UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITY OF PADUA Department of Chemical Sciences

Coding for Chemistry / Machine Learning for Chemistry Academic Year: 2025-2026

PROJECT TITLE: Modeling the HOMO/LUMO gap of an organic molecule

STUDENT NAME: Mohammadreza Lotfi **STUDENT ID: 2186158**

1. Introduction

The electronic structure of a molecule plays a crucial role in shaping its physicochemical properties. The energy difference between the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO) is known as the HOMO-LUMO gap. According to Frontier Molecular Orbital (FMO) theory, this gap is vital for evaluating the kinetic stability, chemical reactivity, and optical properties [1] of a molecule. Accurately determining this gap is essential for designing next-generation materials, especially in organic photovoltaics (OPV), organic light-emitting diodes (OLEDs), and drug discovery [2]. Typically, the HOMO-LUMO gap is calculated using quantum mechanical methods, particularly density functional theory (DFT), which is based on the Kohn-Sham equations [3]. While DFT provides a high level of chemical accuracy, it has a significant computational limitation. The computational cost of standard DFT calculations increases roughly with the number of basis functions (N) at a rate of O(N3) or O(N4), making high-throughput screening of large chemical libraries quite challenging [4].

To address this problem, the field of chemical informatics has increasingly adopted data-driven approaches. Machine learning (ML) algorithms offer a promising alternative by identifying statistical patterns from existing quantum mechanical data to predict molecular properties at a lower computational cost [5]. By converting molecular structures into numerical representations (fingerprints), ML models can avoid the iterative process of solving the Schrödinger equation, thereby speeding up screening from hours to milliseconds.

In this project, a specific subset of the QM9 database was used, which is a benchmark dataset containing geometric, energetic, and electronic properties for around 134,000 stable organic molecules. Our goal is to develop and compare three regression models (random forest, artificial neural network, and kernel ridge regression) to predict the HOMO-LUMO gap based on 2D Morgan fingerprints. This method aims to create a fast-screening pipeline that balances computational efficiency with prediction accuracy.

2. State of the art

Prediction of the molecular properties has relied on quantum chemistry from the begining. Since Cohen and Sham developed density functional theory (DFT) [3], it has become the standard for calculating electronic properties like the HOMO-LUMO gap. Software like Gaussian and ORCA has enabled chemists to calculate these properties with high accuracy. However, as noted by Ratcliffe et al. [4], the computational complexity of these methods increases poorly, creating a barrier to exploring large chemical spaces. To overcome this issue, research has shifted towards data-driven high-throughput screening. A significant milestone was the release of the QM9 database [6] by Ramakrishnan and colleagues in 2014. This dataset provided high-quality quantum mechanical calculations for nearly 134,000 organic molecules and served as a benchmark for training machine learning models. Recently, statistical learning methods, especially quantitative structure-activity relationship models, have received attention. Studies have shown that converting molecules to two-dimensional descriptors, like extended connectivity fingerprints (ECFPs) or Morgan fingerprints [7], allows algorithms to bypass the Schrödinger equation. Butler and colleagues [5] emphasized that modern algorithms, such as random forests and neural networks, can predict quantum properties with accuracy on par with DFT, but in a fraction of the time. Current research focuses on enhancing these models using new architectures like graph neural networks (GNNs). However, for rapid screening on standard hardware, feature-based approaches, like Morgan fingerprints used in this project, remain the most efficient and reliable option for small organic molecules.

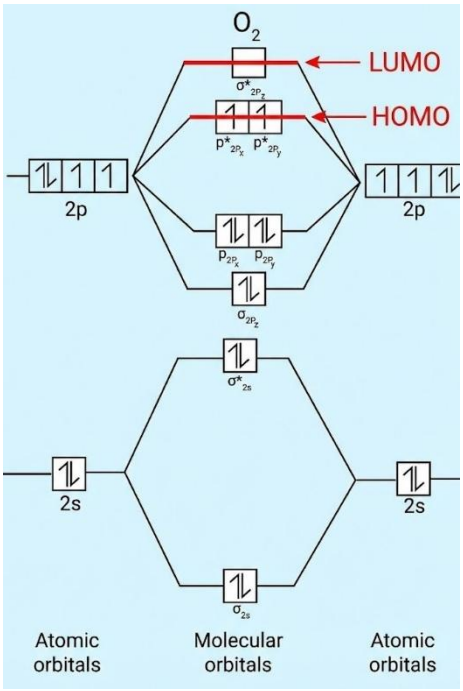


Fig. 1) Molecular orbital (MO) diagram for the oxygen molecule (O_2)

3. Methodology

To predict the HOMO-LUMO gap of organic molecules, a supervised machine learning workflow was created. This process consists of three main steps, data preparation, feature engineering, and model development. All implementations were done using Python (v3.10) and the RDKit library for cheminformatics, along with Scikit-Learn for machine learning tasks [8].

3.1. Dataset Description

The dataset for this study is a subset of the QM9 database, provided as Supplementary Material. It includes 20,000 small organic molecules made up of hydrogen (H), carbon (C), oxygen (O), nitrogen (N), and fluorine (F) atoms. Molecular structures are given in SMILES (Simplified Molecular Input Line Input System) format. The target feature is the HOMO-LUMO gap energy, initially calculated using DFT.

3.2. Feature Engineering

Since standard regression algorithms cannot analyze raw text strings directly, SMILES strings were transformed into numerical feature vectors using Morgan fingerprints (ECFP). Bit vectors were generated with the following parameters, Radius = 2, which considers the central atom and its neighbors up to 2 bonds away, capturing functional groups and local chemical environments. nBits = 1024, with a fixed vector length of 1024 bits was chosen to reduce hash collisions and maintain computational efficiency using RDkit.

3.3. Machine Learning Algorithms

Three separate regression models were trained and evaluated for benchmarking. Firstly, Random Forest Regression (RF), an ensemble learning method that builds a group of decision trees was used [9], then the model with 100 estimators was initialized. Secondly, an Artificial Neural Network (ANN), a Multilayer Perceptron (MLP) regression with two hidden layers (128 and 64 neurons), using the ReLU activation function was implemented.

Finally, Kernel Ridge Regression (KRR), which is based on the Radial Basis Function (RBF) kernel was used. Even though This model requires computing and inverting Gram matrices, which have cubic time complexity and square memory complexity the data sample was the same amount.

2.4. Validation Strategy

To ensure generalizability, the dataset was randomly divided into a training set (80%) and a test set (20%). We evaluated model performance using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). The complete procedure is shown in Fig. 2.

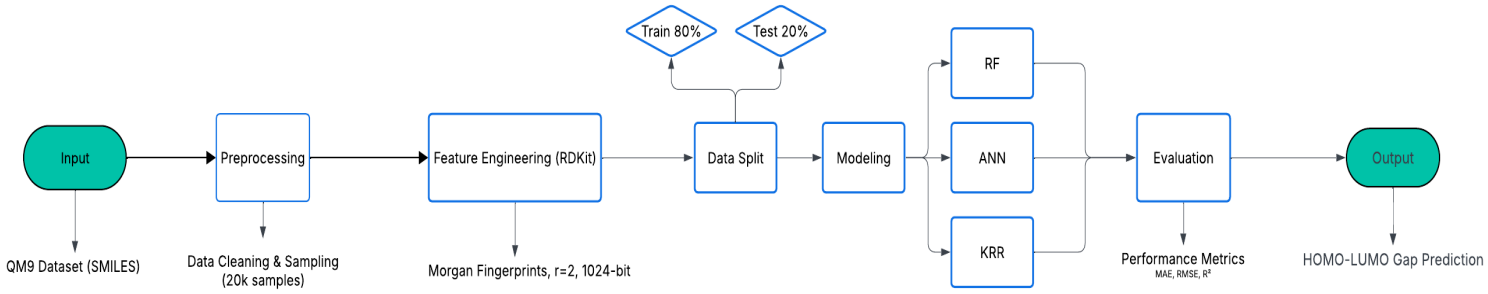


Fig. 2) Flowchart of the process

4. Results and Discussion

4.1. Statistical Performance of Models

Predictive ability of the implemented algorithms was assessed using three standard metrics, coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). The performance of the three models on an independent test set (20% of the data) is summarized in table.1.

Model	R2	MAE	RMSE
RF	0.8795	0.01126	0.0164
ANN	0.868	0.0124	0.01717
KRR	0.8562	0.01327	0.01791

Table.1) Comparison of the results for three models

As observed, the random forest (RF) regression outperformed the other models by achieving the highest R^2 score of 0.88. This indicates that the model can explain approximately 88% of the variance in the HOMO-LUMO gap energies. The mean absolute error (MAE) of 0.01126 Hartree (approximately 0.30 eV) and Root Mean Squared Error (RMSE) of 0.01640 Hartree is low enough for high-throughput screening applications. While ANN and Kernel-Ridge models showed R^2 score of 0.868, 0.856, respectively. And MAE and RMSE of 0.01240, 0.01327 Hartree and 0.01717, 0.01791 Hartree, in that order. These stats shows that all of these three models showed very satisfactory results, and small differences in results can be caused by small dataset number where ANN typically require larger datasets to outperform tree-based methods on tabular data.

4.2. Visual inspection

Fig.3 shows the parity plot for all of the models. The x-axis represents the ground truth values calculated by DFT and the y-axis represents the ML predictions. The red dashed line represents the ideal prediction (y=x). Visual analysis confirms that the data points are tightly clustered around the ideal diagonal line. No significant systematic bias such as overestimation or underestimation is observed over the energy range. The distribution of errors appears to be homoscedastic, meaning that the models performed consistently for both large gap and small gap molecules.

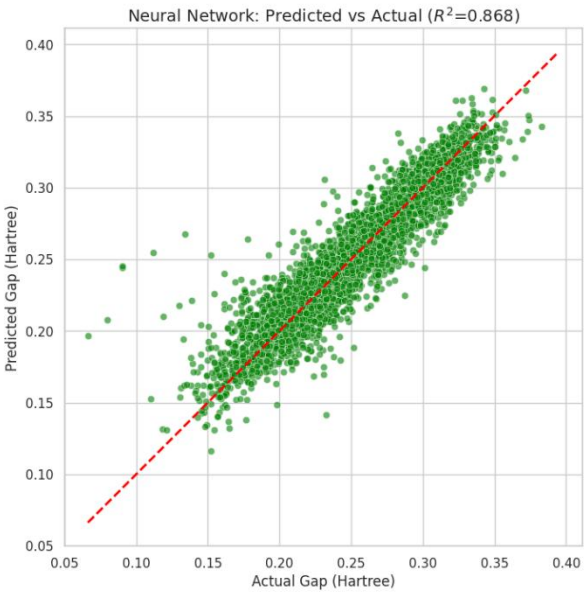


Fig.3.a) parity plot of ANN model

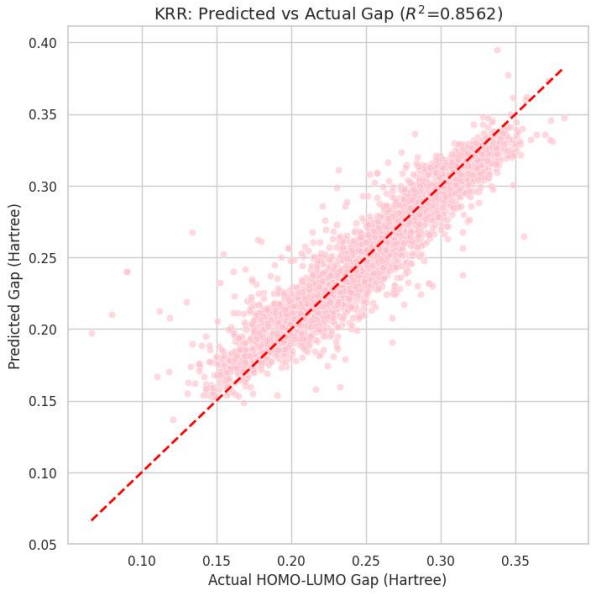
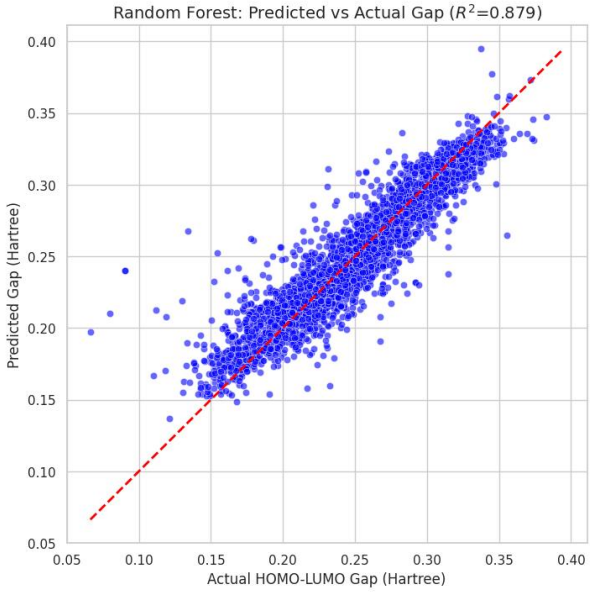


fig.3.b) parity plot of KRR model

Fig.3.c) parity plot of RF model



4.3. Case Study Validation

To assess the chemical validity of the model, we performed predictions on specific molecules with known properties (table.3). This step ensures that the model learns meaningful chemical patterns.

Molecule	Real (eV)	RF (eV)	ANN (eV)	KRR (eV)
Aspirin	5.4	5.3969	4.9037	4.6216
Benzene	6.9	7.0357	6.3904	7.0487
Methane	10.81	11.9825	12.0739	8.3866
Toluene	6.7	6.09	5.6085	5.9026
Phenol	6.15	6.151	6.0554	6.139
Aniline	5.6	5.8878	5.2849	5.582
Ethylbenzene	6.7	6.1419	6.2173	6.1158
Acetaminophene	5	5.3338	5.3888	5.1439
caffeine	5.3	5.1387	4.6959	5.156

Table.3) Comparison of predicted HOMO-LUMO gaps versus actual gaps for selected molecules.

5. Conclusions

In this project, we successfully developed a machine learning pipeline to predict the HOMO-LUMO gap of organic molecules, effectively overcoming the computational bottleneck of traditional density functional theory (DFT). There were some key findings such as feasibility, where we used only morgan finger print and Random Forest algorithm and we reached the R^2 score of 0.88. Secondly, the efficiency of the model where predictions can be made instantaneously enables the possibility to rapidly screen massive chemical libraries. Finally, although the model showed a great performance predicting complex organic molecules there were some reduced accuracies for very simple molecules (e.g., Methane).

6.References

[1] Fukui, K. (1982). Role of Frontier Orbitals in Chemical Reactions. *Science*, 218(4574), 747–754. https://doi.org/10.1126/science.218.4574.747

[2] Brédas, J. L., et al. (2009). Molecular Understanding of Organic Solar Cells. *Accounts of Chemical Research*, 42(11), 1691–1699. https://doi.org/10.1021/ar900099h

[3] Kohn, W., & Sham, L. J. (1965). Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A), A1133. https://doi.org/10.1103/PhysRev.140.A1133

[4] Ratcliff, L. E., et al. (2017). Challenges in large-scale electronic structure calculations. *WIREs Comput. Mol. Sci.*, 7(1), e1290. https://doi.org/10.1002/wcms.1290

[5] Butler, K. T., et al. (2018). Machine learning for molecular and materials science. *Nature*, 559, 547–555. https://doi.org/10.1038/s41586-018-0337-2

[6] Ramakrishnan, R., et al. (2014). Quantum chemistry structures and properties of 134 kilo molecules (QM9). *Scientific Data*, 1, 140022. https://doi.org/10.1038/sdata.2014.22

[7] Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5), 742–754. https://doi.org/10.1021/ci100050t

[8] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[9] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Appendix

https://github.com/molotfii/mlotfi-cc2526