Predicting ground-state molecular properties using the quantum-mechanical dataset QM7-X

Austin O. Atsango
atsango@stanford.edu

Fathelrahman Ali
fathali@stanford.edu

Sanjari Srivastava
sanjari4@stanford.edu

Abstract

Developing flexible, reliable, and widely generalizable methods to predict molecular properties at low computational cost is an important goal in computational chemistry. Various machine learning architectures have been developed to achieve this goal, but many training procedures are inflexible or suffer from a low quality of training sets. Here, we use a relatively simple graph convolution network architecture to predict the total and atomization energies for the recently published QM7-X dataset, which contains reliable and tightly converged properties for a comprehensive subset of small organic molecules. Our results show strong correlation between actual and predicted values, with R^2_{dev} values of 0.95 and 0.92 for total and atomization energies respectively.

1 Introduction

The accurate prediction of ground-state molecular properties—such as total energies, atomization energies, and dipole moments—is a key objective in computational chemistry that typically involves conducting computationally costly quantum mechanical calculations [1]. One way to alleviate the associated computational cost is to train a supervised machine learning model to predict desired molecular properties. However, many such attempts are plagued by datasets (such as the GDB datasets [2, 3, 4]) that either contain only a small subset of the relevant chemical space, have unreliable data, or lack vital information that would be useful for the full deployment of machine learning approaches such as graph convolution networks [5].

Here, we present the results of a scheme for predicting total and atomization energies for an extensive subset of small organic molecules using graph convolution networks(GCNs). We utilize the recently published quantum mechanical dataset QM7-X [5], which contains 42 chemical properties of molecules with up to seven non-hydrogen atoms and is freely available online. QM7-X improves on earlier benchmark datasets such as QM9 [6] and ANI-1x [7] by providing a systematic, extensive, and tightly converged set of physical properties mostly computed using the rigorous hybrid density-functional PBE0 with MBD dispersion [5]. Additionally, QM7-X circumvents the issues of earlier datasets, e.g. by providing the chemical structure for each molecule, including structural isomers, stereoisomers, and several nonequilibrium structures.

2 Related Work

The earliest machine learning approaches to predicting molecular properties focused on engineering inputs/features to encode molecular symmetries such as translational and rotational invariance. These approaches ranged from the use of radial and angular symmetry functions by Behler and Parrinello [8] to the so-called Bag of Bonds model used by Hansen et al [9]. They involved feeding hand-designed

CS230: Deep Learning, Fall 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

features as inputs to either high-dimensional neural networks or kernel methods and were applied successfully to the prediction of molecular forces and energies. As large-scale molecular datasets became widespread, more flexible architectures that modeled the invariances inherent in molecular systems became desirable. One attractive candidate was convolutional neural networks (CNNs), which were pursued because of their translational equivariance. A particularly successful architecture in this domain was SchNet, a model that combined recurring embedding and interaction layers with continuous-filter convolutional layers to deliver rotationally invariant energy predictions with state-of-the-art (SOTA) accuracy [10]. However, because CNNs do not naturally contain all molecular symmetries, SchNet still utilized extensive feature optimization through multiple atom-wise and interaction blocks as well as radial basis functions to assert rotational invariance.

Due to the isomorphism between molecules and graphs, graph neural networks, generally classified as Message-Passing Neural Networks (MPNNs) [11], have been pursued for molecular property prediction. In these models, atoms are encoded as nodes while interatomic connections (e.g. molecular bonds) are encoded as edges. One can more easily incorporate into MPNNs the translational and rotational symmetries usually observed in molecules. Various implementations of MPNNs have previously been applied to the prediction of crystal structure energies [12], chemical reactivity [13], and partial charges [14]. MPNN models such as equivariant graph neural networks have been used to predict molecular properties in the QM9 dataset, achieving SOTA accuracy [11, 15]. To our knowledge, MPNNs are yet to be leveraged to predict properties in the recently published QM7-X dataset. Here, we design an architecture based on a message-passing implementation of GCNs and apply it to the prediction of total and atomization energies in the QM7-X dataset.

3 Dataset and Features

The QM7-X dataset contains molecular data for 4.2 million equilibrium and non-equilibrium structures. To collate the data, we utilized the qchem library [16]. Since the QM7-X dataset contains 100 structures per molecule, we first filtered the data to find the optimized (lowest energy) structures for each molecule. This eliminated redundancy and improved the efficiency of sampling by condensing 4.2 million structures into 42,000 representative structures, one for each molecule. This filtering step also enabled us to obtain a computationally tractable subset of the dataset that did not overwhelm computer memory. We then encoded molecular graphs using the parallel schemes described below. Three representative molecules in the dataset are shown in Fig. 1, and the preprocessing step is illustrated in the first panel of Fig. 2.

3.1 Scheme A

The first scheme utilizes a framework that explicitly accounts for atomic positions in the molecular graph. To achieve this, atoms are encoded as nodes with only 4 features: the atomic number and the corresponding X, Y, and Z coordinates. The graph object is also fed a node positions matrix built from interatomic distances. All inter-atomic pairs are encoded as bidirectional edges between their corresponding atoms, and there are no edge features.

3.2 Scheme B

For the second encoding scheme, we converted the filtered structures into SMILES (simplified molecular-input line-entry system) strings, which were subsequently fed into the cheminformatics package RDKit [17] to extract features and labels. We used an initialization scheme inspired by the Open Graph Benchmark (OGB) [18] that utilizes the pytorch geometric package [19] to build molecular graphs. In this scheme, atoms are encoded as nodes with 9 features which include the atomic number, chirality, the number of chemical bonds, the formal charge, the number of attached hydrogen atoms, the number of radical electrons, hybridization, aromaticity, and whether or not the atom exists in a ring. Molecular bonds are encoded as bidirectional edges connecting their corresponding atoms. Finally, graph edges are given three features (bond type, conjugation status, and bond stereochemistry) corresponding to their respective bond. Atomic positions are not explicitly used in modeling.

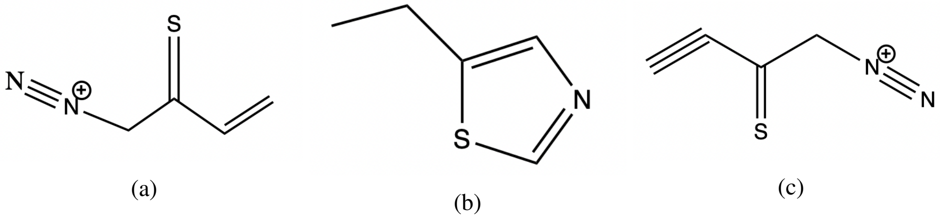


Figure 1: Sample molecules from the QM7-X dataset

4 Methods

Our model is summarized in Fig. 2. This work utilizes graph convolutional networks (GCNs), which were initially built as a robust approach for semi-supervised learning on graph-structured data [20]. The GCN scheme exploits a first-order approximation of localized spectral filters on graphs [21, 22] to derive the following propagation rule across layers:

$$h^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}h^{l}W^{l}) \tag{1}$$

Where σ is an activation function, W^l is a layer-specific weight matrix, h^l are the node embeddings for a layer l, $\tilde{A} = A + I_N$ is an $n \times n$ (n = number of nodes) adjacency matrix ($A_{ij} = 1$ if i and j are connected by an edge, else 0) with added self connections (I_N is the identity matrix), and $\tilde{D}$ is the augmented degree matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. Here, $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is derived from the graph Laplacian, and is a mathematical operator that generalizes traditional image convolution to arbitrary molecular graphs [11]. After GCN propagation, a single vector is obtained per graph using a readout layer that takes a simple mean of the node embeddings:

$$x_G = \frac{1}{|V|}\sum_{v \in V} h_v^{(L)} \tag{2}$$

The graph vector is then passed through a single-layer neural network for final target prediction.

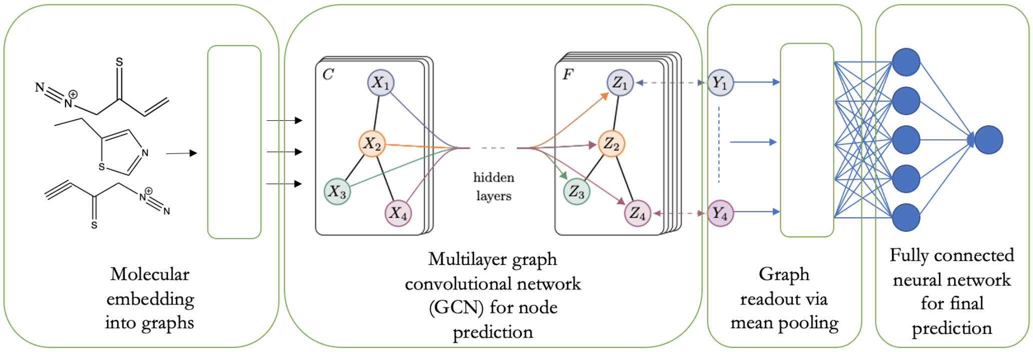

S
N
N
S
N
S
N
N
S
C
X_1
X_2
X_3
X_4
hidden layers
F
Z_1
Z_2
Z_3
Z_4
Y_1
Y_4
Molecular embedding into graphs
Multilayer graph convolutional network (GCN) for node prediction
Graph readout via mean pooling
Fully connected neural network for final prediction

Figure 2: Model architecture. The GCN model was first introduced in [20] and is borrowed from there.

5 Results and Discussion

Our procedure involved training various iterations of the GCN model described in the Methods section to predict for total and atomization energies using the encoding schemes A and B. For both schemes, we used a mean squared error (MSE) cost function and an Adam optimizer for numerical efficiency. We trained on 41,254 molecular structures with a 90/10 train/validation split. The train/dev split was chosen to maximize the size of the training set while simultaneously leaving a substantial number of structures (4000) in the validation set. The labels (total and atomization energies) were 'normalized' by subtracting their respective standard deviation and dividing by the mean in order to make the parameter space more favorable for optimization. This normalization procedure was observed to improve prediction accuracy. The results for both schemes are detailed in the subsections below.

5.1 Scheme A

Here, we trained a baseline model containing 3 GCN layers with 10 hidden channels each. An Adam optimizer with a learning rate of 0.002 was observed to have the best performance. Hyperparameter exploration revealed that this scheme was consistently outperformed by the more detailed Scheme B; a sample of the results from this model is shown in Fig.4a in the Appendix. Due to this gap in performance, all further experiments were conducted with Scheme B, which relies on SMILES strings to extract more descriptive features.

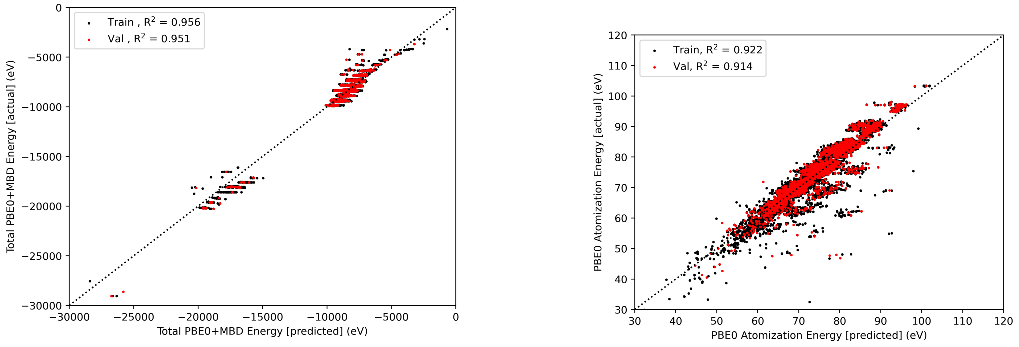
5.2 Scheme B

We conducted a hyperparameter search by varying the number of GCN layers and the number of hidden units per layer. For all architectures, we used a learning rate α =0.005 and trained for 100 epochs because these settings had performed favorably in preliminary experiments. A sample of the results for the search is shown in Table 1, with prediction accuracy measured via R^2 values for the training and dev sets.

Index	Number of GCN hidden layers	Number of GCN hidden units	Total Energy R^2_{train}	Total Energy R^2_{dev}	Atomization Energy R^2_{train}	Atomization Energy R^2_{dev}
1	3	5	0.9246	0.9186	0.7778	0.7606
2	3	10	0.9288	0.9221	0.9045	0.8977
3	3	15	0.9559	0.9512	0.8185	0.8042
4	3	20	0.9414	0.9329	0.8489	0.8410
				⋮		
11	5	15	0.9261	0.9193	0.9090	0.9003
12	5	20	0.9269	0.9211	0.9157	0.9064
13	6	5	0.9110	0.9041	0.7555	0.7414
14	6	10	0.9265	0.9076	0.8818	0.8735
15	6	15	0.9299	0.9231	0.9040	0.8992
16	6	20	0.9271	0.9200	0.9221	0.9143
				⋮		

Table 1: A sample of R^2 values obtained from hyperparameter exploration. The best observed models for total energy and atomization Energy are colored red and blue respectively

Table 1 shows that the GCNs applied here have relatively low variance, as demonstrated by the fact that R^2_{train} and R^2_{dev} are usually separated by only 0.01 and as such, overfitting is not a concern. We also observe that while the prediction of total energy is most accurate with a relatively low number (3, marked red in Table 1) of GCN layers, the prediction of atomization energy is most accurate when there are more (6, marked blue in Table 1) GCN layers. The predictions for both energy values are more accurate with a relatively large number of hidden units (15 for total energy and 20 for atomization energy), which is higher than the number of input channels in both cases (9 features per node). The best-performing architectures for total and atomization energies respectively were isolated for further hyperparameter tuning, but this did not lead to appreciable increases in accuracy. Ultimately, our models perform better on predicting the total energy (R$^2_{train}$, R$^2_{dev}$ = 0.956, 0.951 respectively) than they do predicting the atomization energy (R^2_{train}, R^2_{dev} = 0.927, 0.920 respectively). Correlation plots for the best-performing architectures are shown in Figures 3a and 3b.



(a)

(b)

Figure 3: Comparison of the actual and predicted total energies (left) and atomization energies (right) for the training and validation set, as well as the accuracy as represented by the R^2 values

The observed R^2 values correspond to mean absolute errors (MAE) of 1.03% and 2.17% for total and atomization energies respectively. Although there is strong correlation between the predicted and actual energy values, our observed accuracy is considerably lower than that of SOTA models such as SchNet, which usually report mean absolute errors in the milli-electronvolt range (< 0.002% MAE) [10, 11]. This discrepancy in accuracy can be attributed to the high complexity of molecular encoding schemes used in SOTA models, which typically utilize specially designed basis functions and/or elaborate recurrent networks to optimize internal molecular representations [11, 15]. Our data suggests that we might benefit from a more complex molecular encoding scheme. The stratified nature of the correlation plots in Figures 3a and 3b shows that the GCN model predicts similar energies for many distinct molecules, suggesting that our molecular encoding scheme might not be sophisticated enough to adequately capture subtle differences between input molecules.

The observed high accuracy of SOTA models might be due partly to overfitting. Figure 5a in the Appendix shows the results obtained when a SchNet model trained on the QM9 dataset is used to predict total energies in the QM7-X dataset. Even though the SchNet model yields accurate predictions for a majority of the molecules, it exhibits poor performance for molecules with energies below -10,000 eV. This demonstrates an inability of the SchNet model to generalize beyond a certain energy range.

The results obtained from our GCN model are promising given its simplicity and the fact that it does not utilize atomic positions. They suggest that one way to systematically improve accuracy would be explicitly inlcude atomic positions and adopt a more elaborate molecular encoding scheme.

6 Conclusions and Future Work

Our results demonstrate that GCNs are powerful tools in molecular property prediction, yielding high correlation ($R^2_{ftrain,devg}$ = 0.9559, 0.9512 for total energy and $R^2_{ftrain,devg}$ = 0.9271, 0.9200 for atomization energy) between predicted values and ground truths and low variance despite a relatively simple molecular encoding system. The distinct optimal training schemes—deeper GCNs for atomization energies and shallow GCNs for total energies–highlight the need to tailor GCN architectures to particular target properties, although it would be instructive to investigate whether we can apply multitask or transfer learning. Furthermore, the correlation plots between actual and predicted properties show that the GCN model could benefit from a more elaborate scheme for encoding molecular features. Future work should focus on optimizing the encoding scheme, e.g. by using recurrent neural networks to model atomic positions and their interactions. In this regard, neural fingerprinting architectures [23] are a promising scheme to explore. Our results suggest that employing such a scheme may push our model's performance closer to SOTA.

7 Contributions

- Austin Atsango: Data preprocessing, GCN implementation, hyperparamter search using Scheme B, report preparation.
- Fathelrahman Ali: Data preprocessing, video preparation.
- Sanjari Srivastava: Explored dataset extraction using scripts provided by authors of QM7-X (we finally used qchem instead), implemented and trained the baseline GCN with encoding Scheme A, video preparation.

References

[1] W. Koch and M. C. Holthausen. *A Chemist's Guide to Density Functional Theory*. Weinheim - New York: Wiley - VCH, 2nd edition, 2001.

[2] Jean-Louis Reymond and Mahendra Awale. "Exploring chemical space for drug discovery using the chemical universe database". In: *ACS chemical neuroscience* 3.9 (2012), pp. 649–657.

[3] Lorenz C Blum and Jean-Louis Reymond. "970 million druglike small molecules for virtual screening in the chemical universe database GDB-13". In: *Journal of the American Chemical Society* 131.25 (2009), pp. 8732–8733.

[4] Lars Ruddigkeit et al. "Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17". In: *Journal of chemical information and modeling* 52.11 (2012), pp. 2864–2875.

[5] Johannes Hoja et al. "QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules". In: *Scientific Data* 8.1 (Feb. 2021). DOI: 10.1038/s41597-021-00812-2. URL: https://doi.org/10.1038/s41597-021-00812-2.

[6] Raghunathan Ramakrishnan et al. "Quantum chemistry structures and properties of 134 kilo molecules". In: *Scientific data* 1.1 (2014), pp. 1–7.

[7] Justin S Smith et al. "The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules". In: *Scientific data* 7.1 (2020), pp. 1–10.

[8] Jörg Behler and Michele Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces". In: *Physical review letters* 98.14 (2007), p. 146401.

[9] Katja Hansen et al. "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space". In: *The journal of physical chemistry letters* 6.12 (2015), pp. 2326–2331.

[10] Kristof T Schütt et al. "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions". In: *arXiv preprint arXiv:1706.08566* (2017).

[11] Justin Gilmer et al. "Neural message passing for quantum chemistry". In: *International conference on machine learning*. PMLR. 2017, pp. 1263–1272.

[12] Shubham Pandey et al. "A Graph Neural Network for Predicting Energy and Stability of Known and Hypothetical Crystal Structures". In: (2021).

[13] Connor W Coley et al. "A graph-convolutional neural network model for the prediction of chemical reactivity". In: *Chemical science* 10.2 (2019), pp. 370–377.

[14] Ali Raza et al. "Message passing neural networks for partial charge assignment to metal–organic frameworks". In: *The Journal of Physical Chemistry C* 124.35 (2020), pp. 19070–19082.

[15] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks". In: *arXiv preprint arXiv:2102.09844* (2021).

[16] *qchem*. https://github.com/icanswim/qchem.git. Accessed: 2021-11-01.

[17] Greg Landrum. *RDKit: Open-source cheminformatics*. URL: http://www.rdkit.org.

[18] Weihua Hu et al. "Open Graph Benchmark: Datasets for Machine Learning on Graphs". In: *arXiv preprint arXiv:2005.00687* (2020).

[19] Matthias Fey and Jan E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.

[20] Thomas N. Kipf and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *CoRR* abs/1609.02907 (2016). arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907.

[21] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. “Convolutional neural networks on graphs with fast localized spectral filtering”. In: *Advances in neural information processing systems* 29 (2016), pp. 3844–3852.

[22] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. “Wavelets on graphs via spectral graph theory”. In: *Applied and Computational Harmonic Analysis* 30.2 (2011), pp. 129–150.

[23] Peter C St. John et al. “Message-passing neural networks for high-throughput polymer screening”. In: *The Journal of chemical physics* 150.23 (2019), p. 234111.