SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I4P124

Volume 11 Issue 4, 226-233, April 2024
© 2024 Seventh Sense Research Group®

Original Article

Machine Learning-Based Structure Prediction for QM9 Quantum Datasets

Nahla K[1], Maimoona Ansari[2], Salah Eldeen F. Hegazi[3], Anjali Appukuttan[4], Bincy Vincent[5], Huda Fatima[6]

[1]*School of Data Analytics, Convergence Complex Building, Mahatma Gandhi University, Kerala, India.*
[2]*CIT Technology, Hayalmatar, Jazan, SaudiArabia.*
[3]*Department of Chemical Engineering, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia.*
[4]*Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia.*
[5]*Department of Nursing, Darb College of Nursing, Jazan University, Jazan, Kingdom of Saudi Arabia.*
[6]*Department of Information Technology and Security, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia.*

[6]*Corresponding Author : hsaadullah@jazanu.edu.sa*

Received: 15 February 2024 Revised: 15 March 2024 Accepted: 16 April 2024 Published: 30 April 2024

Abstract - *The inner arrangement of the quantum mechanics dataset QM9 is investigated in this study. The dataset contains 1000 organic molecules as well as being defined in terms of electronic properties. To estimate the atomic composition using inverse molecular design attributes, one must understand the structure and properties of such data. The study used methods for detecting outliers, clustering, and intrinsic dimension analysis. The dataset was found to have descriptive dimensions far higher than their intrinsic dimensionality. Inliner items make up the majority of the inner core area of the QM9 data, whereas outliers dominate the outside region. The atom count in a molecule is strongly related to its outlier or inner character. Despite structural differences, important variables for inverse molecular design are very predictable. The molecular representation was estimated using Graph Neural Network (GNN), a modern Machine Learning (ML) algorithm. This study also did feature extraction and preprocessing before this algorithm. This proposed technique works for the outcomes.*

Keywords - *Feature generation, Feature selection, Machine Learning, Outlier analysis, QM9 data.*

1. Introduction

Computers may discover useful associations from unstructured data without any previous information thanks to ML techniques, which include complex algorithms. Feature extraction in traditional ML models has long relied on extensive domain knowledge and meticulous engineering to convert rare data into a usable illustration or feature vector, from which the learning subsystem could identify input patterns. By analysing raw data, representation learning may automatically find the representations needed for classification or detection [1].

Many branches of chemistry, including physical chemistry, drug discovery, and material science, are starting to rely on ML techniques. Aiming to computationally inexpensively and accurately forecast the governing atomic system properties, including energies along with forces, dipole moments, wave functions, and electron densities, Quantum-based ML (QML) approaches have made substantial strides in the last few years [2]. Assigning reactivity ratings to atoms and molecules may help shed light on the processes of chemical reactions in several fields, such as materials research, medicinal design, atmospheric chemistry, and chemical synthesis. Although measuring reaction rates experimentally has been done for quite some time, it is usually a laborious procedure that becomes more expensive as one attempts to investigate the most difficult processes.

Furthermore, the rates of molecular diffusion limit the actual reaction rates in the solution phase, making it more difficult to measure the extremes of reactivity. The third big, high-quality data set is the foundation of molecular ML models, which are essential for their success and widespread use. New molecular ML models may now be built with the help of the wider community of ML researchers, thanks to easily accessible and machine-actionable datasets. [4].

Some of the most significant advances in ML for molecular property prediction have come from building databases of small-molecule characteristics for use in prototyping and benchmarking new ML architectures. The most recent models have been tested on the QM9 dataset, which is one of several similar datasets [5]. In most MI techniques, there are three main components. The first

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

component consists of data sets comprising information on the assembly of the materials, measurement outcomes linked to these arrangements, and physical attributes pertinent to the material development objectives. Representing the data instances from the first component, the second part gathers a basic report of materials for use in identification as well as analogous extrapolation and then quantitatively characterises them.

The last component is a system that makes use of component algorithms or data mining algorithms to abstract data from the material data sets for particular drives, such as predicting properties or identifying novel material compositions as well as structures [6].

Generating massive volumes of data from electronic structure computations is the first step in applying quantum ML approaches to real-world situations. Therefore, data creation requires a lot of computer resources. Learning how to build appropriate databases to maximise the accuracy and transferability of the models is a crucial problem for expanding the use of ML approaches in chemistry. This stage relies on the human's level of understanding and trust in the link between the cause (the initial database and model) and the outcome (the model's application to a new assignment). This method, which has another name, "interpretability," helps to decipher the connections found in the model's training data or the model's learned associations.

An important aspect of this research is the attempt to establish a correlation between the original chemical databases used to train ML models and the model's ability to forecast a target attribute (tautomerization energy) on a set of samples that have never been seen before. [7] Due to their great mistake sensitivity, very short coherence durations, and overall complexity of manufacture, quantum computers are not yet in a commercially usable condition. When both classical and quantum algorithms fail to address an issue, a hybrid approach called hybrid quantum ML is used. [8]

The goal of achieving quantum dominance in quantum chemistry issues has led to the meticulous development of quantum or hybrid paradigms. The current methods rely heavily on molecular energy quantum simulation for accurate rate prediction in chemical reactions. Rather than demonstrating superiority over classical molecule learning methods, the goal of this work is to construct a complete quantum algorithm and illuminate quantum ML techniques for addressing molecular issues.

Unitary Coupled Cluster (UCC), a popular unsupervised learning technique with a specially constructed circuit for each molecule, is completely different from the proposed method in the quantum world. Graph learning can learn thousands of molecules and predict the characteristics of more complicated molecules, but it is not as accurate as molecular simulation methods for property prediction [9]. Although these methods make use of molecule datasets, research focusing on their structure is lacking. Inverse molecule design and related fields might benefit from understanding the information underlying structure and the properties of the data. This work aims to begin bridging that hole by means of unsupervised ML techniques that have been recycled in inverse molecular design research and that connect quantum mechanical features to atomic composition.

- This study formulates the issue from the perspective of inverse molecule design, builds upon prior research in several ways, and employs a huge dataset.
- Considering that the amount of an element's atoms in a molecule is a ratio variable as discrete values, it generates a multi-target forecast of the entire molecular composition.
- At the same time, taking advantage of the interplay between the chemical elements introduces methods for feature engineering.

This is the remaining structure of the paper: Part 2 explains the sources of the data used in the research, Part 3 details the set of ML techniques that were employed, Part 4 displays the outcomes, and Section 5 draws the final findings.

2. Literature Review

The article makes use of QM9 molecule datasets. They are a component of a bigger set of tools that can be used to speed up the process of creating accurate first-principles simulations of quantum-chemical systems.

Masahiro Sato, Hajime Shimakawa, and Akiko Kumada [10] described the use of ML methods as well as domain knowledge about materials. Data-driven material research has achieved a model shift. Nevertheless, when working with small-scale experimental datasets, ML-based studies have repeatedly disregarded the intrinsic constraint of extrapolative performance-the ability to forecast unknown data. This work provides a full-scale standard for measuring extrapolative performance on twelve different organic molecular characteristics.

When it comes to small-data properties in particular, an important level shows that traditional ML models degrade presentation significantly past the training distribution of property ranges as well as molecular structures. In order to tackle this problem, it provides QMex, a dataset of Quantum-Mechanical (QM) descriptors, and ILR, an interactive linear regression model that uses interaction details among QM descriptors along with statistical data about molecular structures.

The QMex-rooted ILR maintained its interpretability while achieving its existing extrapolative performance. To improve extrapolative predictions with short experimental datasets as well as to uncover new materials or molecules that outperform current candidates, benchmark results, the QMex dataset, as well as the proposed model are useful possessions.

Arghya Bhowmik, Tejs Vegge, Surajit Nandi, and Nandi [4] discussed that in the latest years, there has been a surge in the development of molecular ML methods, which has attracted more and more non-chemists to join the effort. This surge in activity is mostly due to the availability of well-curated, large datasets. Among the large databases of small molecules with B3LYP functional molecular energies, the QM9 dataset stands out.

The energies of these molecules, which were based on G4MP2, were subsequently reported. In order to facilitate a broad range of ML tasks involving QM9 molecules, such as transfer learning, multitask learning, delta learning, etc., a dataset containing QM9 molecule energies is expected using 76 distinct DFT functionals as well as 3 distinct basis sets, resulting in 228 energy numbers per molecule. The reaction energies were given based on these 76 functionals along with basis sets and further included all potential A ↔ B monomolecular interconversions in the QM9 dataset. Finally, it includes the bond modifications for each of the 162 million reactions so that ML-based reaction energy prediction algorithms may take structure and bond information into account.

Alain B. Tchagang and Julio J. Valdés [11] implemented the QM7b and QM9 molecule datasets, which have very distinct overall data structures, as shown via unsupervised analysis. In contrast to the former, which has a distinct two-cluster structure, the latter is divided into an outside area that mostly contains outliers and an inner core region where clusters of inliner items are concentrated. For QM9, the outlier/inliner nature of a molecule is strongly related to its atomic number; thus, molecules with very few or very many atoms tend to be outliers, while molecules with an average number of atoms tend to be inliners and clustered.

When creating predictive models for the de-novo molecules inverse design, it is important to consider these properties. There is a lot of duplication as the intrinsic dimension is much less than the dimension of the descriptor in both datasets. Even though they vary structurally, their qualities provide substantial predictive information on the molecular composition; for example, the original properties allow one to anticipate the atom count in the molecule properly. The predictive powers of embedding spaces with tiny dimensions are preserved.

Gaul, Christopher, and Santiago Cuesta-Lopez [12] highlighted an ML model trained to efficiently along with precisely estimate the energies of a Molecular Structure’s Highest Occupied (HOMO) and Lowest Unoccupied (LUMO) orbitals. It incorporates a “Set2Set” readout module and is based on the SchNet model. When dealing with complicated values, the Set2Set module is superior to sum and average aggregation in terms of expressive capacity.

A large majority of the earlier models have been trained and tested on relatively tiny molecules. As a result, the second contribution is to create a consistent train/validation/test split and to broaden the scope of ML algorithms to include bigger molecules from other sources. A multitask approach is developed as a third contribution to address the issue of sources originating from distinct theoretical levels.

With the combined efforts of all three, the model’s precision approaches that of a chemical model. Because it is trained using the precise molecular geometries derived via DFT geometry optimisations, employing the existing model for such applications presents a problem. The structures produced by the generating algorithm will not have precise geometries, which might lead to the model making inaccurate predictions. Research on how the model’s accuracy is affected by noise in the input coordinates is therefore necessary. Improving the model’s training process by introducing random noise into the input geometries should fix the issue.

Pande, Vijay S., Sinitskiy, and Anton V. [13] developed two methods for molecular system modelling that are quite efficient in practice. A physical technique for estimating the energies and electron densities of molecules is Density Functional Theory (DFT), the commonly employed quantum chemistry technique. The ML of molecular characteristics has also seen a flurry of recent publication activity.

When compared to DFT, ML models have much lower computing costs and may achieve similar accuracy; nevertheless, their lack of physicality, which is a direct connection to quantum physics, restricts their use. The physicality and cheap computing costs of DFT and ML are combined in the proposed method. The generic equations for accurate electron densities and energies may be used to naturally direct ML applications in quantum chemistry. This is achieved by generalising the well-known Hohenberg-Kohn theorems.

Utilising these equations as a foundation, a deep neural network is capable of outperforming existing DFT implementations in terms of speed and accuracy when it comes to computing the electron densities and energies of various organic molecules. Specifically, the average absolute error in the molecules energies containing eight non-hydrogen atoms was as little as 0.9 kcal/mol compared to the values obtained using CCSD (T). This is much lower than the errors seen when using DFT (down to around 3 kcal/mol on the matching molecule set) as well as ML (down to about 1.5 kcal/mol). The proposed method outperforms past DFT functionals created by “human learning” in terms of the prediction of electron densities and energies. As a result, ML grounded in physics has promising prospects for the impending modelling of far bigger molecular systems than is already achievable, with a degree of quantum chemical precision consistent with high-level theoretical predictions.

S. Heinen, A. O. von Lilienfeld, and von Rudorff [14] discussed the optimisation of geometry, searching for transition states using legacy optimizers and using energies as well as forces anticipated in response Operator-based Quantum ML (OQML). Relaxation routes up to 5,500 constitutional isomers with the sum formula $C_7H_{10}O_2$ from the QM9 dataset for geometry optimisations.

Reproducing the lowest geometry with an RMSD of 0.14 Å is achieved by employing the obtained OQML models along with an LBFGS optimizer. On average, the findings from the MP2 reference differ by 14 cm for the convergent equilibrium geometries and 26 cm−1 for the transition state geometries, as determined by the following vibrational normal mode frequency analysis. An Amon-based extension might also be applied to OQML to make it more portable and adaptable to bigger reactants. The production of bigger and more reliable data sets in quantum chemistry, particularly for reaction studies, might benefit from OQML as well.

Hoja et al. [15] described the QM7-X dataset, which comprises 42 physicochemical attributes for about 4.2 million non-equilibriums as well as equilibrium structures of minor organic molecules, including up to seven non-hydrogen atoms (C, O, N, S, and Cl). The global (molecular) as well as local (atom-in-a-molecule) QM7-X properties, which were calculated at the strictly convergent quantum mechanical PBE0+MBD level of theory, range from ground state quantities to response quantities. QM7-X will be an essential component of next-generation ML models for exploring larger areas of CCS and designing molecules with desired properties through the provision of a comprehensive, organised, along tightly converged dataset of physicochemical properties computed by quantum mechanics.

Dominik Lemm, O. Anatole von Lilienfeld, and Mario Falk von Rudorff [16] describe the long-standing issue in physics, biology, chemistry, as well as materials science with the computer prediction of atomic structure. The traditional approach to structure determination, using force fields or ab initio approaches, involves energy reduction, which may be computationally intensive or yield approximations.

Synthetic large data sets that account for chemical space with atomic resolution cannot be generated due to this accuracy/cost trade-off. The Graph-to-Structure (G2S) ML model uses implied correlations between calm structures in training data sets to generalise across compound space and infer interatomic distances for out-of-sample compounds. This allows us to reconstruct coordinates directly, avoiding the traditional energy optimisation problem. Successful predictions for systems that normally need human intervention, enhanced initial estimates for future conventional ab initio-based relaxing, along input creation for the usage of structure-routed quantum ML models are all examples of testing G2S's applicability.

3. Proposed Methodology

Figure 1 shows the system architecture that the study employed for this procedure. Before entering the model search phase, the data undergoes preprocessing and specialised feature engineering. In order to perform feature engineering activities, robust feature selection and generation methods were pre-selected, and the search algorithm did not have to include the simple standardisation that was necessary for preprocessing the QM9 data. Consequently, GNN is used for the prediction of molecular characteristics.

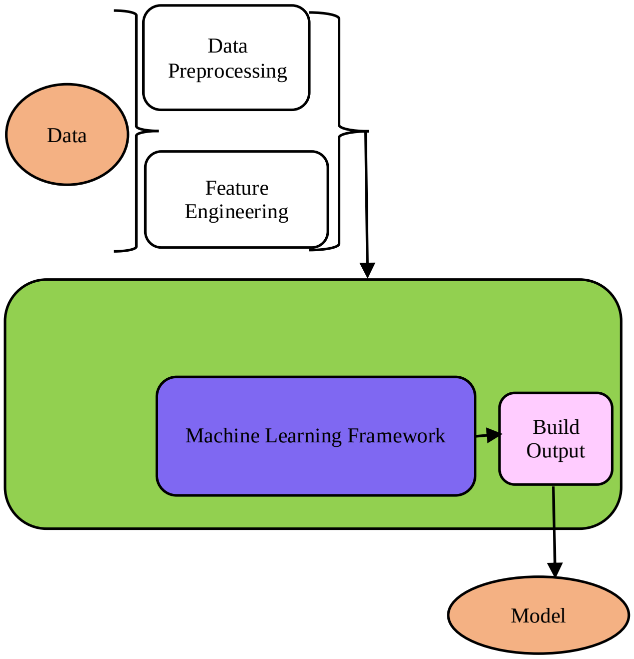

Fig. 1 Proposed model search process architecture

3.1. Data

The chemical formulas for the compounds (QM9) include 133,885 organic molecules, with each molecule containing Nt = 5 of the subsequent elements: Carbon (C), Oxygen (O), Nitrogen (N), Hydrogen (H), as well as Chlorine (Cl). They stand in for the goals of the model in inverse design methods. The molecules are defined by Nv =19 electronic attributes that are calculated using quantum chemistry methods and include geometric, energy, electronic, and thermodynamic properties. In contrast to the QM7b data, which does not include any constitutional isomers, QM9 has 6095 of them out of 134k molecules. Results for a related, consistent, as well as exhaustive chemical space of tiny organic molecules are shown in Table 1 and are believed to be provided by this dataset in terms of highly precise quantum chemical characteristics.

Table 1. QM9: molecules properties

Index	Molecular Property	Explanation
1	g298_atom	298.15K Free Atomisation Energy
2	u298_atom	298.15k Atomization Enthalpy
3	u0_atom	Atomization Energy at 0K
4	cv	Heat Capacity
5	g298	298.15K Free Energy
6	h298	298.15K Enthalpy
7	u298	298.15K Internal Energy
8	u0	Internal Energy at 0K
9	zpve	Zero-Point Vibrational Energy
10	r2	Electronic Spatial Extent
11	Gap	Difference among HOMO as Well as LUMO
12	LUMO	Lowest Unoccupied Molecular Orbital
13	HOMO	Highest Unoccupied Molecular Orbital
14	Alpha	Norm of Static Polarizability
15	Mu	Norm of Dipole Moment
16	C	Rotational Constant
17	B	Rotational Constant
18	A	Rotational Constant

3.2. Preprocessing

QM9 dataset uses different units of measure to convey the original features of the data. A wide range of statistical and ML approaches are susceptible to biases introduced by variables evaluated at various scales, which do not contribute equally to the study. This research preprocessed all datasets by transforming the unique property values into z-scores to ensure the values were similar and to remove any potential bias. This is completed before applying the ML algorithms defined in the next section. By stating all attributes in the same measure unit (variance), eliminate the source of bias by giving the new variables a unit variance and a zero mean. Then, randomly split the 133,885 items from the QM9 dataset into a training set of 90,120,496 objects and a testing set of 13,389 objects to create predictive models using supervised ML techniques.

3.3. Feature Selection

A plethora of descriptive properties characterize the datasets under consideration here, as is true with many real-world datasets. Unexpectedly, many of them are either unrelated to the task at hand (regression or classification) or too noisy. Many ML algorithms are known to show a drop in accuracy when dealing with feature sets that are too big, especially when the number of variables is far larger than what is considered ideal. Not to mention the more mundane concerns of making algorithms slower and consuming more processing resources.

Markov chain-based approaches have successfully solved the rank aggregation issue. The provided (partial) lists form the basis of their transition probabilities, while the chain states signify the candidates requiring ranking. The feature space is the amalgamation of state sets ordered by different selection methods, as well as the method searches for a stable delivery of transition probabilities. When the chain's current state is feature P, the MC4 Markov version employs the following procedure to select the next state:

1. Choose a feature Q uniformly from the amalgamation of every list ranked by the assortment methods;
2. If τ(Q)<τ(P) for the lists majority that rank both P along with Q, then move to Q; otherwise, remain in P.

Copeland proposed selecting candidates according to their win rate in pairwise majority contests; this chain expands on that idea. In addition to improving previously proposed algorithms, the Markov chain technique can handle incomplete lists of candidates (features). It also beats other classical rank aggregation methods on the most popular criteria while being computationally efficient. This article employs the MC4 rank aggregation approach.

3.4. Prediction

GNNs and other new methods have recently made it possible to automatically extract useful characteristics from molecular networks, doing away with the need for the time-consuming and error-prone process of manually creating descriptors. Generally, ML algorithms have been used for this process so far. However, this work makes use of the GNN technique, which effectively predicts the structure.

Beyond simply the end-to-end learning of a data-driven molecular representation, a GNN technique has further advantages. Expand the model to incorporate atomic pairwise distances and other minuscule information. Here, construct a more accurate and reliable molecular predictor. GNNs generally adhere to a recursive neighborhood aggregation approach.

In forward propagation, the hidden states of neighboring nodes are aggregated and transformed iteratively to update the node state. In the forward propagation step of Graph Convolutional Networks (GCN), for instance, a hidden state is updated in Equation (1).

$$h_v^{(l+1)} = \sigma\left(\sum_{w \varepsilon N_V} \frac{1}{C_{vw}} W_1^{(l)} h_w^{(l)} + W_0^{(l)} h_v^{(l)}\right) \qquad (1)$$

In this context, $h_v^{(l)}$ represents the hidden state of node v at the l-th layer, $W^{(l)}$stands for trainable weight, σ suggests an activation function like ReLU, c_{vw} is a normalization constant like $c_{vw} = \sqrt{deg(v)deg(u)}$ and $deg(v)$ is the degree of node v . A lot of work has gone into making GNNs more expressive inside the recursive aggregation framework [17].

4. Results and Analysis

In order to assess how well the proposed strategy works, this part does an analysis along with some comparisons. Table 2 and Figure 2 provide the findings for the QM9 testing sets.

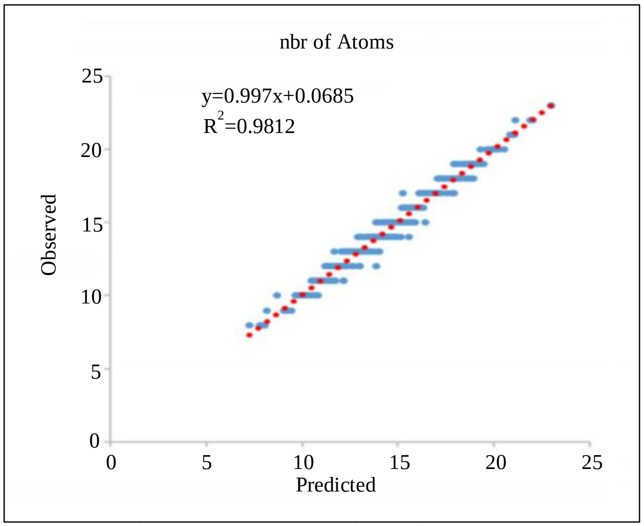

Fig. 2 Relationship among the predicted as well as actual atoms counted in the data for the molecule set with every feature in the QM9 testing set

Table 2. The atom count in a molecule is predicted using regression tree models for QM7b

Features	MSE	MAE	RMSE	R	No. of Rules
Original 19	0.0061	0.0315	0.0781	0.9997	46
Generated 2	0.8917	0.3382	0.9443	0.9475	358

Table 3 demonstrates that the molecular weight predictability is quite good when incorporating all original features. Anticipate the models derived from embedding spaces to be more intricate than those derived from the original features based on the feature space values.

Table 3. Evaluation of the molecular weight's predictability using both the original qualities and the derived features

No. of Features	Gradient	Gamma	Vratio
2	5.7763	8.9940	0.1571
5	136.8410	0.0918	0.0016
10	122.7370	0.0836	0.0015
19 Original	0.1140	0.1160	0.0020

Table 4 includes metrics including the Mean Absolute Error (MAE), coefficient of determination (R2), Root Mean Squared Error (RMSE), as well as Mean Squared Error (MSE) to guarantee compatibility with the proposed technique.

Table 4. Performance metrics with 19 properties

	Metrics (19 Properties)				
MSE	RMSE	MAE	Molecule	R2	Time (h)
0.0032	0.0571	0.0065	97.81%	0.9941	23.50
	Acc (19 Properties)				
Carbon	Hydrogen	Nitrogen	Oxygen	Fluoride	Time (h)
99.72%	98.47%	99.38%	99.79%	95.75%	23.50

Table 5 shows the testing set results for forecasting atomic composition employing all 19 original attributes. The model shows extremely excellent quality since all error measures were tiny, as well as the R2 value in particular. In addition, there was no overfitting, as the testing and training mistakes were almost identical. It calculated the specific accuracies of the chemical elements. All of the element-wise accuracies are quite high; the bottom one is for fluoride (95.75%), which is already very high. This model also obtains very high accuracy (97.81%) for predicting the complete molecule composition. Using the electronic characteristics of molecules as inputs, this method enables highly precise composition calculations.

Table 5. Testing set performance for forecasting atomic composition with 10 characteristics

	Metrics				
MSE	RMSE	MAE	Molecule	R2	Time (h)
0.0025	0.0520	0.0052	99.52%	0.9939	23.40
	Acc				
Carbon	Hydrogen	Nitrogen	Oxygen	Fluoride	Time (h)
99.80%	98.52%	99.48%	99.86%	95.45%	23.40

Table 6 provides the results obtained from the finest discovered model and also illustrates the model search procedure based on the 10 specified features. There are irrelevant properties, noise, and detrimental interactions in the initial collection of properties because the model quality metrics are healthier than the ones produced when considering entire features. There is a statistically significant difference between utilizing all features and using 10 features for most targets, suggesting that employing 10 features improves prediction accuracy. Due to the similarity between the training and testing errors, the improvement achieved while utilizing the specified features cannot be attributed to overfitting.

Table 6. Comparison table

Methods	MSE	RMSE	Acc (%)
Proposed	0.0025	0.0520	99.52
Valdés et al. [18]	0.0027	0.0522	99.49
GNN	0.0030	0.0528	98.04

Various approaches' performance parameters are compared in Table 6. Compared to other models like GNN and the ensemble method in [18], the proposed model seems to perform better across all parameters. On the QM9 dataset, the proposed model has an accuracy of about 99.52%. With a performance gap of 0.03%, this model beats GNN by 1.50% [18].

The MSE of the proposed model is 0.0025, which is 8% better than the model in [18] and 20% better than the GNN model. The RMSE of the proposed model is 0.0520, which is 0.38% better than the model in [18] and 1.53% better than the GNN model. This improvement in the proposed model is due to the use of a feature engineering section along with the prediction model.

5. Conclusion

For QM9, the outlier/inliner nature of a molecule is strongly related to its atomic number; thus, molecules with very few or very many atoms tend to be outliers, while molecules with an average number of atoms tend to be inliners and clustered. When creating predictive models for the inverse design of denovomolecules, it is important to consider these features. There is a lot of duplication as the intrinsic dimension is much less than the descriptor dimension. Even though they vary structurally, their qualities provide substantial predictive information on the molecular composition; for example, the original properties allow one to anticipate the number of atoms in the molecule properly. Predictive powers are preserved even in embedding environments of low dimension.

Supporting the proposed method, models that predicted the atomic composition of molecules for the QM9 dataset utilizing all electronic attributes as predictors performed very well per all major model quality criteria. Hence, the proposed GNN approach with feature selection and preprocessing proved to be better than the other existing approaches.

References

[1] Raghunathan, Shampa, and U. Deva Priyakumar, "Molecular Representations for Machine Learning Applications in Chemistry," *International Journal of Quantum Chemistry*, vol. 122, no. 7, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Clemens Isert et al., "QMugs, Quantum Mechanical Properties of Drug-Like Molecules," *Scientific Data*, vol. 9, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mohammadamin Tavakoli et al., "Quantum Mechanics and Machine Learning Synergies: Graph Attention Neural Networks to Predict Chemical Reactivity," *Journal of Chemical Information and Modeling*, vol. 62, no. 9, pp. 2121-2132, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Surajit Nandi, Tejs Vegge, and Arghya Bhowmik, "MultiXC-QM9: Large Dataset of Molecular and Reaction Energies from Multi-Level Quantum Chemical Methods," *Scientific Data*, vol. 10, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Surajit Nandi, Tejs Vegge, and Arghya Bhowmik, "Large Dataset of Molecular and Reaction Energies from Multi-Level Quantum Chemical Methods," *ChemRxiv*, pp. 1-7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Tien-Sinh Vu et al., "Towards Understanding Structure-Property Relations in Materials with Interpretable Deep Learning," *NPJ Computational Materials*, vol. 9, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Luis Itza Vazquez-Salazar et al., "Impact of the Characteristics of Quantum Chemical Databases on Machine Learning Prediction of Tautomerization Energies," *Journal of Chemical Theory and Computation*, vol. 17, no. 8, pp.4769-4785, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Maria Avramouli et al., "Quantum Machine Learning in Drug Discovery: Current State and Challenges," *Proceedings of the 26th Pan-Hellenic Conference on Informatics*, pp. 394-401, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ge Yan, Huaijin Wu, and Junchi Yan, "Quantum 3D Graph Learning with Applications to Molecule Embedding," *Proceedings of the 40th International Conference on Machine Learning*, pp. 39126-39137, 2023. [Google Scholar] [Publisher Link]

[10] Hajime Shimakawa, Akiko Kumada, and Masahiro Sato, "Extrapolative Prediction of Small-Data Molecular Property Using Quantum Mechanics-Assisted Machine Learning," *NPJ Computational Materials*, vol. 10, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Julio J. Valdés, and Alain B. Tchagang, "Understanding the Structure of qm7b and qm9 Quantum Mechanical Datasets Using Unsupervised Learning," *arXiv*, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Christopher Gaul, and Santiago Cuesta-Lopez, "Machine Learning for Orbital Energies of Organic Molecules Upwards of 100 Atoms," *Physica Status Solidi*, vol. 261, no. 1, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Anton V. Sinitskiy, and Vijay S. Pande, "Physical Machine Learning Outperforms "Human Learning" in Quantum Chemistry," *arXiv*, pp. 1-62, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] S. Heinen, G.F. von Rudorff, and O.A. von Lilienfeld, "Geometry Relaxation and Transition State Search throughout Chemical Compound Space with Quantum Machine Learning," *arXiv*, pp. 1-7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Johannes Hoja et al., "QM7-X, A Comprehensive Dataset of Quantum-Mechanical Properties Spanning the Chemical Space of Small Organic Molecules," *Scientific Data*, vol. 8, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Dominik Lemm, Guido Falk von Rudorff, and O. Anatole von Lilienfeld, "Machine Learning Based Energy-Free Structure Predictions of Molecules, Transition States, and Solids," *Nature Communications*, vol. 12, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Guangyong Chen et al., "Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models," *arXiv*, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Julio J. Valdés, and Alain B. Tchagang, "Novel Machine Learning Insights into the QM7b and QM9 Quantum Mechanics Datasets," *Journal of Computational Chemistry*, vol. 45, no. 15, pp. 1193-1214, 2024. [CrossRef] [Google Scholar] [Publisher Link]