Deep Molecular Graph InfoMax

Boury Mbodj [1] **Supervisor: Prof. William Hamilton** [1]

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. By utilizing mutual information maximization between global level representations of molecular graphs and auxillary molecular descriptors, we demonstrate that the embeddings learned by our model, outperforms the state-of-the art baseline on a property prediction task. Similarly to Deep InfoMax, our approach provides flexible objectives for better generalization.

1. Introduction

Traditional drug development is a complex, costly, and lengthy process that typically spans more than a decade. 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 representation learning models (Kristof T. Schutt, 2016; Joan Bruna, 2014; Justin Gilmer, 2017; Franco Scarselli, 2009; David Duvenaud, 2015), which satisfy the rotational, translational and permutational invariances of molecules naturally, to property prediction tasks. Indeed, representation learning (Yoshua Bengio, 2012) has shown promise to address the shortcomings of conventional quantum mechanical simulations methods such as Density Function Theory (Capelle, 2006), used in both biological and materials science.

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.

(Fan-Yun Sun, 2020) proposed an unsupervised graph-level representation learning model termed InfoGraph. Motivated by the impressive results of Deep InfoMax (R Devon Hjelm, 2019), this model utilizes mutual information 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 SMILES based (Sepp Hochreiter) representations. Our contributions can be summarized as follows:

- We introduce Deep Molecular Graph InfoMax, a flexible unsupervised graph representation learning approach.
- We propose learning efficient molecular representation via mutual information maximization between heterogenous molecular descriptors.
- We show that perfoming global-global information maximization can significantly enhance graph-level embeddings and surpass the state of the art baseline on a property prediction task .

2. Related Work

Our work builds upon recent research based on mutual information maximization and graph representation learning. There are numerous papers on graph neural networks (Franco Scarselli, 2009; Aditya Grover, 2016; William L. Hamilton, 2017; Keyulu Xu, 2019) that demonstrate their ability to learn complex representations of data.

Deep InfoMax (DIM) Mutual information I, is a measure of mutual dependence between two random variables X and Y. It is equivalent to the Kullback Leibler divergence between the joint probability $P(X, Y)$ and the product of marginals $P(X)P(Y)$ (Mohamed Ishmael Belghazi, 2018).

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

COMP396: Undergraduate Research Project

Code: https://github.com/bmbodj/COMP396/tree/master/Winter_2020

$$I(X;Y) = \sum P(x,y) log \frac{P(x,y)}{p(x)P(y)} \quad (1)$$

Founded on the principle of mutual information maximization and self-supervision, Deep InfoMax (DIM) is an approach that learns useful representations by learning a predictive model of both global and local features. (R Devon Hjelm, 2019) proposed three objective functions where mutual information maximization is applied between input data and learned high-level representations. Applied to an image recognition task in the scope of their paper, DIM can also prioritize either global or local information sharing for different applications.

Deep Graph InfoMax (DGI) It is the first general published work on deep information maximization for graphs.Indeed, DGI is an unsupervised node-level learning approach, and it relies on maximizing mutual information between patch representations and corresponding high-level summaries of graphs to obtain meaningful node embeddings. It is based on a noise contrastive learning approach with a binary cross-entropy objective function and has shown competitive performance on both transductive and inductive learning tasks.

InfoGraph Following Deep InfoMax (R Devon Hjelm, 2019) and Deep Graph InfoMax (Petar Velickovic, 2018), (Fan-Yun Sun, 2020), presented another graph representational learning approach. InfoGraph maximizes the mutual information between the graph-level representation and the representations of substructures of different scales. The authors also proposed InfoGraph*, an extension of InfoGraph for semi-supervised scenarios. InfoGraph thus differs from DGI in architecture. The pooling operation, also known as readout, aims to reduce the size of parameters by downsampling the nodes to generate smaller representations and thus avoid overfitting as well as permutation invariance. (Zonghan Wu, 2019).
In contrast to InfoGraph, our method not only involves focusing on the global structure of the representation, but it also incorporates representations of auxiliary molecular descriptors.

3. Proposed Method

In this section, we formulate the problem of learning representations for molecular graphs as an information maximizing problem that we dissect into three parts: global-global mutual information maximization, local-global mutual information maximization, and Deep Molecular Graph InfoMax.

3.1. Objective Function and Motivation

We select the renown Jensen Shannon Divergence (JSD) estimator as the main objective function of our model. Based on Jensen's inequality and the Shannon entropy, this divergence measure quantifies how much one probability distribution differs from another probability distribution. It is a symmetrization of the Kullback Leibler Divergence D_{KL}. For two probability distributions P and Q, the Jensen Shannon Divergence D_{JS} can be defined as :

$$\begin{aligned} &D_{JS}(q||p) = \tfrac{1}{2} D_{KL}(P||\tfrac{1}{2}(P+Q)) \\ &\qquad +1/2 D_{KL}(q||\tfrac{1}{2}(P+Q)) \\ &1. D_{JS}(q||p) = D_{JS}(p||q) \\ &2. D_{JS}(q||q) = 0 \\ &3. D_{JS}(q||p) >= 0 \end{aligned} \quad (2)$$

Three important characteristics of this measure of distance are its symmetric property, convergence, and non-negativity. The generic formula of the JSD we will be following throughout the rest of the paper is :

$$\begin{aligned} max I^{JSD}(X;Y) &:= max D_{JS}(P(X,Y)||P(X)P(Y)) \\ &= max(2log2 + E_{p(x,y)}[-sp(-T(x,y)] \\ &- E_{p(x)p(y)}[sp(T(x^{'},y)] \\ &= max(2log2 + E_{p(x,y)}[log\sigma(T(x,y)] \\ &- E_{p(x)p(y)}[log\sigma(-T(x^{'},y)] \\ &where -> sp(x) = log(1+e^{x}) \\ &\quad -> \sigma(x) = 1\tfrac{1}{1+e^{-x}} \end{aligned} \quad (3)$$

We maximize the lower bound on the JSD the divergence, where T is the discriminator x is an input sampled from P, $x^{'}$ is the negative input pair and sp is the softplus function which is the integral of the sigmoid activation function σ.

3.2. Encoders

SMILES 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 on neural networks such as RNNs (Garre B. Goh) (Zhenqin Wu, 2018) . Thus the main advantages of using SMILES based molecular descriptors instead of solely graph ones, are that:

- the SMILES representation is a basis for interpretability for numerous cheminformatics software packages.
- its use enables a larger chemical compound space due to the higher presence of chemical databases containing millions of molecules represented in this format.
- applying RNNs to SMILES strings can be computationally more efficient than performing training on more sophisticated models such as graphs.

In order to enforce permutational invariance and uniqueness, we use the canonicalized representation of SMILES as a 1-hot-encoding input to a Long Short term memory. Furthermore, to test the effectiveness of the SMILES with the LSTM model, we perform experiments on a supervised learning setting and display the results on Figures (4) and (5).

Graph Model Encoder As in InfoGraph, we choose the Message Passing Neural Networks (MPNNs) (enn-s2) introduced by (Justin Gilmer, 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 (4) and (5) a readout phase (6). The general messaging phase is defined by the following formula:

$$m_v^{t+1} = \sum_{w \epsilon N(v)} M_t(h_v^t, h_w^t, e_{vw}) \tag{4}$$

$$h_v^i = U_i(h_v^{i-1}, m_v^{t+1}) \tag{5}$$

Where the message m_v^{t+1} is the transition function that propagates information and h_v^{t+1} 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.

$$\hat{y} = R(\{h_v^t | v \epsilon G\}). \tag{6}$$

As we mentioned above, we select the (enn-s2s) MPNN variation, which has achieved state-of-the-art results with the continuous edge network message function, a gated recurrent unit (Junyoung Chung, 2014) update function, and a Set2Set (Oriol Vinyals, 2015) readout operator. The power of this encoder can be attributed to the inclusion of the edge attributes, which is an essential feature for molecules.

3.3. Notation and Preliminaries

Given a dataset of multiple different graphs $G = \{G_1, G_2, ...\}$ which represent molecules, our objective is to learn meaningful representations, by utilizing the gradients from a discriminator to help train the encoder network. The obtained representation can be used for downstream classification or regression tasks such as molecular property prediction.

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 in the graph. Finally, the feature vector for the whole SMILES based descriptor and entire graph is expressed as Y.

3.4. Global-Global Mutual Information Maximization

We exemplify our model on the figure (1) below and write our objective function for our single estimator pair as :

$$max_{\phi\omega_1} I_{\phi\omega_1}(Y_{1\phi}(G); Y_{2\phi}(S)) \tag{7}$$

Let I denote the mutual information between the global representation $Y1$ and $Y2$ after applying the readout operations of the graph and LSTM encoders, respectively.

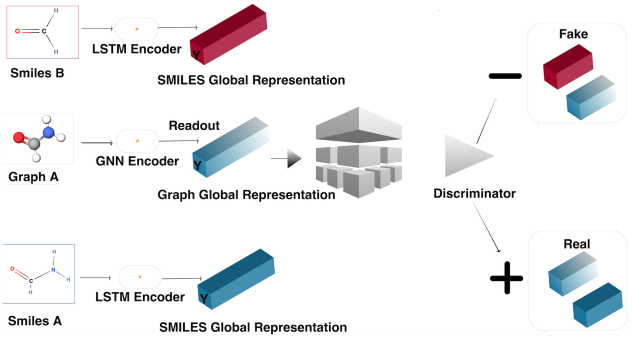

Figure 1. Global-DMGI(G-S): Molecule A (Formamide) is encoded into SMILES Y(S) and Graph Y(G) global representations, and represent a positive pair sample. The discriminator is trained to distinguish the positive real pair from the negative/fake pair consisting of Graph A and SMILES B (Formaldehyde).

Following the DIM and f-GAN (Sebastian Nowozin, 2018) notation, the general f-divergence formulation of the JSD mutual information estimator can be expressed as:

$$\begin{aligned} I_{\phi\omega}^{JSD}(Y_\phi(G); Y_\phi(S)) := \\ E_p[-sp(-T_{\phi\omega}(Y_\phi(x); Y_\phi(x)))] \\ -E_{PXP\sim}[sp(T_{\phi\omega}(Y\phi(x^{'}); Y_\phi(x)))] \end{aligned} \tag{8}$$

Our model learns useful representations by training the discriminator T to estimate the JSD divergence, then training the encoder to minimize this estimate.

For all of our models we generate negative samples by permuting the batch. Our approach differs to that of DIM in that our global DMGI estimator is based on passing both high-level feature vectors Y outputted from the readout function, rather than maximizing information between the graph feature vector Y and its summary feature map H which corresponds to the sum of its local patched representations, centered around every node hi. This can be analogous to a reconstruction task that ensures the preservation of detailed structures that are important to this task.

In order to measure the robustness of our approach, we also investigate the use of global-global mutual information maximization with two graph global representations.

3.5. Local-Global Mutual Information Maximization

The local-global mutual information maximization approach illustrated on figure(2) is equivalent to InfoGraph and DIM methodology to some extent. I designates the mutual information between the patch representation hi centered at an arbitrary node i of G and the global representation Y after applying the readout operation.

$$max_{\phi\omega} \sum_{G\epsilon G} \frac{1}{|G|} I_{\phi\omega}(h_\phi(G); Y_\phi(G)) \tag{9}$$

In this scenario, the mutual information estimator of the JSD can be defined as :

$$\begin{aligned} I^{JSD}_{\phi\omega}(h_\phi(G); Y_\phi(G)) := \\ E_P[-sp(-T_{\phi\omega}(h_\phi(x); Y_\phi(x)))] \\ -E_{PXP\sim}[sp(T_{\phi\omega}(h_\phi(x^{'}); Y_\phi(x)))] \end{aligned} \tag{10}$$

We note that most, if not all existing deep learning work on mutual information maximization, present the local-global mutual information maximization as a stronger, or more suitable objective than global or combinations of global and local objectives.

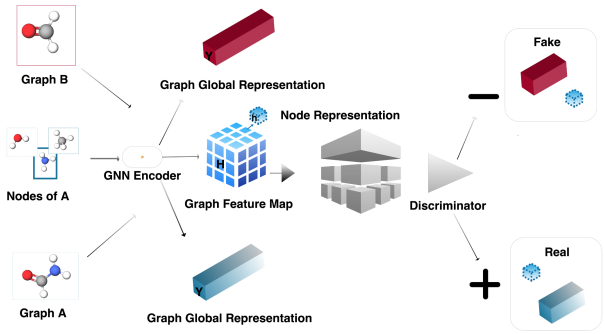

Figure 2. Local-DMGI: Molecule A (Formamide) is transformed into a feature map Hcorresponding to summary of the patch representations every node hi of G. Along with the graph feature map H, the GNN encoder provides the graph A summary vector, computed from its readout operator. A node and global graph embedding of the molecule A represent a positive pair sample. The discriminator is trained to distinguish the positive real pair from the negative/fake pair consisting of graph A and an arbitrary node of B (Formaldehyde) obtained from the permuted batch.

3.6. Deep Molecular Graph InfoMax

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

$$\begin{aligned} &\alpha\, max_{\phi\omega_1} I_{\phi\omega_1}(Y_{1\phi}(G); Y_{2\phi}(G)) \\ &+\beta\, max_{\phi\omega_2} \sum_{G\epsilon G} \frac{1}{|G|} I_{\phi\omega_2}(h_\phi(G); Y_\phi(G)) \end{aligned} \tag{11}$$

where α and β are hyperparameters for the global and local objectives, respectively. We use 0.5 for both α and β in our experiments. Similarly to DIM, integrating the hyperparameters allows us to exercise control over the information, either locally or globally, that we want our model to focus on. However, in contrast to DIM, we do not impose structural constraints on high-level representations with prior matching.

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 Quantum Machine 9 (QM9) dataset (Lars Ruddigkeit, 2012; Raghunathan Ramakrishnan, 2014), as our benchmark. It is composed of 134k molecules with their corresponding equilibrium geometries, frontier orbital eigenvalues, dipole moments, harmonic frequencies, polarizabilities, and thermochemical energetics. All molecules are modeled using Density 8 Functional Theory, and their properties are related to atomization energies, fundamental vibration frequency, states of electrons and measures of spatial distributions of the molecules.
All graph and LSTM based models are implemented using respectively 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. Model Configuration

Our SMILES-based text encoding is composed of 19 different characters for QM9, with the biggest molecule size (25) representing the maximum length of characters fed to the LSTM. In the case of the 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 other features, and the edges are initialized as the bonds.
We use the PyTorch Geometric initialization implementation of the QM9 data, and integrate the one-hot encoding of smiles as part of the data processing. 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 Adam optimizer (Diederik P. Kingma, 2014) and constant learning rate 1e-2. We apply kernel ridge regression with a Laplacian kernel to evaluate our embeddings on 11000 molecules.

4.3. Experiment Evaluation

Table 1 shows the results on the QM9 property prediction task. We observe that both global DMGI objectives out-

Figure 3. DMGI : Deep Molecular Graph InfoMax is a combination of both global-global mutual information maximization **(left)** and local-global Mutual information maximization **(right)** schemes. The influence of both global and local objectives is determined by the choice of hyperparameters α and β.

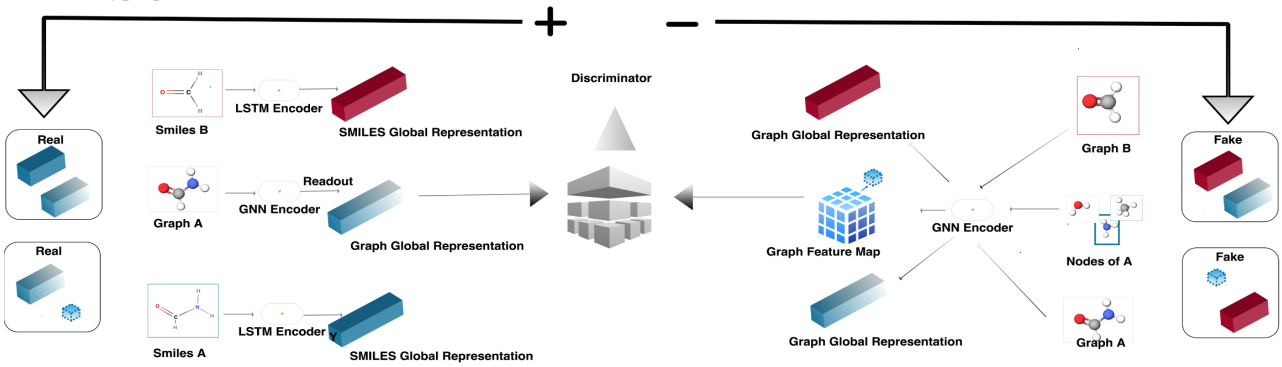

Table 1. Mean absolute error (MAE) accuracy results on QM9 molecular property prediction regression task. GLOBAL-DMGI(G-S) refers to the global mutual information maximization objective based on the combined Graph and SMILES global representation. GLOBAL-DMGI(G-G) denotes the global mutual information maximization objective based on the purely graph global representations.

TARGET	GLOBAL-DMGI(G-S)	LOCAL-DMGI/INFOGRAPH	DMGI	GLOBAL-DMGI(G-G)
MU	0.7226	0.7236	**0.6887**	0.7160
ALPHA	1.6321	1.8344	1.8302	**1.5856**
HOMO	0.0091	0.0079	0.0084	**0.0072**
LUMO	0.0113	0.01197	0.0119	**0.0098**
GAP	**0.0118**	0.0137	0.01176	**0.0118**
R2	**18.9869**	35.3079	33.8895	34.3555
ZVPE	0.0030	0.0037	0.0035	**0.0025**
U0	**7.1710**	9.9567	7.1960	7.7295
U	**7.6604**	9.8407	8.8177	7.9643
H	9.0135	10.17930	8.9315	7.1450
G	7.4572	10.7346	7.5102	**6.5595**
CV	**0.6383**	0.8595	0.8843	0.8214

perform all other unsupervised learning methods. More specifically, the global DMGI (G-G) based on solely Graphs achieves the highest performance on 7 out of 12 target properties. Likewise, our heterogenous global DMGI (S-G) surpasses all other models in 5 out of 12 target properties. We can logically infer that there is more mutual information between the high-level Graph representations of the global DMGI (G-G) as it is based in the same structure.

Interestingly, our global DMGI (G-S) outperforms InfoGraph, despite being a different molecular descriptor. It is a crucial factor in the practical use of machine learning in drug discovery, as it allows us to take advantage of the available resources more efficiently. In this experiment, the standard DMGI objective operates as an average of both local and global DMGI.

As our results are task-specific, this formulation can be more advantageous in more general settings. Our results demonstrate that prioritizing global structure is more suitable in this scenario, as the local propagation scheme might not fully utilize the relational information between nodes. Indeed, it is possible that at training time, edge attributes are discarded or mistaken for 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 datasets, we demonstrate that graph embeddings acquired by our approach outperform the latest state of the art model InfoGraph. Although our experiments are task-specific, they still demonstrate the effectiveness of the introduced framework and prove that it is a solution worth investigating.

References

Aditya Grover, J. L. node2vec: Scalable feature learning for networks. *DOI: http://dx.doi.org/10.1145/2939672.2939754*, 2016.

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

David Duvenaud, Dougal Maclaurin, J. A.-I. R. G.-B. T. H. A. a. A.-G. R. P. A. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292*, 2015.

Diederik P. Kingma, J. L. B. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980v9*, 2014.

Fan-Yun Sun, Jordan Hoffmann, V. V. J. T. Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000v3*, 2020.

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

Franco Scarselli, Marco Gori, A. C. T. M. H.-G. M. The graph neural network model. *IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 20, N*, 2009.

Garre B. Goh, Nathan Hodas, C. S. A. V. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties.

Joan Bruna, Wojciech Zaremba, A. S. Y. L. Spectral networks and deep locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2014.

Junyoung Chung, Caglar Gulcehre, K. C. Y. B. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Justin Gilmer, Samuel S. Schoenholz, P. F. R. O. V. G. E. D. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017.

Keyulu Xu, Weihua Hu, J. L. S. J. How powerful are graph neural networks. *arXiv preprint arXiv:11810.00826*, 2019.

Kristof T. Schutt, Farhad Arbabzadah, S. C. K. R. M. A. T. Quantum-chemical insights from deep tensor neural networks. *arXiv preprint arXiv:1609.08259*, 2016.

Lars Ruddigkeit, Ruud van Deursen, L. C. B. 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.

Mohamed Ishmael Belghazi, Aristide Baratin, S. R. S. O.-Y. B. A. C. R. D. H. Mutual information neural estimation. *arXiv preprint arXiv:1801.04062v4*, 2018.

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

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.

Petar Velickovic, William Fedus, W. L. H. P. L.-Y. B. R. D. H. Deep graph infomax. *arXiv preprint arXiv:1809.10341v2*, 2018.

R Devon Hjelm, Alex Fedorov, S. L.-M. K. G. P. B. A. T. Y. B. Learning deep representations by mutual information and maximization. *arXiv preprint arXiv:1808.06670v5*, 2019.

Raghunathan Ramakrishnan, Pavlo O. Dral, M. R.-O. A. v. L. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data 1, 140022 (2014) doi:10.1038/sdata.2014.22*, 2014.

Sebastian Nowozin, Botond Cseke, R. T. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1801.04062v4*, 2018.

Sepp Hochreiter, J. S. Long short-term memory.

William L. Hamilton, Rex Ying, J. L. Inductive representation learning on large graphs. *arXiv preprint arXiv:706.02216*, 2017.

Yoshua Bengio, Aaron Courville, P. V. Representation learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538v1*, 2012.

Zhenqin Wu, Bharath Ramsundar, E. N. F.-A. V. J. G. C. G. A. S. P.-K. L. V. P. Moleculenet: A benchmark for molecular machine learning. *arXiv preprint arXiv:1703.00564v3*, 2018.

Zonghan Wu, Shirui Pan, F. C. G. L. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596v4*, 2019.

6. Appendix

Figure 4. Epoch Function vs Train loss. To test the effectiveness of the SMILES with the LSTM model, we performed supervised molecular property prediction experiments.

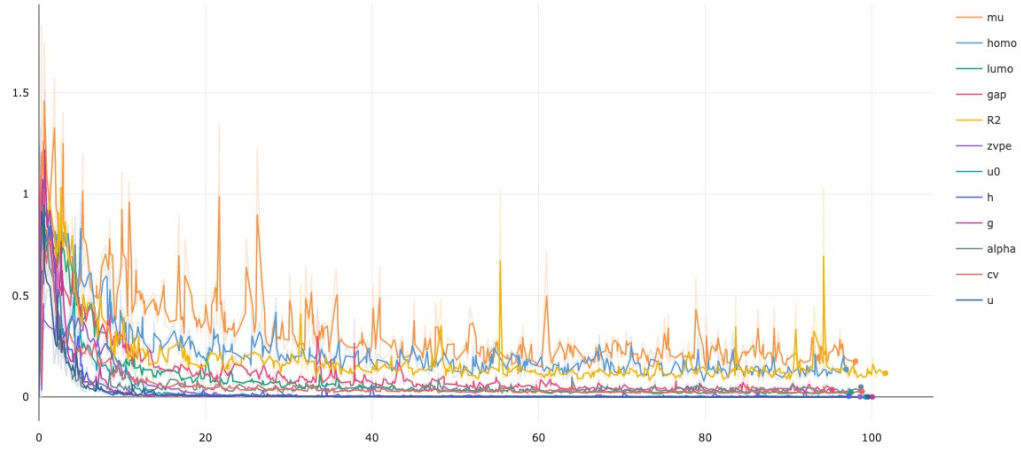

Figure 5. Epoch Function vs Test Error.To test the effectiveness of the SMILES with the LSTM model, we performed supervised molecular property prediction experiments.