# ENSEMBLE MACHINE LEARNING FOR PREDICTIVE MODELING OF ASTROCHEMICAL BINDING ENERGIES: ADVANCING ACCURACY IN COMPLEX SYSTEMS

Supported by tetfund TERTIARY EDUCATION TRUST FUND

**Emmanuel E. Etim[1*], Nahum Bako[2], Humphrey S. Samuel[1], Oko Emmanuel Godwin[3]**
[1]Department of Chemical Sciences, Federal University Wukari
[2]Department of Computer Science, Federal university Wukari
[3]Astrochemistry & Astrophysics Center, Faculty of Engineering, Institute of Applied Chemical Sciences, Universidad Autonoma de Chile, Av. Pedro de Valdivia 425, Providencia, Santiago, Chile
*Corresponding author: emmaetim@gmail.com

**Received: February 14, 2025, Accepted: April 28, 2025**

**Abstract:** Stemming from the relevance of binding energy in understanding the formation, reactivity, and stability of molecular species, particularly in astrochemical environments, this current study leverages an ensemble model integrating six machine learning algorithms: Bagging, Linear Regression, Random Forest, Gradient Boosting, Bayesian Ridge, and Ridge Regression to predict the binding energies of astrochemically relevant molecules. Gradient Boosting demonstrated superior performance in capturing predictive accuracy and variance among individual models. The ensemble model surpassed the predictive power of single algorithms, offering a robust framework for complex chemical systems. The correlation between predicted binding energies and desorption temperatures provides insight into molecule-surface interaction strengths. The ensemble approach illustrates the potential of machine learning techniques in solving intricate astrochemical problems. The ensemble methods effectively capture complex relationships within the molecular data, leading to more accurate and reliable predictions. The results obtained here can be applied in astrochemistry and material sciences and further stress the relevance of machine learning in predictive modeling in Chemistry and other related fields.

**Keywords:** Machine learning, complex systems, Energies, Model, computational techniques

## Introduction

Space may be the source of prebiotic chemicals and compounds with properties related to the origin(s) of life on Earth (Petrignani & Candian 2022). Cations make up around 10% of all known molecular species in the circumstellar and interstellar regions. According to Etim and Arunan (2015), about 80% of these cations are protonated species whose neutral counterparts are likewise recognized molecular species in the interstellar medium (ISM). Since hydrogen has an unquestionably great cosmic abundance, it serves as the main reactant in the majority of ISM chemical reactions. The ionized forms of hydrogen, $H^+$ and $H_3^+$, are the fundamental components of ion-molecule reactions, which are the predominant gas phase chemistry activities in ISM and occur with little to no activation barrier (Etim *et al.*, 2017). This suggests that any neutral molecular species in the ISM can be protonated via gas-phase reaction, given the enormous abundance of $H_3^+$ in the molecular clouds. In the gas phase, protonated species are primarily formed by the reaction of neutral species with a variety of interstellar ions, including $H^+$, $H_3^+$, $C^+$, $He^+$, $HCO^+$, $CH_3^+$, and $H_3O^+$. The neutral species naturally arises from the protonated species as well. Proton binding energies (PBEs) and protonated species that correspond to the same neutral species might differ when a proton binds to a neutral molecule at multiple locations within the molecular structure. The PBE's magnitude provides strength and stability to the protonated species. The PBE's magnitude provides the protonated species. Simply put, a greater PBE value indicates a stronger proton-neutral connection and therefore a higher degree of stability for the protonated species. Astronomical observations are influenced by the direct correlation between a molecular species' interstellar abundance and stability (Etim *et al.*, 2020).

The most stable species are known to be the most abundant in the ISM when compared to their less stable counterparts, with the exception of situations in which other factors such as distinct formation routes, interstellar hydrogen bonding, etc., play a major role (Etim *et al.*, 2018; Etim and Arunan, 2016a; Etim and Arunan, 2017). Given a neutral molecule that gives birth to two distinct protonated species, each with a different PBE, the protonated species with the highest PBE is therefore the most stable relative to the other with the lower PBE. In comparison to its counterpart with lesser stability, the most stable protonated species should be easier to identify due to their greater abundance (assuming that reaction routes do not differ much). This is because, in contrast to its isomer, which arises from a higher PBE value for the identical neutral species and tends to remain protonated, the species with a lower PBE can readily transfer its proton and revert to its neutral state. Because the protonated substances with low PBE have low stability and also have high reactivity, which lowers their interstellar abundance and makes astronomical detection of them challenging (Etim*et al.*, 2018).

One way of understanding the complex chemical processes driving the formation, evolution, and interactions of molecules in the astrochemical environment is by the computation of their binding energies (Villadsen *et al.*, 2022). The accuracy of this binding energy is critical in the determination of molecular stability, reactivity, and the potential for life in space. These energies play a key role in the fate of molecules on interstellar dust grains, influencing processes like molecule formation, desorption, and

subsequent reactions (Siebenmorgen & Zacharias, 2020). Desorption, in particular, is the process by which molecules detach from surfaces, often driven by temperature changes, and is crucial for understanding how molecules transition from solid to gas phases in space (Bovolenta *et al.*, 2020). Accurately predicting desorption energies is vital for modeling the dynamic chemical environments found in astrophysical settings.

Although traditional methods like Density Functional Theory (DFT) (Spiegelman *et al*., 2020) and Hartree-Fock (Hirao *et al*., 2023) are highly accurate for calculating binding energies, they require substantial computational resources, making them impractical for large-scale studies involving diverse molecular species. This makes them impractical for large-scale studies involving diverse molecular species (Andrew *et al.*, 2018).

One of the main experimental techniques used to measure binding energies is temperature-programmed desorption (TPD) (Johnson *et al*., 2024). TPD assesses the energy needed to desorb molecules from a surface as temperature increases, offering valuable insights into molecular stability under various astrophysical conditions (Smith & Kay, 2018). Despite their accuracy, TPD and other experimental methods are often limited by their complexity, high costs, and the time needed to conduct extensive studies.

Due to these challenges, interest in alternative methods that strike a balance between accuracy and efficiency has emerged. Machine learning (ML) has proven itself as the latest improvement, which has opened new possibilities for modeling complex relationships in high-dimensional data, with the prospect of serving as an alternative to traditional approaches (Tufail *et al*., 2023; Samuel *et al.*, 2024). Villadsen and his team, who showed up as pioneers in this field (Villadsen *et al.*, 2022), utilized machine learning approach via Gaussian Process Regression to predict the binding energies of molecules and achieved results with an accuracy of less than ±20% deviation from literature values. Although machine learning techniques are highly effective at managing large datasets and are particularly well-suited for predicting binding energies, relying on a single model may have limitations that can impact accuracy and robustness. To overcome these challenges, ensemble methods, which integrate multiple models, offer improved performance by leveraging the strengths of different predictive approaches and addressing the weaknesses of individual models (Mohammed & Kora, 2023; Etim *et al.*, 2018; Etim *et al.*, 2020).

Ensemble-based Bayesian methods were used in the assessments, such as the uncertainty quantification in computational fluid dynamics (CFD) problems (Zhang *et al*., 2020). Ensemble methods have been shown to achieve superior performance in predictive studies. This investigation is aimed at combining six different machine learning algorithms, including bagging, linear regression, random forest, gradient boosting, and bayesian ridge regression, to predict the binding energy of molecules of potential relevance to astrochemistry, enabling us to to hypothesize at this point that the outcome of the combined ensemble machine learning methods will perform better in predicting the binding energies of these molecular species compared to the individual methods. The objectives of this research are threefold:

1. To construct an ensemble model that combines the strengths of Bagging, Linear Regression, Random Forest, Gradient Boosting, Bayesian Ridge, and Ridge Regression for predicting binding energies of astrochemically relevant molecules.
2. To evaluate the performance of the ensemble model and compare it with individual models using appropriate metrics.
3. To provide insights into the potential applications of the ensemble model in astrochemistry and highlight future research directions.

**Methods and Data**

Our study employs an ensemble approach combining six machine learning algorithms to predict binding energies (BEs) of astrochemically relevant molecules. The ensemble model integrates bagging, linear regression, random forests, gradient boosting, Bayesian ridge, and ridge regression to leverage their distinct strengths and enhance predictive accuracy. These algorithms are trained on a dataset containing binding energies obtained from temperature-programmed desorption (TPD) experiments and molecular features such as surface category, atomic composition, and functional groups. The training data enables the model to generalize and make accurate predictions for molecules and surfaces not seen during training (Samuel et al., 2024; Samuel *et al.*, 2023; Shinggu *et al.*, 2023).

The model's performance was evaluated using bootstrapping and five-fold cross-validation. The former involves repeated sampling with replacement, creating multiple training and testing sets. The latter involves dividing the dataset into five sub-parts, when one of the subset is used as a test; the remaining four were used for training. The model's accuracy was ascertained via three metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared ($R^2$). Afterwards, we used the models to estimate the binding energies of 21 molecules present in the ISMs and compared with known experimental data (Oba *et al.*, 2009; Samuel *et al.*, 2023).

**Ensemble Method**

The ensemble method is employed in this study to enhance the predictive accuracy of binding energies (BEs) by combining the strengths of multiple models. By aggregating the predictions from various models, ensemble learning produces a more accurate and robust prediction than any single model could achieve independently (Mienye & Sun, 2022; Samuel *et al.*, 2024; Osigbemhe *et al.*, 2022). The ensemble in this work includes bagging, linear regression, random forest, gradient boosting, Bayesian ridge, and ridge

regression models, each contributing uniquely to the final prediction.

**Bagging(Bootstrap Aggregating)**

This is a powerful ensemble technique that builds multiple versions of a model by training each version on a different bootstrap sample of the original data (Awujoola *et al.,* 2020; Etim e*t al.,* 2023). These models are then averaged for regression or majority-voted for classification. For a regression problem, the final prediction is computed as the mean of all individual model predictions from the bagged ensemble:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots eqn\,(1)$$

…where y is the number of models in the ensemble.

This reduces variance and improves model stability and accuracy by training multiple base models on different subsets of the data and averaging their predictions.

**Random Forest**

Random Forest is an ensemble learning method primarily used for classification and regression tasks which leverages the concept of combining multiple decision trees to enhance predictive performance and mitigate over-fitting (Zhu, 2020; Oladimeji, et al., 2024; Samuel et al., 2023; Samuel et al., 2024). It enhances the ensemble by creating a multitude of decision trees using different subsets of the data and features. Each tree contributes to the final prediction, which is the average of the individual tree predictions. The Random Forest prediction is given by:

$$\hat{y}_{RF} = \frac{1}{N}\sum_{i=1}^{N} T_i(X) \quad \ldots\ldots\ldots\ldots\ldots\ldots eqn(2)$$

…where each represents the prediction from the th decision tree.

**Linear Regression** is a fundamental statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables. The primary goal of linear regression is to identify the best-fitting linear relationship that predicts the dependent variable based on the values of the independent variables. (James *et al*., 2023). It provides a simple, yet effective baseline in the ensemble. It models the relationship between input features and output as a linear combination of the input features:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \quad \ldots\ldots\ldots\ldots\ldots\ldots eqn(3)$$

…where is the intercept, are the coefficients, and is the number of features.

**Gradient Boosting**

Gradient Boosting is an advanced ensemble learning technique that builds predictive models by combining multiple weak learners, typically decision trees, to create a robust and accurate model. The method iteratively improves the model by focusing on the errors made by previous iterations, making it highly effective for both classification and regression tasks. (Bentéjac *et al*., 2021; Etim *et al.,* 2017). It takes a sequential approach by building models that correct the errors of their predecessors. Each new model is trained to minimize the residual errors from the combined previous models:

$$\hat{y}_m(x) = \hat{y}_{m-1}(x) + \eta . h_m(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots eqn(4)$$

…where is the learning rate, controlling the contribution of each learner to the final prediction.

**Bayesian Ridge Regression**

Bayesian Ridge Regression provides a probabilistic framework for linear regression by incorporating prior distributions on the regression coefficients, which helps manage over-fitting and quantify prediction uncertainty (Imane *et al*., 2022). It assumes a prior distribution over the model parameters and computes a posterior distribution given the data (Etim*et al.,* 2022). The prediction is made using the mean of the posterior distribution:

$$\hat{\beta} = \underset{\beta}{argmin}\left\{\sum_{i=1}^{n}\left(y_1 - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad \ldots\ldots\ldots\ldots eqn(5)$$

…where is the regularization parameter that controls the trade-off between fitting the data and keeping the coefficients small.

**Ridge Regression**

In situations when the independent variables are highly corr elated, ridge regression is a techniquefor estimating the coe fficients of multiple-regression models and is often referred to as Tikhonov regularization, after Andre y Tikhonov. In this study, hyper-parameters for each model in the ensemble were carefully tuned using cross-validation to achieve optimal performance. The ensemble approach leverages the diversity of its constituent models—combining the simplicity of Linear Regression, the robustness of Ridge Regression, the probabilistic nature of Bayesian Ridge, and the powerful tree-based methods of Random Forest and Gradient Boosting.

The final ensemble prediction is formed by averaging the predictions from all models, providing a balanced and accurate prediction of BEs. The variance across the models in the ensemble offers an uncertainty estimate, indicating the confidence level in the predictions, thus enhancing the model's reliability when applied to new, unseen data.

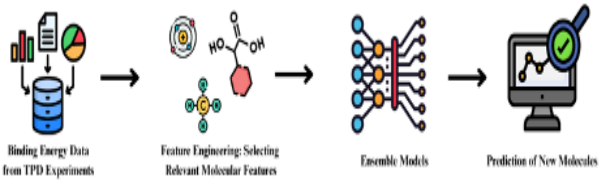


Fig. 1: Workflow: First, molecular data is collected from temperature-programmed desorption experiments and categorized based on binding energies. Next, relevant molecular features are selected and engineered, including atomic composition, functional groups, and valence electrons.

**Data preparation and Feature Engineering**

For this study, we utilized a dataset sourced from Villadsen *et al*. (2022), which integrated data from various laboratory studies to create a comprehensive binding energy (BE) dataset. To avoid redundancy, 354 monolayer and 167 multilayer datasets were merged, streamlining it to 117 single datasets of unique molecules ranging from simple diatomics to complex organic molecules (COMs) such as $N_2$, CO, hydrocarbons (e.g $C_8H_{18}$), ethanol ($C_2H_5OH$), glycolaldehyde ($HOCH_2CHO$) etc.

Feature engineering was crucial in transforming this dataset for effective modeling. By extracting molecular features

based on atomic compositions and functional groups via RDKit python module, the molecules were converted to SMILES strings and various properties which are fundamental for molecular binding were computed such as valence electrons, topological polar surface area (TPSA) which captures molecular interactions like hydrogen bonding, Van der Waals forces etc, hydrogen bond donors and acceptors.

Table 1: Overview of Molecular Features Used in the Dataset

| Feature Category | Feature | Description | Examples |
|---|---|---|---|
| **Atoms** | Carbon | Presence of carbon atoms in the molecule. | Graphene, graphite, highly oriented pyrolytic graphite |
| | Hydrogen | Presence of hydrogen atoms in the molecule. | |
| | Nitrogen | Presence of nitrogen atoms in the molecule. | |
| | Oxygen | Presence of oxygen atoms in the molecule. | |
| | Chlorine | Presence of chlorine atoms in the molecule. | |
| | Cyanide | Presence of cyanide (–CN) groups in the molecule. | |
| **Functional Groups** | Alcohol (–OH) | Presence of hydroxyl (–OH) groups in the molecule. | |
| | Amide (–NC(O)–) | Presence of amide (–NC(O)–) groups in the molecule. | |
| | Amine ($–NH_2$) | Presence of amine ($–NH_2$) groups in the molecule. | |
| | Carbonyl (–C(O)–) | Presence of carbonyl (–C(O)–) groups in the molecule. | |
| | Carboxyl (–COOH) | Presence of carboxyl (–COOH) groups in the molecule. | |
| | Ester (–C(O)O–) | Presence of ester (–C(O)O–) groups in the molecule. | |
| | Ether (–O–) | Presence of ether (–O–) groups in the molecule. | |
| **RDKit& Misc.** | Number of H-bond Acceptors | Number of hydrogen bond acceptor atoms in the molecule. | |
| | Number of H-bond Donors | Number of hydrogen bond donor atoms in the molecule. | |
| | Number of Valence Electrons | Total number of valence electrons in the molecule. | |
| | Topological Polar Surface Area | Sum of the surface area of polar atoms in the molecule, measured in $Å^2$. | |
| | Molecular Mass | The total mass of the molecule. | |
| | Number of Atoms | Total number of atoms in the molecule. | |
| **Surface** | Carbon | Surface type involving carbon-based materials. | Graphene, graphite, highly oriented pyrolytic graphite |
| | Metal | Surface type involving metals. | Gold, nickel |
| | Silicate | Surface type involving silicate materials. | Amorphous silicate, forsterite |
| | Water | Surface type involving water. | Amorphous solid water, crystalline water |

To manage the diverse range of surfaces and their effects, the merged dataset categorized surface features into four primary groups: Carbon (e.g., graphene, graphite), Metal (e.g., gold), Silicate (e.g., amorphous silicate), and Water (e.g., amorphous and crystalline water). These categories were numerically encoded using one-hot encoding for input into machine learning models, simplifying the handling of various surface types while ensuring model clarity. This categorization aimed to balance the need for detailed surface characteristics with the practicalities of model training.

As the dataset now consists of a unified collection of data, we did not need to distinguish between monolayer and multilayer entries in our feature engineering process. However, it is important to acknowledge that variations in pre-exponential factors used in some studies and the potential distribution of BEs on certain surfaces could still impact the accuracy of our models. This underscores the importance of future refinements in dataset completeness and model accuracy to enhance the reliability of our predictions.

**Results and Discussion**

In the following sections, we will assess the model's performance through two distinct evaluation techniques.

**Bootstrapping**

Bootstrapping involves creating multiple resamples of the dataset with replacement and then training and testing the model on these resampled datasets (Vrigazova, 2021). This technique helps to come up with a distribution of the performance metrics e.g MAE, RMSE, and $R^2$, across the various bootstrap samples. From the results of the mean and standard deviation calculations to obtain these metrics, a comprehensive understanding of the model's performance can be obtained which may vary with different data subsets, helping to ascertain its stability and generalization capabilities.

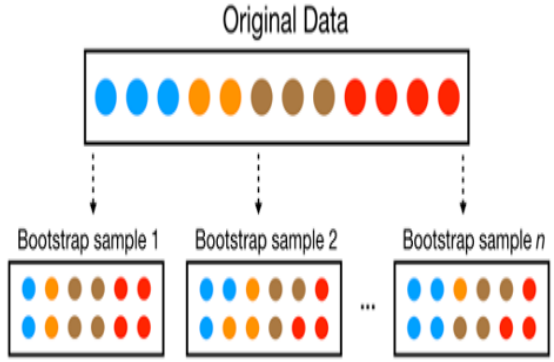


Fig. 2: Visual representation of the bootstrapping process, detailing the original data and the bootstrap samples.

The bootstrapping analysis revealed that the ensemble model, combining predictions from multiple regression models, consistently performed well with a MAE of 0.0327, RMSE of 0.0418, and an $R^2$ of 0.9370. Individual models such as Gradient Boosting demonstrated even lower MAE (0.0125) and RMSE (0.0180), showcasing their precision. The bootstrapping process also highlighted areas of model uncertainty and potential biases, which are crucial for understanding the model's robustness. This thorough evaluation ensures that the model's performance is both reliable and generalizable, guiding further refinements and improving predictive accuracy for practical applications. The corresponding parity plots are shown in the figures below.

**Leave-one-molecule-out cross validation**

Leave-one-molecule-out cross-validation (LOMO-CV) was also employed to assess the model's performance in predicting the BEs. This approach involves excluding each molecule sequentially from the dataset, training the model on the remaining molecules, and then evaluating the model's predictions on the excluded molecule. This method provides a robust measure of how well the model generalizes to new, unseen data.

Table 2 details the predicted binding energies and the corresponding literature values for various molecules, including their molecular formulas. For example, the model predicts a binding energy of 0.1367 eV for methane ($CH_4$) on carbon, with a deviation of 0.0277 eV from the literature value of 0.109 eV. Similarly, the predicted value for Ammonia ($NH_3$) on Carbon is 0.2583 eV, with a minimal deviation of 0.0007 eV from the literature value of 0.259 eV. Water ($H_2O$) on carbon has a predicted binding energy of 0.4644 eV, deviating by 0.0504 eV from the literature value of 0.414 eV. Across all tested molecules, the average deviation of predicted binding energies from the literature values is within ±10%, reflecting the model's overall accuracy while highlighting areas for further refinement. This detailed evaluation through LOMO-CV ensures a comprehensive understanding of the model's predictive capabilities and its reliability in various scenarios.

**Ensemble Model Performance with Mean Predictions:**

The ensemble model, combining predictions from six distinct regression models (Bagging, Linear Regression, Random Forest, Gradient Boosting, Bayesian Ridge, and Ridge Regression), demonstrates robust performance across multiple evaluation metrics, underscoring its efficacy in predicting astrochemical binding energies.

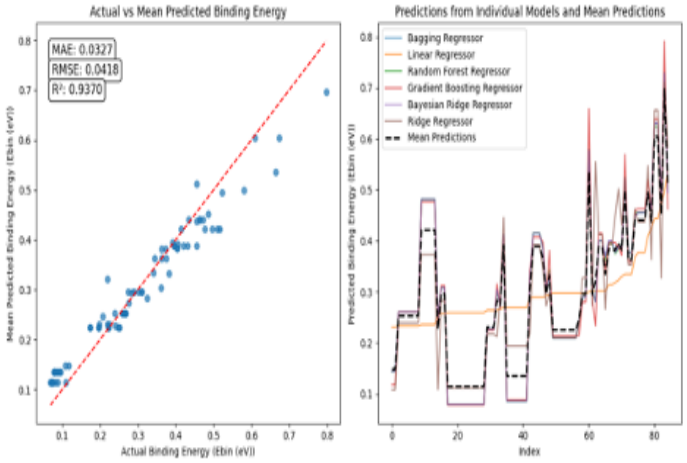


Fig. 3: Performance and Predictions of the Multi-Model Ensemble Approach for Astrochemical Binding Energies

Figure 3 above illustrates the efficacy of a multi-model ensemble approach in predicting astrochemical binding energies with high accuracy. The left subplot features a scatter plot comparing actual binding energies (x-axis) with mean predicted values (y-axis) from the ensemble model. Each blue dot represents a data point, with proximity to the red dashed line indicating minimal deviation from ideal predictions. Model performance metrics include a Mean Absolute Error (MAE) of 0.0327, Root Mean Squared Error (RMSE) of 0.0418, and an R-squared ($R^2$) value of 0.9370, these highlights precision and reliability. The right subplot shows predictions from individual models like Bagging, Linear, Random Forest, Gradient Boosting, Bayesian Ridge, and Ridge regressors, with their mean depicted by a black dashed line. This aggregation smooths variability, enhancing prediction robustness and reducing uncertainty, demonstrating superior accuracy in astrochemical binding energy predictions.

- **Mean Absolute Error (MAE):** 0.0327
- **Root Mean Squared Error (RMSE):** 0.0418
- **R-squared ($R^2$):** 0.9370

**Individual Model Performance:**

**Ridge Regression:**

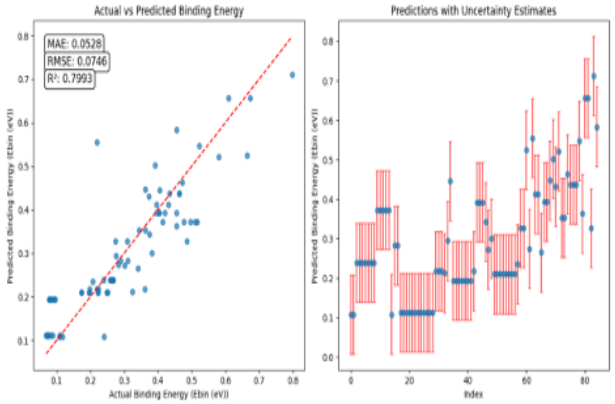


Fig. 4: Ridge Regression Performance and Predictions with Uncertainty Estimates for **Astrochemical Binding Energies**

The performance of the Ridge Regression model in predicting the binding energies are shown in figure 4, along with uncertainty estimates for its predictions. In the plot, the left subplot shows a comparison between actual binding energies ( in the x-axis) and predicted values (in the y-axis) and each blue dotted point represents a data point, the red dashed line corresponds to perfect prediction alignment. The outcome of the model revealed the following:

- MAE: 0.0528
- RMSE: 0.0746
- $R^2$: 0.7993

This indicates that $\sim$ 79.93% of the variance in actual binding energies are captured by the model's predictions. The right subplot shows individual predictions (blue dots) and their associated uncertainties (red error bars) across data points (x-axis). This visualizes the model's predictive uncertainty, showing that while many predictions are accurate, some exhibit significant variability, emphasizing areas where the model's confidence is lower.

Ridge regression performs adequately with moderate error metrics and a reasonable fit ($R^2$). It serves as a comparative baseline for evaluating other models.

**Bagging Regressor:**

Figure 5 depicts the performance of the Bagging Regressor. The left subplot is a scatter plot comparing actual binding energies (x-axis) with predicted binding energies (y-axis). Each blue dot represents a data point. The tight clustering of the blue dots along the red dashed line, which signifies perfect prediction, indicates that the model's predictions are very close to the actual values. This alignment suggests a high level of accuracy, further supported by the metrics: a Mean Absolute Error (MAE) of 0.0185, a Root Mean Squared Error (RMSE) of 0.0256, and an R-squared ($R^2$) of 0.9764. These metrics imply that the model has a low average prediction error and that 97.64% of the variance in actual binding energies is accounted for by the predictions, reflecting a strong predictive capability.

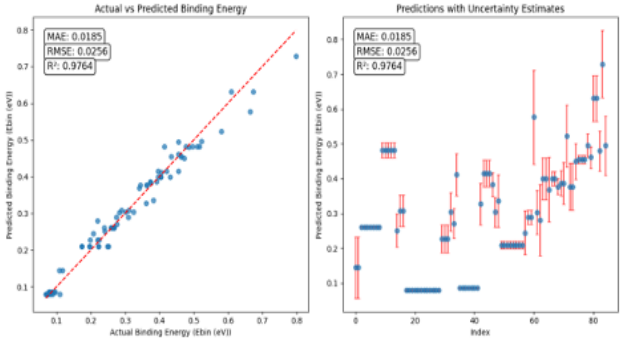


Fig. 5: Performance and Uncertainty Estimates of the Bagging Regressor for Predicting Astrochemical Binding Energies

- **MAE: 0.0185**
- **RMSE: 0.0256**
- **$R^2$: 0.9764**

The bagging regressor significantly outperforms Ridge regression, exhibiting substantially lower MAE and RMSE, indicative of superior predictive accuracy and a higher explained variance ($R^2$).

**Gradient Boosting Regressor:**

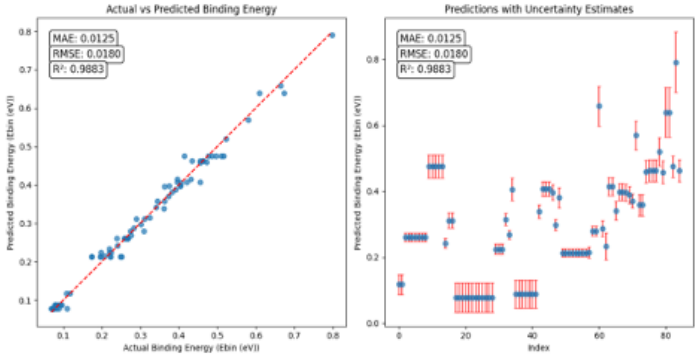


Fig. 6: Performance and Uncertainty Estimates of the Gradient Boost Regressor for Predicting Astrochemical Binding Energies

- **MAE: 0.0125**
- **RMSE: 0.0180**
- **$R^2$: 0.9883**

Gradient boosting exhibits higher predictive performance in comparison with all other individual models, achieving the lowest MAE and RMSE, and the highest $R^2$. This model excels in capturing the intricate relationships within the dataset, yielding precise predictions with minimal error.

**Random Forest Regressor:**

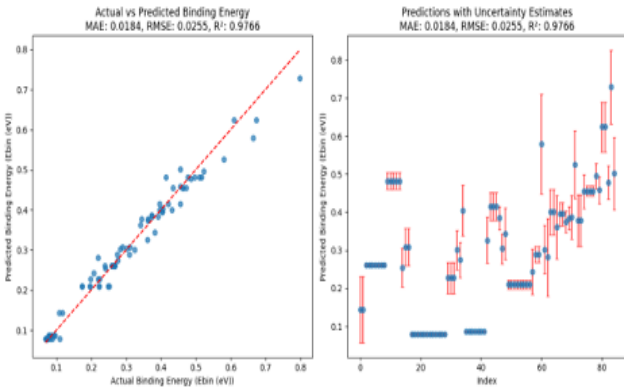


Fig. 7: Performance and Uncertainty Estimates of the Random Forest Regressor for Predicting Astrochemical Binding Energies

- **MAE: 0.0184**
- **RMSE: 0.0255**
- **R²: 0.9766**

Similar to bagging, random forest demonstrates strong performance with low error metrics and a high $R^2$, indicating robust predictive capabilities suitable for complex data patterns.

**Lasso Regression:**

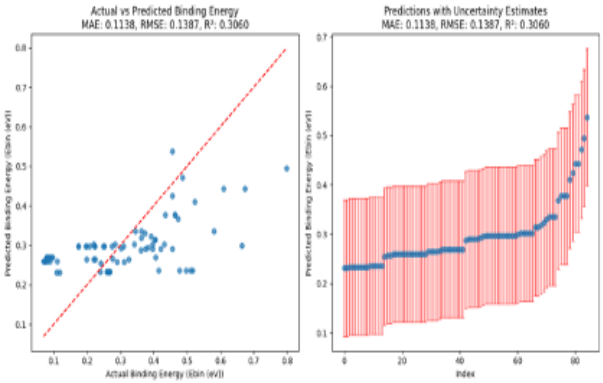


Fig. 8: Performance and Uncertainty Estimates of the Lasso Regression for Predicting Astrochemical Binding Energies

- **MAE: 0.1138**
- **RMSE: 0.1387**
- **R²: 0.3060**

Lasso regression exhibits the weakest performance among the models tested, characterized by higher MAE and RMSE, and a lower $R^2$. This model struggles to effectively explain variance compared to more flexible approaches.

**Bayesian Ridge Regression:**

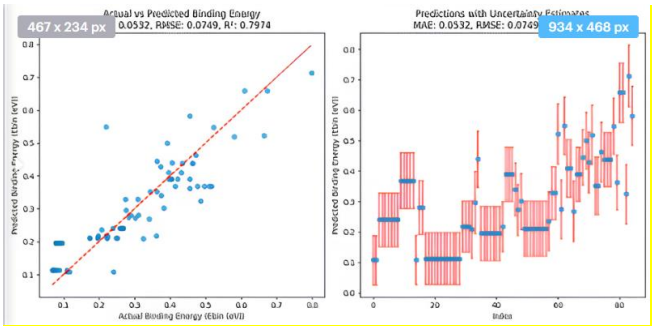


Fig. 9: Performance and Uncertainty Estimates of the Bayesian Ridge Regression for Predicting Astrochemical Binding Energies

- **MAE: 0.0532**
- **RMSE: 0.0749**
- **R²: 0.7974**

Bayesian ridge regression performs similarly to Ridge regression, providing moderate error metrics and a reasonable fit to the data.

**Discussion**

**Model Comparison**

Gradient boosting stands out as the top performer among individual models, demonstrating exceptional predictive accuracy and variance capture. This model is known to perform better due to its ability to correct errors of previous predictions (Hastie *et al.*, 2009). Its ability to minimize errors and maximize $R^2$ underscores its suitability for precise astrochemical binding energy predictions.

**Ensemble Model Performance:**

The ensemble model leverages the strengths of individual models, achieving performance metrics (MAE, RMSE, $R^2$) comparable to or exceeding those of its constituent parts. This amalgamation mitigates model-specific weaknesses, enhancing overall predictive robustness.

**Uncertainty and Confidence:**

Lower MAE and RMSE values in individual models correlate with greater prediction confidence. Models like gradient boosting and random forest, with superior performance metrics, offer reliable predictions crucial for scientific and engineering applications.

**Desorption**

The figure 10 below illustrates the desorption profiles for eight of the astronomically relevant molecules adsorbed onto various surfaces, analyzed under a heating rate of 1 K century$^{-1}$. Each subplot shows the desorption rate as a function of temperature, with the peak desorption temperature (T□) indicated by a green dashed line.

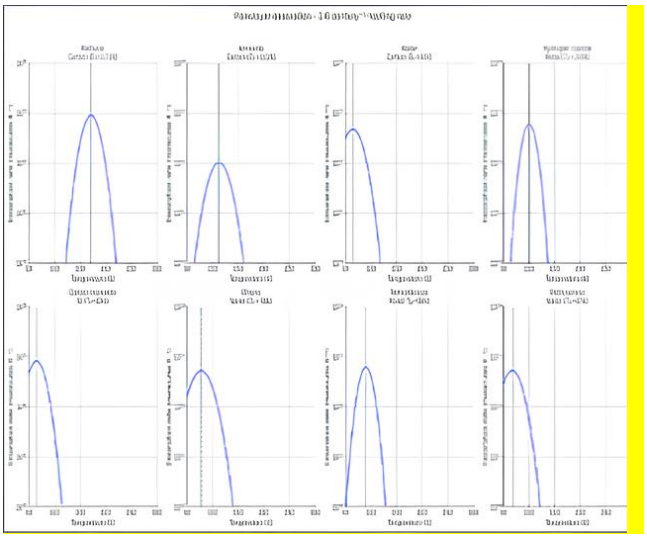

Fig 10: Desorption profiles for eight of the astronomically relevant molecules adsorbed onto various surfaces, analyzed under a heating rate of 1 K century$^{-1}$.

**Table 2: Key Observations from the peak desorption temperature predictions**

| S/N | Molecule | Peak Desorption Temperature $(T_p)$, (K) | Implication/Description |
|---|---|---|---|
| 1 | Methane ($CH_4$) on carbon | ~171 | Relatively easy desorption, sharp peak indicating narrow temperature range. Suggest homogenous binding energy on carbon surface |
| 2 | Ammonia ($NH_3$) on Carbon | 113 | Lower $T_p$ compared to methane, suggest weaker interaction with the carbon surface, narrow peak implies uniform adsorption sites |
| 3 | Water ($H_2O$) on Carbon | ~65 | Low $T_p$, reflecting weak van der Waals interactions, sharp peak indicates a homogeneous interaction |
| 4 | Hydrogen Cyanide ($HCN$) on Metal | ~100 | Slightly broader peak than methane and ammonia, variation in binding energy on the metal surface |
| 5 | Carbon Monoxide (CO) on Silicon (Si) | 65 | Low $T_p$, i.e weak physisorption on the silicon surface, sharp peak suggests uniform adsorption energy. |
| 6 | Ethane ($C_2H_6$) on Water | 78 | Relatively low $T_p$ implies weak interactions, narrow peak indicates minimal variation in binding energies. |
| 7 | Formaldehyde ($H_2CO$) on Metal | 90 | Relatively sharp peak implies uniform binding sites and consistent interaction strength on the metal surface |
| 8 | Methylamine ($CH_3NH_2$) on Water | 67 | $T_p$ slightly higher than water but lower than most other molecules weak physisorption with possible hydrogen bonding to the water surface. The sharp peak suggests uniform interaction |

**Implications**

The desorption temperatures (T□) provide insights into the interaction strength between each molecule and its respective surface. Higher T□ values (e.g., methane on carbon) indicate stronger binding and require higher temperatures for desorption, whereas lower T□ values (e.g., water on carbon, CO on silicon) suggest weaker interactions.

Table 3: Predicted Binding Energies and Actual Binding Energies of some of the molecules

| Name | Surface | Predicted Ebin (eV) | Literature Ebin (eV) | Deviation (eV) |
|---|---|---|---|---|
| Methane | Carbon | 0.1367479 | 0.109 | 0.0277479 |
| Methane | Si | 0.1367479 | 0.118 | 0.0187479 |
| Ammonia | Carbon | 0.2582816 | 0.259 | 0.0007184 |
| Ammonia | Metal | 0.2582816 | 0.264 | 0.0057184 |
| Water | Carbon | 0.4644318 | 0.414 | 0.0504318 |
| Acetylene | Water | 0.239338 | 0.241 | 0.001662 |
| Hydrogen cyanide | Metal | 0.3057397 | 0.311 | 0.0052603 |
| Carbon monoxide | Si | 0.0869211 | 0.069 | 0.0179211 |
| Ethane | Metal | 0.2268072 | 0.198 | 0.0288072 |
| Formaldehyde | Metal | 0.2990792 | 0.324 | 0.0249208 |
| Methylamine | Carbon | 0.2725481 | 0.276 | 0.0034519 |
| Methanol | Metal | 0.4061696 | 0.405 | 0.0011696 |
| Oxygen | Si | 0.1002881 | 0.078 | 0.0222881 |
| Methylacetylene | Water | 0.3213736 | 0.362 | 0.0406264 |
| Acetonitrile | Si | 0.4069568 | 0.396 | 0.0109568 |
| Isocyanic acid | Metal | 0.3470626 | 0.383 | 0.0359374 |
| Carbon dioxide | Water | 0.2136935 | 0.174 | 0.0396935 |
| Nitric oxide | Si | 0.3353052 | 0.337 | 0.0016948 |

These findings enhance our understanding of surface chemistry and adsorption processes, particularly in the contexts of astrochemistry and environmental science where low-temperature desorption is significant. The uniform desorption peaks imply that the surfaces studied offer relatively homogeneous adsorption sites, simplifying the modeling of desorption kinetics in larger systems.

## Conclusion

The ensemble of diverse regression models effectively predicts astrochemical binding energies, surpassing the predictive capacity of individual models. This study highlights the prospects of ensemble techniques in advancing predictive modeling within complex chemical systems, offering insights and methodologies applicable across various scientific domains. Understanding the complex chemical processes driving molecule formation, evolution, and interactions in the astrochemical environment is crucial for determining molecular stability, reactivity, and potential for life in space. Accurately predicting desorption energies is vital for modeling dynamic chemical environments in astrophysical settings. Traditional methods like Density Functional Theory (DFT) and Hartree-Fock are highly accurate but require substantial computational resources, making them impractical for large-scale studies involving diverse molecular species. Temperature-programmed desorption (TPD) is one of the main experimental techniques used to measure binding energies but is often limited by complexity, high costs, and time. Machine learning (ML) has emerged as an alternative method that strikes a balance between accuracy and efficiency, offering improved performance by leveraging the strengths of different predictive approaches and addressing the weaknesses of individual models. Gradient boosting is the top performer in individual models for predictive accuracy and variance capture, making it suitable for precise astrochemical binding energy predictions. Ensemble models, which combine the strengths of individual models, achieve performance metrics comparable to or exceeding their constituent parts, enhancing overall predictive robustness. Models like gradient boosting and random forests offer reliable predictions crucial for scientific and engineering applications.

**Declaration**: The authors declare no conflict of interest.

## References


Andrew, C., Etim E. E, Ushie, O. A.; Khanal. G. P. (2018). Vibrational-Rotational Spectra of Normal Acetylene and Doubly Deuterated Acetylene: Experimental and Computational Studies. *Chemical Science Transactions7*(1), 77-82. DOI:10.7598/cst2018.1432

AwujoolaOlalekan, J., Ogwueleka, F., &Odion, P. (2020). Effective and accurate bootstrap aggregating (Bagging) ensemble algorithm model for prediction and classification of hypothyroid disease. International Journal of Computer Applications, 975, 8887.https://doi.org/10.5120/975-8887

Bentéjac, C., Csörgő, A., &Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.https://doi.org/10.1007/s10462-020-09896-5

Bovolenta, G., Bovino, S., Vöhringer-Martinez, E., Saez, D. A., Grassi, T., & Vogt-Geisse, S. (2020). High level ab initio binding energy distribution of molecules on interstellar ices: Hydrogen fluoride. Molecular Astrophysics, 21, 100095.https://doi.org/10.1016/j.molap.2020.100095

Etim, E. E., Benchmark Studies on the Isomerization Enthalpies for Interstellar Molecular Species *J. Nig. Soc. Phys. Sci.* 2023,5, 527. https://doi.org/10.46481/jnsps.2023.527https://arxiv.org/abs/2302.05911

Etim, E. E., Gorai, P., Das, A., &Arunan, E. (2017). Interstellar protonated molecular species. Advances in Space Research, 60(3), 709-721. https://doi.org/10.1016/j.asr.2017.04.003

Etim, E. E., Gorai, P., Das, A., Chakrabarti, S., and Arunan, E. (2018). Interstellar Hydrogen Bonding. Advances in Space Research, 61(11): 2870-2880, https://doi.org/10.1016/j.asr.2018.03.003.

Etim, E. E., Mbakara, I. E., Khanal, G. P., Inyang, E. J., Ukafia, O. P., Sambo, I. F,. Coupled Cluster Predictions of Spectroscopic Parameters for (Potential) Interstellar Protonated Species. *Elixir Computational Chemistry*,2017, 111: 48818-48822.

Etim, E. E., Benchmark Studies on the Isomerization Enthalpies for Interstellar Molecular Species. J. Nig. Soc. *Phys. Sci*. 2023**,** 5, 527. https://doi.org/10.46481/jnsps.2023.527https://arxiv.org/abs/2302.05911

Etim, E.E, AkpanNdemIkot, Ruth O. Adelagun, Usman Lawal.. Deuterated Interstellar and Circumstellar Molecules: D/H Ratio and Dominant Formation Processes. *Indian Journal of Physics*2020,.https://doi.org/10.1007/s12648-020-01747-x

Etim, E.E, and E. Arunan. Interstellar Isomeric Species: Energy, Stability and Abundance Relationship. *European Physical Journal Plus*, 2016,131:448. DOI 10.1140/epjp/i2016-16448-0

Etim, E.E., Onudibia, M. E., Asuquo, J. E., Ukafia, O. P., Andrew, C., Ushie, O.A.(2017). Interstellar $C_3S$: Different Dipole Moment, Different Column Density, Same Astronomical Source, *FUW Trends in Science and Technology Journal*, 2 (1B): 574-577.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). Boosting and additive trees. *InThe elements of statistical learning* (2nd ed., pp. 337–384). Springer.https://doi.org/10.1007/978-0-387-84858-7

Hirao, K., Nakajima, T., & Chan, B. (2023). Core-level 2s and 2p binding energies of third-period elements (P, S, and Cl) calculated by Hartree–Fock and Kohn–Sham ΔS CF theory. The Journal of Physical Chemistry A, 127(38), 7954-7963.https://doi.org/10.1021/acs.jpca.3c04781

Imane, M., Aoula, E. S., &Achouyab, E. H. (2022). Using Bayesian ridge regression to predict the overall equipment effectiveness performance. In 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) (pp. 1-4). IEEE.https://doi.org/10.1109/IRASET55029.2022.9769327

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Springer International Publishing.https://doi.org/10.1007/978-3-031-18305-5_3

Johnson, K. N., Li, Y., Ezell, M. J., Lakey, P. S., Shiraiwa, M., & Finlayson-Pitts, B. J. (2024). Elucidating gas–surface interactions relevant to atmospheric particle growth using combined temperature programmed desorption and temperature-dependent uptake. Physical Chemistry Chemical Physics.https://doi.org/10.1039/d3cp04561a

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129-99149.https://doi.org/10.1109/ACCESS.2022.3202724

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 35(2), 757-774.https://doi.org/10.1016/j.jksuci.2022.06.006

Oba, Y., Miyauchi, N., Hidaka, H., Chigai, T., Watanabe, N., and Kouchi, A. (2009). Formation of compact amorphous $H_2O$ ice by codeposition of hydrogen atoms with oxygen molecules on grain surfaces, *Astrophys. J.* 701, 464–470. doi:10.1088/0004-637X/701/1/464

Oladimeji, E. O., Etim, E. E., Umeh, E. C., Shinggu, J. P., Oluwadare, O. J., Odeyemi, O. M., & Samuel, H. S. (2024). Exploring the Thermodynamic Characteristics of Isoelectronic Diatomic Interstellar Molecular Species: Oxygen and Sulfur Containing Specie. *UMYU Scientifica*, *3*(2), 146-158.

Osigbemhe, I.G., Louis, H., Khan, E.M., Etim, E. E., Odey, D. O., Oviawe, A. P., Edet, H. O., Obuye, F. (2022a). Synthesis, characterization, DFT studies, and molecular modeling of 2-(-(2-hydroxy-5-methoxyphenyl)-methylidene)-amino) nicotinic acid against some selected bacterial receptors. *J IRAN CHEM SOC*https://doi.org/10.1007/s13738-022-02550-7

Petrignani, A., &Candian, A. (2022). Astrochemistry: Ingredients of life in space. In New Frontiers in Astrobiology (pp. 49-66). Elsevier.https://doi.org/10.1016/B978-0-12-824162-2.00007-5

Samuel, H. S., Etim, E. E., Oladimeji E.O., Shinggu J.P., &Bako B. (2023). Machine Learning in Characterizing Dipole-Dipole Interactions. *FUW Trends in Science & Technology Journal*, *8*(3), 070-082

Samuel, H. S., Etim, E. E., Shinggu, J. P., &Bako, B. (2024). Machine Learning in Thermochemistry: Unleashing Predictive Modelling for Enhanced Understanding of Chemical Systems. Communication in Physical Sciences, 11(1), 47-75

Samuel, H. S., Etim, E., Nweke-Maraizu, U., Bako, B., Shinggu, J. P. (2023). Advances in Experimental Techniques for Corrosion Inhibition Studies: Insights and Applications. J. Appl. Sci. Environ. Manage. 27 (12) 2957-2966

Samuel, H. S., Nweke-Maraizu, U., and Etim, E. E. (2023). Machine learning for characterizing halogen bonding interactions. Faculty of Natural and Applied Sciences Journal of Scientific Innovations, 5(1), 103–115.https://www.fnasjournals.com/index.php/FNAS-JSI/article/view/208

Samuel, H. S., Nweke-Maraizu, U., and Etim, E. E. (2024). Unleashing the Potential of Machine Learning in Chalcogen Bonding Research. *Eurasian Journal of Science and Technology*, 4(3), 133-164. doi: 10.48309/EJST.2024.416374.1091

Samuel, H.S,Etim, E.E., Ugo Nweke-Maraizu., Shinggu, J.P., Bako B (2023). Machine Learning of Rotational Spectra analysis in Interstellar medium. Communication in Physical Sciences, 10(1): 172-203.

Samuel, H.S., Etim, E.E., Nweke-Maraizu, U., &Yakubu, S. (2024). Machine Learning in Chemical Kinetics: Predictions, Mechanistic Analysis, and Reaction Optimization. Applied Journal of Environmental Engineering and Science, 10(1), 36-6.https://www.researchgate.net/publication/380291626_Machine_Learning_in_Chemical_Kinetics_Predictions_Mechanistic_Analysis_and_Reaction_Optimization

Shinggu, J. P., Etim, E. E., and Onen, A. I., (2023). Quantum Chemical Studies on C2H2O Isomeric Species: Astrophysical Implications, and Comparison of Methods. *Communication in Physical Sciences*, **2023**, 9(2): 93-105.

Siebenmorgen, T., & Zacharias, M. (2020). Computational prediction of protein–protein binding affinities. Wiley Interdisciplinary Reviews: Computational Molecular Science, 10(3), e1448.https://doi.org/10.1002/wcms.1448

Smith, R. S., & Kay, B. D. (2018). Desorption kinetics of benzene and cyclohexane from a graphene surface. The Journal of Physical Chemistry B, 122(2), 587-594.https://doi.org/10.1021/acs.jpcb.7b09618

Spiegelman, F., Tarrat, N., Cuny, J., Dontot, L., Posenitskiy, E., Martí, C., Simon, A., &Rapacioli, M. (2020). Density-functional tight-binding: basic concepts and applications to molecules and clusters. Advances in Physics: X, 5(1), 1710252.https://doi.org/10.1080/23746149.2019.1710252

Tufail, S., Riggs, H., Tariq, M., &Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. Electronics, 12(8), 1789.https://doi.org/10.3390/electronics12081789

Villadsen, T., Ligterink, N., & Andersen, M. (2022). Predicting binding energies of astrochemically relevant molecules via machine learning. arXiv preprint arXiv:2207.03906.https://doi.org/10.48550/arXiv.2207.03906

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy, 12(1), 228-242.https://doi.org/10.2478/bsrj-2021-0015

Zhu, T. (2020). Analysis on the applicability of the random forest. In Journal of Physics: Conference Series (Vol. 1607, No. 1, p. 012123). IOP Publishing.https://doi.org/10.1088/1742-6596/1607/1/012123