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

Sterling G. Baird\textsuperscript{a}, Tran Diep\textsuperscript{b}, Taylor D. Sparks\textsuperscript{a,}* \\
\textsuperscript{a}Department of Materials Science and Engineering, University of Utah, Salt Lake City, UT 84108, USA \\
\textsuperscript{b}Department of Chemical Engineering, Brigham Young University, Provo, UT 84604, USA \\

Abstract

We present Descending from Stochastic Clustering Variance Regression (DiSCoVeR), a Python tool for identifying 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-
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 successful synthesis 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. XY₂ as in SiO₂, AlO₂) 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] 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)):

\[ e^{-\frac{1}{2}(X-\mu)\Sigma^{-1}(X-\mu)} \]

where \( X \), \( \mu \), \( \Sigma \), 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):

\[
\begin{pmatrix}
  e^r & 0 \\
  0 & e^r
\end{pmatrix}
\]

where \( r \) represents extracted DensMAP radius.

By evaluating the sum of densities contributed by all of the training points evaluated at each of

---

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

the validation locations (Eq. (3)):

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

where $X_{v,j}$, $\mu_{t,i}$, $\Sigma_{t,i}$, ... , and $n_{\text{train}}$ 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. 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.

Additionally, by performing leave-one-cluster-out cross-validation, 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 neighbor (kNN) 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 structure.

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 kNN within the specified radius.

2.2.3. Cluster Properties

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

$$f_k = \frac{n_{\text{val},k}}{n_{\text{val},k} + n_{\text{train},k}}$$

where $f_k$, $n_{\text{val},k}$, and $n_{\text{train},k}$ 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):

$$E_{\text{avg},k} = \frac{1}{n_k} \sum_{i=1}^{n_k} E_{k,i}$$

where $n_k$, $E_{\text{avg},k}$, 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 10,710 unique chemical formulae and associated bulk moduli from Materials Project [62, 63] to test whether DiSCoVeR can find new classes of materials with high performance. The highest bulk modulus was chosen when considering identical formulae.

We split the data into training, validation, and test sets in accordance with materials informatics best practices [64] using a 0.7/0.2/0.1 train/val/test split as well as via leave-one-cluster-out cross-validation (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 multiobjective Pareto front properties (target vs. chemical uniqueness proxy). The target is weighted at twice that of the proxy property to favor consideration of high performing candidates (Eq. (6)):

$$\frac{1}{w_E + w_p} \left( \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} (E_{\text{true},i} - E_{\text{pred},i})^2 + w_p \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} (p_{\text{true},i} - p_{\text{pred},i})^2 \right)$$

where $w_E$, $w_p$, $n_{\text{val}}$, $E_{\text{true},i}$, $E_{\text{pred},i}$, $p_{\text{true},i}$, and $p_{\text{pred},i}$ 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.

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 scaled target and proxy properties (Eq. (7)):

$$\frac{1}{w_E + w_p} (w_E E_i + w_p p_i)$$

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

\footnote{During initial prototyping, regrettably, the data was split into training and validation sets using a 0.8/0.2 train/val split before moving to a train/val/test scheme.}
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()` [65] as follows:

```python
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 code was formatted via an online formatter (https://black.vercel.app/). The use of a cumulative sum causes the positions of high cluster bulk modulus averages to be further spaced apart and therefore 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).

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.

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), 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 24 clusters, plus a non-cluster of unclassified points comprising a small percentage of the data. 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 compositions via a low density proxy. We discuss this further in Section 3.2.

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
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Splits</th>
<th>Notion of best fit</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>train/val/test</td>
<td>0.7/0.2/0.1</td>
<td>Weighted RMSE</td>
<td>target vs. density*</td>
</tr>
<tr>
<td>train/val/test</td>
<td>0.7/0.2/0.1</td>
<td>Weighted RMSE</td>
<td>target vs. k-neighbors average</td>
</tr>
<tr>
<td>train/val/test</td>
<td>0.7/0.2/0.1</td>
<td>Weighted RMSE</td>
<td>target vs. cluster validation fraction†</td>
</tr>
<tr>
<td>LOCO-CV</td>
<td>24 clusters</td>
<td>Weighted CDF Distance</td>
<td>cluster target mean</td>
</tr>
</tbody>
</table>

Figure 2: Summary of cluster properties. (a) DensMAP embeddings based on EIMD distances between compounds colored by cluster. (b) Histogram of number of compounds vs. cluster ID, colored by cluster.
Figure 3: Density and bulk modulus. (a) DensMAP densities of both training and validation points summed at gridded locations in DensMAP space. (b) 10,710 bulk moduli of training and validation points embedded in DensMAP space.

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 empha-
sizes compounds which have much higher predicted
bulk modulus than that of its neighbors. In addi-
tion 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 neigh-
borhood.

In terms of discovering materials which are chem-
ically distinct from existing materials, train con-
tribution 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 Fig-
ure 4a, cluster shapes exhibit similar orientations.
 As expected (Section 3.1), unclassified points ap-
pear 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 in-
frequently at or near the latter Pareto front. Ad-
ditionally, 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 per-
formance superiority over chemically similar com-
pounds. 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 2 and 3, respectively. An outer
merge of these two lists is given in Table 4.

It is interesting that while the ranking order and
scores are different for the two lists, 10 of the com-
pounds are shared with only 3 that are uncommon
between the two lists. In other words, both prox-
ies identify similar compounds but with differing
priority. This suggests that low-density regions of
the space also exhibit low average bulk modulus
of the surrounding neighborhood (which necessar-
ily is a larger region). It is likely that by replacing
the peak proxy (i.e. kNN average proxy) with a
radius-based neighborhood average, a very differ-
ent set of rankings will result. We expect that by
a proper choice of radius, many of the previously

<table>
<thead>
<tr>
<th>Formula</th>
<th>$E_{\text{pred}}$</th>
<th>$\rho$</th>
<th>$s_\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W2C</td>
<td>328.346</td>
<td>6.767</td>
<td>1.000</td>
</tr>
<tr>
<td>MoC</td>
<td>345.601</td>
<td>16.175</td>
<td>0.980</td>
</tr>
<tr>
<td>WN</td>
<td>345.331</td>
<td>16.086</td>
<td>0.980</td>
</tr>
<tr>
<td>BW</td>
<td>329.552</td>
<td>16.080</td>
<td>0.944</td>
</tr>
<tr>
<td>Ta4C3</td>
<td>308.621</td>
<td>9.050</td>
<td>0.941</td>
</tr>
<tr>
<td>TaC</td>
<td>313.343</td>
<td>16.243</td>
<td>0.907</td>
</tr>
<tr>
<td>TaMoN</td>
<td>285.417</td>
<td>6.603</td>
<td>0.904</td>
</tr>
<tr>
<td>TaN</td>
<td>311.559</td>
<td>16.217</td>
<td>0.903</td>
</tr>
<tr>
<td>Re3Ru</td>
<td>351.544</td>
<td>31.966</td>
<td>0.894</td>
</tr>
<tr>
<td>TeOs3</td>
<td>347.322</td>
<td>30.700</td>
<td>0.892</td>
</tr>
</tbody>
</table>

Table 2: Top-10 ranked high performing, density-proxy candidates. Formula, predicted bulk modulus ($E_{\text{pred}}$) (GPa), kNN average bulk modulus ($E_{\text{pred,kNN}}$) (GPa), and weighted, scaled discovery score based on train contribution to validation density proxy ($s_\rho$).

<table>
<thead>
<tr>
<th>Formula</th>
<th>$E_{\text{pred}}$</th>
<th>$E_{\text{pred,kNN}}$</th>
<th>$s_{kNN}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OsO2</td>
<td>329.073</td>
<td>28.591</td>
<td>1.000</td>
</tr>
<tr>
<td>NbNO</td>
<td>233.005</td>
<td>14.297</td>
<td>0.982</td>
</tr>
<tr>
<td>Ta4C3</td>
<td>308.621</td>
<td>9.050</td>
<td>0.963</td>
</tr>
<tr>
<td>MnIr3</td>
<td>296.835</td>
<td>50.203</td>
<td>0.933</td>
</tr>
<tr>
<td>W2C</td>
<td>328.346</td>
<td>6.767</td>
<td>0.932</td>
</tr>
<tr>
<td>WN</td>
<td>345.331</td>
<td>16.086</td>
<td>0.927</td>
</tr>
<tr>
<td>MoC</td>
<td>345.601</td>
<td>16.175</td>
<td>0.926</td>
</tr>
<tr>
<td>TaN</td>
<td>311.559</td>
<td>16.217</td>
<td>0.900</td>
</tr>
<tr>
<td>TeOs3</td>
<td>347.322</td>
<td>30.700</td>
<td>0.897</td>
</tr>
<tr>
<td>TaC</td>
<td>313.343</td>
<td>16.243</td>
<td>0.892</td>
</tr>
</tbody>
</table>

Table 3: Top-10 ranked high performing, peak-proxy candidates. Formula, predicted bulk modulus ($E_{\text{pred}}$) (GPa), kNN average bulk modulus ($E_{\text{pred,kNN}}$) (GPa), and weighted, scaled discovery score based on average kNN bulk modulus proxy ($E_{\text{pred,kNN}}$).
Table 4: Outer merge of top-10 ranked high performing, density-proxy and peak-proxy candidates. Formula, density discovery score ($s_\rho$), and peak discovery score ($s_{kNN}$).

<table>
<thead>
<tr>
<th>Formula</th>
<th>$s_\rho$</th>
<th>$s_{kNN}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W2C</td>
<td>1.000</td>
<td>0.932</td>
</tr>
<tr>
<td>MoC</td>
<td>0.980</td>
<td>0.926</td>
</tr>
<tr>
<td>WN</td>
<td>0.980</td>
<td>0.927</td>
</tr>
<tr>
<td>BW</td>
<td>0.944</td>
<td>0.827</td>
</tr>
<tr>
<td>Ta4C3</td>
<td>0.941</td>
<td>0.963</td>
</tr>
<tr>
<td>TaC</td>
<td>0.907</td>
<td>0.892</td>
</tr>
<tr>
<td>TaMoN</td>
<td>0.904</td>
<td>0.804</td>
</tr>
<tr>
<td>TaN</td>
<td>0.903</td>
<td>0.900</td>
</tr>
<tr>
<td>Re3Ru</td>
<td>0.894</td>
<td>0.881</td>
</tr>
<tr>
<td>TcOs3</td>
<td>0.892</td>
<td>0.897</td>
</tr>
<tr>
<td>OsO2</td>
<td>0.864</td>
<td>1.000</td>
</tr>
<tr>
<td>NbNO</td>
<td>0.738</td>
<td>0.982</td>
</tr>
<tr>
<td>MnIr3</td>
<td>0.656</td>
<td>0.933</td>
</tr>
</tbody>
</table>

overlapping candidates will drop out in the averaging proxy due to lack of neighbors for low-density candidates. Additionally, it might be of interest to use gridded points in DensMAP space to define the neighborhood rather than the original data.

3.3. Cluster Pareto Front

We also present Pareto fronts for cluster-wise properties. For the ordinate, we use predicted cluster average bulk modulus Figure 5. 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 unsurprising, 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.

3.4. Leave-one-cluster-out Cross-validation

Finally, we perform LOCO-CV to evaluate the utility of the DiSCoVeR method in identifying clusters with high average cluster bulk modulus. We accurately sort the list of validation clusters by their average performance with a weighted scaled sorting error (Section 2.3) of approximately 1%. In other words, the out-of-cluster regression with a larger weight placed on higher performance 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 10,710 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. This work was supported by the National Science Foundation under Grant Nos. DMR-1651668 and DMR-1950589.
Figure 5: 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.

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, Writing - Original Draft

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. The code required to reproduce these findings is hosted at https://github.com/sparks-baird/discover.

References


