Machine Learning Project 2: Quantum Machine Learning

Joachim Koerfer, 293815, joachim.koerfer@epfl.ch[1]
Luca Viano, 294418, luca.viano@epfl.ch[1]
Jannik Reichert, 300198, jannik.reichert@epfl.ch[2]

[1]Master in Computational Science and Engineering, EPFL
[2]Bachelor in Computer Science, EPFL/Technische Universität Berlin

***Abstract*— The aim of this work is the construction of a method to find significant clusters of molecules used in catalysis. At first, the meaningful principal components were selected using a statistical analysis to circumvent the curse of dimensionality in oreder to avoid overfitting. The dimensionality was reduced by Principal Component Analysis. Then, the clusters were found by the density-based method DBSCAN. Further analysis reveals similar structure in each of the clusters regarding to different molecular characteristics. Furthermore, the clusters show to be built by atomic-bond and atomic charateristics of the ligands.**

I. Introduction

The provided dataset contains physical descriptors of 25116 compounds in the form $L_1 - M - L_2$ obtained considering all possible combinations between 6 metals (M) and a pair of ligands (L_1, L_2) chosen among 91 possibilities. The Laboratory for Computational Molecular Design has already successfully applied a supervised machine learning approach on different matrix representations of these molecules in order to predict the corresponding x-axis energy of the volcano plot (a visualisation technique to discriminate whether a molecule behaves as a good catalyst or not) [2]. This work consists, instead, in applying an unsupervised machine learning algorithm on the same compounds to find clusters of molecules with similar characteristics. In this work, the molecule descriptors are not matrices as in [2] but scalar quantities described in I-A. Section II provides the reasons why a particular Machine Learning approach was chosen and, eventually, section III outlines the similarities of the molecules belonging to the same clusters and dicusses the factors causing the clusters split.

A. Presentation of the data

Each compound is described by eight scalar features. They are molecular weight, shape index, molecular volume, number of carbon atoms in the molecules and, finally, the ratio between atoms of the element $*$ bound to the metal and the length of these bonds evaluated for the elements P, N, C, O.

II. Choice of the model

The chosen model should ideally find a compromise between two opposite requirements: Indeed, on the one hand it is well known that each clustering algorithms becomes less efficient when a high number of dimensions is considered. On the other hand, it may be reasonable assuming that assessing the closeness of two molecules requires a large amount of physical quantities and, possibly, of interaction terms among them. The different criteria that have been used to assess the proper number of dimensions, are described in II-A. The next step is assessing the best clustering technique to group the molecules in the chosen space. Notice that the number of existing clusters is unknown and that the convexity of the sets cannot be guaranteed. These reasons make "centroid" approaches like K-means ineffective. In II-C, the detailed selection process is provided.

A. Ideal number of dimensions

The provided features may suffer from collinearity. As an example, it can be supposed that the molecular length is dependent on the molecular weight. In addition, it may be reasonable considering that the energy of a molecule is determined not only by the linear independent effects of quantities like number of atoms, shape index, molecular weight etc. Hence a polynomial expansion including both the power of the columns of X and the pairwise interactions among them could be considered. On the other hand, a feature expansion operation can make the data so sparse that no clustering algorithms can offer meaningful results. Considered both these reasons, it is necessary to go through all the features and verify which truly deserve to be considered in the model. The features which do not significantly add information are discarded from the data matrix X. The proper number of dimensions can be inspected using cross validation techniques adapted to an unsupervised problem. A suitable way to face this problem is the technique known as Gabriel holds-out. [4].

1) Gabriel holds out: This technique is based on a cross validation approach applied to a supervised problem derived from the unsupervised one splitting the data into virtual training and test data. It proceeds as follows.

Firstly, a row-wise partition P_{rows} of K subsets that contain equal numbers of row indices of X is defined. The same is done considering the columns indices, defining a L subsets partition of $P_{columns}$. Then, consider a row-indices subset $k \in P_{rows}$, a column-indices subset $l \in P_{columns}$. Adopting the notation $X[\cdot, \cdot]$, let $X[k,l]$ be the submatrix defined considering the rows and columns indices in k and l respectively. According to the introduced notation, rename $X[k,l]$ as Y_{test},

$X[k, P_{columns} \setminus l]$ as X_{test}, $X[P_{rows} \setminus k, P_{columns} \setminus l]$ as X_{train} and, finally, $X[P_{rows} \setminus k, l]$ as Y_{train} (see table I). Notice that, in this context, the symbol $\setminus$ defines the set subtraction.

X_{train}	Y_{train}	X_{train}
X_{test}	Y_{test}	X_{test}
X_{train}	Y_{train}	X_{train}

TABLE I: Block representation of the submatrices defined, partitioning the matrix X, at each step in the Gabriel cross validation. The column occupied by the submatrices Y_{test} and Y_{train} is the one defined by the indices included in the subset l. The row where X_{test} and Y_{test} lie represents the row indices in k. Notice that the four submatrices are not in general composed by contiguous cells.

Now, the two matrices Y_{test} and Y_{train} can be considered labels of a supervised problem whose inputs are the rows of X_{test} and X_{train} respectively.

The Gabriel strategy consists on training a linear predictor on the Singular Value Decompsition (SVD) of X_{train} truncated to the dimension d using Y_{train} as labels. Then the prediction of this model considering X_{test} as an input is compared with Y_{test}. If the variance along all the d principal components is significantly high, the prediction error PE defined as $\|Y_{predicted} - Y_{test}\|_2$ is expected to decrease drastically when the number of dimensions d is increased. When, on the contrary, the data are not expanded along a dimension, including this dimension in the model does not improve the PE remarkably. Even worse, it can happen that the prediction error increases in heavily noisy data since the added dimension increases the risk of overfitting [4]. Iterating this procedure for all subsets $k \in P_{rows}$ and $l \in P_{columns}$, a mean prediction error $\overline{PE}$ can be defined averaging over the prediction errors $PE(k,l,d)$ computed for all the possible Gabriel configurations defined by k and l:

$$\overline{PE}(d) = \frac{\sum_{k,l} PE(k,l,d)}{k \cdot l} \tag{1}$$

This quantity does not depend any longer on a specific choice of submatrices.

Consequently, a plot of the $\overline{PE}$ versus the number of dimension d can be considered (see appendix). If $d^* = \arg\min_d \overline{PE}(d)$ exists, d^* is the number of dimensions that should be considered. If instead $\overline{PE}$ is monotonously decreasing, d^* is equal to the maximum number of dimensions D (number of columns of X). The latter case took place considering the molecules data and, so, it can be concluded that noise is not present on a sensible level. Although, before claiming that all the features deserve to be included in the model, another check should be performed. Indeed, a hypothesis testing for the null hypothesis H_0 : "*The weight that the linear predictor of the configuration (k, l) assigns to the d-th principal component is equal to zero.*" can be considered. When the null hypothesis holds, it can be shown that:

$$\frac{PE(k,l,d) - PE(k,l,d-1)}{\frac{PE(k,l,D')}{N-D'}} \sim F_{1,N-D'} \tag{2}$$

where F is the probability density function of the F-distribution and D' is the number of columns of X_{train}. Hence, the p-value, seen as a function of the number of the considered features d (eq. 3), can be used to assess whether the model with d dimensions should be preferred over the model with $d-1$ dimensions or not.

$$p(k,l,d) = 1 - cdfF_{1,\ N-D'}\left(\frac{PE(k,l,d) - PE(k,l,d-1)}{\frac{PE(k,l,D')}{N-D'}}\right) \tag{3}$$

Although p-values close to 1 are evidences in support of the null hypothesis, when $p(k,l,d)$ exceeds an onset threshold θ, it can be deduced that the d-th principal component is nearly unnecessary to predict the labels Y_{test} and so it does not add consistent important information to the $d-1$ principal components.

Notice that (2) is valid only if the prediction error of a single Gabriel configuration is considered. It does not hold for $\overline{PE}$. As a consequence, $p(k,l,d)$ must be evaluated for each combination of k and l and the model of dimension d is discarded when the ratio between $\sum_{k,l} \mathbb{1}\{p(k,l,d) > \theta\}$ and the total number of Gabriel configurations $K \cdot L$ is significantly larger than zero.

The concrete application of this criterion on the provided dataset is shown in figure 1. The high number of significant p-values at the third principal component is a hint of clustering the data in a bi-dimensional space. Reducing the data dimensions following this criterion can be harmful when a supervised regression problem is taken into account but it is useful for alleviating the curse of dimensionality that affects each clustering algorithm, thus reducing the risk of overfitting.

B. Dimension reduction

The techniques previously described provide the optimal number of dimensions D^*. The dimensionality reduction down to D^* was done applying the PCA algorithm. The transformed data matrix will be denoted as X_{PCA}.

C. Choice of the clustering algorithm

It is reasonable expecting clusters containing molecules with similar energies along the volcano plot x-axis and with similar chemical composition of the ligands. Although, the former is continuous and so an arbitrary partition into classes should be considered before applying a "centroid" based clustering algorithm. Furthermore, it is desirable that the algorithm that is used does not depend on how many groups of similar ligands are taken into account. It follows that the number of existing clusters cannot be estimated and, consequently, it is necessary to apply a method that does not require this information as an input. A method that respects these demands is DBSCAN [1]. A further advantage

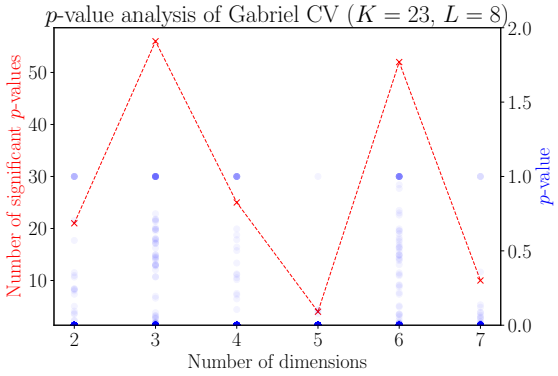

Fig. 1: The blue transparent circles represent the p-values for each combination of k and l where high opacity represents overlapping p-values. The red line represents the number of p-values above the onset threshold of 0.05. It can be seen that when three dimensions are considered, approximately 55 Gabriel configurations (30% of the performed $K \cdot L = 184$ tests) yield a significant p-value. This suggests that only the two principal components of the PCA decomposition of X should be considered to cluster the data.

provided by this technique is the possibility to detect non-convex clusters. This is needed because no information about the clusters shape can be extrapolated a priori. In the following, both a description of the hyperparameters setting procedure and a deeper explanation of the algorithm are given.

1) DBSCAN: This approach does not require a number of initial "centroids" because it does not detect clusters as *sets of points*[1] which are close to each other but as *regions* with a high data density. These are detected accordingly to the input parameters `NumPoints` and `Eps` as follows:

- Each point that includes at least `NumPoints` points in its surrounding hypersphere of radius `Eps`[2] is defined a core point. In addition, a point A is defined *directly density-reachable* from B iff B is a core point and A belongs to the neighbourhood of B with radius `Eps`. Then, C and D are defined *density reachable* iff it exists a sequence $C, P_1, \ldots, P_i, P_{i+1}, \ldots, P_n, D$ where the pairs C, P_1 and P_i, P_{i+1} and P_n, D are directly density reachable for all $1 \leq i < n$.
- Two points A and B are defined *density-connected* iff there exists a core point from which both A and B are density-reachable.
- Finally, a cluster can be defined as a maximum set of points with respect to the notion of density-connectivity.

The critical issue to be faced when using DBSCAN is the hyperparameters setting. This was accomplished following the method proposed by [1] based on the sorted k-th neighbour distance graph. It is generated computing the distance to the k-th neighbour distance $dist_k$ for all the points in the dataset and plotting it versus the data-points sorted in descending order of their k-dist values. A sudden change of slope is expected when the transition between outliers and points belonging to a cluster takes place. Thus, the k-distance value at which this transition is observed can be picked as `Eps` parameter for DBSCAN where `NumPoints` is set to k. The choice of `NumPoints` must be performed guaranteeing an efficient detection of both outliers and clusters. If a too large value of `NumPoints` is considered, DBSCAN can fail to recognise some cluster points, thus labelling them as outliers. On the other hand, an excessively low `NumPoints` leads too groups of outliers erroneously detected as a cluster. As a consequence, the number of detected clusters would be much higher. than the number of the really existing ones. This trade-off can be seen in table II. The quality of the resulting clustering was determined using the Silhouette score [5] that ranges between -1 and 1. The higher Silhouette is, the nearer each point is to the ones in the same cluster, so a high score is desirable. DBSCAN was applied using the library Scikit [3].

[1]In this context the terms "points" and "data" are interchangeable considering that a point is the representation of a datum (a row of X_{PCA}) in the space of dimensions D^*.

[2]The Euclidean norm has been used as concept of distance.

`NumPoints`	`Eps`	Silhouette	Found clusters
2	0.1	−0.156	42
3	0.15	0.116	14
4	0.15	0.253	10
5	0.20	0.266	8
6	0.20	0.277	6
7	0.20	0.277	6
8	0.21	0.277	6
12	0.27	0.277	6
30	0.35	0.254	5

TABLE II: For each choice of `NumPoints`, the parameter `Eps` is chosen visually with respect to the k-distance sorted graph. It can be seen that the Silhouette metric reaches the maximum for the couples for which DBSCAN detects 6 clusters. Hence, this can be considered the number of existing meaningful clusters. It can be noticed that for too low values of `NumPoints` too many clusters are detected, i.e. groups of outliers are considered as significant groups. On the opposite, for the extreme value `NumPoints` = 30 DBSCAN fails in classifying two close clusters as separate.

D. Comments on the expanded features model

Rank estimation, dimensions reduction, DBSCAN and Silhouette validation have been performed on the degree 2 polynomially expanded matrix. DBSCAN, run on the truncated matrix to the estimated rank equal to 5, detects 6 clusters when `NumPoints` = 13 and `Eps` = 1.59 are chosen. As in the not-expanded case, the parameters were set looking at the knee of the k-distance graph. The Silhouette score revealed to be higher but, as Section III explains, the results interpretation becomes less clear.

III. Results and Discussion

Figure 2 shows six distinct clusters. It is known that the $\Delta G(\text{Rxn } A)$ free energy is mainly determined by the metal present in the molecule and fine tuned by the ligands [2]. The 2D PCA projection shows that the energies are ordered along the y direction inside each cluster ("rainbow effect").

In addition with figure 3, it is possible to say that these clusters are characterised by ligands belonging to the same family (see supplemental information of [2]).

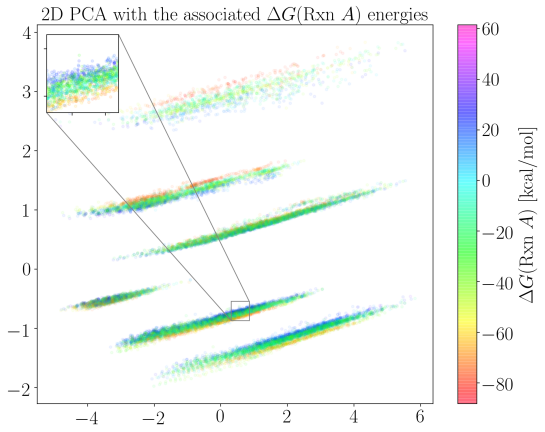

Fig. 2: This figure shows the 2-dimensional PCA projection of the molecule data. The data points are coloured by the ΔG(Rxn A) energies described in [6]. Row shuffling on X and a low opacity level reveal energy gradient directions in each cluster.

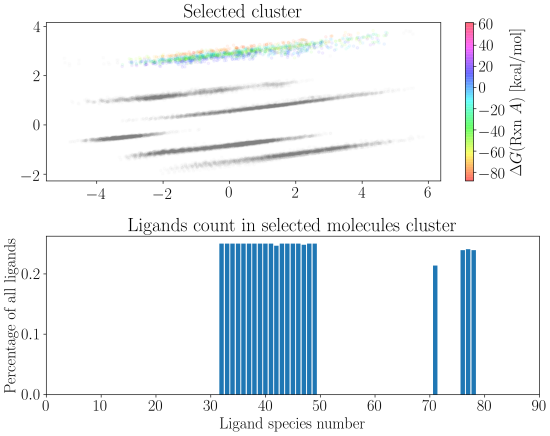

Fig. 3: As it is done here for the upmost cluster, a count of the ligands present in each cluster reveals a very clear separation of the clusters by the ligand families. (See appendix for the other clusters.)

Figure 4 allows to further understand how the points have been clustered. Indeed the points have been coloured by their molecular weight and it is immediately clear that the x-axis is linked to the molecular weight (the colours are coherent between all clusters). When doing an additional plot and colouring the points by molecular volume (which is physically linked to molecular weight), the same kind of plot appears. It thus confirms the supposition.

One thing that remains to be understood is what the clusters represent. The general explanation is that each cluster contains the molecules that have a certain combination of the P, N, C and O atoms as ligands. When colour-coding these combinations for each point in the plot with the values of the last four features in the data set, each of the clusters contains a unique uniform colour, thus supporting this explanation. A further confirm that the cluster splitting is determined by the ligand families is found when the same steps are applied

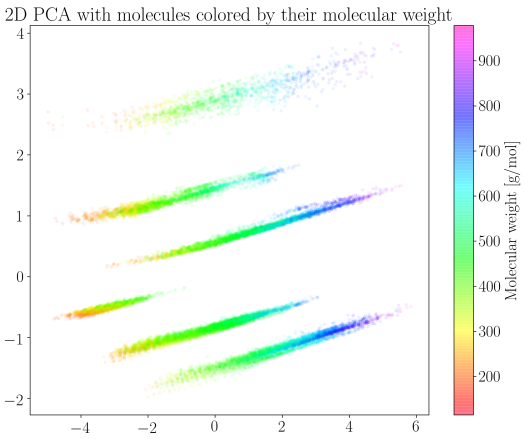

Fig. 4: This figure is similar to figure 2 except that the points have been coloured by their molecular weight. Again, it reveals a clear gradient of the characteristic.

to a reduced dataset containing only one of the 6 metals. The same clusters are observed and the ligands hystograms match the ones obtained from the full dataset. In addition, observing the energy coloured graph, it can be noticed that only a narrow window of the energies range is present. This fact could have been supposed remembering that ΔG(Rxn A) is strongly dependent to the considered metal.

When the analogous plots (Figures 2, 3, 4) for the polynomially expanded matrix are considered, the same interpretations seem to be no longer valid. Most of the molecules fall into the same cluster where almost all the ligands are represented. The other detected clusters are by far less populated and they contain high percentages of particularly small ligands (labelled 68, 69, 81 in [2]). As far as the molecular weight dependence is concerned, an increasing trend along the second principal component (y-axis) can be observed. Finally, the energies' "rainbow effect" becomes less visible. Apart from these minor points, there is no evidence to speculate about the causes of the clusters splitting. This is, probably, also due to the fact that DBSCAN does not manage to recognise different groups in the cluster hosting the larger amount of points. A better DBSCAN parameters tuning and finding a chemical reason behind the clusters separation are open issues for this case.

IV. Conclusion

All the plots present in this report are really clean (no noise, precise colouring). This is because the data provided are not some statistical measured samples but well known features of the molecules that have either been found by expensive computational methods or simple observation (e.g. number of carbon atoms). Thus, the machine learning techniques used here managed to show in such a clear manner the underlying structure of the data.

V. Acknowledgments

We would like to express our appreciation to the the Laboratory for Computational Molecular Design that hosted us, in particular Alberto Fabrizio and Benjamin Meyer for their support.

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters 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, pages 226–231. AAAI Press, 1996.

[2] Benjamin Meyer, Boodsarin Sawatlon, Stefan Niklaus Heinen, Anatole von Lilienfeld, and Clémence Corminboeuf. Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts. *Chemical Science*, 9, 07 2018.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] Patrick O. Perry. Cross-Validation for Unsupervised Learning. *arXiv e-prints*, September 2009.

[5] Peter Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987.

[6] Matthew D. Wodrich, Boodsarin Sawatlon, Michael Busch, and Clémence Corminboeuf. On the generality of molecular volcano plots. *ChemCatChem*, 10(7):1586–1591, 2018.