JOURNAL OF
EMERGING INVESTIGATORS

Article

Prediction of molecular energy using Coulomb matrix and Graph Neural Network

Rishi Hazra[1], Prasenjit Hazra[1]
[1] Skyline High School, Sammamish, Washington

SUMMARY

Molecular energy is used to estimate the overall favorability of the formation of a molecule, which can be measured either directly or indirectly. However, both methods require a chemical reaction. Molecular energy has also been investigated computationally. Yet, these calculations require a lot of expensive computer resources. The purpose of this project is to apply machine learning (ML) techniques to create models from known molecular properties. Once the model is created, the molecular properties of new molecules can be predicted quickly. Feedforward Neural Network (FNN) models were trained with two different representations of the molecules. In the first model, the molecules were represented by the Coulombic force between each atom; in the second model, the molecular structure, the bonding between atoms, and the electron configuration of each atom were provided as inputs. Prediction errors of the two models were compared statistically to determine the better representation of the molecule. The FNN network was able to predict the molecular energy with low errors in both models. Though, the model with the Coulomb matrix performed better for molecules that do not contain high electronegative atoms and performed poorly for molecules with three or more fluorine atoms. The effect of fluorine (or other electronegative atoms) on the Coulomb matrix needs to be further investigated.

INTRODUCTION

Molecular energy is useful in understanding the properties of molecules and the behavior of chemical systems (1). Molecular properties can tell us which isomer is favored at equilibrium and what the rate of reaction is (2). Molecular energy can be measured either directly from the chemical reaction or indirectly from other known molecules. However, both direct and indirect methods require the chemical reaction to take place and are also time-consuming and expensive (3). The reaction requires varying lengths of time for completion depending on the chemical kinetics and involves expensive reagents. Molecular energy has also been investigated computationally. There are several methods currently in use: 1) Ab Initio Calculations; 2) Semi-Empirical Calculations; and 3) Density Functional Theory (DFT). Yet, these calculations demand the use of limited computer resources, such as memory and disk space (2).

The purpose of this study is to design a machine learning model, which takes the molecular structure as the input and returns "molecular energy" as the output ("molecular energy" refers to the potential energy of a system of atoms that can be calculated using the Force Field method (1). These models can be created in a few days with cost-effective cloud-based computing resources and prediction can be done in a few seconds.

The Machine Learning (ML) model was trained with known input and output values, and the network was able to build the function (F) to determine the molecular energy of unknown molecules (**Figure 1**). The molecular energy used for training the model was calculated using the Force Field method (1).

Once the model is created, the molecular properties (e.g., the molecular energy, the reaction rate, etc.) of new molecules can be predicted quickly. These models can be used in computational design and will assist in the discovery of new molecules, compounds, and ultimately, drugs.

Feedforward neural networks (FNN) are one of the common deep learning models. The goal of a feedforward network is to approximate some function F. A feedforward network defines a mapping $y = F(x;\theta)$ and learns the value of the parameters, θ, from the known input and output that result in the best function approximation. These models are called feedforward because information flows through the function being evaluated from x, through the intermediate computations used to define f, and finally to the output y. There are no feedback connections in which outputs of the model are fed back into itself.

In this study, FNN models were trained with two different representations of the molecule. In Model A, the molecules were represented by the Coulomb matrix (the Coulombic force between each atom); in Model B, the molecular structure, the bonding between atoms (adjacency matrix), and the electron configuration of each atom (atom vector) were given as inputs. The prediction error of molecular energy from two models

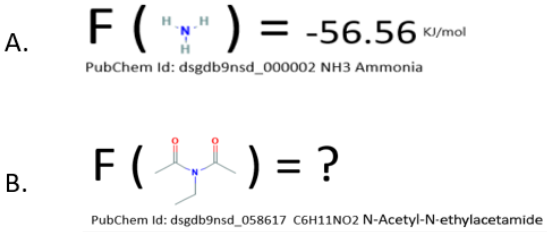


Figure 1: Schematic diagram of the ML model: molecular energy as a function of molecular structure. A) ML model represented by function F is trained with molecular structure as input and known molecular energy as output. B) Function F is used to estimate the molecular energy for a new molecule with unknown molecular energy.

was compared statistically.

We hypothesized that Model B would predict the molecular properties more accurately than Model A, as the input to Model B (valence electron occupancies in atomic orbital) retained the properties of each constituent atom. The Coulomb matrix, as used in Model A, does not consider the individual bonds or type of bonds, but instead assumed electrostatic interactions of each atom in the molecule. With proper design of the network, the model can learn about features like sigma and pi bonding, which is possible in Model B given that our inputs discriminate the presence of electrons in orbitals that are available for sigma (σ) and pi (π) bonding.

We also compared the results of Model B with Graph Neural Network (GNN), which is more suitable for inputs such as the molecular structure of each atom. We hypothesized that Model B and GNN performance will be comparable for predicting molecular energy as both use the electronic configuration but GNN will be more useful for predicting molecular structure as the spatial position of each atom is maintained in the GNN as a connected graph.

RESULTS

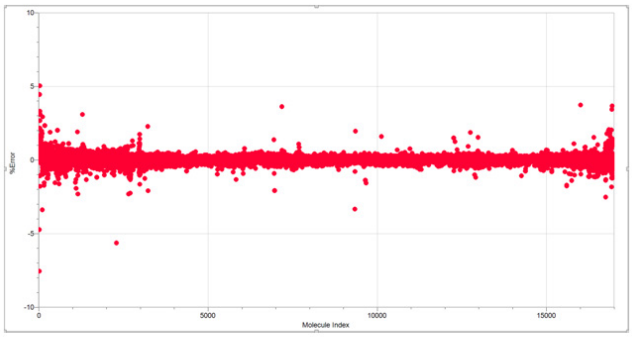
In our experiment, Molecular energy is being predicted using the ML technique using FNN and GNN network. For the FNN network, we used two types of input to represent the molecule: Coulomb matrix and atom vector. The error of each prediction was calculated from the known output. Root Mean Square (RMSE) and Mean Absolute Percent Error (MAPE) were calculated for both models (**Table 1**).

The FNN network was able to predict the molecular energy with low errors for both types of inputs: Coulomb matrices and atom vectors. The FNN with atom vector (Model B) performed better than the FNN with Coulomb matrix (Model A), as we hypothesized. For Model A, the absolute value of the prediction error varies from 0.03 KJ/mol to 30.31 KJ/mol, while the percentage error varies from 0.01% to 29.28%. But the majority of the error lies between +10% to -10% (**Figure 2A**). For Model B, the absolute value of the prediction error varies from 0.03 KJ/mol to 43.08 KJ/mol, while the percentage error varies from 0.01% to 37.22%. But the majority of the error lies between +10% to -10% (**Figure 2B**). The prediction errors using Model A and B for a few sample molecules are given in Tables S1 and S2 (see appendix).

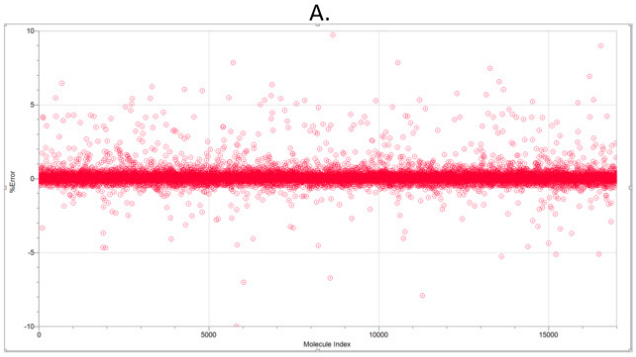
When we compared two different networks for the same set of inputs, the FNN performed better than the GNN. Both RMSE and MAPE are low for the FNN compared to the GNN (**Table 2**). For GNN, the highest prediction error (absolute value) was 90.93 KJ/mol, while the percentage error was as high as 113.9%. But the majority of the error lies between +40% to -40% (**Figure 3, Table S3**).

Model	Input	RMSE	MAPE
A	Coulomb matrix	1.88	0.19
B	Adjacency matrix and atom vector	1.78	0.07

Table 1: Comparison of prediction error with FNN with different inputs. RMSE = Root Mean Square Error, MAPE = Mean Absolute Percent Error.



A.



B.

Figure 2: Scatter plot of %Error with FNN. A) FNN with Coulomb matrix (Model A). The scatter plot is symmetric along the x-axis indicating an equal number of positive (actual > predicted) and negative (actual < predicted) errors. Absolute errors are within the 10% range. B) Scatter plot of %Error for ML test data: FNN with atom vector (Model B). The number of predictions with positive errors (actual > predicted) is more than the number of predictions with negative errors (actual < predicted). Absolute errors are within the 10% range.

Network	RMSE	MAPE
FNN	1.78	0.07
GNN	15.75	2.79

Table 2: Comparison of prediction error (FNN vs. GNN) with adjacency matrix and atom vector input. RMSE = Root Mean Square Error, MAPE = Mean Absolute Percent Error.

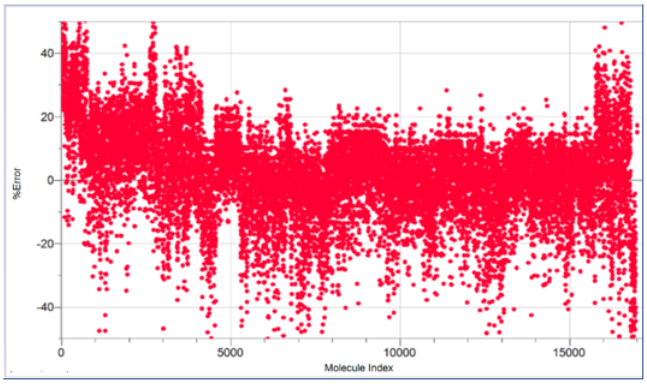


Figure 3: Scatter plot of %Error with GNN. The scatter plot is symmetric along the x-axis indicating an equal number of positive (actual > predicted) and negative (actual < predicted) errors. Absolute errors are within the 40% range.

DISCUSSION

Overall, the FNN network performed better with the atom vector and adjacency matrix compared to the Coulomb matrix. Though, for certain molecules, the Coulomb matrix performed better. For example, FNN with atom vector had higher errors compared to FNN with Coulomb matrix when molecules do not contain fluorine (F) atoms, but we saw large errors with the Coulomb matrix for molecules with three or more fluorine atoms, and the model underestimated the molecular energy (**Figures 4-5).** The top five out of ten atoms with the highest errors are for molecules with three or more fluorine atoms (**Figure 5**). The effect of fluorine (or other electronegative atoms) on the Coulomb matrix needs to be further investigated.

There is another limitation of the Coulomb matrix. One of the premises of using the Coulomb matrix is that no two molecules will have the same Coulomb matrix unless they are identical. But enantiomers will have the same Coulomb matrix, even though the molecules are not chemically identical, can exhibit very different chemical reactivity, and can have different underlying structures. Thus, the model will predict the same energy for the enantiomers, whereas they are not the same (4).

While the GNN results were less accurate than those from the FNN, the GNN is useful to predict the geometry of the molecule. Valence-shell electron-pair repulsion (VSEPR) theory can predict the geometry of the molecule. Yet, this approach fails for systems with lone pairs that have some

Molecular Formula	IUPAC Name	2D Structure	Molecular Energy (Actual) KJ/mol	Molecular Energy (Estimated) KJ/mol	Absolute Error KJ/mol	% Error
C3H2F6	1,1,1,2,3,3-hexafluoropropane		-714.63	-616.67	97.95	13.71
C4H8N2O2	N-[(Z)-3-nitrosobut-2-en-2-yl]hydroxylamine		-417.22	-469.54	52.32	12.54
CH4O	(113C)methanol		-115.73	-72.65	43.08	37.22
C5H6N2O	2-methoxypyrazine		-378.89	-419.16	40.27	10.63
C5H9NO2	N-methyl-3-oxobutanamide		-401.09	-440.78	39.69	9.90
C5H6N2O2	5-methyl-1H-pyrimidine-2,4-dione		-454.08	-484.84	30.75	6.77
C7H7NO	benzamide		-400.79	-429.99	29.20	7.29
C6H7NO	2-aminophenol		-362.72	-390.47	27.75	7.65
C3H5N5O	2,6-diamino-(2,4,6-13C3)1H-1,3,5-triazin-4-o		-466.24	-439.92	26.31	5.64
C5H10N2O2	(3R)-3-aminopyrrolidine-3-carboxylic acid		-456.54	-437.24	19.30	4.23
C3H5N	cyclopropanimine		-172.08	-187.26	15.18	8.82
C6H8N2O	2,3-diaminophenol		-418.08	-432.92	14.84	3.55
C6H7F3	4,4,5-trifluorocyclohexene		-532.37	-518.22	14.15	2.66

Figure 4: Molecules with highest prediction error for FNN with atom vector (Model B). Molecules are sorted by absolute prediction error and top 13 molecules are presented in this figure. The molecules with the highest prediction errors did not necessarily contain fluorine atoms.

Molecular Formula	IUPAC Name	2D Structure	Molecular Energy (Actual) KJ/mol	Molecular Energy (Estimated) KJ/mol	Absolute Error KJ/mol	% Error
C3H2F6	1,1,1,2,3,3-hexafluoropropane		-714.63	-592.29	122.33	17.12
C6H7F3	4,4,5-trifluorocyclohexene		-532.37	-461.89	70.47	13.24
C5H6N2O	2-methoxypyrazine		-378.89	-413.49	34.60	9.13
CH4O	(113C)methanol		-115.73	-81.85	33.88	29.28
C5H9NO2	N-methyl-3-oxobutanamide		-401.09	-434.71	33.62	8.38
C4H8N2O2	N-[(Z)-3-nitrosobut-2-en-2-yl]hydroxylamine		-417.22	-448.85	31.63	7.58
C7H7NO	benzamide		-400.79	-431.11	30.31	7.56
C3H2F3NO2	3,3,3-trifluoro-2-nitroprop-1-ene		-620.28	-592.05	28.23	4.55
C4H7F3O2	(3S)-4,4,4-trifluorobutane-1,3-diol		-606.64	-581.37	25.27	4.17
C4H5F3O2	ethyl 2,2,2-trifluoroacetate		-605.45	-581.62	23.83	3.94

Figure 5: Molecules with highest prediction error for FNN with Coulomb Matrix (Model A). Molecules are sorted by absolute prediction error and top 10 molecules are presented in this figure. Most of the molecules with the highest prediction errors contain fluorine atoms.

other factors, particularly heavy, electronegative atoms. Instead, the exact bond angle and bond length are determined experimentally. Still, this is expensive and time-consuming. We would like to expand our GNN model to predict the molecular geometry – specifically the bond length and bond angle of heavy, electronegative molecules. The GNN will be a good fit for this study as the spatial position of each atom is maintained in the GNN as a connected graph. Looking at the scatter plot for the % error in the GNN (**Figure 3**), it is apparent that non-Gaussian noise is present, which can be improved further by tuning the ML model.

Knowing the molecular energy is essential for the drug discovery method and the traditional methods are time-consuming and computationally intensive. As the prediction error is small (RMSE=1.78 and MAPE=0.07) for the FNN network (Model B), we anticipate this model can be used to predict molecular energy for unknown molecules, which will be important to researchers for drug discovery. Once the model is created with training data, the prediction of an unknown molecule can be done in a few seconds as it was done in this experiment.

MATERIALS AND METHODS

The molecule structure and potential energy of each molecule were downloaded from the Kaggle project, which was originally obtained by scraping the PubChem Database (5,7). Each file contains the properties of molecules, which were made up of hydrogen, carbon, nitrogen, oxygen, fluorine, silicon, phosphorous, sulfur, chlorine, bromine, and iodine. The potential energy obtained from the database was calculated using the Force Field method (1).

FNN:

There are three interconnected hidden layers and one single node as the output layer that represents the molecular

energy (**Figure S1**). The Input layer is fully connected to the first hidden layer.

Model A: Coulomb matrix as the input

The Coulomb matrix describes the features of each molecule. As every molecule has a different number of atoms and the position of each atom is unique in every molecule, each molecule should have a unique Coulomb matrix. The Coulomb matrix for a given molecule is defined by

$$C_{IJ} = \frac{Z_I Z_J}{|R_I - R_J|}, \quad (I \neq J) \qquad C_{IJ} = Z_I^{2.4}, \quad (I = J)$$

where Z_I, Z_J are the electronic charge, and $|R_I - R_J|$ is the distance between two atoms I and J (6). The input layer has an array of 1024 descriptors of the molecule. 900 of these descriptors are from the Coulomb matrix (30x30).

Model B: Adjacency matrix and atom vector as the input

Model B uses the adjacency matrix and atom vector as the input (**Figure S2**). Let us take $C_4H_4N_2$ as an example: The molecular structure is represented as the adjacency matrix in the GNN (1=> Nodes are connected; 0=> Nodes are not connected). The suffix after the atomic symbol (C, N, H) represents the specific atom in the molecule (**Table 3**). The atom vector represents the atomic number of each atom and the number of electrons in each atomic orbital (**Table 4**).

GNN: Predict molecular energy using adjacency matrix and atom vector as the input

In GNN, each edge is represented by a Neural Network Function (f) (**Figure S3**) (6). All nodes get messages from their neighbors passing through the edge (NN function f). At time t + 1, all nodes get messages from neighbors at time t. The message at time t + 1 is given by a function F of node messages at time t and the sum of all the messages received from neighbors. This process was repeated for multiple time instances and stopped after a fixed number of iterations. After this, the results converged. Our network was iterated 10 times

	C1	C2	C3	N1	C4	N2	H1	H2	H3	H4
C1	C1	1	1	0	0	0	1	1	0	0
C2	1	C2	0	1	0	0	0	0	0	0
C3	1	0	C3	0	1	0	0	0	1	1
N1	0	1	0	N1	0	0	0	0	0	0
C4	0	0	1	0	C4	1	0	0	0	0
N2	0	0	0	0	1	N2	0	0	0	0
H1	1	0	0	0	0	0	H1	0	0	0
H2	1	0	0	0	0	0	0	H2	0	0
H3	0	0	1	0	0	0	0	0	H3	0
H4	0	0	1	0	0	0	0	0	0	H4

Table 3: Adjacency matrix of $C_4H_4N_2$.

	Atomic Number	1s	2s	2p	3s	3p	4s	5s	4d
C1	6	2	2	2	0	0	0	0	0
C2	6	2	2	2	0	0	0	0	0
C3	6	2	2	2	0	0	0	0	0
N1	7	2	2	3	0	0	0	0	0
C4	6	2	2	2	0	0	0	0	0
N2	7	2	2	3	0	0	0	0	0
H1	1	1	0	0	0	0	0	0	0
H2	1	1	0	0	0	0	0	0	0
H3	1	1	0	0	0	0	0	0	0
H4	1	1	0	0	0	0	0	0	0

Table 4: Atom vector of $C_4H_4N_2$.

(t = 0, 1,…,9), which is based on the maximum distance (in terms of the number of atoms) between any two atoms. The information stored at each node was used to determine the target attribute (molecular energy).

Nodes represent atoms and edges represent bonds. Each blue oval shape represents an atom, and the suffix represents the position of the specific atom in the molecule as represented in the adjacency matrix (e.g., C, H, N represents carbon, hydrogen, and nitrogen atoms, respectively, and C2 represents the specific carbon atom which is connected to a nitrogen atom and another carbon atom).

Each edge on the graph contains a function f, which was generated by the ML model. Each node sent information to the neighboring node through this function, which varies from edge to edge. v-hat represents the transformation of the information by function f. v at time t+1 represents the information at each node, which is the function F of all the information transformed from the neighboring nodes and that node at time t.

Each node in the molecular structure was represented as a recurrent node in the GNN. Each node was initialized with the atom vector, which has the features of the atom it represents (atomic number and electron configuration).

Machine Learning Training

The models were trained with 80% of the 85,000 molecules (randomly selected) downloaded from the Kaggle Project, and the network was evaluated with the remaining molecules to determine the accuracy of the model using MAPE and RMSE metrics (7). During the training phase, atom vector, adjacency matrix, and molecular energy were used as the input. Once the model was trained, the molecular energy was predicted and the accuracy of different models was compared.

ACKNOWLEDGMENTS

I would like to thank my brother, Neil Hazra, for his help in setting up the Google Cloud. We would also like to thank the scientific reviewers and The Journal of Emerging Investigators Editors for their thoughtful and helpful support while publishing this project.

Received: Dec 31, 2020
Accepted: Sep 23, 2021
Published: Feb 01, 2022

REFERENCES

1. "Force Field (Chemistry)." *Wikipedia: The Free Encyclopedia*, Wikimedia Foundation, 18 Dec. 2019, en.wikipedia.org/wiki/Force_field_(chemistry).
2. Lewars, Errol G. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. 3rd ed., Springer, 2016.
3. Battley, Edwin H. "The advantages and disadvantages of direct and indirect calorimetry." *Thermochimica Acta*, vol. 250, no. 2, 1995, pp. 337 – 352, doi.org/10.1016/0040-6031(94)01963-H.
4. Montavon, Grégoire, *et al*. "Machine learning of molecular electronic properties in chemical compound space." *New Journal of Physics,* vol. 15, no. 9, 2013, pp. 095003., doi.org/10.1088/1367-2630/15/9/095003.
5. Kim, Sunghwan, *et al*. "PubChem in 2021: new data content and improved web interfaces." *Nucleic Acids Res.*, vol. 8, no. 49(D1), 2021, pp. D1388-D1395, doi:10.1093/nar/gkaa971.
6. Gaunt, Alex. "Graph Neural Networks: Variations and Applications." *YouTube*, uploaded by Microsoft Research, Apr 21, 2018, www.youtube.com/watch?v=cWIeTMklzNg.
7. "Potential Energy." Predicting Molecular Properties, *Kaggle*, 29 May 2019, www.kaggle.com/c/champs-scalar-coupling/. Dataset.

APPENDIX

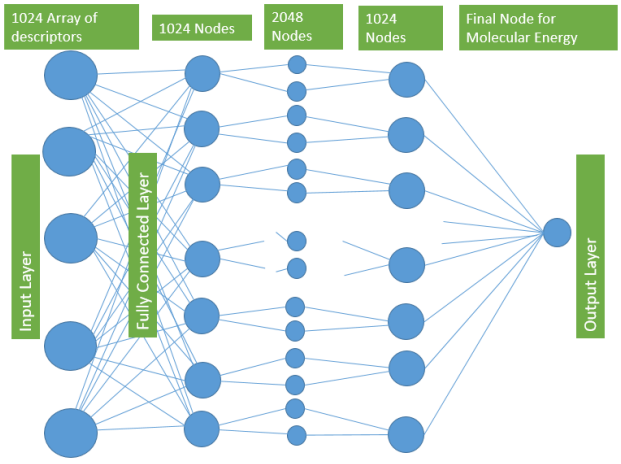


Figure S1: Feedforward Neural Network (FNN). In this network, there are three interconnected hidden layers and one single node as the output layer that represents the molecular energy. The Input layer is fully connected to the first hidden layer.

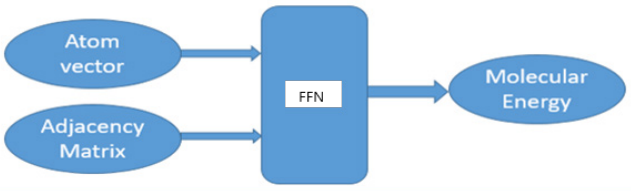


Figure S3: GNN representation of the molecule $C_4H_4N_2$. Each node in the molecular structure is represented as a recurrent node in the GNN. Each recurrent node is initialized with the features of the atom it represents (atom vector).

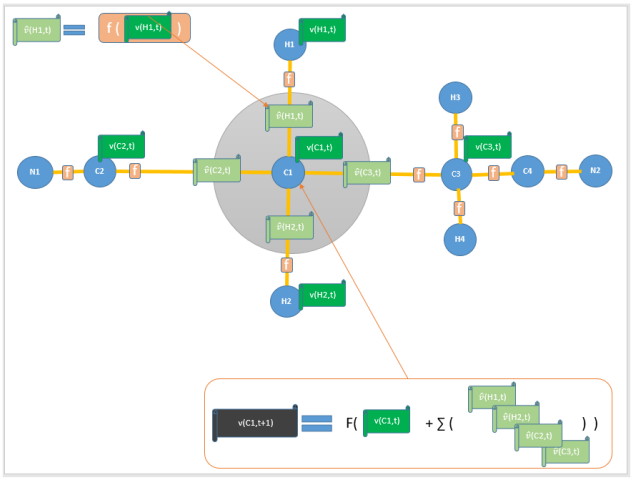


Figure S3: GNN representation of the molecule $C_4H_4N_2$. Each node in the molecular structure is represented as a recurrent node in the GNN. Each recurrent node is initialized with the features of the atom it represents (atom vector).

Molecular Formula	IUPAC Name	Molecular Energy (Actual) KJ/mol	Molecular Energy (Estimated) KJ/mol	Absolute Error KJ/mol	% Error
C6H10N2O	3-cyanopentanamide	-419.40	-418.53	0.87	0.21
C6H6O3	1-(3-hydroxyfuran-2-yl)ethanone	-457.85	-457.64	0.21	0.04
C6H8N2O	2,3-diaminophenol	-418.17	-417.47	0.70	0.17
C6H8O3	propane-1,1,3-tricarbaldehyde	-459.08	-457.86	1.21	0.26
C6H9NO	N-prop-2-enylprop-2-enamide	-363.92	-363.95	0.03	0.01
C6H9NO2	1-nitrocyclohexene	-439.20	-438.46	0.75	0.17
C7H11N	hept-6-enenitrile	-328.03	-328.08	0.04	0.01
C7H12O2	cyclohexanecarboxylic acid	-475.14	-473.58	1.56	0.33
C7H14O	heptan-2-one	-350.40	-350.73	0.33	0.09
C7H14O2	ethyl pentanoate	-425.64	-425.45	0.18	0.04
C7H7NO	benzamide	-400.79	-431.11	30.31	7.56
C7H8O2	2-methoxyphenol	-421.98	-422.66	0.67	0.16
C8H12O	1-cyclohex-3-en-1-ylethanone	-387.22	-387.13	0.09	0.02
C9H14	1-butylcyclopenta-1,3-diene	-351.38	-351.60	0.23	0.06
CH4O	(113C)methanol	-115.73	-81.85	33.88	29.28

Table S1: Predicted molecular energy and associated errors for few sample molecules: FNN with Coulomb matrix - Model A.

Molecular Formula	IUPAC Name	Molecular Energy (Actual) KJ/mol	Molecular Energy (Estimated) KJ/mol	Absolute Error KJ/mol	% Error
C6H10N2O	3-cyanopentanamide	-419.40	-419.37	0.03	0.01
C6H6O3	1-(3-hydroxyfuran-2-yl)ethanone	-457.85	-457.96	0.11	0.02
C6H7F3	4,4,5-trifluorocyclohexene	-532.37	-518.22	14.15	2.66
C6H7NO	2-aminophenol	-362.72	-390.47	27.75	7.65
C6H8N2O	2,3-diaminophenol	-418.08	-432.92	14.84	3.55
C6H8N2O	2,3-diaminophenol	-418.17	-418.13	0.04	0.01
C6H8O3	propane-1,1,3-tricarbaldehyde	-459.08	-459.27	0.20	0.04
C6H9NO	N-prop-2-enylprop-2-enamide	-363.92	-363.79	0.13	0.04
C6H9NO2	1-nitrocyclohexene	-439.20	-439.32	0.11	0.03
C7H11N	hept-6-enenitrile	-328.03	-327.43	0.61	0.18
C7H12O2	cyclohexanecarboxylic acid	-475.14	-475.48	0.34	0.07
C7H14O	heptan-2-one	-350.40	-350.33	0.06	0.02
C7H14O2	ethyl pentanoate	-425.64	-426.08	0.45	0.10
C7H7NO	benzamide	-400.79	-429.99	29.20	7.29
C7H8O2	2-methoxyphenol	-421.98	-421.89	0.10	0.02
C8H12O	1-cyclohex-3-en-1-ylethanone	-387.22	-387.19	0.03	0.01
C9H14	1-butylcyclopenta-1,3-diene	-351.38	-351.30	0.08	0.02
CH4O	(113C)methanol	-115.73	-72.65	43.08	37.22

Table S2: Predicted molecular energy and associated errors for a few sample molecules: FNN with atom vector – Model B.

Molecular Formula	IUPAC Name	Molecular Energy (Actual) KJ/mol	Molecular Energy (Estimated) KJ/mol	Absolute Error KJ/mol	% Error
C2H6	ethane	-79.83	-170.76	90.93	113.90
CH4O	trideuteriomethanol	-115.73	-203.55	87.82	75.88
C3H6O	propan-2-one	-193.17	-252.73	59.56	30.83
C2HNO	formyl cyanide	-206.74	-263.99	57.25	27.69
C2H2O2	ethyne-1,2-diol	-227.83	-282.30	54.47	23.91
C3H4O	cyclopropanone	-191.87	-233.77	41.90	21.84
C3H8O	propan-1-ol	-194.36	-258.65	64.29	33.08
C4H8	but-1-ene	-157.22	-234.11	76.89	48.91
C4H4O	but-2-ynal	-229.99	-256.12	26.13	11.36
C3H4O2	2-oxopropanal	-267.16	-308.50	41.34	15.47
C2H6N2O	methyl-methylimino-oxidoazanium	-264.59	-309.29	44.70	16.89
C5H8	2-methylbuta-1,3-diene	-195.27	-258.99	63.72	32.63
C3H3NO	prop-2-ynamide	-246.07	-292.27	46.20	18.78
C2H3NO2	2-cyanoacetic acid	-283.21	-330.85	47.64	16.82
C3H7NO	propanamide	-248.90	-300.76	51.86	20.84
C4H6O	cyclobutanone	-231.18	-260.26	29.08	12.58
C3H6O2	1-hydroxypropan-2-one	-268.37	-314.43	46.06	17.16
C5H6	2-methylbut-1-en-3-yne	-194.05	-232.41	38.36	19.77
C4H5N	1*H*-pyrrole	-210.08	-243.15	33.07	15.74
C5H10	pent-1-ene	-196.54	-257.30	60.76	30.91
C4H8O	butanal	-232.44	-287.53	55.09	23.70
C2H6O2	ethane-1,2-diol	-268.36	-318.44	50.08	18.66
C2H3N3	1*H*-1,2,4-triazole	-242.25	-304.85	62.60	25.84
C3H5NO2	2-nitroprop-1-ene	-322.43	-366.69	44.26	13.73
C6H14	hexane	-237.10	-278.78	41.68	17.58

Table S3: Predicted molecular energy and associated errors for a few sample molecules: GNN with atom vector.