Utilization of Genetic Programming for Estimation of Molecular Structures Ground State Energies

Nikola Anđelić[1], Sandi Baressi Šegota[1], Ivan Lorencin[1], Matko Glučina[2], Jelena Musulin[1], Daniel Štifanić[1], Zlatan Car[1]

[1] University of Rijeka, Faculty of Engineering Vukovarska 58, 51000 Rijeka Croatia
Email: nandelic@riteh.hr;sbaressisegota@riteh.hr;ilorencin@riteh.hr;jmusulin@riteh.hr; dstifanic@riteh.hr;car@riteh.hr
[2] University of Rijeka Trg Braće Mažuranića 10, 51000 Rijeka, Croatia
Email: matko.glucina@uniri.hr

Abstract

In this paper, we shall be utilizing genetic programming (GP) to predict ground-state energies of molecules made up of C, H, N, O, P, and S (CHONPS) atoms. The GP was trained and tested on a publicly available dataset which consist of 16242 molecules where ground state energies were computed using the density functional theory (DFT). The optimal parameters of GP were chosen using the random parameter search method. After multiple GP executions, the best symbolic expression was chosen using a coefficient of determination (R^2), mean absolute error (MAE), and root mean square error ($RMSE$). The best symbolic expression achieved R^2, MAE, and $RMSE$ of 0.9434, 0.48, and 0.86, respectively.

Key words: CHNOPS dataset, genetic programing, ground state energies

1 Introduction

The idea of predicting electronic structure was investigated using neural networks [1], support vector regression [2] and tree-based machine learning methods [3]. Unlike the aforementioned methods, the benefit of using genetic programming (GP) are the symbolic expressions we get as a result, and they correlate the input variables with the output variables. Examples of GP implementation to various problems is documented in [4, 5, 6, 7]. In this paper, the idea is to investigate the possibility of implementing GP on a publicly available CHNOPS dataset [8] to find symbolic expressions which could estimate the ground state energy of molecules with high accuracy.

2 Materials and Methods

2.1 Dataset

As we previously mentioned, a publicly available dataset was used. This dataset consists of intermolecular Coulomb repulsion operators (Coulomb matrices) for each molecule and the corresponding ground state energies obtained with density functional theory (DFT) simulations. However, it should be noted that the Coulomb matrices of each molecule are symmetric matrices so for each molecule only the upper triangle data of the matrix is provided in the dataset. So for each instance (molecule) in the dataset there are 1275 input variables and one input variable. Due to a large number of input variables the correlation heat map which is used to investigate the correlation between variables in the

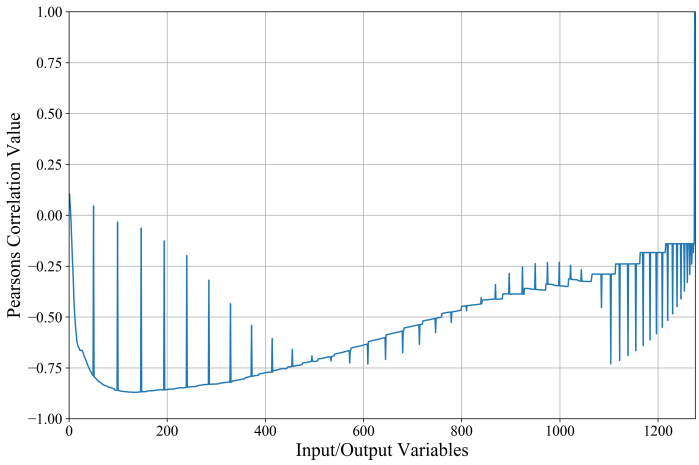

Figure 1: Pearsons Correlation Value of 1275 input variables to output variable (Ground State Energy)

dataset is almost impossible. So only correlation values between each input variable and the output variable (ground state energy) is shown in Figure 1. The Pearson's correlation method is used to investigate the correlation between input and output variables. As seen in Figure 1 the majority of input variables have a weak correlation (-0.5 - 0.5) to the output variable. This could potentially indicate that these variables won't end up as one of the variables of the final symbolic expression obtained with GP. Initially, in GP all input variables will be included.

2.2 Genetic Programming

The genetic programming (GP) was utilized to obtain a symbolic expression for the estimation of ground-state energies with a random selection of GP parameters before each algorithm execution. The initial population was crated using a ramped half-and-half method which was evolved for a randomly selected number of generations from a predefined range. The fitness function used in each generation to evaluate each population member was a mean absolute error (MAE). In each generation of tournament selection winners 4 genetic operations were performed (crossover and subtree/hoist/point mutation). If a predefined fitness value (stopping criteria) was achieved by one of the population members the GP execution would be terminated. To control the size of population members the definition of parsimony coefficient is required which is also responsible for the prevention of the bloat phenomenon. The range of all GP parameters is given in Table 1.

The metrics used to evaluate symbolic expressions are coefficient of determination (R^2), MAE, and root mean squared error ($RMSE$).

3 Results and Discussion

The best symbolic expression was selected based on the highest R^2 value and the lowest MAE and $RMSE$ values. Due to its size, the symbolic expression will not be shown here. However, in Table 2, GP parameters as well as R^2, MAE, and $RMSE$ values

GP Parameters	Lower Bound	Upper Bound
Population size	100	200
Number of Generations	100	200
Tournament Selection	10	20
Initial Tree Depth	3 - 7	7 - 12
Crossover	0.9	1
Subree Mutation	0.1	1
Hoist Mutation	0.1	1
Point Mutation	0.1	1
Stopping Criteria	$1 \cdot 10^{-6}$	$1 \cdot 10^{-4}$
Range of Constants	-10000	10000
Parsimony Coefficient	$1 \cdot 10^{-6}$	$1 \cdot 10^{-4}$

Table 1: The range of GP parameters used in random parameters selection process

obtained on the train/test dataset are shown instead of the best symbolic expression. From the presented results shown in Table 2 it can be noticed that overfitting occurred

GP Paramters	R^2	MAE	$RMSE$
1970 147 108 (4,7) 0.37, 0.37, 0.098, 0.14, 0.00032, 0.87, (-1984.21, 4130.14), $9.51 \cdot 10^{-5}$	0.965/0.9434	0.469/0.48	0.68/0.86

Table 2: GP parameters and evaluation metric values of the best symbolic expression

since the R^2, MAE, and $RMSE$ values are higher on the training set than those achieved on the testing dataset. The graphical comparison of the ground state energies from the test dataset and the values calculated using symbolic expression is shown in Table 2.

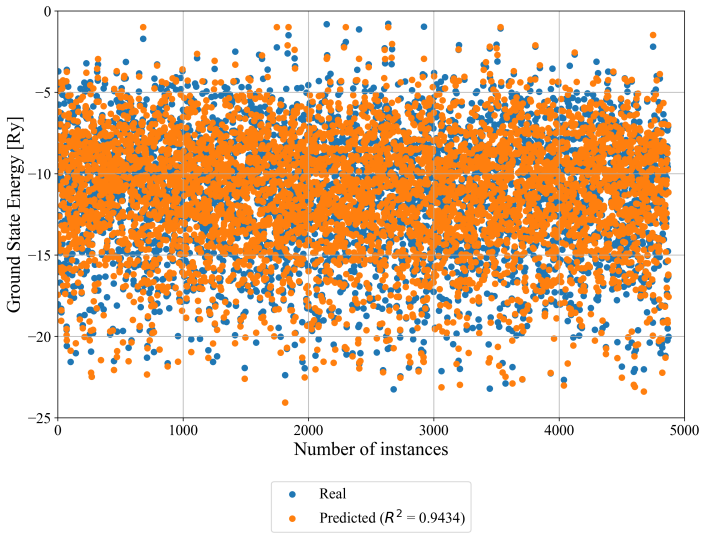

Figure 2: The comparison of the ground state energies from test dataset and the values calculated by the obtained symbolic expression

4 Conclusions

The GP algorithm was utilized on a publicly available dataset to investigate if it is possible to obtain the symbolic expression for estimation of ground state energy with high accuracy. The proposed approach showed that symbolic expressions for estimation of ground state energies could be obtained with GP. The only problem with this approach is overfitting which can be avoided using cross-validation techniques and different data scaling/normalizing methods.

Acknowledgments

This research has been (partly) supported by the CEEPUS network CIII-HR-0108, European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS), project CEKOM under the grant KK.01.2.2.03.0004, Erasmus+ project WICT under the grant 2021-1- HR01-KA220-HED-000031177 and University of Rijeka scientifc grant uniri-tehnic-18-275- 1447.

References

[1] Behler, Jörg. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. Physical Chemistry Chemical Physics 13.40 (2011): 17930-17955.

[2] Balabin, Roman M., and Ekaterina I. Lomakina. Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?. Physical Chemistry Chemical Physics 13.24 (2011): 11710-11718.

[3] Himmetoglu, Burak. Tree based machine learning framework for predicting ground state energies of molecules. The Journal of chemical physics 145.13 (2016): 134101.

[4] Anđelić, Nikola, et al. Estimation of gas turbine shaft torque and fuel flow of a CODLAG propulsion system using genetic programming algorithm. Pomorstvo 34.2 (2020): 323-337.

[5] Anđelić, Nikola, et al. Estimation of COVID-19 epidemic curves using genetic programming algorithm. Health informatics journal 27.1 (2021): 1460458220976728.

[6] Anđelić, Nikola, et al. Estimation of covid-19 epidemiology curve of the united states using genetic programming algorithm. International Journal of Environmental Research and Public Health 18.3 (2021): 959.

[7] Anđelić, Nikola, et al. Use of Genetic Programming for the Estimation of CODLAG Propulsion System Parameters. Journal of Marine Science and Engineering 9.6 (2021): 612.

[8] Himmetoglu, Burak. Tree based machine learning framework for predicting ground state energies of molecules. The Journal of chemical physics 145.13 (2016): 134101.