Guided Clustering for Selecting Representatives Samples in Chemical Databases

Felipe V. Calderan[1], João Paulo A. de Mendonça[2], Juarez L. F. Da Silva[2], and Marcos G. Quiles[1(✉)]

[1] Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil
{fvcalderan,quiles}@unifesp.br

[2] São Carlos Institute of Chemistry, University of São Paulo, PO Box 780, 13560-970 São Carlos, SP, Brazil
jpalastus@gmail.com, juarez_dasilva@iqsc.usp.br

Abstract. Machine Learning (ML) methods, from unsupervised to supervised algorithms, have been applied to solve several tasks in the Materials Science domain, such as property prediction, design of new chemical compounds, and surrogate models in molecular dynamics simulations. ML methods can also play a fundamental role in screening materials by reducing the number of compounds under scrutiny. This reduction assumes that compounds similarly represented by a given descriptor might have similar properties; thus, an unsupervised ML method, such as the K-Means algorithm, can cluster the data set and deliver a set of representative samples. However, this selection depends on the molecular representation that might not directly relate to the target property. Here, we propose a framework that lets the specialist select a set of representative samples in a guided fashion. In particular, a loop between a clustering algorithm (k-means) and an optimization method (Basin-Hopping) is implemented, which allows the system to learn feature weights to form more homogeneous clusters given the target property. The framework also offers other visual and textual functionalities to support the expert. We evaluate the proposed framework in two scenarios, and the results show that the guidance enhances clustering formations, both in coarse (few and big clusters) and fine (many small clusters) analyses.

Keywords: Guided Data Clustering · Data Visualization · Representative Chemical Sample Selection

1 Introduction

Materials Science research has a direct impact on human civilization development, from the building of tools of iron to the creation of the semiconductors we use on every electronic device today. In order to achieve these advancements, Material Science requires its own set of technological leaps, and one of the most

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14111, pp. 133–149, 2023.
https://doi.org/10.1007/978-3-031-37126-4_10

promising techniques today is the use of Machine Learning. This branch of Artificial Intelligence has provided sophisticated algorithms for regression analyses [16], classification [41], clustering [3], dimensionality reduction [7] and other useful tools to support the study of existing materials and the discovery of new ones [9,22,26].

A particularly interesting application of machine learning to Materials Science arise from the challenge of grouping similar items together (i.e., data clustering [17]) and electing one sample of each group to represent everyone else in that group. This problem is known as representative selection, and many examples of this type of algorithm to solve it exist in the literature, as they are flagship methods of filtering massive amounts of data in order to acquire insights. Bayada, Hamersma, and van Geerestein [5] studied different methods such as Ward agglomerative clustering [39], Kohonen maps [20] and partitioning methods to select representative samples of compounds for more specialized testing. Likewise, Engels et al. [13] utilized Ward to obtain homogeneous subsets (and closely related candidates) of Janssen's corporate database. McGregor and Pallai [24] used the Jarvis-Patrick algorithm to get representative samples from the centroids computed to make compound acquisition more efficient. Brockherde et al. [8] extracted 2000 representative conformers selected from classical MD trajectories by K-Means sampling. Batista et al. [3,4] used the same clustering algorithm to choose representative samples for the Ab-initio investigation of CO_2 adsorption on Transition Metals 13-atom 4d clusters and to study the energy decomposition to access the stability changes induced by CO adsorption on Transition Metals 13-atom clusters. Finally, Zibordi-Besse et al. [42] applied the Euclidean Similarity Distance (EDS) [11] to get a representative set of structures for the $(MO_2)_n$ cluster for different metals M and values n in order to calculate the properties of those chemical clusters.

Clustering algorithms are typically unsupervised Machine Learning methods, which can lead to a significant limitation when trying to examine specific properties of a chemical system. The feature vector describes how they see the universe in which the items reside, and since clustering formations are directly tied to the chosen representation, they might not reveal the best structure according to the property under research, causing a suboptimal representative selection. We propose that guidance be introduced into the clustering process, in order to gradually change the weights to optimize the clusters formed concerning a given property. This guidance consists of a loop with two main blocks: a clustering algorithm and an optimization one. When the clustering algorithm produces a configuration, it is directed to the optimization algorithm to compute the sum of the desired property intracluster variance of all clusters. After that, new weights for the features are selected in an attempt to feed them to the clustering algorithm to generate clusters that reduce the variance. This approach has the potential to create more cohesive groups that will emphasize the particular property under study.

To make guided representative selection accessible to the community, we developed a user-friendly tool called GuidedClustering where the user fill in the

input data and essential parameters, and the software does the rest. In summary, the user provides a data set with the features and properties of interest, then a property is specified as a target and not used as a feature during clustering. The tool can automatically cluster the data into K groups, where K is a suggested value based on clustering quality measures (like the Silhouette score [33] or the Calinski-Harabasz index [10]). Later on, the clustering-optimization loop starts, and when it is finished, the software reports the results as textual and graphical information. The space-representing scatter plots are projected in two dimensions using PCA [1] and t-SNE [23].

The approach was tested with the use of two sets of molecular structures (nanoclusters). These data sets differ not only in their chemical systems, but also in how they were built. One of them was generated through Molecular Dynamics and the other through Density Functional Theory. Our results demonstrate that the strategy is versatile to different types of chemical data and is promising in enhancing the qualitative analyses of the clustering process in both cases. This allowed interesting thermodynamic insights on the structure-property relation in our test cases.

2 Theoretical Approach for Data Clustering in Quantum-Chemistry Problems

2.1 Clustering Algorithms

Clustering algorithms seek to group entries from a data set based on a similarity measure, bringing forth a final result where items belonging to one cluster (group of items) are more similar between themselves than to items from external clusters [19]. There are many of these algorithms [18] and they can be very different, but at the same time, share many common characteristics.

A simple mathematical representation for a generic clustering method is:

$$K(\mathbb{X}) = \overrightarrow{\mathbb{L}} \tag{1}$$

where K is the clustering method, $\mathbb{X} \in (\mathbb{R}^n \times \mathbb{R}^m)$ is the input data set with n entries and m features, and $\overrightarrow{\mathbb{L}} \in \mathbb{Z}_+^n$ is the label vector, which contains the cluster assigned for each one of the n items.

There must be a way of calculating the similarity between entries in the feature space. This can be done by measuring the distance between the values of each of the items features. A method that has predominated for millennia [11] to execute this measurement is the *Euclidean distance*, but other methods, like the *Manhattan Distance* [12] can also be used. It all depends on the data set being clustered and the expected results.

K-Means typically uses the *Euclidean Distance* and is the clustering algorithm of choice for GuidedClustering, specifically scikit-learn's implementation due to good results obtained by the QTNano group using it in past researches

Algorithm 1. K-Means Clustering

```
1: Centroids ← K random items ∈ 𝕏
2: while STOP CRITERION not satisfied do
3:     for each x ∈ 𝕏 do
4:         ASSIGN(x, closest c ∈ Centroids)
5:     end for
6:     for each c ∈ Centroids do
7:         c ← AVERAGE(x's assigned to c)
8:     end for
9: end while
```

[4]. It is a partitional method and is very simple and commonly used in the literature, due to its simplicity and swiftness. Algorithm 1 describes how K-Means operates.

One of the most notable K-Means limitation is that it needs to know beforehand the amount K of clusters to be generated. The proposed toolbox suggests a sensible amount by applying a simple brute-force optimization. It calculates the *Silhouette Score* [33] for a range of possible number of clusters informed by the user, then returns the best one.

2.2 Optimization Algorithms

Mathematical optimization tries to find the best available values for a function, given certain constraints and objectives. Problems with explicit objectives can be expressed as:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} / \max f(x),\ x = \begin{bmatrix} x_1\ x_2 \cdots x_n \end{bmatrix}^T \in \mathbb{R}^n, \\ \text{subject to } \phi_j(x) = 0,\ j \in [1..M], \\ \psi_k(x) \geq 0,\ k \in [1..N][40] \end{aligned} \tag{2}$$

where $f(x)$ is the cost function, x is the decision vector (and $x_1, \cdots, x_n$ is a state), $\phi_j(x)$ are constraints in terms of M equalities and $\psi_k(x)$ are constraints in terms of N inequalities.

After the amount K has been defined, GuidedClustering uses a global-search optimization algorithm to minimize the total intracluster variance. Global optimization tries to escape local optima by adding an item of randomness to the search and/or an acceptance criterion. These keep the algorithm from being so easily convinced that a current best-found solution is the global optimum [21].

We chose Basin-Hopping [38] as the global optimizer. It not only does local search, but also perturbs the current search position in a randomized, but controlled manner. Algorithm 2 shows Basin-Hopping in its simplest form, which always store the best found value up to the current point and reject anything that is not as good (therefore, being a monotonic optimizer). Generally, operation 7 is replaced by the Metropolis criterion of the Monte Carlo algorithms.

Algorithm 2. Monotonic Basin-Hopping [29]

```
i ← 0
X_i ← random initial point in variable space
Y_i ← LOCALSEARCH(X_i)
while STOP CRITERION not satisfied do
    X_{i+1} ← PERTURB(Y_i)
    Y_{i+1} ← LOCALSEARCH(X_{i+1})
    if f(Y_{i+1}) < f(Y_i) then
        i ← i + 1
    end if
end while
```

2.3 Guided Clustering

What we call Guided Clustering are clustering algorithms that use additional background knowledge information about the data to produce better clusters. In the literature, it is common to see explicit definitions of *Must-link* vs *Cannot-link* constraints [37] or *Positive label* and *Negative label* constraints (of course, there are algorithms that support both [2]). GuidedClustering does not use explicit constraints, rather, it implicitly generates the constraints as a side effect of the optimization, which tries to minimize the sum of intracluster variances.

Let $\mathbb{X} \in (\mathbb{R}^n \times \mathbb{R}^m)$ be the input data set and $\overrightarrow{\mathbb{B}} \in \mathbb{R}^n$ the bias property vector, with n values (one for each entry in the input matrix). Also, let $\overrightarrow{\mathbb{W}} \in \mathbb{R}^m$ be the weights vector. The first step to generate a guided clustering is to apply the weights to $\mathbb{X}$. This is done by multiplying the values of each column j of the data set by the value of the index j in the weights vector. To write this procedure mathematically, let's define an operator $\bullet$:

$$\begin{aligned} &A \bullet \overrightarrow{B} = A \cdot diag(\overrightarrow{B}), \\ &diag(\overrightarrow{B}) = \sum_{i=1}^{m} (e_i e_i^T)(e_i^T \overrightarrow{B}) \in (\mathbb{R}^m \times \mathbb{R}^m), \\ &A \in (\mathbb{R}^n \times \mathbb{R}^m),\ \overrightarrow{B} \in \mathbb{R}^m,\ e_i \text{ is the } i\text{th basis} \end{aligned} \tag{3}$$

Therefore:

$$\mathbb{X} \bullet \overrightarrow{\mathbb{W}} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \bullet \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} x_{11}w_1 & \cdots & x_{1m}w_m \\ \vdots & \ddots & \vdots \\ x_{n1}w_1 & \cdots & x_{nm}w_m \end{bmatrix} \tag{4}$$

The next step is to feed the clustering method with the newly generated weighted features matrix to obtain the labels vector. Thus, according to Eq. 1: $K(\mathbb{X} \bullet \overrightarrow{\mathbb{W}}) = \overrightarrow{\mathbb{L}} \in \mathbb{Z}_+^n$. With this, the labels vector $\overrightarrow{\mathbb{L}}$ and the bias property vector $\overrightarrow{\mathbb{B}}$ are known, so it is possible to get the sum of the intracluster variance

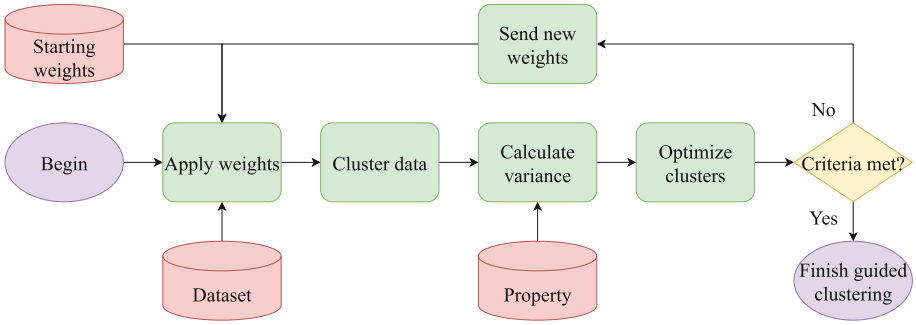

Fig. 1. Guided clustering flowchart. The first step is to apply the starting weights (normally 1) in each column of the data set. The next step is to cluster the data and then use the property data to calculate the property intracluster variance. Having the variance value, the next part of the program tries to reduce the variance by rerunning the previous steps with different weights. Finally, when a criteria is met (maximum iterations number, for instance), the program is finished.

of $\overrightarrow{\mathbb{B}}$:

$$V(\overrightarrow{\mathbb{L}}, \overrightarrow{\mathbb{B}}) = \sum_{j=1}^{k} \sigma^2 \left(\left[c_1 \cdots c_n \right]^T \right),$$
$$c_i = \begin{cases} \overrightarrow{\mathbb{B}}_i, \text{ if } \overrightarrow{\mathbb{L}}_i = j \\ 0, \ otherwise \end{cases}, \ i \in [1..n], \tag{5}$$
$$k \text{ is the number of clusters}$$

Then, all the pieces are ready to run the global optimization algorithm. Following the Set of Eq. 2, it is obtained that:

$$\underset{\overrightarrow{\mathbb{W}} \in \mathbb{R}^m}{min/max} \ f(\overrightarrow{\mathbb{W}}) = V(\overrightarrow{\mathbb{L}}, \overrightarrow{\mathbb{B}}) \in \mathbb{R}_+, \tag{6}$$
$$\text{subject to the algorithm's parameters}$$

Finally, the optimization algorithm is run until the maximum number of iterations is reached or the acceptance criteria is met. Figure 1 shows a diagram that summarizes the entire process. Although the implementation used to generate the results for this work uses K-Means as the clustering algorithm and Basin-Hopping as the global optimization one, they could be replaced.

3 Quantum-Chemistry Databases

3.1 Cu_{55} Nanoclusters

This database was built to expand the knowledge about morphology on metallic nanoclusters. The geometries obtained via Revised Basin-Hopping Monte Carlo (rBHMC) in-house tool [31] for Cu_{55} nanoclusters, where interactions are

described via ReaxFF [34] with parameters from Nielson *et al.* [27]. In ten runs of rBHMC, a total of 1048 Cu_{55} geometries were generated, all corresponding to local minima on the potential energy surface of those systems. The focus of the clustering application in the original paper was to generate samples for additional analysis, but here we will look directly at the guided clustering process as a way to extract physico-chemical insights and strengthen the specialist's understanding of the data set itself.

3.2 $CeZrO_4$

This database was previously used to investigate the energetic, structural and electronic properties of mixed cerium and zirconium nanoclusters, which are very important in nanocatalysis [14]. In this original context, the data was used as it is, for direct analysis of the phisico-chemical quantities and the relation between them. Our idea here is to exploit the fact that the data can be labeled by composition (i.e., we already know that the data can be clustered using the Ce/Zr ratio as label) to see how the guidance can show new clusters that are different from this trivial one and bring up new facts about the stability of these mixed oxide nanoclusters.

The set contains mixed oxides from $Ce_{15}ZrO_{30}$ to $CeZr_{15}O_{30}$ with 1646 structures generated from the variation in the amount of Ce and Zr atoms, exploring different patterns of substitution of Ce and Zr, but also different geometrical arrangements and nanocluster shapes for each concentration.

4 Technical Details of the Computational Framework

4.1 Number of Clusters by Visualization Approach

It is possible to visually approximate the amount of clusters in the system by looking at a two-dimensional projection of the data. Take Fig. 5 as an example: looking at the chart **d**, even if there were no distinct colors, it is easy to visually reach the conclusion that there are around 16 clusters of data.

In this particular example, the number of true clusters (defined by the specialist) coincides with the projection. This does not mean that this is always the case, because projections are not perfect representations of the distances between the items in higher dimensions, they just provide an idea on how the entries are distributed.

To better meet the needs of the specialist, GuidedClustering provides two different projection solutions: the tested and trusted PCA (Principal Component Analysis) and the state-of-the-art t-SNE (t-distributed Stochastic Neighbor Embedding). Both implementations are from the scikit-learn library.

PCA receives the multidimensional table as input, then find a new set of orthogonal variables (the *principal components*) which explain most of the variance in the data, by removing inter-correlated columns and performing other mathematical procedures. Since it is possible to obtain an arbitrary number (up

to the amount of dimensions that originally exists) of principal components, selecting the two most important allows the construction of more meaningful 2D charts [1].

The other method, t-SNE [23], is the actual state-of-the-art in projections. It differs from PCA in two main points: it is a non-linear method (which lets it understand more complex underlying patterns) and it cares about the local neighbors (since it tries its best to keep similar items close together).

There are two different measures to evaluate a PCA projection: Explained Variance Ratio and Trustworthiness. The explained variance ratio is a metric that reflects the percentage of the variance that is attributed to each component. Trustworthiness expresses the preservation of local structures and is given by the Eq. 7:

$$\begin{aligned} T(k) &= 1 - \alpha(k)\beta(k), \text{ such that} \\ \alpha(k) &= \frac{2}{nk(2n - 3k - 1)}, \\ \beta(k) &= \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} (r(i,j) - k) \end{aligned} \tag{7}$$

where $r(i, j)$ represents the rank of the output space item j according to the pairwise relation with the other output space items. The set of k-nearest neighbors in the output space but not in the input space [36] is given by $U_i^{(k)}$.

The Trustworthiness metric is also able to evaluate t-SNE's projection performance, while the Explained Variance Ratio is something more attached to the way PCA creates the projection.

Thus, having ran GuidedClustering for a given k, it is possible to verify if k is a satisfactory value by visual inspection, calculation of the evaluation measures or by deep inspecting which entries belong to which cluster (looking into their features).

4.2 The Importance of the Silhouette to Obtain the Number of Clusters

K-Means will always divide the data set into the specified number K of clusters, given that there are enough entries to cluster, but the problem is to identify if the division is meaningful (of high quality). Over the years, some methods were developed to address this issue, silhouette being one of them.

Silhouette uses the mean intracluster distance a and the mean distance b of the closest group for each item (note that b refers to the group closest to the item under analysis, other than the group it belongs to). The result of this measure for an item i is inside the range $[-1, 1]$, where, -1 indicates that the item has been incorrectly assigned and 1 is the best possible value. Given a data set $\mathbb{X}$, the silhouette score over all samples is given by Eq. 8.

$$\mathbb{S} = \frac{\sum_{i=1}^{|\mathbb{X}|} \frac{b_i - a_i}{max(a_i, b_i)}}{|\mathbb{X}|}, \; 2 \leq K \leq |\mathbb{X}| - 1 \tag{8}$$

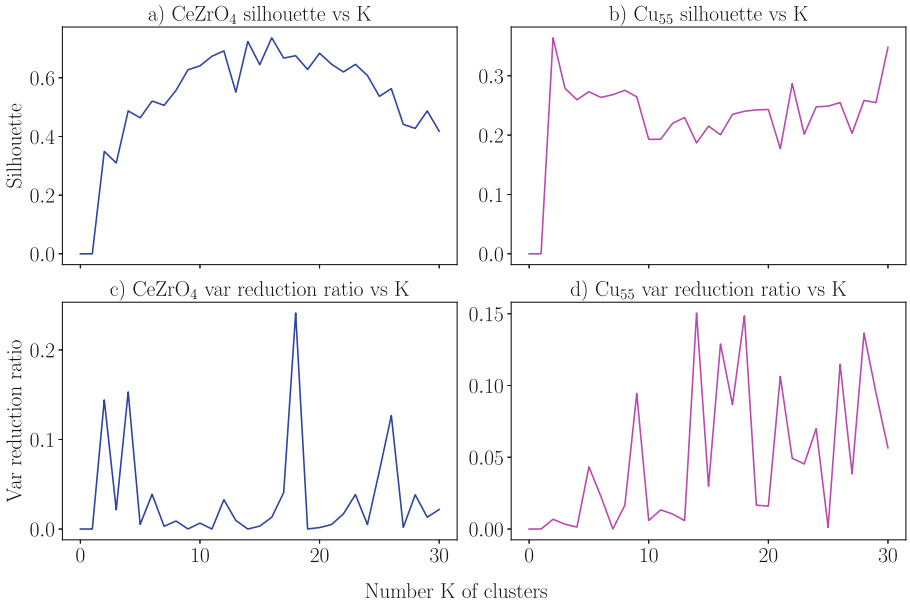

Fig. 2. In **a** and **b**, it is demonstrated the silhouette score, as K increases, for the initial unguided clustering of the data sets $CeZrO_4$ and Cu_{55}, respectively. The charts **c** and **d** show the percentage of overall intracluster variance reduction, as K increases, for the data sets $CeZrO_4$ and Cu_{55}, respectively.

The silhouette coefficient depends solely on the partition of the data and not on the clustering algorithm used, and is a powerful tool to help the interpretation and validation of cluster formations [33]. Since it can quantitatively express the clustering quality, GuidedClustering can use it to automatically choose a value K of clusters, if the user wishes to. The provided value is only a suggestion and should be further analyzed by the specialist, since there is no right answer to the correct amount of clusters (it depends on the context and what are the expected results). Figures 2(a) and (b) show the silhouette scores for 2 different chemical data sets.

4.3 The Role of the Variance Minimization in the Clustering

It is typical for clustering algorithms to minimize the distance between the entries within a cluster (consequently reducing the within-cluster variance), in fact, that is precisely what K-Means tries to do. Unfortunately, K-Means does not have any explicit physical-chemical knowledge, so its procedure does not attribute weights to the different features, which can result in suboptimal results in the point-of-view of the specialist.

The intracluster variance of a given property shows how heterogeneous clusters are in respect to that property. By guiding the clustering procedure with the proposed optimization algorithm, it is possible to deform the space where the items co-exist in order to weight the features to potentially move the entries from one cluster to another. This is done in an attempt to form more homogeneous clusters, highlighting the given property. By generating more homogeneous clusters, it becomes easier to analyze the data (since it is more coherent and cohesive) and verify the underlying patterns related to the property used in the guidance.

This is a way to introduce domain-specific knowledge into an algorithm that is agnostic to the domain it is being applied and can provide improvements in the cluster structures, as will be discussed in the next section.

5 Chemical Trends Obtained by Guided Clustering Analysis

The philosophy behind minimizing the variance of a physico-chemical descriptor along a guided clustering lies on getting clusters that represent well this descriptor accumulation points on the configuration space. Clustering can be a powerful tool to explore the Potential Energy Surface, highlighting local minima and basins that are better labeled by a single energy value. In this section, we present two test cases where we explore this feature to perform clustering in sets of nanoclusters.

5.1 Cu_{55} Nanoclusters

The structures on this data set were used in a previous work from Mendonça et al. [25], where a default clustering strategy, without any guidance, was used. The structures on the data set were originally obtained via Revised Basin-Hopping Monte Carlo Algorithm (rBHMC) [31] using previously parameterized ReaxFF potential energy [27,34]. A total of 1048 nanocluster structures were obtained with respective energies also reported.

Here, the structural data was processed to obtain the eigenvalues of the Coulomb matrix for each structure. In the Basin-Hopping approach, we try to explore the configuration space, jumping between different local minima and trying to explore different basins of the potential energy surface. These basins are regions of concentration around local minima points on the potential energy, and we expect a concentration of structures with similar energy value inside the same basin. Based on that, we propose the use of the potential energy to guide the clustering towards those minima, aiming to identify them. Visualization of the distribution of the potential energy, as well as clustering results, can be seen in Fig. 3.

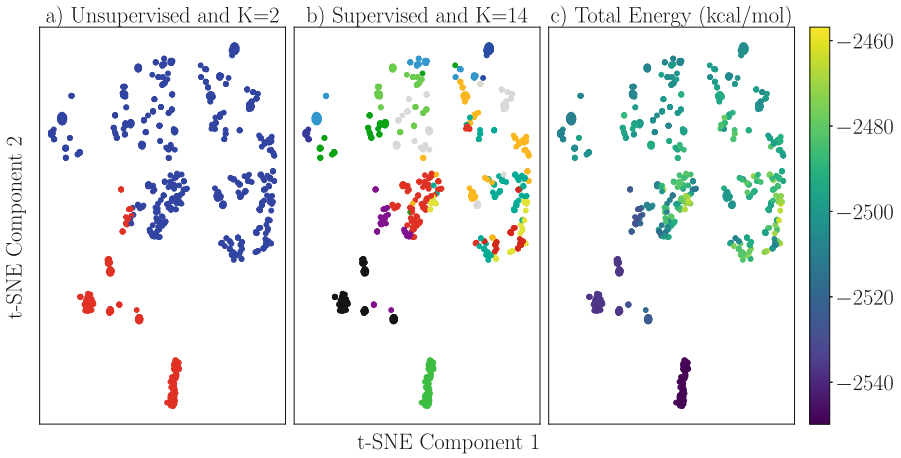

Fig. 3. All the charts show the 2D projection of the 55D data using the t-SNE method. Chart **a** has 2 colors, representing the label of each item given by the unguided clustering procedure for K=2. Chart **b** has 14 colors, each one representing a label attributed by the guided clustering procedure for K=14. Finally, chart **c** is colored by the value of total energy property. t-SNE perplexity was set as 30.0.

Observing the unguided clustering, we diagnosed that two clusters is the best configuration, in respect of silhouette metric, for this data set. The structures were separated in sets, one with higher energy and bigger variance and another with lower energy and lower variance. Those sets describe the difference between structures with high ordering (associated with the solid phase) and low ordering (associated with amorphous or liquid nanoclusters). Not by coincidence, those two clusters showcased potential energy variance of 102.86 and 86.45 kcal/mol, the equivalent (via $3/2Nk_bT$) to 627 and 527 K, in the range of the phase transition temperatures the authors listed on the original work. This can be seen as the clustering algorithm is grouping structures that are in basins accessible at these temperatures. We also highlight that for $k = 2$, the guided and unguided clusterings are very similar (Normalized Mutual Information (NMI) evaluated as 0.986), showing that this is a very natural configuration that does not demand a strong distortion of the descriptors space to be captured.

In another hand, this analysis was not enough to identify the different basins inside the solid phase structures, that we know exist and are explored via additional molecular dynamics on the original publication. A completely different scenario is seen in the case where we select k based on the optimal guidance performance. Here, we identified that $k = 14$ minimizes the potential energy variance per cluster, which indicates that this configuration is far more adapted to the detection of small local minima on the PES explored by the rBHMC. In fact, three of the four clusters with lower average energies can be directly associated to the isometric I_h, pseudo-tetragonal (C_2, but close to T_d) and the capped decahedron (D_{5h}) shapes of nanoclusters, already known as highly stable

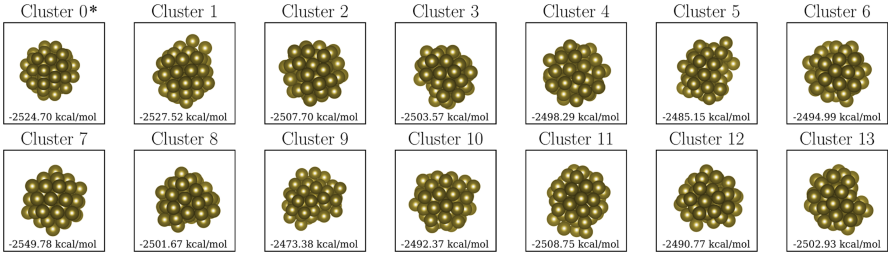

Fig. 4. The charts show every representative sample of the clusters formed by guided clustering applied to the Cu_{55} data set. The respective potential energies are displayed right below the 3D representations. In the case of cluster 0 (*), we to choose show the more symmetrical structure (D_{5h}) observed in this class, manually selected by the specialist. The automatic representative selected was a deformed I_h structure.

for Cu_{55}. The latter was not automatically selected as the representative of its cluster but has been included in Fig. 4, for ease of viewing. Here, we also saw that the weights on the coordinates of the configuration space greatly impacted the clustering process, having NMI= 0.730. Additionally, we observed that the highly symmetrical I_h and D_{5h} nanoclusters are present in very specialized clusters, with variance inferior to 0.15 kcal/mol, which can be understood as those structures being located in a very narrow basin and reinforces the stability associated to those geometries. Another structure was identified in a very narrow group (Cluster 3 in Fig. 4, variance of 0.16 eV), composed of a I_h Cu_13 core and a disorganized surface with [555] and [666] CNA signatures. This was not reported yet, and since this structure can be a solid liquid intermediate, further investigations are necessary to understand it in depth.

5.2 $CeZrO_4$

The set we are calling $CeZrO_4$ is actually a set of nanoclusters of ceria-zirconia with chemical formula $(CeO_2)_n(ZnO_2)_{15\text{-}n}$. This set was originally build by Felicio-Sousa et al. [14] to explore the stability and electronic properties of those systems. The quantum chemistry simulations were performed using numerical atom-centered orbitals [6,15] (NAOs) in the Fritz-Haber Institute Ab-initio molecular simulations [6,28] (FHI-aims) package. Perdew-Burke-Ernzenhof functional [30] (PBE) with Zeroth-order relativistic approximation [35] (ZORA) was used. Since the data was already refined in a quantum calculation, the clustering here is exploited to improve our understanding of the structure-property relation.

As the nuclei charge plays an important part in the Coulomb matrix definition, we visually observed that the proportion of each nucleus in the nanocluster (or even n itself) is a good candidate for clustering label for the data, since it looks well clustered in this representation space. We expect as a consequence that both methods will be able to cluster the data similarly. In fact, the natural

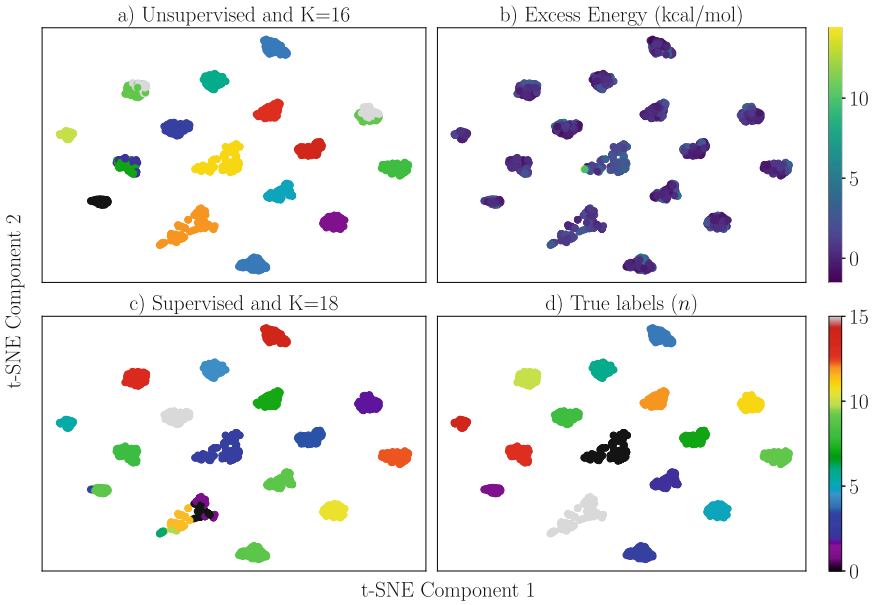

Fig. 5. All the charts show the 2D projection of the 45D data (Coulomb matrices eigenvalues) using the t-SNE method. Chart **a** has 16 colors, which represent the final configuration obtained by the unguided clustering procedure with $k = 16$. Chart **b** is colored by the value of excess energy property. Chart **c** has 18 colors (clusters) and was clustered by the guided procedure with $k = 18$. Finally, Chart **d** also has 16 colors, but this time they represent the true labels given by the amount of Ce and Zr on each entry. t-SNE perplexity was set as 30.0.

clustering in 16 groups was the one to minimize the silhouette metric, and the weighting on the features reduced the average variance per cluster by only 1% (NMI= 0.953).

In Figure 5(a), we show the t-SNE projection of the data for the unguided clustering when $k = 16$. We see that the clusters (gray and lime in Figure 5(a)) share together two visually distinct clusters on the unguided clustering. Those visual clusters correspond to $n = 12$ and $n = 11$, and were correctly separated by the guidance process. In the other hand, we see that both the guided and unguided labeling case clusters together two other visual clusters, associated to $n = 3$ and $n = 4$, that are the most stable systems reported by Felicio-Sousa et al. [14].

The guidance was done with the variance on the excess energy, and we observed that the number of clusters that minimized this metric (reducing it by 24%) was 18. Again, we see that most of the cluster labeling was similar between the guided and unguided case, with NMI= 0.940. The guided version ends up associating five representatives for $n = 15$, which can be explained by the natural lower variance on excess energies on those nanoclusters even with

large geometrical changes. It also ends up clustering together $n = 1$, 2, and 3, combining three high variance clusters (in the unguided case) in the same cluster, reducing the overall average variance per cluster. The unguided clustering ends up splitting $n = 2$ and $n = 15$. This points to the fact that the clusters have very similar stability behavior, independent of the different composition. Local features of the geometries are more important here to determine stability than the Ce/Zr ratio, and this is why the natural clusters ended up mixed in the biased training.

We also observed that traditional metrics on guided clustering point towards the fact that the guided version for $k = 18$ is the one that better depicts the expected natural clustering in n. In particular, we measured homogeneity and completeness [32], which are 0.967 and 0.983, respectively, and the NMI in relation to the labeling with n, which is 0.975.

6 Conclusion

Here, we proposed a clustering guidance strategy based on feature scaling informed by the value of an auxiliary observable to enhance the relationship between clustering labels and the nature of the experiments that provided the data. In this approach, the k-means clustering is guided in two ways: (i) Global silhouette metric is used to suggest the number of clusters. (ii) Basin-Hopping optimization is used to find weights for each feature that minimize the average intracluster variance on the informed observable. It is also worth mentioning that projections are valuable tools for obtaining insights into how high-dimensional data is structured.

The proposed framework was evaluated on two sets of molecular structures represented by their respective Coulomb matrix eigenvalues. The combination of the suggested k and the guided variance strategy was shown to be a path to both explore clustering with a more broad thermodynamic sense (like in solid-liquid-like clusters for $Cu_5 5$, when $k = 2$) and detailed explorations on more complex structure-property relations (like the clustering relation to the number of Ce atoms per nanocluster in the $CeZrO_4$ case). Finally, our results with this procedure might inspire future automated sample selection and configuration space exploration strategies.

7 Data and Software Availability

Data can be obtained directly with the authors under request.

Acknowledgements. The authors gratefully acknowledge support from FAPESP (São Paulo Research Foundation) and Shell, projects No. 2017/11631−2, 2018/21401−7 and 2022/09285−7, and the strategic importance of the support given by ANP (Brazil's National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation. The authors also thank for the infrastructure provided to our computer cluster by the Department of Information Technology - Campus São Carlos.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. WIREs. Comput. Statist. **2**(4), 433–459 (2010). https://doi.org/10.1002/wics.101. https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101
2. Bai, L., Liang, J., Cao, F.: Semi-supervised clustering with constraints of different types from multiple information sources. IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1 (2020). https://doi.org/10.1109/TPAMI.2020.2979699
3. Batista, K.E.A., Ocampo-Restrepo, V.K., Soares, M.D., Quiles, M.G., Piotrowski, M.J., Da Silva, J.L.F.: Ab Initio investigation of co_2 adsorption on 13-atom $4d$ clusters. J. Chem. Inf. Model. **60**(2), 537–545 (2020). https://doi.org/10.1021/acs.jcim.9b00792. https://doi.org/10.1021/acs.jcim.9b00792
4. Batista, K.E.A., Soares, M.D., Quiles, M.G., Piotrowski, M.J., Da Silva, J.L.F.: Energy decomposition to access the stability changes induced by co adsorption on transition-metal 13-atom clusters. J. Chem. Inf. Model. **61**(5), 2294–2301 (2021). https://doi.org/10.1021/acs.jcim.1c00097. https://doi.org/10.1021/acs.jcim.1c00097. pMID: 33939914
5. Bayada, D.M., Hamersma, H., van Geerestein, V.J.: Molecular diversity and representativity in chemical databases. J. Chem. Inf. Comput. Sci. **39**(1), 1–10 (1999)
6. Blum, V., et al.: Ab initio molecular simulations with numeric atom-centered orbitals. Comput. Phys. Commun. **180**(11), 2175–2196 (2009). https://doi.org/10.1016/j.cpc.2009.06.022. https://doi.org/10.1016/j.cpc.2009.06.022
7. Boubchir, M., Boubchir, R., Aourag, H.: The principal component analysis as a tool for predicting the mechanical properties of perovskites and inverse perovskites. Chem. Phys. Lett. **798**, 139615 (2022)
8. Brockherde, F., Vogt, L., Li, L., Tuckerman, M.E., Burke, K., Müller, K.R.: Bypassing the Kohn-sham equations with machine learning. Nat. Commun. **8**(1), 1–10 (2017)
9. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A.: Machine learning for molecular and materials science. Nature **559**(7715), 547 (2018)
10. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Statist. **3**(1), 1–27 (1974). https://doi.org/10.1080/03610927408827101. https://www.tandfonline.com/doi/abs/10.1080/03610927408827101
11. Cha, S.H.: Comprehensive survey on distance/similarity measures between probability density functions. City **1**(2), 1 (2007)
12. Craw, S.: Manhattan Distance, p. 639. Springer, US, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8_506
13. Engels, M.F., Thielemans, T., Verbinnen, D., Tollenaere, J.P., Verbeeck, R.: Cerberus: a system supporting the sequential screening process. J. Chem. Inf. Comput. Sci. **40**(2), 241–245 (2000)
14. Felício-Sousa, P., et al.: Ab initio insights into the structural, energetic, electronic, and stability properties of mixed $ce_nzr_{15-n}o_{30}$ nanoclusters. Phys. Chem. Chem. Phys. **21**(48), 26637–26646 (2019). https://doi.org/10.1039/c9cp04762j. https://doi.org/10.1039/c9cp04762j
15. Havu, V., Blum, V., Havu, P., Scheffler, M.: Efficient integration for all-electron electronic structure calculation using numeric basis functions. J. Comput. Phys. **228**(22), 8367–8379 (2009). https://doi.org/10.1016/j.jcp.2009.08.008. https://doi.org/10.1016/j.jcp.2009.08.008
16. Hkdh, B.: Neural networks in materials science. ISIJ Int. **39**(10), 966–979 (1999)

17. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999). https://doi.org/10.1145/331499.331504. https://dx.doi.org/10.1145/331499.331504
18. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Comput. Surv. **31**(3), 264–323 (Sep1999) 10.1145/331499.331504, https://doi.org/10.1145/331499.331504
19. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010). https://doi.org/10.1016/j.patrec.2009.09.011. https://www.sciencedirect.com/science/article/pii/S0167865509002323. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)
20. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
21. van Laarhoven P.J.M., A.E.: Simulated annealing. In: Simulated Annealing: Theory and Applications, vol. 37, pp. 7–15. Springer, Dordrecht (1987). https://doi.org/10.1007/978-94-015-7744-1_2
22. Lo, Y.C., Rensi, S.E., Torng, W., Altman, R.B.: Machine learning in chemoinformatics and drug discovery. Drug Discov. Today **23**(8), 1538–1546 (2018)
23. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008). https://www.jmlr.org/papers/v9/vandermaaten08a.html
24. McGregor, M.J., Pallai, P.V.: Clustering of large databases of compounds: using the mdl "keys" as structural descriptors. J. Chem. Inf. Comput. Sci. **37**(3), 443–448 (1997)
25. de Mendonça, J.P.A., Calderan, F.V., Lourenço, T.C., Quiles, M.G., Da Silva, J.L.F.: Theoretical framework based on molecular dynamics and data mining analyses for the study of potential energy surfaces of finite-size particles. J. Chem. Inf. Model. **62**(22), 5503–5512 (2022). https://doi.org/10.1021/acs.jcim.2c00957. https://doi.org/10.1021/acs.jcim.2c00957. pMID: 36302503
26. Morgan, D., Jacobs, R.: Opportunities and challenges for machine learning in materials science. Annu. Rev. Mater. Res. **50**(1), 71–103 (2020). https://doi.org/10.1146/annurev-matsci-070218-010015
27. Nielson, K.D., van Duin, A.C.T., Oxgaard, J., Deng, W.Q., Goddard, W.A.: Development of the ReaxFF reactive force field for describing transition metal catalyzed reactions, with application to the initial stages of the catalytic formation of carbon nanotubes. J. Phys. Chem. A **109**, 493–499 (2005)
28. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2006). https://doi.org/10.1007/978-0-387-40065-5
29. Olson, B., Hashmi, I., Molloy, K., Shehu, A.: Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules. Advances in Artificial Intelligence 2012 (2012). https://doi.org/10.1155/2012/674832
30. Perdew, J.P., Ernzerhof, M., Burke, K.: Rationale for mixing exact exchange with density functional approximations. J. Chem. Phys. **105**(22), 9982–9985 (1996). https://doi.org/10.1063/1.472933
31. Rondina, G.G., Da Silva, J.L.F.: Revised basin-hopping Monte Carlo algorithm for structure optimization of clusters and nanoparticles. J. Chem. Inf. Model. **53**(9), 2282–2298 (2013). https://doi.org/10.1021/ci400224z
32. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 410–420 (2007)

33. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7. https://www.sciencedirect.com/science/article/pii/0377042787901257
34. van Duin, A.C.T., Dasgupta, S., Lorant, F., Goddard, W.A.: ReaxFF: a reactive force field for hydrocarbons. J. Phys. Chem. A **105**, 9396–9409 (2001)
35. van Lenthe, E., Snijders, J.G., Baerends, E.J.: The zero-order regular approximation for relativistic effects: the effect of spin-orbit coupling in closed shell molecules. J. Chem. Phys. **105**(15), 6505–6516 (1996). https://doi.org/10.1063/1.472460
36. Venna, J., Kaski, S.: Local multidimensional scaling. Neural Netw. **19**(6), 889–899 (2006). https://doi.org/10.1016/j.neunet.2006.05.014. https://www.sciencedirect.com/science/article/pii/S0893608006000724. Advances in Self Organising Maps - WSOM2005
37. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al.: Constrained k-means clustering with background knowledge. In: ICML, vol. 1, pp. 577–584 (2001)
38. Wales, D.J., Doye, J.P.K.: Global optimization by basin-hopping and the lowest energy structures of Lennard-jones clusters containing up to 110 atoms. J. Phys. Chemis. A **101**(28), 5111–5116 (1997). https://doi.org/10.1021/jp970984n
39. Ward, J.H., Jr.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
40. Yang, X.S.: Introduction to Mathematical Optimization: From Linear Programming to Metaheuristics. Cambridge International2 Science Publishing (2008)
41. Zheng, J., Lu, T., Lian, Z., Li, M., Lu, W.: Machine learning assisted classification of post-treatment amines for increasing the stability of organic-inorganic hybrid perovskites. Mater. Today Commun. **35**, 105902 (2023)
42. Zibordi-Besse, L., Seminovski, Y., Rosalino, I., Guedes-Sobrinho, D., Da Silva, J.L.F.: Physical and chemical properties of unsupported $(mo_2)_n$ clusters for $m = ti$, zr, or ce and $n = 1--15$: A density functional theory study combined with the tree-growth scheme and euclidean similarity distance algorithm. J. Phys. Chem. C **122**(48), 27702–27712 (2018). https://doi.org/10.1021/acs.jpcc.8b08299