**PREDICTING THE PROPERTIES OF ORGANIC MOLECULES WITH NEURAL NETWORKS**

Background

Organic molecules are a promising building block for material because they can be light, transparent, flexible, and scalable. The attractive organic electronics applications include transparent solar windows, portable energy sources, and foldable displays. However, many technologies do not exist yet due to several issues such as performance, stability, economics, etc. Many of these aspects can be improved if the “right” molecule is selected. Instead of performing expensive and time-consuming calculations and/or experiments, it would be tremendously valuable if the properties could be rapidly predicted based on given molecules as inputs using a machine learning model, accelerating the materials discovery process.

Dataset

The dataset called QM9 was downloaded from quantum-machine.org.1,2 The dataset contains 134K stable small organic molecules (i.e., elements of C, H, O, N, F) with up to 9 heavy atoms.3,4 Several molecular properties are included in this dataset, such as geometries, harmonic frequencies, dipole moments, polarizabilities, enthalpies, and atomization energies. Each entry was calculated with density functional theory using the B3LYP/6-31G(2df,p) level of theory. The QM9 dataset is often used to benchmark machine learning models and study structure-property relationships of organic systems.

Data wrangling and preprocessing.

In the dataset, each molecule has its xyz file where all of the information associated with that particular molecule are included. The 17 properties came as a single line containing (1) tag, (2) index, (3-5) rotational constant A, B, C, (6) mu - dipole moment, (7) alpha - isotropic polarizability, (8) homo - the energy of highest occupied molecular orbital, (9) lumo - the energy of lowest occupied molecular orbital, (10) gap - the difference between homo and lumo, (11) r2 - electronic spatial extent, (12) zpve - zero-point vibrational energy, (13) U0 - internal energy at 0 K, (14) U - internal energy at 298.15 K, (15) H - enthalpy at 298.15 K, (16) G - free energy at 298.15 K, (17) Cv - heat capacity at 298.15 K. A python script was used to extract data and aggregate output into a table called data.out. Because this is a relatively clean operation, the data types were correct, and no missing value was found. However, the open-source cheminformatics code called RDKit could not work with 746 molecules which must be eliminated. The final total number of entries is 133,139 molecules.

Exploratory data analysis (EDA)

The histograms of molecular properties are shown below (Figure 1). It can be observed that most of these properties contain extreme values that must be treated before modeling. Additionally, the scales of these properties vary significantly, and proper normalization is likely required if one wishes to model them.

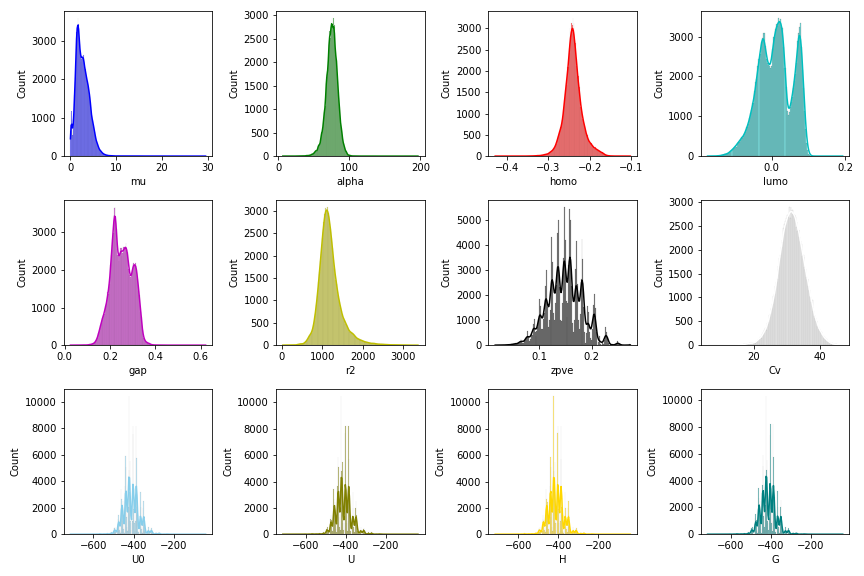


Figure 1 - histograms of molecular properties on the QM9 dataset.

The following procedures were performed to check the structure and property spaces (Figure 2). For structure space, each smile string was converted to molecular fingerprints. Note that the smile string stands simplified molecular-input line-entry system, a line notation that describes a molecule. Subsequently, the molecular fingerprint supplies a molecule for the computer as a bit vector. These fingerprints are used as inputs for principal component analysis (PCA) to obtain the structure space. To get the property space, all of the scaled molecular properties for each molecule were used as inputs for PCA.

For the structure space (left), the 133K molecules appear to be nicely distributed. It can be noticed that there is a region near (0.75, 0.25) where additional samples could be helpful. For property space (right), the properties of these molecules are also distributed nicely. However, a significant amount of data points contain extreme values (top region of this plot).

A picture containing text, plant

Description automatically generatedA picture containing text, screenshot

Description automatically generated

Figure 2 - structure space based on fingerprints (left) and properties space based on molecular properties (right) in this dataset. The dimensionality reductions were performed with principal component analysis (PCA).

Modeling

In this work, the goal is to use the chemical structure to predict molecular properties. Several classes and functions were adapted from an example provided on Keras website. Here, a type of graph neural network called message passing neural network (MPNN) will be used to represent molecules and predict properties.5 An example of the network is shown in Figure 3. It can be observed that MPNN requires atom features, bond features, and pair information fed into the message passing unit. The outputs of the message passing unit and molecule indicator function are used as inputs for transformer encoder and neural networks.

Diagram

Description automatically generated

*Figure 3 - An example of MPNN architecture from Keras.*5

Parameters

For this dataset, the train/test ratio was 80/20. Within the training set, 20% was used as a validation set. For the MPNN, the hidden layers were varied between 1 and 2 layers where each layer could have 128, 256, or 512 nodes (6 models total). The output layer has 12 nodes for molecular properties shown in Figure 1. A total of 50 epochs was used. A more extended training of 100 epochs was also tested to ensure that the validation score remains relatively constant (MPNN-long model). The loss function and metric were based on mean squared error, and used with Adam optimizer. The entire models with their parameters were saved under *the model* directory.

|  |  |  |
| --- | --- | --- |
| **Model** | **Hidden layers** | **Nodes per hidden layer** |
| 1 | 1 | 512 |
| 2 | 1 | 256 |
| 3 | 1 | 128 |
| 4 | 2 | 512 |
| 5 | 2 | 256 |
| 6 | 2 | 128 |

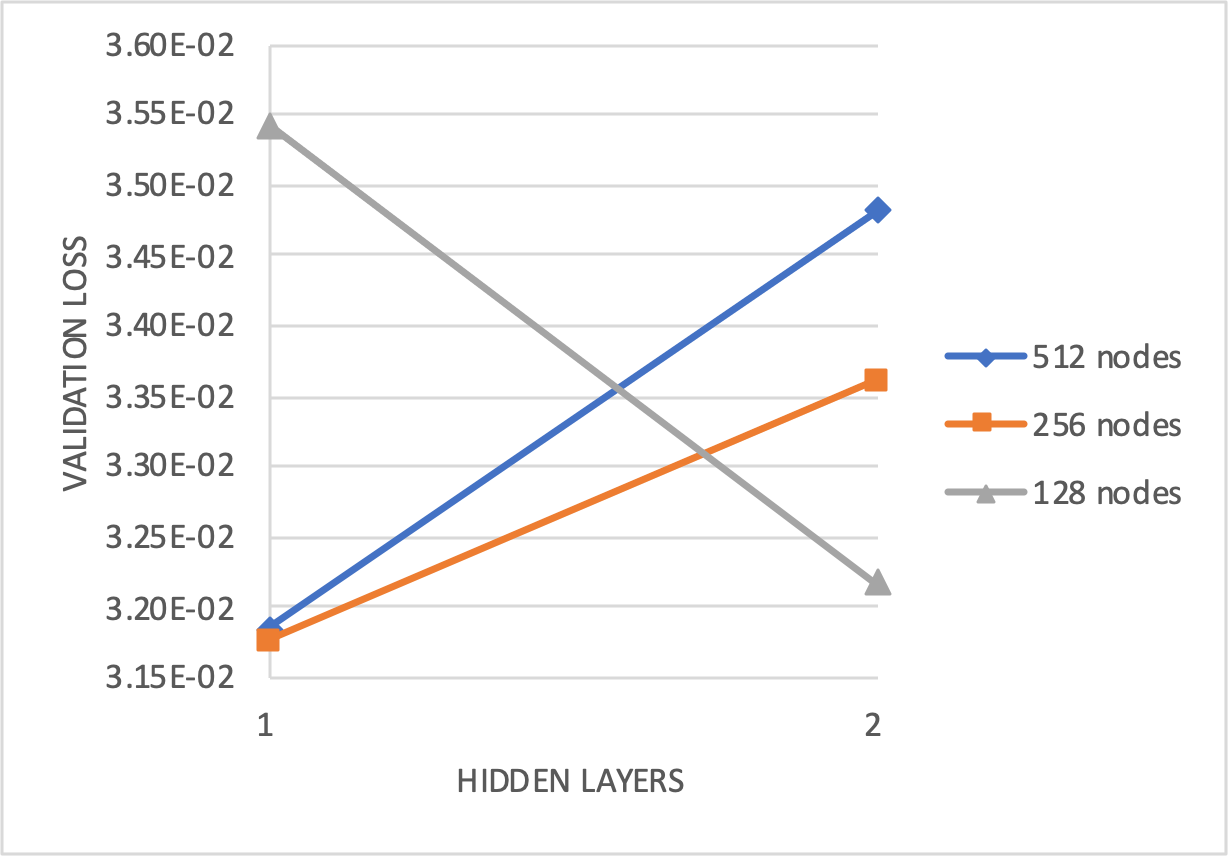
Results

For this model, the 50 epochs were sufficient for training (Figure 4). With cross-validation (Figure 5), it can be observed that an additional hidden layer is undesirable for 256 and 512 nodes. However, an improvement can be observed for 128 nodes, which is reasonable because a small network can benefit from additional parameters from the extra layer. However, the extensive architecture of 256 or 512 nodes might already have sufficient parameters to make good predictions. Thus, additional layers might not improve or even reduce the performance. From this cross-validation, the models with one hidden layer with 256 or 512 nodes have the best architecture (Model 1 & 2).

Chart

Description automatically generated

*Figure 4 - Training history for Model 1 as an example.*



*Figure 5 - Cross validation using validation loss for the 6 models.*

When applying Model 1 to the test set, it can be observed that the MPNN model is effective for properties such as alpha, zpve, U0, U, H, G, and Cv. Other properties such as mu, homo, lumo, gap, and r2 are more difficult to predict. The model has the most difficulty with mu prediction (dipole moment). Dipole moment may be a property that requires information about a molecule at a long range where MPNN might not be able to handle effectively. The coefficient of determination for mu, alpha, homo, lumo, gap, r2, zpve, U0, U, H, G, and Cv are 0.79, 0.99, 0.93, 0.98, 0.97, 0.96, 0.99, 0.99, 0.99, 0.99, 0.99, and 0.99, respectively. The errors obtained from these MPNN are comparable to recent academic literature.6,7

A picture containing building, window

Description automatically generated

*Figure 6 - Correlation of true versus predicted values for 12 properties obtained from Model 1.*

Conclusions and recommendations

The exercise confirms that the MPNN can accurately predict molecular properties based on smiles strings as input. The main success came from the coupling of molecular graph representation with neural networks. With this level of performance, the machine learning model can explore a large space of molecules that exist in reality, given that the new molecule falls within the training data distribution. While the performance is considered acceptable, further optimization can be explored, including atom features, bond features, MPNN unit, transformer unit, network architectures, optimizer, and loss functions. To expand the generalizability of MPNN, the model can be retrain with more diverse chemical dataset via transfer learning.

References

1. <http://quantum-machine.org/datasets/>
2. <https://figshare.com/collections/Quantum_chemistry_structures_and_properties_of_134_kilo_molecules/978904>
3. L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, J. Chem. Inf. Model. 52, 2864-2875, 2012.
4. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Scientific Data 1, 140022, 2014.
5. <https://keras.io/examples/graph/mpnn-molecular-graphs/>
6. Montavon, Grégoire, et al. "Machine learning of molecular electronic properties in chemical compound space." New Journal of Physics 15.9 (2013): 095003.
7. Gilmer, Justin, et al. "Neural message passing for quantum chemistry." International conference on machine learning. PMLR, 2017.