributable to factors such as indentation size effects (ISE), localized structural heterogeneities, and compositional variations across samples. Such scatter also underscores the importance of accounting for experimental parameters during model training, as load-dependent hardness behavior may inform the latent space representations extracted *via* VIB. Precisely modeling these load-dependent variations is essential, allowing the neural network to generalize across different loading conditions robustly, and enhancing the reliability and applicability of hardness predictions in experimental validations.

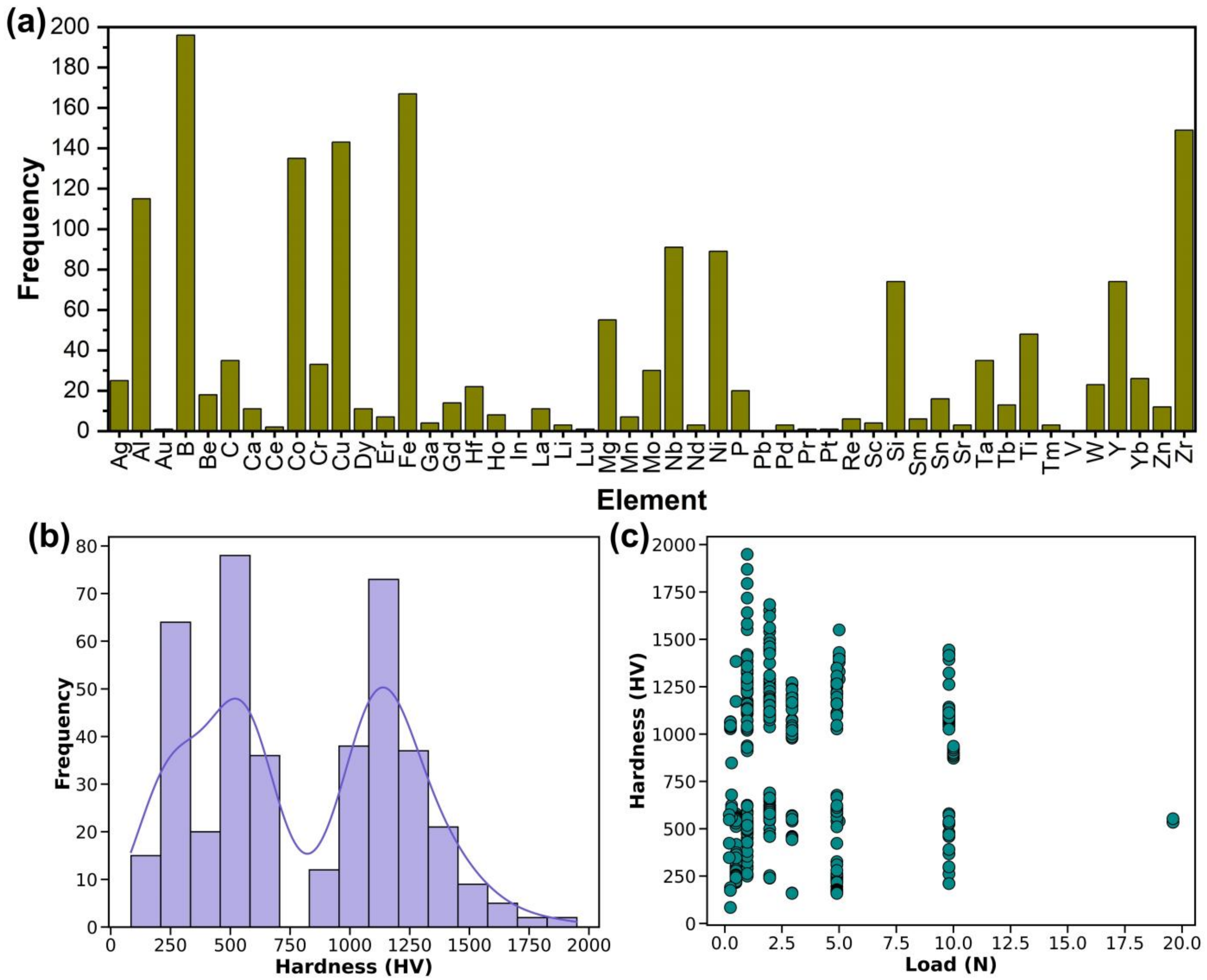

Figure SF1 - Exploratory data analysis of the multicomponent metallic glasses dataset. (a) Frequency distribution of elements present in the dataset. (b) Hardness (HV) distribution across the dataset, showing a multimodal pattern indicative of diverse alloy compositions and properties. (c) Scatter plot of hardness versus load, demonstrating the variation in hardness measurements with applied load.

S2. Supplementary Explanation on Variational Information Bottleneck and Attention Mechanisms

The Information Bottleneck principle, originally proposed by Tishby *et al*., formalizes the idea that a good representation of input data should capture only the information essential for predicting an output, while discarding irrelevant details.[1] In our setting, the input variable X represents the alloy composition and load, and the output variable Y is the predicted Vickers hardness. The goal is to learn a latent variable Z that compresses X as much as possible while preserving information relevant to Y.[2] This is achieved by minimizing the functional:

$$\mathcal{L}_{IB} = I(Z;X) - \beta I(Z;Y)$$

where $I(\cdot;\cdot)$ denotes mutual information and β is a trade-off parameter controlling the strength of the bottleneck.[1] Since the exact computation of mutual information is intractable in deep neural networks, the Variational Information Bottleneck (VIB) approximates this objective using variational inference.[3] In the VIB approach, the encoder network maps each input x to a Gaussian distribution in latent space, defined by its mean $\mu(x)$ and standard deviation $\sigma(x)$. A latent vector z is sampled using the reparameterization trick:

$$z = \mu(x) + \sigma(x) \cdot \epsilon, \qquad \epsilon \sim N(0, I)$$

This stochastic representation is then passed to the decoder network, which predicts the output property. The VIB loss function is defined as:

$$\mathcal{L}_{VIB} = \mathbb{E}_{q(z|x)}[-\log p(y|z))] + \beta D_{KL}(q(z|x)||p(z)),$$

The first term ensures predictive accuracy by minimizing the negative *log-likelihood* (typically mean squared error) of the predicted property.[3] The second term is the Kullback–Leibler divergence between the learned posterior $q(z|x)$ and a simple prior $p(z)$, usually taken as a standard normal distribution. This KL regularization penalizes overly complex latent encodings and encourages compression.[4] The result is a model that learns to represent the compositional space in terms of a small number of abstract latent variables, each of which carries physically useful information relevant to the target property. Importantly, as shown in the main text (Figure 5), we find that only a few latent dimensions contribute significantly to predicted hardness, indicating that the network has learned a disentangled, interpretable latent structure.

To further enhance interpretability, we incorporate an attention mechanism that allows the model to selectively focus on the most important elements in the alloy composition. The attention mechanism in neural networks was originally developed in the context of sequence modeling[5] but has found broad applicability across domains, including materials science, due to its ability to selectively focus on important features.[6] In the VIBANN framework, attention is applied directly to the input composition vector to learn which elements in a multicomponent alloy contribute most significantly to the target property, such as hardness.

In this context, the input vector $x \in \mathbb{R}^d$ represents the atomic fractions of d different elements in the alloy. The attention mechanism computes a set of weights $a = [a_1, a_2, \dots, a_d]$, where each weight $a_i \in [0,1]$ reflects the importance of the element i for the prediction task. These weights are computed using a learnable function, often a shallow neural layer followed by a softmax operation to normalize the values[5]:

$$a = softmax(W_a \cdot x + b_a)$$

Here, $W_a \in \mathbb{R}^{d\times d}$ and $b_a \in \mathbb{R}^d$ are trainable parameters, and the softmax function is defined as:

$$a_i = \frac{\exp(h_i)}{\sum_{j=1}^{d} \exp(h_j)}, \text{ where } h = W_a \cdot x + b_a$$

The result is a probability distribution over the input features, where the elements with higher attention scores are emphasized more strongly. These weights are then applied to the input *via* element-wise multiplication, producing an attended input vector:

$$x_{att} = a \odot x$$

where $\odot$ denotes the Hadamard (element-wise) product. The attended input x_{att} is passed downstream into the encoder of the VIB framework. By assigning high weights to composition components that are important for hardness prediction, and low or near-zero weights to irrelevant or redundant elements, the model effectively filters its input in a task-specific manner.

Mathematically, the attention mechanism can be seen as a feature-wise relevance map conditioned on the input itself. Unlike static feature selection techniques, attention is fully differentiable and adaptable: it automatically tunes its focus as the model sees new data, while simultaneously providing interpretable weightings that can be visualized and analyzed.[7] Importantly, the attention scores are learned jointly with the rest of the model through backpropagation, meaning they adapt dynamically to the training data. This leads to several advantages. First, the attention mechanism enables global interpretability: the average attention weight for each element across all training samples reveals which elements the model considers most consistently important. Second, attention supports context-dependent interpretation: for a given alloy composition, attention scores may shift depending on which element combinations are present, thereby capturing interaction effects that are difficult to model using fixed feature importance scores.

By combining VIB and attention, our VIBANN model learns a compressed, physically structured latent representation of composition–property relationships, while simultaneously highlighting which elements contribute most to the target property. This dual interpretability not only facilitates scientific understanding but also enables guided inverse design by allowing researchers to trace how small changes in input composition or latent representation affect the predicted outcome.

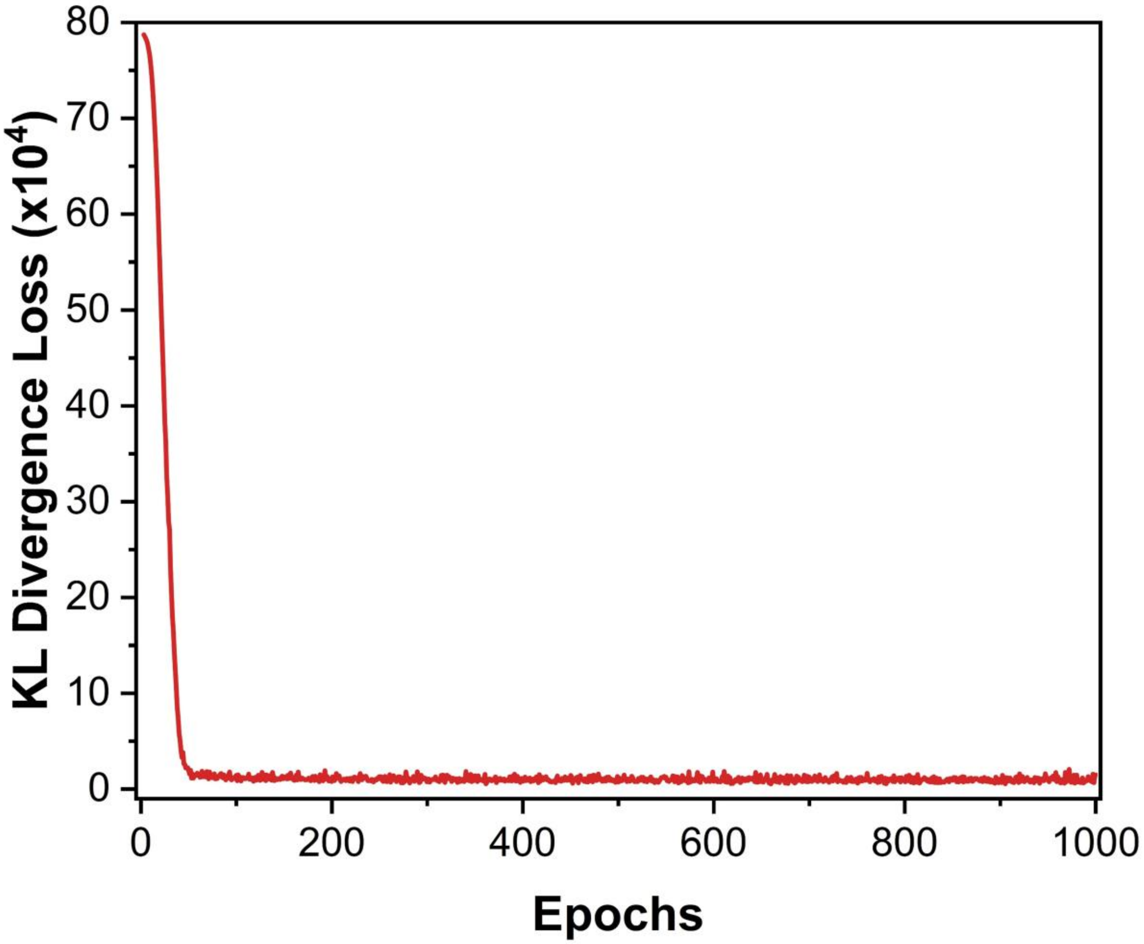

Figure SF2 - KL divergence loss evolution across epochs, demonstrating effective optimization of the latent space regularization. The significant drop within the first 100 epochs signifies the efficient balancing of reconstruction accuracy and latent space compression in the VIB framework.

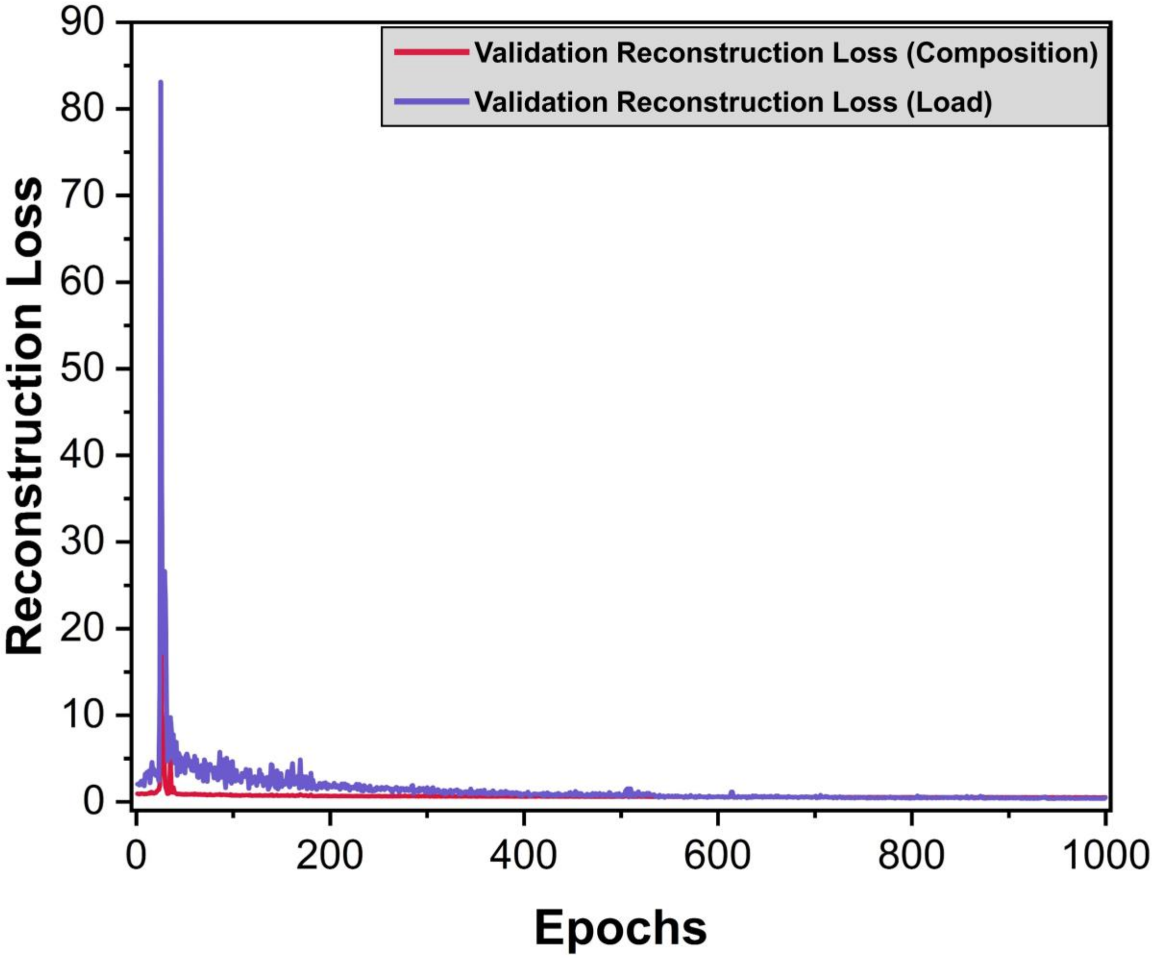

Figure SF3 - Validation reconstruction loss as a function of epochs for both composition (red) and load (blue) inputs during training of the VIBANN model. The sharp initial decrease followed by stabilization indicates rapid learning and convergence of the reconstruction model.

S3. Comparative Performance of the VIBANN model with benchmark regressors

Figure SF4 presents a comprehensive benchmarking analysis of our Variational Information Bottleneck–Attention Neural Network (VIBANN) framework against a suite of widely adopted ML regression algorithms for predicting the hardness (HV) of MMGs. The models considered include Lasso regression (Figure SF4(a)), Ridge regression (Figure SF4(b)), Random Forest (RF) (Figure SF4(c)), k-Nearest Neighbors (KNN) (Figure SF4(d)), and Gradient Boosting (GB) (Figure SF4(e)). The performance is quantitatively summarized in Figure SF4(f) using the coefficient of determination (R^2) and mean absolute error (MAE) as metrics. Among the classical linear models, both Lasso and Ridge regressions demonstrate limited predictive capacity with R^2 scores of 0.853 and 0.854, respectively, and relatively high MAE values (~84 HV). These models assume linear additive relationships among features and therefore fail to capture the inherent nonlinearity and high-order interactions present in the multivariate composition-load-hardness landscape of MMGs. Such simplifications are inadequate in the context of disordered systems like MMGs, where localized atomic environments, electronic structure, and loading conditions contribute nonlinearly to mechanical properties. The ensemble-based Random Forest ($R^2 = 0.913$, MAE = 63.7 HV) and Gradient Boosting models ($R^2 = 0.934$, MAE = 52.9 HV) provide notable improvements by leveraging nonlinearity and feature interactions through hierarchical tree structures. Gradient Boosting outperforms RF, possibly due to its sequential error-correcting nature, which is more effective at minimizing bias. The KNN model achieves a strong R^2 of 0.921 and MAE of 60.3 HV, indicating good local consistency but poor global generalization. Its performance is highly sensitive to data density and suffers in sparsely populated regions of the feature space—an important limitation when extrapolation beyond known alloy compositions is required. Crucially, the VIBANN model surpasses all benchmarks, delivering the highest R^2 and lowest MAE values, while offering additional advantages that extend beyond conventional performance metrics. Moreover, while some models exhibit comparable predictive accuracy, they fail to offer insight into the epistemic uncertainty of the predictions—a critical aspect when proposing novel, untested compositions for synthesis. The VIBANN model, by contrast, enables principled uncertainty quantification through its variational inference framework and Monte Carlo dropout during inference. This capability is particularly vital in materials discovery applications, where false positives can lead to significant experimental costs. It aligns with the fundamental objectives of modern materials informatics—not merely to predict properties, but to understand and guide the design of materials with targeted functionalities through interpretable and data-efficient representations.

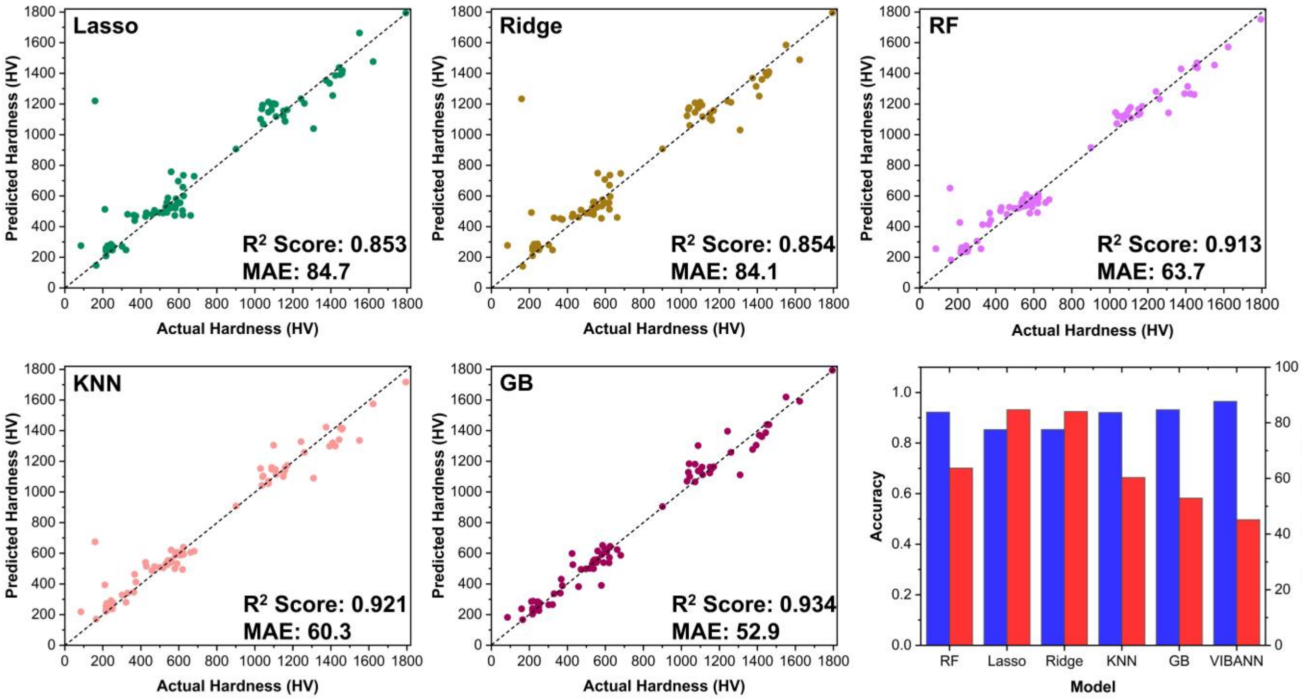

Figure SF4 - Performance comparison of various machine learning models in predicting the hardness (HV) of multicomponent metallic glasses. Predicted vs. actual hardness plots for (a) Lasso regression, (b) Ridge regression, (c) Random Forest (RF), (d) k-Nearest Neighbors (KNN), and (e) Gradient Boosting (GB); (f) Summary plot comparing accuracy (R^2) and mean absolute error (MAE) for all models, with the VIBANN model achieving the best performance among the models.

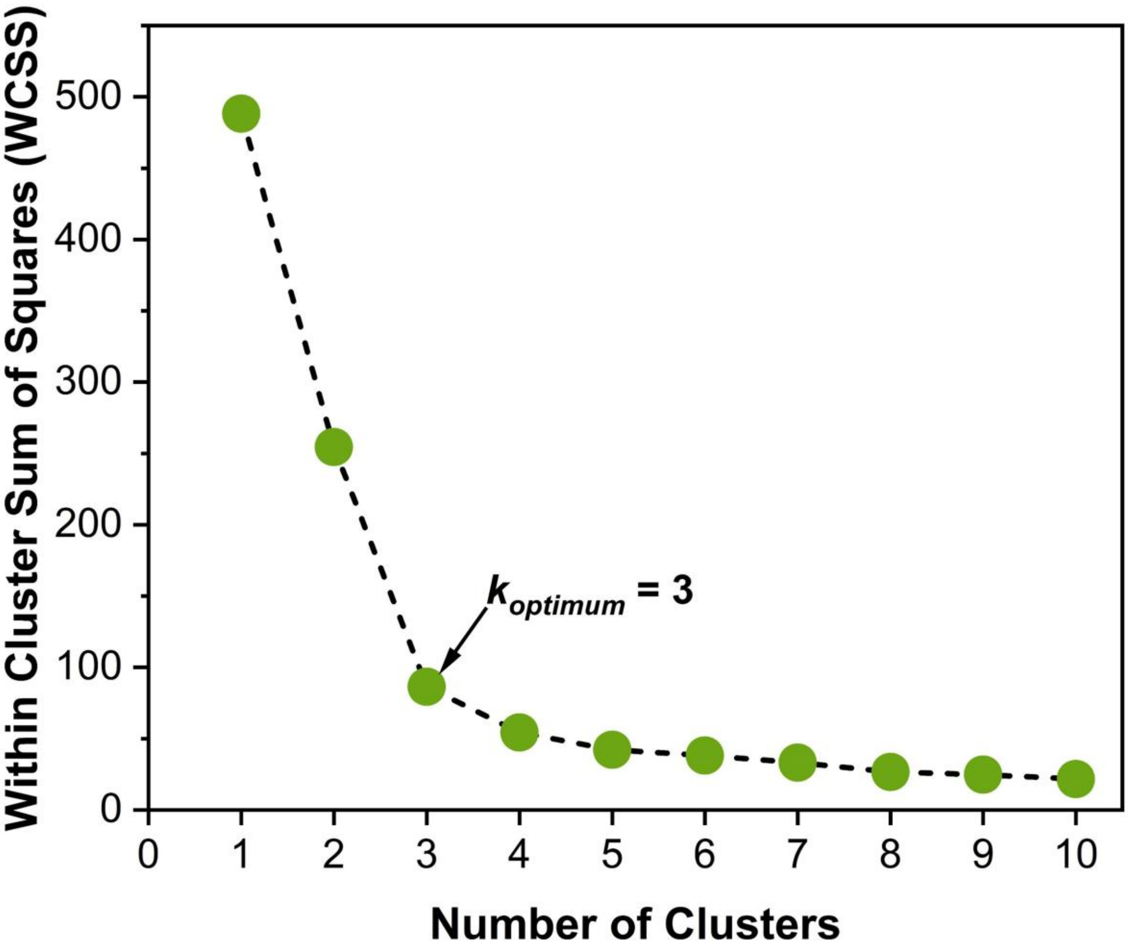

Figure SF5 - Elbow plot showing the within-cluster sum of squares (WCSS) as a function of the number of clusters for K-means clustering. The optimum number of clusters ($k_{optimum}$ = 3) is identified at the "elbow" point, where the rate of decrease in WCSS slows, indicating the most appropriate number of clusters.

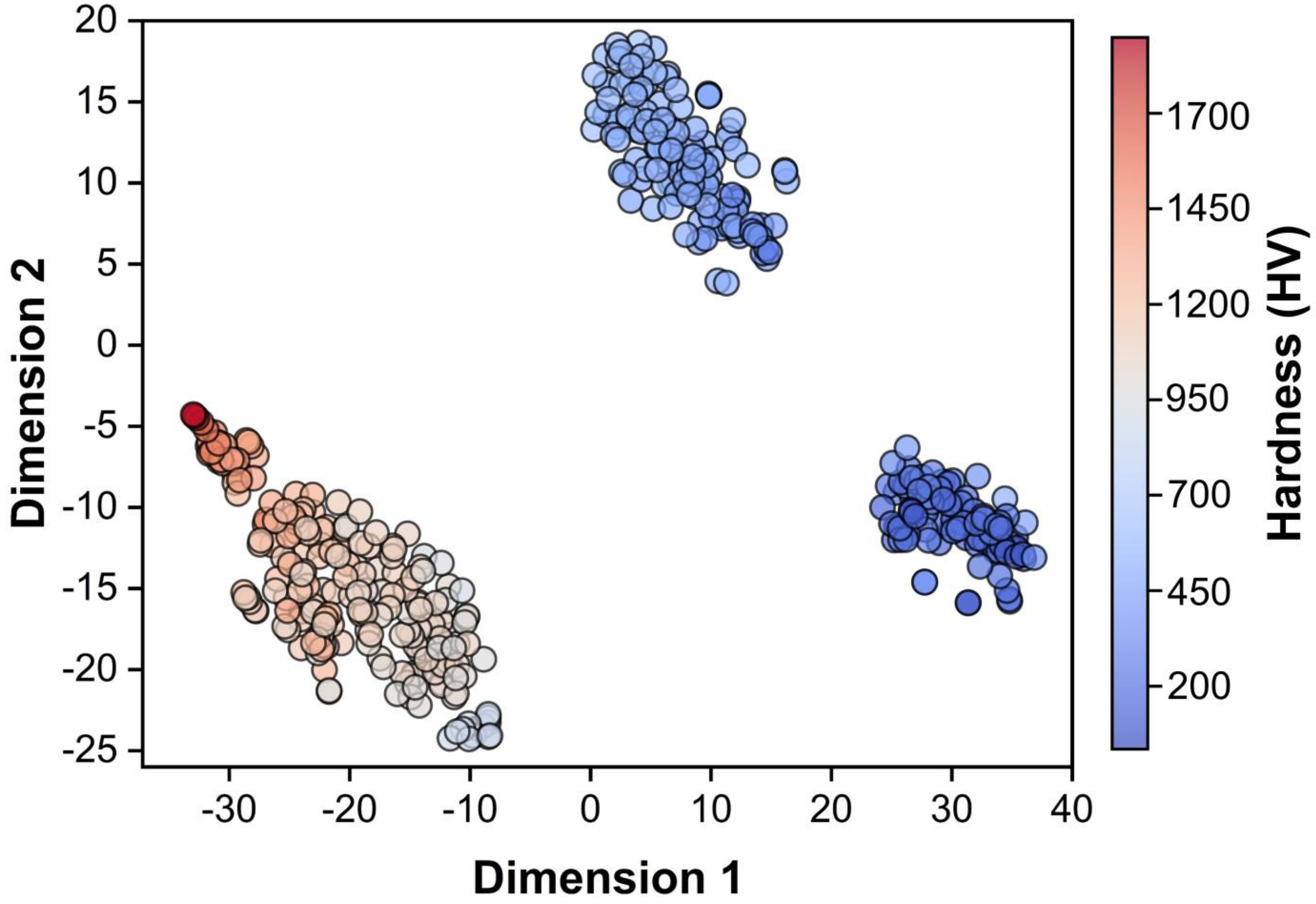

Figure SF6 - t-SNE visualization of the latent space for multicomponent metallic glasses, with hardness (HV) values represented as a color gradient. Distinct regions in the latent space correspond to clusters of alloys with similar hardness values, with higher hardness values highlighted in red and lower values in blue.

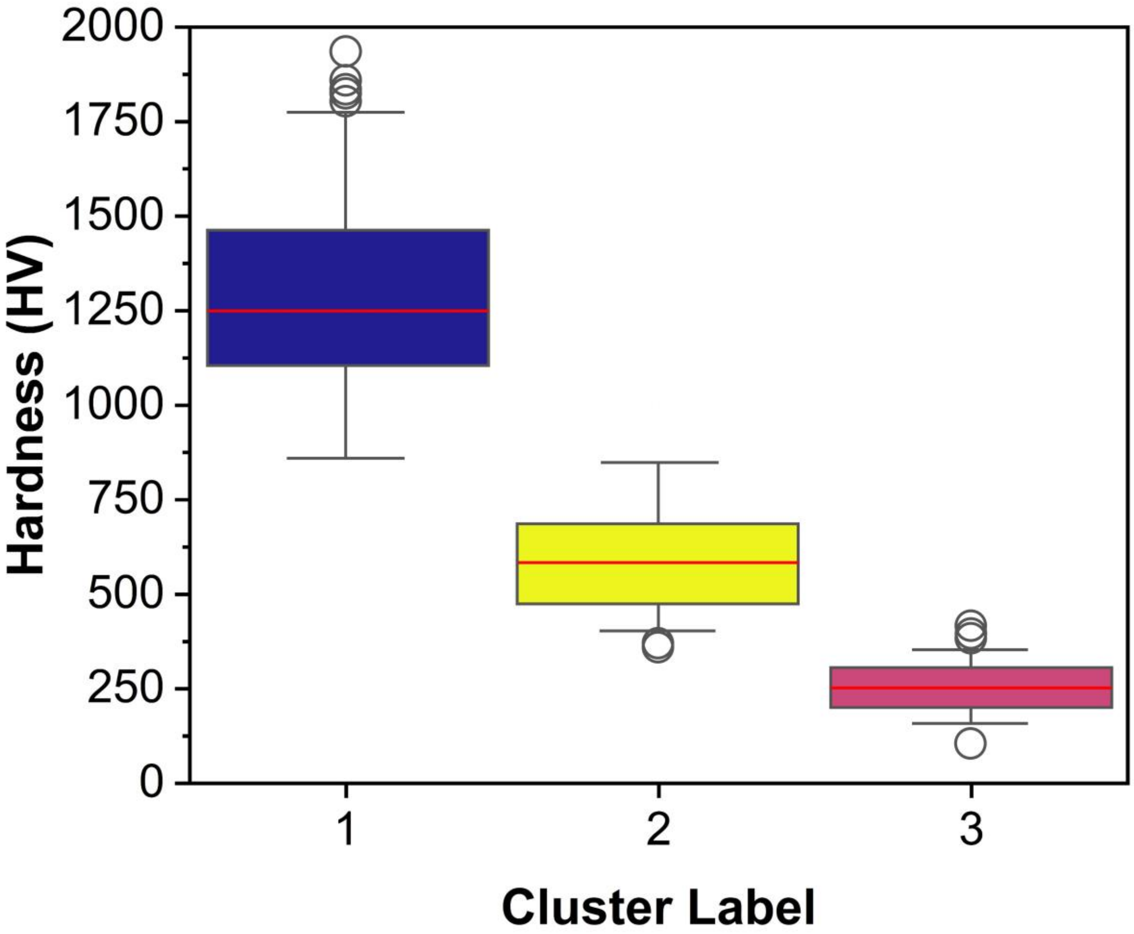

Figure SF7 - Boxplot illustrating the distribution of measured hardness (HV) values for the three distinct clusters. The horizontal line within each box denotes the median hardness, while the upper and lower edges of the box represent the 75th (Q3) and 25th (Q1) percentiles, respectively. The whiskers extend to 1.5 times the interquartile range (IQR) beyond Q1 and Q3, and any points lying outside these whiskers are plotted as outliers.

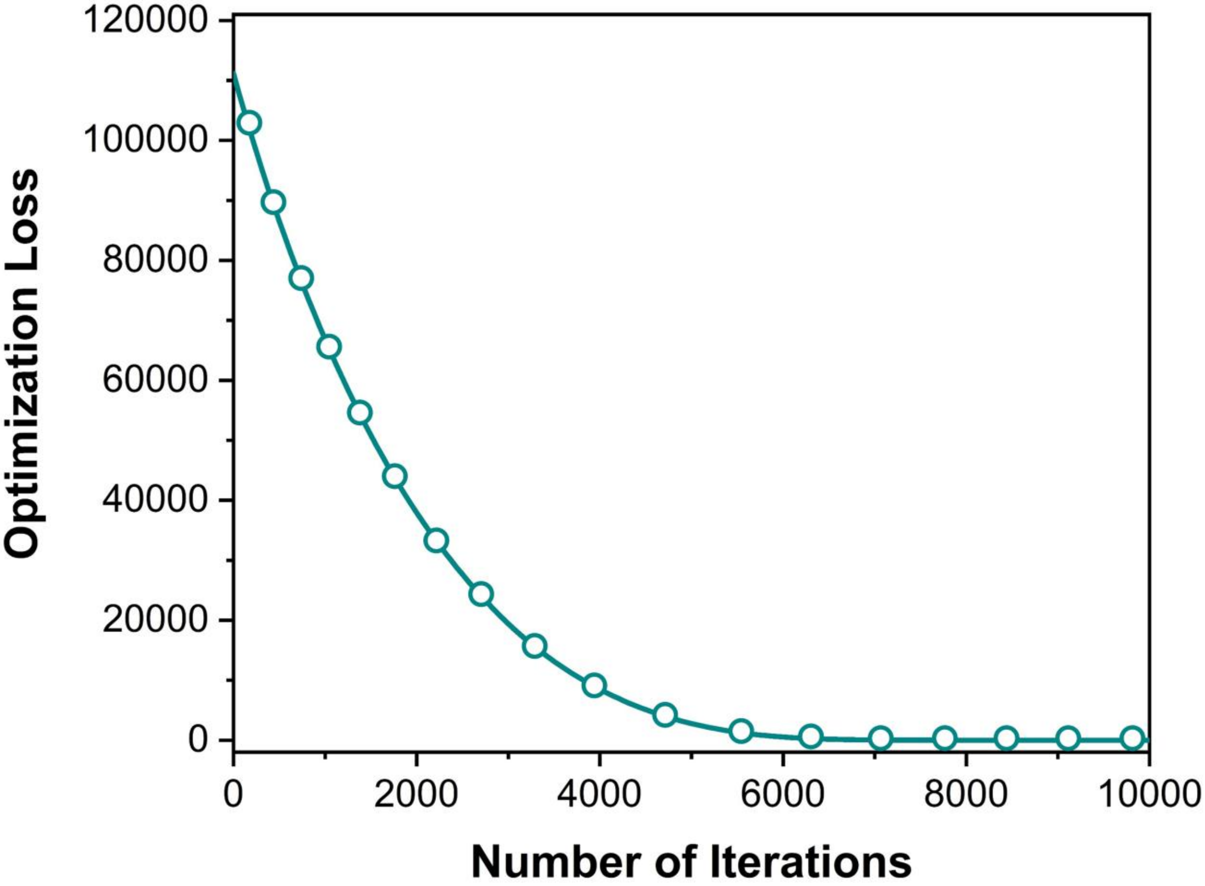

Figure SF8 - Optimization loss as a function of epochs during the training process for inverse alloy design using gradient based optimization.

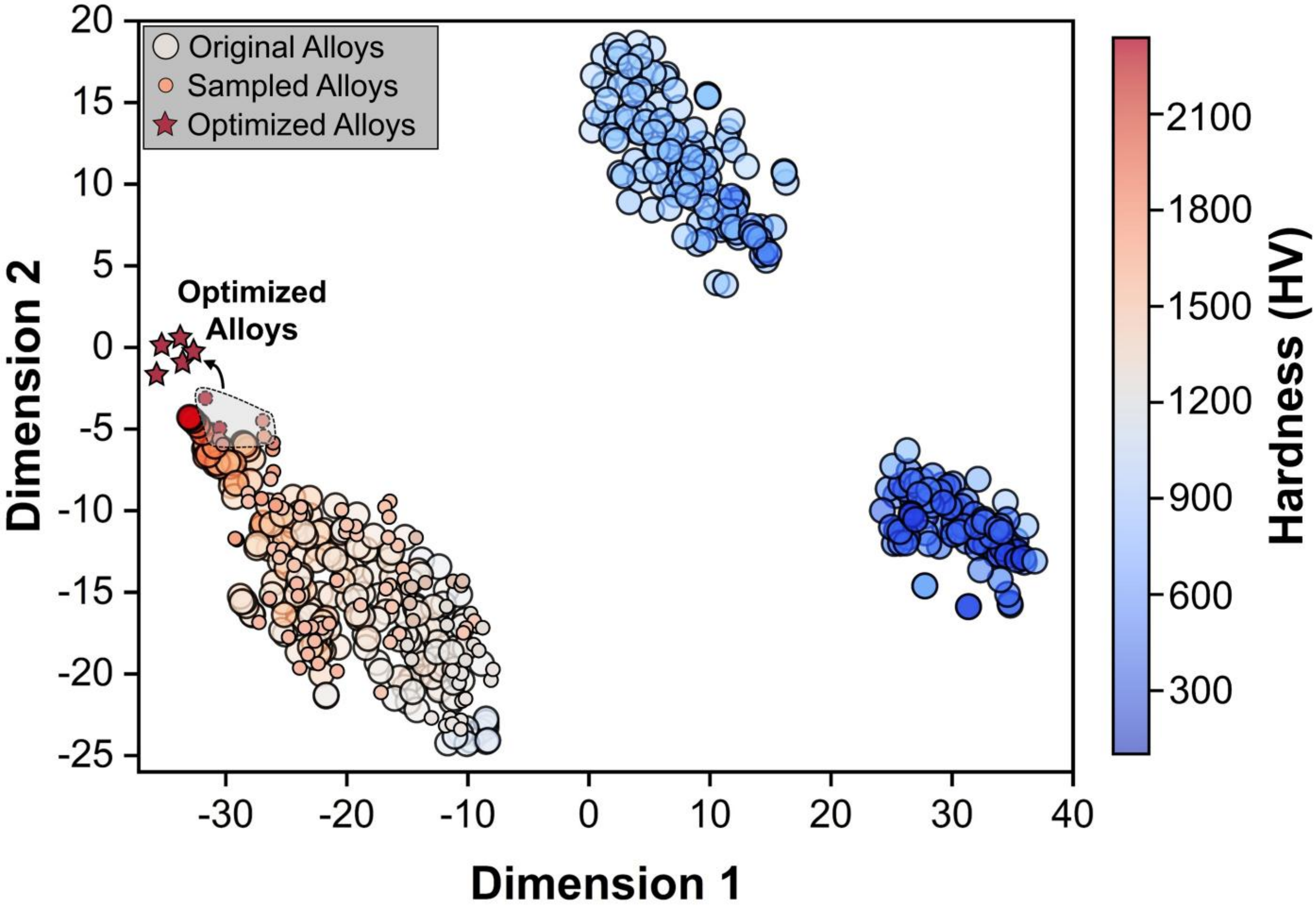

Figure SF9 - t-SNE projection of the VIBANN latent space showing the original dataset alloys, randomly sampled latent points, and inverse-designed optimized alloys (stars). The color map indicates the predicted Vickers hardness (HV), with a gradient from low (blue) to ultra-high (red) hardness values.

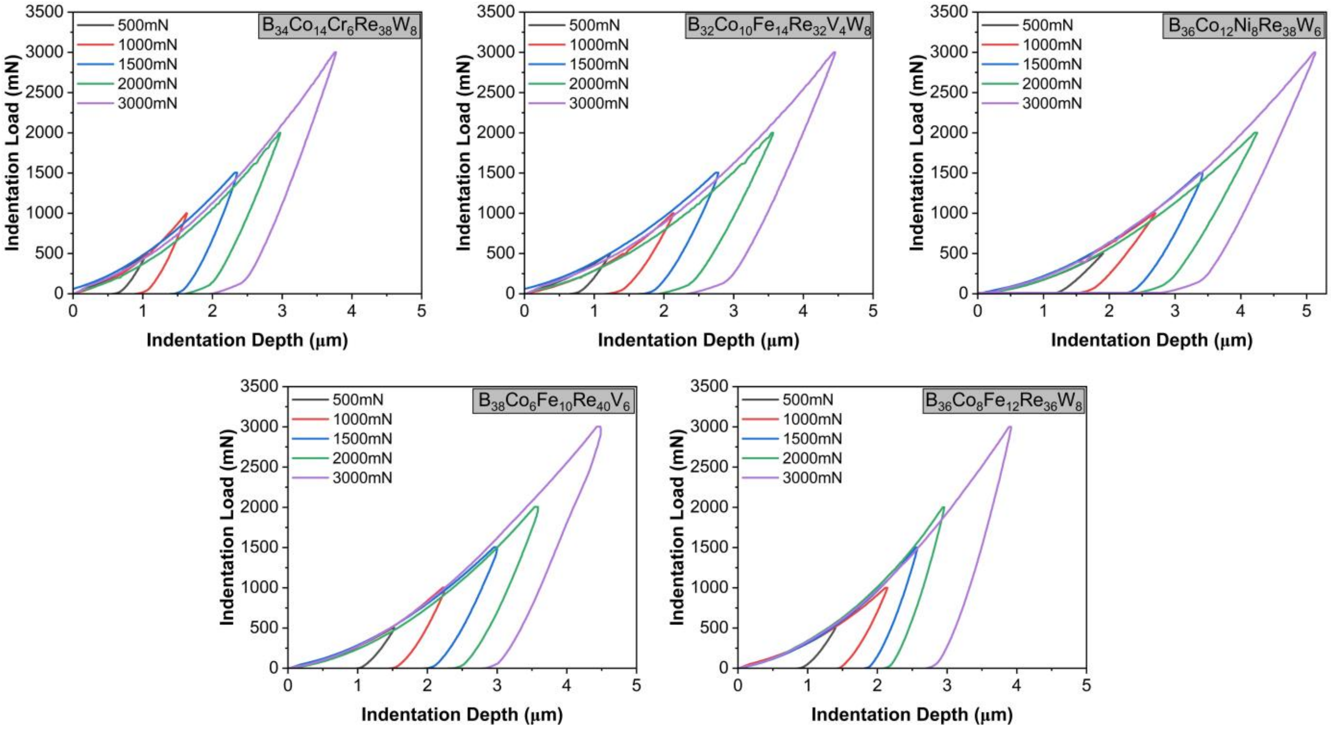

Figure SF10 - Load versus indentation depth curves for multicomponent metallic glasses tested at various loads.

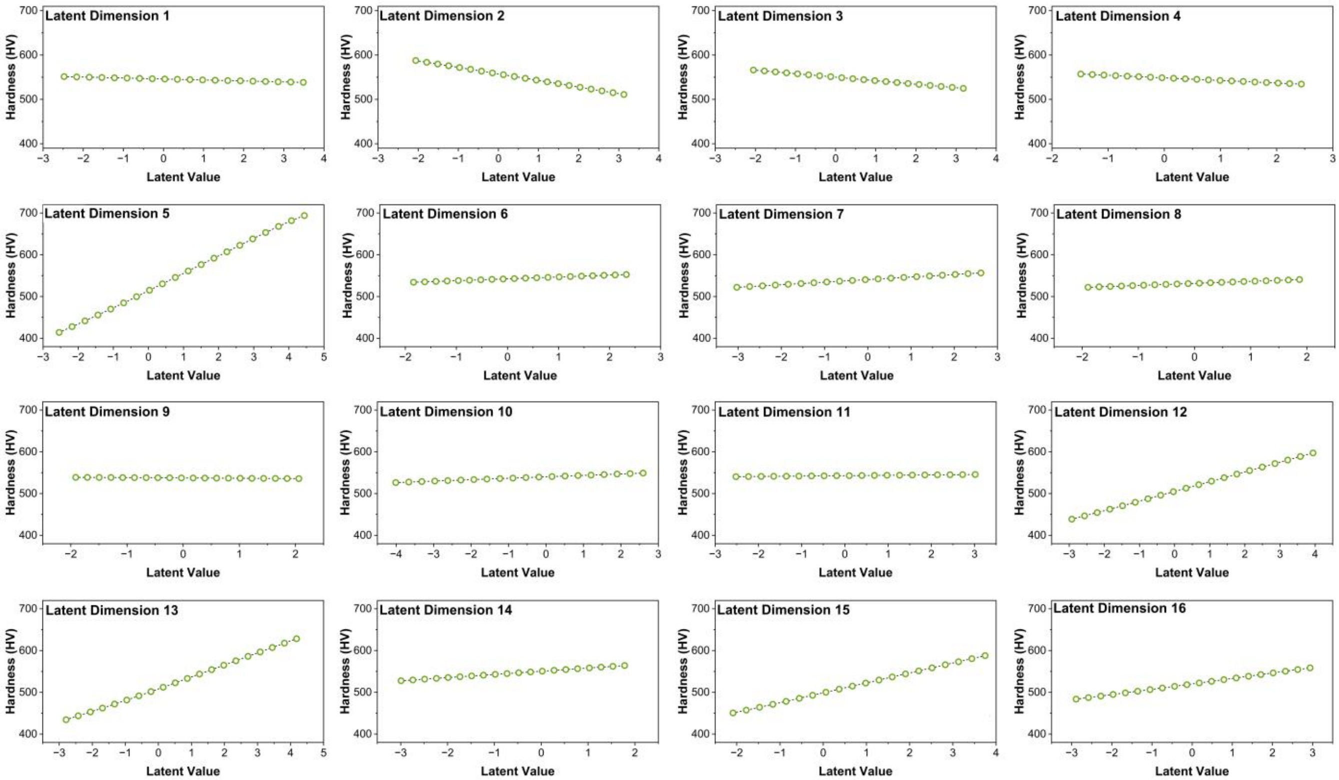

Figure SF11 - Hardness variation as a function of latent perturbations, obtained by traversing a representative latent vector from the center of the latent space. Most latent dimensions demonstrate predominantly linear hardness responses, suggesting a smoothly organized latent representation. Minor nonlinearities, notably in dimensions 5, 12, 13 and 15, indicate regions of property saturation or softening, consistent with the localized complexity.

Table ST1 – Search space for all relevant hyperparameters and the best-performing hyperparameters for the various ML models (the best hyperparameters are chosen based using Bayesian optimization and 5-fold cross-validation).

Model	**Hyperparameter**	**Hyperparameters Grid**	**Best Hyperparameters**
K-Nearest Neighbours (KNN)	Number of Neighbours	2, 30	4
	Weight	uniform, distance	distance
	Metric	euclidean, manhattan	manhattan
Random Forest (RF)	Number of estimators	10-500	124
	Criterion	gini, entropy	Entropy
	Maximum depth	2-30	19
	Minimum sample split	1-20	2
	Minimum sample leaf	1-20	3
Lasso Regression (LR)	Alpha	0.00001-1.0	0.01
	Selection	cyclic, random	cyclic
Ridge Regression (RR)	Alpha	0.00001-1.0	0.01
	Solver	auto, svd, cholesky, lsqr, sparse_cg	svd
Gradient Boosting (GB)	Number of estimators	10-200	128
	Learning rate	0.00001-1.0	0.001
	Maximum depth	1-20	13
	Minimum sample split	2-20	6

Table ST2 – Bulk chemical compositions of the final optimized MMGs measured by energy dispersive spectroscopy.

$\mathrm{B_{34}Co_{14}Cr_{6}Re_{38}W_{8}}$	B (at.%)	Co (at.%)	Cr (at.%)	Re (at.%)	W (at.%)	
	33.4	15.1	6.5	37.2	7.8	
$\mathrm{B_{32}Co_{10}Fe_{14}Re_{32}V_{4}W_{8}}$	B (at.%)	Co (at.%)	Fe (at.%)	Re (at.%)	V (at.%)	W (at.%)
	30.1	10.9	14.4	32.2	4.8	7.6
$\mathrm{B_{36}Co_{12}Ni_{8}Re_{38}W_{6}}$	B (at.%)	Co (at.%)	Ni (at.%)	Re (at.%)	W (at.%)	
	34.7	12.4	7.8	37.4	7.7	
$\mathrm{B_{38}Co_{6}Fe_{10}Re_{40}V_{6}}$	B (at.%)	Co (at.%)	Fe (at.%)	Re (at.%)	V (at.%)	
	37.2	6.3	10.6	39.2	6.7	
$\mathrm{B_{36}Co_{8}Fe_{12}Re_{36}W_{8}}$	B (at.%)	Co (at.%)	Fe (at.%)	Re (at.%)	W (at.%)	
	35.4	7.7	12.8	36.3	7.8	

References

1 Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, (2000).

2 Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, (2016).

3 Kingma, D. P. & Welling, M. (Banff, Canada, 2013).

4 Achille, A. & Soatto, S. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**, 2897-2905, (2018).

5 Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, (2014).

6 Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30**, (2017).

7 Jain, S. & Wallace, B. C. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, (2019).