COMS 4772 Project Report: Topic Modeling of Chemical Molecules Using Latent Dirichlet Allocation

Yixuan Li (yl3803), Erica Chenchen Wei (cw3137)

Abstract

Latent Dirichlet Allocation is applied on chemical molecules for topic modeling. Each molecule is treated as a document and be converted to a vector using a pretrained unsupervised learning approach, which inspired by Word2Vec technique in NLP. The topics are constrained as properties of molecules and document of each molecule focus on representing structure and elements it contains. Our result shows that the LDA can identify the correct topic on molecules by comparing their structures and properties.

1. Introduction

Topic modeling is a probabilistic framework originally developed to extract the hidden thematic structure of a collection of text documents. The extracted thematic structure is quite compatible with the concepts humans would use to organize the texts. Specifically, topic modeling imagines a fixed set of topics. Each topic represents a set of words. And the goal is to map all the documents to the topics in a way, such that the words in each document are mostly captured by those imaginary topics.

With large amount of data-sets on chemical structures, there has been increased interest in seeking approaches to organize and explore this data. Traditional clustering approaches like K-means usually rely on a similarity measure to divide the molecules into clusters. However, defining a similarity between compounds is often not straightforward and the resulting clusters are also often hard to interpret. Topic modeling does not require a similarity measure and allows interpretability.

In this project, we implemented topic modeling to organizing large molecule sets into "chemical topics", where they can be seen as patterns of co-occurring fragments that recurs across a set of molecules. We used molde Mol2Vec (Jaeger et al., 2018), which is inspired by Word2Vec technique in Natural Language Processing algorithms. Mol2Vec coming up with a vector embedding representation that clusters Morgan substructure in molecules with similar meanings, depending on their frequency of co-occurrence on neighbors. Then we applied the vector representation of molecules on LDA with different number of topics, as well as using unsupervised clustering on vectors. By comparing the molecule properties and substructures in the sample topic group and different topic group to understand if LDA and other unsupervised clustering can identify the correct topics by the Morgan substructure information.

Our focus is on evaluating the performance, interpretability and robustness of the new method. Since it is difficult to quantitatively measure the performance of a topic model, we used labeled data and quantify how well the model could reconstruct series of chemical compounds from a set of molecules. By interpreting these chemical topics, we would be able to understand certain properties of these molecules, i.e. "Melting Point" and we should also be able to discover the substructures of the molecules that lead to these properties. On the other hand, we believe that with the correct topic modeling on molecules, we can group molecules on their similar properties and substructures. This can be very helpful for chemical scientists to do experiments on a certain grouped molecules. It will reduce their work and time spending on without knowing any pre-grouped molecules.

2. Related Work

In (Schneider et al., 2017), they present the first chemistry-related implementation of using topic modeling on chemical molecules. They defined molecules as documents and the substructures or fragments derived from these molecules as words. After generating the substructures, they constructed a matrix where each row corresponds to a molecule and each column is the count of a certain substructure in that molecule. The matrix was used as the input to the LDA algorithm and the output contained two matrices: the topic-fragment matrix and the molecule-topic matrix. Based on these matrices the topics along with the most probable fragments per topic can be retrieved and visualized.

Inspired by word2vec representation, Mol2vec(Jaeger et al., 2018) is an unsupervised machine learning approach to learn vector representations of molecular substructures. It learns vector representations of molecular substructures that point in similar directions for chemically related substruc-

tures. We believe using Mol2vec embedding for representing molecules in our topic modeling task could improve our LDA performance and uncover more meaningful substructures for interpretation.

3. Dataset

We mainly used the meting point datasets from Jean-Claude Bradley's Legacy Dataset of Open Melting Points(Lang, 2014), which contains 28,645 measurements. After we cleaned and processing data into Mol2Vec, we have 20725 smiles string valid for our project.

To get more properties of each molecules, we wrote a script to dowload molecule properties from the OpenChemLib JS(cheminfo, 2015-2017) API by using SMILES string. We got 9 additional properties such as LogP, logS, Polar Surface Area, relative Weight, absolute Weight, rotatable Bond Count. Donor count, Center count, Accepted Bound. Adding the melting point, we have 10 properties in total.

Following table shows of our property labels with corresponding detail description.

LogP	octanol partition coefficient
LogS	aqueous solubility
psa	Polar Surface Area
mw	relative weight
em	absolute weight
DonorCount	sum of atoms having H donor
CenterCount	tetrahedral atoms
rotatableBondCount	Num of bonds with free rotation
acceptorCount	Num of molecular acceptor atom
mpC	melting point in Celsius degree

To test the stability of our method with a large-scale dataset, we download raw CID-SIMILES from PubChem, which contains 10 million smiles string and we converted part of them into 300 vector representations and got corresponding properties from OpenChemLib JS(cheminfo, 2015-2017) API. However, we couldn't finish the experiment on this large dataset due to time and computational constraints.

4. Methodology

Like the Word2vec models in Natural Language Processing, where vectors of closely related words are in close proximity in the vector space, Mol2vec(Jaeger et al., 2018) learns vector representations of molecular Morgan substructures that point in similar directions for chemically related substructures. The model Mol2Vec, which we used to convert molecules into vector representation, are pre-trained on 19.9 million molecules(so-called corpus) from ChEMBL version 23 and ZINC 15, two databases of small molecules, one storing bioactivity data, and the other commercially-available compounds, with skip-gram, window size 10 and dimension 300, feeding into two-layers neural network to train for the underlying substructure vector embedding.

The workflow is similar to Word2Vec, Mol2vec converts each SMILES string of molecules to a set of Morgan substructure, with each Morgan substructure having a unique integer string ("words"), then training the neural network with those substructures ("words") in different molecules ("sentences"). Once it gets underlying substructure vector embedding, it converts SMILES string(input data for a molecule) to a set of Morgan substructure, and then input each substructure into the model, it will give a 300-dimension vector. The model finally sums up all 300-dimension vectors to get a complete representation for a molecule.

LDA is a very popular topic model in recent years, among the fields of Natural Language Processing, computational biology, machine learning and artificial intelligence. LDA is introduced first by David Blei, Andrew Ng and Michael I. Jordan (Blei, 2012), a generative statistical model where the document topics are embedded in the mixture collections of words. Documents comes from a random process where each document is produced from a set of topics, and these topics are distributions over a fixed vocabulary. LDA is a hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics(Blei, 2012). It's a very similar process in understanding the hidden properties of molecules in chemistry while we observe certain elements and substructures of compounds. In chemistry, the properties of compound depend on certain elements it contains and certain substructure within that compound.

From this intuition, we want to experiment if LDA can also detect the properties of molecules(e.g. topics in documents) for given molecules(documents). We will need to generate compatible datasets as input of LDA, which are supposed to be like sentences in language while preserving the structures and elements in molecules. Inspired from Mol2Vec introduced by J. Chem. Inf. Model., which is an unsupervised machine learning approach to learn vector representations of molecular substructures(Jaeger et al., 2018), we are confident that we can convert substructures of molecules into vectors. Then we want to combine embedding vectors with other information of molecules such as bond number, elements counting and angle between bonds to get a full representation of molecules.

4.1. Vector Representation

Embedding molecules to a vector representation is the key premise in our project, our project are based on such representation available, which from Mol2Vec(Jaeger et al.,

2018). The strategy to get a vector representation has been mentioned on Method section, but here we want to analyze why this vector embedding is useful and convincing. Molecular fingerprints are a way to represent molecules as mathematical objects, it extracts features of the molecule, based on atom and radius around the atom, hash them, and use the hash to determine bits that should be set. Therefore, we can be certain that each Morgan substructure is represented by a unique hash value. Here's an example, showing as Figure 1., by using radius 1, we can convert an amino acid 'CC(N)C(=O)O' to a set of integer strings. As

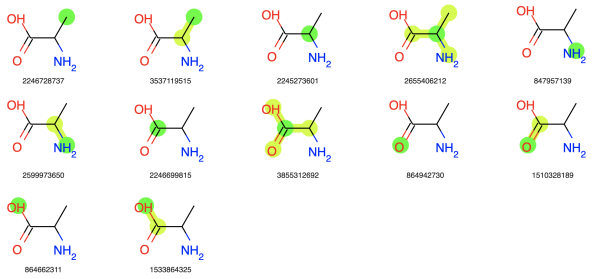

Figure 1. Identify morgan substructure

we mentioend before, each substructure will have a unique 300-dimensional vectors, then summing them up will give the complete representation for a molecule.

From Figure 2., by using t-SNE 2d projection, we plot the each substructure of 300 vector as well as the total 300 vectors presenting the example molecule 'CC(N)C(=O)O'.

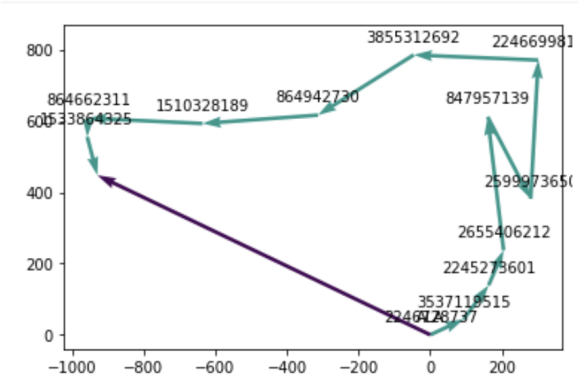

Figure 2. 2D projection on substructure and complete vector

Furthermore, from(Jaeger et al., 2018), they illustrate the 2D projections (t-SNE) of Mol2vec vectors of all amino acids. As showing in Figure 3., the bold arrows representing the total vectors as well as small arrows are the substructures in corresponding amino acids. As the arrow magnitudes reflect importance, the amino acids with similar substructures have the same direction and magnitudes on the plots.

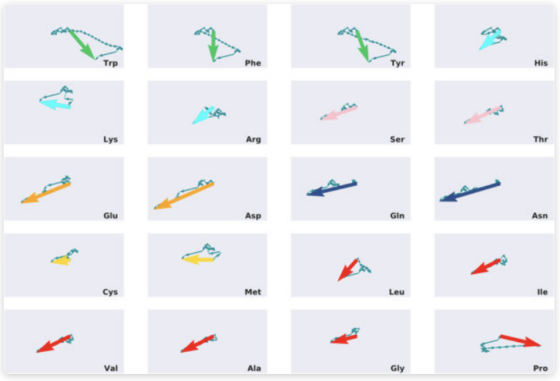

Figure 3. 2D projection on all amino acids

4.2. Chemical Topic Modeling

In chemical modeling, we define molecules as documents and the vector representation of substructures as words. As the input for the LDA algorithm, we take the vector representation of the molecules and we scale it for each molecule. The final LDA returns the molecule topic matrix as the output. We use the scikit-learn LDA implementation which is based on an online variational Bayes algorithm. And we adapt the following parameters for our chemical topic models: number of topics ([20, 40, 60]), learning method (we use "online" as default), maximum number of iterations for the optimization to converge (increased to 100), and the random state (to obtain a reproducible model). For online learning the model is optimized incrementally by running the optimization with chunks of the data set. The online approach is necessary when building models on large data sets that cannot be kept in memory at once. For the molecules, a topic profile – showing the probabilities of a molecule to be associated with a certain topic – can be extracted or alternatively, the most likely topic can be assigned to the molecules. This is a distinct advantage of the topic modeling approach compared to clustering methods like K-means where interpretation is difficult because the clusters are defined solely by the compounds that compose them up and their similarities.

5. Evaluation Criteria

In order to asses whether topic modeling is applicable to detect useful properties for chemical data, we designed several evaluations for our experiment: First we want to evaluated if the method is able to get "reasonable" topics from a set of compounds, for example, if the compounds being grouped together have similar chemical series which developed from same organic rings. After confirming that LDA can capture reasonable topics, we want to further evaluate if LDA can

detect chemical properties of compounds such as solubility and melting point. We evaluate the perform by collecting the distribution of the property values of the molecules in the same topic. As we believe that topic modeling offers better interpretability, we are also interested in comparing the performance of topic modelling with clustering methods.

Furthermore, we vary the number of topics to checking whether changing this parameter could drastically affecting the performance of LDA, as it is usually hard to define it without a thorough understanding of the dataset. This is a crucial aspect as we want LDA model to provide stable interpretation of the topics.

6. Results

6.1. Meaningful Substructures

We first train the LDA algorithm with 40 topics, and we sort the molecules in each topic with its probability of belonging to that topic in descending order. The distribution of molecules over the topics is shown in figure 4. In order to

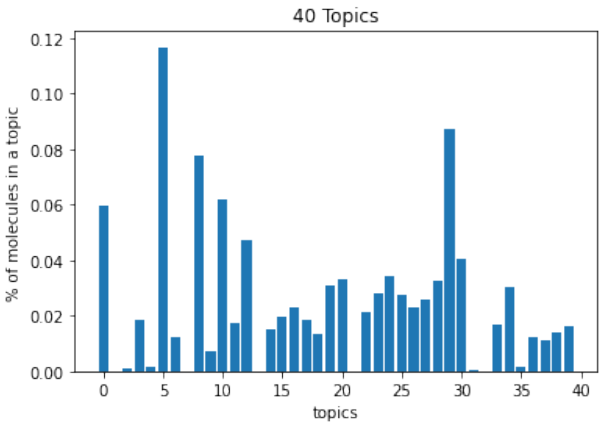

Figure 4. Topic modeling with 40 topics

interpret the meaning of the topics, we choose three topics at random and plot the chemical structures based on the SMILE information with the top 15 probabilities. In figure 5 and 6, we are showing the chemical structures in two of the topics, and we discover that most of the molecules with the double bond of oxygen are grouped in topic 0 and also molecules with one or more carbon circles are grouped in topic 5. In addition to the similar substructure of molecules in each topic, we also plot the distribution of the chemical properties of these molecules, for example, melting point, lipophilicity, and solubility of a compound (figure 7 and 8). In these histograms, we can still see that in some chemical properties, value are mostly centered around the mean value. For example, in topic 0, most of the molecules have the lipophilicity between 0 and 5, and in topic 5, the majority of the molecules have the solubility between -5 and 0. These figure are showing that LDA could indeed provide interpretable topics and by assigning meaning to these topics, it could help our understand certain properties of the molecules.

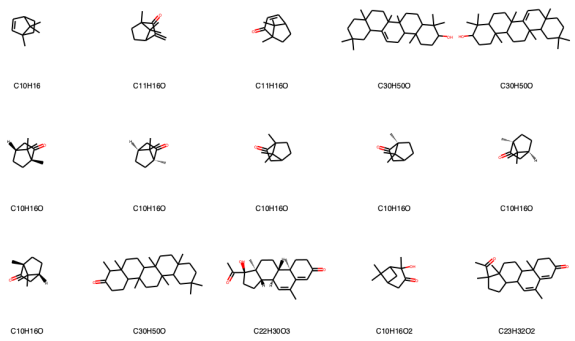

Figure 5. Chemical structures in topic 0

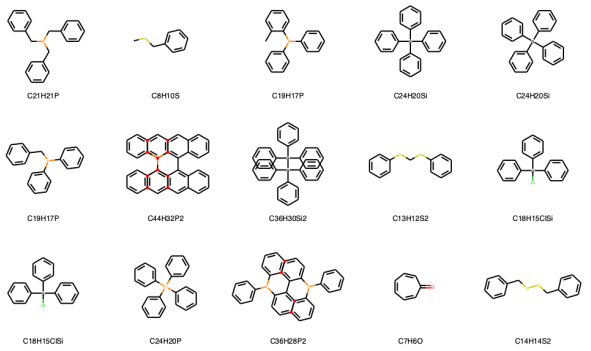

Figure 6. Chemical structures in topic 5

6.2. Comparison with Clustering Algorithms

We also want to compare the difference of the topic modelling approach and clustering methods. We choose Gaussian mixture models and hierarchical clustering as the comparing algorithms and we set the number of clusters to the same as the number of topics used in the LDA algorithm. The distributions of molecules across clusters of the two clustering algorithms are shown in figure 9 and 10. Compared to the clustering algorithm, the molecules in topic modeling are more evenly distributed and while with clustering algorithm, there is always one cluster with significantly more molecules.

To further investigate the interpretability of the clusters, we sample 15 molecules of a random topic and draw the corresponding chemical structures (figure 11 and 12). Although some of the molecules appear to have similar substructures, it is harder to find a general substructure for all the molecules

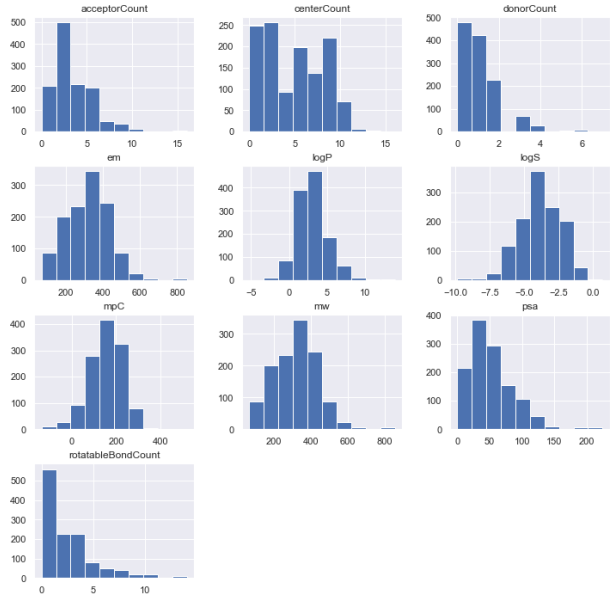

Figure 7. Properties of Topic 0

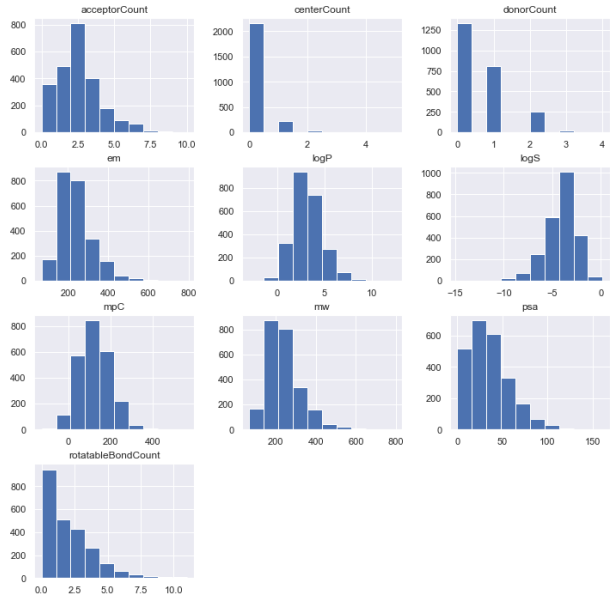

Figure 8. Properties of Topic 5

within the same cluster. For example, in figure 12, the bond between two carbon circles has various shapes and there are also molecules having one or three carbon circles. Moreover, when examining the quality of the clusters using silhouette coefficient, which is calculated as

$$s = \frac{b - a}{\max(a, b)}$$

where a is the mean distance between a sample and all other points in the same class, and b is the mean distance between a sample and all other points in the next nearest cluster, we

Figure 9. Gaussian mixture model with 40 clusters

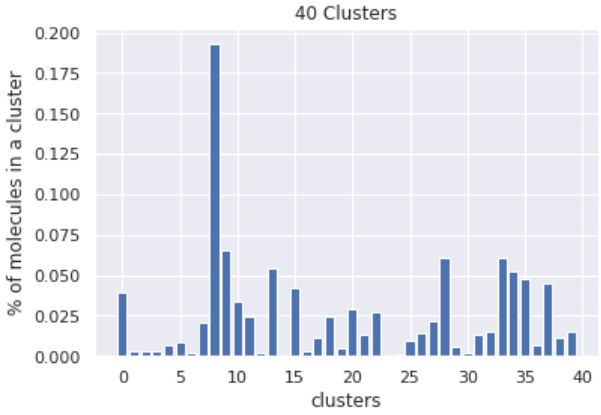

Figure 10. Hierarchical clustering with 40 clusters

Hierarchical Clustering	-0.31
Gaussian Mixture Model	-0.28

Table 1. Silhouette Coefficient for Clusterings

found both of the clustering have values below zero, implying that the clusters are not dense and well separated enough. Therefore, topic modeling provides more consistency of the substructures and therefore, more interpretability.

6.3. With Different Number of Topics

One of the challenges of using LDA for topic modeling is the uncertainty of the number of the topics. Selecting the number of topics may seem to be arbitrary without any previous knowledge of a certain dataset. Therefore, we want to experiment with different number of topics and see if these models could also generate meaningful topics. Figure 13 and 14 show the chemical structures in two random topics generated by a LDA with 20 topics and figure 15 and 16 are the chemical structures in two random topics with a LDA on 60 topics. As the figure shows, there are also similar

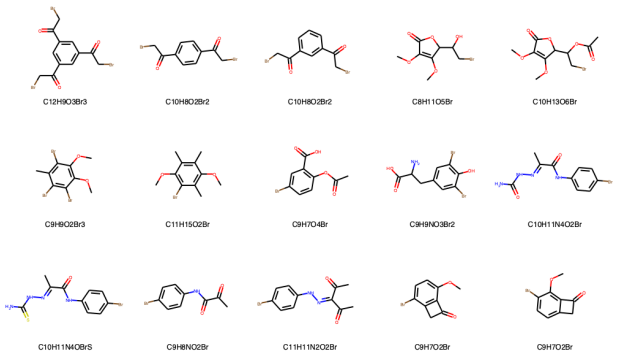

Figure 11. Chemical structures with GMM

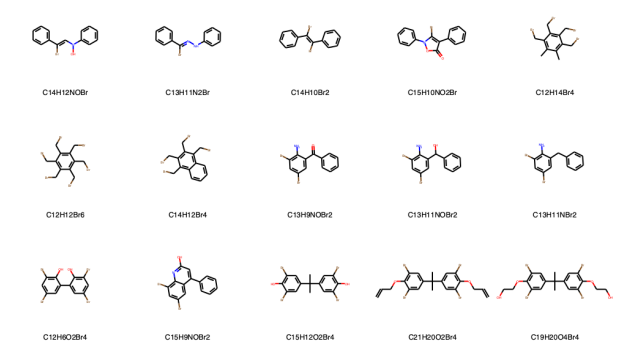

Figure 12. Chemical structures with hierarchical clustering

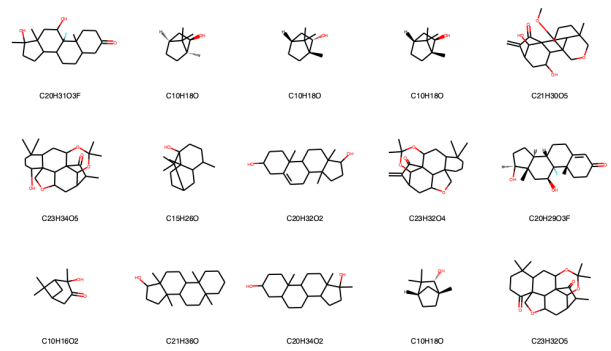

Figure 13. Chemical structures in topic 1 out of 20 topics

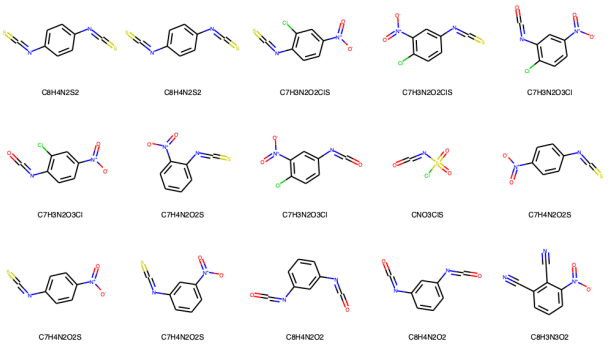

Figure 14. Chemical structures in topic 9 out of 20 topics

substructures within a same topic. For example, in topic 40 of the LDA model trained with 60 topics, each molecule has either a triple bond or a double bond with the ammonium element, and in topic 11, each compound has one or more single bond with the hydroxide. Although the number of topics is chosen arbitrarily, we can still observe meaningful groups of the elements. As chemical modeling is robust to this parameter, it could be easily used on other dataset to assist the understanding of it. However, it is not clear of how to decide the optimal number of topics for the LDA model as a quantitative evaluation of performance is challenging. Such problem may need further investigation.

7. Discussion

Topic modeling with LDA offers a lot of interesting advantages. In many of the examples above the chemical topic model allows intuitive visualization of the topics mapped directly onto the molecules. Investigating the top fragments of the topics of the model enables quick identification of "interesting" topics to analyze further. Comprehending the model is a huge benefit for researchers and the ability of the chemical topic model to reproduce human-assigned concepts makes it a great tool to explore sets of molecules. Besides, the topic model describes a generative process which could be used to generate novel molecules: having a chemical topic model of different series of active molecules against a certain target, the model could be used to create novel molecules by combining the fragments and the topics. Our code can be found at: https://github.com/EricaWei053/Topic-Modeling-on-molecules

8. Future work

There are many future work can be done on this topic. For example, we may use largw-scale dataset to train our model for the stability. We may use different neural network architectures to predict topics, similar as sentimental analysis in Natural Language Processing since we found there is a similarity between vector representing on molecules and sentences. We may also consider finding a way to go backward, using known properties to predict the molecule substructures. Also, improving computational efficiency is also an aspect we want to investigate.

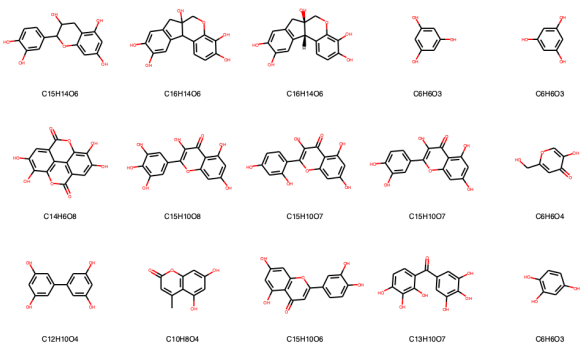

Figure 15. Chemical structures in topic 11 out of 60 topics

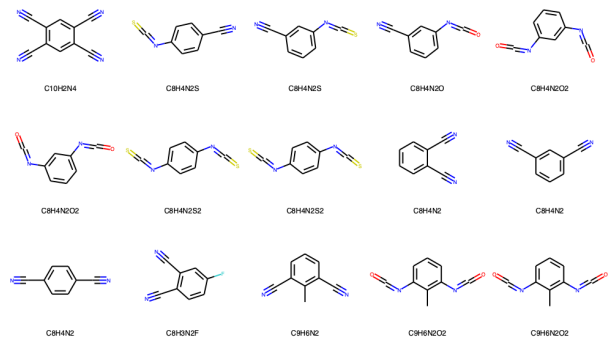

Figure 16. Chemical structures in topic 40 out of 60 topics

References

Blei, David M.; Ng, A. Y. J. M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2012. doi: 10.1162/jmlr.2003.3.4-5.993.

cheminfo. Openchemlib js. *Github*, 2015-2017. URL https://github.com/cheminfo/openchemlib-js.

Jaeger, S., Fulle, S., and Turk, S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 58(1):27–35, 2018. doi: 10.1021/acs.jcim.7b00616.

Lang, J.-C. B. A. W. A. Jean-claude bradley's legacy dataset of open melting points. *Figshare*, 2014. URL https://figshare.com/articles/Jean_Claude_Bradley_Open_Melting_Point_Datset/1031637.

Schneider, N., Fechner, N., Landrum, G. A., and Stiefl, N. Chemical topic modeling: Exploring molecular data sets using a common text-mining approach. *Journal of Chemical Information and Modeling*, 57(8):1816–1831, 2017. doi: 10.1021/acs.jcim.7b00249.