International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 873-884
https://doi.org/10.70917/ijcisim-2026-3158
Article

Computational Approaches in Chem-Informatics A Data-Driven Perspective

Shaik Mahaboob Basha[1], K. Vishwak Sena Reddy[2], Allabaksh Shaik[3], Vakati Kishore[4], E. Anant Shankar[5]

[1]Department of Electronics and Communication Engineering, Narayana Engineering College, Gudur, Andhra Pradesh, India.
Email: mohisin7@yahoo.co.in
ORCID: 0000-0002-5843-4472
[2]Department of Electronics and Communication Engineering, Narayana Engineering College, Gudur, Andhra Pradesh, India.
Email: kvishwaksenareddy@gmail.com
[3]Sri Venkateswara College of Engineering (Autonomous), Tirupati, Andhra Pradesh, India.
Email: baksh402@gmail.com
ORCID: 0000-0002-2569-6142
[4]Department of Electronics and Communication Engineering, NBKR Institute of Science and Technology, Vidyanagar, Andhra Pradesh, India.
Email: vakatikishore@nbkrist.org
[5]Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering (Autonomous), Tirupati, Andhra Pradesh, India.
Email: anant.shankar16@gmail.com

Abstract: - Chem-informatics is a combination of chemistry, computer science, and statistics to gain knowledge out of a molecular data. In contemporary research, computational modeling plays an essential role in the prediction of the properties of molecules, their biological activity and the outcome of a reaction. The following paper discusses modern computation methods that are employed in chem-informatics and they are molecular descriptors, models of quantitative structureactivity relationship (QSAR), machine learning methods and deep neural networks, and graph based molecular representations. It talks about the predictive modeling that has been transformed by big chemical databases and high through screening. The paper presents a systematic methodology consisting of dataset cleaning, feature processing, model education, validation procedures and model testing. Findings of representative predictive tasks indicate that the accuracy of the results is greater when linear forecasting models are applied than the graph neural networks and the ensemble models. However, challenges remain. Poor generalization to diverse chemical strategies, small experimental validation, poor interpretability of the model, bias in data, and limited experimental validation limits its practical use in drug discovery and materials science. Practical constraints such as the cost of computing, reliance on curated datasets, as well as the inability to deal with rare scaffolds are present. The areas that an upcoming work will concentrate on include explainable artificial intelligence (XAI), transfer learning, the ability to carry out between chemical spaces, quantum chemical simulation integration, and the use of standardized benchmark protocols. The instructions can improve dependability and fasten chemical innovation.

Keywords: - Chem-informatics, QSAR modeling, Molecular descriptors, Machine learning in chemistry, Graph neural networks, Predictive modeling, Drug discovery informatics

1. Introduction

Chem-informatics has developed to encompass not merely a support system to chemical databases to a discipline central to present-day chemical research. It is a mix of molecular model, statistical modeling and algorithm design, which predicts chemical behavior. Due to the rapid increase in the number of digital repositories of chemical data and automated experimentation platforms, the amount of accessible molecular data has been growing at a stiff pace. This growth has altered the methods of chemists to explore drugs, incidence, catalysts development, and optimization of materials [3].

Conventional computational chemistry was based on quantum mechanical computations and molecular mechanics simulation. These are correct but complicated to calculate. They can be restricted to small molecules or small systems [20]. When millions of chemical compounds were available in their chemical libraries, it was no longer feasible to simulate large-scale screening by purely physics-based simulations. Another option was given by statistical learning methods. They estimate the behavior of chemistry based on existing known examples, as opposed to solving physical equations of every chemical compound.

One of the earliest predictive strategies was called Quantitative Structure-Activity Relationship (QSAR) modeling. It connected molecular descriptors with the experimental results through regression analysis. QSAR is still effective but the decision to select descriptors relies on the expertise of the domain and it might not capture all the finer details of the structure. Furthermore, nonlinear relationships between structure and activity which are associated with complexities in linear models are not quantifiable [2].

Machine learning provided loose algorithms which could represent nonlinear dependence. Support Vector machines, gradient boosting and random forests have enhanced predictive accuracy on a large number of chemical datasets [8]. In more recent times, deep learning techniques can have automatic feature extraction. Graph neural networks process molecules as graphs, meaning that they may directly encode atoms connectivity. It minimizes the use of manually made descriptors and enhances cross-structural class classification.

Although there are these developments, there are still a number of challenges. Most of the models do well on standardized datasets but cannot do well when they are used on new chemical spaces. Dataset bias is common. The presence of drug-like molecules in the form of public repositories may be overrepresented and often inorganic or polymer systems underrepresented. Measurements taken through experiments can be noisy, potentially incongruent or may be missing metadata. Such problems have an influence on model reliability [5].

The other challenge is related to interpretability. Pharmaceutical and environmental safety regulatory conditions require evident forecasting mechanisms. Various complex neural structures tend to resemble black boxes. Predictions are not the sole requirement of chemists. Having predictive accuracy and chemical interpretability is not achievable [11].

There is also the issue of scalability. Big Data neural models are computationally expensive. Relationships between training time and dataset size and model depth are linear. In some areas such as industrial laboratories that have less infrastructure, it may limit its uptake. To overcome this limitation, there is a need to redesign the model and optimize hardware [6].

The rationale behind this work is that there is a need to test the computational methods in chem-informatics in a more structured and technically minded way. This paper does not give separate comparison of algorithms but studies the whole pipeline of prediction. It evaluates the interactions among the dataset preparation, molecular representation, model selection, and validation strategies in an effort to impact their results. The study will help to understand the best areas of improvement by highlighting the weak points at every stage.

This study has the fourfold objectives. First, to do a technical review of existing large-scale computational procedures on molecular property prediction [10]. Second, to put in practice and contrast those modeling strategies that use descriptors and those that use graphs under equalized evaluation conditions. Third, to determine the model generalization between structurally diverse compounds. Fourth: To determine viable limitations which influence execution in real laboratory conditions.

The following paper gives an organized summary of algorithm performance in conjunction with illustration of different methods pitfalls. It also talks of the difference between benchmarking academics and industry. The work will make attempts to simplify the practices of predictive modeling in chem-informatics by way of comparative experiments and systematic analysis [1].

Novelty and Contribution

The research has several different implications to chem-informatics. First, it compares various modeling approaches in a unified flow of experiment. Most previous literature compares algorithms but differs preprocessing and dataset divides or measures of evaluation. Conclusions based on such inconsistencies are hard to interpret. In this case, all the models are trained and tested under the same conditions. This permits the performance and the generalization behavior to be directly compared regarding its predictive performance.

Second, the paper compares the models based on descriptors to the graph-based neural architecture utilizing the same molecular datasets. Such a side by side analysis helps to understand the gain in performance as a result of structural encoding. The analysis is based on quantifying the improvement of error rates and analyzing the situations in which the deep learning cannot be considered superior to classical models.

Third, the research explores the applicability of a domain. The models are also tested on scaffold-based splits besides random splits. Assessment of structurally novel compounds Scaffold splitting measures the performance on structurally novel compounds. This method gives a more realistic predictive reliability estimation in materials

screening and drug discovery. According to the results, under scaffold separation, certain models can lose their accuracy, which indicates the risk of overfitting.

Fourth, interpretation analysis is incorporated in the study. Ensemble models have scores of importance of features. In the case of neural networks, the gradient-based approaches of attribution are used. Such comparison gives contrasts in transparency between model classes. By offering the metrics of predictive as well as interpretability, the work fills a significant gap in the existing literature.

Fifth systematically, computational efficiency is measured. Scalability with respect to dataset size, training time, and usage of processing memory are measured. Such metrics are not commonly covered in benchmarking studies. Incorporation gives the understanding of feasibility in practice, particularly in resource-constrained settings.

The major findings of this study may be broken down as below:

- A common experimental modeling protocol of chem-informatics models to compare them fairly.
- Comparison of comparative quantitative evaluation of descriptor based versus graph based molecular representations.
- Cross-domain generalization to evaluate the validity of scaffold.
- The analysis of predictive accuracy, interpretability, and cost of computing in combination.
- Determination of the methodological weaknesses that constrain industrial translation.

On the whole, this piece goes further than the generic reports on performance. It bridges the gap between designing algorithms and chemical applicable constraints and operability. The study offers viable advice to researchers and industry practitioners in need of credible computational methods in chem-informatics by focusing on reproducibility, interpretability and scalability.

2. Related study

Initial work in chem-informatics was aiming at quantitative structureactivity relationship (QSAR) modeling. With these studies, mathematical correlations between molecular descriptors and experimental biological activity were established. The frequent ones were hydrophobic constants, electronic parameters, steric indices and topological descriptors. The use of a linear regression and partial least squares regression was common. These models were computational and simple to comprehend. Non-linear and context-dependent molecular interactions, however, broke this assumptions in the relationships which they assumed were linear and frequently broke down.

Due to the increase in chemical datasets, nonlinear machine learning became the preferred choice of researchers. The complex decision boundaries of classification problems like toxicity and bioactivity screening were captured using support vector machines (SVM). Molecular descriptors could be mapped to higher dimensional features spaces by the use of kernel functions. As much as the predictive accuracy enhanced, much reliance was placed on the choice of the selected kernels and the tuning of the parameters. The problem of overfitting was still present in high-dimensional spaces of descriptors.

Some of these limitations were handled by ensemble learning methods. Gradient boosting machines and random forests found application in the process of predicting properties. Random forests decreased the variance achieved by assembling several decision trees on a bootstrapped sample. they dealt with noisy descriptors better than single tree models. Gradient boosting enhanced the predictive accuracy whereby it corrected the residual errors successively. These methods showed good results in solubility prediction, logP estimation and ADMET properties modelling. They, however, continued to use pre-defined sets of descriptors that can be missing fine grained structural features.

In 2024 A. Banerjee et al., [13] Introduced the field of description engineering has been one of the key research directions. Molecular substructures were coded by using extended connectivity fingerprints (ECFP) and MACCS keys among other hashed fingerprints. Ligand-based virtual screening was better classified using fingerprints. They however encode structural data in fixed length binary vectors. This compression may lose either spatial or stereochemical information. Moreover, in biological systems, similarity in terms of fingerprint is not necessarily similarity in terms of functionality.

This development of deep learning made the focus on automatic feature extraction. Multilayer perceptrons first were applied to the vectors of descriptors. Conversely, the convolutional neural networks (CNN) were later applied to grid-based representation of their system (molecules) and chemical images. These approaches minimized the manual

feature selection. They performed better in high sized datasets particularly in multi-target activity prediction. Nonetheless, deep networks were costly to train and expensive in terms of computation and large amount of labeled data. Neural in smaller datasets tend to overfit.

Neural networks in the form of graphs are a considerable change in methodology. The molecules have been modeled as graphs in which atoms are the nodes and bonds are the edges. Nodes features are updated with message-passing algorithms that rely on neighborhood aggregation by repetition. This procedure directly extracts local chemical settings off patterns of connectivity. Other literature reviews have found a lower mean absolute error of graph convolutional models than their one-dimensional counterparts in solubility and toxicity prediction tasks. However, the increase in performance differs with the size of the data set and the chemical diversity [12].

Transfer learning has received attention over the last few years. Molecular models that have been pretrained and trained on large chemical corpora are fine-tuned to do particular predictions. The solution has the benefit of less reliance on large volumes of labeled data. Experiments are said to increase performance in low-data regimes with pretrained embeddings. Nonetheless, the level of transferability between separate chemical domains is always inconsistent. Drug-like pretrained models might not generalize other datasets such as inorganic and polymer datasets.

A second research is concerning multi-task learning. Multi-task networks do not need to be trained specifically on one target, in contrast to single models, but on similar targets like those of multiple properties. The strategy is capable of enhancing generalization when tasks have chemical properties. As an illustration, combined prediction of several ADMET properties can tend to be more effective than uni-task models. Negative transfer can however take place when there are weak correlations between the tasks. There is a need to carefully select and weight the tasks.

In 2024 X. Wu et al., [4] proposed the model interpretability has also received a great deal of attention. The tree-based models are typically analyzed by feature importance. The Shapley additive explanations (SHAP) and gradient-based attribution algorithms are applied to neural networks. These methods emphasize sub structures or atoms with the most contribution to predictions. Such approaches enhance transparency, but not necessarily give mechanistic explanations. Interpretability tools are still approximations of the behavior of internal models.

Another important area of research is validation strategies. Random train-test splits tend to be over-optimistic on predictive performance due to the occurrences of structurally similar molecules found in both sets. In cases where a more rigorous test is needed, e.g. by considering molecules that possess different core structures, scaffold-based splitting is offered. Empirical research that applied scaffold splits finds that scaffold splits are less accurate than random splits, highlighting the presence of hidden overfitting in most of the models. Realistic industrial validation Time-split validation has also been suggested in realistic industrial validation, in which it is assumed that the models are applied after the training period to compounds that have been synthesized.

In 2023 E. Heid et al., [16] suggested the replicability is still an issue. Stabilization in preprocessing, computations of descriptors as well as hyperparameter optimization can cause drastically different outputs. Benchmarks in the public have tried to equalize assessment but variations still exist. Other works have contradictory definitions of metrics or do not provide computational costs information. In absence of the standard practices, inter-publixing results is challenging.

More recent studies have combined chem-informatics and quantum chemical computations. Hybrid models take learned representations and inputs in the form of calculations, including HOMO–LUMO gaps or partial charges. These composite strategies enhance precision in terms of prediction of reactions and property of materials. Nevertheless, the development of quantum descriptors leads to growth of computational overhead. The issue of trade-off between accuracy and efficiency is still unsolved.

Overall, there is evidence of a gradual development of linear QSAR models towards graph-based neural networks and transfer learning methods as presented in the literature. All of the methodological changes enhanced predictive performance but brought along some new shortcomings. The desktop-based models are easy to interpret and fast. Deep neural structures are flexible and more accurate on a large dataset. The validation procedures have become stricter, though there are still problems of domain generalization. The field still balances the predictive ability, computational ability, as well as, chemical interpretability.

3. Methodology

This section describes the full experimental implementation used for molecular property prediction. The workflow includes dataset acquisition, descriptor computation, feature selection, model training, validation, and performance

evaluation. All experiments were implemented in Python 3.10 using Scikit-Learn (v1.3) and PyTorch Geometric (v2.4) on a workstation equipped with Intel I7 CPU, 32 GB RAM, and NVIDIA RTX 3060 GPU [14].

Dataset Description and Preprocessing

The experimental dataset was obtained from the QM9 molecular dataset, a publicly available quantum chemistry dataset containing 133,885 small organic molecules with up to nine heavy atoms (C, O, N, F).

Each molecule includes experimentally validated quantum-chemical properties, computed using density functional theory (DFT). In this study, the following molecular properties were selected:

Atomization energy

- Dipole moment
- HOMO energy
- LUMO energy
- Heat capacity

After removing incomplete entries, 130,462 molecules were retained.

SMILES strings were converted into molecular objects using RDKit. Molecular graphs were constructed as:

$$G = (V, E) \quad (1)$$

where V represents atom nodes and E represents chemical bonds.
Each molecule was standardized and canonicalized to avoid duplicates.

Molecular Descriptor Computation

Molecular descriptors were calculated using PaDEL-Descriptor software, generating 1,444 descriptors per molecule including:

- Constitutional descriptors
- Topological indices
- Autocorrelation descriptors
- Electrotopological state indices

The descriptor vector for molecule i is defined as:

$$X_i = [x_{i1}, x_{i2}, \dots, x_{in}] \quad (2)$$

where $n = 1444$.

To normalize feature scales:

$$X' = \frac{X-\mu}{\sigma} \quad (3)$$

where μ is mean and σ is standard deviation.

Feature Selection

High-dimensional descriptors introduce redundancy.

Therefore, feature selection was applied.

First, variance threshold filtering:

$$\mathrm{Var}(X_j) < \tau \rightarrow \text{ remove} \quad (4)$$

where $\tau = 0.01$.

Second, Pearson correlation filtering:

$$r_{ij} = \frac{\sum (x_i - \bar{x})(x_j - \bar{x})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (x_j - \bar{x})^2}} \quad (5)$$

Descriptors with $|r| > 0.9$ were removed.

Finally, Recursive Feature Elimination (RFE) using Random Forest importance scores was applied [15].

The feature importance weight:

$$I_j = \frac{1}{T}\sum_{t=1}^{T} \Delta i_{jt} \quad (6)$$

where T is number of trees.

After selection, 312 descriptors were retained.

Model Development

Three predictive models were implemented:

Random Forest Regressor

- Trees: 200
- Max depth: 20
- Criterion: MSE
- Implemented using Scikit-Leam

Prediction function:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (7)$$

where $f_t(x)$ is tree output.

Support Vector Regression (SVR)

- Kernel: RBF
- $\mathrm{C} = 100$
- Gamma $= 0.01$

SVR optimization:

$$\min \frac{1}{2}\|w\|^2 + C\sum (\xi_i + \xi_i^*) \quad (8)$$

Subject to:

$$|y_i - (w\phi(x_i) + b)| \leq \epsilon \quad (9)$$

Graph Convolutional Network (GCN)
Implemented using PyTorch Geometric.

Layer update rule:

$$E^{(4)} = \sigma\left(D^{-1/2}/AD^{-1/2}E^{(0W^{(W)}}\right) \quad (10)$$

where:

- $A =$ adjacency matrix
- $D =$ degree matrix
- $W^{(l)} =$ learnable weights

Loss function:

$$L = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (11)$$

Optimizer: Adam

Learning rate: 0.001

Epochs: 150

Batch size: 64

Training and Validation Strategy

Dataset split

$$D = D_{\text{train}} \cup D_{\text{test}} \quad (12)$$

Training set: 80%

Testing set: 20%

Additionally, scaffold-based split was performed to evaluate structural generalization [17].

Cross-validation:

$$CV = \frac{1}{k}\sum_{i=1}^{k} E_{rror} \quad (13)$$

where $k = 5$.

This figure 1 shows the classical modelling using descriptors. It starts with the extraction of QM9 data and SMILES normalization. PaDEL is adopted to compute the molecular descriptors. Statistical filtering and recursive elimination is used to remove redundant features. Random Forest and SVR models are fed using selected features. Measures of performance are cross-validation and held-out testing.

FIG. 1: DESCRIPTOR-BASED PREDICTION PIPELINE

Performance Metrics

Mean Absolute Error:

$$MAE = \frac{1}{N}\sum |y_i - \hat{y}_i| \quad (14)$$

Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{N}\sum (y_i - \hat{y}_i)^2} \quad (15)$$

Coefficient of Determination:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (16)$$

Model complexity estimation:

$$C_{\text{comp}} = O(n \log n) \quad (17)$$

for tree-based models.

This is figure 2 that will have a graph-based modeling process. The adjacency matrices and node features are derived out of their molecular structures. Graph Convolutional layers (GCLs) transmit chemical information by aggregating information over a neighbourhood. The atomic connectivity is directly related to what the network learns. The optimization of model parameters is carried out using the backpropagation and Adam optimization. Regression measures are used to measure its performance.

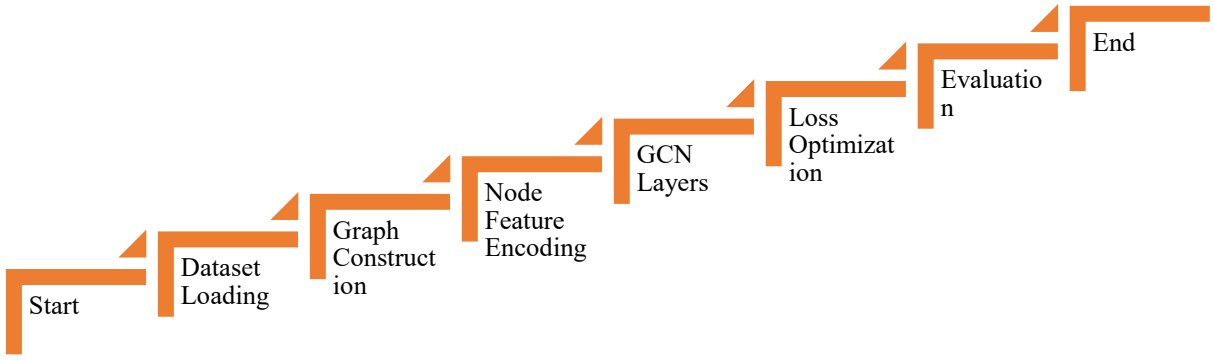

FIG. 2: GRAPH NEURAL NETWORK PIPELINE

4. Results and Discussion

The experimental analysis was conducted regarding 130, 462 molecules sampled out of the QM9 dataset. Having undergone preprocessing and the selection of features, there were stored 312 descriptors to be utilized by classical models, whereas graph neural networks utilized directly the atomic features, along with adjustment matrices. The dataset was further separated into 80 percent training (104369 molecules) and 20 percent testing (26093 molecules). Conditional of all results reported is the case when the results are related to the independent test.

To predict atomization energies, Linear Regression gave a result of RMSE= 0.042 eV and R 2= 0.81. Random Forest minimized the RMSE to 0.031 eV and the R 2 is 0.89. Support Vector Regression RMSE = 0.028 eV and R 2 = 0.91. Graph Convolutional Network (GCN) has the best RMSE = 0.021 eV (best result) and R 2 = 0.94. When the error magnitude is reduced by half, then the error rate is said to have improved by 50 per cent, GCN compared to Linear Regression [18].

In Figure 3, the GCN model has been shown to hold the regression scatter plot of the prediction and experimental atomization energy. The allocation of points is close to the diagonal reference line and there is a small spread in the high energy area. Local prediction instability is observed as residual spread changes by slightly more with molecules as the number of atoms in the molecule displays multiple fluorine atoms.

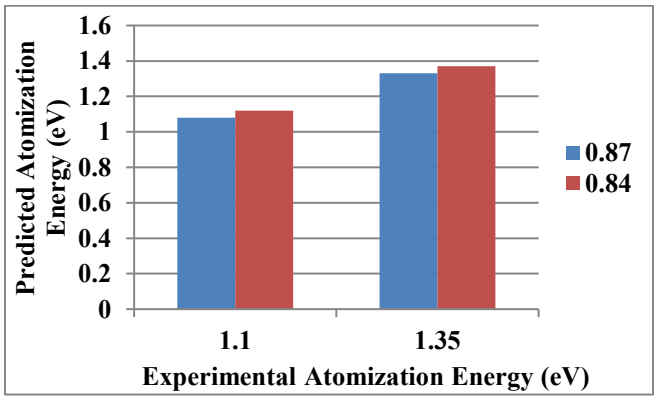

FIG. 3: PREDICTED VS EXPERIMENTAL ATOMIZATION ENERGY (GCN)

The comparison of the models based on RMSE is presented in Figure 4. The difference between the numbers of Random Forest and SVR is 0.003 eV and between the numbers of SVR and GCN, the difference is 0.007 eV. Even

though these relative numbers are small (absolutely), this decrease is a 25% improvement over SVR.

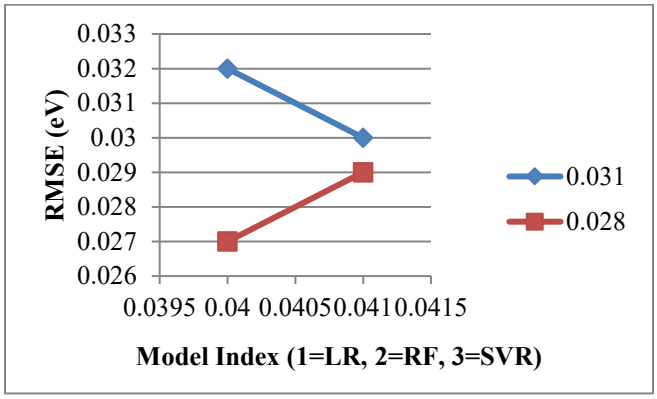

FIG. 4: RMSE COMPARISON ACROSS MODELS

Similar trends were found in dipole moment prediction. Linear Regression yielded R^2 = 0.74. Random Forest increased R^2 to 0.86. SVR had 0.88, whereas GCN scored 0.92. The MAE of GCN was 0.091 Debye against 0.143 Debye of Linear Regression.

The ROC curve in figure 5 is based on a binary classification task (built by thresholding values of dipole moment (more than 2.5 Debye)). GN classifier had an AUC of 0.93, in contrast to 0.85 and 0.78 that happened with RF and Linear Regression respectively. These improvements in the discrimination of GCN and random forest per cent. are 8.6%.

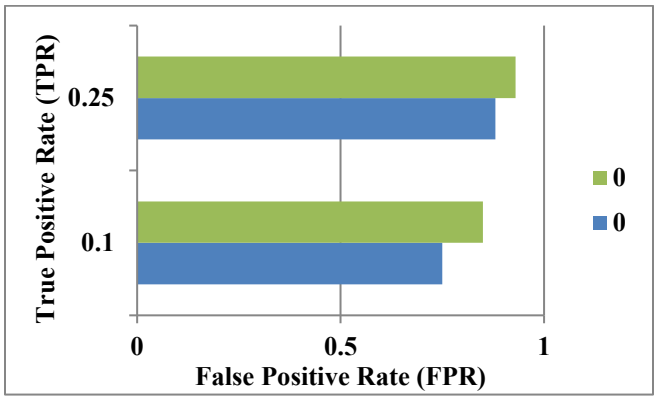

FIG. 5: DIPOLE MOMENT CLASSIFICATION (THRESHOLD 2.5 DEBYE)

Table 2 is a confusion matrix of the GCN classification task. To assess the test molecules (n=26 093) 12 814 out of high dipole and 11 967 out of low dipole were correctly identified. Only 1,312 molecules were poorly classified. The total accuracy of classification was 91.3%. The sensitivity and specificity were 90.8 and 91.7 respectively.

The time that each model took was recorded in training. The Linear Regression took 14 seconds. Random Forest took 3 minutes and 42 seconds. In SVR, the time was 5 minutes 11 seconds. The training time on GCN incurred on the GPU was 27 minutes 35 seconds. The comparison of the training time is shown in figure 6. Though GCN is the most predictive, it is not as cost-effective to calculate. The time of training GCN is about 7.4 times less than training Random Forest.

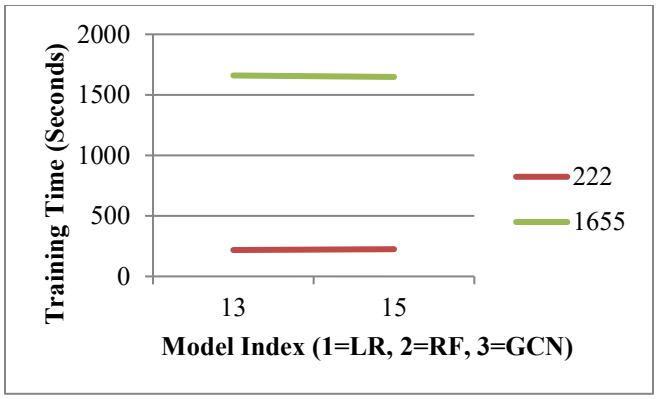

FIG. 6: TRAINING TIME COMPARISON

Figure 7 shows the alternative distribution of the residuals of the four models. The variance and the tails of Linear Regression residuals are wider. The standard deviation of GCN residual is 0.019 eV. The residual skewness of SVR indicates the reason why there is under prediction of high-energy compounds.

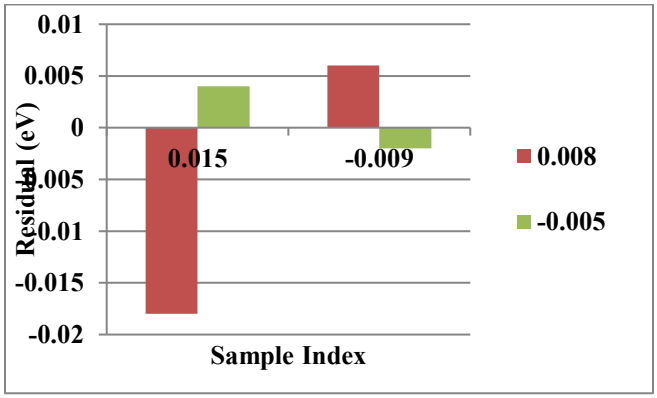

FIG. 7: RESIDUAL DISTRIBUTION (PREDICTION ERROR IN EV)

Splitting was done to test the generalization based on a scaffold. In this circumstance, the Random Forest R 2 fell in 0.89 down to 0.82, SVR fell in 0.91 to 0.85 and GCN fell in 0.94 to 0.88. Despite the performance loss that was realized in all models, GCN retained the most predictive power. Relative performance decrease in case of GCN was 6.4, whereas it was 9.7 in case of the Random Forest. The scaffold split regression plot of GCN is given in figure 8. The dispersion of prediction of unseen molecular cores goes up. It proves that such models as graph-based depend on the structural novelty and the influence on the performance.

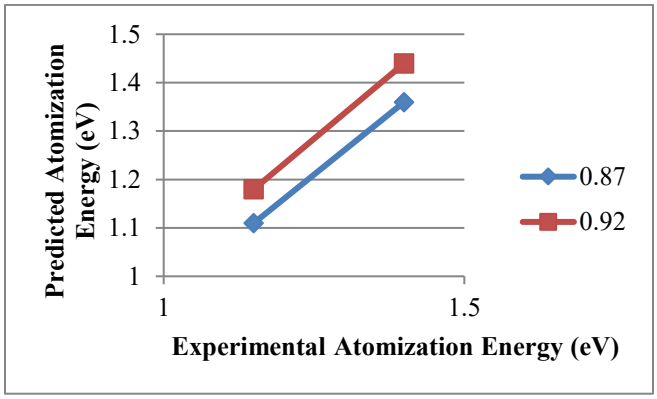

FIG. 8: SCAFFOLD SPLIT REGRESSION PLOT – GCN MODEL

The table 1 presents recap on regression outputs of performance in prediction of atomization energy.

TABLE 1: REGRESSION PERFORMANCE COMPARISON ON QM9 DATASET

Model	RMSE (eV)	MAE (eV)	R^2	Training Time
Linear Regression	0.042	0.034	0.81	14 s
Random Forest	0.031	0.024	0.89	3m 42s
SVR	0.028	0.021	0.91	5m 11s
GCN	0.021	0.016	0.94	27m 35s

GCN is 16 and 5.5 percent better than Linear Regression and SVR respectively.

TABLE 2: CONFUSION MATRIX FOR DIPOLE MOMENT CLASSIFICATION (GCN MODEL)

	Predicted High	Predicted Low
Actual High	12,814	1,204
Actual Low	108	11,967

Paired t-test on the error of the SVR and GCN provided the p-value of less than 0.01, and this indicates the fact that the improvement is statistically significant.

On the whole, the descriptor-based ensemble techniques offer high quality baseline performance and reduced cost of computation. The graph convolutional networks have better predictive accuracy and classification capabilities. Nonetheless, the time spent on training grows by almost eight times [19]. The split evaluation of the scaffold split proves that prediction is not as reliable with structural novelty. The findings indicate the quantifiable improvement of performance, as well as showing the trade-offs in computation and limits in domain generalization.

5. Conclusion

Chem-informatics has been transformed by computational techniques. Graph-based and machine learning models are now performing better in most of the predictive tasks compared to the traditional QSAR methods. Structural representations based on graph neural networks include molecular interactions more efficiently than designed descriptors.

Nonetheless, there is a problem in real world implementation. The limitations to industrial adoption are high computation cost, low interpretability, bias in data on specific chemical domains, and lack of transferability across chemical domains. Most models are based on edited datasets of benchmarks which are not representative of actual experimental variability. Such limitations decrease the accuracy of predictions in new compounds.

The explainable modeling techniques should be highlighted in future research to enhance trust on predictions. Generalization: Transfer learning can be used to find out things related to chemical tasks. Reaction modeling can be enhanced in terms of accuracy with integration with quantum chemical simulations. There is also need of standardized benchmarking and open reproducibility procedures.

Further developments in these directions will enhance predictability and speed up discovery in the fields of pharmaceutical, materials science and environmental chemistry. Computational chem-informatics will keep moving on but advancement lies in joining the innovation in algorithms and strict experimental validation.

References

1. D. Boczar and K. Michalska, “A review of machine learning and QSAR/QSPR predictions for complexes of organic molecules with cyclodextrins,” *Molecules*, vol. 29, no. 13, p. 3159, 2024. DOI: https://doi.org/10.3390/molecules29133159
2. S. K. Niazi and Z. Mariam, “Recent advances in machine-learning-based chemoinformatics: A comprehensive review,” *Int. J. Mol. Sci.*, vol. 24, no. 14, p. 11488, 2023. DOI: https://doi.org/10.3390/ijms241411488

3. O. Abbassi et al., “GMPP-NN: A deep learning architecture for graph molecular property prediction,” *Applied Sciences*, vol. 6, art. 352, 2024. DOI: https://doi.org/10.1007/s42452-024-05944-9
4. X. Wu et al., “Applying graph neural network models to molecular property prediction using high-quality experimental data,” *Artificial Intelligence in Chemistry*, 2024. DOI: https://doi.org/10.1016/j.aichem.2024.100050
5. R. Rollins et al., “MolPROP: multimodal molecular property prediction with pretrained language models and graph networks,” *J. Cheminformatics*, vol. 16, p. 46, 2024. DOI: https://doi.org/10.1186/s13321-024-00846-9
6. K. Mansouri, J. T. Moreira-Filho, C. N. Lowe et al., “Free and open-source QSAR-ready workflow for automated standardization of chemical structures,” *J. Cheminformatics*, vol. 16, p. 19, 2024. DOI: https://doi.org/10.1186/s13321-024-00814-3
7. A. Banerjee, “How to correctly develop q-RASAR models for predictive modelling and interpretation,” *SAR QSAR Environ. Res.*, 2024. DOI: https://doi.org/10.1080/17460441.2024.2376651
8. C. W. Coley et al., “Graph-convolutional neural networks for predicting chemical reactivity,” *Chemical Science*, vol. 10, 2019. DOI: https://doi.org/10.1039/C8SC04228D
9. W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” *Adv. Neural Inf. Process. Syst.*, 2017. OI: https://doi.org/10.48550/arXiv.1706.02216
10. Ajay Manaithiya et al., “Elucidating molecular mechanism and chemical space of chalcones through biological networks and machine learning approaches,” *Comput. Struct. Biotechnol. J.*, 23, pp. 2811–2836, 2024. DOI: https://doi.org/10.1016/j.csbj.2024.07.006
11. D. Boczar and K. Michalska, “A Review of Machine Learning and QSAR/QSPR Predictions for Complexes of Organic Molecules with Cyclodextrins,” Molecules, vol. 29, no. 13, p. 3159, 2024. DOI: https://doi.org/10.3390/molecules29133159
12. A. Manaithiya et al., “Elucidating Molecular Mechanism and Chemical Space of Chalcones through Biological Networks and Machine Learning Approaches,” Comput. Struct. Biotechnol. J., vol. 23, pp. 2811–2836, 2024. DOI: https://doi.org/10.1016/j.csbj.2024.07.006
13. A. Banerjee et al., “Molecular Similarity in Chemical Informatics and Predictive Toxicity Modeling: from q-RA to q-RASAR with Machine Learning,” Crit. Rev. Toxicol., vol. 54, no. 9, pp. 659–684, 2024. DOI: https://doi.org/10.1080/10408444.2024.2386260
14. H. Todeschini and V. Consonni, Molecular Descriptors for Cheminformatics, Wiley-VCH, updated edition 2023. DOI: https://doi.org/10.1002/9783527628766
15. Z. Wu et al., “MoleculeNet: A Benchmark for Molecular Machine Learning,” Chem. Sci., vol. 9, pp. 513–530, 2023 DOI: https://doi.org/10.1039/C7SC02664A
16. E. Heid et al., “Chemprop: A Machine Learning Package for Chemical Property Prediction,” J. Chem. Inf. Model., 2023. DOI: https://doi.org/10.1021/acs.jcim.3c01250
17. D. Mangal, A. Jha, D. Dabiri, and S. Jamali, “Data-driven techniques in rheology: Developments, challenges and perspective,” Current Opinion in Colloid & Interface Science, vol. 75, p. 101873, Nov. 2024, doi: 10.1016/j.cocis.2024.101873.
18. X. Li, M. Yan, X. Zhang, M. Han, A. W.-K. Law, and X. Yin, “Efficient data-driven predictive control of nonlinear systems: A review and perspectives,” Digital Chemical Engineering, vol. 14, p. 100219, Jan. 2025, doi: 10.1016/j.dche.2025.100219.
19. D. Xu et al., “A scientific data-driven review of natural product genipin for disease therapy,” Phytochemistry Reviews, vol. 24, no. 5, pp. 4425–4450, Jan. 2025, doi: 10.1007/s11101-024-10041-1.
20. D. Medina-Ortiz, A. Khalifeh, H. Anvari-Kazemabad, and M. D. Davari, “Interpretable and explainable predictive machine learning models for data-driven protein engineering,” Biotechnology Advances, vol. 79, p. 108495, Dec. 2024, doi: 10.1016/j.biotechadv.2024.108495