Research Article

Comparing Simple Neural Network and Ensemble Learning Models in Predicting Hydration Energy of Molecules Represented by RDKit and Mordred Descriptors

Ibrahim Irfan

Alpharetta High School, 3595 Webb Bridge Rd, Alpharetta, GA, 30005, USA

ABSTRACT

Predicting molecular properties is a crucial task in the chemical sciences. Recently, there has been an enormous focus on leveraging machine learning to predict various molecular properties ranging from solubility to reaction rates. This study develops machine learning models to predict the hydration energy of molecules, comparing the performance of a neural network (Multi-Layer Perceptron) and an ensemble learning model (Random Forest). Using descriptors generated by RDKit and Mordred, we aimed to identify the optimal molecular representations for predictive accuracy. The FreeSolv database of 642 molecules provided experimental hydration energy data for training and testing. The models were evaluated using mean squared error (MSE) and the coefficient of determination (R^2), with the Multi-Layer Perceptron achieving an R^2 above 0.9, outperforming the Random Forest model. Results suggest that the neural network model, in combination with RDKit descriptors, offers a strong balance between accuracy and computational efficiency. This study demonstrates the potential for simpler machine learning models to accurately predict molecular properties, supporting broader applications in chemistry where computational resources are limited.

Keywords: machine learning; chemistry; cheminformatics; solvation; modelling; hydration energy

INTRODUCTION

Artificial Intelligence (AI) has become a powerful tool for innovation in the field of chemistry, with numerous applications that are transforming how chemists approach complex challenges. From retrosynthesis predictions to molecule discovery, AI is playing an integral role in streamlining and enhancing research and development. According to Mitsubishi Tanabe Pharma, scientists have been able to use AI to analyze vast amounts of data, coming from complex databases of AI's transformative role in chemistry, enabling breakthroughs in drug discovery, molecular property prediction, and structural analysis, revolutionizing R&D efficiency.

However, such large-scale modeling requires substantial compute time and data to develop. These models are costly to run and are hosted on powerful systems that drain both power and space, reducing the deployability of

Corresponding author: Ibrahim Irfan, E-mail: ibi12341@icloud.com.

Received November 24, 2024**; Accepted** December 15, 2024
https://doi.org/10.70251/HYJR2348.24183189

AI in industry. Experts predict that by 2030, the average production-ready model would take 200X (5.4 Gigawatts) of the stated power, or 30% of all power currently used by data centers (1). Powerful, large-scale AI models require substantial computational resources to operate on a regular basis, posing deployability challenges. To address this, simpler models like Random Forest and Multi-Layer Perceptron (MLP) can balance performance with efficiency. A Random Forest Regressor is known for its performance enhancement for datasets that have both numerical and categorical features (2). An MLP model is a fully connected neural network, which is theoretically able to model any non-linear patterns. In the context of chemistry, the representation of molecules is just as critical to the success of predictive models as the choice of algorithm. In the field of computational chemistry, there are two leading descriptor calculators: RDKit (3) and Mordred (4).

A specific application of AI in chemistry is molecular property prediction, where a molecular structure is mapped to an output property. Recent work has explored predicting numerous properties, including solubility, ionization potential, electron affinity, and hydration energy. (5–8) Hydration energy refers to the energy released when ions dissolve in water, forming intermolecular bonds with water molecules and creating a hydration shell around ions. This process is pivotal in understanding the behavior of aqueous ions in pharmaceutical and environmental contexts.

Overall, this paper aims to create machine learning models that are able to predict the hydration energy of molecules, evaluating the prediction accuracy of both neural networks and an ensemble learning model , while also evaluating the use of two RDKit and Mordred featurizers on generating the molecular representations that provide optimal performance. To accomplish this comparison of both models and descriptor calculators, we will leverage the FreeSolv database, a comprehensive account of 642 molecules, with accounts of their experimental hydration energies; The database is used often in research in development of models that predict hydration energy using alternate methods - typically more complex ones, in order to achieve a more accurate result, with one such paper assessing the property based on Molecular Density Functional Theory (9). The objective of this comparison and the development of the model in general is to create a model that aims to utilize either one of these models to create an accurate extrapolation prediction model, with a coefficient of determination value of above 0.9.

MATERIALS AND METHODS

Dataset and Tools

This study used the FreeSolv database (642 molecules) with molecular weights from 5 to 500 g/mol (Figure 1A) and hydration energies ranging from –35 to 5 kcal/mol (Figure 1B).

To achieve the goal of finding the best model to predict these hydration energies, I used Python's Sci-kit Learn package to appropriately manage the dataset as well as the model training. To generate a list of features, I selected Python's RDKit package, a popular choice in the community for its efficiency among other options. For the sake of comparison, I also utilized Mordred for feature calculation, which is a less popular choice due to its inefficiency, as evidenced by its calculation of 1800 two and three dimensional features.

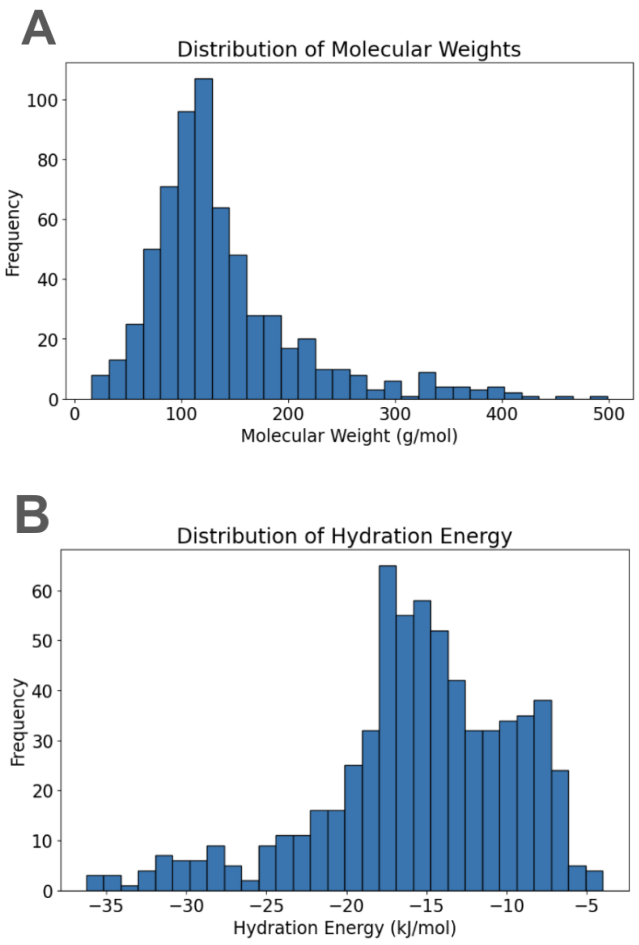

Figure 1. Distributions of (A) Molecular Weight and (B) Hydration Energy within FreeSolv.

Data Splitting and Preparation

The dataset was split into 80% for training the models, and 20% for testing the models, ensuring that the models were tested on unseen data. A Bemis-Murcko Scaffold split ensured diversity in training and testing data, with SMILES strings converted to molecule objects, each scaffold uniquely represented. Finally, all the unique scaffolds were mapped back to their original molecule objects, which were then fed into a featurizer method that returned a list of 208 features for the molecule (Figure 2).

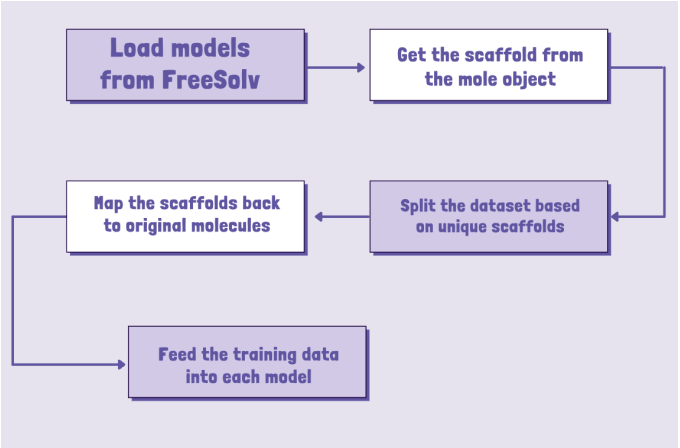

Figure 2. Preprocessing workflow.

Feature Selection

Feature selection was performed using a combination of Variance Threshold and SelectKBest within GridSearchCV to refine the dataset by identifying the most relevant features for predicting Hydration Energy. The Variance Threshold is a simple feature selection technique that removes features with low variance, meaning those that do not vary significantly across samples. Features with very low variance often do not contribute useful information to the model and can introduce noise. The method eliminates features below a certain variance threshold, ensuring only those with sufficient variability are retained (10). SelectKBest is another feature selection method that ranks features based on their correlation with the target variable and retains only the top K features. It uses statistical tests like ANOVA F-values or chi-squared tests (for classification problems) to assess how strongly each feature correlates with the target variable. The f_classification scoring metric was employed in both methods to evaluate and select features with the highest predictive power.

Metrics

Model performance was assessed using mean squared error (MSE), which measures the average squared differences between predicted and actual values, penalizing larger errors for robustness. The Coefficient of Determination (R^2) was also used to evaluate correlation strength between predicted and actual values, offering insight into overall model accuracy.

RESULTS

Model Performance

The MLP model performed best, achieving an MSE of 5.35 and an R^2 of 0.91, outperforming more complex models like Density Functional Theory predictions. Its R^2 value indicates that 91% of the variance in hydration energy was explained, showing a strong correlation between predicted and actual values, with parity plots closely clustering around the 45-degree line. The MLP's higher feature selection through grid search enhanced its accuracy compared to Random Forest, which tested fewer features. While R^2 indicates correlation strength, MSE was also used to highlight performance differences more effectively (Figure 3 and Figure 4).

Featurizer Comparison

In addition, the top performing model was tested with features calculated with both RDKit as well as Mordred, and its inclusion and description are discussed above. However, as seen in the figure above, we can see no significant difference between the two featurizers and their performance in enhancing the model. While this does not show the superiority of one feature selection tool over the other - despite RDKit taking significantly less time (Table 1), it does provide evidence that there is no feature that contributes to the prediction of hydration energy that is so specialized that it is only available in one of the two feature selection sets.

Feature Importance

According to the beeswarm graphs, the most important feature that the MLP regressor considered when making its predictions is the PEOE_VSA1 feature - as seen by its highest mean absolute SHAP value. The feature is one that is impossible to decipher through just the codename it is given without specific background understanding on naming conventions, as well as complex chemical concepts. Essentially, VSAX features were one of three libraries created by Paul Labute, that would be useful in Quantitative Structure-Activity Relationship (QSAR)

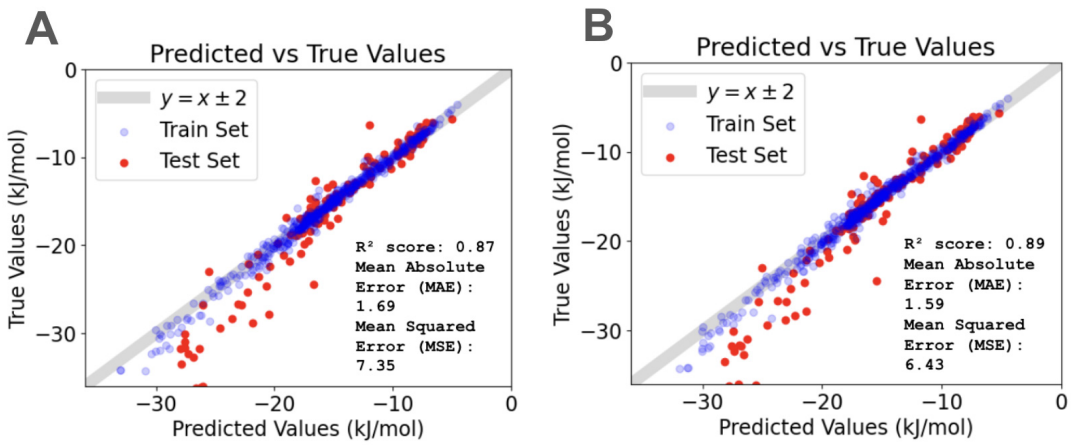

Figure 3. Results of Random Forest Models. Left (A) has an R^2 of 0.87. Right (B) has an R^2 of 0.89.

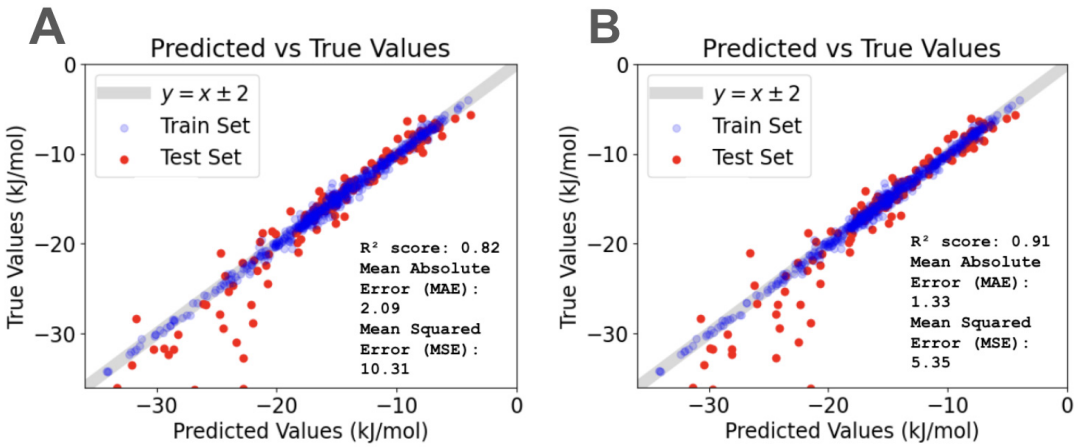

Figure 4. Results of the MLP models. Left (A) has an R^2 of 0.82. Right (B) has an R^2 of 0.91.

Table 1. Comparison of RDKit and Mordred descriptor calculators

Metric	RDKit	Mordred
Accessibility	Imported through conda or pip	Can only be imported through pip
Number of Descriptors	~1800	~640
Performance (on 4 cores)	**5 minutes 37 seconds** to calculate the descriptor list for all FreeSolv Molecules	**24 hours 10 minutes** to calculate the descriptor list for all FreeSolv Molecules.
Documentation	Extensively maintained	Limited

Studies (11). VSA is short for Van der Waals Surface area, while PEOE stands for partial equalization of orbital electronegativity. PEOE-VSA descriptors describe the properties of chemical bonds rather than the atomic numbers or weights in molecular structures (12). Overall, the feature aims to quantify the surface area contributed by different atoms, considering their partial charges x. This connects to a molecule's Surface Area, to consider its interactions with water. Solvent Accessible Surface Area (SASA) is defined by the area of the surface of the molecule that is available to react with its corresponding solvent. This means that a larger SASA value indicates a great surface area available to interact with the solvent. This directly correlates to how easily a water shell can form around the molecule. If it is more difficult to form a water shell, more energy will be required to form it, resulting in less energy that can be released. Because this applies to all molecules, this is generated as one of the most important factors considering hydration energy, as agreed with the MLP results.

According to both plots in Figure 4, the second most valuable feature is BCUT2D_LOGPHI. This feature comes from the BCUT2D descriptor set, which is a set of molecular descriptors used to represent the two-dimensional structure and topological features of a molecule in cheminformatics (12). According to the Royal Society of Chemistry, the LOGPHI feature is described as the highest eigenvalue weighted by crippen logP (13). In this context, an eigenvalue represents a scalar that indicates how much a given feature or vector is stretched or scaled during a linear transformation. When applied to molecular data, eigenvalues help identify key characteristics, like the most influential chemical properties, that are important for predicting behaviors such as hydrophobicity (14). Essentially, the feature describes the spatial distribution of hydrophobicity within the molecule, or a molecule's ability to repel water. This is an interesting next-best valued feature, since the PEOE-VSA feature was related to how the molecule interacts with water. However, this analysis may be inaccurate as hydrophobicity complements hydrophilicity, which describes a water-attracting property. A low LOGPHI value indicates a distribution that favor's water attraction, increasing interactions between molecules and the water solvent they are dissolved in, once again greatly varying the hydration energy, through the explanation of Solvent Accessible Surface Area seen above.

In addition, the emphasis on valence electrons and accountancy for natures of all bonds is able to record the distribution of partial charges across the molecules, and brings in the importance of partial charge and dipole moments, a pivotal factor when considering hydration energy. The significance of Dipole Moments comes from its interaction with water. The property itself highlights its polarization and separation of negative and positive charges that concentrate at each end of the molecule. Similarly, water is known to be a common polar solvent, signifying a more significant interaction with another polar molecule through their dipole-dipole interactions. Therefore, higher dipole-moment-molecules tend to dissolve in water more easily, and water molecules have an easier and less energy-inducing experience surrounding the molecule. This significantly affects the hydration energy, as it is primarily determined by how much energy is required to break intramolecular forces, as well as the difficulty to surround the molecule.

The model's ability to capture these complex interactions through features like Chi1v and PEOE-VSA

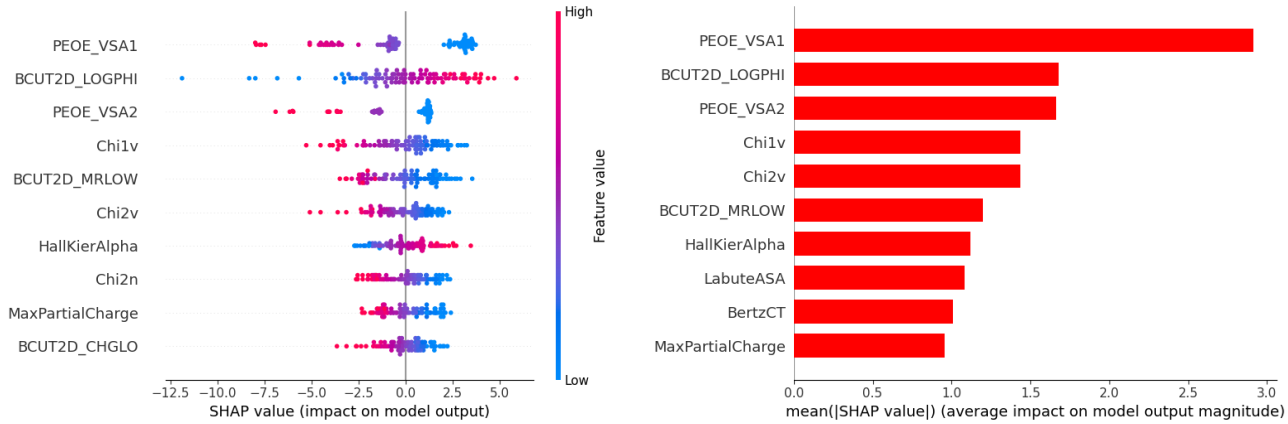

Figure 5. Beeswarm plot of features, as well as a plot of absolute mean impact.

demonstrates its effectiveness in predicting hydration energy. Specifically, it utilized features that were expected, beforehand, to have a decently strong correlation with hydration energy, with logical explanations on the molecular level that detail the interactions between molecules and water.

DISCUSSION

Other work has attempted to predict hydration energy on FreeSolv using quantum mechanical methods, with high accuracy (9, 15). For example, utilizing Hypernetted Chain Approximation (HCA) a Pearson R value of 0.93 was achieved, matching the accuracy of free energy calculations, while reducing the computation time from hundreds of CPU hours to two CPU minutes per molecule (9). Another work used Molecular Density Functional Theory, also achieved a Pearson R value of 0.9 (16). Despite being vastly more complex in the methodologies chosen, these models perform similarly to the MLP model made above. To compare, R^2 indicates correlation strength, while MSE penalizes larger errors, offering complementary insights into model performance. Although the MLP model shows a weaker MSE than more complex techniques, it processes data significantly faster—taking only a fraction of a second per molecule, compared to models like the HCA model at 2 CPU minutes per molecule. Even within the realm of more elementary models, The MLP's interconnected neurons capture complex, non-linear patterns, enhancing accuracy for intricate datasets. In contrast, decision tree-based models, which average outputs from multiple trees, are better suited for simpler relationships but may miss deeper complexities (17).

Our analysis revealed weaker model predictions for lower hydration energies, a trend seen across studies using the FreeSolv dataset, such as one utilizing an implicit solvent model done by the Technical University of Munich, which clearly demonstrates weaker predictions in lower ranges due to a lack of data (18). As shown in the left-skewed distribution of Panel B, Figure 1.1, and highlighted by the parity plots in Figures 2.1 and 2.2, fewer molecules are reported in lower hydration ranges, limiting predictive accuracy for these values. This data limitation, rather than model performance, accounts for the reduced accuracy in lower ranges, underscoring the need for an expanded dataset. Updating FreeSolv or creating larger datasets would improve model accuracy and address data scarcity challenges in chemical research. Advanced models, like the Graph Neural Networks (GNNs) from Shimakawa et al., show promise for capturing complex molecular interactions across a wider range, especially for low-value predictions (19). By incorporating interaction terms between descriptors, as seen in models like QMex-ILR, these next-generation models could overcome limitations faced by simpler approaches.

Despite this, simplistic models like the one showcased here can achieve comparable results to more complex ones while using fewer computational resources. This emphasizes the portion of AI models in industry that tend to be overlooked—those focused on fundamental machine learning algorithms available through Python packages. The stability of molecular properties, which do not change as frequently as data in other fields, supports the use of these simpler models in chemistry research, as they can be built on pre-existing datasets without needing new training sets from scratch. For example, researchers recently used an AI-driven approach to search for molecules with large polarizability and electronic gaps, identifying new pathways in chemical space with unexpected molecular structures (20). However, the limitations of simpler models become evident when predicting more complex molecular properties, which leads us to consider more advanced techniques.

CONCLUSION

Overall, I found that the MLP regressor had the overall best performance out of the two models tested. It has a coefficient of determination value of over 0.9, and an MAE < 2, meeting our original performance targets. These results demonstrate the superiority of neural network-based models over ensemble learning methods such as Random Forest Regressor.

For future research, models can be taken from other packages such as TensorFlow, which while not having the same fundamental models that are available in Scikit Learn. The use of simpler models in industry is a concept that should be explored further as a result of the performance discussed above, and the wide array of models available across multiple packages and languages, that may surprise the industry with their efficiency and effectiveness. Furthermore, OpenAI is one of the largest developers of AI currently, so tools developed by them could be very real competitors when evaluating the use of AI in Chemistry. By comparing these advanced models with foundational methods from Scikit-Learn or custom-built models, researchers can evaluate both the accuracy and scalability of AI in real-world experimental R&D, potentially driving significant advancements in the field.

REFERENCES

1. Dorrier J. AI Models Scaled Up 10,000x Are Possible by 2030, Report Says [Internet]. Singularity Hub. 2024. Available from: https://singularityhub.com/2024/08/29/ai-models-scaled-up-10000x-are-possible-by-2030-report-says/ (Accessed on 2024-5-15)
2. Hancock JT, Khoshgoftaar TM, Johnson JM. Evaluating classifier performance with highly imbalanced Big Data. *J Big Data*. 2023; 10 (1): 1-31. https://doi.org/10.1186/s40537-023-00724-5.
3. Landrum G. RDKit [Internet]. [cited 2024 Nov 7]. Available from: https://rdkit.org/.
4. Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *J Cheminform*. 2018; 10 (1): 4. https://doi.org/10.1186/s13321-018-0258-y.
5. Hewitt M, Cronin MTD, Enoch SJ, Madden JC, Roberts DW, Dearden JC. In silico prediction of aqueous solubility: the solubility challenge. *J Chem Inf Model*. 2009; 49 (11): 2572-2587. https://doi.org/10.1021/ci900286s.
6. Mazouin B, Schöpfer AA, von Lilienfeld OA. Selected machine learning of HOMO-LUMO gaps with improved data-efficiency. *Mater Adv*. 2022; 3 (22): 8306-8316. https://doi.org/10.1039/D2MA00742H.
7. Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. *Acc Chem Res*. 2021; 54 (2): 263-270. https://doi.org/10.1021/acs.accounts.0c00699.
8. Zhang Y, Xu W, Liu G, Zhang Z, Zhu J, Li M. Bandgap prediction of two-dimensional materials using machine learning. *PLoS One*. 2021; 16 (8): e0255637. https://doi.org/10.1371/journal.pone.0255637.
9. Luukkonen S, Belloni L, Borgis D, Levesque M. Predicting hydration free energies of the FreeSolv database of drug-like molecules with molecular density functional theory. *J Chem Inf Model*. 2020; 60 (7): 3558-3565. https://doi.org/10.1021/acs.jcim.0c00526.
10. API Reference. scikit-learn. Available from: https://scikit-learn.org/stable/api/index.html (Accessed on 2024-7-15)
11. RDKit blog - What are the VSA Descriptors?. 2023. Available from: https://greglandrum.github.io/rdkit-blog/posts/2023-04-17-what-are-the-vsa-descriptors.html (Accessed on 2024-7-15)
12. Zhang Z-X, Cao Y-L, Chen C, Wen L-Y, Ma Y-D, Wang B-Z, et al. Machine learning-assisted quantitative prediction of thermal decomposition temperatures of energetic materials and their thermal stability analysis. Energetic Materials Frontiers. 2023; Available from: http://dx.doi.org/10.1016/j.enmf.2023.09.004 (Accessed on 2024-7-15).
13. Gou Y, Shen L, Cui S, Huang M, Wu Y, Li P, et al. Machine learning based models for high-throughput classification of human pregnane X receptor activators. *Env Sci Adv*. 2023; 2 (2): 304-312. https://doi.org/10.1039/D2VA00182A.
14. 15.7: Eigenvalues and Eigenvectors. Chemistry LibreTexts. Libretexts; 2018. Available from: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Mathematical_Methods_in_Chemistry_(Levitus)/15%3A_Matrices/15.07%3A_Eigenvalues_and_Eigenvectors (Accessed on 2024-7-23)
15. Riquelme M, Lara A, Mobley DL, Vestraelen T, Matamala AR, Vöhringer-Martinez E. Hydration free energies of organic molecules in the FreeSolv database calculated with polarized atom in molecules atomic charges and the GAFF force field [Internet]. ChemRxiv. 2018. Available from: https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/60c73d97469df448fcf42762/original/hydration-free-energies-of-organic-molecules-in-the-free-solv-database-calculated-with-polarized-atom-in-molecules-atomic-charges-and-the-gaff-force-field.pdf (Accessed on 2024-7-24). https://doi.org/10.26434/chemrxiv.6015611.
16. Kiely E, Zwane R, Fox R, Reilly AM, Guerin S. Density functional theory predictions of the mechanical properties of crystalline materials. *Cryst Eng Comm*. 2021; 23 (34): 5697-5710. https://doi.org/10.1039/D1CE00453K.
17. Just How Flexible are Neural Networks in Practice? Available from: https://ar5iv.labs.arxiv.org/html/2406.11463 (Accessed on 2024-8-1)
18. Predicting solvation free energies with an implicit solvent machine learning potential. Available from: https://arxiv.org/html/2406.00183v2 (Accessed on 2024-9-15)
19. Shimakawa H, Kumada A, Sato M. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. *Npj Comput Mater*. 2024; 10 (1): 1-14. https://doi.org/10.1038/s41524-023-01194-2.
20. Towards computational design of molecules with desired properties. Available from: https://www.alcf.anl.gov/news/towards-computational-design-molecules-desired-properties (Accessed on 2024-9-15)