Artificial Intelligence in Molecular Physics

[Vol-01, Issue-02,July 2026]
DOI: XX
ISSN: XXXX-XXXX

Machine Learning-Based Prediction of Electronic and Thermodynamic Properties of Small Organic Molecules Using the QM9 Database

Hyunsung Cho

Department of Materials Science and Engineering, Seoul National University, Seoul 08826, Korea

Received: 06/06/2026 **Revised:** 12/07/2026
Accepted: 20/07/2026 **Published:** 27/07/2026

Abstract—*The ability to accurately predict the electronic and thermodynamic properties of molecules is crucial in computational chemistry, molecular screening and materials discovery. Conventional quantum-chemical calculations provide reliable estimates but require substantial computational resources, creating a need for efficient and interpretable predictive approaches. This study developed and compared descriptor-based machine-learning models for predicting HOMO energy, HOMO–LUMO gap, free energy at 298.15 K, and heat capacity at 298.15 K using the QM9 database. A curated dataset comprising 133,885 small organic molecules described by 20 numerical molecular descriptors was analysed. Ridge Regression, Random Forest, and Extra Trees models were trained using an 80:20 train–test split and evaluated through mean absolute error, root mean square error, and the coefficient of determination (R^2). Feature-importance analysis and residual evaluation were performed to assess model interpretability and prediction stability. Random Forest achieved the highest accuracy for HOMO energy (R^2 = 0.8227) and HOMO–LUMO gap (R^2 = 0.8992), Ridge Regression produced near-perfect prediction of free energy at 298.15 K (R^2 = 1.0000), and Extra Trees yielded the best performance for heat capacity (R^2 = 0.9748). Exact molecular weight, molecular weight, oxygen atoms, carbon atoms, and hydrogen-bond acceptors emerged as the most influential descriptors, while residual distributions centred near zero indicated minimal systematic prediction bias. The proposed framework offers a reliable, interpretable, and computationally fast solution for initial prediction of molecular properties and high-throughput molecular screening, but further testing with a more chemically diverse set of molecules is encouraged to expand its applicability.*

Keywords— ***QM9, molecular property prediction, Random Forest, thermodynamic properties, molecular descriptors***

1. INTRODUCTION

Molecular property prediction has emerged as a key objective in computational chemistry as the electronic and thermodynamic properties dictate the chemical behaviour, stability and functional performance of organic molecules in pharmaceutical, catalytic and materials science applications. A number of functional molecules are based on small organic molecules, and their fast characterization is crucial for speeding up the discovery of molecules and at the same time lowering the computational cost of traditional quantum chemical calculations. Synthetic Chemistry advancements have allowed the preparation of structurally diverse small organic molecules in an efficient manner, making opportunities to incorporate computational tools with experimental design for the rapid evaluation of properties (Achar et al., 2017). The advent of machine learning has also revolutionized this area with the development of predictive models that can learn complex relationships between molecular structure and physicochemical properties from large datasets and hence aid in high-throughput molecular screening and rational compound design (Wei et al., 2019). Computational prediction has been done traditionally using density functional theory (DFT), which is able to give reliable estimation of molecular electronic and thermodynamic properties, but the computation of large molecular libraries requires significant computational resources. With the number of candidate molecules still growing, there is increasing interest in supplementing quantum mechanical simulations with data-driven algorithms that can yield accurate results in much less time. The application of machine learning in computational chemistry thus has emerged as a promising research area, where the accurate prediction, optimization and interpretation of molecular behavior, while preserving good agreement with theoretical calculations, are possible. (Westermayr et al., 2021) Meanwhile, advances in deep learning have further boosted predictive modelling by achieving better structural extraction from molecular representations and better scalability of computational materials research (Choudhary et al., 2022).

The frontier molecular orbitals are among the most important molecular properties of scientific interest and useful in the understanding of reactivity and electronic behaviour. The highest occupied molecular orbital (HOMO) demonstrates the electron-donating ability of a molecule, while the HOMO–LUMO energy gap is generally used as a measure of chemical stability, optical activity, and charge-transfer ability. Accurate prediction of these electronic descriptors is thus relevant to molecular design for various applications including organic electronics, photovoltaics and medicinal chemistry. There has been significant improvement in the prediction of DFT-derived HOMO and LUMO energies using machine learning approaches, which is an efficient method when compared to repeated electronic structure calculations (Pereira et al., 2017). Descriptor-based approaches have also revealed that structural features of molecules can be properly used to estimate frontier orbital energies and electronic gaps (Huang et al., 2017). But the electronic properties are not sufficient to characterize the behaviour of the molecule. Thermodynamic properties such as free energy and heat capacity offer valuable complementary information on the molecular stability, energetic feasibility, and behaviour with temperature. These properties are essential for the assessment of reaction pathways, molecular conformation and physicochemical properties under different conditions. In molecular characterization, frontier orbital analysis studies are still found to be important, in addition to the studies of molecular thermodynamic functions (Yoğurtçu & Ersanlı, 2025). Also, the accurate determination of free energy-related properties has become more important in computational drug discovery that relies on energetic calculations for ligand optimization and molecular interaction analysis (Abel et al., 2017).

The rapidly increasing number of curated molecular databases has significantly improved the use of machine learning in chemistry by offering high-quality data sets with quantum mechanically calculated molecular properties. These help to develop predictive models that can estimate several molecular properties at once without relying on the simulation that can be quite costly. In recent reviews, the use of machine learning has been highlighted not only for predicting interatomic interactions, but also for the accurate estimation of various molecular properties (Fedik et al., 2022). The development of new open-source frameworks for chemical property prediction has further enhanced model development, model reproduction and access (Heid et al., 2023). Ensemble learning techniques are also receiving a lot of interest as these combine the predictions of several decision trees and are suitable for molecular datasets, where nonlinear relationships exist (Davronov & Adilova, 2021). Previous studies have largely concentrated on individual electronic properties, specific thermodynamic properties or particular modelling approaches and have not studied several electronic and thermodynamic properties in a common framework based on a descriptor. Comparative studies of conventional regression and ensemble learning models with the same set of molecular descriptors to concurrently predict HOMO energy, HOMO–LUMO gap, free energy, and heat capacity are still scarce. Moreover, there are relatively few works that integrate predictive performance with systematic interpretation of the importance of the descriptors despite the rising demand for interpretable machine learning models. This is in line with the general problem of combining knowledge from quantum chemistry with machine learning to successfully predict thermodynamic stability and other molecular properties

(Schmidt et al., 2017).

In this context, the present study proposes and evaluates descriptor-based machine learning models to predict a set of key electronic and thermodynamic properties of small organic molecules using the QM9 database, and compares the models with each other. The study compares the predictive power of Ridge Regression, Random Forest, and Extra Trees models for estimation of HOMO energy, HOMO–LUMO gap, free energy and heat capacity and also uses feature importance analysis to identify the molecular descriptors that contribute the most to the prediction. The study will combine the use of comparative modelling with descriptor interpretation to offer an efficient and interpretable molecular property prediction tool for future applications in computational chemistry and data-driven molecular discovery.

2. METHODOLOGY

2.1 Research Design

The aim of the present study was to predict the electronic and thermodynamic properties of small organic molecules based on molecular descriptors using a machine learning technique with quantitative data. The analytical workflow comprised data acquisition, preprocessing of descriptors, evaluation of data quality, model development, predictive assessment and feature importance analysis. Three machine-learning algorithms (Ridge Regression, Random Forest, and Extra Trees) were then implemented to estimate four target properties: HOMO energy, HOMO–LUMO gap, free energy at 298.15 K and heat capacity at 298.15 K and the performance of the models was evaluated using the standard regression metrics; the descriptor importance was analyzed to determine which molecular properties contribute most strongly to the prediction accuracy. The methodological framework was designed in order to ensure reproducible model development and objective comparison of the different learning approaches.

2.2 Data Source

The molecular data for this investigation were taken from a publicly available scientific repository that supports computational molecular modelling and machine-learning research (Iacovella et al., 2025). The collection includes carefully selected quantum chemical data for a wide variety of structurally different small organic molecules, and a set of molecular descriptors and physicochemical properties. The curated resource gives a consistent representation of molecules for predictive modelling and to evaluate algorithms for comparison, which allows consistency in the analytical workflow (Iacovella et al., 2025).

2.3 Data Preparation and Feature Engineering

Prior to the construction of the models, the downloaded dataset was checked to ensure it was complete, consistent and suitable for predictive modelling. Molecular entries were duplicated and observations were missing, both of which were evaluated and only valid molecules and observations were used for further analysis. The final data set was clean from missing values and redundancy, and comprised 133,885 molecules. A set of twenty numerical molecular descriptors was adopted as predictor variables and the HOMO energy, HOMO–LUMO gap, free energy at 298.15 K, and the heat capacity at 298.15 K were used as prediction targets. All the predictor variables were restructured and presented as a continuous dataset for analysis, but the numbers were kept to perform training and testing of the model.

2.4 Machine Learning Model Development

Three regression algorithms were developed with different learning characteristics, which are complementary to each other, to predict the selected molecular properties. Ridge Regression was used to obtain a linear baseline model that can be used for correlated numerical descriptors. Ensemble tree-based methods were used to account for the nonlinear relationships between molecular descriptors and target properties: Random Forest and Extra Trees. The prepared dataset was split into training and testing sets (80:20), and each algorithm was trained independently for each of the target properties. To allow direct comparison of the algorithm performance with the various molecular properties, separate predictive models were derived for the energy of HOMO and HOMO–LUMO gap, free energy and heat capacity.

2.5 Model Evaluation and Descriptor Interpretation

Each of the models was tested for predictive ability using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2), which was used to measure prediction accuracy and goodness of fit. The most appropriate algorithm for each molecular property was selected using comparative performance analysis with these regression metrics. In order to gain insight into the behavior of the models, feature importance analysis was conducted on the ensemble-based models, which led to the quantitative estimation of the contribution of the different molecular descriptors to the prediction performance. Residual distributions were further analyzed to evaluate prediction bias, error symmetry and model robustness for the 4 properties being targeted, thereby giving further indications of model reliability.

3. RESULTS

3.1 Dataset Characteristics and Property Distribution

The processed QM9 dataset consisted of 133,885 small organic molecules with 4 target properties which include

both electronic and thermodynamic properties, and 20 numerical molecular descriptors. Preprocessing did not result in any observations being lost or in duplicate molecules being recorded, ensuring the integrity of the data for machine learning analysis. The distributions of numbers were different for the target variables, with an average for the HOMO energy of −630.06 kJ mol−1, a calculated average for the HOMO–LUMO gap of 659.26 kJ mol−1, a calculated average of the free energy of –1.08 × 106 kJ mol−1, and a calculated average of the heat capacity of 0.1322 kJ mol−1 K−1. The different variations suggest the predictive framework was tested on significantly different numerical scales and physicochemical behaviour for properties. The overall quality of the processed QM9 data set, along with various information about the descriptors and statistical summary of the four target properties, is presented in Table 1.

Table 1. Characteristics of the QM9 Dataset and Target Properties

Characteristic	Value	Unit or Description
Total Molecules	133,885	Small organic molecules
Molecular Descriptors	20	Numerical predictor variables
Target Properties	4	Electronic and thermodynamic properties
Missing Values	0	Across modelling variables
Duplicate Molecules	0	Based on molecular identity
HOMO Energy	−630.0587 ± 58.1061	kJ mol^{-1}
HOMO–LUMO Gap	659.2639 ± 124.7608	kJ mol^{-1}
Free Energy at 298.15 K	−1080596.3078 ± 105179.4604	kJ mol^{-1}
Heat Capacity at 298.15 K	0.1322 ± 0.0170	kJ mol^{-1} K^{-1}

3.2 Predictive Performance of Machine-Learning Models

Three supervised learning algorithms were tested for prediction of the four properties of the molecules. The model results showed that prediction accuracy was found to be dependent on the physicochemical property to be estimated. The best predictive performance by Random Forest was obtained for HOMO energy (R^2 = 0.8227) and HOMO-LUMO gap (R^2 = 0.8992) while the best prediction for free energy was obtained by Ridge Regression (R^2 = 1.0000). Extra Trees gave the best performance with respect to heat capacity (R^2 = 0.9748). All properties exhibited the same trends of ensemble tree-based methods that consistently showed greater nonlinear molecular relationships with MAE and RMSE, and the linear Ridge model showed greater predictive ability for free energy. Table 2 provides an overview of prediction results for all four target properties obtained by the machine-learning models evaluated.

Table 2. Performance of the Machine-Learning Models

Target Property	Machine-Learning Model	Mean Absolute Error	Root Mean Square Error	R Squared
HOMO Energy	Random Forest	17.3384	24.3773	0.8227
HOMO Energy	Extra Trees	17.8676	25.6815	0.8032
HOMO Energy	Ridge Regression	29.9891	39.9579	0.5237
HOMO–LUMO Gap	Random Forest	27.6236	39.7220	0.8992
HOMO–LUMO Gap	Extra Trees	28.4705	42.2509	0.8860
HOMO–LUMO Gap	Ridge Regression	61.0643	76.6034	0.6252
Free Energy at 298.15 K	Ridge Regression	53.0528	69.7610	1.0000
Free Energy at 298.15 K	Random Forest	145.6130	3446.0049	0.9989
Free Energy at 298.15 K	Extra Trees	84.6862	3783.2514	0.9987
Heat Capacity at 298.15 K	Extra Trees	0.0017	0.0027	0.9748

Heat Capacity at 298.15 K	Random Forest	0.0022	0.0031	0.9677
Heat Capacity at 298.15 K	Ridge Regression	0.0041	0.0053	0.9042

As shown in Figure 1A, the Random Forest model accurately predicted HOMO energy with the predicted energy values being well aligned with the line of perfect agreement. The predictions made by Random Forest in the case of the HOMO-LUMO gap were very uniform, as can be seen in Figure 1B, except for the slightly greater dispersion at higher values. In Figure 1C, we see that the Ridge Regression gave almost perfect correspondence between the observed and the predicted free-energy values. Figure 1D shows that the Extra Trees model was able to accurately estimate molecular heat capacity with little deviation from the diagonal line.

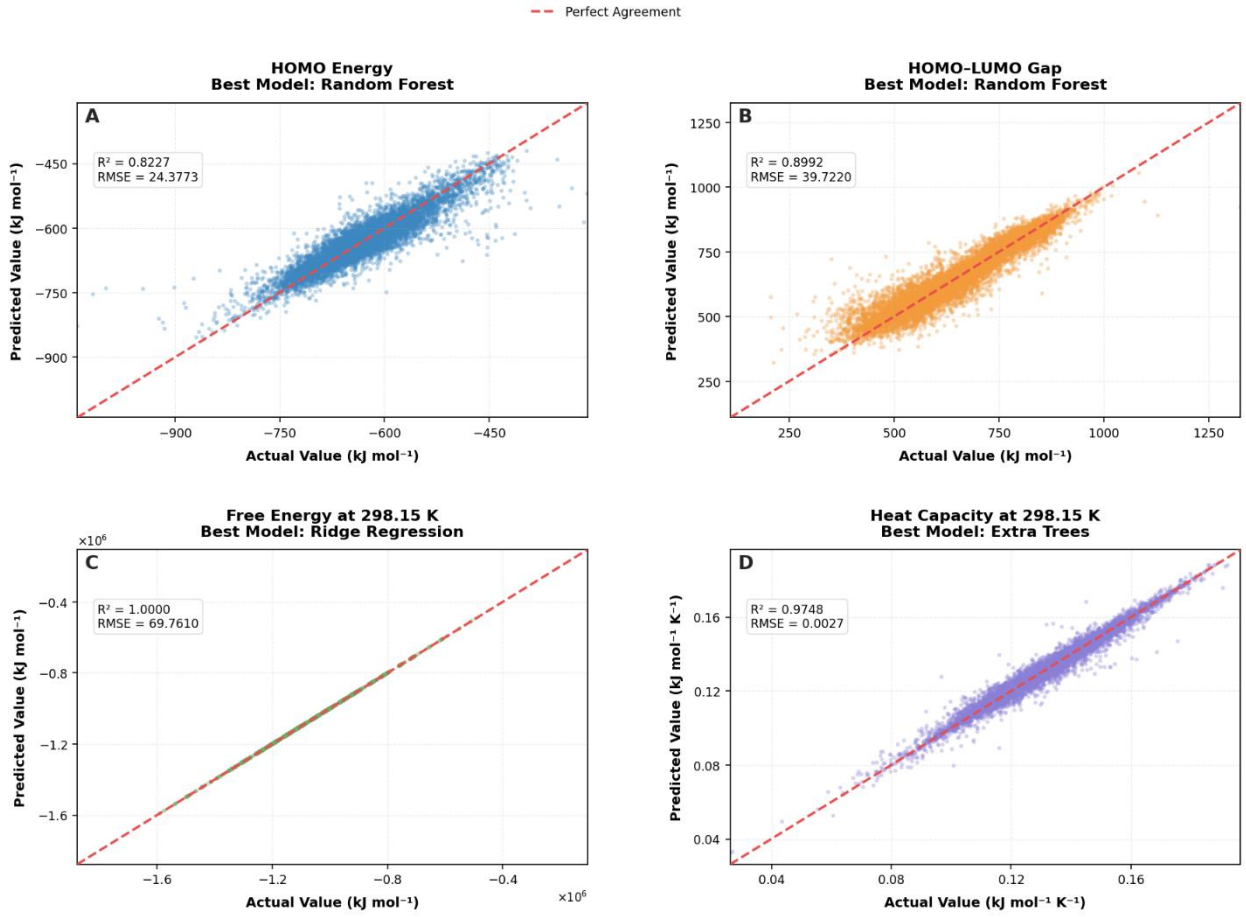

Figure 1. Actual and Predicted Electronic and Thermodynamic Properties. (A) HOMO Energy prediction using Random Forest. (B) HOMO–LUMO Gap prediction using Random Forest. (C) Free Energy at 298.15 K prediction using Ridge Regression. (D) Heat Capacity at 298.15 K prediction using Extra Trees.

3.3 Molecular Descriptor Importance and Error Analysis

Descriptor analysis showed that the most important descriptors were all related to the molecular size and atomic composition of the molecules, thus supporting the interpretability of the models. The average importance of each molecular property is shown, with molecular weight (0.1923) being the most significant, followed by Exact molecular weight (0.2131), Oxygen atoms (0.1463), Carbon atoms (0.1223), and Hydrogen-bond acceptors (0.0991). The other descriptors which were lower ranked (Bertz complexity, Fraction Csp3, aromatic ring count, Log P), had comparatively small contributions to the predictive performance, suggesting that the electronic and thermodynamic properties investigated in this study were primarily influenced by molecular size and elemental composition. Table 3 shows the relative importance of the molecular descriptors in the two models, Random Forest and Extra Trees.

Table 3. Ranked Importance of Molecular Descriptors

Rank	Molecular Descriptor	Random Forest Importance	Extra Trees Importance	Average Importance
1	Exact Molecular Weight	0.2996	0.1265	0.2131
2	Molecular Weight	0.2884	0.0963	0.1923
3	Oxygen Atoms	0.0775	0.2151	0.1463
4	Carbon Atoms	0.1299	0.1146	0.1223
5	Hydrogen-Bond Acceptors	0.0608	0.1374	0.0991
6	Heavy Atoms	0.0032	0.1051	0.0541
7	Total Atoms	0.0879	0.0155	0.0517
8	Fluorine Atoms	0.0231	0.0631	0.0431
9	Topological Polar Surface Area	0.0153	0.0658	0.0406
10	Nitrogen Atoms	0.0109	0.0241	0.0175
11	Valence Electrons	0.0012	0.0302	0.0157
12	Bertz Complexity	0.0005	0.0018	0.0012
13	Log P	0.0010	0.0013	0.0012
14	Fraction Csp3	0.0005	0.0011	0.0008
15	Aromatic Ring Count	0.0001	0.0012	0.0006

As shown in Figure 2, descriptors that were related to the molecular weight provided the highest contribution in model prediction, while those related to molecular topology and complexity showed a relatively small contribution.

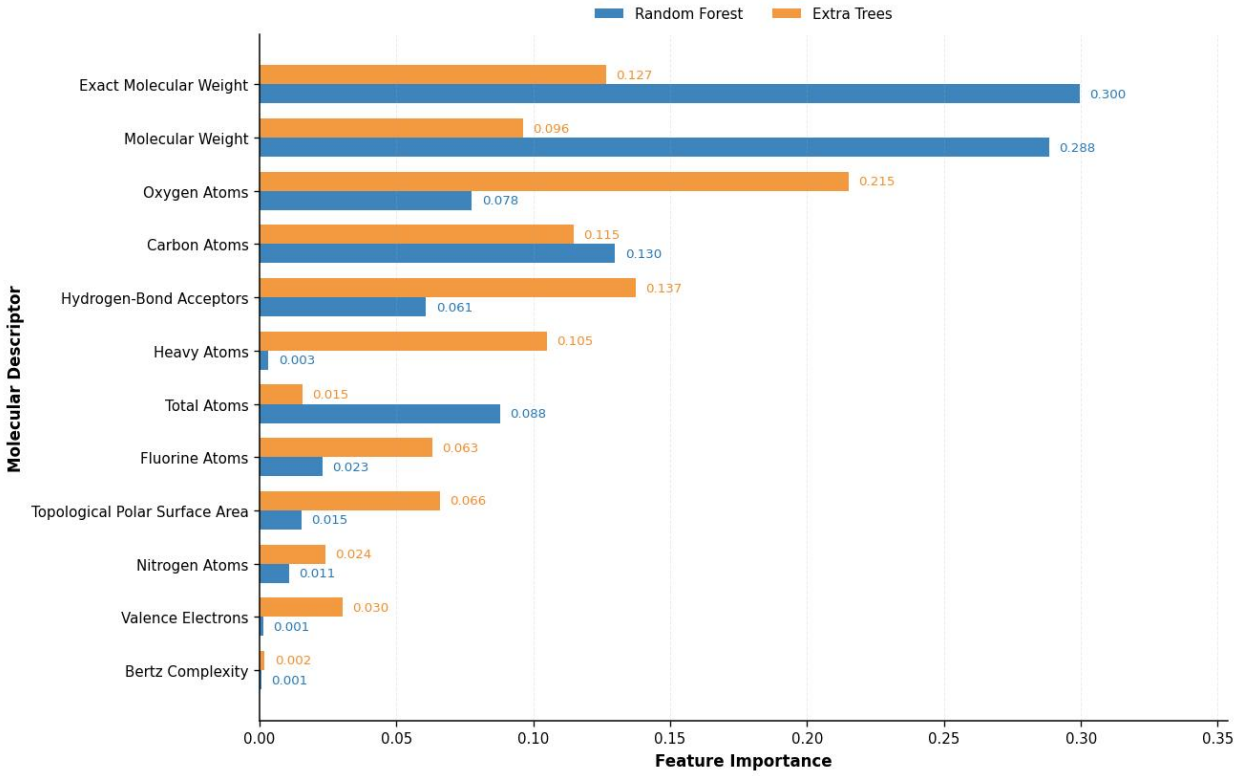

Figure 2. Comparative Importance of Molecular Descriptors Derived from the Random Forest and Extra Trees Models.

Further, the best-performing models were found to be robust by performing residual analysis. There were relatively small systematic prediction biases for the residual distributions of all target properties with the majority centred around zero. The residual distributions of HOMO energy and HOMO-LUMO gap were approximately symmetric, while the distribution of errors in heat capacity had no symmetric shape and was quite small with a very small prediction variance. Free energy also showed residuals around zero, although this was on a much greater numerical scale than the others, indicating that model behavior was quite uniform throughout the data set. The obtained residual distribution for the prediction of HOM-related energy by the Random Forest model is shown in Figure 3A. The residual distribution for the prediction of HOMO-LUMO gap is shown in Figure 3B, which shows that there is little systematic prediction error. The residual distribution for the Ridge Regression for free-energy prediction is shown in Figure 3C. Figure 3D represents the residual distribution for the Extra Trees model in predicting heat capacity, which shows the very low prediction variability.

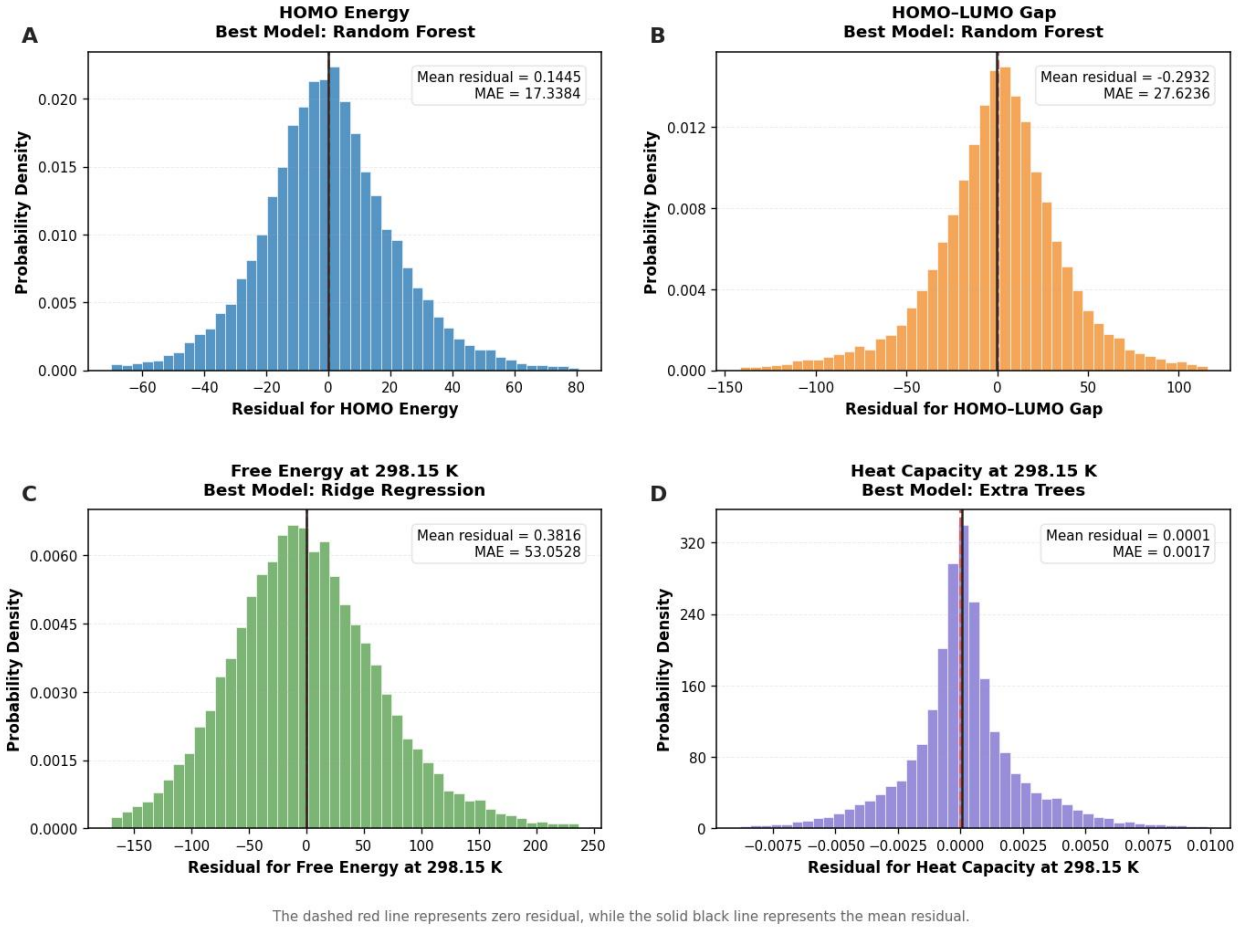

Figure 3. Residual Distributions of the Best-Performing Machine-Learning Models. (A) HOMO Energy residual distribution. (B) HOMO-LUMO Gap residual distribution. (C) Free Energy at 298.15 K residual distribution. (D) Heat Capacity at 298.15 K residual distribution.

4. DISCUSSION

The results indicate that the prediction performance varied depending on the type of chemical and physical properties of the target property. The highest accuracy with R^2 value was obtained for HOMO energy and HOMO–LUMO gap by the model of Random Forest, which gave R^2 values of 0.8227 and 0.8992, respectively. These results suggest that the electronic properties are dependent on the non-linear interaction between the molecular size, atomic composition, polarity and structural properties. This is reflected by the apparent lack of accuracy for Ridge Regression on both electronic targets, indicating that linear relationships alone were not enough of a model to capture the complexity of frontier-orbital behaviour. Ridge Regression gave the

highest R^2 value of 1.0000 and resulted in considerably lower prediction errors when compared to the ensemble models, which predicted the free energy best at 298.15K. The near-perfect result suggests that the descriptors were strongly and systematically related to the free energy in the QM9 chemical space. Much of the size-dependency of the energetic variation was likely accounted for by molecular weight, atom numbers and variables related to valence. The result is therefore likely to be very accurate for the present data set, but not necessarily for broader and more chemically diverse molecular spaces.

Extra Trees showed the highest R^2 (0.9748) and lowest RMSE (0.0027 kJ mol^{-1} K^{-1}) prediction for heat capacity. The performance indicates that the heat capacity was strongly dependent on the molecular composition and structure, but with some degree of non-linearity. The best-performing models had residual distributions which were centred near zero, suggesting little systematic bias. The residual variation in HOMO energy and HOMO–LUMO gap was larger than the heat capacity residuals, indicating that it was more difficult to model the electronic structure with molecular descriptors in the conventional sense. Molecular weight, oxygen atoms, carbon atoms and hydrogen-bond acceptors were the most important features for prediction, according to feature-importance analysis. The dominance suggests that the size of the molecules and the elements in the molecule were important factors in the properties examined. The contributions from aromatic ring count, fraction Csp3, Log P and Bertz complexity were comparatively small, with no indication that the molecular topology was unimportant. In contrast, these descriptors contributed provided only limited additional information after highly influential compositional variables explained a significant amount of the target variance.

The good performance of descriptor-based models confirms previous work showing that multiple QM9 properties can be predicted from minimal molecular representations, without performing prohibitively expensive quantum-chemical calculations anew (Pinheiro et al., 2020). Recent regression studies have also revealed that structure–property relationships can be found using alternative data-driven approaches, albeit the quality of the results is sensitive to the level of complexity of the target molecule and the way it is represented (Jacobs et al., 2024). The ensemble models' excellent performance for the electronic properties aligns with deep-learning studies that prove the utility of nonlinear modelling for the prediction of quantum-chemical properties as well as downstream molecular properties (Lim et al., 2022). Analyses of the QM7b and QM9 sets have also been made comparative, and the difficulty of the properties to be predicted was found to differ between these sets (Valdés & Tchagang, 2024), along with the fact that a single model could not be identified as the most successful for predicting all target properties. This is reflected in the model rankings for each property, as reported below. QM9 data have been employed to build chemistry-informed neural networks that have shown that including physical knowledge can enhance the prediction of molecular properties (Wayo et al., 2025). The present descriptor-based approach offers a more comprehensible and less complex alternative, especially in cases where computational efficiency and the transparency of the features' contribution to the descriptor are crucial. Similarly, the effectiveness of Random Forest and Extra Trees is correlated with ensemble methods in chemical reactivity prediction, where the ensemble of learners is able to represent the nonlinear relationships better than any of the individual learners (Cuesta et al., 2023).

The properties of the quantum-chemical databases may have a significant effect on the performance of the models, particularly when the distribution of properties and the composition of the molecules used for training and testing are similar (Vazquez-Salazar et al., 2021). This is specifically important for the almost ideal prediction of free energy. To illustrate its ability to learn quantum chemical patterns from structured inputs, the application of random forest to complex molecular quantities like the Hessian matrices is also discussed (Domenichini & Dellago, 2023). Further comparison with other methods like neural networks and classical algorithms demonstrated that the strength and regularity of correlations between molecular descriptors and calculated targets determine the performance of the algorithm (Normatov et al., 2025). The high accuracy attained for the HOMO–LUMO gap is also suited to the graph-based methods that explicitly incorporate the molecular connectivity for the prediction of the band gap (Abbassi & Ziti, 2025). The present results show that the performance of conventional descriptors can be quite good even though the graph models may include more structural information. Research on hybrid modelling also demonstrates the advantages of combining quantum-chemical data with other algorithms for accurate and meaningful predictions (Goshu, 2025).

The results show that the relatively simple machine-learning models can give fast estimates of important electronic and thermodynamic properties. These models can help in a preliminary screening of molecules, prioritize candidate molecules and minimize the need for repeated quantum-chemical calculations in the molecular design process. It is also worth noting from the results that selection of the model is important for the target property. For nonlinear electronic relationships, ensemble methods proved to be more appropriate, while regularized linear regression worked very well for the highly additive free-energy target. Descriptor rankings also offer chemical insight into prediction that can be understood.

There are a number of restrictions to be taken into account. The analysis was only applied to small molecules from the

QM9 database and external validation was carried out with no independent dataset. The exceptionally high free-energy performance might have been due partly to strong correlations among the size-related descriptors. Furthermore, the models were mostly based on the conventional molecular descriptors and did not necessarily include explicit 3D geometry.

The framework needs to be validated on chemically diverse datasets in future and try to use scaffold-aware/composition-aware data splitting. The redundancy of the descriptors needs to be explored further, especially for the molecular weight and atom-count descriptors. They would help determine if more complex representations offer any advantages and would provide a benchmark in the form of comparisons with graph neural networks, geometry-informed models and hybrid models. Estimation of the uncertainty and applicability domain should also be integrated to define the reliability of the predictions outside the original QM9 chemical space.

5. CONCLUSION

The machine-learning models were used to accurately and efficiently predict key electronic and thermodynamic properties of 133,885 small organic molecules of the QM9 database. Predictive performance was found to be dependent on the property; Random Forest achieved the best performance for both HOMO energy and the HOMO–LUMO gap, Ridge Regression was almost perfect for predicting free energy at 298.15 K, and Extra Trees predicted the heat capacity values best. These differences demonstrate the need to choose among models based on the linear and physicochemical nature of the target as opposed to a single algorithm for all properties. Descriptor analysis also indicated that the most important descriptors for prediction were exact molecular weight, molecular weight, oxygen atoms, carbon atoms, and hydrogen-bond acceptors. The significance suggests that the selected QM9 properties were greatly influenced by the molecular size and elemental composition. Residuals with low values around zero also contributed to low systematic bias and stability of the best models. In general, the framework provides a comprehensible alternative to the repeated quantum-chemical calculations for fast molecular screening and initial property estimation. Its applicability can be further enhanced, though, when model validation is performed on larger, chemically broader datasets in the future, as well as the geometry-aware modelling and the uncertainty analysis.

REFERENCES

1.Abbassi, O., & Ziti, S. (2025). QMGBP-DL: a deep learning and machine learning approach for quantum molecular graph band-gap prediction. Molecular Diversity, 29(4), 3501-3515.

2.Abel, R., Wang, L., Harder, E. D., Berne, B. J., & Friesner, R. A. (2017). Advancing drug discovery through enhanced free energy calculations. Accounts of chemical research, 50(7), 1625-1632.

3.Achar, T. K., Bose, A., & Mal, P. (2017). Mechanochemical synthesis of small organic molecules. Beilstein journal of organic chemistry, 13(1), 1907-1931.

4.Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., ... & Wolverton, C. (2022). Recent advances and applications of deep learning methods in materials science. npj Computational Materials, 8(1), 59.

5.Cuesta, S. A., Moreno, M., Lopez, R. A., Mora, J. R., Paz, J. L., & Marquez, E. A. (2023). ElectroPredictor: An application to predict Mayr's electrophilicity E through implementation of an ensemble model based on machine learning algorithms. Journal of Chemical Information and Modeling, 63(2), 507-521.

6.Davronov, R., & Adilova, F. (2021, July). A comparative analysis of the ensemble methods for drug design. In AIP Conference Proceedings (Vol. 2365, No. 1, p. 030001). AIP Publishing LLC.

7.Domenichini, G., & Dellago, C. (2023). Molecular Hessian matrices from a machine learning random forest regression algorithm. The Journal of Chemical Physics, 159(19).

8.Fedik, N., Zubatyuk, R., Kulichenko, M., Lubbers, N., Smith, J. S., Nebgen, B., ... & Tretiak, S. (2022). Extending machine learning beyond interatomic potentials for predicting molecular properties. Nature Reviews Chemistry, 6(9), 653-672.

9.Goshu, B. S. (2025). Enhancing Molecular Machine Learning with Quantum-Chemical Descriptors and Hybrid Modeling. International Journal of Advanced Nano Computing and Analytics, 4(2), 129-142.

10.Heid, E., Greenman, K. P., Chung, Y., Li, S. C., Graff, D. E., Vermeire, F. H., ... & McGill, C. J. (2023). Chemprop: a machine learning package for chemical property prediction. Journal of chemical information and modeling, 64(1), 9-17.

11.Huang, Y., Rong, C., Zhang, R., & Liu, S. (2017). Evaluating frontier orbital energy and HOMO/LUMO gap with descriptors from density functional reactivity theory. Journal of molecular modeling, 23(1), 3.

12.Iacovella, C., Chodera, J.& Yan, S. (2025). modelforge curated dataset: QM9 (Version full_dataset_v1.2) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.17536462

13.Jacobs, R., Polak, M. P., Schultz, L. E., Mahdavi, H., Honavar, V., & Morgan, D. (2024). Regression with large language models for materials and molecular property prediction. arXiv preprint arXiv:2409.06080.

14.Lim, M. A., Yang, S., Mai, H., & Cheng, A. C. (2022). Exploring deep learning of quantum chemical properties for absorption, distribution, metabolism, and excretion predictions. Journal of Chemical Information and Modeling, 62(24), 6336-6341.

15.Normatov, S., Nesterov, P. V., Aliev, T. A., Timralieva, A. A., Novikov, A. S., & Skorb, E. V. (2025). Search for Correlations Between the Results of the Density Functional Theory and Hartree–Fock Calculations Using Neural Networks and Classical Machine Learning Algorithms. ACS omega, 10(6), 5919-5933.

16.Pereira, F., Xiao, K., Latino, D. A., Wu, C., Zhang, Q., & Aires-de-Sousa, J. (2017). Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals. Journal of chemical information and modeling, 57(1), 11-21.

17.Pinheiro, G. A., Mucelini, J., Soares, M. D., Prati, R. C., Da Silva, J. L., & Quiles, M. G. (2020). Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset. The Journal of Physical Chemistry A, 124(47), 9854-9866.

18.Schmidt, J., Shi, J., Borlido, P., Chen, L., Botti, S., & Marques, M. A. (2017). Predicting the thermodynamic stability of solids combining density functional theory and machine learning. Chemistry of Materials, 29(12), 5090-5103.

19.Valdés, J. J., & Tchagang, A. B. (2024). Novel machine learning insights into the QM7b and QM9 quantum mechanics datasets. Journal of Computational Chemistry, 45(15), 1193-1214.

20.Vazquez-Salazar, L. I., Boittier, E. D., Unke, O. T., & Meuwly, M. (2021). Impact of the characteristics of quantum chemical databases on machine learning prediction of tautomerization energies. Journal of chemical theory and computation, 17(8), 4769-4785.

21.Wayo, D. D. K., Noor, M. Z. B. M., Ganji, M. D., Saporetti, C. M., & Goliatt, L. (2025). Q-DFTNet: A Chemistry-Informed Neural Network Framework for Predicting Molecular Dipole Moments via DFT-Driven QM9 Data. Journal of Computational Chemistry, 46(22), e70206.

22.Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., ... & Lei, M. (2019). Machine learning in materials science. InfoMat, 1(3), 338-358.

23.Westermayr, J., Gastegger, M., Schütt, K. T., & Maurer, R. J. (2021). Perspective on integrating machine learning into computational chemistry and materials science. The Journal of Chemical Physics, 154(23).

24.Yoğurtçu, H. N., & Ersanlı, C. C. (2025). Structural parameters, NLO, HOMO, LUMO, MEP, chemical reactivity descriptors, Mulliken-NPA, thermodynamic functions, hirshfeld surface analysis and molecular docking of 1, 3-Bis (4-methylphenyl) triazine. International Scientific and Vocational Studies Journal, 9(1), 130-144.