

Cite this: DOI: 00.0000/xxxxxxxxxx

Visualizing High Entropy Alloy Spaces: Methods and Best Practices

Brent Vela,^a Trevor Hastings,^{a†} Marshall Allen,^{a,b} and Raymundo Arróyave^a

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

1 tSNE and UMAP Algorithms

t-distributed Stochastic Neighbor Embedding (tSNE) is an unsupervised machine learning dimensionality reduction technique that has gained notoriety in the machine learning community for its ability to preserve both the global and local structure of the data. Since Van der Maaten and Hinton introduced tSNE in 2008¹, it has been used in various fields such as genetics², astronomy³, and of particular importance to this work, materials science^{4,5}. tSNE seeks to map the location of a data point i in high-dimensional space, x_i , to a location in lower dimensional space, y_i . This mapping must preserve the local structure of the high-dimensional data and thus must be aware of points neighboring point i in high-dimensional space, x_j , such that y_i and y_j are near each other in the embedding if x_i and x_j are near each other in high-dimensional space. Likewise, if x_i and x_j are far from each other in high-dimensional space, their corresponding embedded locations, y_i and y_j , are to be far from each other. To achieve this, the algorithm converts high-dimensional Euclidean distances between each pair of data points x_i and x_j into conditional probabilities $P_{j|i}$ via mapping them onto a Gaussian distribution. This Gaussian distribution is centered at x_i with a standard deviation σ_i . These conditional probabilities represent how *similar* two data points are to each other. This Gaussian distribution is used to calculate the conditional probability that x_i would be neighbors with x_j , $P_{j|i}$. However, this conditional probability $P_{j|i}$ must be normalized to account for imbalanced cluster populations. This normalization is achieved by dividing the Gaussian distribution by the sum of all conditional probabilities associated $P_{k|i}$ with the Gaussian centered at x_i .

$$P_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)} \quad (1)$$

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n} \quad (2)$$

The next step in the algorithm requires randomly projecting the high-dimensional space to 2-dimensions. Analogous to the high-dimensional case, the algorithm determines the Euclidean distance between every pair of points y_i and y_j in the low-dimensional embedding of the dataset and maps these to a Cauchy distribution centered at y_i . The use of the Cauchy distribution as opposed to the normal distribution is deliberate as using the Cauchy distribution solves *the crowding problem*¹ which was common in SNE, the predecessor to tSNE. The cost function for this minimization is C .

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq i} \exp\left(1 + \|x_k - x_l\|^2\right)^{-1}} \quad (3)$$

Points in the 2D embedding are then incrementally rearranged such that the Kullback-Leiber divergence between the joint probability distribution of the 2D-dimensional embedding approaches the joint probability distribution of the high dimensional representation of the data. This is achieved by minimizing the Kullback-Leibler divergence of the two distributions. The Kullback-Leibler divergence of the joint probabilities in the original space and the embedded space will be minimized by gradient descent.

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log\left(\frac{p_{j|i}}{q_{j|i}}\right) \quad (4)$$

Similar to t-SNE, UMAP⁶ is a non-linear dimensionality reduction technique. The UMAP algorithm has its foundations in Riemannian geometry. The Uniform Manifold Approximation and Projection (UMAP) algorithm and t-SNE both create a high-dimensional representation of the data, a 2-dimensional representation of the data, and then optimize the 2-dimensional embedding to be as similar to the high-dimensional representation as possible. In t-SNE this similarity-representation is the Kullback-Leibler divergence, while in UMAP this representation is a *fuzzy simplicial complex*; this fuzzy simplicial complex is used to learn the manifold structure of the high-dimensional data to be projected. This fuzzy simplicial complex can be understood as a

^a Materials Science and Engineering Department, Texas A&M University, College Station, TX.

† Corresponding Author, Trevor Hastings, trevorhastings@tamu.edu

weighted graph where edge weights represent the likelihood that two points are connected. The simplicial complex is defined by spheres of radius r extending from each point. In this graph, points with overlapping spheres are considered connected. However, the selection of this radius is not arbitrary. If the radius is too small there will be few connections in the graph representation of the data, resulting in small isolated clusters in the 2-dimensional embedding of the data. If the radius r is too large, unrelated data points will have connections between them and a meaningful embedding cannot be created. An appropriate radius r can easily be found in the ideal case when the data is uniformly distributed on the high-dimensional manifold, however, this is the ideal case. It is often the case that data is undersampled in some regions, and oversampled in other regions. Furthermore, this data is often noisy. The UMAP algorithm overcomes this issue by assuming that the data is uniform distribution on the manifold and assuming that the space expands or contracts where data is sparse or dense, respectively. To realize this change in the local metric of distance in practice, the radius r extending from each point is set individually, thus approximating a local metric of distance per point. As mentioned, the local notion of distance is defined by the density or sparsity of data in a particular area of the manifold. The radius is set as the distance between a point and its n th nearest neighbor. This technique results in an approximation of the manifold structure of the high dimensional data that preserves global structure in the data. Regarding the “fuzziness” of the simplicial complex and the weight of these connections, the UMAP algorithm decreases the likelihood of connection as the radius is extended. Finally, to ensure the local structure is preserved the algorithm requires that each point must be connected to at least its 1st nearest neighbor.

For both t-SNE and UMAP, once a polygonal embedding of a barycentric design space is generated for a system with a specific dimensionality and resolution, it can be reused for other design spaces of the same dimensionality and resolution. For instance, an embedding created for the Co-Cr-Fe-Ni-Mn system sampled at 5 at.% could be applied to the W-Re-Ta-Nb-Hf system also sampled at 5 at.%. However, if one wishes to project an alloy space sampled at 2 at.% and only has an embedding at 5 at.% (of the same dimensionality), a machine learning model could be trained to interpolate the x and y coordinates of the existing projection for the new resolution. However, in light of affine projections, these points are moot as the affine projection method is able to project design spaces sampled at non-uniform intervals to 2D.

2 Comparison of DRAs

To create an embedding that encompasses the entire alloy space, we use a similar method for both tSNE and UMAP. We grid sample compositions from the MPEA space to generate a hyper-tetrahedron of compositions. This means that the n -dimensional MPEA composition space is sampled at uniform increments, including unary, binary, up to and including $(n + 1)$ -nary. This hyper-tetrahedron of compositions is then processed by the tSNE and/or UMAP algorithm, projecting the high dimensional composition space to 2D. For a 2D composition space (ternary), an ideal 2D embedding is an equilateral triangle. A 3D composition space (quaternary) should have a square 2D reference embed-

ding, and a 4-dimensional embedding (quinary) should resemble a pentagon, and so forth.

tSNE DRA can also embed a hypercube similarly to UMAP, as shown in Figure 1. In the case of tSNE, the resultant projections also result in polygonal compositional embeddings. Depending on the perplexity and `n_iter` parameters chosen (analogous to UMAP’s nearest neighbors and number of epochs), it embeds high dimensional composition spaces into regular polygons of n sides or concave polygons with $2n$ sides. These tSNE projections can be interpreted in the same way as UMAPs, i.e. alloys mapped near vertices are rich in a particular element, alloys on the edge connecting two vertices are rich in those 2 elements, and alloys in the central regions are chemically complex.

However, tSNE has a significant shortcoming. tSNE tends to skew compositions with a majority constituent element towards the edges, resulting in overcrowding, as shown in Figure 1. While UMAP also skews compositions with a majority constituent element towards the edges, it better preserves both local and global structures. This allows chemically complex alloys rich in a particular element to be plotted closer to central regions resulting in less overcrowding. UMAP’s resistance to overcrowding is well-documented and is attributed to its superior preservation of structure⁷.

Barycentric design spaces for alloy compositions form simplices. For example, in the case of binary alloys, this corresponds to a Gibbs triangle, while for ternary alloys, it corresponds to a Gibbs tetrahedron, which is a 3D simplex. Importantly, the entire n -simplex, including its interior, constitutes a well-defined manifold. For instance, a 2D simplex (triangle) is a 2-manifold embedded in 3D space, and a 3D simplex (tetrahedron) is a 3-manifold embedded in 4D space^{8–10}. In higher dimensions, the simplex similarly spans an n -dimensional manifold embedded in an $(n + 1)$ -dimensional Euclidean space.

Because simplices are concave polytopes and concave polytopes can be decomposed using simplicial complexes (a mesh of smaller simplexes) by definition¹¹, they are particularly amenable to projection with UMAP. The UMAP algorithm⁶ employs simplicial complexes to approximate the structure of high-dimensional data. Specifically, UMAP constructs a fuzzy topological representation by building a simplicial complex, first forming local neighborhoods and then connecting them via shared simplices to capture the global data manifold. In our case, UMAP is applied to the high-dimensional simplex of alloy compositions, which allows the algorithm to preserve both local and global structures during the dimension reduction process. Importantly, UMAP leverages the inherent manifold structure of simplices (n -dimensional objects embedded in $(n + 1)$ -dimensional Euclidean space), embedding the n -dimensional compositional data into a lower-dimensional space while retaining topological relationships.

When applying UMAP to the high-dimensional simplex of alloy compositions, the algorithm works by preserving the local distances and relationships between neighboring points while projecting the n -dimensional manifold onto a 2D plane. UMAP exploits the inherent manifold structure of simplices, as they are n -dimensional manifolds embedded in $(n + 1)$ -dimensional Euclidean space⁶.

When hyperparameters are tuned correctly, in the 2D projec-

tion produced by UMAP, the vertices of the simplex, which correspond to pure elementals, are mapped such that the distances between them are maximized, often resulting in a regular polygonal shape. Although the original high-dimensional space is fully n -dimensional, UMAP "flattens" this manifold into a 2D representation while preserving the relative distances between compositions. This enables a clearer visual interpretation of the relationships between different compositions and their correlations with material properties.

Linear DRA techniques, such as PCA, do not project barycentric design spaces in a symmetric, geometric manner that resembles a polygonal embedding. As shown in Figure 1, while PCA has some clustering ability, it does not intuitively embed compositions in a polygonal embedding like tSNE and UMAP and affine projections. Specifically, PCA fails to capture global structure in the data. By global structure, we mean how clusters or groups of data points that are far apart in high-dimensional space are arranged relative to each other in the lower-dimensional projection. The global structure in a barycentric design space is well represented by a polygon, as unaries should be separated as much as possible, and higher entropy alloys can be represented in central regions.

PCA cannot represent this nonlinear global structure. In contrast, UMAP preserves both local structure (i.e., relationships between neighboring points) and global structure (i.e., the spatial arrangement of clusters relative to one another). UMAP, for instance, embeds chemically complex alloys in the center of the projection while alloys with a majority element are positioned toward the corners, creating a geometrically intuitive 2D map that is particularly useful for alloy design spaces.

These points about DRAs are moot however, as upon review we have found that affine projections, as described in the main manuscript, accomplish a more regular polygonal embedding at a significantly lower computational cost. However, we reiterate, the method of projecting the high-dimensional barycentric alloy spaces is just a means to an end. The contribution of this work lies in how these projections are used for alloy design.

3 Information Loss and Overcrowding in DRAs

No projection is without some information loss, whether it be tSNE, UMAP, or affine projection. Therefore there will always be some degree of overcrowding present in a 2D projection of a high dimensional compositions space. DRAs are only appropriate for plotting *major* alloying constituents i.e. alloying agents that are present in the alloy at concentrations near 5 at.%. When considering the effect of *minor* alloying agents such as Boron, Silicon and Carbon that generally appear in alloys at concentrations below 1 at.% these projection techniques are not appropriate as the effect of the major alloying components will overshadow the effect of any minor alloying component. To better visualize the effect of minor alloying compositions we recommend using multiple projections as shown in Figure 2. In this Figure 2a all points represent an alloy, but all points have 0.5% carbon added. In Figure 2b all points represent an alloy but all points have 1 % carbon added.

4 Topology of Alloy Design Spaces

Topologically, an n -simplex is homeomorphic to an n -dimensional ball¹⁰. While this may not appear topologically interesting on its own, the topology can become more complex when constraints are applied. For example, consider applying a constraint that restricts the design space shown in Figure 3. Any alloy that does not have a single BCC phase at 2000°C is considered infeasible. Additionally, applying a solidification range constraint, where only alloys with a solidification range narrower than 100 K are considered feasible, further restricts the design space. These properties were predicted using Thermo-Calc's TCHEA6 database.

As a result, multiple feasible regions emerge that are not topologically connected within the design space—there is no path to incrementally move from one feasible region (or "island") to another via small composition changes (e.g., 5 at.% increments, which is the resolution of this design space). This discontinuous topology has important implications for functionally graded material design¹², which is a topic of ongoing research.

For instance, the diagram highlights that 13 V-rich alloys form a connected set that can be compositionally graded to one another. Similarly, a larger set of feasible alloys can be compositionally graded within their region, though they remain isolated from other parts of the design space.

Therefore, while the topology of the entire barycentric space is homeomorphic to an n -ball, the feasible space within it can exhibit more complex, disconnected topologies. Representing these topologies on an interpretable 2D projection helps us intuitively understand the connectivity within the design space and its dependence on alloy chemistry.

5 Visualizing Databases

High entropy alloys are perhaps the material class most amenable to UMAP construction as their definition of disorder naturally lends them to a visual capable of integrating several elements. The HEA space is so large that any materials discovery project needs to apply judicious constraints to the design space for practicality. A sensible starting point is to identify a crystal structure of interest, as that will dictate the possible constituents, mechanical procedures like grinding and polishing, and reasonable properties of interest. Unique to HEAs is its definition based on atomic fractions. A HEA design project almost certainly carries atomic fraction maximums at its onset. UMAP embeddings are well suited for overlaying several sets of data points, each of which that are subsets of the previous, based on some project constraint.

Using the database curated by Borg et al.¹⁴, a selection of BCC HEAs was identified for a UMAP embedding of the 8-element Al-Ti-V-Zr-Nb-Mo-Hf-Ta HEA space. In addition, a list of example elemental maximums was applied to the composition space. For example, candidate alloys in Figure 4 cannot have Zr present above 40 at.%. Alloys that pass these compositional constraints are shown in light grey whereas alloys that fail these constraints are shown in dark grey. In Figure 4, experimental compressive yield strength data is plotted on top of an AS-UMAP. The color-axis represents the measured yield strength. This case study showcases a single objective problem: the maximization of ultimate strength under compression. With multiple objectives, an array of materials becomes relevant as they dominate all others, the

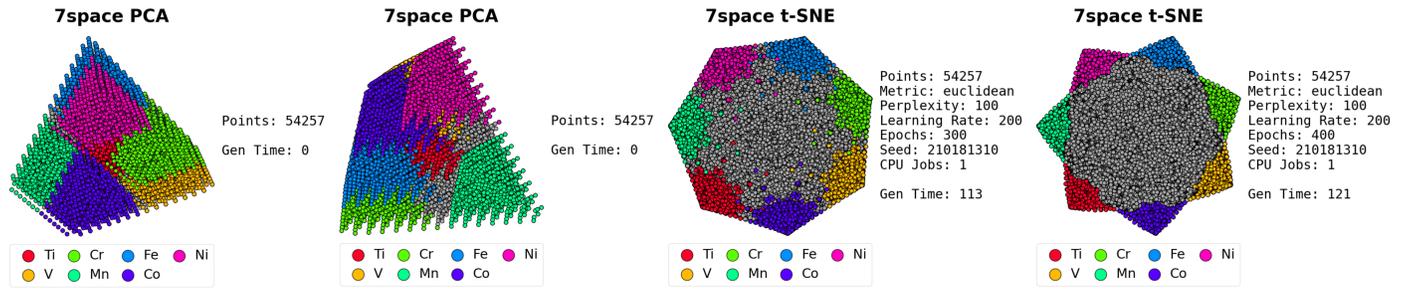


Fig. 1 Alternative embeddings: Two PCA examples with different seeds; two tSNE examples with varying perplexity and iterations. Each dot in these projections represent a unique composition. If a dot is colored it signifies the concentration of a particular element is ≥ 50 at.%.

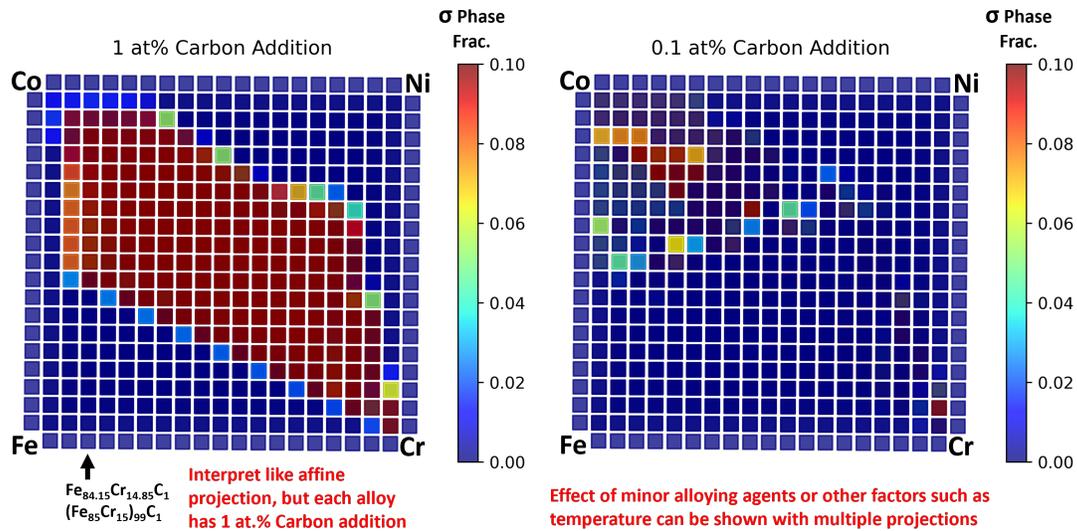


Fig. 2 In order to view the effect of minor alloying agents or other factors such as temperature, multiple projections can be used. a) The FeCrCoNi alloy space sampled at 5 at.% with 1 at.% C additions in each alloy. b) The FeCrCoNi alloy space sampled at 5 at.% with 0.1 at.% C additions in each alloy.

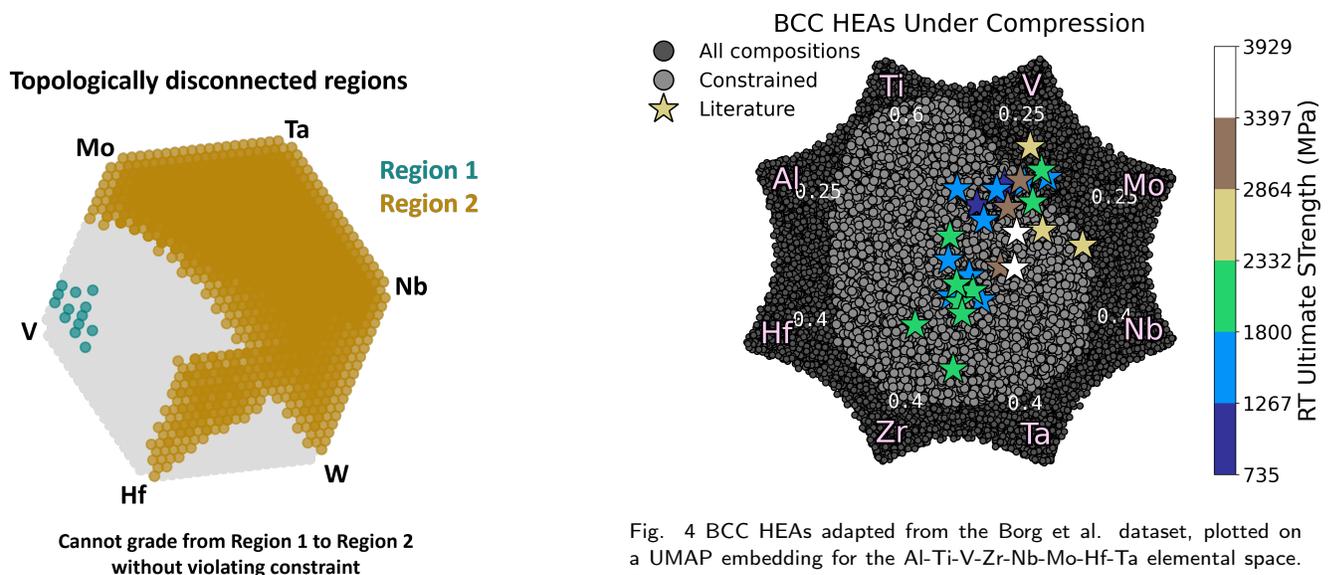


Fig. 3 A affine projection of the WNbTaMoVHf alloy space with 2 constraints applied to it: 1) Alloys must have a single BCC phase at 2000°C and 2) Alloys must have a solidification range narrower than 100K. These constraints are derived from Ref¹³.

Fig. 4 BCC HEAs adapted from the Borg et al. dataset, plotted on a UMAP embedding for the Al-Ti-V-Zr-Nb-Mo-Hf-Ta elemental space. The constrained dataset has maximum atomic fractions according to the annotations near each element vertex. Experimental data is sorted by color according to ultimate strength (MPa) from room temperature compression tests.

Pareto front of the dataset. Using UMAPs similar to Figure 4, one can showcase where the Pareto points are over multiple iterations of experiments. Figure 4 also provides a visual for the degree to which literature data violates the self-imposed constraints for a project, such as a maximum atomic fraction for an element that would easily oxidize.

In Figure 4, a database is used to provide information for an existing project. It provides a literature reference in its chosen objective manifold that the project may or may not intersect. What if we want to inform a project at the design level based on existing databases? This is already a common strategy amongst physics disciplines: cataloging existing experiments across multiple teams to determine which regions can be of scientific interest merely by virtue of being *unexplored*. UMAPs can be an excellent tool for these works - one such example is in the field of shape memory alloys.

Shape memory alloys (SMAs) have seen a large movement towards phase engineering studies, ever since a link was derived between their hysteresis and their crystallographic parameters; by utilizing a formal approach to SMAs with a basis in linear algebra, authors have been able to craft SMAs with abnormally long fatigue lives¹⁵. However, current approaches to find new ‘near-zero hysteresis’ SMAs are dependent on existing literature¹⁶. The large number of compositions and possible tertiary elements can make it difficult to determine which regions of SMA research are lacking in experimental data. Furthermore, a notable roadblock in high-temperature SMA (HTSMA) research has been the lack of experimental data in certain alloy systems, due to the difficult nature of some of the tests. Both of these systems could be supplemented with UMAP visualizations, to gauge a *formal literature assessment* of the field (beyond what is typically limited to long review papers or supplemental tables of compositions).

In Figure 5, high-temperature SMAs (HTSMAs) are plotted as an example subset of SMA research. Nine commonly seen elements in HTSMA research were applied to a UMAP embedding based on reported transformation temperatures from a comprehensive HTSMA review¹⁷. Figure 5 (top) includes two databases of reported HTSMAs. A select number of SMAs have been annotated with their elemental subset, a note of the author, and the year of publication. Following the tutorial for UMAPs in *Methods*, a reader can recognize probable regions for compositions on the graph despite not possessing a database of direct compositions. Ti-Ni-X alloys are in a cluster, a binary Ti-Pd example is directly between the Pd and Ti vertices, an HTSMA primarily consisting of uranium is directly by its vertex, and an HTSMA with 5 constituents is closer to the center of the AS-UMAP than any of the other data points.

Note that the positions of components around the vertices are completely arbitrary and *selected by the author*. UMAP generation provides a random solution to the embedding problem built on a specific seed. For typical *asymmetric* embeddings, this arrangement would have to be changed at the embedding level (e.g. an author would have to run the script multiple times with different seeds to obtain one they desire aesthetically). However, in these types of case studies involving atomic fractions, it makes more sense to plot the entire phase space of possibilities (given a maximum atomic resolution). This results in every column of

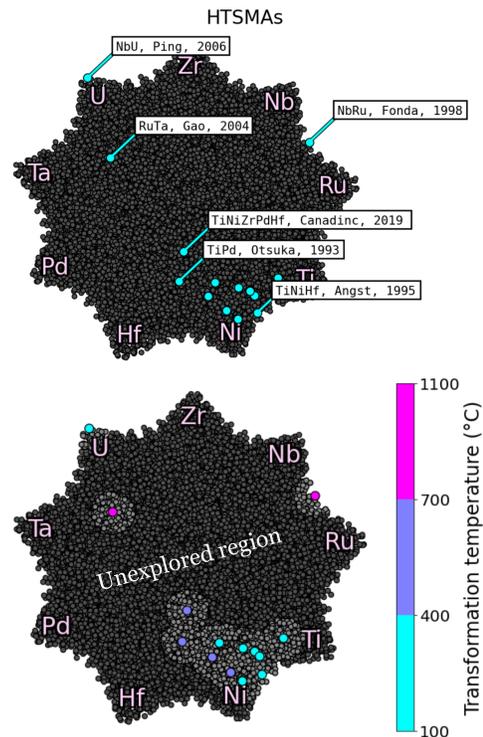


Fig. 5 High-temperature shape memory alloys embedded in a UMAP, adapted with data from Karaman et al. and Canadinc et al. (top) The space of Ti-Ni-Zr-Nb-Ru-Pd-Hf-Ta-U HTSMAs with highlighted authors. (bottom) Regions of compositions similar to known alloys in literature are highlighted, and author data points are color-coded according to their three regimes of austenite-martensite transformation temperatures.

data being *symmetric* i.e. every composition from 0% to 100% is somewhere on the graph. After the UMAP embedding has finished, if an author would rather change vertices, they can simply rename and rearrange the columns to their liking. A study specifically focusing on Ni-Ti, for example, might intentionally choose to separate those vertices to better visualize small changes away from a $Ti_{50}Ni_{50}$ composition.

In Figure 5 (bottom), the HTSMAs have been categorized by their austenite-martensite transformation temperatures. In addition, data points within the UMAP embedding have been lightly colored within a certain distance of known experimental data, to accentuate compositional regions that are lacking in experiments. Note that this distance is Euclidean *in the UMAP embedding*, using only x and y coordinates and not an Euclidean distance calculated from the dimensionality of the compositions themselves. This visualization has the large advantage of revealing unexplored regions even accounting for additional elements that may not be part of the original dataset.

Figure 5 illustrates several unexplored regions in the SMA design space that authors could pursue. A sufficiently large database of HTSMAs over the past 25 years could be split up into 3 temperature regions across the same UMAP embedding (removing the color bar and using 3 graphs). This would provide a multidimensional view of composition, status in literature, transformation temperature, and a property of choice such as hysteresis width in (°C)—like that shown in Figure 4—a very efficient conflagration of information at a glance.

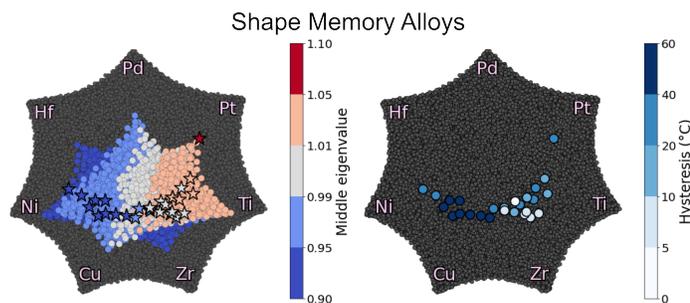


Fig. 6 Shape memory alloys embedded in a UMAP including a predictive model for middle transformation eigenvalues, adapted from Zadeh et al. (left) The SMA space with experimental data points overlaid on top of model predictions. (right) Values of Hysteresis widths for those experimentally reported SMAs.

Figure 6 (left) uses a 7-element UMAP representation of NiTi SMAs with several common tertiary additions. In this example the elements' vertices were intentionally separated. Experimental data points (the stars) have been plotted against a background of model approximations of austenite-martensite transformation matrix eigenvalues (the technical indicator for low hysteresis), taken from a database of Shape Memory Alloys¹⁸. Like before, it is readily apparent which regions of the embedding are lacking in experimental data. In this example, a UMAP provides a novel view of the model's predictions in conjunction with experimental data. Each of the modeled regions in this case overlap, due to the nature of the embedding. Figure 6 (right) is a similar UMAP with reported hysteresis values of the same experimental SMAs. The strategy in the UMAPs in Figure 6 for SMAs could equally be applied to copper-based SMAs, iron-based SMAs, manganese-nickel-based magnetic SMAs, or simply reduced portions of the extensive NiTi-based SMA literature, focusing on tertiary additions of interest.

6 Visualizing Combinatorics: Polymers

Determining the shape or conformation of polymers has been pivotal to their development for applications; this is typically combined with molecular weight measurements from chromatography and radii of gyration from light scattering experiments^{19,20}. As shown in *Methods*, the vertices of a UMAP are inherently unrelated to their materials science concept. They are simply fractions of a whole: this makes the visualization readily accessible to distributions of data, like that of particle sizes, grain sizes, or in this case, molecular weight spreads.

A simple expression for the radius of gyration based on the freely jointed chain model for linear polymers is as follows:

$$\langle R_g \rangle = \sqrt{\frac{1}{6}Nb^2} \quad (5)$$

One can then express this as a root mean square of the radius of gyration for a distribution of polymer chains in solution by summing over their fractional occurrences. As an example, fractionated polyethylene, composed of 6 monodisperse fractions, makes up a solution with molecular weights of 10,000 through 60,000 (g/mol). The vertices on a UMAP would then correspond to a completely monodisperse solution, utilizing the equation above, with increasing values of R_g with increasing molec-

ular weight. What isn't as obvious at a glance is how this value would change as the distribution does. Using these 6 molecular weights and a resolution of 5% fractions, Figure 7 shows how the root mean square radius of gyration changes with modifications to the distribution. For the purposes of this model, the persistence length is equal to the length of the monomeric unit (0.154 nm, 28 g/mol).

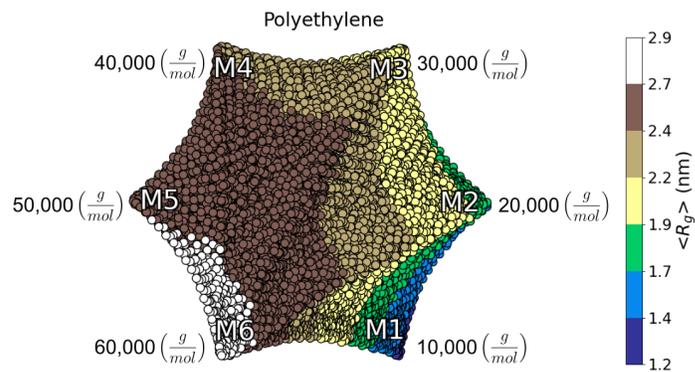


Fig. 7 A polyethylene polymer composed of 6 monodisperse fractions, with root mean square radii of gyration calculated across its distribution of molecular weights. The data points at the vertices are equivalent to the calculation of R_g for a single chain using that molecular weight.

In cases like these where the visual of a distribution is all that's desired, the ability to precisely label the vertices becomes less pivotal, making this an amenable application for UMAPs of a hypercube with a large number of dimensions, approaching the shape of a circle. Rather than label the monodisperse fractions explicitly, a gradient of molecular weights with radial tick marks could describe many types of polymers. Furthermore, soft matter authors could express properties in terms of other models that better represent polymers with cross-linking or other conformal geometries, via common modifications to the freely jointed chain model²¹. Otherwise identical distributions could be plotted side-by-side, with varying chain models, providing insight into how the choice of model affects predictions of a property over various molecular weight distributions.

Author contributions

BV: Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing Original Draft, Writing Review & Editing. TH: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing Original Draft, Writing Review & Editing. MA: Visualization. RA: Funding Acquisition, Project Administration, Resources, Supervision, Validation, Writing Review & Editing

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code associated with this work is available at the following repository: DOI: 10.24433/CO.7775216.v1

Acknowledgements

BV acknowledges Grant No. NSF-DGE-1545403 (NSF-NRT: Data-Enabled Discovery and Design of Energy Materials, D³EM) and Grant No. 1746932. RA acknowledge support from NSF through Grant No. DMREF-2323611. High-throughput CALPHAD calculations were carried out in part at the Texas A&M High-Performance Research Computing (HPRC) Facility.

References

- 1 L. Van der Maaten and G. Hinton, *Journal of machine learning research*, 2008, **9**, 2579–2605.
- 2 W. Li, J. E. Cerise, Y. Yang and H. Han, *Journal of bioinformatics and computational biology*, 2017, **15**, 1750017.
- 3 F. Anders, C. Chiappini, B. X. Santiago, G. Matijević, A. B. Queiroz, M. Steinmetz and G. Guiglion, *Astronomy & Astrophysics*, 2018, **619**, A125.
- 4 B. Vela, S. Mehalic, S. Sheikh, A. Elwany, I. Karaman and R. Arroyave, *Additive Manufacturing Letters*, 2022, **3**, 100085.
- 5 T. Z. Khan, T. Kirk, G. Vazquez, P. Singh, A. Smirnov, D. D. Johnson, K. Youssef and R. Arroyave, *Acta Materialia*, 2022, **224**, 117472.
- 6 L. McInnes, J. Healy and J. Melville, *arXiv preprint arXiv:1802.03426*, 2018.
- 7 M. Njue and B. Franklin, 2020.
- 8 J. A. Hillman, *Mathematische Zeitschrift*, 1996, **223**, 473–481.
- 9 J.-D. Boissonnat, R. Dyer, A. Ghosh, A. Lieutier and M. Wintraecken, *Discrete & Computational Geometry*, 2021, **66**, 666–686.
- 10 A. Hatcher, *Algebraic topology*, Cambridge University Press, Cambridge, 2002, pp. xii+544.
- 11 B. Büeler, A. Enge and K. Fukuda, *Polytopes—combinatorics and computation*, 2000, pp. 131–154.
- 12 M. Allen, T. Kirk, R. Malak and R. Arroyave, *Journal of Mechanical Design*, 2022, **144**, 041704.
- 13 C. Acemi, B. Vela, E. Norris, W. Trehern, K. C. Atli, C. Cleek, R. Arroyave and I. Karaman, *Acta Materialia*, 2024, 120379.
- 14 C. K. Borg, C. Frey, J. Moh, T. M. Pollock, S. Gorsse, D. B. Miracle, O. N. Senkov, B. Meredig and J. E. Saal, *Scientific Data*, 2020, **7**, 430.
- 15 H. Gu, L. Bumke, C. Chluba, E. Quandt and R. D. James, *Materials Today*, 2018, **21**, 265–277.
- 16 W. Trehern, R. Ortiz-Ayala, K. Atli, R. Arroyave and I. Karaman, *Acta Materialia*, 2022, **228**, 117751.
- 17 J. Ma, I. Karaman and R. D. Noebe, *International Materials Reviews*, 2010, **55**, 257–315.
- 18 S. H. Zadeh, C. Cakirhan, D. Khatamsaz, J. Broucek, T. D. Brown, X. Qian, I. Karaman and R. Arroyave, 2024.
- 19 L. Yuan, N. Hamidi, S. Smith, F. Clemons, A. Hamidi and C. Tang, *European Polymer Journal*, 2015, **62**, 43–50.
- 20 K. Terao and J. W. Mays, *European Polymer Journal*, 2004, **40**, 1623–1627.
- 21 I. W. Hamley, *Introduction to soft matter : synthetic and biological self-assembling materials.*, John Wiley & Sons, 2007.