UNIVERSIDADE FEDERAL DE SÃO PAULO

BACHARELADO EM CIÊNCIA E TECNOLOGIA

CARLOS GUILHERME MORAES

Análise de agrupamentos em bases de compostos químicos: Detecção de moléculas anômalas

São José dos Campos, SP

Agosto de 2021

CARLOS GUILHERME MORAES

Análise de agrupamentos em bases de compostos químicos: Detecção de moléculas anômalas

Relatório de conclusão referente à pesquisa feita durante a bolsa de Iniciação Tecnológica PIBIT/CNPq.

Orientador: Prof. Dr. Márcio Porto Basgalupp

Co-orientador: Prof. Dr. Marcos Gonçalves Quiles

Universidade Federal de São Paulo - UNIFESP

Instituto de Ciência e Tecnologia (ICT) - Campus São José dos Campos

São José dos Campos, SP

Agosto de 2021

Resumo

A Aprendizagem de máquina ou aprendizagem automática, é um ramo da inteligência artificial capaz de aprender com dados, reconhecer padrões, tomar decisões. Suas aplicações estão presentes em diversas áreas, como por exemplo em ciências de materiais. Neste cenário, o aprendizado de máquina tem contribuído com a solução de diversos problemas, dentre eles, a predição de propriedades moleculares. Treinar um modelo para realizar a predição de propriedades não é uma tarefa fácil e envolve a aquisição de um bom conjunto de dados. Contudo, em muitas ocasiões, a base de dados pode conter moléculas específicas que prejudicam a acurácia do modelo gerado, denominadas moléculas anômalas. Neste projeto, investigamos, a partir de agrupamentos de dados em grafo, o conjunto de dados de moléculas orgânicas QM9 com o objetivo de identificar tais moléculas.

Palavras-chaves: aprendizado não-supervisionado; ciências de materiais; moléculas orgânicas.

Lista de Ilustrações

Sumário

1 Introdução

O aprendizado de máquina pode ser dividido em dois grandes paradigmas: aprendizado supervisionado e aprendizado não-supervisionado. Métodos pertencentes ao paradigma supervisionado requer, como entrada, um conjunto de dados no qual todos os exemplos estejam rotulados. O objetivo principal de tais métodos está no mapeamento das entradas (vetor de características) nos seus respectivos rótulos. Se o rótulo for discreto, temos um problema de classificação; se o rótulo for um dado contínuo, temos um problema de regressão [1].

Por outro lado, no aprendizado não-supervisionado, não há uma demanda obrigatória de rótulos. Nesse cenário, o método pertencente a este paradigma tem como objetivo encontrar padrões existentes nos dados (espaço de atributos). Muitas vezes usado para tarefas de agrupamentos em dados, o aprendizado não-supervisionado pode ser utilizado como pré-processamento de dados em problemas de classificação ou de regressão para realização de filtragem, aprendizado de novas representações e agrupamento dos exemplos [2] utilizando, por exemplo, métodos baseados em grafos nos quais se destacam por apresentar algumas vantagens.

O método em grafo é caracterizado pela forma como os dados são organizados visto que as amostras são representadas como nós no grafo e as arestas representam a semelhança entre as amostras. O método apresenta vantagens, pois, comparado ao K-*means*, por exemplo, e levando em conta a sua organização, é possível representar estruturas não-triviais no espaço de atributos que estão além das estruturas gaussianas comumente detectadas por algoritmos como o K-means [3-5].

O projeto em questão tem como objetivo aplicar métodos de agrupamento baseados em grafos para analisar dados de compostos químicos investigando um dos principais conjuntos químicos (QM9) utilizados pela ciência dos materiais na perspectiva do aprendizado de máquina.

O conjunto de dados QM9 fornece coordenadas em 3D de cada átomo da molécula (XYZ), representação InCHI [6], e o SMILES, *Simplified Molecular Input Line Entry System* [7], que representa estruturas químicas através de caracteres ASCII. E o projeto, neste caso, utiliza a representação SMILES para codificar moléculas em um vetor de características.

2 Objetivos

O conjunto utilizado QM9 é um conjunto público amplamente utilizado em pesquisas que envolvam aprendizado de máquina [7-13], são em torno de 134 mil moléculas orgânicas que o compõem que variam de 3 a 29 átomos compostos por C, H, O, N e F.

O conjunto, além de apresentar representações XYZ, InChI e SMILES também disponibiliza 15 propriedades físico-químicas e embora diversos trabalhos tenham empregado o QM9, algumas dessas características ainda não foram completamente exploradas. Por exemplo, em trabalhos envolvendo predição de propriedades é possível ver que diversas moléculas apresentam erro de predição bastante superior à média do conjunto. Além disso, a estrutura topológica do conjunto também não está descrita na literatura. Esta informação pode ser útil na construção de preditores e também na identificação de moléculas problemáticas.

Nesse contexto, este projeto de pesquisa tem como objetivo principal realizar uma investigação não supervisionada do QM9 a partir de uma perspectiva de agrupamentos em grafos. Especificamente, como resultados desse estudo, espera-se responder às seguintes questões:

1. Os exemplos do QM9 estão organizados em grupos (*clusters*)?

2. Moléculas problemáticas ou anômalas, que apresentam alto erro de predição, podem ser isoladas a partir dessa estrutura de agrupamentos?

3 Metodologia

3.1 Obtenção do QM9

O projeto inicia-se com a obtenção do conjunto QM9 no qual está disponível de forma pública onde diversas pesquisas que envolvem aprendizado de máquina podem obter acesso. O conjunto, como vimos, é composto por aproximadamente 134 mil moléculas orgânicas que variam de 3 a 29 átomos e são compostas por C, H, O, N e F. Além disso, também apresenta dados em relação à disposição do átomo em coordenadas em 3D, representações InChI, propriedades físico-químicas e o principal para esse projeto, os SMILES.

3.2 Extração dos Descritores

Com a obtenção do QM9, deseja-se então fazer a extração de descritores. A partir de uma dada molécula e de sua representação, diversos atributos (*features*/*descriptors*) podem ser extraídos. Nesse projeto, exploramos a representação SMILES como representação padrão de cada molécula. A adoção do SMILES é motivada pelo baixo custo computacional em comparação com outras representações, como a XYZ. No SMILES, cada molécula é representada por uma cadeia de símbolos contendo seus átomos (símbolos atômicos) e suas respectivas ligações; os átomos de hidrogênio podem ser omitidos nessa representação [14].

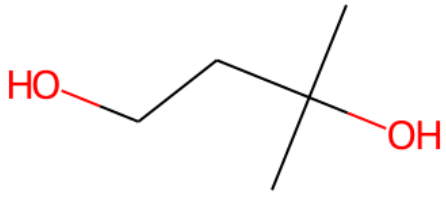

Figura 1: Fórmula estrutual obtida com o RDKit do SMILES: “CC(C)(O)CCO”.

Para a extração de descritores obtidos através do SMILES é utilizado o pacote Mordred disponível em linguagem Python no qual permite o cálculo de 1826 descritores topológicos e geométricos de moléculas [15]. O Mordred é utilizado juntamente com o pacote RDKit, também disponível em linguagem Python, no qual é utilizado para a leitura e manipulação dos dados químicos disponíveis.

3.3 Pré-processamento

O pacote utilizado Mordred, como vimos, disponibiliza um valor elevado de descritores, desta forma, é necessário um pré-processamento dos dados antes da geração da rede.

Para isso, é feita, primeiramente, a eliminação de atributos com dados faltantes ou que a variância seja igual a zero. Uma vez que seja encontrado um atributo ou dado faltante de uma coluna representado por algum descritor específico no dataframe, toda a coluna, ou seja, o descritor, é removida.

Em seguida, é feita a estandardização dos dados utilizando o pacote preprocessing, subpacote de Sklearn, no qual disponibiliza uma rotina, StandardScaler, capaz de realizar o processo resultando em dados no qual a média é igual a zero e o desvio padrão igual a 1.

Também é utilizado o PCA [16] que realiza a redução da dimensionalidade linear usando SVD dos dados para projetá-los em um espaço dimensional inferior [17].

E, por último, com os dados redimensionados, é obtido um novo conjunto de dados com as variâncias a fim de encontrar as quais somadas ultrapassam 95%, assim, separando-as dos dados redimensionados com o intuito de utilizá-las, então, para a criação da rede.

3.4 Criação da Rede

Após o tratamento e pré-processamento dos dados é possível, então, ser feita a geração da rede. A rede (grafo) é formada por um conjunto de vértices e um conjunto de arestas nos quais ligam os vértices, sua composição é possível ser comparada a um mapa rodoviário idealizado onde os vértices são cidades e as arestas são estradas de mão única. A entrada pode ser matematicamente representada como $X = \{x_1, x_2, \ldots, x_m\}$, no qual *m* representa o número de exemplos (número de nós) e $x_i \in R^n$ representa o vetor de atributos do exemplo *i* de dimensão *n*. O conjunto de dados também pode disponibilizar um conjunto de labels $L = \{y_1, y_2, \ldots, y_m\}$. Ao utilizar o aprendizado não-supervisionado, foco deste projeto, apenas o conjunto *X* é considerado.

Para gerar a rede, então, primeiro define-se um conjunto de nós $V = \{v_1, v_2, \ldots, v_m\}$, que representa as moléculas contidas no conjunto *X* redimensionado e o conjunto de ligações (arestas) *E* representa a similaridade entre os pares de moléculas. Existem diversas maneiras de se gerar o conjunto de arestas, para este projeto, é adotado a regra do corte (ϵ−cut). Seja $d_{ij} \in D$ onde *D é* o conjunto composto pela distância *Euclidiana* entre as moléculas cada aresta $a_{ij} \in E$, que representa a relação entre os exemplos *i* e *j* é definida pela seguinte regra:

$$a_{ij} = \begin{cases} 1 \; if \; e^{-\frac{d_{ij}}{max\{D\}}} > \epsilon \; and \; i \neq j \\ 0 \; if \; e^{-\frac{d_{ij}}{max\{D\}}} \leq \epsilon \; and \; i = j \end{cases} \quad (1)$$

no qual ϵ representa o limiar de conexão para criar a respectiva aresta entre as moléculas *i* e *j*. A distância *Euclidiana* mencionada é obtida da seguinte forma:

$$d_{ij} = ||x_i - x_j|| \quad (2)$$

A configuração do limiar ϵ é feita de forma empírica, pois não há uma maneira analítica de se obter um valor específico para ϵ onde se busca uma rede que seja conectada (poucos componentes) porém, esparsa (grau médio baixo).

3.5 Detecção de Moléculas Anômalas

Existem diversos algoritmos para realizar a análise de agrupamentos em grafos/redes. Nesse projeto usamos a rotina de seleção de comunidade denominada *greedy modularity communities* pertencente ao pacote Networkx presente na linguagem Python. Ele, a partir de uma rede já criada, obtém, de forma gulosa, a maximização da modularidade, medida capaz de avaliar a conectividade de partições de uma rede a partir, assim, do estudo de sua configuração, representada por *Clauset-Newman-Moore* [18].

Para identificar as moléculas anômalas em relação à predição de propriedades, são considerados resultados de predição de propriedades do QM9 já realizados em nosso grupo de pesquisa. Especificamente, são isoladas as moléculas que apresentam erro elevado (e.g. superior a dois desvios). Na sequência, investiga-se, então, a posição topológica de tais moléculas nas redes geradas a fim de conseguir correlacionar características topológicas dessas moléculas no grafo e, utilizando tais características, automatizar o processo de isolamento e exclusão de moléculas problemáticas.

5 Resultados e Discussões

Primeiramente, com a obtenção dos dados QM9, foi separado as moléculas nos quais apresentassem 19 átomos visto que a quantidade total de moléculas (134 mil) tornaria o tempo de execução muito alto. Filtrando apenas para moléculas com 19 átomos obteve-se 18336 moléculas, mantendo-se ainda um valor bastante elevado, com uma distribuição de átomos representados pela Figura 2.

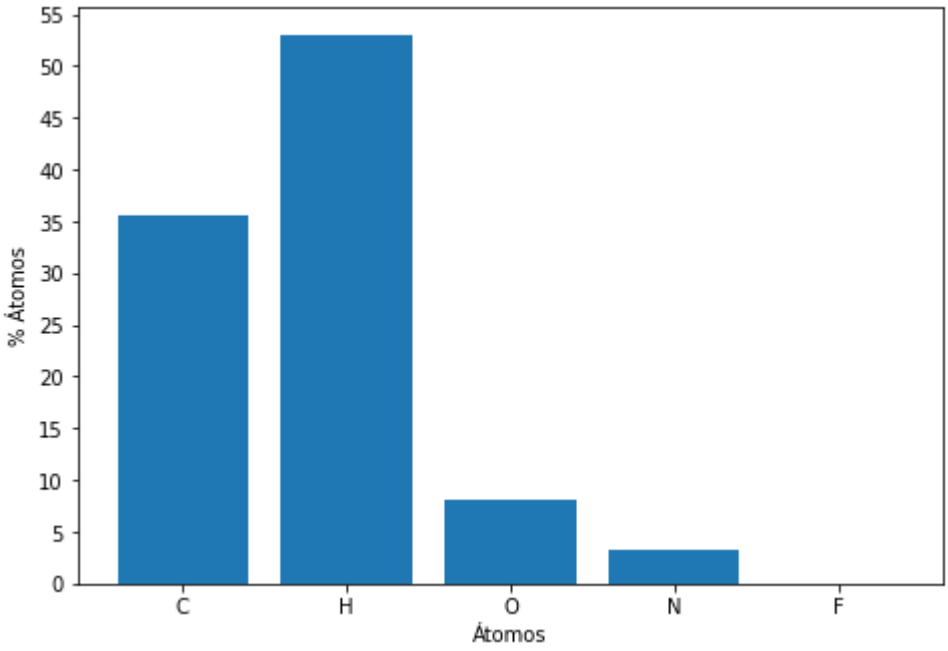

Figura 2: Distribuição dos átomos C, H, O, N, F em porcentagem.

Dos 348384 átomos, 18336 moléculas vezes 19 átomos, o mais abundante é o Hidrogênio com 53,08% seguido do Carbono com 35,64%, Oxigênio com 8,09%, Nitrogênio com 3,19% e o Flúor com somente 0,01%.

Com a obtenção dos descritores juntamente com o uso do Mordred, RDKit e a realização do pré-processamento, foi obtido, então, a matriz com a distância *Euclidiana* que posteriormente foi usada para a geração da rede. A matriz, com dimensões de

18336 x 18336, é uma matriz quadrada no qual apresentou uma média de 43,2 nas distâncias obtidas. Sua distribuição pode ser melhor vista pela Figura 3.

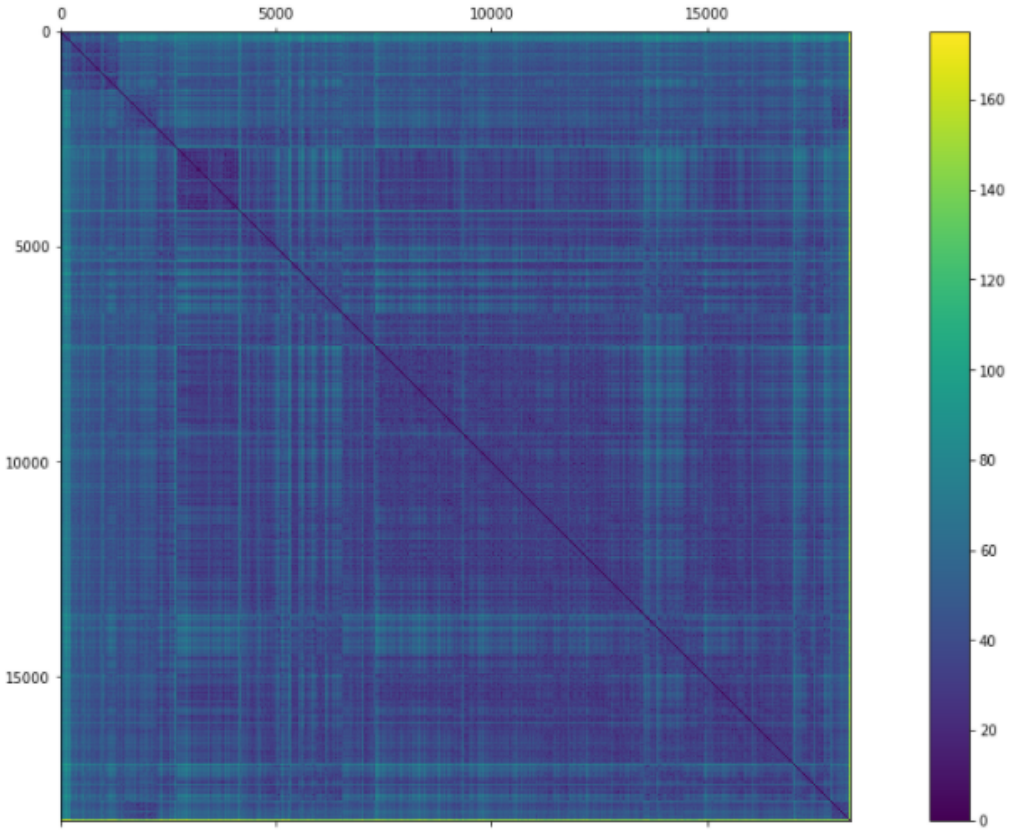

Figura 3: Distância *Euclidiana* entre as moláculas.

A Figura 3 apresenta uma linha com a cor referente ao valor zero pois na diagonal da matriz estão presentes a distância da molécula com ela mesma, desta forma, o valor será sempre zero.

Em seguida foi obtido a matriz de adjacência utilizando o limiar, como já visto, com valor de aproximadamente 0,79 e o resultado da matriz pode ser visto representado pela Figura 4.

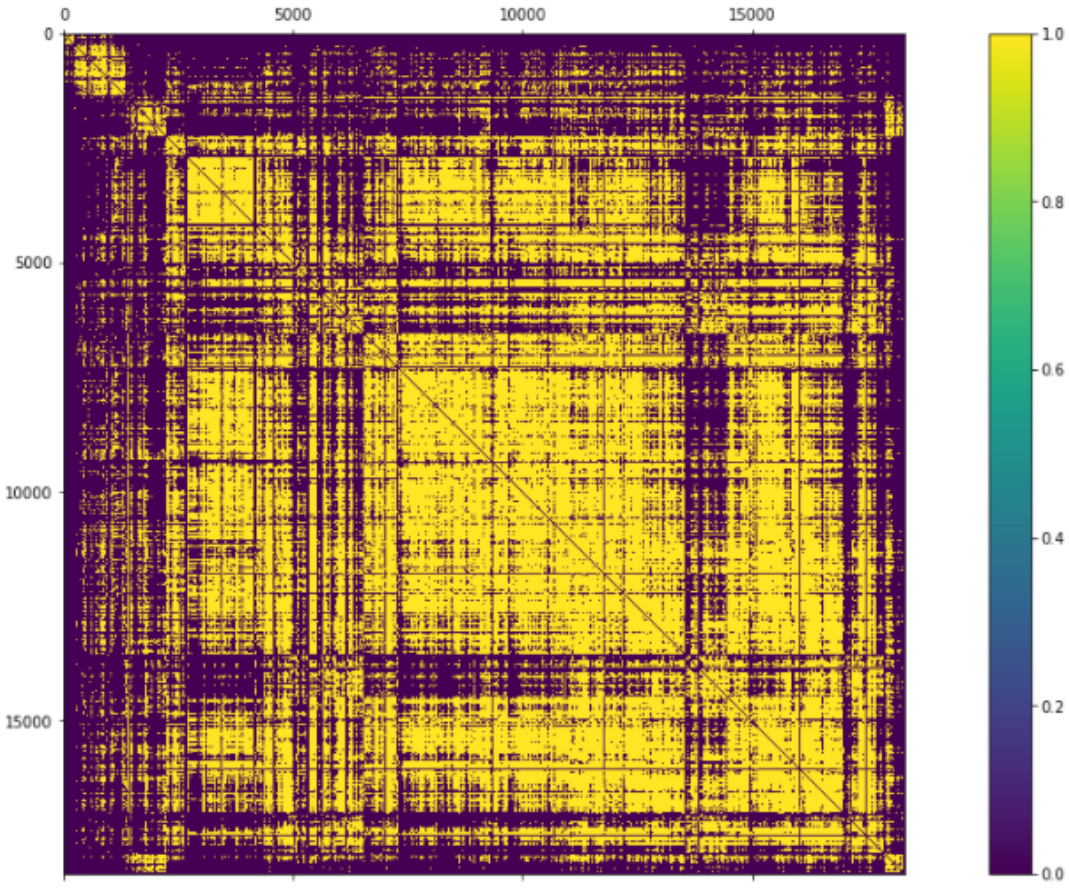

Figura 4: Matriz de adjacência obtida através da matriz de distâncias.

A média de valores com 1 foi de aproximadamente 48,44% dos valores e a diagonal, como pode ser vista, se manteve com 0.

Após a obtenção da matriz de adjacência, houve problemas com o tempo de execução para a obtenção da rede, como vimos, temos 18336 moléculas com 19 átomos e inicialmente os dados com os quais estávamos trabalhando apresentavam 18336 linhas referentes às moléculas e no máximo 2000 colunas referentes aos descritores, o que já eram valores bastante elevados, com a obtenção da matriz de adjacência, os dados passaram a apresentar 18336 linhas por 18336 colunas o que tornou o tempo de execução muito grande, desta forma, para a obtenção da rede, foi usado uma parcela da matriz de adjacência original e a rede resultante pode ser conferida pela Figura 5.

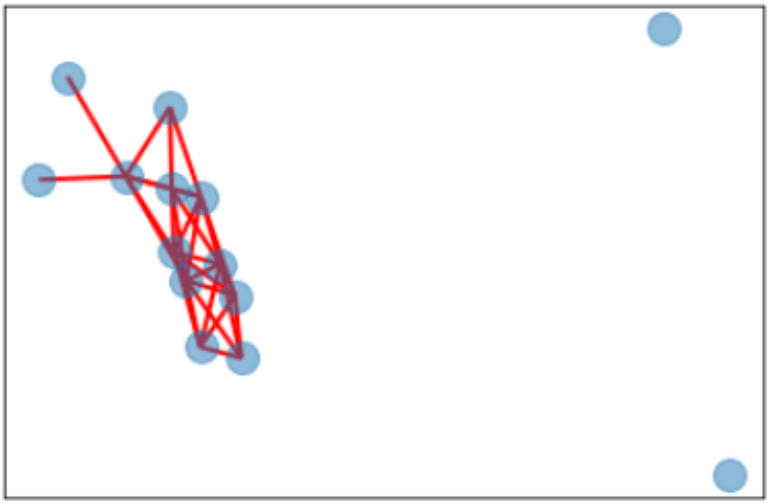

Figura 5: Grafo gerado por uma parcela da matriz de adjacência.

O grafo obtido apresentou uma densidade de aproximadamente 48%, levando em consideração que a parcela utilizada apresentava dimensões de aproximadamente 1000 x 1000 valores.

Também foi obtido a relação das moléculas com os seus respectivos graus representado pela Figura 6.

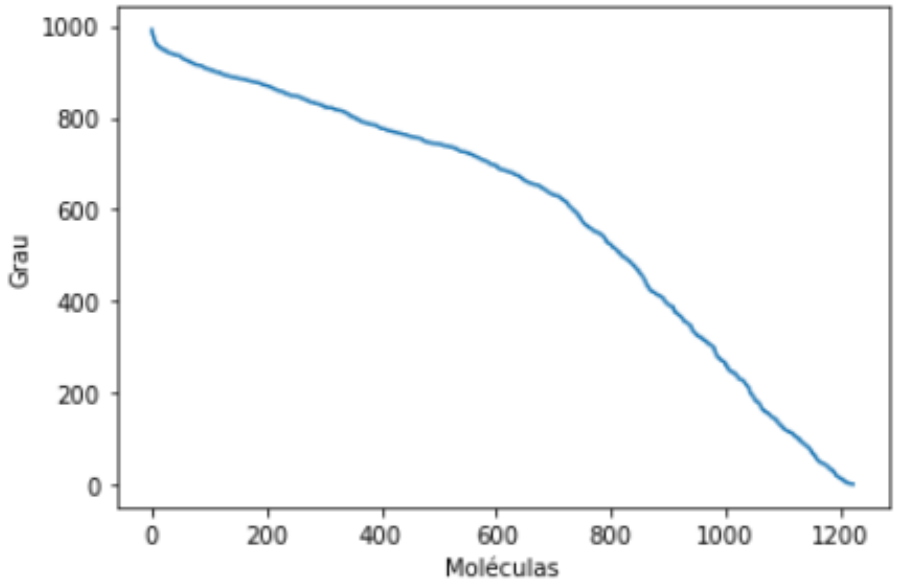

Figura 6: Grau de cada vértice da rede.

É possível observar, com a Figura 6, que a relação de moléculas com os seus graus se aproxima de uma linha reta com uma certa elevação somente no centro da reta, mostrando que o valor do grau mais comum está em torno de 600.

Por último, foi aplicado, então, o algoritmo de seleção de comunidade para a rede gerada. O resultado, então, foi a obtenção de 5 comunidades nos quais apresentavam 673, 546, 2, 1 e 1 moléculas como podemos ver pela Figura 7.

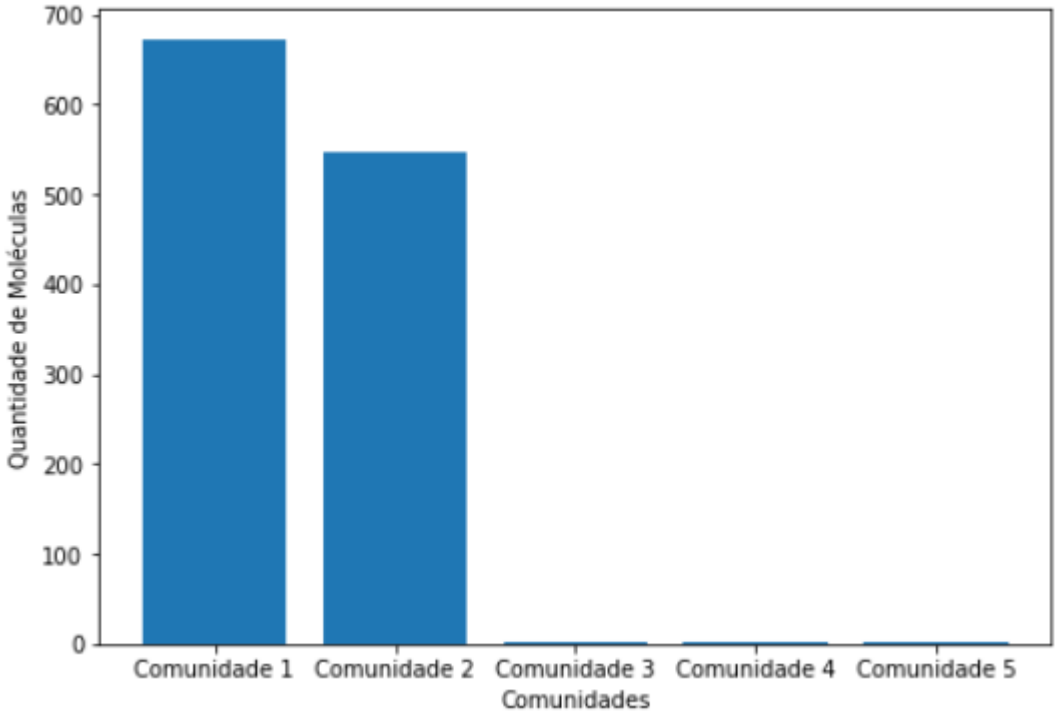

Figura 7: Comunidades obtidas após a aplicação da seleção de comunidades.

6 Considerações Finais

A obtenção das comunidades através do algoritmo de detecção de comunidades foi bem sucedida, 3 das 5 comunidades apresentaram uma quantidade de moléculas muito inferior ao restante tendo potencial para serem moléculas anômalas, porém, como mencionado, o uso da matriz de adjacência original se tornou inviável considerando o tempo de execução e o hardware disponível, e o projeto, desta forma, não pôde ser finalizado pela questão do tempo.

Em trabalhos futuros o projeto será melhor otimizado com o foco justamente no tempo de execução para que então possa ser cumprido com o objetivo de responder às perguntas propostas: Os exemplos do QM9 estão organizados em grupos (clusters)?; Moléculas problemáticas ou anômalas, que apresentam alto erro de predição, podem ser isoladas a partir dessa estrutura de agrupamentos?

7 Referências

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second ed., 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Comput. Surv*., vol. 41, no. 3, 2009.

[3] X. Zhu, *Semi-supervised learning with graphs*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2005.

[4] D. J. Cook and L. B. Holder, “Graph-based data mining,” *IEEE Intelligent Systems and their Applications*, vol. 15, pp. 32–41, March 2000.

[5] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, “Graph networks as a universal machine learning framework for molecules and crystals,” *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

[6] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, “Quantum chemistry structures and properties of 134 kilo molecules,” *Scientific data*, vol. 1, p. 140022, 2014.

[7] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, “Moleculenet: a benchmark for molecular machine learning,” *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[8] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, “Graph networks as a universal machine learning framework for molecules and crystals,” *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

[9] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, “Schnet–a deep learning architecture for molecules and materials,” *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241722, 2018. 9

[10] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, and P. Rinke, “Chemical diversity in molecular orbital energy predictions with kernel ridge regression,” *The Journal of chemical physics*, vol. 150, no. 20, p. 204121, 2019.

[11] B. Huang and O. A. Von Lilienfeld, “Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity,” 2016.

[12] P. B. Jørgensen, M. Mesta, S. Shil, J. M. García Lastra, K. W. Jacobsen, K. S. Thygesen, and M. N. Schmidt, “Machine learning-based screening of complex molecules for polymer solar cells,” *The Journal of chemical physics*, vol. 148, no. 24, p. 241735, 2018.

[13] W. Pronobis, K. T. Schutt, A. Tkatchenko, and K.-R. Müller, “Capturing intensive and extensive dft/tddft molecular properties with machine learning,” *The European Physical Journal B*, vol. 91, no. 8, p. 178, 2018.

[14] D. Weininger, “Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules,” *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[15] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, “Mordred: a molecular descriptor calculator,” *Journal of cheminformatics*, vol. 10, no. 1, p. 4, 2018.

[16] I. T. Jolliffe, *Principal Component Analysis*. Springer, second ed., 2002.

[17] sklearn.decomposition. PCA. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html> Acesso em 20 de ago. de 2021.

[18] Clauset, A., Newman, ME, & Moore, C. “Finding community structure in very large networks.” *Physical Review* E 70 (6), 2004.