Machine learning for HOMO-LUMO-gap Prediction and Inverse Molecular Design

Mattice Criel

Yarno De Jaeger

Thibo Van Eeckhoorn

Abstract

Accurate predictions of molecular electronic properties such as the HOMO–LUMO gap and inverse design of molecules based on such properties are critical for materials discovery and molecular engineering, but remain costly and time consuming with experimental methods and computationally expensive with quantum-chemical methods. While machine learning has gained increased relevance in such applications as a less expensive alternative, modern models still fail to reliably produce chemically viable results. In this study, we benchmark a LightGBM model against a GCN model for property prediction and a VAE-based model for molecular design. Our results show that the LightGBM and GCN model achieve comparable HOMO-LUMO gap prediction performance, while the VAE model shows good reconstruction capabilities, but limited chemical validity and property control for the generation of new molecules. Only need to add one sentence for **Outlook**Rewrite abstract to only contain 5 sentences: **Context: Problem: Solution: Results: Outlook:**

1 Introduction

The HOMO-LUMO gap is the energy difference between the highest occupied and lowest unoccupied molecular orbitals and is a fundamental descriptor of a molecule's electronic structure. It strongly depends on molecular structure and functional groups and governs key properties such as chemical reactivity, optical absorption, charge transport, and conductivity, determining whether a material behaves as a conductor, semiconductor, or insulator (Dwivedi et al., 2025). The HOMO-LUMO gap is also critical for the performance of organic photovoltaics (OPVs) and organic light-emitting devices (OLED) (Liu et al., 2015), making its accurate prediction and targeted design central to molecular engineering and materials science.

Experimentally, HOMO-LUMO gaps are typically estimated using optical spectroscopy or voltammetry(Costa et al., 2016; Sworakowski, 2018). However, these measurements require synthesized, purified materials making large-scale exploration of hypothetical chemical structures costly and time consuming. Computational quantum chemistry methods such as Hartree-Fock and Density functional theory offer accurate alternatives but are computationally expensive and scale poorly with molecular size and dataset volume, limiting their applicability for large-scale molecular exploration.

Machine learning provides an efficient alternative by enabling rapid prediction of electronic properties(Hasan et al., 2025). Crucially, large-scale quantum-chemical datasets like QM9, contain extensive information about molecular structures and properties of organic molecules, enabling the development and benchmarking of ML models with the data volume required for robust learning. Beyond property prediction, machine learning also enables generative molecular design, allowing in silico exploration and optimization of hypothetical molecules, thereby reducing experimental cost

and accelerating molecular discovery.

In this work, we develop a unified framework for HOMO-LUMO gap prediction and inverse molecular design. We first construct predictive models using a descriptor-based LightGBM approach and a graph convolutional network (GCN) that learns directly from molecular graph topology, achieving comparable accuracy and demonstrating the complementary strengths of traditional ensemble methods and graph-based deep learning. Building on the GCN, we further explore conditional molecular generation by integrating it into a conditional variational autoencoder (cVAE) to generate molecules targeting specified HOMO-LUMO gaps. While the predictive models demonstrate strong performance and generalization, the generative results highlight persistent challenges in achieving reliable property control and valid molecular generation, underscoring the need for improved training strategies, more expressive latent representations, and more effective decoding mechanisms.

2 Related work

In recent years, significant progress has been made in predicting HOMO-LUMO gaps using various machine learning techniques, including deep learning approaches, ensemble methods, and graph-based models. Early descriptor-based regression models achieved mean absolute errors (MAEs) in the range of approximately 0.10-0.25 eV using methods such as Random Forests and LightGBM trained on molecular connectivity and handcrafted descriptor features (Goh et al., 2022). One notable example is the study by Hasan et al. (2025), who used a database of molecular properties computed with Hartree–Fock theory and reduced the MAE of a LightGBM model to 0.1675 eV.

In this context, LightGBM is adopted in the present work due to its proven effectiveness for molecular property prediction using descriptor-based representations. As a gradient boosting decision tree algorithm, LightGBM is well suited to modeling complex, non-linear relationships between molecular descriptors and electronic properties while maintaining strong generalization performance on medium-sized datasets. Moreover, LightGBM offers advantages in terms of computational efficiency, robustness to feature scaling, and interpretability through feature importance analysis. These characteristics make it a strong and well-established baseline, enabling fair comparison with prior descriptor-based approaches while providing competitive accuracy without the added complexity of deep or graph-based models.

Graph neural networks (GNNs) have emerged as a powerful tool to predict molecular properties from graph representations of molecular structures. As there is still a lack of 3D graph-based and 3D grid-based methods for molecular property prediction(Li et al., 2022), we will focus on a 2D graph-based model. One such model that is common for molecular property predicted, but also specifically for HOMO-LUMO prediction, is the graph convolutional network (GCN)(Choi et al., 2022; Therrien et al., 2025; Xia et al., 2023).

Xia et al. (2023) demonstrated that deep-learning models generally are unable to outperform non-deep ones for molecular property prediction. However, Xia et al. (2023) also report MAEs of the predicted HOMO-LUMO gap (using the QM9 dataset) that indicate graph-based deep-learning models as more accurate than traditional models. To more clearly assess whether graph-based deep-learning models outperform traditional models for HOMO–LUMO gap prediction, we chose to compare the performance of a LightGBM model with that of a graph convolutional network (GCN).

In the context of generative models, variational autoencoders (VAEs) have emerged as a stan-

dard framework for de novo molecular design(Walters & Barzilay, 2021). One of the more recent influential works is done by Gómez-Bombarelli et al. (2018) applied VAEs to construct a continuous and differentiable latent space from SMILES representations of molecules, enabling gradient-based optimization of molecular properties using a Gaussian process. This approach has inspired numerous subsequent studies (Blaschke et al., 2018; Kusner et al., 2017; Yoshikai et al., 2024). Notably, Lim et al. (2018) introduced a conditional VAE framework for property-controlled molecular generation, demonstrating the advantages of incorporating molecular property information into the encoding process and manipulating it during decoding. However, many studies have highlighted limitations of SMILES-based representations, suggesting that graph-based encoders and decoders better capture molecular structure and improve the validity of generated molecules. These findings motivate the graph-based generative approach adopted in our work.

3 Methodology

The source code for this project is available on Github(ThiboVE, 2025).

3.1 Dataset

For all models in this study, we used the QM9 dataset(Team, 2024), a widely adopted quantum chemistry dataset comprising 130,831 small organic molecules with associated properties, including the HOMO–LUMO gap. Only molecules with valid SMILES representations were retained, leaving 129,012 molecules, can be seen in Figure 1. SMILES (Simplified Molecular Input Line Entry System) is a text-based notation that encodes chemical structures as ASCII strings, enabling efficient computational processing; for example, benzene is represented as `C1=CC=CC=C1`. While SMILES strings can be interpreted by computers, they are not directly suitable as input for most machine learning models, which require numerical representations. Consequently, preprocessing is necessary to transform molecular structures into informative, fixed-length numerical features suitable for model training.
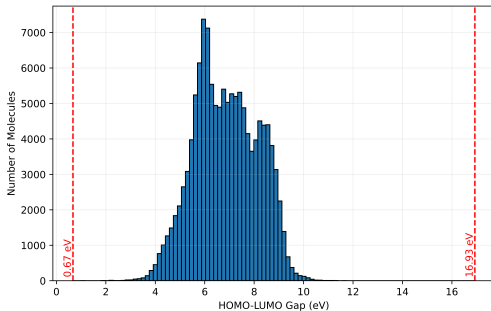
Figure 1: Illustration of the bandgap distribution in the QM9 dataset where only the valid molecules are kept. The vertical dashed lines indicates the minimum and maximum bandgap values in the dataset.

3.2 Molecular Representations and Preprocessing

In this study, we focus on two different models for HOMO-LUMO gap prediction: LightGBM, a classical machine learning model and GCN, a deep learning model, which require different types of molecular representations. Therefore, two types of molecular representations were generated from the SMILES strings.

3.2.1 DESCRIPTOR-BASED FEATURES FOR LIGHTGBM

Using RDKit, a set of 217 molecular descriptors was generated for each molecule from the SMILES representation(RDKit: Open-Source Cheminformatics, 2025). These descriptors capture various aspects of molecular structure and properties, including molecular size, shape, electronic characteristics, surface area, and the presence of functional groups. Together, they provide a fixed-length numerical representation of each molecule suitable for machine learning models such as LightGBM.

3.2.2 GRAPH-BASED REPRESENTATION FOR GCN AND CVAE:

For graph neural networks, molecules are represented as graphs where atoms are nodes and bonds are edges. As the QM9 dataset represents molecules as SMILES, all molecule SMILES were translated to a node and edge matrix using the *Chem* package of RDKit(RDKit Developers, 2025). Node features include atomic number, formal charge, hybridization, aromaticity, and whether the atom is in a ring structure. The edge matrix was defined as an adjacency matrix where the diagonal elements were set to 1, indicating a self-connection, which makes the matrix amenable to convolutions(Deshmukh, 2023). The distance of each bond was included in the edge matrix as 1 over the bond distance. The representation of a molecule as a node and edge matrix allows the GCN to learn structural features directly from molecular graphs without relying on precomputed descriptors.

3.3 Data Splitting

To ensure unbiased evaluation, the dataset was split into training, validation, and test sets. For all models, nested cross-validation was used:

- **Outer loop:** 5-fold cross-validation to estimate generalization performance

- **Inner loop:** 10-fold cross-validation for hyperparameter tuning

The splits were stratified based on the distribution of the HOMO-LUMO gap values into 10 quantile bins to preserve this distribution across the different folds.

3.4 Evaluation metrics

Since both the LightGBM and GCN models are used to predict HOMO-LUMO gaps and are directly compared, the same evaluation metrics are applied to both. Commonly used metrics for HOMO-LUMO gap prediction include the mean absolute error (MAE, Equation (2)), root mean squared error (RMSE, Equation (3)), and the coefficient of determination (R^2, Equation (4)) between the predicted and true HOMO-LUMO gaps.

The cVAE model performance is evaluated using the average character-level reconstruction accuracy ($\overline{CRA}$, Equation (5)), the chemical validity of newly generated SMILES strings using RDKit (Equation (6)), and the agreement between the target HOMO-LUMO gap and the GCN-predicted gap of the generated molecules, quantified using MAE. Each metric is calculated as

3.5 LightGBM

LightGBM is a gradient-boosted decision tree algorithm optimized for speed and high-dimensional data. It grows trees in a leaf-wise rather than level-wise manner, allowing it to capture complex, non-linear relationships more efficiently than many traditional boosting methods. This makes it well-suited for tabular molecular descriptor data, as it is robust to feature scaling, multicollinearity, and missing values.

Hyperparameter optimization for the LightGBM model was carried out using Optuna. The search space included parameters controlling model complexity and regularization: the number of leaves (num_leaves, 16–256), the learning rate (learning_rate, 1×10^{-3}–0.1, log-scaled), feature subsampling (feature_fraction, 0.7–1.0), sample bagging (bagging_fraction, 0.7–1.0; bagging_freq, 1–5), and the minimum number of samples per leaf (min_data_in_leaf, 10–100). Other LightGBM settings, such as the boosting type (gbdt) and objective function (regression_l1), were kept fixed throughout. Optuna selected the optimal hyperparameter configuration by minimizing the mean absolute error (MAE) obtained from the inner loop of the nested cross-validation, using a total of 50 trials.

After nested cross-validation, the best-performing hyperparameters were selected (either by averaging over outer folds or by majority vote). The final LightGBM model was then trained on the entire dataset to produce predictions for comparison with the GCN model.

3.6 GCN

The implementation of the GCN model was based on a tutorial about creating a simple Pytorch-based GCN(Deshmukh, 2023). Some edits to the tutorial code were made, for example the addition of node features, the addition of an R^2 metric, and correcting a mistake with the standardization of the loss.

Starting from the node and edge matrix representations of a graph, the GCN model first applied a graph convolution using multiple convolution layers. Each convolution layer gives a node information about its neighbors using the matrix multiplication shown in Figure 2. After the convolution layers *pooling* is applied, which turns the 2-dimensional matrix into a 1-dimensional vector that contains the means of every column of the node matrix. This vector can then be passed to the neural network, which is a simple multilayer perceptron (MLP), a fully connected feedforward neural network.

Figure 2: Graph convolution for an acetamide molecule(Deshmukh, 2023).

Hyperparameter optimization for the GCN model was performed using a simple grid search, as this allowed easy parallelization of hyperparameter tuning. The parameters and their possible values included in the grid search are: batch_size (64, 128, 256, 384), hidden_nodes (64, 96, 128), n_conv_layers (1-5), n_hidden_layers (1-3), learning_rate (0.001, 0.003, 0.005, 0.007, 0.01). Other parameters such as the maximum dimensions of the node and edge matrix, and the number of epochs

were kept constant. After nested cross-validation, the same best-performing hyperparameters were obtained for each outer fold: batch_size = 256, hidden_nodes = 128, n_conv_layers = 4, n_conv_layers = 2, learning_rate = 0.003, the number of epochs was kept constant at 50.

3.7 cVAE

The conditional variational autoencoder (cVAE) is chosen for its probabilistic nature, which enables sampling of new molecules with higher validity than a standard autoencoder (Kingma & Welling, 2019). Our implementation is partially inspired by (Gómez-Bombarelli et al., 2018; Lim et al., 2018), except for the encoder. The model consists of a graph-based GCN encoder, a continuous conditioned latent space, and a GRU sequence decoder (Cho et al., 2014; Yuan et al., 2020), combining the richer information from graph representations with the simplicity of SMILES decoding.

Training maximizes the conditional ELBO (equivalently, minimizes the negative ELBO), composed of a categorical cross-entropy reconstruction loss and a Kullback-Leibler divergence regularizing the latent space(Beckham, 2023):

$$\mathcal{L}_{cVAE} = \mathbb{E}[\log\ p(x|z,y)] - \beta D_{KL}(q(z|x,y)\|\mathcal{N}(0,I)) \tag{1}$$

where $q(z|x,y)$ is the encoder's approximate posterior, mapping a molecule and its HOMO-LUMO gap to a latent distribution, and $p(x|z,y)$ is the decoder's likelihood, giving the probability of reconstructing a SMILES sequence from a latent vector and conditioning value.

Hyperparameters are optimized using Bayesian optimization with a TPE sampler (Watanabe, 2023), although only a limited search is feasible, consisting of only 15 trials, due to the large parameter space and time constraints. For the same reason, five inner folds are used for cross-validation instead of ten, but given the large size of the QM9 dataset, this reduction is not expected to significantly affect the results. An overview of the selected optimal hyperparameters is provided in Table 3.

The optimal set of hyperparameters is used to generate new molecules using two approaches: random sampling of latent vectors from the latent space and perturbation of latent vectors corresponding to molecules in the dataset by adding noise drawn from a normal distribution scaled by a factor of 0.5. In both cases, the latent vectors are decoded while conditioning on target HOMO-LUMO gaps.

For each experimental pathway, ten molecules with distinct target HOMO-LUMO gaps are selected. For each molecule, 1,000 generations are performed, and the average chemical validity is computed, as described in Section 3.4. For the valid generations of each molecule, the mean absolute error (MAE) is calculated by comparing the HOMO-LUMO gap predicted by the GCN model with the corresponding target gap, yielding a per-molecule MAE.

4 Results

4.1 LightGBM

For the performance of the model we first looked to the mean absolute error (MAE), root mean square error (RMSE), and R^2 between the predicted and true HOMO-LUMO gap which are calculated for each outer cross-validation fold. These results are shown in Table 1 and Figure 3. We see that the values for the different metrics does not change much with each fold which is in line whit the chosen stratified splits. The mean MAE of 0.1418 eV is in line with the work of Hasan et al. (2025).

The feature importance stability plot (Figure 4) summarizes the mean and standard deviation of LightGBM feature importances across the five outer cross-validation folds for the top 20 descriptors. Points indicate the average importance of each descriptor, while error bars reflect variability across folds, providing insight into both feature relevance and robustness. Several descriptors consistently dominate the model, most notably MinAbsEStateIndex, BCUT2D$_{\text{MWLOW}}$, and BCUT2D$_{\text{MRHI}}$, which exhibit high mean importance with low variability. This indicates that these features are both highly informative and stable predictors of the HOMO–LUMO gap.

Overall, feature importance shows a gradual decay, with approximately the top ten descriptors accounting for most of the predictive power. The low standard deviations observed for the most important features suggest that the model relies on a robust core of chemically meaningful descriptors that remain consistent across different training subsets.These results highlight the relevance of descriptors capturing electronic and topological properties, such as BCUT metrics and EState indices, for accurate HOMO-LUMO gap prediction, while lower-importance features appear to contribute only marginally to overall performance.

4.2 GCN

Just as for the LightGBM model, the mean absolute error (MAE), root mean square error (RMSE), and R^2 between the predicted and true HOMO-LUMO gap were determined for the GCN model. All evaluation metrics of the GCN model (Table 2) have slightly higher values than those of the LightGBM model (Table 1). The consistent appearance of the same group of outliers across all outer folds (upper left half of each R^2 plot in Figure 5) indicates that the stratified outer cross-validation split was balanced.

4.3 cVAE

The reconstruction performance of the cVAE model, quantified by the average character-wise reconstruction accuracy $\overline{CRA}$, across the five outer cross-validation folds is reported in Table 4. The model achieves an overall mean reconstruction accuracy of 0.852, indicating that it is generally capable of accurately reconstructing individual molecules. Notably, outer folds three and five exhibit substantially higher reconstruction accuracies than the remaining folds. This improvement can be attributed to more favorable hyperparameter configurations identified during their respective tuning procedures (Table 3), in particular a higher learning rate, a larger GRU hidden dimension, and a lower β parameter.

Beyond reconstruction performance, the primary metric of interest is the generative capability of the cVAE. Using the best-performing trained model (outer fold 3), the chemical validity of molecules generated from both randomly sampled latent vectors and perturbed latent vectors is evaluated (Figure 6a and Table 5). Across the ten randomly sampled target molecules, validities are obtained between 1.1% and 6.6% with an average of 4.5% while for the perturbed samples, equally low validities are obtained between 1.8% and 6.1% with an average of 4.2%.

For each target molecule, the mean absolute error (MAE) between the HOMO-LUMO gaps predicted by the GCN model and the corresponding target gaps used to condition the cVAE (Section 3.7) is computed and reported in Figure 6b and Table 5. It is important to note that this MAE is calculated per molecule and depends on the number of valid generations, as only valid molecules contribute to the error calculation. Consequently, the validity indirectly influences the MAE. The perturbed and random latent sampling methods yield MAEs averaging on 0.739 eV and 0.763 eV respectively. For

7

both methods, that is considerably larger than the errors reported for the standalone GCN model, meaning there is a significant difference between the target and generated HOMO-LUMO gap.

To further explore the perturbed latent space vectors, we visualize the generated chemical structures to assess how effectively the cVAE can produce molecules with the same HOMO-LUMO gap while exhibiting slight structural variations. Figures 7 to 9 present three starting molecules alongside four of their respective perturbed generated molecules.

5 Discussion

5.1 LightGBM and GCN have similar performances

The LightGBM model achieves a mean absolute error (MAE) of 0.14 eV and an R^2 of 0.97 for HOMO-LUMO gap prediction across all cross-validation folds (Table 1). The consistency of these metrics across folds indicates that the stratified cross-validation splits are well balanced. Our results can be directly compared to those of Hasan et al. (2025), as they employ the same descriptor-based approach using RDKit-generated molecular descriptors as input to their machine learning models. In their study, a LightGBM model achieved a MAE of 0.1675 eV, which is higher than the error obtained in the present work. One likely contributing factor is the hyperparameter optimization strategy: while Hasan et al. (2025) rely on a grid search, we employ Optuna, a Bayesian optimization framework that enables more efficient exploration of the hyperparameter space.

Feature importance analysis further reveals that approximately 50 descriptors contribute meaningfully to the prediction of the HOMO-LUMO gap, with around 20 descriptors exhibiting particularly strong influence, while the remaining features provide only minor refinements to the predictions. The most influential descriptors predominantly capture topological and electronic properties of the molecules, whereas atom-level features are less important. This observation is expected, as variations in atomic composition are implicitly reflected in global electronic and topological descriptors.

The GCN model achieves a mean MAE of approximately 0.16 eV for HOMO-LUMO gap prediction, with a consistently high R^2 of about 0.97 across all cross-validation folds (Table 2), indicating a balanced split. This performance improves upon a very recent GCN result reported for the PCQM4Mv2 dataset, where a MAE of around 0.20 eV was observed(Hu et al., 2021), and is competitive with other GCN based approaches reporting MAEs of 0.14 eV(Choi et al., 2022) and 0.12 eV(Therrien et al., 2025). Although the present model does not surpass the best-performing architectures in the literature, the relatively small performance gap suggests that a standard 2D GCN can capture most of the relevant structure–property relationships governing the HOMO–LUMO gap. Overall, these results reinforce the effectiveness of graph convolutional networks for molecular electronic property prediction while highlighting that further gains likely require more expressive models or additional structural information.

Although prior work has shown that deep-learning models do not consistently outperform non-deep models for molecular property prediction(Xia et al., 2023), our results show that the GCN attains comparable accuracy to that of the LightGBM model. While the GCN model learns chemically meaningful structural relationships through a graph, the LightGBM model uses descriptors of the molecule to learn mappings between these features and the target property. Rather than clearly favoring one approach over the other, our findings suggest that deep-learning and traditional models address the the prediction of HOMO-LUMO gap from complementary perspectives while achieving similar overall performance. A more comprehensive comparison across a wider range of models and

datasets, similar to the analysis in Xia et al. (2023) but focused specifically on HOMO-LUMO gap prediction, would help clarify the relative advantages of traditional versus deep-learning approaches for this specific task.

5.2 Molecular Generation with cVAE

The cVAE model achieves a mean character-wise reconstruction accuracy of $\overline{CRA} = 0.852$Table 4. While reconstruction accuracies for cVAEs on the exact same dataset are not available, Blaschke et al. (2018) report 0.862 with teacher forcing (0.963 without) for a standard VAE, and Mollaysa et al. (2024) achieved 0.961 with a Transformer VAE. These results show that our cVAE performs just below literature standard, however, our limited hyperparameter tuning suggests further improvements are possible.

Despite this reconstruction performance, the chemical validity of both the randomly sampled and perturbed generated molecules is low and comparable (Figure 6a), which is surprising given prior work showing that local latent sampling often increases validity (Lim et al., 2018). A likely contributing factor to the overall low validity is the high teacher forcing ratio employed during training (Table 3), which improves reconstruction by conditioning the decoder on ground-truth tokens but reduces its robustness during free generation. Consequently, the model learns to reconstruct known sequences well but struggles to generate valid novel molecules when sampled freely.

In addition, the model struggles to generate molecules closely resembling structures from the dataset (Figures 7 to 9). Although the model captures certain chemical motifs, such as the ring structures observed in Figure 9, these patterns are not consistently preserved across the generated samples. This contrasts Lim et al. (2018) and Gómez-Bombarelli et al. (2018), who reported structurally similar generated molecules. The lack of structural preservation observed here suggests that the learned latent space does not encode fine-grained molecular structure and a deeper analysis of the quality of the latent space is required to find the origin of this issue.

Finally, the large mean absolute errors observed for the HOMO-LUMO gaps of generated molecules further indicate limited conditional control (Table 5). The errors are substantially larger than those reported in literature (Gómez-Bombarelli et al., 2018), highlighting that the conditional encoding is insufficient to reliably enforce target electronic properties, especially given low chemical validity.

6 Conclusion

From a predictive modeling perspective, the LightGBM approach demonstrated strong and reliable performance for HOMO–LUMO gap estimation. By leveraging RDKit-based molecular descriptors and Bayesian hyperparameter optimization with Optuna, the model achieved a mean absolute error of 0.14 eV and an R^2 of 0.97, which is competitive with, and in some cases superior to, recent literature results (Hasan et al., 2025). The stability of the performance across cross-validation folds and the consistency of feature importance rankings indicate that the model captures robust structure-property relationships rather than overfitting to specific subsets of the data. Furthermore, the dominance of topological and electronic descriptors among the most influential features highlights the physical relevance of the learned mappings. These results confirm that well-tuned, descriptor-based ensemble models such as LightGBM remain highly effective baselines for molecular electronic property prediction, particularly when large, high-quality datasets are available.

In this work, a conditional variational autoencoder was developed to generate molecular SMILES conditioned on the HOMO-LUMO gap. While the model achieved a reasonable reconstruction accuracy of 0.852, the generated molecules exhibited low chemical validity, limited structural similarity to source molecules and large deviations from the target HOMO-LUMO gaps. Further improvements should focus on increasing the reconstruction accuracy through more extensive hyperparameter tuning, enhancing generative robustness by, among other factors, more careful handling of the teacher forcing ratio, and systematically assessing the quality of the learned latent space using methods proposed in prior work (Gómez-Bombarelli et al., 2018; Lim et al., 2018). Future research should also explore improving the encoder architecture and replacing the SMILES-based decoder with a graph-based decoder, which could explicitly address graph isomorphism and prevent the generation of strings that do not correspond to valid molecular graphs (Gómez-Bombarelli et al., 2018).

7 Disclaimers

Portions of this report were improved using ChatGPT version 5 to refine writing style and enhance clarity. Furthermore, ChatGPT was also used to adapt existing code for the implementation of the machine learning models. All scientific decisions, experiments, analyses and interpretations were done by the authors themselves.

Appendix A. Evaluation metrics

$$MAE = \frac{1}{N} \sum^{N} |y_{pred} - y_{true}| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N} \sum^{N} (y_{pred} - y_{true})^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum^{N} (y_{true} - y_{pred})^2}{\sum^{N} (y_{true} - \bar{y})^2} \tag{4}$$

$$\overline{CRA} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{T_n} \sum_{t=1}^{T_n} \mathbf{1}(\hat{x}_{n,t} = x_{n,t}) \tag{5}$$

with N the total number of test molecules and T_n the number of characters per SMILES string.

$$Validity = \frac{1}{n_{trials}} \sum_{i=1}^{n_{trials}} \mathbf{1}(isValid(\hat{x}_i)) \tag{6}$$

Appendix B. LightGBM

Table 1: LightGBM : mean absolute error (MAE), root mean square error (RMSE), and R^2 between predicted and target HOMO-LUMO gaps per outer fold. Mean values across all five folds are also reported.

	MAE (eV)	RMSE (eV)	R^2
Outer Fold 1	0.1417	0.2160	0.9715
Outer Fold 2	0.1408	0.2122	0.9727
Outer Fold 3	0.1411	0.2200	0.9705
Outer Fold 4	0.1430	0.2183	0.9709
Outer Fold 5	0.1423	0.2227	0.9696
Mean	0.14	0.22	0.97
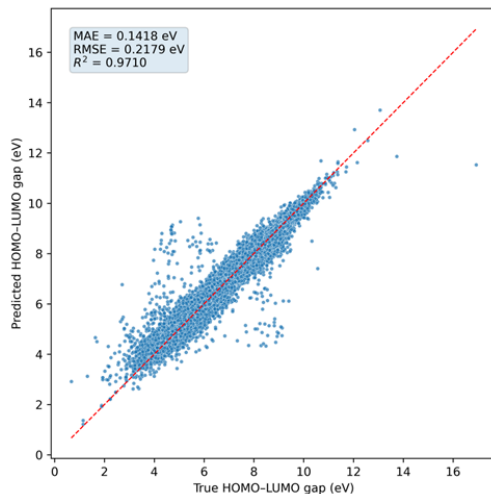
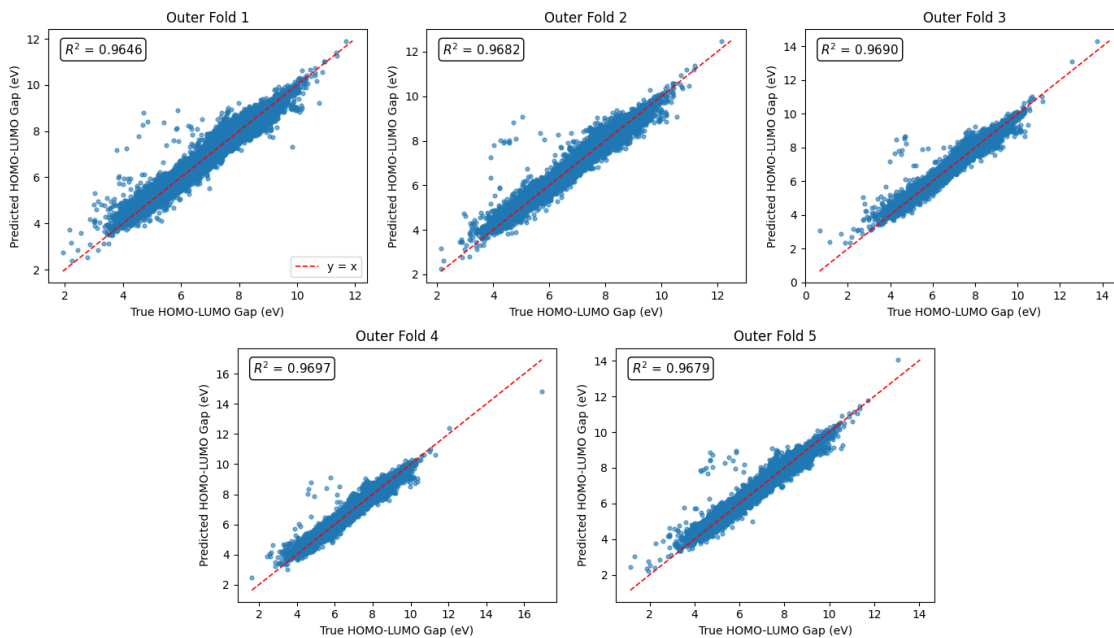**Figure 3:** Predicted vs true HOMO-LUMO gap for all 5 outer folds together of the LigthGBM model.
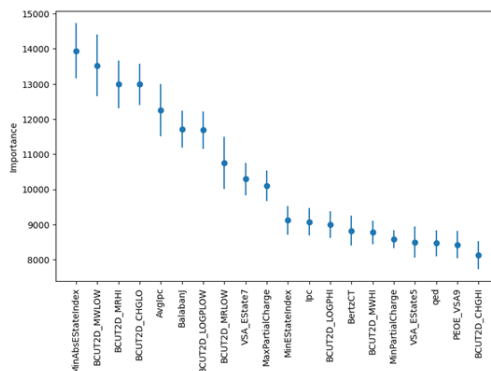
Figure 4: The top 20 most important features across the 5 outer folds where the error bar reflect variability across folds.

Appendix C. GCN

Table 2: GCN: Mean absolute error (MAE), root mean square error (RMSE), and R^2 between predicted and target HOMO-LUMO gaps per outer fold. Mean values across all five folds are also reported.

	MAE (eV)	RMSE (eV)	R^2
Outer Fold 1	0.1677	0.2403	0.9646
Outer Fold 2	0.1585	0.2280	0.9682
Outer Fold 3	0.1567	0.2255	0.9690
Outer Fold 4	0.1586	0.2231	0.9697
Outer Fold 5	0.1587	0.2299	0.9679
Mean	0.16	0.23	0.97

Figure 5: Predicted vs true HOMO-LUMO gap for all 5 outer folds of the GCN model.

Appendix D. cVAE

D.1 Hyperparameters

Table 3: The optimized cVAE hyperparameters for each outer fold achieved after running 15 trials in a Bayesian optimization algorithm using a TPE sampler. The number of epochs and batch size were fixed to narrow the searching space of the Bayesian algorithm.

|	fold 1	fold 2	fold 3	fold 4	fold 5
learning rate	0.00133	0.00133	0.00548	0.00133	0.00548
latent dimension	62	62	50	62	50
#hidden nodes	79	79	66	79	66
GRU dimension	42	42	64	42	64
#convolution layers	1	1	1	1	1
#hidden layers	1	1	2	1	2
#GRU layers	1	1	1	1	1
#FC layers	3	3	2	3	2
embedding dimension	18	18	14	18	14
teacher forcing ratio	0.876	0.876	0.888	0.876	0.888
β	4.62	4.62	1.88	4.62	1.88
batch size	1000	1000	1000	1000	1000
#epochs	15	15	15	15	15

D.2 Results

Table 4: The average character-wise reconstruction accuracy, as calculated by Equation (5), of the test set per outer fold of the cross-validation.

	$\overline{CRA}$
Outer fold 1	0.842
Outer fold 2	0.840
Outer fold 3	0.871
Outer fold 4	0.846
Outer fold 5	0.863
Mean	0.852

Table 5: The validity and MAE of the random samples and perturbed samples, obtained as described in Section 3.7.

| Molecule	Random samples		Perturbed samples	
	Validity	gap MAE (eV)	Validity	gap MAE (eV)
1	0.066	0.605	0.042	0.918
2	0.054	0.543	0.018	0.926
3	0.048	0.886	0.045	0.620
4	0.011	1.135	0.045	0.748
5	0.066	0.726	0.049	0.796
6	0.018	0.992	0.034	0.609
7	0.044	0.467	0.055	0.806
8	0.036	0.817	0.061	0.707
9	0.048	0.904	0.026	0.890
10	0.061	0.850	0.042	0.369
Mean	0.045	0.763	0.042	0.739
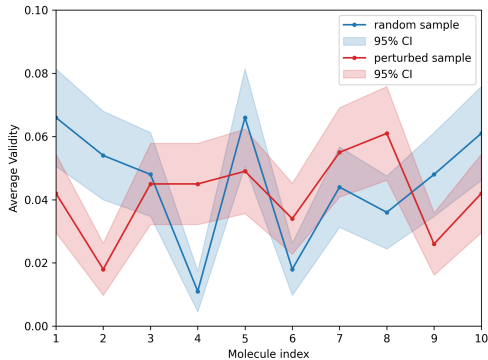
(a) The average validity for both 10 randomly sampled latent vectors and 10 perturbed latent vectors with their respective 95% confidence intervals.
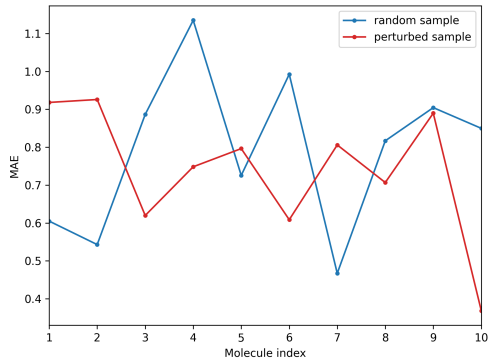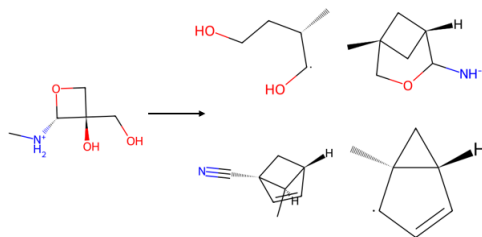
(b) The MAEs of the valid generated molecules for each of the 10 random and 10 perturbed samples.
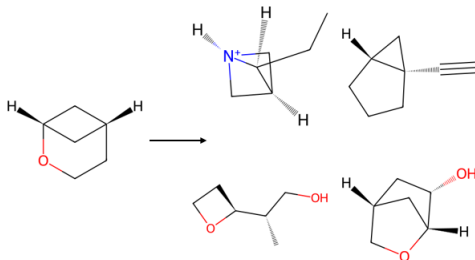
Figure 6: A visual representation of the data from Table 5.
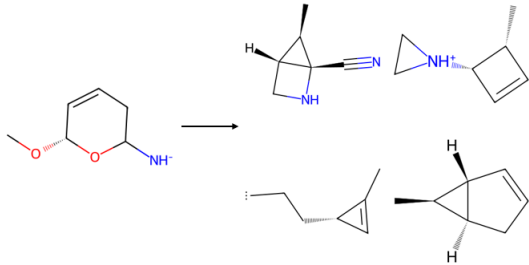
Figure 7: Starting molecule 1 from Table 5 and four of its perturbed generated molecules.

Figure 8: Starting molecule 5 from Table 5 and four of its perturbed generated molecules.

Figure 9: Starting molecule 10 from Table 5 and four of its perturbed generated molecules.

References

Beckham, C. (2023). Conditional vaes. https://beckham.nz/2023/04/27/conditional-vaes.html

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., & Chen, H. (2018). Application of generative autoencoder in de novo molecular design. *Molecular Informatics*, *37*(1-2), https://onlinelibrary.wiley.com 1700123. https://doi.org/https://doi.org/10.1002/minf.201700123

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. https://arxiv.org/abs/1406.1078

Choi, J. Y., Zhang, P., Mehta, K., Blanchard, A., & Pasini, M. L. (2022). Scalable training of graph convolutional neural networks for fast and accurate predictions of homo-lumo gap in molecules. https://arxiv.org/abs/2207.11333

Costa, J. C., Taveira, R. J., Lima, C. F., Mendes, A., & Santos, L. M. (2016). Optical band gaps of organic semiconductor materials. *Optical Materials*, *58*, 51–60. https://doi.org/https://doi.org/10.1016/j.optmat.2016.03.041

Deshmukh, G. (2023). *Building a graph convolutional network for molecular property prediction* [Accessed: 2025-12-15]. Medium – TDS Archive. https://medium.com/data-science/building-a-graph-convolutional-network-for-molecular-property-prediction-978b0ae10ec4

Dwivedi, A. D. D., Maji, D., & Chakrabarti, P. (2025). Introductory Chapter: Organic Electronics – From Fundamentals to Applications [Section: 1]. In A. D. D. Dwivedi, D. Maji, & P. Chakrabarti (Eds.), *Organic Electronics - From Fundamentals to Applications*. Section: 1. London, IntechOpen. https://doi.org/10.5772/intechopen.1011119

Goh, K. L., Goto, A., & Lu, Y. (2022). Lgb-stack: Stacked generalization with lightgbm for highly accurate predictions of polymer bandgap. *ACS Omega*, *7*(34), 29787–29793. https://doi.org/10.1021/acsomega.2c02554

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of

molecules [PMID: 29532027]. *ACS Central Science*, *4*(2), https://doi.org/10.1021/acscentsci.7b00572, 268–276. https://doi.org/10.1021/acscentsci.7b00572

Hasan, M. M., Tarkhaneh, O., Bungay, S. D., Poirier, R. A., & Islam, S. M. (2025). Predicting homo-lumo gaps using hartree-fock calculated data and machine learning models [Epub 2025 Sep 10, PMID: 40929702]. *J. Chem. Inf. Model.*, *65*(18), 9497–9515. https://doi.org/10.1021/acs.jcim.5c01412

Hu, W., Fey, M., Ren, H., Nakata, M., Dong, Y., & Leskovec, J. (2021). Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*. https://ogb.stanford.edu/docs/lsc/leaderboards/#pcqm4mv2

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *CoRR*, *abs/1906.02691* arXiv 1906.02691. https://arxiv.org/abs/1906.02691

Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017). Grammar variational autoencoder (D. Precup & Y. W. Teh, Eds.). In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning*, PMLR. https://proceedings.mlr.press/v70/kusner17a.html

Li, Z., Jiang, M., Wang, S., & Zhang, S. (2022). Deep learning methods for molecular representation and property prediction. *Drug Discovery Today*, *27*(12), 103373. https://doi.org/https://doi.org/10.1016/j.drudis.2022.103373

Lim, J., Ryu, S., Kim, J. W., & Kim, W. Y. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, *10*(1), 31. https://doi.org/10.1186/s13321-018-0286-7

Liu, X., Chen, H., & Tan, S. (2015). Overview of high-efficiency organic photovoltaic materials and devices. *Renewable and Sustainable Energy Reviews*, *52*, 1527–1538. https://doi.org/10.1016/j.rser.2015.08.032

Mollaysa, A. Et al. (2024). Conditional variational autoencoders for molecular generation and property control. *arXiv preprint arXiv:2402.11950*. https://doi.org/10.48550/arXiv.2402.11950

RDKit Developers. (2025). *Rdkit.chem module documentation* [Accessed: 2025-12-15]. RDKit. https://www.rdkit.org/docs/source/rdkit.Chem.html

RDKit: Open-Source Cheminformatics. (2025). Rdkit documentation and python tutorials [Accessed December 6, 2025].

Sworakowski, J. (2018). How accurate are energies of homo and lumo levels in small-molecule organic semiconductors determined from cyclic voltammetry or optical spectroscopy? *Synth. Met.*, *235*, 125–130. https://doi.org/10.1016/j.synthmet.2017.11.013

Team, P. (2024). Torch_geometric.datasets.qm9 — pytorch geometric documentation [Accessed: 2025-12-09].

Therrien, F., Sargent, E. H., & Voznyy, O. (2025). Using gnn property predictors as molecule generators. *Nature Communications*, *16*(1), 4301. https://doi.org/10.1038/s41467-025-59439-1

ThiboVE, M., YarnoDJ2409. (2025). Project-machine-learning [GitHub repository, accessed December 18, 2025]. https://github.com/ThiboVE/project-machine-learning

Walters, W. P., & Barzilay, R. (2021). Applications of deep learning in molecule generation and molecular property prediction [PMID: 33370107]. *Accounts of Chemical Research*, *54*(2), https://doi.org/10.1021/acs.accounts.0c00699, 263–270. https://doi.org/10.1021/acs.accounts.0c00699

Watanabe, S. (2023). Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*. https://doi.org/10.48550/arXiv.2304.11127

Xia, J., Zhang, L., Zhu, X., Liu, Y., Gao, Z., Hu, B., Tan, C., Zheng, J., Li, S., & Li, S. Z. (2023). Understanding the limitations of deep models for molecular property prediction: Insights and solutions (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine, Eds.). In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems*, Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/cc83e97320000f4e08cb9e293b12cf7e-Paper-Conference.pdf

Yoshikai, Y., Mizuno, T., Nemoto, S., & Kusuhara, H. (2024). A novel molecule generative model of vae combined with transformer for unseen structure generation. *arXiv*. https://doi.org/10.48550/arXiv.2402.11950

Yuan, Q., Santana-Bonilla, A., Zwijnenburg, M. A., & Jelfs, K. E. (2020). Molecular generation targeting desired electronic properties via deep generative models. *Nanoscale, 12*, 6744–6758. https://doi.org/10.1039/C9NR10687A