Deep Molecular Graph InfoMax

Boury Mbodj [1] **William L. Hamilton** [1 2]

Abstract

Machine learning has shown promise in tackling various problems in drug discovery; however, the shortage of annotated data is still a major obstacle to its success. To address this limitation, we introduce *Deep Molecular Graph InfoMax*, a framework for the completely unsupervised learning of molecular graph-level representations. Unlike previous work on graphs that leverage mutual information between homogeneous views of data, we focus on an approach that additionally utilizes auxiliary molecular descriptors and provides flexible objectives for better generalization. We demonstrate that our model is competitive with strong baselines when generating representations in the fully unsupervised setting.

1. Introduction

Traditional drug development is a complex, costly, and lengthy process that typically spans more than a decade (DiMasi et al., 2018). To this extent, machine learning techniques have been of particular interest in the application of several stages of the drug discovery pipeline, such as in prediction mechanisms and target identification. Recently, there have been significant advances in applying supervised graph neural networks (Bruna et al., 2014; Duvenaud et al., 2015; Gilmer et al., 2017; Schutt et al., 2016), which satisfy the rotational, translational and permutational invariances of molecules naturally, to property prediction tasks.

Indeed, representation learning has shown strong potential to address the shortcomings of conventional quantum mechanical simulations methods such as Density Function Theory (Capelle, 2006), used in both biological and materials science. It treats the problem of representing graph structure as a machine learning task itself by using a data-driven approach to learn embeddings (Hamilton et al., 2017). However, the tradeoff between accuracy and computational cost, and the limited amount of labeled training data remains a challenge. The latter is partly due to legal and privacy constraints on work with sensitive health records in the pharmaceutical industry. As such, learning good representations without relying on annotations is an essential step towards improving machine learning models for drug development.

Sun et al. (2020) proposed an unsupervised graph-level representation learning model termed InfoGraph. The impressive results of Deep InfoMax (Hjelm et al., 2018) motivated the creation of this model which utilizes mutual information (MI) maximization for unsupervised learning between graph-level representations and the representations of substructures of different granularity. Inspired by this recent work on DIM and InfoGraph, we present a novel method for learning representations in an unsupervised manner between molecular graphs and Simplified Molecular-Input Line-Entry System (SMILES) strings based representations (Daylight Chemical Information Systems, 2019; Weininger, 1988). Our contributions can be summarized as follows:

- We propose learning efficient molecular representation via MI maximization between heterogeneous views of molecular data.
- We show that performing MI maximization between high-level representations can significantly enhance graph-level embeddings and surpass strong baselines on a property prediction task.

2. Related Work

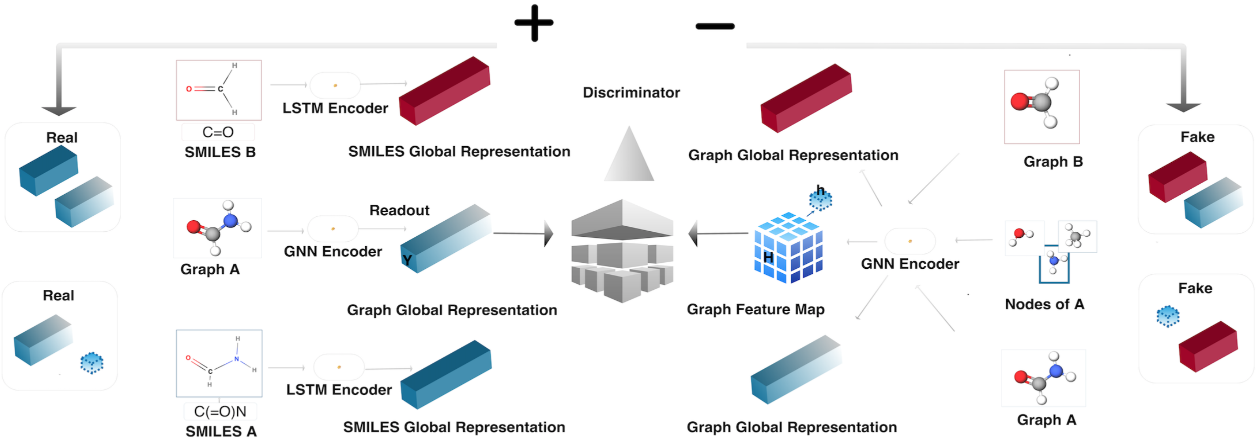
Our framework builds upon recent research based on deep learning on molecular data, particularly graph representation learning.

A diverse set of work on molecular deep learning models has been presented to advance the drug discovery process. Early work involved fingerprinting methods, used in the training of neural networks, with one of the most prevalent techniques being the Extended Connectivity FingerPrinting (Rogers & Hahn, 2010). Later on, other architectures that directly operate on graphs such as neural fingerprints (Duvenaud et al., 2015) and molecular graph convolutions

[1] McGill University, Quebec, Canada [2] Mila. Correspondence to: Boury Mbodj <mbodj.boury@gmail.com>, William L. Hamilton <wlh@cs.mcgill.ca>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

Figure 1. **Model Architecture**: Deep Molecular Graph InfoMax is a combination of both global mutual information maximization **(left)** and local mutual information maximization **(right)** schemes. The influence of both global and local objectives is determined by the choice of hyperparameters α and β.



(Kearnes et al., 2016) emerged. These approaches have been included in the message passing neural networks (MPNNs) framework which consists of other supervised graph neural networks (Defferrard et al., 2016; Li et al., 2016).

Deep Graph InfoMax (DGI) and InfoGraph are two notable unsupervised graph representation learning approaches that have been recently proposed. The former is a node-level learning technique that relies on applying the objective function between patch representations and corresponding high-level summaries of graphs. Whereas the latter focuses on learning graph-level representations in both unsupervised and semi-supervised scenarios via MI maximization.

InfoGraph differs from DGI in architecture as it is also constructed with a pooling operator. Also known as readout, the pooling operator aims to reduce the size of parameters by down-sampling the nodes to generate smaller representations as well as avoid overfitting and ensure permutation invariance (Wu et al., 2019). In contrast to InfoGraph, our method not only offers the flexibility to focus on the global structure of the representation, but it also incorporates representations of auxiliary molecular descriptors.

3. Proposed Method

In this section, we discuss the intuition behind our framework and formulate the problem of learning representations for molecular graphs. Then, we expose the unsupervised learning setup of graph-level representations that focuses on utilizing high-level representations of heterogeneous views of molecular data. We then outline the setting based on maximizing MI between representations of subsets of the graph such as the nodes and the information content of the entire graph.

3.1. Model Intuition

Given a dataset of multiple different molecules represented as both graphs G $= \{G_1, G_2, ...\}$ and SMILES S $= \{S_1, S_2, ...\}$, our objective is to learn meaningful embeddings by utilizing the gradients from the discriminators to help train their respective encoder's network.
Each graph representation, with n total number of nodes, is composed of a set of node embeddings h such that H $= \{h_1, h_2, ..h_n\}$ denotes the summary representation of all nodes/patch representations of the graph. The feature vector of a molecular SMILES based representation or graph is expressed as Y. In the case that entirely prioritizes global information content, our model maximizes MI by discriminating the summarized graph representation from the global SMILES representation. Equivalently, in the localized version, the discriminator is trained to distinguish the summarized graph representation from its node representations.We implement our framework with the encoders and discriminators specified in the subsection below.

3.2. Encoders and Discriminators

SMILES Strings Encoder The Simplified Molecular-Input Line-Entry System (SMILES) is a string-based representation of molecules that was introduced in the late 1980s and represents a universal standard for many software applications. Numerous studies have demonstrated the effectiveness of models based on SMILES fed to neural networks such as RNNs (Goh & N. Hodas; Wu et al., 2018). In order to enforce permutational invariance and uniqueness, we use the canonicalized representation of the molecule's

SMILES representation as a 1-hot-encoding input to a Long Short term memory. The LSTM is used both as part the DMGI framework, but also as a competitive baseline (i.e LSTM autoencoder) in order to gauge the quality of our model.

Graph Model Encoder As in InfoGraph, we choose a Message Passing Neural Networks (MPNNs) *(enn-s2)* introduced by Gilmer et al. (2017) as our main graph encoder. On an undirected graph $G = (V, E)$ with node features x_v, edge features e_{vw} and a neighborhood defined as $N(v)$, a message passing process is composed of two phases, a messaging phase (2) and (3) a readout phase (4). The general message passing phase is defined by the following formula:

$$\mathrm{m}_v^{t+1} = \sum_{w\epsilon N(v)} \mathrm{M}_t(\mathrm{h}_v^t, \mathrm{h}_w^t, \mathrm{e}_{vw}). \tag{1}$$

$$\mathrm{h}_v^{t+1} = \mathrm{U}_t(\mathrm{h}_v^t, \mathrm{m}_v^{t+1}). \tag{2}$$

Where the message m_v^{t+1} is the transition function that propagates information and h_v^t denotes the hidden states that are updated for T iterations. The readout phase (6), which is computed with a readout function R, generates a representation of the entire graph based on node and edge hidden representations.

$$\hat{\mathrm{y}} = \mathrm{R}\Big(\{h_v^t | \mathrm{v}\epsilon G\}\Big). \tag{3}$$

The *(enn-s2s)* MPNN variation is constructed with the continuous edge network message function, a gated recurrent unit (Chung et al., 2014) update function, and a Set2Set (Vinyals & S. Bengio, 2015) readout operator. The power of this encoder can be attributed to the inclusion of edge attributes which is an essential feature for molecules.

Discriminators Both graph and SMILES string-based descriptor discriminators are implemented as feed-forward neural networks.

3.3. Objective Function

We select the Jensen Shannon Divergence (JSD) estimator as the main objective function of our model. The generic formula of the JSD we will be following throughout the rest of the paper is :

$$\max I^{JSD}(X;Y) := \max D_{JS}\left(P(X,Y)\,||P(X)\,P(Y)\right) \tag{4}$$

$$= \max\Bigg(2\log 2 + E_{p(x,y)}\Big[-\,\mathrm{sp}\Big(-\,\mathrm{T}\left(x,y\right)\Big)\Big] - E_{p(x)p(y)}\Big[\mathrm{sp}\Big(\mathrm{T}\left(x^{'},y\right)\Big)\Big]\Bigg) \tag{5}$$

Here, we maximize the lower bound on the JSD the divergence, where T is the discriminator, x is an input sampled from P and $x^{'}$ is the negative input pair. More specifically, we generate useful representations by training the discriminator T to estimate the JSD divergence and train the encoder to minimize this estimate. Similarly to InfoGraph, negative samples are generated by permuting the batch.

3.4. Global Mutual Information Maximization

Our model architecture is illustrated on the figure (1), where the global MI maximization is on the left of the discriminator. We can express the objective function for as :

$$\max_{\phi\omega_1} \mathrm{I}_{\phi\omega_1}\big(Y_{1\phi}\left(\mathrm{G}\right); Y_{2\phi}\left(\mathrm{S}\right)\big). \tag{6}$$

Let I denote the mathematical notation of mutual information between the global representation Y_1 and Y_2 after applying the readout operations of the graph (G) and LSTM (S) encoders, respectively.

3.5. Local Mutual Information Maximization

The local MI maximization approach illustrated on the right of figure (1) is equivalent to InfoGraph and DIM methodology to some extent. I designates the mutual information between the patch representation h centered at an arbitrary node of G and the global representation Y after applying the readout operation.

$$\max_{\phi\omega} \textstyle\sum_{G\epsilon G} \frac{1}{|G|} \mathrm{I}_{\phi\omega}\big(\mathrm{h}_\phi\big(\mathrm{G}\big); \mathrm{Y}_\phi(\mathrm{G})\big). \tag{7}$$

We note that most, if not all existing graph representation learning work on MI maximization, present the local MI maximization as a more robust, or suitable objective than global or combinations of global and local objectives.

3.6. Deep Molecular Graph InfoMax

As the generic version of our model, we define the complete objective for DMGI below:

$$\alpha \max_{\phi\omega_1} \mathrm{I}_{\phi\omega_1}\big(Y_{1\phi}\big(\mathrm{G}\big); Y_{2\phi}(\mathrm{S})\big) + \beta \max_{\phi\omega_2} \textstyle\sum_{G\epsilon G} \frac{1}{|G|} \mathrm{I}_{\phi\omega_2}\big(\mathrm{h}_\phi(\mathrm{G}); Y_\phi(\mathrm{G})\big). \tag{8}$$

where α and β are hyper-parameters for the global and local objectives, respectively. Both α and β are arbitrarily set to 0.5 in our experiments. Similarly to DIM, integrating the hyper-parameters allows us to exercise control over the information, either locally or globally. However, in contrast to DIM, we do not impose structural constraints on high-level representations with prior matching.

Table 1. **Mean absolute error (MAE) accuracy results on QM9 molecular property prediction regression task**. We compare our model with an LSTM autoencoder (LSTM-AE), and InfoGraph which is the same as DMGI-Local. DMGI-Global refers to the global mutual information maximization objective as the name implies, and DMGI denotes the generic model with parameters α and β arbitrarily set to 0.5 each.

Target	LSTM-AE	InfoGraph	DMGI-Global	DMGI
mu	1.2639	0.8231	**0.8168**	0.8450
alpha	5.0834	2.2483	**1.8585**	2.5098
Homo	0.0169	0.0108	0.0111	**0.0106**
Lumo	0.0380	0.0153	0.0154	**0.0150**
gap	0.0357	**0.0182**	0.0194	0.0184
R2	204.9099	52.3440	**42.4115**	46.4030
ZVPE	0.0239	0.0052	0.0050	**0.0047**
U	24.4618	11.4668	**10.3556**	11.9785
U0	30.9940	11.6948	**9.8761**	12.1456
H	22.7233	11.3416	**10.0502**	12.0853
G	28.8239	11.5395	**9.9134**	11.2979
Cv	2.6136	1.1697	**0.7131**	1.0836

4. Experiments and Analysis

In this section, we empirically evaluate the performance of our model on a molecular property prediction task.

4.1. Dataset and Setup

We use the Quantum Machine 9 (QM9) dataset (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012) as our benchmark. It is composed of 134k molecules which are modeled using Density Functional Theory, and their properties are related to atomization energies, fundamental vibration frequency, states of electrons and measures of spatial distributions of the molecules (Gilmer et al., 2017).

All graph and LSTM based models are implemented using the PyTorch Geometric (Fey & Lenssen, 2019) and PyTorch (Paszke et al., 2017) deep learning libraries. Finally, we run the experiments using a Google Colab GPU, and train the models over different configurations.

4.2. Results and Discussion

Table 1 shows the results of the QM9 property prediction task. We observe that DMGI-Global outperforms all other unsupervised learning methods on 8 out of 12 target properties. More specifically, the methodology with the global objective distinctly achieves a lower MAE on all targets related to atomization energies (U, U0, H, G), spatial distributions of electrons in a molecule (alpha, R2) and the heat capacity (Cv). We find that the results obtained for properties related to fundamental vibrations of the molecule (HOMO, LUMO, and gap) are slightly better with local or combined objectives of our framework. Property prediction is a complex task; however, our results demonstrate that prioritizing global structure can be more suitable in some scenarios. It is possible that the model based on the local objective discards, or misinterprets edge attributes as noise. Further analysis will need to be conducted in future work, in order to determine an exact theoretical explanation.

5. Conclusion

In this paper, we presented Deep Molecular Graph InfoMax, an unsupervised representation learning technique to learn graph-level embeddings of molecules of arbitrary sizes. Through our experiments involving the QM9 benchmark dataset, we demonstrated that our framework surpasses two strong baselines, namely an LSTM autoencoder and InfoGraph. Although our experiments are task-specific, they still demonstrate the effectiveness of the introduced method, and prove that it is a methodology worth investigating.

References

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral networks and deep locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2014.

Capelle, K. A bird's-eye view of density-functional theory. *arXiv preprint arXiv:0211443*, 2006.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Daylight Chemical Information Systems, I. Smiles: Simplified molecular input line entry system, 2019.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Conference on Neural Information Processing Systems (NIPS)*, 2016.

DiMasi, J., Hansen, R. W., and Grabowski, H. G. The price of innovation: new estimates of drug development costs. *Journal of Health Economics 22 151–185*, 2018.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gomez-Bombarelli, R., Hirzel, T., and Al an Aspuru-Guzik, R. P. A. Convolutional networks on graphs for learning molecular fingerprints. *Conference on Neural Information Processing Systems (NIPS)*, 2015.

Fey, M. and Lenssen, J. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. *International Conference on Machine Learning (ICLR)*, 2017.

Goh, G. B. and N. Hodas, C. Siegel, A. V. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties.

Hamilton, W. L., Ying, R., and Leskovec, J. Inductive representation learning on large graphs. *Conference on Neural Information Processing Systems (NIPS)*, 2017.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information and maximization. *International Conference on Machine Learning (ICLR)*, 2018.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Chemical Information and Modeling DOI: 10.1021/ci100050t*, 2016.

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980v9*, 2014.

Li, Y., Zemel, R., Brockschmidt, M., and Tarlow, D. Gated graph sequence neural networks. *International Conference on Learning Representations (ICLR)*, 2016.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Pytorch, team pytorch. 2017.

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

Ramakrishnan, R., Dral, P. O., Rupp, M., Anatole, O., and Lilienfeld, V. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data 1, 140022 (2014) doi:10.1038/sdata.2014.22*, 2014.

Rogers, D. and Hahn, M. Convolutional networks on graphs for learning molecular fingerprints. *Journal of Chemical Information and Modeling DOI: 10.1021/ci100050t*, 2010.

Ruddigkeit, L., van Deursen, R., and L. C. Blum, J.-L. R. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *J. Chem. Inf. Model. 2012, 52, 11, 2864-2875*, 2012.

Schutt, K. T., Arbabzadah, F., Chmiela, S., Muller, K. R., and Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *arXiv preprint arXiv:1609.08259*, 2016.

Sun, F., Hoffmann, J., Verma, V., and Tang, J. Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000v3*, 2020.

Vinyals, O. and S. Bengio, M. K. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391v4*, 2015.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences 28 (1), 31-36, DOI: 10.1021/ci00057a005*, 1988.

Wu, Z., Ramsundar, B., Feinberg, E. N., Vishnu, A., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: A benchmark for molecular machine learning. *arXiv preprint arXiv:1703.00564v3*, 2018.

Wu, Z., Pan, S., Chen, F., and Long, G. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596v4*, 2019.

6. Appendix

6.1. Model Configuration

Our SMILES-based text encoding is composed of 19 different characters for QM9, and the largest molecule size (i.e 26) is the maximum length of characters fed to the LSTM. In the case of the graph encoding, the input representation at the node level consists of different atomic properties such as the atomic nuclear charge, the hybridization state, and the features at the edge level consists of the number of bonds.

We use Infograph and the PyTorch Geometric QM9 intialization implementations as basis for constructing our model. For all of the tasks, we first standardize the target values using Scikit-learn StandardScaler (Pedregosa et al., 2011), so that all targets have a mean of zero and unit variance.

All of our models are trained with the Adam optimizer (Kingma & Ba, 2014) for 10 epochs, at a constant learning rate of 1e-3, and a batch of size 64. The performance of our models are assessed on a molecular property prediction task with 16000 molecules (due to resource contsraints), where we use the embeddings learned on the completely unsupervised setting. We train a kernel ridge regression classifier with 50% of the samples, select a Laplacian kernel, and set alpha to 10^{-3}. For each model, we evaluate the performance on 25% of the data and report the mean absolute error on the test set (i.e remaining samples).