ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

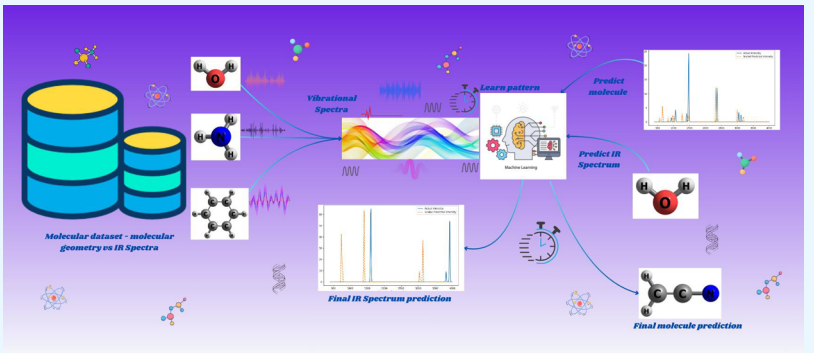
Unlocking the Potential of Machine Learning in Enhancing Quantum Chemical Calculations for Infrared Spectral Prediction

Adithya Ranjith Kartha, Dhanush P. Ajayakumar, Muhammad Idris, and Gopi Ragupathy*

Cite This: *ACS Omega* 2025, 10, 19224−19234

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



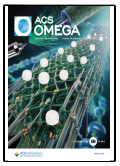
ABSTRACT: Infrared (IR) spectroscopy is a fundamental tool for analyzing molecular structures and chemical interactions by identifying the vibrational modes of molecules. Traditional quantum mechanical methods, such as density functional theory, are highly accurate but computationally expensive and impractical for large-scale molecular systems. This project investigates the integration of machine learning (ML) techniques to predict IR spectra, offering a promising alternative that significantly reduces computational costs while maintaining high accuracy. Additionally, the project explores the utilization of IR spectra for molecular identification and classification into molecular families, enhancing the practical utility of spectral data in various scientific applications. Using TensorFlow-based ML frameworks, models were developed and trained on a data set derived from high-quality computational chemistry analyzers. These data sets, sourced from computationally optimized geometry and IR spectrum from the Gaussian 16 Program Suite, include extensive molecular geometry data, bond lengths, vibrational modes, and other quantum mechanical properties. The models aim to predict key IR spectral features, such as vibrational frequencies and intensities, while maintaining interpretability by linking chemical and quantum mechanical principles to predictions. The integration of ML with IR spectroscopy provides a scalable as well as accelerated solution for analyzing complex molecular systems. This approach holds potential in fields such as drug discovery, materials science, and chemical engineering, where rapid and accurate spectral predictions are critical. This perspective highlights the advancements achieved, the current challenges, and the future potential of ML in the context of IR spectroscopy, providing a solid foundation for further exploration at the intersection of chemistry and data science.

1. INTRODUCTION

Infrared (IR) spectroscopy is a fundamental analytical tool in chemistry, offering profound insights into molecular vibrations, chemical bonds, and structural interactions.[1,2] Its applications span diverse fields such as materials science, biochemistry, and drug discovery, where rapid and accurate molecular characterization is crucial.[3−5] However, traditional quantum mechanical methods like Density Functional Theory (DFT) and Hartree−Fock (HF), while delivering high accuracy in predicting IR spectra, are computationally intensive and time-consuming, especially for large and complex molecular systems.[6] These limitations present a significant bottleneck, particularly in high-throughput applications and industries that demand scalable solutions.

Ab initio quantum mechanical methods have traditionally been the gold standard for predicting infrared (IR) spectra, with Density Functional Theory (DFT) and Møller-Plesset perturbation theory (MP2) being widely applied for gas-phase predictions.[7−10] Commonly used functionals, such as B3LYP and M05, are effective but often exhibit quantitative discrepancies with experimental data due to factors like anharmonicity, finite temperature effects, and the need for scaling factors. These limitations can result in frequency errors of up to 30 cm^{-1}.[11−14] Higher-level methods, including CCSD(T) and QCISD, offer improved accuracy but are

Received: March 14, 2025
Revised: March 21, 2025
Accepted: March 27, 2025
Published: April 28, 2025



https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

computationally prohibitive and restricted to smaller molecular systems.[15,16]

For condensed-phase spectra, ab initio molecular dynamics (AIMD) has gained popularity for first-principles IR predictions. However, AIMD is computationally intensive and highly dependent on the quality of the underlying force fields. Both gas-phase and condensed-phase predictions demand substantial computational resources and significant user expertise, which pose challenges to their broader applicability in practical and high-throughput scenarios.[17,18]

The emergence of ML offers a groundbreaking alternative to these computational challenges.[19] By training on precomputed quantum mechanical data sets, ML models can predict molecular properties, including IR spectra, with remarkable speed and efficiency.[20] These models leverage intricate molecular features—such as bond lengths, vibrational modes, dipole moments, and atomic charges—to predict properties that traditionally require computationally intensive quantum mechanical calculations. By learning the complex relationships between these features and target outputs, ML models eliminate the need for direct quantum calculations, offering a more time-efficient and cost-effective solution. This data-driven approach ensures that accuracy is preserved, as the models are trained on high-quality data sets derived from rigorous quantum mechanical simulations.[21] Furthermore, ML allows for flexibility in exploring molecular properties across diverse chemical families, making it an invaluable tool for tasks like IR spectra prediction, where the relationship between molecular structure and vibrational characteristics is often nonlinear and highly complex.[21]

Tools like TensorFlow have facilitated the development of cutting-edge neural network architectures that are particularly well-suited for handling molecular data sets.[22,23] Convolutional Neural Networks (CNNs) are designed to detect spatial patterns and hierarchical structures, making them ideal for analyzing molecular geometries, bond lengths, and angle distributions. By focusing on localized features and their combinations, CNNs can effectively model the spatial dependencies within a molecule. On the other hand, Graph Neural Networks (GNNs) represent molecules as graphs, where atoms serve as nodes and bonds act as edges.[24] This representation allows GNNs to model both the local and global interactions within molecules, capturing subtle atomic relationships and bond connectivity that influence vibrational properties. Moreover, these architectures excel in handling the multidimensional and highly interdependent nature of molecular data sets.[25] For instance, CNNs can analyze molecular descriptors encoded as grids or matrices, while GNNs offer a natural way to interpret molecules as dynamic and flexible systems. Both models benefit from TensorFlow's scalability, allowing for efficient processing of large data sets containing thousands of molecules and their corresponding vibrational properties. Additionally, the integration of custom loss functions and activation layers in TensorFlow provides the ability to optimize the networks for specific tasks, such as predicting vibrational frequencies or intensities with high precision.[24−26]

These advanced neural network architectures also enable generalization across diverse molecular systems. For example, once trained, the models can predict properties for new or previously unseen molecules, extending their utility to practical applications like drug discovery, materials design, and chemical engineering.[27,28] The ability to accurately predict IR spectra and classify molecular families using these architectures has opened new possibilities in leveraging spectral data for both theoretical and experimental studies. As a result, ML has emerged as a transformative tool in computational chemistry, bridging the gap between quantum mechanical rigor and real-world scalability.[29]

Researchers stand to benefit immensely from this paradigm shift toward ML in molecular analysis. ML models enable rapid and efficient analysis of complex molecular properties, reducing the time traditionally required for computational simulations and facilitating faster insights into molecular behavior.[20] By automating predictions of molecular properties like IR spectra, ML models free up valuable computational and human resources for experimental validation and the exploration of new hypotheses, accelerating the pace of scientific discovery.[30,31] This shift is particularly significant in fields where experimental resources are often limited, allowing researchers to focus more on testing innovative ideas rather than being bogged down by the computational challenges of large-scale data processing.[32]

Furthermore, by eliminating the computational bottleneck inherent in traditional quantum mechanical methods, ML models allow researchers to handle much larger molecular libraries than ever before. These models can predict molecular properties for thousands of compounds in a fraction of the time it would take using conventional approaches, enabling scientists to focus on discovery rather than the constraints of time-consuming calculations.[33] This enhanced capability extends beyond academia to industries such as pharmaceuticals, materials science, and environmental chemistry, where large-scale data analysis and molecular screening are critical.[34] These sectors are increasingly willing to invest in ML technologies, recognizing their potential to revolutionize research and development processes.

The ability to save time, resources, and computational costs while scaling operations makes ML-driven IR spectroscopy not only a breakthrough in scientific research but also a commercially viable innovation. In industries like pharmaceuticals, the speed and accuracy of ML models can significantly expedite drug discovery processes, enabling researchers to screen vast libraries of compounds more efficiently. Similarly, in materials science, ML models can help accelerate the characterization of new materials, identifying promising candidates for specific applications without the need for extensive experimental testing. In environmental chemistry, these models can be applied to monitor pollutants, analyze contaminants, and predict chemical behaviors in various ecosystems. The commercial advantages of ML in these fields are clear, offering industries the ability to streamline processes, reduce costs, and ensure faster time-to-market for new products.[35]

Moreover, the scalability of ML techniques further enhances their value to industries, allowing them to process and analyze ever-expanding data sets with ease. This capacity to handle large volumes of data efficiently supports not only research and development but also regulatory compliance, quality control, and predictive maintenance in manufacturing processes. As these technologies mature, machine learning-driven IR spectroscopy has the potential to become a cornerstone in the industries it serves, driving innovation, improving decision-making, and delivering more cost-effective solutions for businesses and consumers alike.[36,37]

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

The scope of this project involves developing, training, and validating ML models designed to predict IR spectra based on quantum mechanical data. This process includes curating robust data sets derived from high-quality computational calculations, designing sophisticated ML architectures tailored to the problem, and optimizing model performance for accurate predictions. The goal is to establish machine learning as a reliable and efficient tool for IR spectroscopy, one that can streamline molecular analysis and reduce computational costs. Rigorous validation against both experimental data and theoretical benchmarks is essential to ensure the models' accuracy, scalability, and generalizability across a wide range of molecular systems. Furthermore, this study explores various potential applications across scientific and industrial domains, demonstrating how these innovations can address real-world challenges in fields like drug discovery, materials science, and environmental monitoring.

By uniting the computational rigor of quantum chemistry with the efficiency and scalability of ML, this project not only advances the capabilities of IR spectroscopy but also highlights its transformative potential for both researchers and industries alike. The synergy between these two fields promises to redefine how molecular properties are analyzed, accelerating innovation and discovery across multiple disciplines. This work lays the foundation for future advancements in both fundamental research and practical applications, offering a novel approach to molecular characterization that could shape the future of industries reliant on chemical analysis.

2. OVERVIEW OF ML TECHNIQUES USED

2.1. Multioutput Regressor. The Multioutput Regressor is a wrapper model that extends single-output regressors to handle multiple target variables simultaneously. As you can see in Figure 1, a separate model is trained for each output variable, making it well-suited for multidimensional regression problems. For instance, in predicting IR spectra, the frequencies and intensities of different vibrational modes are distinct but related outputs. By training individual regressors for these outputs and combining their results, the Multioutput Regressor enables efficient and independent prediction of each target variable. Although the models for different outputs are independent, this technique allows for parallel processing and easier handling of multi-output tasks without requiring the development of a specialized algorithm.[38]

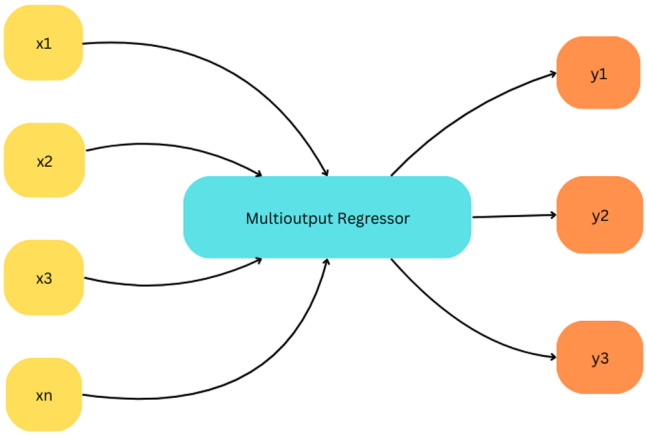


Figure 1. Multi-output regressor model data flow of inputs and outputs.

2.2. Random Forest Regressor. The Random Forest Regressor is an ensemble ML model that builds multiple decision trees during training and combines their predictions to produce a final output. Each decision tree trains on a random subset of the data, which introduces diversity into the learning process and reduces overfitting. In Figure 2, we see the model averaging the outputs of all trees in the ensemble improves predictive accuracy and enhances robustness against noise compared to a single decision tree. Random Forest works well for regression tasks, like predicting continuous variables (e.g., IR frequencies and intensities), because it can model complex, nonlinear relationships in the data. The model is also interpretable, allowing insight into feature importance, which can help identify key factors influencing predictions.[39]

The Random Forest Regressor serves as the base model within the Multioutput Regressor, leveraging an ensemble of decision trees to predict continuous variables like IR frequencies and intensities. By averaging predictions from multiple trees trained on random subsets of the data, it provides high accuracy and robustness against overfitting. Random Forest effectively models the complex, nonlinear relationships between molecular features and spectral properties, while its interpretability allows for insights into feature importance, highlighting key factors that influence predictions.

Spectroscopy prediction has been explored using various machine learning methods, each with its strengths and limitations. Deep learning captures complex spectral correlations but requires large data sets, while SVMs depend on careful kernel selection. We chose Random Forest for its balance of accuracy, efficiency, and interpretability. Our model predicts IR spectra, infers molecular identity, and classifies molecular families, expanding ML applications in spectral analysis. Future work should explore hybrid models or direct comparisons to enhance performance.

3. MODEL ARCHITECTURE AND DESIGN

3.1. Start and Data Collection. The process begins with data acquisition: The data sets are sourced from computationally optimized geometry and IR spectrum from Gaussian 16 Software Package.[72] Details of the selected molecules including their SMILES notation is included in the Supporting Information as Table 2.7, page number 15. The monomer geometries were initially optimized at the B3LYP level of theory using the 6−311++G (d, p) basis set. The reported IR spectra are based on the raw computed frequencies without any adjustments. However, we exported the data using ChemCraft visualization software with a constant bandwidth of 30 for each molecule. These optimized geometries were then used as starting points for further calculations without imposing any constraints on the structural parameters. Vibrational frequency calculations were subsequently performed to verify that the geometries correspond to minima on the potential energy surface and to assist in assigning experimental frequencies.[9,10]

Vibrational data is loaded from Vibrational_Data.csv, which contains molecular vibrational frequencies and their corresponding IR intensities.

Molecular properties, such as atomic coordinates (X, Y, Z), dipole moment, and point group, are extracted from Molecular_Coordinates.xlsx.

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

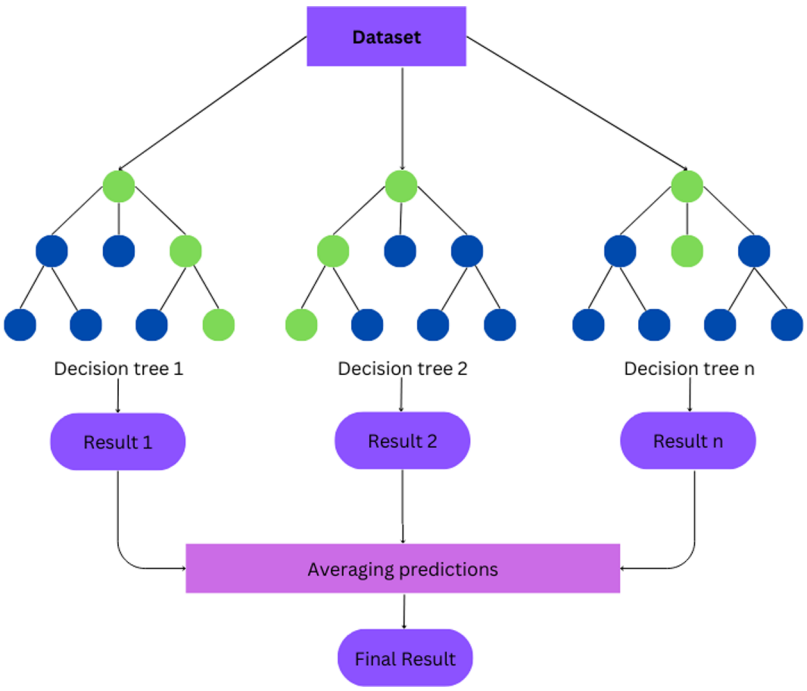


Figure 2. Random forest model decision tree structures and working of final analysis to produce result.

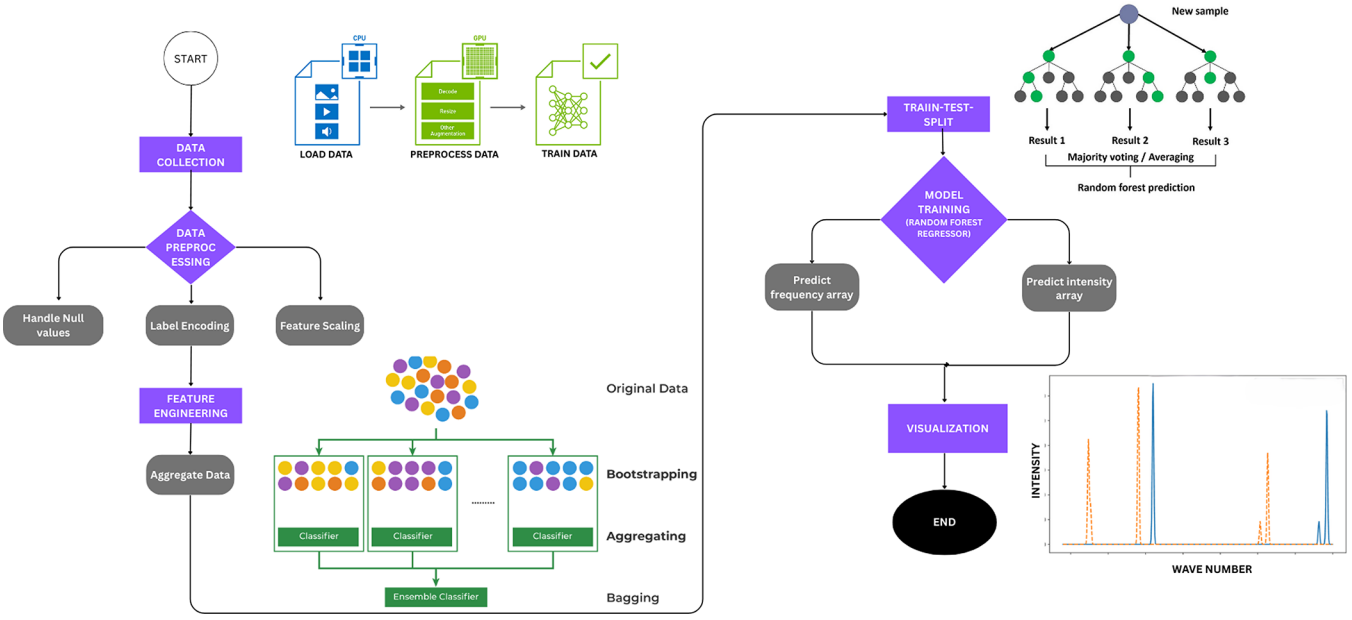


Figure 3. Flowchart representing an end-to-end machine learning pipeline indicating stepwise processing and data manipulation.

The features chosen for predicting IR spectra, such as molecular coordinates,[40] number of bonds, dipole moment, and point group symmetry, are directly tied to the physical and quantum properties[41] influencing vibrational modes and their IR activity. Molecular coordinates provide a 3D representation of atomic arrangements, which determine bond lengths and angles, critical for vibrational frequency calculation.[42] The number of bonds captures the molecule's structural complexity, influencing the total vibrational modes. Dipole moment is key for determining IR activity, as only vibrational modes causing a dipole change are IR-active, and its magnitude affects absorption intensity.[43] Point group symmetry encodes molecular symmetry, simplifying the analysis by indicating which modes are IR-active based on selection rules.[44]

These data sets form the foundation for the predictive model as shown in Figure 3.

3.2. Data Preprocessing. Once the data was collected, it underwent a series of cleaning and transformation steps to ensure it was ready for analysis. First, we handled null values. Missing data was addressed to avoid inconsistencies in training. Next, feature scaling normalized numerical features like coordinates and dipole moments to ensure uniform input ranges. Lastly, label encoding converted categorical features (e.g., Point Group) into numerical values to make them usable

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

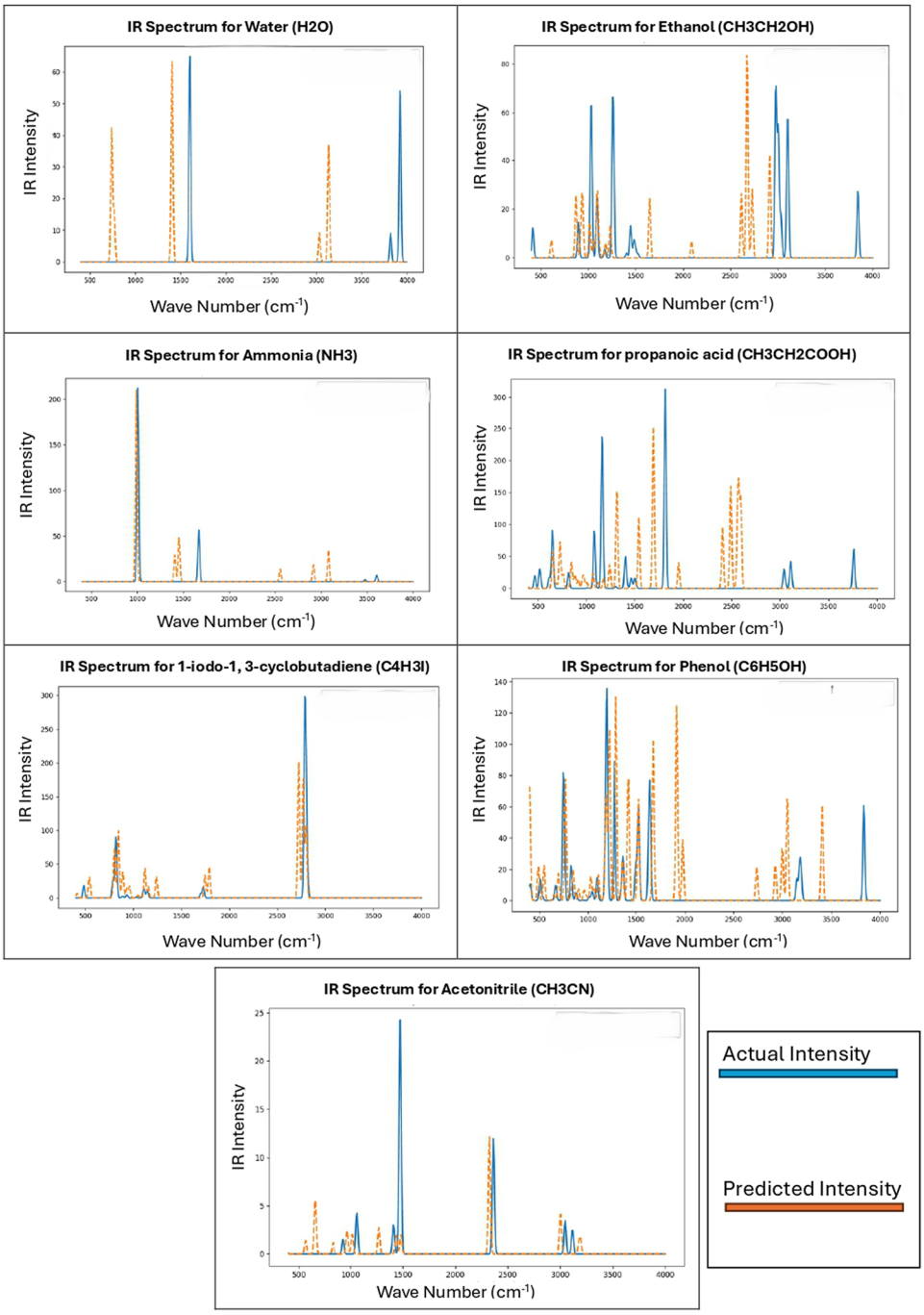


Figure 4. IR spectra prediction vs actual spectra comparison results graph of molecules (i) water (H_2O), (ii) ethanol (CH_3CH_2OH), (iii) ammonia (NH_3), (iv) propanoic acid (CH_3CH_2COOH), (v) 1-iodo-1,3-cyclobutadiene (C_4H_3I), (vi) phenol (C_6H_5OH), and (vii) acetonitrile (CH_3CN).

by ML models. This step ensured that the data set was clean, well-structured, and ready for feature engineering.[45]

3.3. Feature Engineering. Feature engineering involved extracting meaningful information from the raw data. Aggregation involved molecular properties being statistically weighted (e.g., calculating mean *X*, *Y*, *Z* coordinates, summing the number of bonds). The vibrational and molecular data sets were merged on the Molecule column, creating a unified data set that associated each molecule with its structural and vibrational features. This enriched data set contained both input features (e.g., molecular properties) and target outputs (e.g., vibrational frequencies and intensities).

3.4. Train-Test Split. The merged data set was split into training and testing subsets to evaluate the model's performance. For this model, a split of 80% training and 20% testing data. Features (X) included molecular coordinates, number of bonds, dipole moment, and encoded point group. Target (Y) contained the vibrational frequencies and IR intensities. The split ensured that the model was trained on one portion of the data and tested on unseen data to evaluate its generalization ability. 112 molecules were allocated toward training while 28 was set aside for testing. The split was random but it maintained the presence of smaller and bigger molecules to avoid bias.

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

3.5. Model Training. The pipeline employed Random Forest Regressors wrapped in a Multioutput Regressor to predict 2 outputs: first, the frequency array, which contained the vibrational frequencies for each molecule, and second, the intensity array, which had the corresponding IR intensities. Random Forest was chosen for its robustness, ability to handle nonlinear relationships, and feature importance analysis. Using separate models for frequency and intensity allowed for more precise predictions of each target.[46]

3.6. Prediction and Visualization. After training, the pipeline predicted IR spectra for new or existing molecules. For a given molecule, the model outputted predicted frequency and intensity arrays. In terms of visualization, Gaussian curves were generated for each predicted and actual frequency-intensity pair to simulate the IR spectrum.[47] A plot was created comparing actual vs predicted IR spectra, showcasing the model's accuracy and performance. This provided a clear, intuitive way to evaluate the quality of predictions and validate the model.

3.7. Overall Design Principles. Data Integration combines structural and vibrational data for comprehensive analysis. Scalability allows for easy inclusion of new molecules or additional features. Finally, the model's modularity means that each step (data collection, preprocessing, feature engineering, etc.) is self-contained, making the pipeline flexible and extensible.

4. RESULTS AND DISCUSSION

This study focuses on optimizing predictive models for vibrational frequencies and molecular characteristics using enriched data sets and advanced ML techniques. Key enhancements included increasing data resolution from 100 to 500 points per molecule and introducing features such as "Dominant Atom," point group symmetry, and dipole moments to improve clustering and prediction accuracy.

By leveraging supervised learning for vibrational frequency predictions and clustering for molecular family analysis, we achieved reliable alignment between predicted trends and experimental data. TensorFlow was used for efficient model training and validation, with comparative analysis confirming the importance of molecular features in enhancing prediction fidelity. This section explores these optimizations, evaluates model performance, and highlights their implications for molecular spectroscopy.

4.1. Selection of Representative Molecules for Model Validation. The most critical aspect of accurately predicting IR spectra is the quality of the input data used to train the model. Thus, a thoughtful and methodical approach to selecting this data is crucial to ensure that the model performs with optimal precision. As a result, to validate our predictive model and evaluate its accuracy, we selected a diverse set of molecules that are structurally distinct and chemically significant. The chosen molecules include a mix of simple organic and inorganic compounds, each exhibiting unique IR activity. In Figure 4, we have shown the predicted vs actual IR spectra for the chosen molecules. These molecules are (i) Water (H_2O): A fundamental inorganic molecule with strong IR-active bending and stretching vibrations. Its simplicity and ubiquitous presence make it an ideal baseline for testing vibrational predictions. (ii) Ethanol (CH_3CH_2OH): Representing the alcohol functional group (−OH), ethanol features prominent IR signals from its hydroxyl and alkyl group vibrations. Its role in numerous chemical and biological processes makes it an essential test case. (iii) Propanoic Acid (CH_3CH_2COOH): As a representative of carboxylic acids, this molecule is characterized by strong IR activity due to its C═O and O−H stretches. Its structural complexity provides a good test for distinguishing functional group contributions. (iv) Acetonitrile (CH_3CN): A nitrile compound with a distinct CN stretching vibration, acetonitrile is both industrially and spectroscopically significant, making it a valuable addition to the data set. (v) Phenol (C_6H_5OH): A cyclic aromatic molecule with a hydroxyl group, phenol combines the vibrational contributions of aromatic rings and alcohols. Its structural complexity tests the model's ability to predict multiple vibrational modes accurately. (vi) Halogenated Furan (C_4H_3I): This molecule introduces the effects of halogen substitution, with iodine contributing unique vibrational characteristics due to its mass and electronegativity. vii) Ammonia (NH_3): A fundamental inorganic molecule with characteristic bending and symmetric/asymmetric stretching vibrations, ammonia's simple geometry contrasts with the complexity of organic molecules in the set.

The main reasons for selection are as follows: (i) Structural diversity: These molecules span a range of functional groups (hydroxyl, carboxylic acid, nitrile, aromatic, halogenated, and inorganic). This diversity allows for a comprehensive evaluation of the model's ability to distinguish and predict functional group-specific vibrational frequencies. (ii) IR activity: Each molecule was chosen for its strong and distinct IR-active vibrations. These signals are crucial for validating the alignment of predicted and experimental vibrational trends. (iii) Chemical significance: The molecules represent chemically significant classes with broad applications in industrial, biological, and environmental contexts, making the findings relevant to real-world scenarios. (iv) Technical testing: The data set includes both simple and structurally complex molecules to test the model's adaptability. For example, phenol and halogenated furan challenge the model with their complex vibrational interactions, while water and ammonia provide simpler benchmarks. These carefully selected molecules provide a robust foundation for testing our predictive model, ensuring its reliability across a wide spectrum of chemical structures and IR activities.[48−50]

4.2. Analysis of IR Spectra Prediction Model Performance and Data Set Adaptability. Initially, using a smaller data set primarily consisting of smaller or midsized molecules, our predictive model achieved a high level of accuracy. The mean absolute error (MAE) was 7.53 cm^{-1}, the mean squared error (MSE) was 251, and the coefficient of determination (R^2) was 0.90, indicating strong predictive performance. These results highlighted the model's effectiveness in capturing vibrational trends for simpler molecular systems. When a larger data set was introduced, including more structurally complex and larger molecules, the overall accuracy of the model decreased. The R^2 value dropped to 0.65, reflecting the increased variability and complexity of the expanded data set. Despite this reduction, the inclusion of larger molecules brought significant advantages. First, the expanded data set offered greater structural diversity, encompassing a wider range of vibrational modes and enhancing the model's generalizability. This improved its ability to predict vibrational trends across a broader spectrum of molecular systems. Additionally, the inclusion of larger and more complex molecules made the model more representative of real-world applications, increasing its relevance to industrial and research contexts. Finally, the

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

performance drop provided valuable insights into the model's limitations, highlighting areas such as feature engineering and algorithmic refinement that could be targeted for future optimization. These benefits collectively underscore the value of expanding the data set, despite the trade-offs in predictive accuracy. These results demonstrate the trade-off between achieving high accuracy on simpler data sets and building a robust model capable of handling more diverse and complex molecular systems. This expanded approach sets a solid foundation for further improving the model's adaptability and real-world applicability.

4.3. Clustering of Molecules Based on Vibrational and Structural Features. Using clustering techniques, molecules were grouped into distinct families based on their vibrational and structural properties as shown in Figure 5. The clustering process utilized features such as vibrational data (mean, median, and standard deviation of frequency and IR intensity), molecular dipole moment, encoded point group,[51] and dominant atom. A weighted approach was employed to emphasize point group and dominant atom features, ensuring that these critical parameters influenced the clustering process effectively.[52]

CLUSTER	MOLECULES
Cluster 0	$(C_2H_5O)_2$, $(CH_3)_2O$, $C_{10}H_7CN$, $C_{10}H_7PH_2O$, C_2H_3NO, C_2H_4O, C_2H_6SO, C_3H_3CN, $C_3H_3CONH_2$, C_3H_3COOH, C_3H_3I, C_3H_3NO, $C_3H_3PH_2O$, C_2H_3SHO, C_3H_5CHO, C_2H_5CN, C_3H_5COOH, C_3H_5I, C_3H_5NO, $C_3H_7PH_2O$, C_3H_5SHO, C_3H_5SOOH, C_4H_3CHO, C_4H_3CN, $C_4H_3CONH_2$, C_4H_3COOH, C_4H_3I, C_4H_3NO, C_4H_4O, C_4H_3SHO, C_4H_7CN, C_4H_7NO, C_4H_7SOOH, C_4H_8O, C_4NH_5, C_5H_5CN, C_5NH_5, C_6H_5CN, C_6H_5OH, CH_2CHNH, CH_3, CH_3CH_2COOH, CH_3CH_2OH, CH_3CN, CH_3COCH_3, CH_3OH, CNH_3O, H_2O_2, H_2SO, HBO_2, HCN, HCOH, $HCONH_2$, HNO, HOCN
Cluster 1	CH_2
Cluster 2	BH_3, C_2H_2, C_2H_4, C_2H_6, C_3H_8, $C_4H_3C_2H$, $C_4H_4N_2$, C_5H_4, C_6H_{12}, C_6H_6, CCl_4, CF_4, CH_4, CO_2
Cluster 3	C_3H_3Br, C_3H_3Cl, C_3H_3F, C_3H_5Br, C_3H_5Cl, C_3H_5F, C_4H_3Br, C_4H_3Cl, C_4H_3F, C_4H_4S, C_5H_6, C_5H_8, $C_6H_5CH_3$, $CHBr_3$, $CHCl_3$, CHF_3
Cluster 4	CH, H_2O, H_2S, HCl, HF, NH_3

Figure 5. Table containing each molecule within the 5 clusters; molecular prediction and family identification facilitated by this clustering process.

As you can see in Figure 6, the analysis resulted in 5 distinct clusters, each representing molecules with similar structural and vibrational characteristics. The optimal number of clusters was determined to be 5 by the silhouette square and the elbow method. The clusters and their representative molecules are as follows: (i) Cluster 0: Contains larger, more complex organic molecules, such as ethers, nitriles, and aromatic compounds. Examples include $(C_3H_2O)_2$, $(CH_3)_2O$, and $C_{10}H_7CN$. These molecules are characterized by complex vibrational interactions, contributing to their grouping. (ii) Cluster 1: Represents smaller, simpler hydrocarbons, such as CH_2. These molecules show minimal vibrational complexity and serve as a baseline for simplicity in the data set. (iii) Cluster 2: Encompasses small hydrocarbons and similar molecules, such as BH_3, C_2H_2, C_2H_6, and C_3H_8. This group highlights molecules with straightforward, symmetrical vibrational modes and minimal IR complexity. (iv) Cluster 3: Includes halogenated hydrocarbons, such as C_3H_3Br, C_3H_3Cl, and C_3H_5F. These molecules share vibrational patterns influenced by the presence of halogens, which significantly affect their IR activity. (v) Cluster 4: Comprises simple inorganic molecules, such as H_2O, NH_3, and HCl. These are characterized by well-defined bending and stretching vibrations, reflecting their unique properties compared to organic molecules.

4.4. Evaluation of Clustering Results and Technical Insights. Clustering techniques provided significant insights into the molecular data set, highlighting key aspects of functional group identification, diversity within clusters, and improved molecular categorization. The clusters aligned well with functional group categorizations, effectively capturing molecular similarities through vibrational and structural features. While the clusters primarily grouped molecules with shared properties, they also exhibited diversity, particularly in

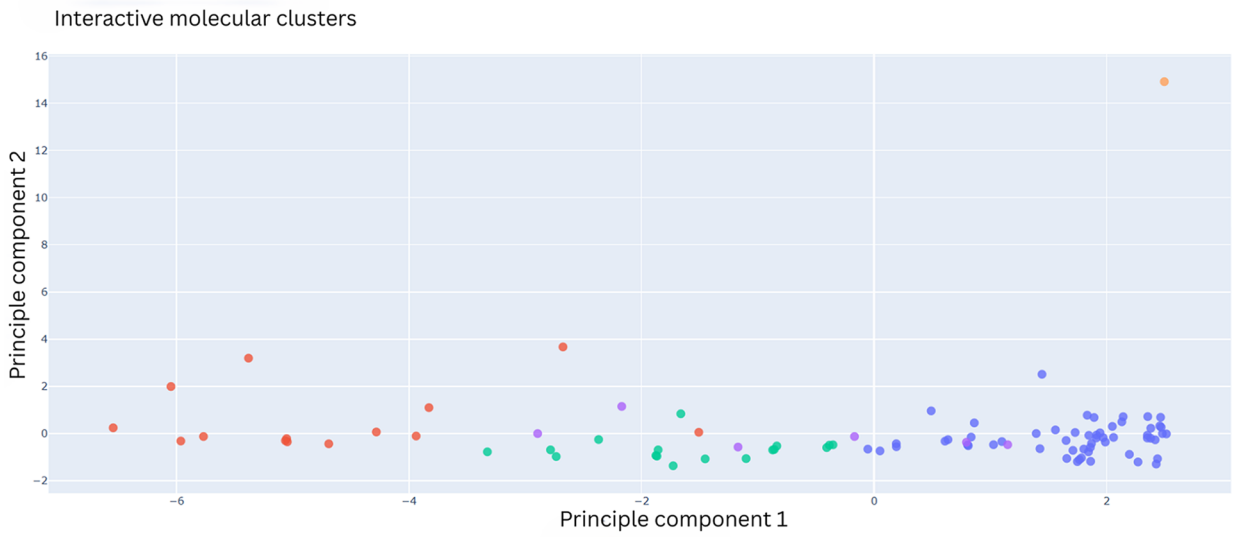


Figure 6. Interactive molecular clustering chart for 100 test molecules. Molecules are grouped into 5 cluster families with weightage given on their vibrational data, molecular dipole moment, encoded point group and dominant atom.

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

Cluster 0, which included molecules with complex structures and interactions. This diversity underscores the robustness of the clustering approach in handling a variety of molecular characteristics. Additionally, the clustering results offered a reliable framework for identifying and categorizing molecules based on vibrational data, proving useful for applications in spectroscopy and chemical analysis.

The inclusion of features such as point group symmetry and dominant atom encoding was instrumental in distinguishing between molecular families with overlapping vibrational properties. Principal Component Analysis (PCA) further enhanced the analysis by reducing the feature space, facilitating clear visualization of clusters and their separation. However, challenges such as overlapping boundaries, particularly between Clusters 2 and 3, indicated areas for refining the feature set. While the clustering method is scalable to larger data sets, incorporating additional features like bond angles and molecular topology could further enhance differentiation and improve the robustness of the clustering model. This balance of strengths and challenges points to opportunities for future optimization and broader applicability.

4.5. Overview of Results. The results and discussion of this study illustrate the effectiveness of the implemented predictive and clustering models in analyzing molecular vibrational and structural properties. The predictive model demonstrated strong performance with smaller data sets, achieving a mean absolute error (MAE) of 7.53/cm and a coefficient of determination (R^2) of 0.90. This is mainly due to the selected features such as corresponding vibrational frequency, quantum properties of the molecules, and 3D structural data. The model uses interatomic distances, bond angles, and torsional angles instead of Cartesian coordinates to eliminate translational disambiguation problems. However, the inclusion of larger and more structurally complex molecules revealed the trade-off between accuracy and generalization, with the R^2 dropping to 0.65.

The decline in model accuracy for larger and more structurally complex molecules arises from both machine learning constraints and fundamental chemical principles. From a computational perspective, increasing molecular size results in a higher number of vibrational degrees of freedom, expanding the feature space and complicating generalization in tree-based models. Furthermore, spectral congestion and the sparsity of distinct vibrational signatures hinder precise pattern recognition. Chemically, extensive vibrational coupling, spectral broadening, and anharmonic perturbations obscure individual vibrational contributions, reducing spectral resolution and interpretability. While our Random Forest model demonstrates robust performance for small to mid-sized molecules, future advancements may involve transfer learning, ensemble approaches, or deep learning architectures optimized for high-dimensional spectral representations to enhance predictive accuracy across diverse molecular frameworks. Cross-validation was done during training and out-of-bag (OOB) estimation provided validation without the need for a separate validation set. The final test set was kept distinct to test the model's performance and versatility in adaptability across different sizes of molecules which was essential for fairness.

The clustering results demonstrate the model's ability to classify molecules into meaningful families based on vibrational and structural characteristics. Features like point group symmetry and dominant atom encoding, combined with PCA visualization, proved critical in distinguishing molecular families and ensuring effective cluster separation. While slight overlaps in some clusters and scalability challenges were observed, the clustering framework successfully provided valuable insights for molecular identification and categorization.[53,54]

Overall, these findings underscore the utility of vibrational data combined with molecular features for predictive modeling and clustering, paving the way for further refinements and broader applications in spectroscopy, molecular design, and chemical analysis.

5. FUTURE SCOPE

The future scope of this project presents numerous opportunities to enhance and expand the model's capabilities, making it even more versatile and applicable to real-world challenges. One key avenue for improvement is the integration of larger and more diverse data sets. By incorporating molecular data from a wider range of systems, such as biomolecules, polymers, and large organometallic complexes, the model's prediction capabilities can be significantly enhanced. These molecules often exhibit intricate vibrational modes that require more advanced modeling techniques to predict accurately.[55] In addition, diversifying the data set to include a broader range of chemical environments, such as varying temperatures, pressures, or solvent conditions, would allow the model to provide more robust predictions applicable to real-world scenarios where environmental factors play a significant role.[56] Expanding the scope of the data set in these ways would not only improve the model's performance but also ensure it is more adaptable and reliable across various chemical domains, such as pharmaceuticals, materials science, and environmental chemistry.

As the field of machine learning continues to evolve, integrating more advanced modeling techniques could further refine the model's accuracy. Architectures like convolutional neural networks (CNNs) and graph neural networks (GNNs) are particularly promising for this purpose.[57,58] CNNs are highly effective at identifying spatial patterns, making them ideal for analyzing molecular geometries and understanding how atoms and bonds influence the overall structure. Meanwhile, GNNs treat molecules as graphs, where atoms are nodes and bonds are edges, enabling a deeper understanding of molecular interactions and properties that are otherwise difficult to capture using traditional methods.[59,60] These neural network architectures can handle the complexity of molecular data more effectively, allowing the model to capture subtle relationships within the data and improve the accuracy of IR spectra predictions. Additionally, the incorporation of reinforcement learning techniques could facilitate dynamic model improvement. By continuously learning from experimental feedback, the model could adapt to new data and challenges, ensuring that it remains state-of-the-art and capable of refining its predictions over time.[61−63] Moreover, multi-objective optimization methods could be employed to strike a balance between accuracy, interpretability, and computational efficiency. This would make the model more practical for both academic research and industrial applications, where these factors are crucial for deployment.[64]

A promising path forward is the integration of quantum chemistry calculations with ML models. Quantum chemistry methods offer highly accurate predictions of molecular properties, but their computational cost can be prohibitive

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

for large data sets.[65] A hybrid approach, combining quantum mechanical calculations with ML models, could provide a solution that leverages the strengths of both approaches. This would allow for efficient yet accurate predictions, making the model an essential tool for both theoretical research and practical applications.[66] Furthermore, the model could be integrated into automated systems designed for identifying unknown compounds based on their IR spectra. Such automation would streamline workflows in industries like forensic science, environmental monitoring, and quality control, where rapid compound identification is essential. This would also reduce the reliance on manual interpretation, enabling faster decision-making and improving operational efficiency across various sectors.[67]

Also, the current ML approach relies solely on computationally derived IR spectra and has not been tested against experimental data. Computationally derived IR spectra contain only the essential number of peaks, whereas experimentally obtained IR spectra often exhibit additional peaks due to noise. Moreover, molecular interactions can lead to peak broadening, making it more difficult for the machine learning model to detect the desired peak. However, we acknowledge this as a limitation of the current work and will explore ways to incorporate this aspect in our future studies.

In addition to these enhancements, the model could be adapted to support real-time molecular design workflows. Researchers could input proposed molecular structures and receive immediate predictions of their IR spectra, drastically speeding up the screening process for molecules with desired properties. This capability would be especially valuable in fields such as drug discovery, materials science, and catalysis, where fast, reliable predictions can significantly accelerate innovation.[68] Furthermore, the methodology developed for predicting IR spectra could be extended to other spectroscopic techniques like Raman, UV−vis, and NMR spectroscopy. By incorporating multiple types of spectroscopic data, the model could provide a more comprehensive analysis of molecular behavior, offering valuable insights for a wide range of analytical chemistry applications.[69] Finally, the model holds great potential for real-time environmental and industrial applications. In industries like chemical manufacturing and environmental monitoring, it could be used to detect and analyze compounds instantly, enabling real-time monitoring of processes such as industrial emissions or pollutant tracking. These applications align with broader goals of sustainability, safety, and regulatory compliance. The ability to deploy such a model in real-time would support efforts to improve environmental standards, reduce harmful emissions, and ensure compliance with stringent regulations, thus contributing to a more sustainable future for both industry and society.[70,71]

6. CONCLUSIONS

This study successfully integrates ML with computational chemistry to predict IR spectra, showcasing a novel and efficient approach to modeling molecular vibrational behavior. By utilizing detailed molecular descriptors such as Cartesian coordinates, bond counts, dipole moments, and point group symmetries, the ML models, powered by Random Forest Regressors and MultiOutput Regressors, effectively capture the nonlinear relationships between molecular geometry and spectroscopic properties. This methodology demonstrates significant advantages over traditional quantum chemistry calculations, offering rapid, scalable, and accurate spectral predictions while reducing computational overhead. The study establishes a foundation for future developments, including the extension of predictive capabilities to other spectroscopic techniques, the adoption of advanced architectures like graph neural networks, and the real-time application of these models in molecular design and industrial processes. This work underscores the potential of AI-driven approaches to revolutionize molecular spectroscopy, enabling deeper insights and broader applications in both research and industrial domains.

■ ASSOCIATED CONTENT

Data Availability Statement

To ensure full transparency and reproducibility, a detailed "Data and Software Availability" section is provided in the Supporting Information. This section outlines the data sets, computational tools, and ML methodologies used in this study. The data sets utilized in this study are available in the form of Google Spreadsheets, the links to which are also provided under the same subsection. These files include details on molecular geometries, number of bonds, dipole moments, point groups, frequencies, IR intensities, and vibrational modes. The computational tools subsection includes all the softwares and packages used in the model's life cycle. These primarily include softwares Gaussian 16[72] and Chemcraft,[73] and packages such as TensorFlow and SciKit-Learn.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.5c02405.

Diagrams and models offering a visual explanation of the implemented modules: molecule prediction, including molecule information, along with the SMILES notation, and their IR graphs; and implementation of Code 4, data and software availability (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Gopi Ragupathy − *Department of Chemistry, School of Advanced Sciences, Vellore Institute of Technology, Vellore 632014, India;* orcid.org/0000-0002-9435-5712; Email: r.gopi@vit.ac.in

Authors

Adithya Ranjith Kartha − *School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India*

Dhanush P. Ajayakumar − *School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India*

Muhammad Idris − *School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.5c02405

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.R.K., D.P.A., and M.I. express their gratitude to VIT Vellore. G.R. acknowledges the management of VIT Vellore and the

seed grant (Sanctioned Order No: SG20230052) for providing a sophisticated computing facility.

REFERENCES

(1) Ditler, E.; Luber, S. Vibrational spectroscopy by means of first-principles molecular dynamics simulations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12* (5), No. e1605.
(2) Agarwal, T.; Prasad, A. K.; Advani, S. G.; Babu, S. K.; Borup, R. L. Infrared Spectroscopy for Understanding the Structure of Nafion and its Associated Properties. *J. Mater. Chem. A* **2024**, *12*, 14229−14244.
(3) Wrobel, T. P.; Bhargava, R. Infrared spectroscopic imaging advances as an analytical technology for biomedical sciences. *Analytical chemistry* **2018**, *90* (3), 1444−1463.
(4) Hunt, N. T. Using 2D-IR Spectroscopy to Measure the Structure, Dynamics, and Intermolecular Interactions of Proteins in H2O. *Acc. Chem. Res.* **2024**, *57* (5), 685−692.
(5) Czarnecki, M. A.; Morisawa, Y.; Futami, Y.; Ozaki, Y. Advances in molecular structure and interaction studies using near-infrared spectroscopy. *Chem. Rev.* **2015**, *115* (18), 9707−9744.
(6) Shao, X.; Paetow, L.; Tuckerman, M. E.; Pavanello, M. Machine Learning electronic structure methods based on the one-electron reduced density matrix. *Nat. Commun.* **2023**, *14* (1), 6281.
(7) Fan, L.; Ziegler, T. Application of density functional theory to infrared absorption intensity calculations on main group molecules. *J. Chem. Phys.* **1992**, *96*, 9005−9012.
(8) Halls, M. D.; Velkovski, J.; Schlegel, H. B. Harmonic Frequency Scaling Factors for Hartree-Fock, S-VWN, B-LYP, B3-LYP, B3-PW91 and MP2 with the Sadlej pVTZ Electric Property Basis Set. *Theor. Chem. Acc.* **2001**, *105*, 413−421.
(9) Gopi, R.; Ramanathan, N.; Sundararajan, K. Experimental Evidence for Blue-Shifted Hydrogen Bonding in Fluoroform-Hydrogen Chloride Complex: A Matrix Isolation Infrared and ab Initio Study. *J. Phys. Chem. A* **2014**, *118* (29), 5529−5539.
(10) Gopi, R.; Ramanathan, N.; Sundararajan, K. Acetonitrile-Water Hydrogen-Bonded Interaction: Matrix-Isolation Infrared and ab initio Computation. *J. Mol. Struct.* **2015**, *1094*, 118−129.
(11) DeBlase, A. F.; Bloom, S.; Lectka, T.; Jordan, K. D.; McCoy, A. B.; Johnson, M. A. Origin of the Diffuse Vibrational Signature of a Cyclic Intramolecular Proton Bond: Anharmonic Analysis of Protonated 1,8-Disubstituted Naphthalene Ions. *J. Chem. Phys.* **2013**, *139*, No. 024301.
(12) Malik, M.; Wysokinski, R.; Zierkiewicz, W.; Helios, K.; Michalska, D. Raman and Infrared Spectroscopy, DFT Calculations, and Vibrational Assignment of the Anticancer Agent Picoplatin: Performance of Long-Range Corrected/Hybrid Functionals for a Platinum(II). *Complex. J. Phys. Chem. A* **2014**, *118*, 6922−6934.
(13) Katari, M.; Nicol, E.; Steinmetz, V.; Van-Der-Rest, G.; Carmichael, D.; Frison, G. Improved Infrared Spectra Prediction by DFT from a New Experimental Database. *Chem.—Eur. J.* **2017**, *23*, 8414−8423.
(14) Alcolea Palafox, M. Scaling Factors for the Prediction of Vibrational Spectra. I. Benzene Molecule. *Int. J. Quantum Chem.* **2000**, *77*, 661−684.
(15) Maltseva, E.; Petrignani, A.; Candian, A.; Mackie, C. J.; Huang, X.; Lee, T. J.; Tielens, A. G. G. M.; Oomens, J.; Buma, W. J. HighResolution IR Absorption Spectroscopy of Polycyclic Aromatic Hydrocarbons: The Realm of Anharmonicity. *Astrophys. J.* **2015**, *814*, 23.
(16) Gaw, J. F.; Willetts, A.; Green, W. H.; Handy, N. C. In *Advances in Molecular Vibrations and Collision Dynamics*, Bowman, J. M.; Ratner, M. A., Eds., 1991; Vol. *1*, p 169.
(17) Gaigeot, M.-P.; Sprik, M. Ab Initio Molecular Dynamics Computation of the Infrared Spectrum of Aqueous Uracil. *J. Phys. Chem. B* **2003**, *107*, 10344−10358.
(18) Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. Fast and Accurate Quantum Chemical Modeling of Infrared Spectra of Condensed-Phase Systems. *J. Phys. Chem. B* **2020**, *124*, 6664−6670.
(19) Ezugwu, A. E.; Greeff, J.; Ho, Y. S. A comprehensive study of groundbreaking machine learning research: analyzing highly cited and impactful publications across six decades. *J. Eng. Res.* **2023**, *13*, 371−383.
(20) Morawietz, T.; Artrith, N. Machine learning-accelerated quantum mechanics-based atomistic simulations for industrial applications. *Journal of Computer-Aided Molecular Design* **2021**, *35* (4), 557−586.
(21) Ullah, A.; Chen, Y.; Dral, P. O. Molecular quantum chemical data sets and databases for machine learning potentials. *Machine Learning: Science and Technology* **2024**, *5* (4), No. 041001.
(22) TensorFlow Team. Graph Neural Networks in TensorFlow; TensorFlow Blog, **2024**. https://blog.tensorflow.org/2024/02/graph-neural-networks-in-tensorflow.html.
(23) Kensert, A. MolGraph: Molecular Machine Learning with TensorFlow and Keras. *GitHub* [Online]. https://github.com/akensert/molgraph.
(24) Reiser, P.; Neubert, M.; Eberhard, A.; et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3*, 93.
(25) Robb, E. W.; Munk, M. E. A neural network approach to infrared spectrum interpretation. *Mikrochim. Acta* **1990**, *100* (3/4), 131−155.
(26) Fessenden, R. J.; Györgyi, L. Identifying functional groups in IR spectra using an artificial neural network. *J. Chem. Soc., Perkin Trans.* **1991**, *2* (11), 1755−1762.
(27) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharmaceutics* **2019**, *13*, 2524.
(28) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2022**, No. 145301.
(29) Novic, M.; Zupan, J. Investigation of infrared spectra-structure correlation using kohonen and counterpropagation neural network. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (3), 454−466.
(30) Shilpa, S.; Kashyap, G.; Sunoj, R. B. Recent applications of machine learning in molecular property and chemical reaction outcome predictions. *J. Phys. Chem. A* **2023**, *127* (40), 8253−8271.
(31) Singh, S.; Sunoj, R. B. Molecular machine learning for chemical catalysis: Prospects and challenges. *Acc. Chem. Res.* **2023**, *56* (3), 402−412.
(32) Ganthavee, V.; Trzcinski, A. P. Artificial intelligence and machine learning for the optimization of pharmaceutical wastewater treatment systems: a review. *Environ. Chem. Lett.* **2024**, *22*, 2293−2318.
(33) Pfau, D.; Spencer, J. S.; Matthews, A. G. D. G.; Foulkes, W. M. C. Ab-Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks. *Phys. Rev. Res.* **2019**, *2*, No. 033429.
(34) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; et al. QM40: A Dataset for Machine Learning in Drug Discovery Representing 88% of FDA-Approved Drug Chemical Space. *Sci. Data* **2024**, *11*, 1376.
(35) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning molecular dynamics for the simulation of infrared spectra. *Chemical science* **2017**, *8* (10), 6924−6935.
(36) Kim, U. J.; Furtado, C. A.; Liu, X.; Chen, G.; Eklund, P. C. Raman and IR spectroscopy of chemically processed single-walled carbon nanotubes. *J. Am. Chem. Soc.* **2005**, *127* (44), 15437−15445.
(37) Rao, C. N. R. Chemical applications of infrared spectroscopy. In *Chemical Applications of Infrared Spectroscopy*, 1963; pp 681−681.
(38) Paliwal, A.; Subramanian, G.; Ramsundar, B.; Pande, V. MolPROP: Predicting Multiple Molecular Properties Simultaneously using Language and Graph Representations. *J. Cheminf.* **2024**, *16* (1), 46.
(39) Zhu, L.; Yao, Q. A Quantum-Inspired Ensemble Method and Quantum-Inspired Forest Regressors. *arXiv preprint arXiv:1711.08117* 2017. https://arxiv.org/abs/1711.08117.

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234

(40) Al Saleh, A.; Allouche, A.-R. Neural network approach for predicting infrared spectra from 3D molecular structure. *Chem. Phys. Lett.* **2024**, *856*, No. 141603.
(41) Tang, X.; Li, S. Quantitative relationships between bond lengths, stretching vibrational frequencies, bond force constants, and bond orders in the hydrogen-bonded complexes involving hydrogen halides. *Struct. Chem.* **2018**, *29* (1), 43−50.
(42) Gonze, X.; Lee, C. Infrared intensities and Raman-scattering activities within density-functional theory. *Phys. Rev. B* **1996**, *54* (11), 7830−7838.
(43) University of Illinois. *Applications of Group Theory to Spectroscopy*. https://xuv.scs.illinois.edu/516/lectures/chem516.04.pdf.
(44) Kang, X.; Zhang, Y. Adapting Differential Molecular Representation with Hierarchical Prompts for Multi-label Property Prediction. *arXiv preprint arXiv:2405.18724* 2024. https://arxiv.org/abs/2405.18724.
(45) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9* (2), 513−530.
(46) Domenichini, G.; Dellago, C. Molecular Hessian matrices from a machine learning random forest regression algorithm. *J. Chem. Phys.* **2023**, *159* (19), 194111.
(47) Jägerfeld, P. J. *Gauss Fitting of DFT Derived Infrared Spectra*; CRC Network Catalysis, 2024.
(48) Krzyḋanowski, M.; Matyszczak, G. Machine Learning Prediction of Organic Moieties from the IR Spectra, Enhanced by Additionally Using the Derivative IR Data. *Chem. Pap.* **2024**, *78*, 3149−3173.
(49) Pracht, P.; Pillai, Y.; Kapil, V.; Csányi, G.; Gönnheimer, N.; Vondrák, M.; Margraf, J. T.; Wales, D. J. Efficient Composite Infrared Spectroscopy: Combining the Doubly-Harmonic Approximation with Machine Learning Potentials. *arXiv preprint arXiv:2408.08174* 2024. https://arxiv.org/abs/2408.08174.
(50) Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry. *arXiv preprint arXiv:2407.17492* 2024. https://arxiv.org/abs/2407.17492.
(51) D'Costa, R.; Nguyen, T. Prediction of Vibrational Spectra of a Molecule of C Point Group with Group Theory. *Math. Stat. Eng. Appl.* **2024**, *39* (3), 125−138.
(52) Akgüller, Ö.; Balcı, M.A.; Cioca, G. Clustering Molecules at a Large Scale: Integrating Spectral Geometry and Deep Learning. *Molecules* **2024**, *29* (16), 3902.
(53) Kim, S.; Park, J. Deep Clustering of Small Molecules at Large-Scale via Variational Autoencoder. *BMC Bioinf.* **2022**, *23*, 132.
(54) Harris, P.; Nyman, L. *Group Theory Applications for Molecular Vibration Analysis*, 2023.
(55) Zhang, Y.; Lin, H. Infrared spectra prediction using attention-based graph neural networks. *Digital Discovery* **2024**, *3*, 254−267.
(56) Gastegger, M.; Schütt, K. T.; Müller, K. R. Machine Learning of solvent effects on molecular spectra and reactions. *arXiv preprint arXiv:2010.14942* 2020.
(57) Hochuli, J.; Helbling, L.; Skaff, S.; Schneider, G. TopologyNet: Topology-based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **2017**, *13* (11), No. e1005690.
(58) Zhou, Y.; Wang, H.; Xu, X. FragNet: A Graph Neural Network for Molecular Property Prediction with Four Layers of Interpretability. *arXiv preprint arXiv:2410.12156* 2024. https://arxiv.org/abs/2410.12156.
(59) Wieder, O.; et al. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technol.* **2020**, *37*, 1−12.
(60) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. In *Proc. Adv. Neural Inf. Process. Syst.*, 2019; Vol. *31*, pp 6412−6422.
(61) Devata, S.; Sridharan, B.; Mehta, S.; Pathak, Y.; Laghuvarapu, S.; Varma, G.; Priyakumar, U. D. DeepSPInN−deep reinforcement learning for molecular structure prediction from infrared and 13C NMR spectra. *Digital Discovery* **2024**, *3*, 818−829.
(62) van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double Q-learning, Accessed on: August 8, 2022. [Online]. Available: https://arxiv.org/abs/1509.06461.
(63) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **2019**, *9* (1), 10752.
(64) Lee, D.; Wang, Y. Bayesian multi-objective optimization for computational efficiency in chemical modeling. *Chem. Eng. Sci.* **2022**, *260*, No. 117403.
(65) Rupp, M.; Tkatchenko, A.; Müller, K. R.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087−2096.
(66) Yao, K.; Herr, J. E.; Toth, D. W.; McIntyre, R.; Parkhill, J. Hybrid Quantum Graph Neural Network for Molecular Property Prediction. *arXiv preprint arXiv:2405.05205* 2024. https://arxiv.org/abs/2405.05205.
(67) Leveraging Infrared Spectroscopy for Automated Structure Elucidation. *ChemRxiv* preprint, 2024.
(68) Huang, K.; Fu, T.; Khan, D.; Abid, A.; Abdalla, A.; Abid, A.; Glass, L. M.; Zitnik, M.; Xiao, C.; Sun, J. MolDesigner: Interactive Design of Efficacious Drugs with Deep Learning. *arXiv preprint arXiv:2010.03951* 2020.
(69) Jahoda, P.; Drozdovskiy, I.; Sauro, F.; Turchi, L.; Payler, S.; Bessone, L. Machine Learning for recognition of minerals from multispectral data. *arXiv preprint arXiv:2005.14324*, 2020.
(70) Parker, F. *Applications of infrared spectroscopy in biochemistry, biology, and medicine*; Springer Science & Business Media, 2012.
(71) Bellisola, G.; Sorio, C. Infrared spectroscopy and microscopy in cancer research and diagnosis. *Am. J. Cancer Res.* **2011**, *2* (1), 1.
(72) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16, Revision C.01*; Gaussian Inc.: Wallingford CT., 2016.
(73) Chemcraft - graphical software for visualization of quantum chemistry computations. Version 1.8, build 682. https://www.chemcraftprog.com.

https://doi.org/10.1021/acsomega.5c02405
ACS Omega 2025, 10, 19224−19234