Data-efficient and interpretable inverse materials design using a disentangled variational autoencoder

Cheng Zeng[1, 2, †, *], Zulqarnain Khan[1, †, *], and Nathan L. Post[2]

[1]Institute for Experiential AI, Northeastern University, Boston, MA 02115, United States
[2]The Roux Institute, Northeastern University, Portland, ME 04101, United States
[†]These authors contribute equally: Cheng Zeng, Zulqarnain Khan.
[*]Corresponding author: c.zeng@northeastern.edu, z.khan@northeastern.edu

Abstract

Inverse materials design has proven successful in accelerating novel material discovery. Many inverse materials design methods use unsupervised learning where a latent space is learned to offer a compact description of materials representations. A latent space learned this way is likely to be entangled, in terms of the target property and other properties of the materials. This makes the inverse design process ambiguous. Here, we present a semi-supervised learning approach based on a disentangled variational autoencoder to learn a probabilistic relationship between features, latent variables and target properties. This approach is data efficient because it combines all labelled and unlabelled data in a coherent manner, and it uses expert-informed prior distributions to improve model robustness even with limited labelled data. It is in essence interpretable, as the learnable target property is disentangled out of the other properties of the materials, and an extra layer of interpretability can be provided by a post-hoc analysis of the classification head of the model. We demonstrate this new approach on an experimental high-entropy alloy dataset with chemical compositions as input and single-phase formation as the single target property. While single property is used in this work, the disentangled model can be extended to customize for inverse design of materials with multiple target properties.

1   Introduction

Materials play a pivotal role in shaping the modern society and many grand technological challenges are materials challenges. These range from lower-cost battery materials for energy storage, to quantum computing materials and bio-compatible materials for healthcare applications [1, 2, 3]. Thanks to advances in high-throughput computing [4], robotics [5], machine learning force fields [4] and open data respositories of materials [6, 7], materials design and discovery have now reached an unprecedented rate and scale [8]. Although advances in algorithm and hardware significantly reduce the computation time for each iteration, it can still take extensive iterations or computation to pinpoint a small range of potential materials candidates with desired properties [9].

Inverse materials design unlocks the potential to optimize new materials towards a target property. In general there are four approaches for inverse materials design [10]; high-throughput virtual screening [11], global optimization [12], reinforcement learning [13] and generative models [14, 15]. Among these, generative models can be very data-efficient as they allow for encoding expert informed information into the model, thereby reducing the amount of data that is required to learn a compact low-dimensional representation. Moreover, the learned representation space can generate new data using the knowledge encoded during training. Comparing various generative models, a study by Türk et al showed that variational autoencoder (VAE) is more robust than reinforcement learning and generative adversarial networks because VAE has a better representation of underlying distributions and training a VAE model is easier [16]. However, current generative models are primarily present as an unsupervised learning approach which learns a latent space assumed to entangle the relationship between materials representations and target properties [17, 10, 18]. It is not ideal as one will need to perform post-optimization to explore materials with better target properties and may even fail to find any useful materials. A recent work by Xie and Tomioka et al introduced diffusion-based generative processes together with a fine tuning process to discover materials with multiple target properties such as magnetic property and supply chain risk [19]. Although a state-of-the-art discovery rate and materials stability were reported, the full periodicity of the crystal structures

restricts the design space for inorganic materials, the complexity of large datasets and diffusion models hinder a wider application and the interpretability of the methods, and the practicality of the new materials discovered in this process is doubtful without an expert insight [20].

Therefore, there is an urgent need to create a workflow for inverse materials design that is data efficient, interpretable and can be geared towards multi-property optimization. The objective of this work is to develop high-entropy alloys (HEAs) which tend to form a single-phase structure. Conventional methods to predict single-phase alloys rely on sophisticated design of experiments, thermodynamic modeling and first-principles calculations, which are not efficient search for high-entropy alloys because the number of points to search grow combinatorially with the increase of elements [21, 22]. For example, a HEA with five possible elements and each element has an integer composition between 5% and 35%, the number of possible compositions is on the order of 10^7, let alone the massive numbers of possible atomic arrangement for each composition. Although machine learning algorithms have proven to speed up the search of corrosion-resistance or mechanically strong HEAs [9, 23, 24], this forward design strategy still needs to scan a wide configuration and composition space. Inverse design is theorized to be a much more robust approach as it learns a probabilistic relationship between materials representation and a compact latent space where new materials candidates can be generated from this relationship. Coupled with an uncertainty estimate, search directions for new materials can be identified with a risk quantification [18]. Here we introduced a disentangled generative model using a semi-supervised variational autoencoder for the inverse design of complex materials. We demonstrated this approach on single-phase high-entropy alloys. Although our demonstration focuses on single materials property (single phase formation), the approach can also be applied to materials discovery with multi-objective targets. This approach can utilize both labelled and unlabelled data sets. The proposed methods are highly versatile and robust, and can be readily extended to other engineering domains where there exists a probabilistic input-output relationship.
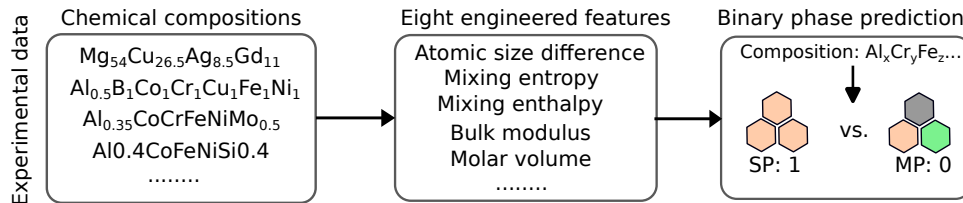
2   Methods & Theories

2.1   Data set

Figure 1: Experimental high-entropy alloy dataset for single phase formation.

This work uses an experimental data set where inputs are chemical compositions of alloys and outputs are binary phase predictions indicating the formation of single phase (**SP**: 1) versus multiple phases (**MP**: 0). A summary of this dataset is shown in Figure 1. It is an up-to-date and well-recognized dataset for HEA single phase prediction collected by Yan et al [23]. Empirical rules suggest that the single phase formation relies more on the synergistic effects of mixing different alloys rather than the independent attributes of each element [23, 25]. Therefore, eight engineered features are used rather than element-wise compositions. The eight features include bulk modulus (k), molar volume (V_m), melting temperature (T_m), valence electron concentration (VEC), atomic size difference (δ), Pauling electronegativity difference (δ_χ), mixing entropy (ΔS_{mix}) and mixing enthalpy (ΔH_{mix}). Methods to calculate these features can be found in the supporting information of our previous work [9]. Those eight features were found to be informative for the prediction of single phase formation [9, 23]. More importantly, this feature engineering provides a general and compact representation of alloys comprising various elements, offering more generalizability for alloy design. Given that an experimental data set is used, the trained ML models can thus give predictions that are more likely to be manufactured in practice, without explicitly considering the manufacturing process and environmental conditions. For the purpose of inverse materials design, we also created another data set where chemical compositions were transformed to element-wise compositions. For simplicity, we restrict the element list to the top-30 frequent elements in the data set. Hence each alloy is converted to a 30-element composition feature vector. This composition feature vector, combined with the binary phase formation, is used as the input for the variational autoencoder.

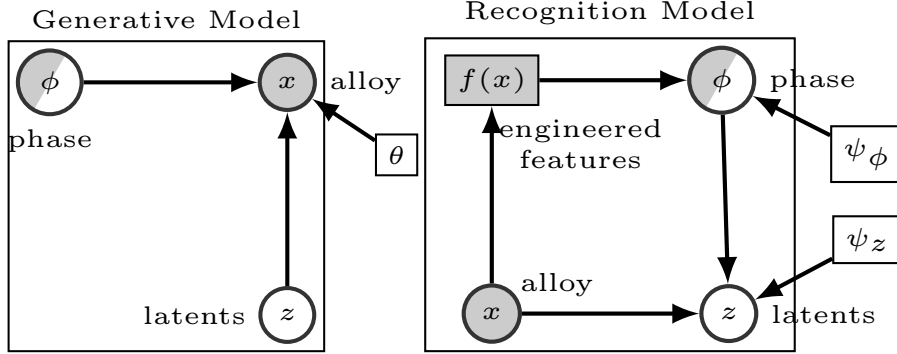## 2.2 Proposed Model: Disentangled VAE

2.3 Generative model

Figure 2: Generative (Left) and recognition model (Right) in the disentangled variational autoencoder for inverse design of single-phase high-entropy alloys.

Assume $x \in \mathbb{R}^+$ represents an alloy and that each alloy has a phase $\phi \in \{0, 1\}$ that is our target property. The phase is assumed known for at least a subset of the training data. In addition assume $z \in \mathbb{R}^l$ as an l- dimensional latent variable, represents any other factors that may be responsible for generating alloys. With this notation in place, we write the generative model (Figure 2, left) for the data in the form of the following joint probability distribution:

$$p_\theta(x, \phi, z) = p_\theta(x|\phi, z)p(\phi)p(z) \tag{1}$$

Where θ parameterizes the likelihood using a neural network that takes as input the sampled instances of ϕ and z. Different choices of priors can be experimented with for each variable, we propose one possible set of choices that best suit the data-space that each variable belongs to. The likelihood of the compositional data is modeled as a multinomial parameterized by the neural network θ :

$$x \sim Multinomial(\theta(\phi, z)) \tag{2}$$

Priors for the binary Phase ϕ are modeled using Bernoulli distribution and lastly, we assume a standard normal prior over the latent variable z as a relatively uninformative prior.

$$\phi \sim Bernoulli(r), z \sim Normal(0, I) \tag{3}$$

Note that informed choices for the hyperparameters of these prior distributions can be made, e.g. the phase prior can be chosen to bias the model towards observing more multiphase alloys vs single phase alloys.

2.3.1 Recognition model

With the above generative model in place, we can now define the recognition model to perform inference. There are different acceptable recognition models given the above generative model. We propose one that conforms best to the our goals of enabling prediction, interpretation, and exploration. The recognition model $q_\psi(\phi, z|x)$ takes as input the data and maps it to the latent representation space, serving as the variational approximation to the otherwise intractable posterior distribution $p(\phi, z|x)$. We assume the recognition model factorizes as follows under a mean-field assumption:

$$q(\phi, z|x) = q_{\psi_\phi}(\phi|f(x))q_{\psi_z}(z|x, \phi) \tag{4}$$

The idea behind this factorization is that, we know from prior knowledge that the phase ϕ is well predicted by a physics informed hand-engineered transformation $f(x)$ [9]. Specifically this transformation takes in composition and outputs eight relevant physical descriptors [23]; atomic size difference, mixing enthalpy,

mixing entropy, Pauli electronegativity difference, molar volume, bulk modulus, melting temperature and valence electron concentration. In turn the neural network ψ_ϕ uses these engineered features to predict the binary phase. Note that $f(x)$ is a pre-specified transformation and not a learnable function. Finally the latent variable z encodes everything else about the alloy x conditioned upon values of phase. This recognition model allows us to encode hand-engineered features that we know are useful for predicting phase, and rely on the expressivity of neural networks to encode other information necessary for the eventual reconstruction of the training data through the decoder.

2.3.2 Model training

We can train both the generative model and the recognition model simultaneously by maximizing the following variational objective function w.r.t the neural network parameters for the generative and recognition model:

$$\sum_{n=1}^{N} \mathcal{L}(\theta, \psi_z, x^n) + \gamma \sum_{m=1}^{M} \mathcal{L}^{sup}(\theta, \psi_\phi, \psi_z; x^m, \phi^m) \tag{5}$$

The first part of this objective function optimizes over all $n = 1, \ldots, N$ data points for which supervision is not available. This is the standard VAE evidence lower bound (ELBO) loss, which can generally be thought of as learning to reconstruct input data with some regularization on the latent space. The second term is the supervised loss and uses M points where supervision is available i.e. points for which values for phase formation are available. Where the constant γ is a hyperparameter that balances prediction accuracy for supervised learning and reconstruction accuracy for unsupervised learning. The objective above can then be approximated using a Monte Carlo estimator as explained in the survey by Luengo et al [26]. We utilize Pyro [27], a pytorch based probabilistic programming language, for model specification as well as training and inference.

The entire data set with 1373 data points was split into 90:10 training/testing. The training set was furthered split into 70:30 labelled/unlabelled, the unlabelled training data is further split into 80:20 training/validation, this results in 864 labelled training examples, 296 unlabelled training examples, 75 validation examples, and 138 test examples. Two hidden layers of size 100 each were used for both recognition and generative models. 'Adam' optimizer was used for optimization. An initial training rate of 10^{-4} was employed with a 0.9 decay rate for momentum and 0.999 decay rate for squared gradients. The learning rate was reduced by a factor of 0.5 if no improvement in training has been observed for 200 consecutive steps. A batch size of 32 was used for training the model with up to 20000 epochs. The model with the highest validation accuracy was saved as the best model for hereafter inference and analysis. To examine the randomness of data splitting, five different random seeds were used to split the train, validation, unlabelled and test datasets. The mean and standard deviation of prediction accuracy and area under the curve for receiving operating characteristics are reported in Section 3.1.1.

2.4 Post-hoc analysis: SHAP feature importance

To analyze the overall impact of each feature on the classifier for the single-phase formation, we utilize the existing post-hoc analysis method implemented in SHAP [28], a game theoretic approach to explain any model output. Specifically, we used the model-agnostic kernel explainer that relies on specially-weighted local linear regression. The training set was used as the background dataset to integrate out features, and the overall feature importance was evaluated and reported on the test data.

3 Results and Discussion

3.1 Semi-supervised machine learning

3.1.1 Classification for single phase formation

Using five random seeds for data splitting, the mean and standard deviation of prediction accuracies for training, validation and test are 0.883±0.027, 0.930±0.026 and 0.829±0.050, respectively. The mean and standard deviation of area under curves (AUC) are 0.954±0.014, 0.955±0.029 and 0.890±0.025 for respective training, validation and test data. Using the model with the highest test accuracy, Figure 3 shows the receiving operation characteristic (ROC) curves for training, validation and test datasets. The area under curves (AUC) are all no less than 0.91, suggesting a reliable prediction of single-phase formation.
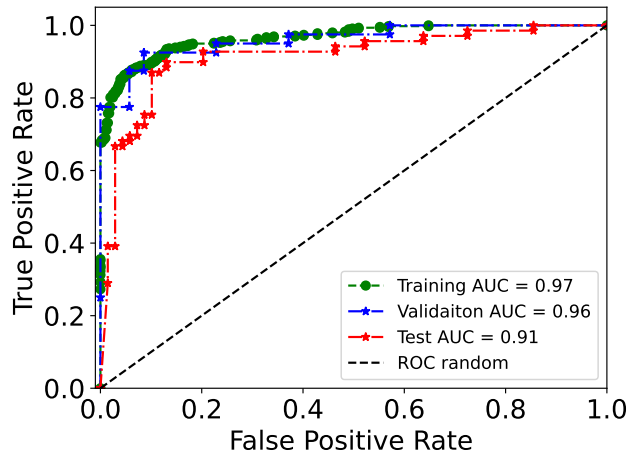
Figure 3: ROC curves for training, validation and test.

3.1.2 Alloy reconstruction

Next, we chose a few HEA examples in the test dataset, and we compared the predicted class to the ground truth labels as well as the reconstructed alloys versus the original ones. We first compute the predicted probability of being single phase using the classifier. We then reconstruct the alloy given the composition features and the predicted probability using the generative model. The results are summarized in Table 1.

- In the first example '$Fe_{19}Ni_{19}Cr_{19}Co_{13}Al_{19}Mo_9$', it shows that the predicted class is highly likely to be single phase, consistent with the ground truth. Plus, the reconstructed alloy comprises of the same elements as the original alloy with slightly different composition (marginally higher Co and Ni compositions and lower Cr and Al compositions).

- The second example '$Al_{11}Ti_{22}V_{22}Nb_{22}Zr_{22}$' is primarily made up of refractory elements with a small amount of Al and it forms a multiple-phase structure in experiments. It is also predicted to form a multiple-phase structure. The reconstruction is largely similar except with trace amounts of Cr and Ta which are not present in the original. While this isn't perfect reconstruction, note that Cr and Ta are in nature also refractory elements, similar to Ti, V, Nb and Zr.

- The third example '$Al_4Ti_{23}Mo_{23}V_{23}Ta_{23}$' is similar to the second one but with less Al. It tends to form a single-phase structure. The predicted class agrees well with the ground truth, and the reconstructed alloy contains three more elements (Cr, Nb and Zr) than the input alloy, which can be explained by the same rationale for the second example.

- In the last example, the predicted probability 0.52 of '$Fe_{20}Ni_{20}Co_{20}Ti_{20}Cu_{20}$' slightly favors the formation of single phase, which turns out to be the ground truth. The reconstructed alloy also has more elements (Al, Cr, V, Mn) and no Ti. The replacement of Ti with Cr and V follows the same logic as the second and third examples. The addition of a small amount of Al may be attributed to the frequent appearance of Al in the training data while Mn, located in the middle of first-row transition metals, reconciles differences between different types of transition metals between (Fe, Co, Ni), (V, Cr) and Cu. To further examine the reconstructed alloy, its predicted probability is 0.51, indicating a self-consistent reconstruction process.

The above examples show that while the reconstructions are (expectedly) not exact, the "errors" in reconstruction happen along materially explainable directions - indicating that the latent space is encoding useful information about the alloy generation process.

5

Table 1: Evaluation of semi-supervised autoencoder for alloy reconstruction. Since formulas for original alloys are varied, we converted all formulas to a standard format where compositions are rounded to integers with a sum close to 100, and element orders are sorted. Total compositions may not be exactly 100 because of the rounding issue.

Original alloy	Reconstructed alloy	Predicted probability	Ground truth label
$Fe_{19}Ni_{19}Cr_{19}Co_{13}Al_{19}Mo_9$	$Fe_{19}Ni_{21}Cr_{15}Co_{21}Al_{15}Mo_9$	0.15	0 (Multiple phase)
$Al_{11}Ti_{22}V_{22}Nb_{22}Zr_{22}$	$Cr_3Al_7Ti_{18}V_{20}Nb_{32}Zr_{16}Ta_4$	0.10	0
$Al_4Ti_{23}Mo_{23}V_{23}Ta_{23}$	$Cr_1Al_2Ti_{29}Mo_{26}V_{11}Nb_{16}Zr_8Ta_7$	0.95	1 (Single phase)
$Fe_{20}Ni_{20}Co_{20}Ti_{20}Cu_{20}$	$Fe_{17}Ni_{23}Cr_{13}Co_{21}Al_2Cu_5V_{10}Mn_{10}$	0.52	1

3.1.3 Data efficiency

In comparison with supervised machine learning only. this proposed semi-supervised framework is more data efficient, and the reason is twofold. First, it learns a probabilistic feature–target relationship which uses prior distributions for features, target and latent variables. This expert-informed knowledge constraints model fitting and is supposed to be superior in generalizing the data points and to show less variance when predicting on unknown data. Second, it leverages all relevant information from both labelled and unlabelled data. Although a classifier solely determines the target property once it is trained, the unsupervised learning is integrated into the supervised learning because the losses are added up and optimized together. The priors also regularize the classifier head, and it enforces a more accurate description of feature–target relationship that influences the reconstruction process.

To verify the data efficiency of the proposed framework, we held out the same 138 data points for testing. The prediction accuracy were compared between semi-supervised learning and conventional supervised learning on the training and test data.Neural network models were used for the supervised learning using hyper-parameters close to ones chosen for the classifer of the semi-supervised learning. We split the remaining data into labelled, unlabelled and validation datasets, and we varied the number of data points for each dataset. We summarized the prediction accuracy on the training and test data in Table 2. When there is more labelled data (first row), the semi-supervised learning performs similar to supervised learning in terms of test accuracy. Significantly, when the size of labelled data is much smaller (second row), the proposed semi-supervised model outperforms a fully supervised model in terms of test accuracy. It shows that the proposed semi-supervised method can perform as well as a fully supervised method when there is a lot of labelled training data available and outperforms fully supervised classification when fewer labelled data points are available. Note that the proposed model also learns a latent space that can be explored and used for generating compositions (aspects that are missing from a fully supervised model).

Table 2: Data efficiency of the semi-supervised learning compared versus supervised learning only. 'SS' and 'SL' refers to respective semi-supervised learning and supervised learning.

Labelled data	Unlabelled data	Validation data	Test accuracy (SS vs SL)
864	296	75	0.877 *vs* 0.870
247	790	198	0.841 *vs* 0.826

3.2 Latent space representation

3.2.1 Disentanglement from target property

Latent space for all data points and their ground truth labels are shown in Figure 4(a), and a kernel density estimate for the data distribution is depicted in Figure 4(b). One can see that single phase alloys are mixed up with multiple phase alloys in the latent space, implying that the single phase formation aspect has been disentangled from this latent representation. Although the input alloys consist of up to 30 different elements and live in a high-dimensional space, more than 84% data points are condensed into a small latent region ($z1 \in [-2.5, 2.5]$ and $z2 \in [-2, 2]$), as indicated by the kernel density plot in Figure 4(b). We also listed two pairs of high-entropy alloys in latent-space regions where data densities are the highest. Each pair of points is of similar latent variables but has different true labels again showing that the space captures properties other than the phase formation.
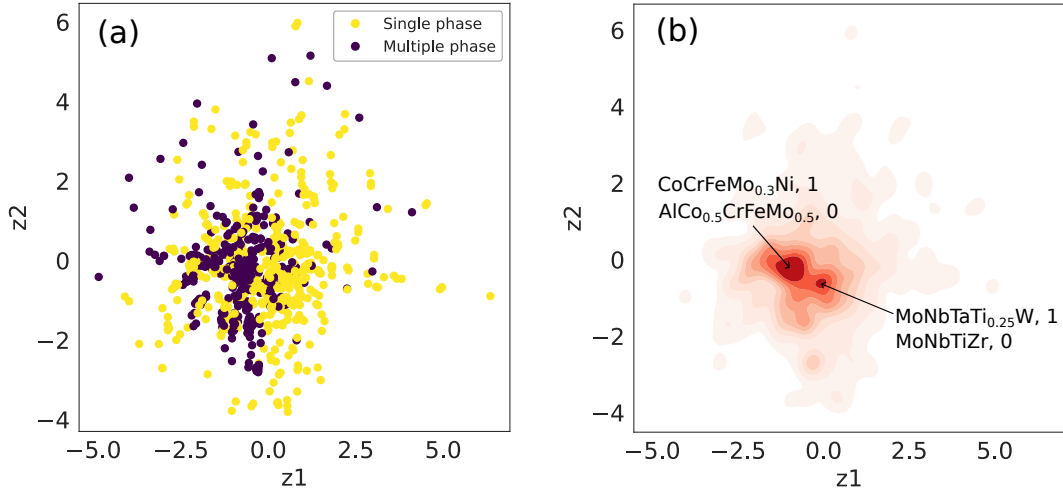
Figure 4: Disentangled latent representation (a) and Data point distribution (b). True labels for the HEAs are given as a binary presentation where '1' and '0' stand for single phase and multiple phase structures respectively.

3.2.2 Association with other properties

The proposed semi-supervised learning methodology allows us to disentangle the target property from the latent space, allowing the latent space to be implicitly associated with other properties. Figure 5 shows the association of learned latent space with two other properties. Figure 5(a) represents the number of elements for each alloy in the full dataset. It is found that HEAs with no less than four elements are concentrated in a smaller region of the latent space compared to simpler alloys with less than four elements. In Figure 5(b), we created three groups of elements based on their positions in the chemical periodic table, including magnetic, noble and refractory elements. One may argue that the exact group some elements belong to may vary by definition, but the purpose here is to group elements that at least share some similarity in atomic features. We then chose four elements from one group and created an equimolar four-element HEA. and generated its latent variables. One can find that HEAs for each group of elements are located in distinct regions in the high data density region of the latent space (shown in Figure 4. Hence, the model learned a general representation for elements across the chemical periodic table, which further explains why reconstructed alloys sometimes contain additional elements from the same group of elements and why high-dimensional composition features can be mapped into the more compact latent space.
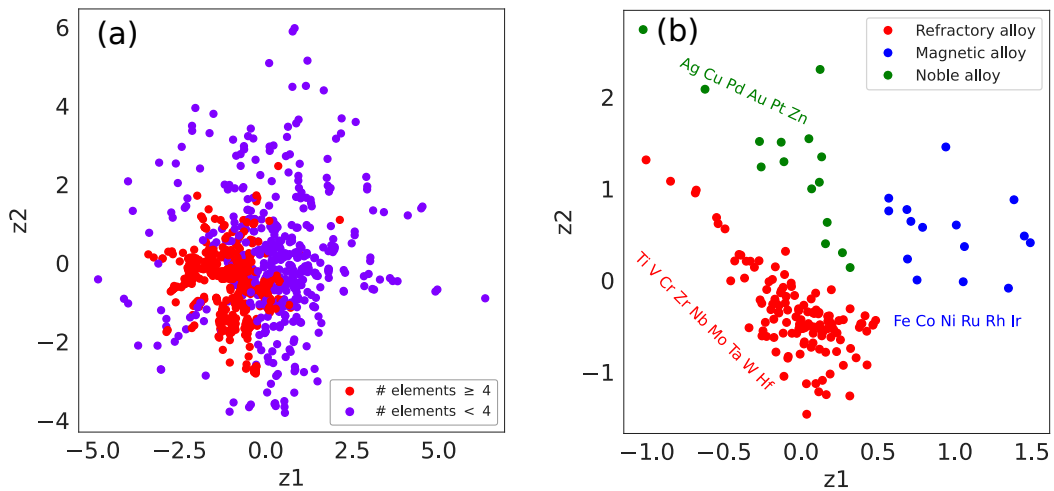

Figure 5: Number of elements distribution for all data points represented in the latent space (a), and Latent variables for four-element HEAs generated from three groups of elements (b). Element lists for each group are indicated by texts close to the corresponding point cloud.

3.3 Inverse materials design

The proposed semi-supervised variational autoencoder is versatile for inverse materials design. With the classifier head in the recognition model, one can carry out high-throughput virtual screening of materials to identity potential candidates with desirable target properties. However, there is a chance that no desired materials can be found even if a wide range of composition space is scanned, making this method inefficient.

Another method is to start from the latent space and a given probability for single-phase formation. The generative model will create an alloy based on latent variables and single-phase formation probability. The generated alloy should however be re-examined by the classifier to ensure the predicted probability agrees with the initial input probability. For example, if we aim to find single-phase refractory HEAs, we first locate a latent point, such as $(z1, z2) = (0.0, -0.8)$, among the refractory alloy blob as shown in Figure 5. For a formation probability of 0.9, the generated alloy is '$Ti_9V_{10}Nb_{19}Zr_{58}Ta_1Hf_2$', and the classifier predicts it to be single phase with a probability close to 1. If we set the probability as 0.1, the alloy generated becomes '$Cr_3Al_{12}Ti_{19}V_{21}Nb_{28}Zr_{13}Ta_5$', and the predicted probability is 0.04. Although the learned latent representation generalizes many different elements, caution should be taken if we would like to generate alloys in regions where labelled data are scarce, which can result in high uncertainty for the predicted target property. As an example, if we start from $(z1, z2) = (0.8, -0.5)$ which belongs to the magnetic alloy blob where data densities are low according to Figure 4(b). The generated alloy with an input probability of 0.1 is '$Fe_2Co_{81}Ta_{14}Pt_2$', yet the predicted probability turns out to be 0.71.
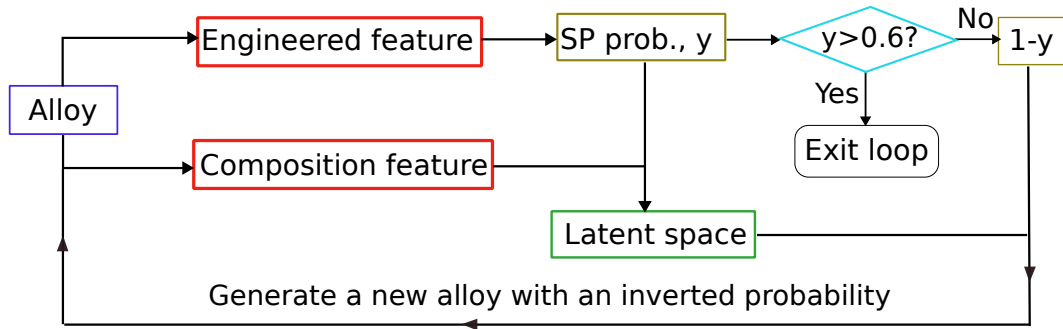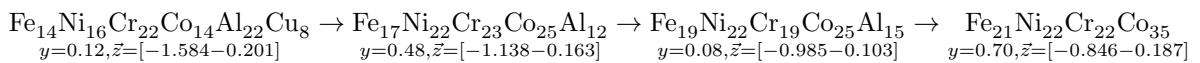

Figure 6: Workflow of the iterative process to search for single-phase alloys using the disentangled variational autoencoder. 'SP' stands for single phase.

This inconsistency between input probability and predicted probability for alloys generated from low-density regions, motivates the third method. In practice, one may have a specific alloy in mind, but the alloy does not give the desired property. The goal is to nudge the alloy composition in directions where the desired property can be achieved without significant changes of alloy constituents. An iterative process is proposed to find a single-phase alloy from an initial multi-phase alloy, as illustrated in Figure 6. The initial alloy is converted to engineered features which are used to predict the single-phase probability. The composition features converted from the alloy are combined the single-phase probability to generate its latent representation. We invert the single-phase probability, and the probability is concatenated with the latent representation to generate a new alloy. The loop ends when the single-phase probability predicted by the classifier is larger than a predefined cutoff (e.g. 0.6). We demonstrate this workflow using an initial alloy '$Al_{1.4}Co_{0.9}Cr_{1.4}Cu_{0.5}Fe_{0.9}Ni_1$' or equivalently '$Fe_{14}Ni_{16}Cr_{22}Co_{14}Al_{22}Cu_8$'. The evolution towards a single-phase alloy follows the sequence:

$$\underset{y=0.12,\vec{z}=[-1.584-0.201]}{Fe_{14}Ni_{16}Cr_{22}Co_{14}Al_{22}Cu_8} \rightarrow \underset{y=0.48,\vec{z}=[-1.138-0.163]}{Fe_{17}Ni_{22}Cr_{23}Co_{25}Al_{12}} \rightarrow \underset{y=0.08,\vec{z}=[-0.985-0.103]}{Fe_{19}Ni_{22}Cr_{19}Co_{25}Al_{15}} \rightarrow \underset{y=0.70,\vec{z}=[-0.846-0.187]}{Fe_{21}Ni_{22}Cr_{22}Co_{35}}$$

where the corresponding predicted single-phase probability y and latent variables $\vec{z}$ are shown right below the chemical formula of the alloy. One can see that after three iterations, the alloy eventually transforms to an alloy with a single-phase probability 0.70. In the first step, the alloy moves into a region without Cu, which enhances its single-phase probability. The second step detours around the target, transitioning to an alloy with reduced probability before it was finally inverted to a single-phase structure. The alloy evolution

occurs in latent regions close to the initial point, hence it preserves many original elements or introduces new elements that are learned to be similar to existing ones. This probabilistic process offers more control over the inverse design process and provides intuitive interpretability of inversion of materials for desired properties. One can see that with principal elements of Fe, Ni, Cr and Co, it is unfavorable to form single-phase structures when Al and Cu are added. Moreover, the compositions of Al and Cu are almost evenly replaced by Fe, Ni and Co compositions while Cr compositions remain almost unchanged in the alloy evolution process. It suggests that Cr also reduces the probability of single-phase formation when mixed with the three magnetic elements. This observation agrees very well with our previous work, indicating that Al and Cr are adverse to the single-phase formation for high-entropy alloy AlCrNiCoFe [9].

3.3.1 Post-hoc analysis for improved interpretability

The semi-supervised learning approach is intended to be interpretable by design to the extent that it disentangles the latent space into the properties of interest. We can add additional one layer of interpretability by through post-hoc analysis methods. In the recognition model, the classifier head takes as input the engineered features and predicts the single-phase formation. That is still a black-box process. We utilize existing post-hoc explainability approaches, specifically Shapley values (SHAP) — a widely used approach from cooperative game theory [28] — to provide insight which input features can be attributed to the predictions made by the classification head for each input alloy. Figure 7(a) shows the aggregated feature importance values across all samples in the test dataset. One can see that lower mixing entropy and atomic size differences, and higher melting temperature and bulk modulus are more likely to form a single-phase alloy. Figure 7(b) compares the eight engineered features for an original multi-phase alloy '$Fe_{14}Ni_{16}Cr_{22}Co_{14}Al_{22}Cu_8$' and the inverted single-phase alloy '$Fe_{21}Ni_{22}Cr_{22}Co_{35}$'. It is clear that the inverted alloy is pushed toward the direction where it is of a much smaller atomic size difference, lower mixing entropy and higher melting temperature and bulk modulus, in satisfactory agreement with the feature importance analysis.
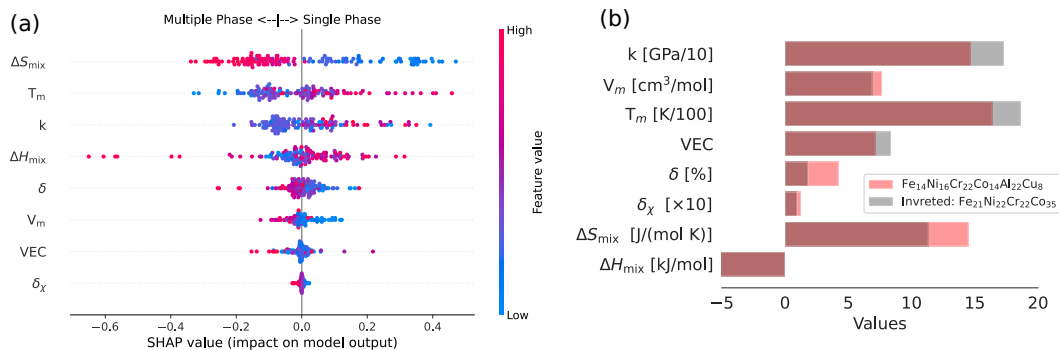

Figure 7: Post-hoc analyses for single-phase formation. (a) SHAP values for all data points and (b) Feature values for an original multi-phase alloy and its inverted single-phase alloy.

4 Conclusion and Outlook

We utilized probabilistic modeling recently developed in computer science community to inverse materials design. We introduced a disentangled autoencoder trained in a semi-supervised manner. This model can achieve better prediction accuracy when only a small amount of labelled data are available because it encodes expert-informed priors and leverages all useful information in the model. This disentangled framework is by design more interpretable since it offers a separate understanding for materials representations and the target property. Additional levels of explainability can be achieved via post-hoc analyses.

This approach was demonstrated on an experimental high-entropy alloy dataset. We focused on the inverse design of high-entropy alloys that are likely to form single-phase structures. The learned latent representations were found to be compact and disentangled from the target property, opening up endless design space with a separate and tunable target representation. With similar latent variables, it allows transformations across alloys with similar alloy constituents but distinct materials properties. The latent space is implicitly entangled with other properties, including the number of elements and alloy types. Using a well-trained disentangled variational autoencoder, inverse materials design can be conceived in three different ways, including high-throughput virtual screening using the classification/regression head, design from a

latent-space region and an iterative design process as benchmarked in this work.

Although a single target property is used in this work, this disentangled probabilistic modeling approach can be easily adapted for multiple properties as long as reasonable prior distributions for each property can be identified. The approach implemented in this work may lead to alloys with harmful or expensive elements. Hence, new methods to constrain material representations while searching for new materials are crucial to develop cost-effective materials with desired properties. Uncertainty and inconsistency in the inverse materials design necessitates an active learning approach where reliable materials validation should be incorporated, and uncertainty estimate and retraining algorithms should be developed. In this line of research, a human-in-the-loop intelligent interface may be beneficial to guide the search of new materials and identification of new data to improve the model, in particular for non-expert users. We would also expect to see extensions of this approach to other types of materials representations (e.g. atomic structures, microstructures and crystal graphs etc) and models for inverse design (e.g. diffusion models and generative adversarial networks).

Acknowledgments

References

[1] Li, Y.; Lu, Y.; Zhao, C.; Hu, Y.S.; Titirici, M.M.; Li, H.; Huang, X.; Chen, L. Recent advances of electrode materials for low-cost sodium-ion batteries towards practical application for grid energy storage. *Energy Storage Materials* **2017**; 7, 130–151.

[2] de Leon, N.P.; Itoh, K.M.; Kim, D.; Mehta, K.K.; Northup, T.E.; Paik, H.; Palmer, B.S.; Samarth, N.; Sangtawesin, S.; Steuerman, D.W. Materials challenges and opportunities for quantum computing hardware. *Science* **2021**; 372, eabb2823.

[3] Tahmasebi, E.; Alam, M.; Yazdanian, M.; Tebyanian, H.; Yazdanian, A.; Seifalian, A.; Mosaddad, S.A. Current biocompatible materials in oral regeneration: a comprehensive overview of composite materials. *Journal of Materials Research and Technology* **2020**; 9, 11731–11755.

[4] Merchant, A.; Batzner, S.; Schoenholz, S.S.; Aykol, M.; Cheon, G.; Cubuk, E.D. Scaling deep learning for materials discovery. *Nature* **2023**; 624, 80–85.

[5] Szymanski, N.J.; Rendy, B.; Fei, Y.; Kumar, R.E.; He, T.; Milsted, D.; McDermott, M.J.; Gallant, M.; Cubuk, E.D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C.J.; Persson, K.; Zeng, Y.; Ceder, G. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**; 624, 86–91.

[6] Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K.A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**; 1, 011002.

[7] Choudhary, K.; Garrity, K.F.; Reid, A.C.E.; DeCost, B.; Biacchi, A.J.; Hight Walker, A.R.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A.G.; Centrone, A.; Davydov, A.; Jiang, J.; Pachter, R.; Cheon, G.; Reed, E.; Agrawal, A.; Qian, X.; Sharma, V.; Zhuang, H.; Kalinin, S.V.; Sumpter, B.G.; Pilania, G.; Acar, P.; Mandal, S.; Haule, K.; Vanderbilt, D.; Rabe, K.; Tavazza, F. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput Mater* **2020**; 6, 1–13. Publisher: Nature Publishing Group.

[8] Pyzer-Knapp, E.O.; Pitera, J.W.; Staar, P.W.J.; Takeda, S.; Laino, T.; Sanders, D.P.; Sexton, J.; Smith, J.R.; Curioni, A. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **2022**; 8, 1–9.

[9] Zeng, C.; Neils, A.; Lesko, J.; Post, N. Machine learning accelerated discovery of corrosion-resistant high-entropy alloys. *Computational Materials Science* **2024**; 237, 112925.

[10] Wang, J.; Wang, Y.; Chen, Y. Inverse Design of Materials by Machine Learning. *Materials* **2022**; 15, 1811.

[11] Afzal, M.A.F.; Haghighatlari, M.; Ganesh, S.P.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *J. Phys. Chem. C* **2019**; 123, 14610–14618.

[12] Geng, Y.; van Anders, G.; Dodd, P.M.; Dshemuchadse, J.; Glotzer, S.C. Engineering entropy for the inverse design of colloidal crystals from hard shapes. *Science Advances* **2019**; 5, eaaw0514.

[13] Xian, Y.; Dang, P.; Tian, Y.; Jiang, X.; Zhou, Y.; Ding, X.; Sun, J.; Lookman, T.; Xue, D. Compositional design of multicomponent alloys using reinforcement learning. *Acta Materialia* **2024**; 274, 120017.

[14] Ma, W.; Cheng, F.; Xu, Y.; Wen, Q.; Liu, Y. Probabilistic Representation and Inverse Design of Metamaterials Based on a Deep Generative Model with Semi-Supervised Learning Strategy. *Advanced Materials* **2019**; 31, 1901111.

[15] Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **2018**; 4, eaap7885.

[16] Türk, H.; Landini, E.; Kunkel, C.; Margraf, J.T.; Reuter, K. Assessing Deep Generative Models in Chemical Composition Space. *Chem. Mater.* **2022**; 34, 9455–9467.

[17] Chen, C.T.; Gu, G.X. Generative Deep Neural Networks for Inverse Materials Design Using Backpropagation and Active Learning. *Advanced Science* **2020**; 7, 1902607. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.201902607.

[18] Rao, Z.; Tung, P.Y.; Xie, R.; Wei, Y.; Zhang, H.; Ferrari, A.; Klaver, T.; Körmann, F.; Sukumar, P.T.; Kwiatkowski da Silva, A.; Chen, Y.; Li, Z.; Ponge, D.; Neugebauer, J.; Gutfleisch, O.; Bauer, S.; Raabe, D. Machine learning–enabled high-entropy alloy discovery. *Science* **2022**; 378, 78–85.

[19] Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Shysheya, S.; Crabbé, J.; Sun, L.; Smith, J.; Nguyen, B.; Schulz, H.; Lewis, S.; Huang, C.W.; Lu, Z.; Zhou, Y.; Yang, H.; Hao, H.; Li, J.; Tomioka, R.; Xie, T. MatterGen: a generative model for inorganic materials design **2024**.

[20] Cheetham, A.K.; Seshadri, R. Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery. *Chem. Mater.* **2024**; 36, 3490–3495. Publisher: American Chemical Society.

[21] Yao, H.W.; Qiao, J.W.; Hawk, J.A.; Zhou, H.F.; Chen, M.W.; Gao, M.C. Mechanical properties of refractory high-entropy alloys: Experiments and modeling. *Journal of Alloys and Compounds* **2017**; 696, 1139–1150.

[22] Feng, R.; Zhang, C.; Gao, M.C.; Pei, Z.; Zhang, F.; Chen, Y.; Ma, D.; An, K.; Poplawsky, J.D.; Ouyang, L.; Ren, Y.; Hawk, J.A.; Widom, M.; Liaw, P.K. High-throughput design of high-performance lightweight high-entropy alloys. *Nat Commun* **2021**; 12, 4329.

[23] Yan, Y.; Lu, D.; Wang, K. Accelerated discovery of single-phase refractory high entropy alloys assisted by machine learning. *Computational Materials Science* **2021**; 199, 110723.

[24] Tandoc, C.; Hu, Y.J.; Qi, L.; Liaw, P.K. Mining of lattice distortion, strength, and intrinsic ductility of refractory high entropy alloys. *npj Comput Mater* **2023**; 9, 1–12.

[25] Pei, Z.; Yin, J.; Hawk, J.A.; Alman, D.E.; Gao, M.C. Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules. *npj Comput Mater* **2020**; 6, 1–8.

[26] Luengo, D.; Martino, L.; Bugallo, M.; Elvira, V.; Särkkä, S. A survey of Monte Carlo methods for parameter estimation. *EURASIP J. Adv. Signal Process.* **2020**; 2020, 25.

[27] Bingham, E.; Chen, J.P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; Goodman, N.D. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research* **2018**; .

[28] Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In I. Guyon; U.V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan; R. Garnett, eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. **2017**; pp. 4765–4774.