CHEMICAL ENGINEERING TRANSACTIONS
VOL. 117, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Fabrizio Bezzo, Flavio Manenti, Gabriele Pannocchia, Almerinda di Benedetto

ISBN 979-12-81206-17-5; **ISSN** 2283-9216
DOI: 10.3303/CET25117171

VAE-Assisted Data Augmentation for Improved Molecular Prediction with Graph Neural Networks (GNNs) in Low-Data Regimes

Gabriela C. Theis Marchan, Pegah Naghshnejad, Andrew Okafor, José A. Romagnoli*

Department of Chemical Engineering, Louisiana State University, Baton Rouge, LA 70803.
jose@lsu.edu

This study presents a novel approach to enhancing molecular property prediction through variational autoencoder (VAE)-assisted data augmentation in low-data regimes. The methodology combines graph neural networks (GNNs) with VAEs to improve predictive accuracy on molecular datasets from MoleculeNet, specifically ESOL (water solubility) and FreeSolv (hydration-free energy). By generating chemically valid molecules that align with the original dataset's chemical space, the approach enhances model performance, particularly for graph attention networks (GATs). Results show significant improvements in prediction accuracy, with GAT models demonstrating increased R^2 values from 0.879 to 0.918 for FreeSolv and 0.873 to 0.885 for ESOL when trained on augmented datasets. The study validates the effectiveness of VAE-generated molecules through chemical space analysis and property distribution comparisons, offering a promising solution for molecular property prediction in data-limited scenarios.

1. Introduction

The accurate prediction of molecular properties is a longstanding challenge in computational chemistry and materials science. Conventional machine learning (ML) methods often rely on fixed feature representations, such as molecular descriptors or fingerprints, which fail to capture the full structural and relational complexity of molecules. Graph neural networks (GNNs) have revolutionized the field by treating molecules as graphs, where atoms are nodes and bonds are edges (Zhang et al. 2019). Through message-passing frameworks, GNNs, such as graph convolutional Networks (GCNs) (Kipf and Welling 2016) and graph attention networks (GATs) (Velickovic et al. 2017), iteratively aggregate and update node features, enabling the extraction of both local and global molecular properties.

While GNNs outperform traditional ML methods in various applications, including the prediction of ion activity coefficients in ion exchange membranes and virtual screening(Naghshnejad et al. 2024), their performance is often hindered in low-data regimes. Molecular datasets are inherently limited, as experimental measurements and quantum chemical simulations are resource intensive. Many benchmarks, such as MoleculeNet, contain few labeled molecules for several properties, leading to overfitting and reduced generalizability in GNN models. To address data scarcity, researchers have explored approaches like few-shot learning, transfer learning, and the use of synthetic data. Among these, generative models offer a promising solution by directly increasing the volume and diversity of training datasets. Variational autoencoders (VAEs) have gained attention for their ability to encode molecular structures into continuous latent spaces and generate chemically valid samples. Recent advancements, such as integrating VAEs with SELFIES (self-referencing embedded strings) (Ather 2024), ensure the chemical validity of generated molecules, making them suitable for augmenting training datasets in molecular property prediction tasks.(Nnadili et al. 2024)

This study aims to combine VAEs and GNNs to address data limitations in molecular property prediction. By leveraging synthetic data generated by VAEs, this study aims to enhance the generalization ability and

Paper Received: 5 January 2025; Revised: 3 February 2025; Accepted: 29 May 2025
Please cite this article as: Theis Marchan G.C., Naghshnejad P., Okafor A., Romagnoli J.A., 2025, VAE-Assisted Data Augmentation for Improved Molecular Prediction with Graph Neural Networks (GNNs) in Low-Data Regimes , Chemical Engineering Transactions, 117, 1021-1026 DOI:10.3303/CET25117171

robustness of GNN models. This framework improves prediction accuracy and lays the groundwork for future advancements in data augmentation and graph-based learning in molecular informatics.

2. Methodology

This study employs a multi-step approach to predict molecular properties using graph-based neural networks augmented by variational autoencoders (VAEs) for data generation. The methodology comprises three main stages: data preprocessing, model architecture development, and training and evaluation. The workflow illustrated in Figure 1 demonstrates a comprehensive pipeline that integrates data augmentation with molecular property prediction. The process begins with the original dataset, initially split into training and test sets to ensure proper evaluation of the final model. The training data, then, follows two parallel paths, the first being the VAE-based data generation path. In this path, the training data is fed into a variational autoencoder (VAE), where the VAE's encoder compresses the molecular information into a latent space representation. The decoder then reconstructs molecular structures from this latent space, and these generated molecules form a new dataset that supplements the original training data.

The second parallel path focuses on initial model development. Here, the training data is used to pre-train a preliminary graph attention network (GAT). This pre-trained model is then used to predict molecular properties for the new molecules generated by the VAE, ensuring that the augmented dataset includes structural information and predicted properties for the synthetic molecules. This critical step helps maintain property prediction consistency across original and generated molecular data.

After obtaining property predictions for the VAE-generated molecules, the augmented dataset combines these molecules with their predicted properties alongside the original training data. This comprehensive dataset is then used to train the final GAT model for molecular property prediction. The final model's performance is evaluated using the previously set-aside test set, kept separate from the training and augmentation processes to ensure an unbiased assessment of the model's predictive capabilities. The green dashed boundary in Figure 1 encompasses the data augmentation process, highlighting its role as a crucial intermediate between the original data and the final model development. This augmentation strategy addresses common challenges in molecular property prediction, such as limited training data and chemical space coverage, by effectively expanding the training set while maintaining the integrity of the test set for final model evaluation.

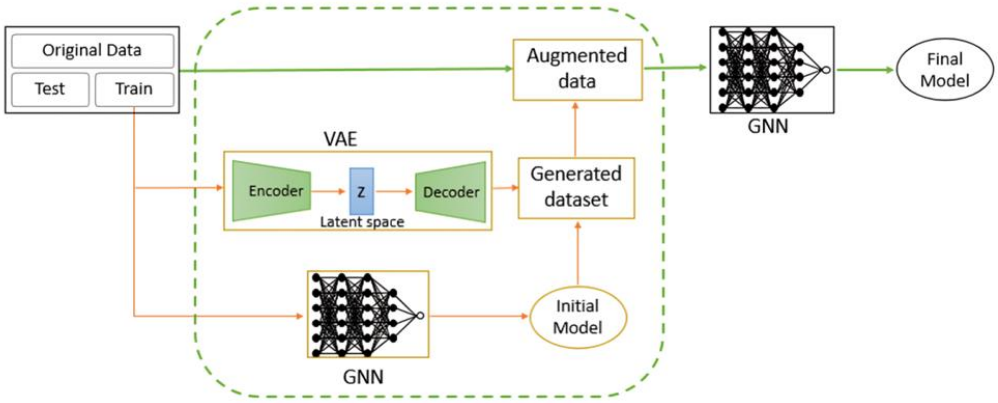

Figure 1: Workflow for VAE-Assisted Data Augmentation and GNN-Based Molecular Property Prediction.

3. Implementation

The datasets used in this study were sourced from MoleculeNet,(Wu et al. 2018) a benchmark platform widely used for molecular machine learning. The analysis was conducted on two datasets: ESOL, which predicts water solubility (1128 molecules), and FreeSolv, which estimates the free energy of solvation (642 molecules). These molecules were represented as SMILES strings for initial input. For molecular representation, both datasets used identical feature engineering when constructing the molecular graphs. This comprises of node features such as atomic numbers, formal charges, degree of atoms, explicit valence, implicit valence, and aromatic atoms (74-dimensional feature vector per atom), along with bond features including types, conjugation, and ring membership (12-dimensional edge feature vector). A functional group analysis revealed distinct chemical spaces between datasets: ESOL contained more complex, drug-like molecules with aromatic rings (988 instances), halogens (835), ketones (582), ethers (340), and alcohols (274), while FreeSolv featured smaller organic molecules with halogens (454), aromatic rings (325), ketones (191), ethers (178), and notably included

carboxylic acids (14) which were absent in ESOL. To ensure robust evaluation, a single test-train split was performed with a ratio of 0.2. This split was conducted only once, ensuring that the test data consisted exclusively of real molecular structures and did not contain any artificially generated molecules, providing a reliable basis for model evaluation. Additionally, the ESOL and FreeSolv datasets contained partial overlap between them datasets, with 196 common molecules in the total dataset and 15 common molecules between their respective test sets. This moderate overlap reflects some commonality in the chemical spaces while maintaining sufficient independence for meaningful evaluation.

The reliability of VAE-generated molecules was validated through multiple checks: chemical validity was enforced using SELFIES representation in the generative model. Additionally, we performed RDKit structure validation and sanity checks. The VAE generated different numbers of molecules for each dataset: 30,000 for ESOL and 20,000 for FreeSolv, which were then filtered for chemical validity and property range consistency. After filtering, approximately 29,502 valid molecules were retained for ESOL and 13,528 for FreeSolv for augmenting the training sets. The generated molecules were added to the original training data in a balanced manner to avoid overwhelming the distribution of the original dataset. The test sets remained unchanged to ensure unbiased evaluation of the models.

The study employed a GAT model for this study, implemented using the GATConv layer. The model comprised two graph convolutional layers and a fully connected feed-forward network (multilayer perceptron) with three layers. The size of each layer (embedding size) in both architectures was set to 128 units, which was optimized to capture molecular graph features effectively. The model architecture and key hyperparameters were optimized, specifically focusing on the learning rate and batch size. After extensive trials, the optimal learning rate was identified as 0.001, and the batch size was determined to be 20. Additionally, a seed value of 42 was used to ensure the reproducibility of the optimization process. This combination of architecture and hyperparameter optimization helped the models to generalize well to unseen data. The optimized models were tested on the reserved test set, which did not contain any data generated through augmentation, ensuring a fair performance comparison. Although the models were trained for a maximum of 1000 epochs, early stopping was implemented to prevent overfitting. This was triggered when no improvement in validation loss was observed over several consecutive epochs. Doing so halted the training process once optimal performance was reached, thus saving computational resources and reducing the risk of overfitting. The training and optimization pipeline yielded models with improved accuracy in predicting molecular properties, as further detailed in the results section.

4. Results and Discussion

To assess the relevance of the molecules generated by the VAE in the context of molecular property prediction, the distribution within the chemical space of the original datasets was analyzed using two dimensionality reduction techniques: principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). The aim of this analysis was to verify whether the generated molecules lie within the chemical space of the training and test datasets, which is crucial for ensuring that the generated molecules are chemically relevant and can meaningfully augment the training set for the downstream predictive model. The VAE-generated molecules must represent the training data distribution for effective data augmentation in low-data regimes. Molecules outside the chemical space of the training data may introduce noise, potentially degrading model performance. Using PCA and t-SNE, the high-dimensional molecular features were projected into a lower-dimensional space, visually comparing the chemical spaces occupied by the training, test, and generated datasets.

Figure 2, presents the chemical space visualizations for both the ESOL and FreeSolv datasets, using PCA and t-SNE to represent the distribution of the training, test, and generated molecules. The PCA plots (Figure 2, top-left and bottom-left) provide a linear approximation of the variance in the datasets. At the same time, the t-SNE plots (Figure 2, top-right and bottom-right) offer a nonlinear perspective, emphasizing the local structure of the data. In the PCA plots, the generated molecules (green-colored) are distributed within the same principal components as the training data (purple-colored) for ESOL and FreeSolv datasets. This indicates that the VAE successfully learned the significant directions of variance in the training data and generated molecules consistent with these directions. The generated molecules do not exhibit extreme deviations, suggesting they lie well within the chemical space defined by the original dataset. Similarly, in the t-SNE plots, which provide a more localized view of the data, the generated molecules cluster within the same regions as the training data, further confirming that the VAE-generated molecules retain the local neighborhood structure of the original chemical space. The test data (red-colored) also occupies overlapping regions with both the training and generated datasets, indicating that the VAE-generated molecules are not outliers but rather valid molecules that exist within the ESOL and FreeSolv chemical spaces.

To further evaluate the utility of the VAE-generated molecules, graph neural networks (GNNs), specifically a GAT, were trained on the training datasets for both the ESOL and FreeSolv properties. This trained model was then used to predict molecular properties for the generated dataset, and the distributions of the predicted properties were compared to those of the training and test datasets. Figure 3 illustrates the density plots of these property distributions for ESOL and FreeSolv, as predicted by the GAT model.

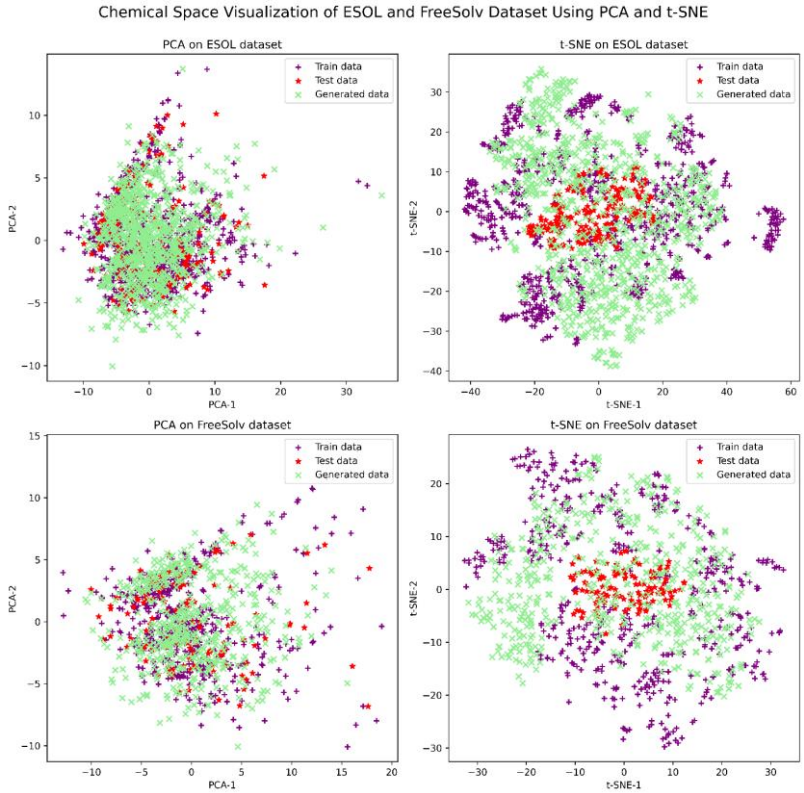

Figure 2: Chemical Space Visualization of Training, Test, and VAE-Generated Molecules for ESOL and FreeSolv Datasets Using PCA and t-SNE.

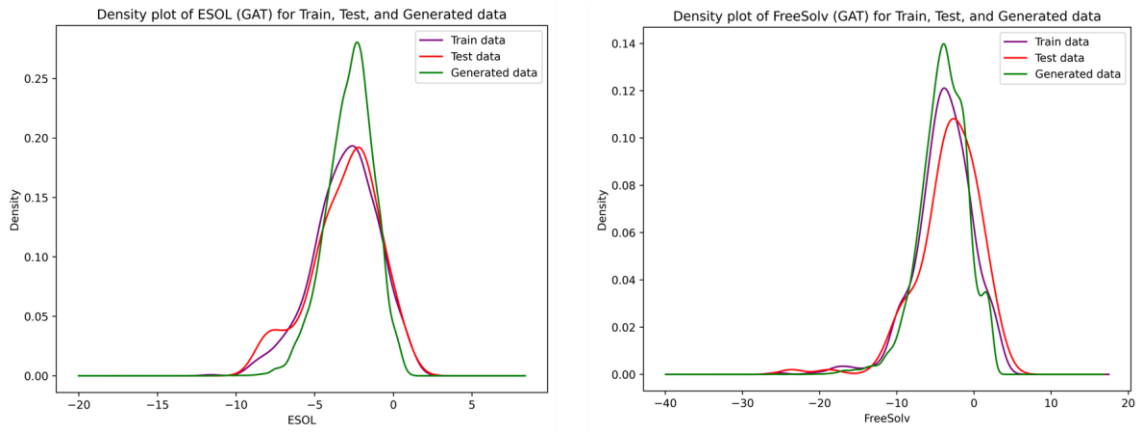

Figure 3: Density plots for ESOL and FreeSolv.

The results indicate that the distributions of the predicted properties for the generated molecules closely align with those of the test and training datasets. In particular, for the GAT model, the generated data (green curves) follow the general shape and peak location of the training (purple curves) and test data (red curves) distributions, with significant overlaps. This suggests that the VAE-generated molecules possess similar property distributions to the original datasets, affirming that they are chemically relevant and consistent with the original data. The fact

that the generated molecules yield property predictions that closely match the original data indicates that the GAT model, when trained on these augmented datasets, can generalize effectively to new, unseen molecules. To assess the performance of the new model, which was trained on a larger dataset consisting of the VAE-generated dataset and the original training dataset, its predictive accuracy was compared to that of the initial model trained solely on the original training data. The GAT model was evaluated on the same dedicated test set, which was not used during training and contained no augmented data. This strict separation was implemented to avoid the data contamination issues highlighted in prior studies, where improper splitting of augmented datasets has been shown to compromise test set integrity and inflate performance metrics.(Maleki et al. 2022) The model was trained with an optimized number of generated molecules in the augmented data, to minimize overfitting. For ESOL, synthetic data added to the 902-molecule base was 902 molecules, while for FreeSolv, the synthetic data added to the 513-molecule base was 2052 molecules. Test sets (226 for ESOL and 129 for FreeSolv) remained unchanged. The evaluation metrics used were the coefficient of determination (R^2) and mean squared error (MSE), which provide insights into the model's ability to generalize to unseen data

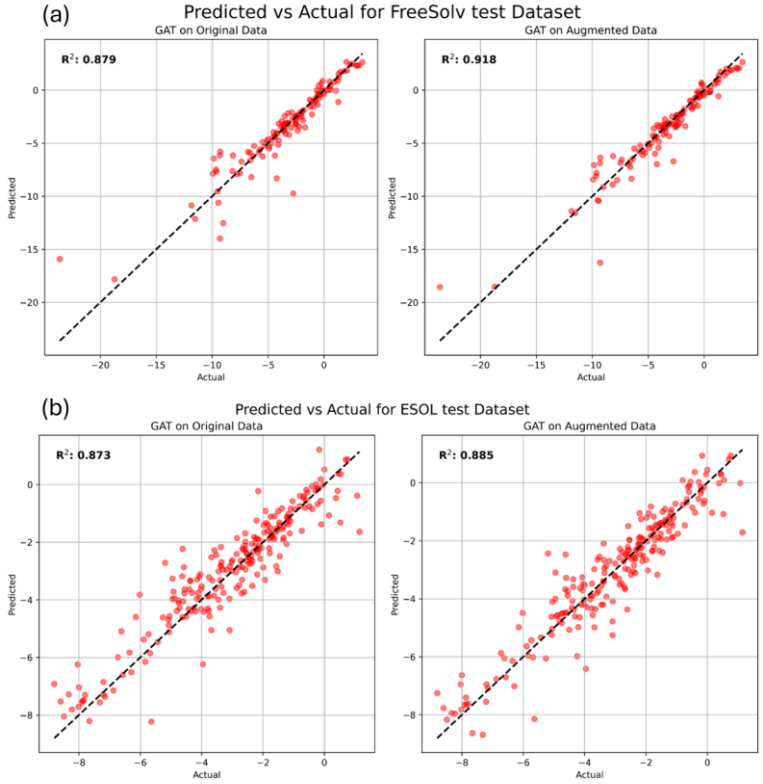

Figure 4: Performance of GAT (a) FreeSolv Dataset and (b) ESOL Dataset.

For the free energy of solvation (FreeSolv) property, the GAT model shows a noticeable improvement in both R^2 and MSE when trained on the augmented dataset, as shown in Table 1 and Figure 4a. Specifically, the R^2 value increases from 0.879 to 0.918, while the MSE decreases significantly from 2.011 to 1.353. This improvement suggests that the VAE-generated molecules help the GAT model generalize better to unseen data, making more accurate predictions for the FreeSolv property. This is likely due to the augmentation providing the GAT model with a more diverse training set, allowing it to capture the underlying molecular representations more effectively.

For the water solubility (ESOL) property, both models exhibit similar trends, as shown in Table 1 and Figure 4b. The GAT model demonstrates an improvement in R^2, increasing from 0.873 to 0.885, and a corresponding reduction in MSE from 0.602 to 0.546. This improvement is consistent with the FreeSolv results, indicating that the GAT model benefits from the additional training examples generated by the VAE, which likely contribute to improved generalization for molecular properties in low-data regimes.

Table 1: Performance Comparison of the GAT Model with and without VAE-Assisted Data Augmentation on FreeSolv and ESOL Properties, Evaluated Using R^2 and MSE.

Property	Without data augmentation with VAE		With data augmentation with VAE	
	R2	MSE	R2	MSE
Free Energy of Solvation (FreeSolv)	0.879	2.011	0.918	1.353
Water Solubility (ESOL)	0.873	0.602	0.885	0.546

5. Conclusion

This study demonstrates the potential of variational autoencoder (VAE)-assisted data augmentation to improve molecular property prediction in low-data regimes. By generating chemically valid molecules that align with the original dataset's chemical space, the performance of the graph attention network (GAT) was enhanced on the ESOL and FreeSolv datasets. Specifically, for the FreeSolv dataset, the GAT model achieved a notable improvement, with R^2 increasing from 0.879 to 0.918 and MSE decreasing from 2.011 to 1.353. Similarly, for the ESOL dataset, the GAT model showed an R^2 increase from 0.873 to 0.885 and a reduction in MSE from 0.602 to 0.546. These results highlight the ability of VAE-generated data to provide additional diversity and improve model generalization in molecular property prediction. The findings emphasize the importance of selecting an appropriate architecture, as the GAT model effectively leveraged the augmented dataset to achieve significant performance gains. This work underscores the value of data augmentation strategies in addressing challenges associated with limited training data in molecular machine learning.

Acknowledgments

This work was supported by the U.S Department of Energy, Office of Science, under the Office of Basic Energy Science Separation Science program under Award No. DE-SC0022304. The authors gratefully acknowledge the computer time allotted by the high-performance computing center (HPC) at LSU and the Louisiana Network initiative.

References

Ather M.M., 2024, The fusion of multilingual semantic search and large language models: A new paradigm for enhanced topic exploration and contextual search. Carleton University.

Kipf T.N., Welling M., 2016, Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907.

Maleki F., Ovens K., Gupta R., Reinhold C., Spatz A., Forghani R., 2022, Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. Radiology: Artificial Intelligence. 5(1):e220028.

Naghshnejad P., Theis Marchan G., Olayiwola T., Kumar R., Romagnoli J., 2024, Graph-based modeling and molecular dynamics for ion activity coefficient prediction in polymeric ion-exchange membranes. Industrial & Engineering Chemistry Research.

Nnadili M., Okafor A.N., Olayiwola T., Akinpelu D., Kumar R., Romagnoli J.A., 2024, Surfactant-specific ai-driven molecular design: Integrating generative models, predictive modeling, and reinforcement learning for tailored surfactant synthesis. Industrial & Engineering Chemistry Research. 63(14):6313-6324.

Velickovic P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y., 2017, Graph attention networks. stat. 1050(20):10-48550.

Wu Z., Ramsundar B., Feinberg E.N., Gomes J., Geniesse C., Pappu A.S., Leswing K., Pande V., 2018. Moleculenet: A benchmark for molecular machine learning. Chemical science. 9(2):513-530.

Zhang S., Tong H., Xu J., Maciejewski R., 2019, Graph convolutional networks: A comprehensive review. Computational Social Networks. 6(1):1-23.