DiSCoVeR: a Materials Discovery Screening Tool for High Performance, Unique Chemical Compositions

Sterling G. Baird[a], Tran Q. Diep[b], Taylor D. Sparks[a,*]

[a]*Department of Materials Science and Engineering, University of Utah, Salt Lake City, UT 84108, USA*
[b]*Department of Chemical Engineering, Brigham Young University, Provo, UT 84604, USA*

Abstract

We present Descending from Stochastic Clustering Variance Regression (DiSCoVeR) (`https://github.com/sparks-baird/mat_discover`), a Python tool for identifying and assessing high-performing, chemically unique compositions relative to existing compounds using a combination of a chemical distance metric, density-aware dimensionality reduction, and clustering. We introduce several new metrics for materials discovery and validate DiSCoVeR on Materials Project bulk moduli using compound-wise and cluster-wise validation methods. We visualize these via multi-objective Pareto front plots and assign a weighted score to each composition where this score encompasses the trade-off between performance and density-based chemical uniqueness. We explore an additional uniqueness proxy related to property gradients in chemical space. We demonstrate that DiSCoVeR can successfully screen materials for both performance and uniqueness in order to extrapolate to new chemical spaces.

Keywords: machine learning, uniform manifold approximation and projection, optimization, earth mover's distance, Wasserstein distance

---

1. Introduction

Guided materials discovery examples have been increasingly prevalent in the literature. Some of these are experimental [1–9] and computational [10, 11] adaptive design schemes using high-throughput experimental [5, 12–19] or computational (e.g. density functional theory (DFT) [20–29] and finite element modeling [30, 31]) methods. Extraordinary predictions, or predictions which perform close to or better than top performers in the training data are rarer [32–34]. Kauwe et al. [35] describes how it is even rarer to discover materials that are fundamentally (as opposed to incrementally) different from existing materials, i.e. discover new chemistries. A suite of regression models are available for use as the backbone for a materials discovery project. A non-exhaustive list ordered from oldest to newest by journal publication year includes GBM-Locfit [36], CGCNN [37], MEGNet [38], wren [39], GATGNN [40], iCGCNN [27], Automatminer [41], Roost [42], DimeNet++ [43], Compositionally-Restricted Attention-Based Network (CrabNet) [44], and MODNet [45], each with varying advantages and disadvantages.

Many of the algorithms used for materials discovery in the literature are Euclidean-based Bayesian optimization schemes which seek a trade-off between high-performance and high-uncertainty regions [4, 9, 11, 29, 34, 46–51], thereby favoring robust models and discovery of better candidates, but not explicitly favoring discovery of novel compounds.

Kim et al. [52] introduced two metrics for materials discovery: predicted fraction of improved candidates and cumulative maximum likelihood of improvement. These metrics are geared at identi-

---

[*] Corresponding author.

Email address: `sparks@eng.utah.edu` (Taylor D. Sparks)

fying "discovery-rich" and "discovery-poor" design spaces in the context of high-performance rather than chemical distinctiveness.

In this work, we introduce the Descending from Stochastic Clustering Variance Regression (DiSCoVeR) algorithm, which unlike previous methods, screens candidates that have a high probability of success while enforcing – through the use of a novel loss function – that the candidates exist beyond typical materials landscapes *and* have high performance. In other words, DiSCoVeR acts as a multi-objective screening where the promise of a compound depends on both having desirable target properties and existing in sparsely populated regions of the cluster to which it's assigned. This approach then favors discovery of novel, high-performing chemical families.

2. Methods

DiSCoVeR depends on clusters exhibiting homogeneity with respect to chemical classes, which we enforce via a recently introduced distance metric: Element Mover's Distance (ElMD) [53]. Dimensionality reduction algorithms such as Uniform Manifold Approximation and Projection (UMAP) [54] or t-distributed stochastic neighbor embeddings [55] can then be used to create low-dimensional embeddings suitable for clustering algorithms such as Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN*) [56] or k-means clustering [57].

Finally, these can be fed into density estimator algorithms such as Density-preserving Uniform Manifold Approximation and Projection (DensMAP) [58] a UMAP variant or kernel density estimation [59, 60] where density is then used as a proxy for chemical uniqueness.

Additionally, we describe our data and validation methods. By combining a materials suggestion algorithm and DiSCoVeR, it is possible to assess the likelihood of a new material existing relative to known materials.

The workflow for creating chemically homogeneous clusters is shown in Figure 1.

2.1. Chemically Homogeneous Clusters

How are chemically homogeneous clusters achieved? The key is in the dissimilarity metric used to compute distances between compounds. Recently, ElMD [53] was developed based on Earth Mover's or Wasserstein Distance; ElMD calculates distances between compounds in a way that more closely matches chemical intuition. For example, compounds with similar composition templates (e.g. XY_2 as in SiO_2, TiO_2) and compounds with similar elements are closer in ElMD space. In other words, clusters derived from this distance metric are more likely to exhibit in-cluster homogeneity with respect to material class which in turn allows in-cluster density estimation to be used as a proxy for novelty.

In this work, we use UMAP for dimensionality reduction and HDBSCAN* for clustering similar to the work by Hargreaves et al. [53][1] which successfully reported clusters of compounds that match chemical intuition.

2.2. Proxies for Chemical Uniqueness

2.2.1. Density-preserving Uniform Manifold Approximation And Projection

A multivariate normal probability density function is assigned to each datapoint embedded in DensMAP space (Eq. (1)):

$$e^{-\frac{1}{2}(X-\mu)\cdot\frac{1}{\Sigma}\cdot(X-\mu)} \tag{1}$$

where X, μ, Σ, and $\cdot$ represent DensMAP embedding position at which to be evaluated, train or validation DensMAP embedding position, covariance matrix, and tensor product, respectively.

The covariance matrix used in this work is given by Eq. (2):

$$\begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix} \tag{2}$$

where r represents extracted DensMAP radius.

We evaluate the sum of densities contributed by all of the training points evaluated at each of the

---

[1] In Hargreaves et al. [53], Density-based Spatial Clustering of Applications with Noise [61] was used instead of HDBSCAN*.

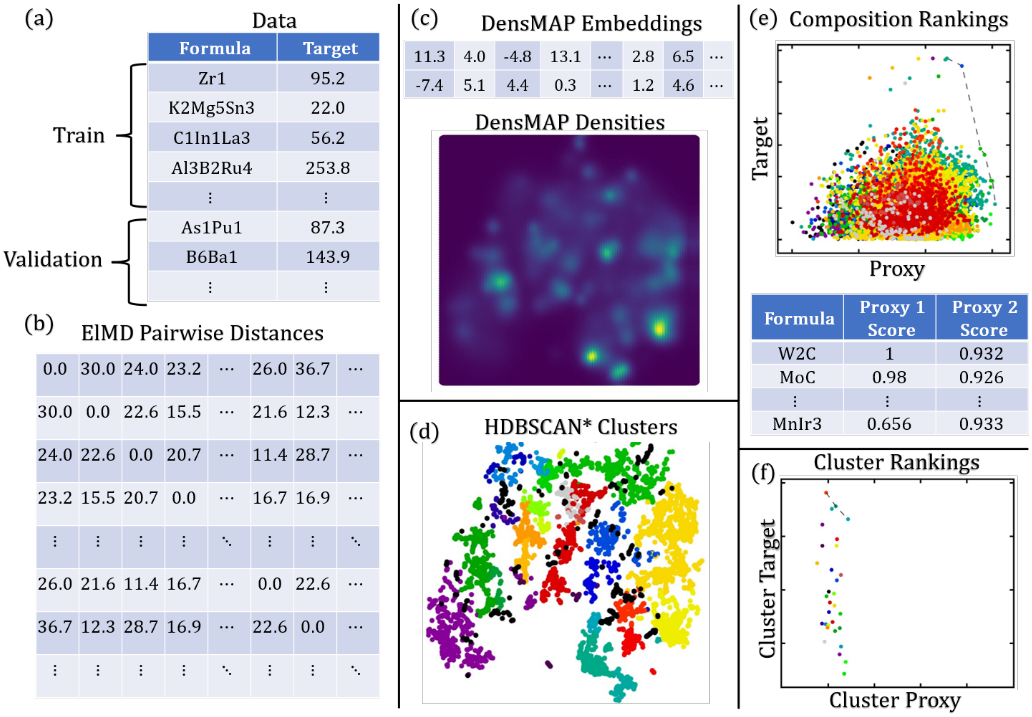


Figure 1: DiSCoVeR workflow to create chemically homogeneous clusters. (a) Training and validation data. (b) ElMD pairwise distances. (c) DensMAP embeddings and DensMAP densities. (d) Clustering via HDBSCAN*. (e) Pareto plot and discovery scores. (f) Pareto plot of cluster properties.

validation locations (Eq. (3)):

$$\sum_{i=1}^{n_{\text{train}}} e^{-\frac{1}{2}\backslash(X_{v,j}-\mu_{t,i})\cdot\frac{1}{\Sigma_{t,i}}\cdot(X_{v,j}-\mu_{t,i})} \tag{3}$$

where $X_{v,j}$, $\mu_{t,i}$, $\Sigma_{t,i}$, $\cdot$, and n_{train} represent j-th validation DensMAP embedding position at which to be evaluated, i-th train DensMAP embedding position, i-th train covariance matrix, tensor product, and total number of train points, respectively. By doing so, we obtain a proxy for chemical uniqueness relative to existing materials. By combining high-fidelity CrabNet predictions of bulk modulus with DensMAP validation densities, we extract a list of promising compounds at the Pareto front – the line or "front" at which the trade-off between performance and chemical uniqueness is optimal. CrabNet predictions have been shown to be comparable to state-of-the-art composition-based materials regression schemes, and since structure is often not known during a materials discovery search, CrabNet is a reasonable model choice. One partial workaround for the limitation of structure being unknown a-priori has been explored in the Bayesian Optimization With Symmetry Relaxation algorithm [62], which may be of interest to incorporate into DiSCoVeR in future work.

Additionally, by performing leave-one-cluster-out cross-validation (LOCO-CV) [63], we accurately sort the list of validation clusters by their average performance with a scaled sorting error of approximately 1 %. This proof-of-concept strongly suggests that DiSCoVeR will successfully identify the most promising compounds when supplied with a set of realistic chemical formulae that partly contains out-of-class formulae produced via a suggestion algorithm. To our knowledge, this is a novel approach that has never been used to encourage new materials discovery as opposed to incremental discoveries within known families.

2.2.2. k-Nearest Neighbor Average

An average of the bulk moduli for the k-nearest neighbors (kNNs) is computed as a poor man's gradient as one type of proxy for chemical uniqueness. In this work, we use $k = 10$ to define the local neighborhood of influence, where kNNs are determined via the ElMD. Compounds which exhibit high predicted target bulk moduli relative to their kNNs are considered unique in terms of property gradient, despite having similar chemical composition.

Because it is based on nearest neighbors rather than a defined radius, compounds which are in relatively sparse UMAP areas may have neighbors from a chemically distant cluster. In this case, if all kNNs come from the same cluster, and this cluster exhibits similar properties, this can skew the measure to some extent. This artifact can be avoided by instead using a defined radius and a variable number of kNNs while ignoring compounds which have no kNNs within the specified radius.

2.2.3. Cluster Properties

Cluster validation fraction is given by Eq. (4):

$$f_k = \frac{n_{\mathrm{val},k}}{n_{\mathrm{val},k} + n_{\mathrm{train},k}} \tag{4}$$

where f_k, $n_{\mathrm{val},k}$, and $n_{\mathrm{train},k}$ represent validation fraction of the k-th cluster, number of validation points in the k-th cluster, and number of training points in the k-th cluster, respectively. This indicates to what extent a given cluster consists of unknown compounds and can be useful in identifying clusters which are chemically distinct from existing compounds.

Cluster target mean is given by Eq. (5):

$$E_{\mathrm{avg},k} = \frac{1}{n_k} \sum_{i=1}^{n_k} E_{k,i} \tag{5}$$

where n_k, $E_{\mathrm{avg},k}$, and $E_{k,i}$ represent number of points in the k-th cluster, mean bulk modulus of k-th cluster, and bulk modulus of the i-th point in the k-th cluster, respectively. This is useful for identifying clusters that exhibit overall high performance.

2.3. Data and Validation

As a proof of concept, we use 10 583 unique chemical formulae and associated bulk moduli from Materials Project [64, 65] to test whether DiSCoVeR can find new classes of materials with high performance. In accordance with materials informatics best practices [66], we also sanitize the data. Materials are filtered to exclude noble gases, Tc-containing compounds, and compounds with an energy above hull value greater than 500 meV. The highest bulk modulus is chosen when considering identical formulae. We use CrabNet [44] as the regression model for bulk modulus which depends only on composition to generate machine learning features; however, one of the other models mentioned in Section 1 could have been used instead.

We split the data into training, validation, and test sets using a 0.8/0.2 train/val split as well as via LOCO-CV. We report two types of validation tests as summarized in Table 1. One of the validation methods uses a weighted root-mean-square error (RMSE) of various multi-objective Pareto front properties (target vs. chemical uniqueness proxy). The target is weighted against the proxy property (Eq. (6)):

$$\begin{aligned} \frac{1}{w_E + w_p} \Biggl(& w_E \sqrt{\frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \left(E_{\mathrm{true},i} - E_{\mathrm{pred},i}\right)^2} \\ & + w_p \sqrt{\frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \left(p_{\mathrm{true},i} - p_{\mathrm{pred},i}\right)^2} \Biggr) \end{aligned} \tag{6}$$

where w_E, w_p, n_{val}, $E_{\mathrm{true},i}$, $E_{\mathrm{pred},i}$, $p_{\mathrm{true},i}$, and $p_{\mathrm{pred},i}$ represent bulk modulus weight, proxy weight, number of validation points, DFT-calculated bulk modulus of the i-th validation point, predicted bulk modulus of the i-th validation point, true proxy property of the i-th validation point, and predicted proxy property of the i-th validation point, respectively. We use $w_E = 1$ and $w_p = 1$.

In the current implementation, however, the chemical uniqueness proxy is determined a-priori and simultaneously using the full dataset; thus, the error contribution from the chemical uniqueness proxy is zero. This approach is reasonable for small- to medium-sized datasets (e.g. <20 000), but can quickly become intractable for large datasets due to memory constraints. We plan to modify DiSCoVeR to be compatible with large datasets in near future work by utilizing the ElMD metric directly within DensMAP rather than computing a pairwise distance matrix in advance.

Likewise, the score for each compound is a weighted sum of the robust-scaled[2] target and proxy properties (Eq. (7)):

$$\frac{1}{w_E + w_p}\left(w_E E_i + w_p p_i\right) \tag{7}$$

where w_E, w_p, E_i, and p_i represent bulk modulus weight, proxy weight, robust-scaled predicted bulk modulus of the i-th validation point, and robust-scaled predicted uniqueness proxy of the i-th validation point, respectively. We use $w_E = 1$ and $w_p = 1$.

The other validation method is a LOCO-CV approach using cumulative density function (CDF) distance (i.e. Earth Mover's or Wasserstein distance) as a metric to determine the sorted similarity of a predicted cluster property vs. a true cluster property using `scipy.stats.wasserstein_distance()` [67] as follows[3]:

```
import numpy as np
from scipy.stats import (
    wasserstein_distance,
)

# positions of weights
nclust = len(avg_true)
u = np.cumsum(np.linspace(0, 1, nclust))
u = np.flip(u)
v = u.copy()

# sort by same indices
sorter = np.flip(avg_true.argsort())
u_weights = avg_true[sorter]
v_weights = avg_pred[sorter]

error = wasserstein_distance(
    u,
    v,
    u_weights=u_weights,
    v_weights=v_weights,
)
```

where `avg_true and avg_pred` represent the 1D array of DFT-calculated average bulk moduli for each cluster and the 1D array of predicted average bulk moduli for each cluster, respectively, given by Eq. (5). The use of a cumulative sum causes the positions of high cluster bulk modulus averages to be further spaced apart and therefore is more costly to "move earth" between the two distributions. In other words, inaccuracies associated with high-performing clusters are weighted more heavily than inaccuracies for low-performing clusters. This weighted error is then scaled by dividing by a "dummy" error, where `v_weights` is replaced by the average bulk modulus of the training data for each of the training splits (as opposed to the predictions on the validation data) during computation of the Wasserstein distance.

3. Results and Discussion

We present characteristics of the DensMAP embedding and clustering scheme (Section 3.1), followed by compound-wise (Section 3.2) and cluster-wise (Section 3.3) Pareto front results. Finally, we discuss results of the LOCO-CV scheme.

3.1. *Density-preserving Uniform Manifold Approximation And Projection Characteristics*

We present a DensMAP clustering of ElMD distances between all pairs of compounds (Figure 2a) and plot the cluster count histogram (Figure 2b). We then sum densities at equally spaced locations across DensMAP space (Figure 3a) and color the points according to bulk modulus values (Figure 3b).

We obtain a total of 27 clusters, plus a non-cluster of unclassified points comprising a small percentage of the data (∼5 %). The number of clusters gives an estimation of the number of distinct chemical classes present in the dataset and is also affected by DensMAP and HDBSCAN* model parameters such as local density regularization strength (`dens_lambda`) and minimum cluster size (`min_cluster_size`). The unclassified points are typically isolated points in DensMAP space. In other words, unclassified points will likely exhibit high chemical contrast relative to other composi-

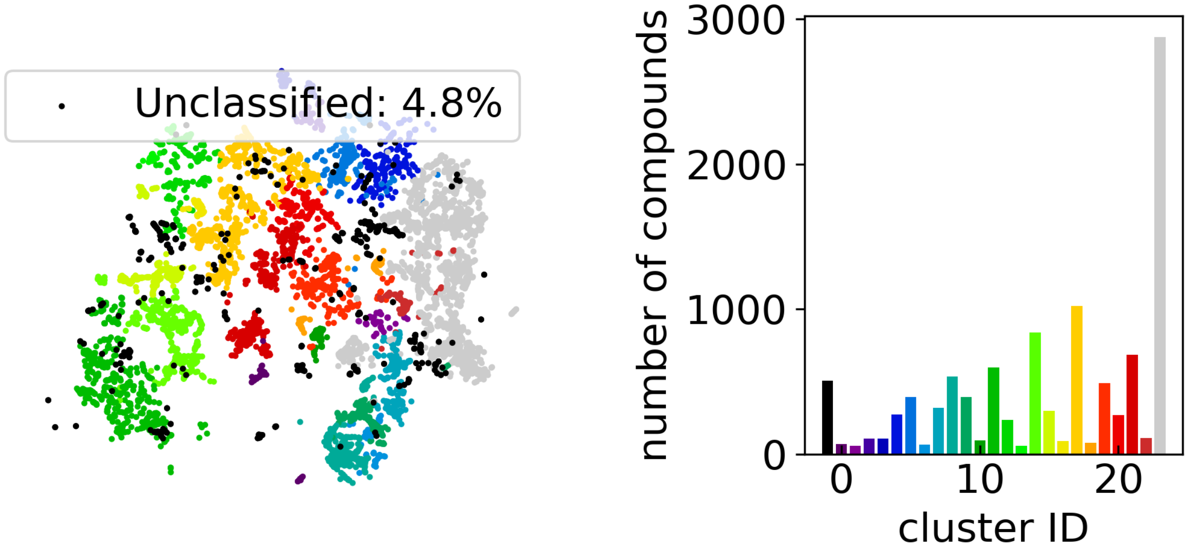
---

[2] See sklearn.preprocessing.RobustScaler.

[3] The code was formatted in Black code style via an online formatter: `https://black.vercel.app/`.

Table 1: Validation methods, splits, notion of best fit, and property used to calculate notion of best fit. *This density is the sum of all training densities evaluated at the validation location in the embedded DensMAP space. For the k-neighbors data, the average of the 10 nearest neighbor properties were used as a proxy. †cluster validation fraction refers to the ratio of number of validation points within a cluster (as opposed to training points) to the total number of points in the cluster. DensMAP densities and cluster fractions are determined simultaneously for both validation and training sets during the DensMAP embedding resulting in computational throughput restrictions. In other words, "predicted" and "true" are identical due to implementation of DiSCoVeR at the time of writing. We plan to address this in future work.

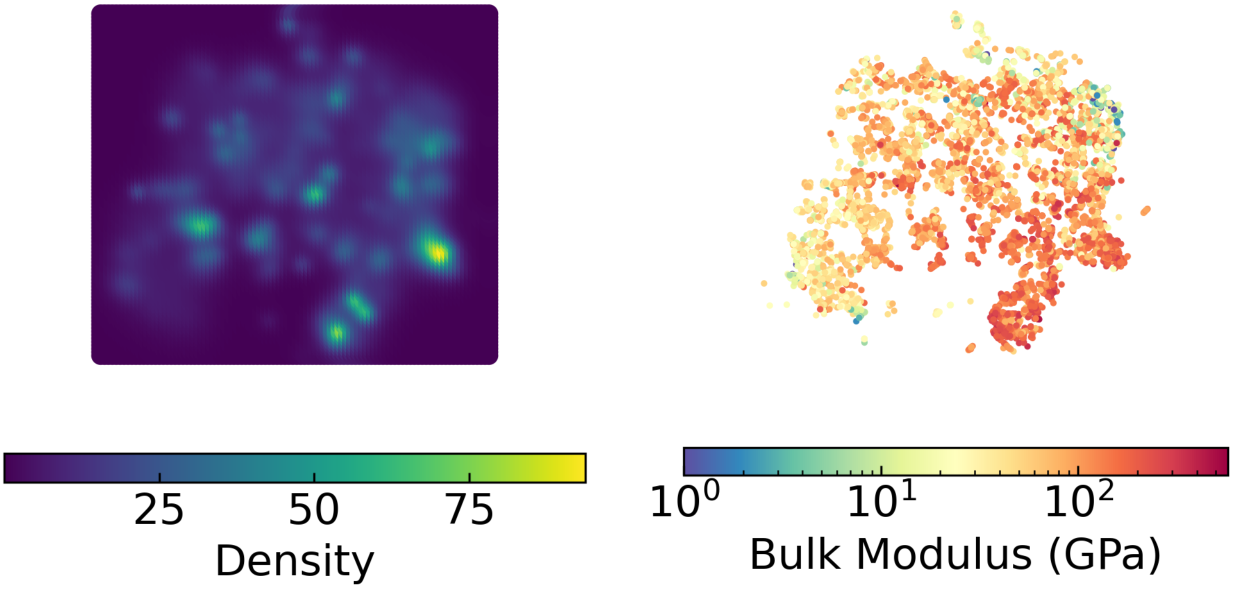
Method	Splits	Notion of best fit	Property
train/val	0.8/0.2	Weighted RMSE	target vs. density*
train/val	0.8/0.2	Weighted RMSE	target vs. k-neighbors average
train/val	0.8/0.2	Weighted RMSE	target vs. cluster validation fraction†
LOCO-CV	27 clusters	Weighted CDF Distance	cluster target mean



(a)

(b)

Figure 2: Summary of cluster properties. (a) DensMAP embeddings based on ElMD distances between compounds colored by cluster. Equal aspect ratio scaling was used. (b) Histogram of number of compounds vs. cluster ID, colored by cluster.



(a)

(b)

Figure 3: Density and bulk modulus. (a) DensMAP densities of both training and validation points summed at gridded locations in DensMAP space. (b) 10 583 bulk moduli of training and validation points embedded in DensMAP space. Equal aspect ratio scaling was used for both (a) and (b).

tions via a low density proxy. We discuss this further in Section 3.2.

A summary of the computational runtimes of the various methods is given in Table 2. Computation of the full pairwise distance matrix takes ∼18 s, which is quite fast due to use of a CUDA/Numba [68] version of the Wasserstein distance that we developed for this work. An NVIDIA GeForce RTX 2060 is used for GPU computations, and an Intel® Core™ i7-10750H CPU @ 2.60GHz is used for CPU computations. All non-GPU calculations are single-threaded.

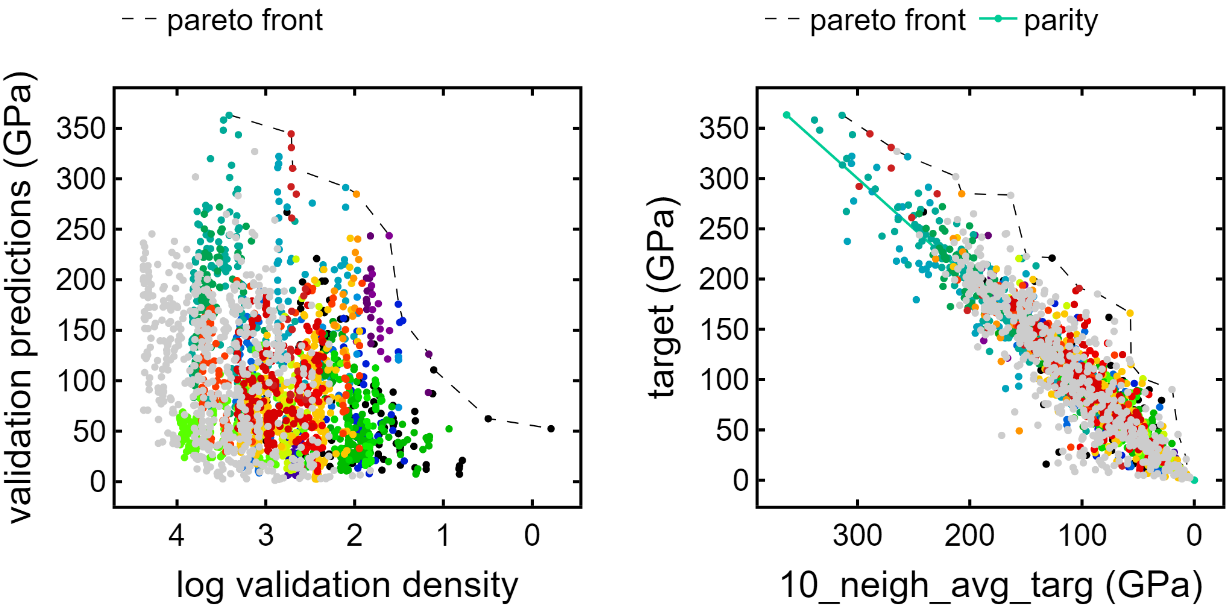
Table 2: Summary of computational runtimes. Procedure, runtime (time), and whether or not a GPU was used (GPU) (Y=Yes, N=No) for various steps in DiSCoVeR. Visualization DensMAP (Vis. DensMAP) and 100 × 100 gridded density summation (100 × 100 grid) are unnecessary steps to produce rankings; however, they are helpful for visualizations presented in this work. Non-GPU calculations are single-threaded. Reported runtimes should be considered approximate, as they are representative of only a single run.

Procedure	Time (s)	GPU
CrabNet	91	Y
ElMD	18	Y
Cluster DensMAP	137	N
Vis. DensMAP	47	N
HDBSCAN*	0.14	N
100 × 100 grid	11	N
Density-proxy	2.7	N
Total	296	—

3.2. Compound Pareto Fronts

We present compound-wise Pareto fronts—a common technique used in multi-objective optimization—with predicted bulk modulus as the ordinate and one of two compound-wise proxies as the abscissa: train contribution to validation log density (Figure 4a) and k-nearest neighbor average (Figure 4a) as described in Section 2.2.

On the other hand, k-nearest neighbor average acts as a poor man's gradient - in other words, used



(a)

(b)

Figure 4: Compound-wise Pareto plots. (a) Pareto plot of validation bulk modulus predictions (GPa) vs. train contribution to validation log density, colored by cluster. The Pareto front is plotted as a dashed line. (b) Pareto plot of training and validation bulk modulus predictions vs. kNN average bulk modulus (GPa) where $k = 10$. The Pareto front is given by a dashed line. A line of parity is given by a solid teal line to emphasize that compounds well above this line are considered unique.

in conjunction with target predictions, it emphasizes compounds which have much higher predicted bulk modulus than that of its neighbors. In addition to the Pareto front, a parity line is also plotted. Compounds which are far above the parity line are high-performing relative to the surrounding neighborhood.

In terms of discovering materials which are chemically distinct from existing materials, train contribution to validation log density is the preferred proxy. We note that each of the proxies produce distinct plots. In the case of Figure 4a, clusters tend to be stratified horizontally, whereas in Figure 4a, cluster shapes exhibit similar orientations. As expected (Section 3.1), unclassified points appear frequently at or near the first Pareto front owing to the fact that unclassified points are likely to have a lower density proxy and therefore higher score. By contrast, unclassified points appear infrequently at or near the latter Pareto front. Additionally, the unique list of clusters present at the Pareto front are different for each plot. In other words, these are two types of chemical uniqueness – the first emphasizing chemical "distance" from other compounds and the latter emphasizing performance superiority over chemically similar compounds. We believe that either may be successfully used in the domain of materials discovery.

Compounds were assigned scaled discovery scores as described in Section 2.3 for each of the chemical uniqueness proxies. The top-10 ranked candidates for the density and peak proxies are given in Tables 3 and 4, respectively. An outer merge of these two lists is given in Table 5.

It is interesting to note the lack of shared compounds between the top-10 lists of the two proxies. By contrast, in previous tests, we found that increasing the weight of E_{pred} (w_E=2) led to significant overlap between the two lists, although with differing priority (i.e. the order of the rankings was different). Because the weights used can have a significant effect on the rankings, it may be worth probing several values for a given study to elucidate and assess behaviors. Indeed, as w_E grows larger, it tends towards a classic approach of searching for high-performance candidates only, yet for very

Table 3: Top-10 ranked high-performing, density-proxy candidates. Formula, predicted bulk modulus (E_{pred}) (GPa), kNN average bulk modulus ($E_{\text{pred,kNN}}$) (GPa), and weighted, scaled discovery score based on train contribution to validation log density proxy (s_ρ).

Formula	E_{pred}	ρ	s_ρ
ReB2	344.735	15.047	2.799
B2W	331.170	15.014	2.666
TaMoN	291.558	8.146	2.573
UB2Os3	285.076	7.213	2.550
MoN	322.441	17.284	2.480
TaN	321.934	17.240	2.477
B2Mo	310.618	14.814	2.471
BMo	315.166	17.322	2.406
Cr2N	272.085	8.150	2.380
Co(BW)2	311.329	17.503	2.360

Table 4: Top-10 ranked high-performing, peak-proxy candidates. Formula, predicted bulk modulus (E_{pred}) (GPa), kNN average bulk modulus ($E_{\text{pred,kNN}}$) (GPa), and weighted, scaled discovery score based on average kNN bulk modulus proxy ($E_{\text{pred,kNN}}$).

Formula	E_{pred}	$E_{\text{pred,kNN}}$	s_{kNN}
WO2	283.530	27.347	5.457
UO3	166.335	10.970	5.443
NiH	185.373	36.430	4.830
V2O3	221.155	11.297	4.415
FeF2	158.245	31.627	4.328
Mg(MoO2)2	162.200	15.936	4.255
ZrSiO	191.061	12.816	4.238
PaB3	188.567	12.512	3.966
NiHO2	148.222	26.699	3.827
CrOF3	90.328	13.641	3.773

Table 5: Outer merge of top-10 ranked high-performing, density-proxy and peak-proxy candidates. Formula, density discovery score (s_ρ), and peak discovery score (s_{kNN}).

Formula	s_ρ	s_{kNN}
ReB2	2.799	1.795
B2W	2.666	2.145
TaMoN	2.573	1.991
UB2Os3	2.550	3.239
MoN	2.480	-0.117
TaN	2.477	2.471
B2Mo	2.471	1.162
BMo	2.406	-0.484
Cr2N	2.380	-0.151
Co(BW)2	2.360	0.519
V2O3	1.737	4.415
WO2	1.655	5.457
ZrSiO	1.372	4.238
PaB3	1.361	3.966
UO3	1.208	5.443
Mg(MoO2)2	0.950	4.255
NiHO2	0.341	3.827
CrOF3	0.337	3.773
NiH	0.284	4.830
FeF2	0.225	4.328

small values of w_E, the performance of the top-ranked compounds may be too low to be of utility in real-world applications.

The weighted RMSE for the validation data is 26.5 GPa; however, as mentioned in Section 2.3, the proxy error contribution is zero in this work.

3.3. *Cluster Pareto Front and Leave-one-cluster-out Cross-validation*

We also present a Pareto front for cluster-wise properties. For the ordinate, we use predicted cluster average bulk modulus Figure 5a. For the abscissa, we use cluster validation fraction as a proxy for chemical distinctiveness of a cluster. In this example, the data is clustered tightly in the abscissa due to a the train/val split being applied randomly without regard to cluster. In a more realistic scenario with much more validation data than training data, where the validation encompasses previously unexplored chemical spaces, there is likely to be a larger spread. Indeed, such a use-case is the intention for this visualization tool. There is a much wider spread in the ordinate, indicating an interesting feature of the clustering results: compositions which are chemically similar to each other also tend to have, on average, similar bulk moduli. This is reasonable, especially since the regression model used is based purely on composition.

In future work, it may be interesting to replace average bulk modulus with best-in-cluster bulk modulus to explore a different type of high-ranking clusters.

Finally, we perform LOCO-CV to evaluate the utility of the DiSCoVeR method in identifying clusters with high average cluster bulk modulus. A LOCO-CV parity plot is given in Figure 5. We accurately sort the list of validation clusters by their average performance with a weighted scaled sorting error (Section 2.3) of $\sim$1.4 %. In other words, the out-of-cluster regression is very accurate. This suggests that CrabNet can successfully extrapolate performance predictions for new chemical spaces in accordance with the goal of DiSCoVeR. In future work, we plan to also test the out-of-cluster extrapolation performance for chemical uniqueness proxies (Section 2.3).

4. Conclusion

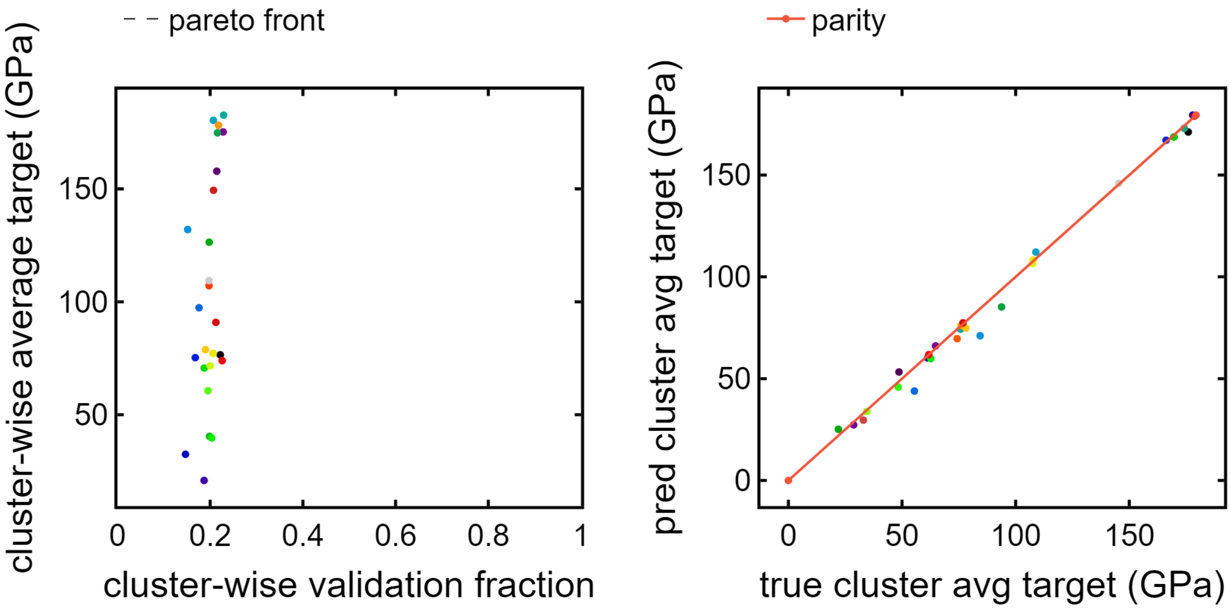
We embedded ElMD distances in DensMAP space and clustered via HDBSCAN* to identify chemically similar clusters for 10 583 compositions. We introduced new proxies (i.e. metrics) for uniqueness-based materials discovery in the form of train contribution to validation log density, k-neighbor averages, and cluster validation fraction. By pairing these with the CrabNet regression model, we visualize Pareto plots of predicted bulk modulus vs. uniqueness proxy and obtain weighted uniqueness/performance rankings for each of the compounds. This reveals a new way to perform materials discovery with a focus towards identifying new high-performing, chemically distinct compositions.

Acknowledgement

The authors thank Dr. Anna Little for useful discussions regarding density- and distance-preserving dimensionality reduction techniques. The authors thank Anthony Wang for assistance with the CrabNet repository. Plots were produced via either Matplotlib [69] or Plotly [70], several tables were formatted via an online formatter [71], and equations were typeset using the Mathematica [72] `TeXport` functionality of `auto-paper` [73]. This work was supported by the National Science Foundation under Grant Nos. DMR-1651668 and DMR-1950589.

CRediT Statement

Sterling G. Baird: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Taylor D. Sparks**: Supervision, Project administration, Funding acquisition, Conceptualization, Formal analysis, Resources, Writing - Review & Editing. **Tran Q. Diep**: Conceptualization, Methodology, Software, Formal analysis



(a)

(b)

Figure 5: LOCO-CV results. (a) Pareto plot of cluster-wise average bulk modulus predictions (GPa) vs. cluster-wise validation fraction. This emphasizes the trade-off between high-performing clusters and chemically unique clusters relative to the original data. (b) Parity plot of predicted cluster-wise average bulk modulus (GPa) vs. DFT-calculated average bulk modulus (GPa).

Data Availability

The raw data required to reproduce these findings are available to download from materialsproject.org. The processed data required to reproduce these findings are available to download from https://dx.doi.org/10.6084/m9.figshare.16786513.v2. The code required to reproduce these findings is hosted at https://github.com/sparks-baird/mat_discover. Interactive Pareto front plots are available at https://sparks-baird.github.io/mat_discover/figures/.

References

[1] P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nature Communications 9 (2018). doi:10.1038/s41467-018-03821-9.

[2] B. Cao, L. A. Adutwum, A. O. Oliynyk, E. J. Luber, B. C. Olsen, A. Mar, J. M. Buriak, How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics, ACS Nano 12 (2018) 7434–7444. doi:10.1021/acsnano.8b04726.

[3] Y. Chen, Y. Tian, Y. Zhou, D. Fang, X. Ding, J. Sun, D. Xue, Machine learning assisted multi-objective optimization for materials processing parameters: A case study in Mg alloy, Journal of Alloys and Compounds 844 (2020) 156159. doi:10.1016/j.jallcom.2020.156159.

[4] K. Homma, Y. Liu, M. Sumita, R. Tamura, N. Fushimi, J. Iwata, K. Tsuda, C. Kaneta, Optimization of a Heterogeneous Ternary Li3PO4-Li3BO3-Li2SO4Mixture for Li-Ion Conductivity by Machine Learning, Journal of Physical Chemistry C 124 (2020) 12865–12870. doi:10.1021/acs.jpcc.9b11654. arXiv:1911.12576.

[5] Z. Hou, Y. Takagiwa, Y. Shinohara, Y. Xu, K. Tsuda, Machine-Learning-Assisted Devel-

opment and Theoretical Consideration for the Al 2 Fe 3 Si 3 Thermoelectric Material, ACS Applied Materials and Interfaces 11 (2019) 11545–11554. doi:10.1021/acsami.9b02381.

[6] X. Li, Z. Hou, S. Gao, Y. Zeng, J. Ao, Z. Zhou, B. Da, W. Liu, Y. Sun, Y. Zhang, Efficient Optimization of the Performance of Mn2+-Doped Kesterite Solar Cell: Machine Learning Aided Synthesis of High Efficient Cu2(Mn,Zn)Sn(S,Se)4 Solar Cells, Solar RRL 2 (2018). doi:10.1002/solr.201800198.

[7] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature 533 (2016) 73–76. doi:10.1038/nature17439.

[8] A. Sakurai, K. Yada, T. Simomura, S. Ju, M. Kashiwagi, H. Okada, T. Nagao, K. Tsuda, J. Shiomi, Ultranarrow-Band Wavelength-Selective Thermal Emission with Aperiodic Multilayered Metamaterials Designed by Bayesian Optimization, ACS Cent. Sci. 5 (2019) 319–326. doi:10.1021/acscentsci.8b00802.

[9] Y. K. Wakabayashi, T. Otsuka, Y. Krockenberger, H. Sawada, Y. Taniyasu, H. Yamamoto, Machine-learning-assisted thin-film growth: Bayesian optimization in molecular beam epitaxy of SrRuO3 thin films, APL Materials 7 (2019). doi:10.1063/1.5123019. arXiv:1908.00739.

[10] S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda, J. Shiomi, Designing Nanostructures for Phonon Transport via Bayesian Optimization, Phys. Rev. X 7 (2017) 021024. doi:10.1103/PhysRevX.7.021024.

[11] A. Talapatra, S. Boluki, T. Duong, X. Qian, E. Dougherty, R. Arróyave, Autonomous efficient experiment design for materials discovery with Bayesian model averaging, Physical Review Materials 2 (2018). doi:10.1103/PhysRevMaterials.2.113803. arXiv:1803.05460.

[12] M. W. Gaultois, T. D. Sparks, C. K. Borg, R. Seshadri, W. D. Bonificio, D. R. Clarke, Data-driven review of thermoelectric materials: Performance and resource onsiderations, Chemistry of Materials 25 (2013) 2911–2920. doi:10.1021/cm400893e.

[13] M. W. Gaultois, A. O. Oliynyk, A. Mar, T. D. Sparks, G. J. Mulholland, B. Meredig, Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties, APL Materials 4 (2016). doi:10.1063/1.4952607.

[14] A. M. Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, J. Brgoch, Machine Learning Directed Search for Ultraincompressible, Superhard Materials, Journal of the American Chemical Society 140 (2018) 9844–9853. doi:10.1021/jacs.8b02717.

[15] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Machine learning assisted design of high entropy alloys with desired property, Acta Materialia 170 (2019) 109–117. doi:10.1016/j.actamat.2019.03.010.

[16] D. Xue, D. Xue, R. Yuan, Y. Zhou, P. V. Balachandran, X. Ding, J. Sun, T. Lookman, An informatics approach to transformation temperatures of NiTi-based shape memory alloys, Acta Materialia 125 (2017) 532–541. doi:10.1016/j.actamat.2016.12.009.

[17] Z. Zhang, A. Mansouri Tehrani, A. O. Oliynyk, B. Day, J. Brgoch, Finding the Next Superhard Material through Ensemble Learning, Adv. Mater. (2020) 2005112. doi:10.1002/adma.202005112.

[18] Y. Iwasaki, R. Sawada, V. Stanev, M. Ishida, A. Kirihara, Y. Omori, H. Someya, I. Takeuchi, E. Saitoh, S. Yorozu, Identification of advanced spin-driven thermoelectric materials via interpretable machine learning, npj Computational Materials 5 (2019) 6–11. doi:10.1038/s41524-019-0241-9.

[19] F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments, Science Advances 4 (2018). doi:10.1126/sciadv.aaq1566.

[20] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, Adaptive Strategies for Materials Design using Uncertainties, Sci Rep 6 (2016) 19660. doi:10.1038/srep19660.

[21] P. V. Balachandran, Data-driven design of B20 alloys with targeted magnetic properties guided by machine learning and density functional theory, Journal of Materials Research 35 (2020) 890–897. doi:10.1557/jmr.2020.38.

[22] P. V. Balachandran, J. Young, T. Lookman, J. M. Rondinelli, Learning from data to design functional materials without inversion symmetry, Nature Communications 8 (2017). doi:10.1038/ncomms14282.

[23] P. V. Balachandran, T. Shearman, J. Theiler, T. Lookman, Predicting displacements of octahedral cations in ferroelectric perovskites using machine learning, Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials 73 (2017) 962–967. doi:10.1107/S2052520617011945.

[24] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, Nat Commun 9 (2018) 3405. doi:10.1038/s41467-018-05761-w.

[25] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad, Machine Learning Strategy for Accelerated Design of Polymer Dielectrics, Sci Rep 6 (2016) 20952. doi:10.1038/srep20952.

[26] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B 89 (2014) 094104. doi:10.1103/PhysRevB.89.094104.

[27] C. W. Park, C. Wolverton, Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery, Phys. Rev. Materials 4 (2020) 063801. doi:10.1103/PhysRevMaterials.4.063801.

[28] A. Seko, H. Hayashi, H. Kashima, I. Tanaka, Matrix- and tensor-based recommender systems for the discovery of currently unknown inorganic compounds, Phys. Rev. Materials 2 (2018) 013805. doi:10.1103/PhysRevMaterials.2.013805. arXiv:1710.00659.

[29] A. D. Sendek, Q. Yang, E. D. Cubuk, K. A. N. Duerloo, Y. Cui, E. J. Reed, Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials, Energy and Environmental Science 10 (2017) 306–320. doi:10.1039/c6ee02697d.

[30] B. B. Hoar, S. Lu, C. Liu, Machine-Learning-Enabled Exploration of Morphology Influence on Wire-Array Electrodes for Electrochemical Nitrogen Fixation, Journal of Physical Chemistry Letters 11 (2020) 4625–4630. doi:10.1021/acs.jpclett.0c01128.

[31] B. Yan, R. Gao, P. Liu, P. Zhang, L. Cheng, Optimization of thermal conductivity of UO2–Mo composite with continuous Mo channel based on finite element method and machine learning, International Journal of Heat and Mass Transfer 159 (2020) 120067. doi:10.1016/j.ijheatmasstransfer.2020.120067.

[32] A. O. Oliynyk, L. A. Adutwum, B. W. Rudyk, H. Pisavadia, S. Lotfi, V. Hlukhyy, J. J. Harynuk, A. Mar, J. Brgoch, Disentangling Structural Confusion through Machine Learning: Structure Prediction and Polymorphism of Equiatomic Ternary Phases ABC, Journal of the American Chemical Society 139 (2017) 17870–17881. doi:10.1021/jacs.7b08460.

[33] J. M. Rickman, H. M. Chan, M. P. Harmer, J. A. Smeltzer, C. J. Marvel, A. Roy, G. Balasubramanian, Materials informatics for the screening of multi-principal elements and high-entropy alloys, Nature Communications 10 (2019) 1–10. doi:10.1038/s41467-019-10533-1.

[34] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, T. Lookman, Accelerated search for BaTiO3-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning, Proceedings of the National Academy of Sciences of the United States of America 113 (2016) 13301–13306. doi:10.1073/pnas.1607412113.

[35] S. K. Kauwe, J. Graser, R. Murdock, T. D. Sparks, Can machine learning find extraordinary materials?, Computational Materials Science 174 (2020). doi:10.1016/j.commatsci.2019.109498.

[36] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, A. Gamst, A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds, Sci Rep 6 (2016) 34256. doi:10.1038/srep34256.

[37] T. Xie, J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, Phys. Rev. Lett. 120 (2018) 145301. doi:10.1103/PhysRevLett.120.145301.

[38] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, Chem. Mater. 31 (2019) 3564–3572. doi:10.1021/acs.chemmater.9b01294.

[39] R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, Wyckoff Set Regression for Materials Discovery, in: Neural Information Processing Systems, 2020, p. 7.

[40] S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu, J. Hu, Graph convolutional neural networks with global attention for improved materials property prediction, Phys. Chem. Chem. Phys. 22 (2020) 18141–18148. doi:10.1039/D0CP01474E.

[41] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm, npj Comput Mater 6 (2020) 138. doi:10.1038/s41524-020-00406-3.

[42] R. E. A. Goodall, A. A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry, Nat Commun 11 (2020) 6280. doi:10.1038/s41467-020-19964-7.

[43] J. Klicpera, S. Giri, J. T. Margraf, S. Günnemann, Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules, arXiv:2011.14115 [physics] (2020). arXiv:2011.14115.

[44] A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, D. Sparks, Compositionally-Restricted Attention-Based Network for Materials Property Predictions, npj Computational Materials (2021) 33. doi:10.1038/s41524-021-00545-1.

[45] P.-P. De Breuck, G. Hautier, G.-M. Rignanese, Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet, npj Comput Mater 7 (2021) 83. doi:10.1038/s41524-021-00552-2.

[46] M. Asahara, R. Fujimaki, An Empirical Study on Distributed Bayesian Approximation Inference of Piecewise Sparse Linear Models, IEEE Trans. Parallel Distrib. Syst. 30 (2019) 1481–1493. doi:10.1109/TPDS.2019.2892972.

[47] T. Baldacchino, E. J. Cross, K. Worden, J. Rowson, Variational Bayesian mixture of experts models and sensitivity analysis for

nonlinear dynamical systems, Mechanical Systems and Signal Processing 66–67 (2016) 178–200. doi:10.1016/j.ymssp.2015.05.009.

[48] R. Eto, R. Fujimaki, S. Morinaga, H. Tamano, Fully-Automatic Bayesian Piecewise Sparse Linear Models, in: International Conference on Artificial Intelligence and Statistics, 2014, p. 9.

[49] W. Hashimoto, Y. Tsuji, K. Yoshizawa, Optimization of Work Function via Bayesian Machine Learning Combined with First-Principles Calculation, Journal of Physical Chemistry C 124 (2020) 9958–9970. doi:10.1021/acs.jpcc.0c01106.

[50] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, COMBO: An efficient Bayesian optimization library for materials science, Materials Discovery 4 (2016) 18–21. doi:10.1016/j.md.2016.04.001.

[51] H. Wahab, V. Jain, A. S. Tyrrell, M. A. Seas, L. Kotthoff, P. A. Johnson, Machine-learning-assisted fabrication: Bayesian optimization of laser-induced graphene patterning using in-situ Raman analysis, Carbon 167 (2020) 609–619. doi:10.1016/j.carbon.2020.05.087.

[52] Y. Kim, E. Kim, E. Antono, B. Meredig, J. Ling, Machine-learned metrics for predicting the likelihood of success in materials discovery, npj Comput Mater 6 (2020) 131. doi:10.1038/s41524-020-00401-8.

[53] C. J. Hargreaves, M. S. Dyer, M. W. Gaultois, V. A. Kurlin, M. J. Rosseinsky, The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions, Chem. Mater. 32 (2020) 10610–10620. doi:10.1021/acs.chemmater.0c03381.

[54] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426 [cs, stat] (2020). arXiv:1802.03426.

[55] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[56] L. McInnes, J. Healy, S. Astels, Hdbscan: Hierarchical density based clustering, JOSS 2 (2017) 205. doi:10.21105/joss.00205.

[57] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inform. Theory 28 (1982) 129–137. doi:10.1109/TIT.1982.1056489.

[58] A. Narayan, B. Berger, H. Cho, Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability, Preprint, Bioinformatics, 2020. doi:10.1101/2020.05.12.077776.

[59] E. Parzen, On Estimation of a Probability Density Function and Mode, The Annals of Mathematical Statistics 33 (1962) 1065–1076. doi:10.1214/aoms/1177704472.

[60] M. Rosenblatt, Remarks on Some Nonparametric Estimates of a Density Function, The Annals of Mathematical Statistics 27 (1956) 832–837. doi:10.1214/aoms/1177728190.

[61] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, AAAI Press, Portland, Oregon, 1996, pp. 226–231.

[62] Y. Zuo, M. Qin, C. Chen, W. Ye, X. Li, J. Luo, S. P. Ong, Accelerating materials discovery with Bayesian optimization and graph deep learning, Materials Today (2021) S1369702121002984. doi:10.1016/j.mattod.2021.08.012.

[63] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, L. Ward, Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery,

Mol. Syst. Des. Eng. 3 (2018) 819–825. doi:10.1039/C8ME00012C.

[64] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. a. Persson, The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials 1 (2013) 011002. doi:10.1063/1.4812323.

[65] M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, M. Asta, Charting the complete elastic properties of inorganic crystalline compounds, Sci Data 2 (2015) 150009. doi:10.1038/sdata.2015.9.

[66] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T. D. Sparks, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, Chem. Mater. (2020) 12. doi:10.1021/acs.chemmater.0c01907.

[67] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, R. Munos, The Cramer Distance as a Solution to Biased Wasserstein Gradients, arXiv:1705.10743 [cs, stat] (2017). arXiv:1705.10743.

[68] S. K. Lam, A. Pitrou, S. Seibert, Numba: A LLVM-based Python JIT compiler, in: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, LLVM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1–6. doi:10.1145/2833157.2833162.

[69] J. D. Hunter, Matplotlib: A 2D Graphics Environment, Computing in Science Engineering 9 (2007) 90–95. doi:10.1109/MCSE.2007.55.

[70] P. T. Inc., Collaborative data science, https://plot.ly, 2015.

[71] Create LaTeX tables online – TablesGenerator.com, https://www.tablesgenerator.com/, 2021.

[72] W. R. Inc., Mathematica, Version 12.2, 2020.

[73] Auto-paper, sparks-baird, 2021.