Continuous Representation of Molecules Using Graph Variational Autoencoder

Mohammadamin Tavakoli, Pierre Baldi
Department of Computer Science,
University of California, Irvine
{mohamadt, pfbaldi}@ics.uci.edu

Abstract

In order to continuously represent molecules, we propose a generative model in the form of a VAE which is operating on the 2D-graph structure of molecules. A side predictor is employed to prune the latent space and help the decoder in generating meaningful adjacency tensor of molecules. Other than the potential applicability in drug design and property prediction, we show the superior performance of this technique in comparison to other similar methods based on the SMILES representation of the molecules with RNN based encoder and decoder.

Introduction

Using machine learning to predict molecular structure properties is a challenging problem [7, 3]. While the governing equations (e.g. Schrodinger equation) are difficult and computationally expensive to solve, the fact that an underlying model exists is appealing for machine learning techniques. However, this problem is difficult from a technical point of view. The space of molecules is discrete and non-numerical. Thus, *"how to best represent molecules and atoms for machine learning problems?"* is still a question.

Despite having numerous ways to represent molecules such as methods introduced in [18, 1], all the representations are suffering from a few shortcomings, such as 1) discrete representation, 2) lengthy representation, 3) non-injective mapping, and 4) non-machine readable representation.

Here, we proposed a new method that borrows the main idea from [5] and [12] and overcomes all the aforementioned shortcomings. Our method which takes the graphical structure of the molecule as the inputs consists of a variational framework with a side predictor to better prune the structure of the latent space. Then an inner product decoder transfers the samples of latent space into meaningful adjacency tensors. To compare with the main benchmark which is a text-based encoding of molecules [9] we performed two experiments on the QM9 dataset [16, 15] and ZINC [11]. Both experiments show the success of this method. Although this work is presenting preliminary results of Graph VAE, further experiments and comparisons are left to future work.

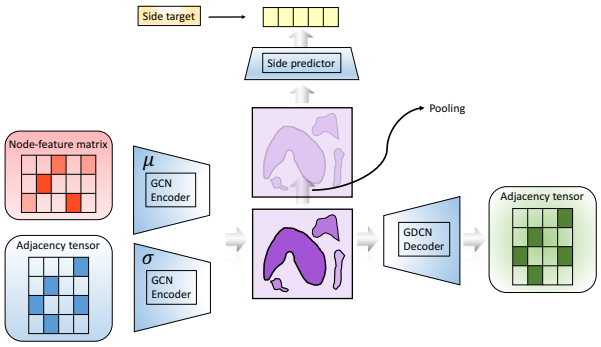

Figure 1: Outline of the model. As depicted above, the model inputs are both node-feature and adjacency tensor, while the model is only outputting the adjacency tensor. The side prediction network is simply using the data points in the latent space as its input.

Method

Molecules and Graphs A molecule can be represented by an undirected graph $G = (V, E, R)$, with nodes (atoms) $v_i \in V$ and labeled edges (bonds) $(v_i, e, v_j) \in E$ where $r \in R$ is an edge type. Since we focus on small molecules with four bond types, R is equal to 4. An n by d node-feature matrix H is also carrying more information about each node. These two tensors, together, represent a molecular structure.

Variational Autoencodes To help ensure that points in the latent space correspond to valid realistic molecules, and to minimize the dead areas of the latent space, we chose to use a variational autoencoder (VAE). To further ensure that the outputs of the decoder are corresponding valid molecules we employed the open-source cheminformatics suite RDKit30 to validate the chemical structures of output molecules in terms of atomic valence. All invalid outputs are discarded. It is necessary to mention that the ordering of the nodes assumed to be unchanged.

VAE and Side Prediction To better learn the graph structure of the molecules, the encoder part of the VAE consists of GCN layers. The same method as [17] has been employed to perform relational update which can be formulated as:

$$h_i^{l+1} = \sigma(\sum_{i \in R} \sum_{j \in N_r^i} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)})$$

where N_r^i denotes the set of nodes connected to node i through the edge type $r \in R$. Since we are focusing on small molecules, we applied three layers of GCN in our encoder model to gather information from 3-hop neighbors of each atom. The structure of encoder consists of two, three-layer GCNs for both mean and the covariance. GCNs of the encoder share the filters of the first two layers. Here we can formulate the encoding and sampling scheme as follows:

$$q(\mathbf{Z}|H, A) = \prod_1^N q_i(z_i|H, A),$$

$$q_i(z_i|H, A) = \mathcal{N}(z_i|GCN_\mu, GCN_\sigma)$$

The GCN_μ and similarly GCN_σ are: $GCN(H, A) = \hat{A}\sigma(\hat{A}\sigma(\hat{A}HW_0)W_1)W_2$, where the $\hat{A}$ is the normalized adjacency tensor, W_i is the filter parameter of each layer, and σ is the activation function [2]. Finally, as suggested in [12] we use the simplest form of the decoder which can be seen as graph deconvolution network. The output of the encoder is simply the inner product between latent variable:

$$p(A|\mathbf{Z}) = \prod_1^N \prod_1^N p(A_{ij}|z_i, z_j),$$

$$p(A_{ij} = 1|z_i, z_j) = \sigma({z_i}^T z_j)$$

For the side prediction part, we employ a simple regression model in the form of a multilayer perceptron (MLP) to the network that predicts the properties from the latent space representation. The input of the side predictor is a vector obtained through a pooling mechanism of the latent representation as follows:

$$G(H^{(L)}) = \sum_{i=1}^N softmax(h_i^L.W_p)$$

Where W_P is the pooling weight matrix and $H^{(L)}$ is the output of the GCN_μ.

Finally, the autoencoder is trained jointly on the reconstruction task and a property prediction task; The joint loss function is the summation of the two losses, as follows:

$$\begin{aligned}\mathcal{L} &= \textit{ELBO} + \textit{negative log likelihood} \\ &= \mathbb{E}_{q(\mathbf{Z}|H,A)} - KL(q(\mathbf{Z}|H, A)||p(Z)) \\ &+ MSE(sidenetwork)\end{aligned}$$

Experiments

We performed two experiments to show the usefulness of continuous representation. In the first experiment, we focus on the prediction of property and the generation of the valid molecules. In the second experiment, we use this continuous representation to propose a new metric for measuring the molecular similarity.

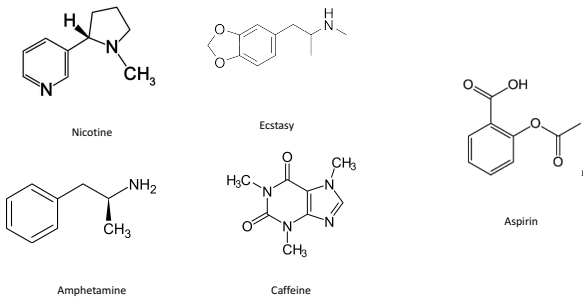

Figure 2: Drugs compare to Aspirin

Table 1: Three models trained with three different side property. As shown below, using *Druglikeliness* better helps predicting *Solubility* and *Synthesizability*

Side property	Valid outcome	*Sol*	*Synt*	*Druglikeliness*
Solubility	**75.3**	97.03	88.7	84.2
Synthesizability	**73.0**	89.8	98.21	86.3
Druglikeliness	**74.6**	91.0	90.7	95.11

Property Prediction

Using a subset of QM9 dataset [15] as the training set, we extract 48,000 molecules covering a broad range of molecules. Each molecule in the training set is chosen to have up to 20 atoms. The training objective on the side predictor was set to be one of the *Solubility*, *Druglikeliness*, and *Synthesizability*. We employ the continuous representation of molecules using each network to predict the other two unseen properties. The performance of each model plus the percentages of validly generated molecules are summarized in Table 1. In order to check the validity of the outcome, we only check for the validity of the atomic valence. As it is shown in Table 1 the accuracy of each property is comparable to the state of the art property predictions mentioned in [8]. Although Graph VAE is not outperforming the predictions based on [8], it shows that using a property as a heuristic to prune the latent space, can help with predicting other molecule properties.

Molecular Similarity Measure

Numerous similarity or distance measures have been used widely to calculate the similarity or dissimilarity between two samples. Since metrics are focusing more on 2-dimensional representation rather than 3-dimensional structure, our model as a "2D structure-aware representation" is an accurate metric for the similarity measure. Normalized Euclidean distance between the latent representation of two molecules after pooling operation is the metric we define to capture the similarity. Here we compare three well-known similarity measures with our technique and also to the methods introduced in [9]. This method which is using the SMILES representation of the molecules as the input employs a VAE with a side predictor. Both encoder

Table 2: Similarity measures between Aspirin and four different drugs. Using Graph VAE as a new metrics, shows consistency with other metrics. The GVAE is trained with the solubility as the side property.

metric	Amphetamine	Ecstasy (MDMA)	Nicotine	Caffeine
Tanimoto	0.398	0.324	0.229	0.258
Dice	0.569	0.490	0.373	0.410
Cosine	0.607	0.490	0.374	0.434
Graph VAE	0.363	0.199	0.147	0.176
SMILES VAE [9]	0.724	0.489	0.340	0.321

and decoder parts of the VAE are based on RRN and sequence to sequence model. Although all the graphical information of the molecule is encoded within the SMILES representation, inferring the graphical structure (e.g., adjacency tensor) from the SMILES string is an exhausting process that is based on several rules. Despite the numerous techniques built upon using the SMILES representation of the molecules [6, 10, 4, 14, 13], it has been shown that it is more efficient to take advantage of the graph structures and employ GCNs to process molecular structures. Here, we chose Aspirin as a sample drug and compare its similarity with four different drugs with four different similarity measures. We compare the performances of our technique with [9], which is using a similar approach but operating on text representation of molecules. Our experiment shows that graph-based hidden representation is carrying more information than only text. Table 2 is summarizing the result of the similarity measure experiment.

As it is shown in table 2, our metric is very well aligned with all other well-known metrics which is another proof for the applicability of our model.

Experiment Details

GVAE consists of two GCNs for the encoder, a pooling mechanism, and a multi-layer perceptron for the side prediction. Both GCNs are three-layer networks with filter matrices W_0, W_1, and W_2 of 32*32, 32*32m and 32*16 respectively. The pooling weight matrix W_p is of size 1*64 which outputs a vector of length 64 to represent the whole molecule. A two-layer MLP with 32 and 1 hidden units is employed to perform the regression task.

In Table 2, we use our own implementation of the SMILES VAE. Both GVA and SMILES VAE are trained using a dataset of 70,000 molecules which are randomly selected from ZINC.

In Table 2, all measures except the continuous representations are calculated with the same fingerprinting algorithm. It identifies and hashes topological paths (e.g. along with bonds) in the molecule and then uses them to set bits in a fingerprint of length 2048. The set of parameters used by the algorithm is - minimum path size: 1 bond - maximum path size: 7 bonds - number of bits set per hash: 2 - target on-bit density 0.3.

Conclusion

We proposed a generative model through which we can find continuous representation for molecules. As shown in the experiments section, this technique can be used in different chemoinformatics tasks such as drug design, drug discovery and property prediction. As future work, one can think of attention based graph convolutions and more complicated decoders. These two extensions can be studied in future works.

References

[1] E. J. Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*, 2017.

[2] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2015.

[3] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10(2):370–377, 2019.

[4] A. Dalke. Deepsmiles: An adaptation of smiles for use in. 2018.

[5] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2224–2232. Curran Associates, Inc., 2015.

[6] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung. Deep learning for molecular design-a review of the state of the art. *Molecular Systems Design & Engineering*, 2019.

[7] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, and P. Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 2018.

[8] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[9] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[10] M. Hirohara, Y. Saito, Y. Koda, K. Sato, and Y. Sakakibara. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC bioinformatics*, 19(19):526, 2018.

[11] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

[12] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, 2016.

[13] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik. Selfies: a robust representation of semantically constrained graphs with an example application in chemistry. *arXiv preprint arXiv:1905.13741*, 2019.

[14] S. Kwon and S. Yoon. Deepcci: End-to-end deep learning for chemical-chemical interaction prediction. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 203–212. ACM, 2017.

[15] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules.

[16] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

[17] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[18] 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*, 9(2):513–530, 2018.