

A Graph-Based Clustering Analysis of the QM9 Dataset via SMILES Descriptors

Gabriel A. Pinheiro[1(✉)], Juarez L. F. Da Silva[2], Marinalva D. Soares[3], and Marcos G. Quiles[3]

[1] Associate Laboratory for Computing and Applied Mathematics, National Institute for Space Research, PO BOX 515, 12227-010 São José dos Campos, SP, Brazil
gabriel.pinheiro@inpe.br

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

[3] Institute of Science and Technology, Federal University of São Paulo, 12247-014 São José dos Campos, SP, Brazil
mdiasoares@gmail.com, quiles@unifesp.br

Abstract. Machine learning has become a new hot-topic in Materials Sciences. For instance, several approaches from unsupervised and supervised learning have been applied as surrogate models to study the properties of several classes of materials. Here, we investigate, from a graph-based clustering perspective, the Quantum QM9 dataset. This dataset is one of the most used datasets in this scenario. Our investigation is two-fold: 1) understand whether the QM9 samples are organized in clusters, and 2) if the clustering structure might provide us with some insights regarding anomalous molecules, or molecules that jeopardize the accuracy of supervised property prediction methods. Our results show that the QM9 is indeed structured into clusters. These clusters, for instance, might suggest better approaches for splitting the dataset when using cross-correlation approaches in supervised learning. However, regarding our second question, our finds indicate that the clustering structure, obtained via Simplified Molecular Input Line Entry System (SMILES) representation, cannot be used to filter anomalous samples in property prediction. Thus, further investigation regarding this limitation should be conducted in future research.

Keywords: Clustering · Graph · Quantum-chemistry

1 Introduction

Machine learning methods can be roughly divided into two main paradigms: supervised and unsupervised [11,17]. Methods belonging to the supervised paradigm demands a dataset in which all samples are labeled, and the objective is the map the input features that describe each sample to its respective label. If the labels are discrete, we have a classification problem; if labels are continuous,

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12249, pp. 421–433, 2020.
https://doi.org/10.1007/978-3-030-58799-4_31

we have a regression problem [11]. On the other hand, unsupervised methods do not require labels. In this scenario, the method tries to find hidden patterns using only the available features. For instance, these methods can be used for clustering the dataset into a set of homogeneous groups; select which features are the most important to represent the samples, or even to perform anomaly detection [5].

When dealing with classification or regression problems, unsupervised methods can be adopted as a preprocessing phase [13]. These methods can be used to filter, learn new representation, and also for grouping samples into clusters. Thus, revealing how data samples are distributed, or clustered, into the feature's space. There are several clustering methods in the literature [14]. Among them, graph-based methods have some compelling features. For example, graph-based methods rely on how data samples are drawn. Samples are represented as nodes in the graph and the edges, or links, represent the similarity between samples (nodes). A graph-based method has some advantages in comparison to other clustering methods, such as the K-means [14]. For instance, the graph might represent non-trivial structures in the feature space that are beyond Gaussian structures commonly detected by the K-means [6,8,23]. Moreover, the relation between patterns, such as paths, are also easily obtained by analyzing the graph structure [23].

Here, we apply a graph-based clustering method for analyzing the QM9 dataset [20], which is one of the most used datasets in machine learning applied to materials science. The QM9 dataset provides there basic representation for each molecule: atom positions represented by the X, Y, Z coordinates (XYZ representation); the International Chemical Identifier, or InChI [20]; and the Simplified Molecular Input Line Entry System representations, or SMILES [22], which represent atoms and bonds using a string with a simple grammar. In contrast to the XYZ representation, the SMILES format does not require an optimization of the molecular structure via Density Functional Theory (DFT) or Molecular Dynamics [7,15,16] or a complete knowledge of the experimental structure, and hence, the SMILES representation has a smaller computational cost to be obtained. Here, we take the SMILES format into account to encode the molecules into a feature vector (see Sect. 2.1).

In this work, by conducting this investigation of the clustering structure of the QM9, we expect to find some insights about the QM9 molecules. Specifically, this study aims to answer the following questions:

1. Are the samples in the QM9 organized in clusters?
2. Regarding the property prediction problem, is it possible to identify molecules that jeopardize the accuracy of predictors (anomalous molecules)?

Our results, as depicted in Sect. 3, show that the molecules in the QM9 dataset are indeed organized into clusters. Thus, this structure might be used to set the cross-validation structure when dealing with supervised property predictions. However, regarding our second question, our results have not highlighted any insight on how to isolate anomalous molecules using the cluster structure

of the data. Consequently, further investigation must be conducted to evaluate how to filter these molecules that expose the property prediction process.

The remain of this paper is organized as follows. Section 2 introduces the QM9 dataset and the methods taken into account in this research. Results and discussion are depicted in Sect. 3. Finally, some concluding remarks are drawn in Sect. 4.

2 Data and Methods

2.1 The QM9 Dataset

The quantum-chemistry data used in this work came from the QM9 quantum-chemistry data, a public dataset composed of ∼134-kilo organic molecules made up of C, H, O, N, and F with up to twenty-nine atoms. This dataset provides 15 properties (geometric, energetic, electronic and thermodynamic) and geometries optimized via the DFT-B3LYP/6-31G($2df, p$) level of theory, as well as the SMILES, XYZ, and InChI representations for every molecule therein [20]. We used the QM9 previously processed by [10], which removed ∼3k molecules that failed in the InChI consistent check [20] and reports the atomization energies of the total energies properties from the original QM9. We also eliminated 612 molecules that the RDKit[1]. Python package was not able to read. Thus, after withdrawn all these samples from the QM9, the dataset was reduced to 130 217 molecules.

2.2 Feature Extraction

Several features can be extract from each molecule. In SMILES, each molecule is represented by a string containing its atoms (atomic symbols) and bonds, hydrogen atoms may be omitted in this representation [21]. In this sense, a Propane molecule can be whether of these string CC#C or C#CC, where # is a triple bond. Thus, without an explicit loading order the algorithm can yield various strings for a particular molecule, as shown in the Propane example. Thus, a general approach to overcome this problem is the use of canonical SMILES, which guarantees uniqueness representation. Nevertheless, recent work with SMILES has explored the noncanonical SMILES as data augmentation of neural network models [1,2].

Herein, we utilize the RDKit to read the SMILES derived from the relaxed geometry and the Mordred[2] Python package for feature extraction. Mordred is open-source software with an extensive molecular descriptor calculator for cheminformatics. It computes up to 1826 topological and geometric descriptors, including those available in the RDkit [18]. As the package has integration with RDKit, it was required to convert the molecule in an RDkit molecule object to perform the descriptor extraction. Moreover, since RDKit can understand

[1] https://www.rdkit.org/.
[2] https://pypi.org/project/mordred/.

noncanonical and canonical SMILES as the same molecule, the canonization of the SMILES was not required. In this work, we focus only on the topological features, which are SMILES-based descriptors, covering modules as diverse as autocorrelation, atom count, Wiener index, and others.

After extracting all features of the QM9 using the Mordred package, we eliminate three subset of features: 3D descriptors (213 features), features with variance equal to zero (539 features), and features with missing value (194 features). Thus, our final dataset consists of 130217 samples with 880 features.

2.3 Graph Generation

In a graph-based method, the first step consists of generating a network (or graph) $G = (V, E)$ for a given data set [4]. The input is represented as $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, in which n represents the number of samples, and x_i is the feature vector of sample i. The dataset might also provide a set of labels $\mathcal{L} = \{y_1, y_2, ...y_n\}$. When dealing with unsupervised learning, only $\mathcal{X}$ is taken into account.

To generate a graph, we define a set of nodes $V = \{v_1, v_2, ..., v_n\}$, which represent the data samples in $\mathcal{X}$, and a set of edges E representing the similarity (or relation) between pair of samples. There are several manners to define the set of edges, e.g. the $\epsilon-$cut and the $k-$NN [23]. Here we considered the $\epsilon-$cut method. Each edge $e_{ij} \in E$, which represent the relation between samples i and j is defined as follows:

$$e_{ij} = \begin{cases} 1 \text{ if } d(x_i, x_j) \leq \theta \text{ and } i \neq j \\ 0 \text{ if } d(x_i, x_j) > \theta \text{ or } i = j \end{cases} \tag{1}$$

where θ is a connection threshold that defines the minimum similarity to create the connection between samples i and j; and $d(x_i, x_j)$ is a distance function. Here we assume $d(x_i, x_j)$ to be the Euclidean distance:

$$d(x_i, x_j) = \|x_i - x_j\|. \tag{2}$$

Once the graph G is generated, the analysis of the dataset might proceed.

3 Results and Discussion

In this section we present a summary of all computer experiments conducted in this research.

3.1 Feature Space Reduction

As defined in Sect. 2.3, each pair of samples is connected by an edge whether the distance between their vector representations, x_i and x_j is shorter than a threshold θ. To compute this distance between a pair of molecules, we have taken two vector representations into account. First, we assume that each molecule is

represented by all available features (880 dimensions). Second, due to the high correlation observed between several pairs of features, we perform the Principal Component Analysis (PCA) and extract the major components to represent the samples.

To evaluate these two representations, we select a subset of 5k random molecules of the QM9 and run several experiments. We found that the two approaches led to quantitatively similar results. Thus, we decided to continue our experiments using only the second approach, which consists in computing the PCA of the data. This selection is three-fold:

1. Accurate representation: with 134 components (out of 880 features), we represent 99 % of the variance of the original data representation;
2. PCA is fast to compute;
3. The generation of the graph demands the computation of $O(n^2)$ pair distances; thus, by reducing the vector representation from 880 to 134 we greatly reduce the computational cost of the process.

3.2 Cluster Structure of the QM9

With each molecule represented by its 134 principal components, we built several graphs varying the threshold. We aim to generate a graph which is sparse albeit connected. Thus, we seek for a threshold that delivers a graph with a lower number of components and with a low average degree ($\langle k \rangle$). If a large number of components are observed, the graph might be too disconnected and an evaluation of the relational structure between molecules is not allowed. Conversely, if the number of components is low, it leads to a very high ($\langle k \rangle$) creating a densely connected graph. In this scenario, the evaluation is also compromised.

To face this trade-off, we run experiments with several thresholds. Table 1 shows some results regarding the generated graph for a given threshold (θ). Specifically, we depict the number of connected components (#-Comp); size of the largest component ($L-$Comp); number of communities (#-Comm); size of the largest community (L-Comm); number of singletons (#-Singletons); average degree ($\langle k \rangle$). The communities were detected with the Louvain Algorithm [3], albeit similar results were found with other community detection algorithms [19]. It is worth noting that community and graph cluster represents, in the context of this study, the same type of structure: a group of densely connected nodes.

Evaluating the results in Table 1, we find that $\theta = 15$ provides a proper threshold, the number of components is low and $\langle k \rangle \approx 51$, which represents a sparse graph. Moreover, the number of singletons is also small meaning that only a set of 1326 molecules are completely isolated from all the components of the graph. With this graph we can perform a deeper evaluation of the QM9 via the graph structure.

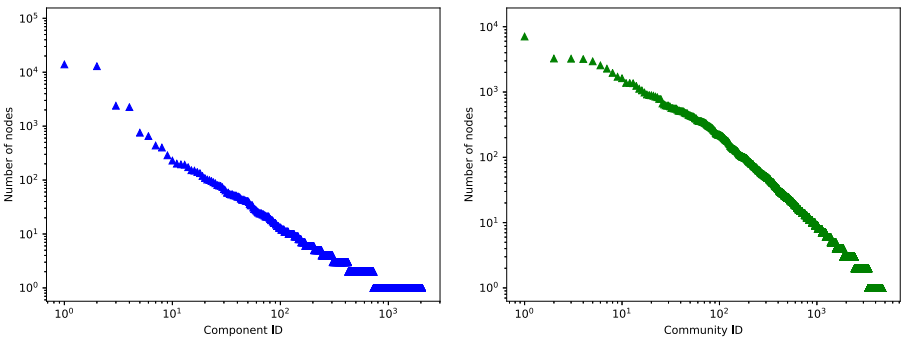
Figure 1 depicts the size of components (a) and communities (b). We can observed that there are a few components and few communities with a large number of nodes and several small components/communities. It represents that the dataset has indeed a cluster structure with a few large homogeneous regions

Table 1. Graph measurements. Number of connected components (#-comp); size of the largest component ($L-$Comp); number of communities (#-Comm); size of the largest community (L-Comm); Singletons; average degree ($\langle k \rangle$) for each graph generating according to the threshold θ.

Threshold (θ)	#-Comp	$L-$Comp	#-Comm	$L-$Comm	#-Singletons	$\langle k \rangle$
1	130009	3	130009	3	129977	≈0
5	114618	16	114903	10	116952	0.27
10	42203	1539	52655	189	58809	3.95
15	2055	87922	4697	12105	1326	51.1
20	264	125957	467	30470	180	734.2
25	59	129939	86	78563	28	6104.0

and many sparse structures composed by a small number of molecules. This cluster structure might be used to address the cross-correlation technique used in supervised machine learning processes, such as in molecular property prediction. In this case, instead of randomly splitting the data set into n folds, the clusters could be taken into account to deliver this division. Thus, we can guarantee that samples from each cluster are represented in the folds, which is not guaranteed when traditional cross-validation algorithm is employed.

Interestingly, in the log-log plot shown in Fig. 1, both curves seem to be fitted by a straight line and might indicate a power-law distribution regarding the size of the components and communities. This type of distribution is widely observed in real graphs (networks) from biology to technological graphs [9].



(a) (b)

Fig. 1. Connected components and communities of the graph build with $\theta = 15$. (a) Components; (b) Communities.

Out of 4697 clusters, only 182 represent clusters with more than 100 samples. Moreover, these 182 communities sum up 98 096 molecules of a total of 130 218.

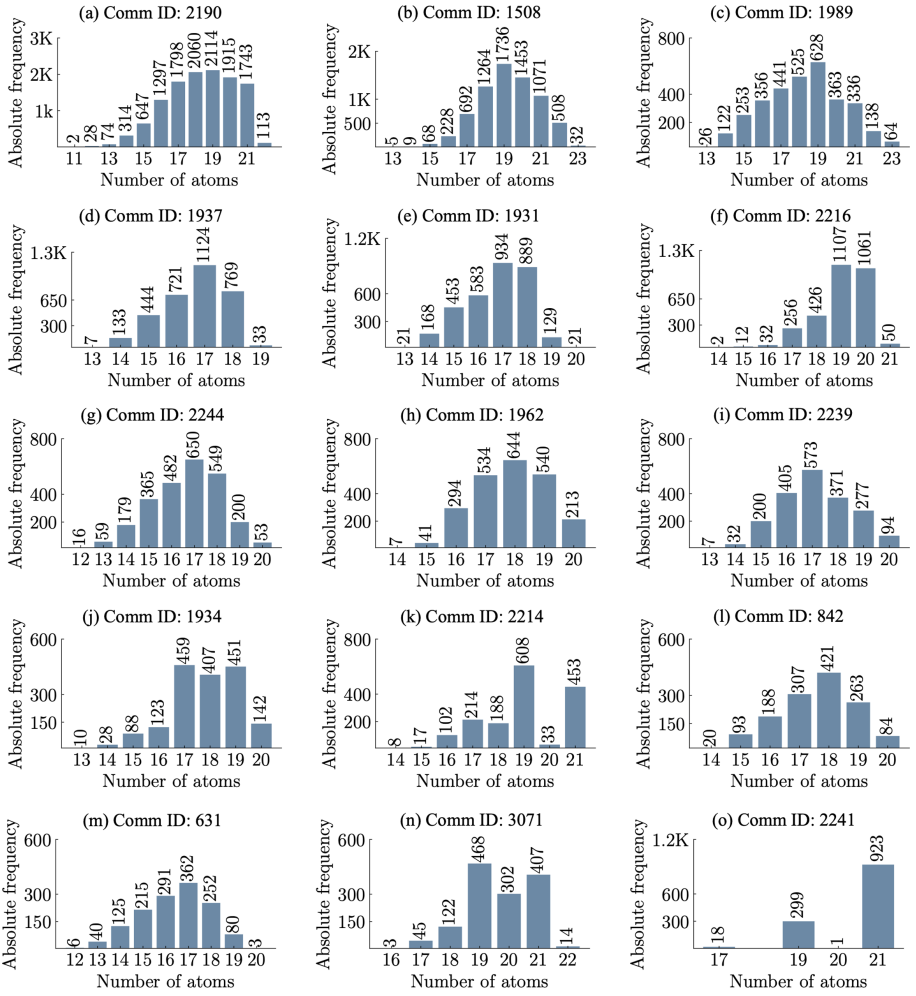


Fig. 2. Histogram of the molecule size grouped into the largest fifteen communities.

The remaining samples are clustered into sets with only a small number of samples. Figures 2 and 3 show histograms of molecule sizes and chemical species of the largest clusters. One could observe that the distribution of molecules regarding their sizes and chemical species is not completely homogeneous, which indicates that each cluster groups molecules with distinct features. For instance, by analyzing the clusters (IDs) 1937 and 2216, depicted in Figs. 2(d) and (f), and 3(d) and (f), we see that the number of atoms and the distribution of chemical species are different between clusters. The same outcome holds for the other clusters. Moreover, the larger the cluster is, the more heterogeneous is its structure. Small clusters contain more similar molecules regarding their sizes and chemical species.

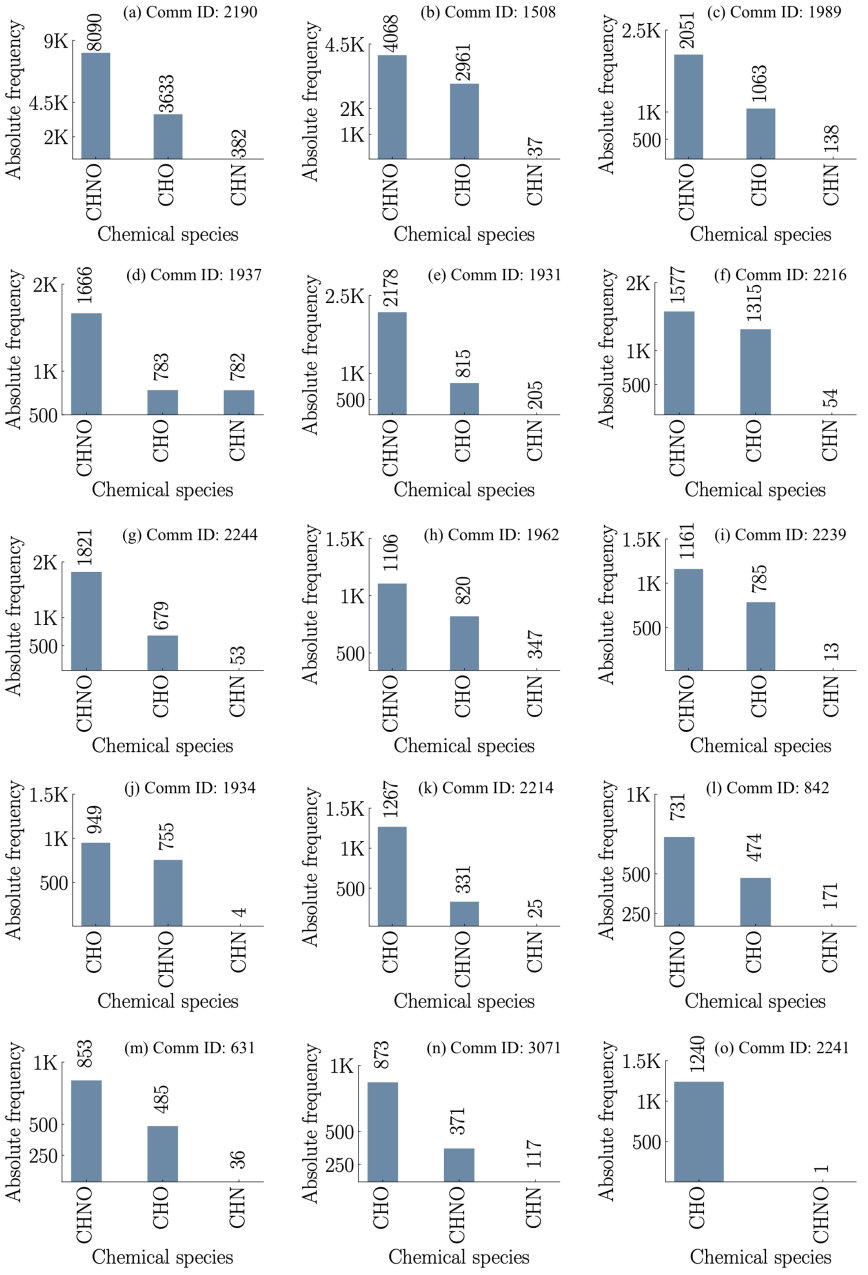


Fig. 3. Histogram of the chemical species grouped into the fifteen largest communities.

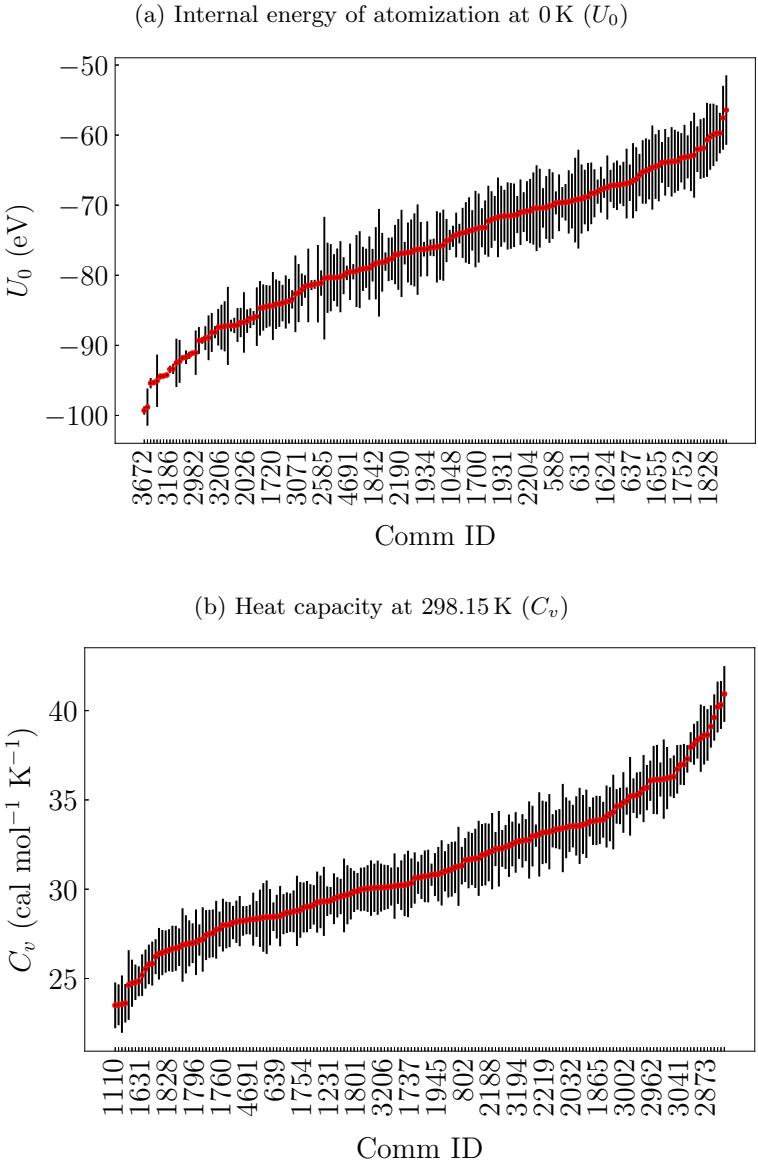


Fig. 4. Internal energy (U_0) and heat capacity C_v in each cluster. Red marker and black line represent the mean and standard deviation of the properties in each cluster, respectively. Only clusters with more than 100 samples are shown. Nevertheless, we labeled only 23 clusters IDs on the x-axis. (Color figure online)

To further investigate the cluster structures of the data, in Fig. 4, we show the mean and the standard deviation of the heat capacity at 298.15 K (C_v), and the internal energy of atomization at 0 K (U_0) for each cluster. To facilitate the visual scrutiny of the figures, we present only clusters with more than 100 samples.

For both properties, we can notice that each cluster is assigned to a specific mean value of C_v and U_0 and a lower standard deviation, which means that the clusters, indeed, group similar molecules in the same fashion as observed for molecule sizes and chemical species.

The cluster structure of the network and its relation to the property prediction problem is discussed in the following section.

3.3 Community Evaluation and Anomalous Molecules Identification

To identify the anomalous molecules regarding property prediction, we train Multilayer Perceptron neural networks (MLP) [12] over the QM9 for several available properties. After the learning phase, we evaluate each molecule of the dataset and annotate their errors. Using a cross-correlation approach [17], we select a subset of 1 % of molecules that presented the largest errors. Thus, our task here was to try to identify those molecules, named anomalous, using the clusters structure described in the last section.

Figure 5 illustrates a graph in which nodes represent communities. In this coarse-grained visualization of the whole network we can see how the communities (or clusters) are related to each other. Size of the nodes indicates the size of the communities, color represents the proportion of anomalous molecules clustered in the community. Nodes in gray color indicate that about 1 % of the samples in the community introduces a high error in the predictor. This amount corresponds to the same proportion of anomalous molecules labeled in the QM9. Thus, this outcome opposes our initial hypotheses that anomalous samples could be grouped into specific clusters (groups of high error molecules). Unfortunately, based only on our clustering analysis with SMILES descriptors, we are not able to highlight these anomalous samples and remove them from the QM9 dataset.

In Fig. 5, communities in purple color represents homogeneous communities consisting of only molecules with low error. Although it might suggest a pattern of very accurate regions of the feature space, these communities are quite small and represent just a small fraction of the molecules of the dataset.

By analyzing each community isolated, we observe that most of the small communities (size between 2 and 100), are quite homogeneous regarding the properties of their molecules. Besides, from a total of 3193 communities in this range, only 213 communities contain molecules with high error prediction (anomalous). On the other hand, when considering larger communities (with more than 100 samples), the number of communities containing anomalous molecules is 95 out of 175. This result is sound assuming that the larger the community is, the more heterogeneous is its structure. Regarding the communities holding a single node, which represents a total of 1326 communities, a set of 53 singletons represent anomalous molecules that should be further scrutinized.

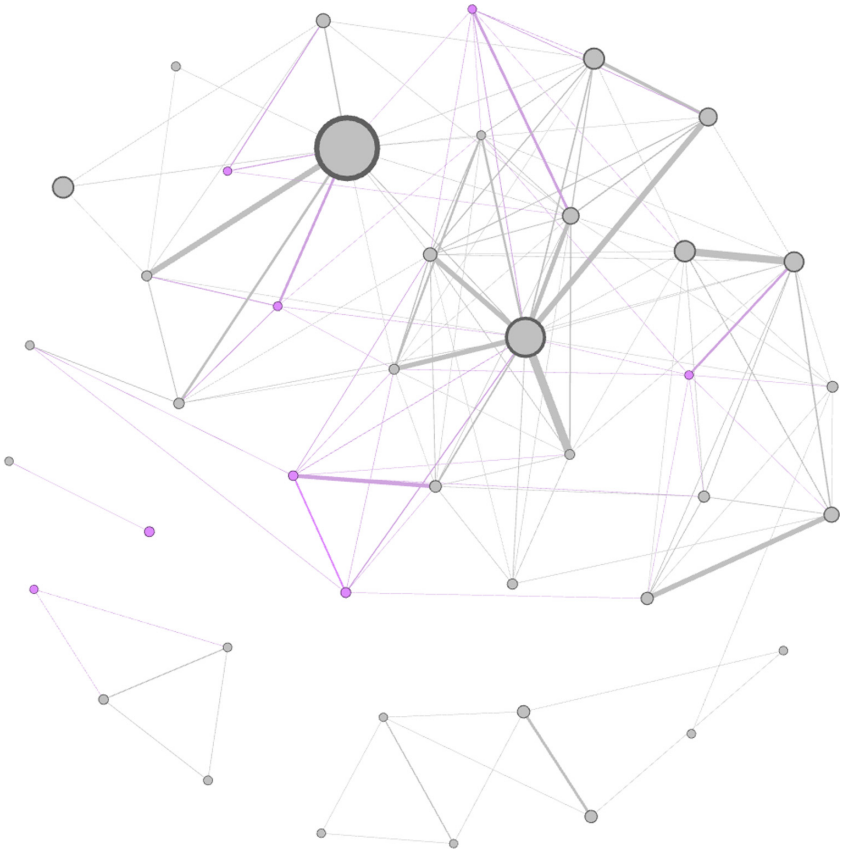


Fig. 5. Graph of the community (cluster) structure. Each node represents a community. Size of the circle informs the size of the community (number of clustered molecules). Gray color point communities in which about 1% of molecules are anomalous. Purple color represents communities with absence of anomalous molecules. (Color figure online)

4 Conclusions

In this work, we investigate, from a graph-based clustering perspective, the Quantum-chemistry QM9 dataset. This study was two-fold: 1) investigate if the QM9 is organized in clusters; and 2) whether high-error molecules, regarding the property prediction problem, could be isolated into specific clusters. It would allow us to remove these noise samples from the dataset.

First, we have mapped the QM9 molecules into a graph by encoding each molecule as a feature vector representing descriptors computed from the SMILES representation. Next, we scrutinize the cluster structure of the graph considering several thresholds and by applying the Louvain Community detection algorithm. Once a suitable graph structure was obtained, several analyses were conducted,

and the results indicated that: 1) the QM9 is composed of several clusters, in which the molecules are grouped according to their properties; and 2) by using this clustered structure we cannot isolate anomalous molecules that reduce the accuracy supervised models, such as the MLP network used for property prediction.

In the next step, we will evaluate other approaches for graph construction and feature representation for the molecules of the QM9 dataset. In particular, our future research will focus on how to detect and isolate anomalous molecules from a dataset composed of molecules. This task is essential to build accurate property prediction models that can be employed as surrogate models in material analysis.

Acknowledgments. The authors gratefully acknowledge support from FAPESP (São Paulo Research Foundation, Grant Numbers 2016/23642-6, 2017/11631-2, and 2018/21401-7), Shell 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. We also thank Johnatan Mucelini for the support with the quantum QM9 dataset.

References

1. Arús-Pous, J., et al.: Randomized smiles strings improve the quality of molecular generative models. J. Cheminformatics **11**(1), 1–13 (2019)
2. Bjerrum, E.J.: Smiles enumeration as data augmentation for neural network modeling of molecules. arXiv preprint arXiv:1703.07076 (2017)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theo. Exp. **2008**(10), P10008 (2008). https://doi.org/10.1088/1742-5468/2008/10/p10008
4. Breve, F.A., Zhao, L., Quiles, M.G.: Particle competition and cooperation for semi-supervised learning with label noise. Neurocomputing **160**, 63–72 (2015). https://doi.org/10.1016/j.neucom.2014.08.082
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3) (2009). https://doi.org/10.1145/1541880.1541882
6. Chen, C., Ye, W., Zuo, Y., Zheng, C., Ong, S.P.: Graph networks as a universal machine learning framework for molecules and crystals. Chem. Mater. **31**(9), 3564–3572 (2019). https://doi.org/10.1021/acs.chemmater.9b01294
7. Chmiela, S., Sauceda, H.E., Müller, K.R., Tkatchenko, A.: Towards exact molecular dynamics simulations with machine-learned force fields. Nat. Commun. **9**(1), 3887 (2018)
8. Cook, D.J., Holder, L.B.: Graph-based data mining. IEEE Intell. Syst. Appl. **15**(2), 32–41 (2000). https://doi.org/10.1109/5254.850825
9. Costa, L.D.F., et al.: Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Adv. Phys. **60**, 329–412 (2011)
10. Faber, F.A., et al.: Prediction errors of molecular machine learning models lower than hybrid DFT error. J. Chem. Theory Comput. **13**(11), 5255–5264 (2017)
11. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7

12. Haykin, S.S.: Neural Networks and Learning Machines, 3rd edn. Pearson Education, Upper Saddle River (2009)
13. Hinton, G.E., Sejnowski, T.J., Hughes, H.: Unsupervised Learning: Foundations of Neural Computation. MIT Press, Cambridge (1999)
14. 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
15. Karplus, M., McCammon, J.A.: Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. **9**(9), 646–652 (2002). https://doi.org/10.1038/nsb0902-646
16. Mardirossian, N., Head-Gordon, M.: Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. Mol. Phys. **115**(19), 2315–2372 (2017). https://doi.org/10.1080/00268976.2017.1333644
17. Mitchell, T.M., Thomas, M.: Machine Learning. McGraw-Hill, Maidenhead (1997)
18. Moriwaki, H., Tian, Y.S., Kawashita, N., Takagi, T.: Mordred: a molecular descriptor calculator. J. Cheminformatics **10**(1), 4 (2018)
19. Quiles, M.G., Macau, E.E.N., Rubido, N.: Dynamical detection of network communities. Sci. Rep. **6**(1), 25570 (2016)
20. Ramakrishnan, R., Dral, P.O., Rupp, M., Von Lilienfeld, O.A.: Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data **1**, 140022 (2014)
21. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. **28**(1), 31–36 (1988)
22. Wu, Z., et al.: MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. **9**(2), 513–530 (2018)
23. Zhu, X.: Semi-supervised learning with graphs. Ph.D. thesis, School of Computer Science, Carnegie Mellon University (2005)