*Original Article*

Check for updates

Available online at
www.heca-analitika.com/ijds

**Infolitika Journal of Data Science**

Vol. 3, No. 1, 2025

# Inductive Biases in Feature Reduction for QSAR: SHAP vs. Autoencoders

**Teuku Rizky Noviandy [1], Ghifari Maulana Idroes [2], Andi Lala [3], Zuchra Helwani [4] and Rinaldi Idroes [3,*]**

[1] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.)
[2] Department of Nuclear Engineering and Engineering Physics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia; ghifarimaulana145@gmail.com (G.M.I.)
[3] School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; andi_lala@usk.ac.id (A.L.); rinaldi.idroes@usk.ac.id (R.I.)
[4] Department of Chemical Engineering, Universitas Riau, Pekanbaru 28293, Indonesia; zuchra.helwani@lecturer.unri.ac.id (Z.H.)

* Correspondence: rinaldi.idroes@usk.ac.id

**Article History**

Received 7 March 2025
Revised 14 May 2025
Accepted 22 May 2025
Available Online 30 May 2025

**Keywords**:
Feature selection
Interpretability
Generalization
LightGBM
Autoencoder

**Abstract**

Machine learning models in drug discovery often depend on high-dimensional molecular descriptors, many of which may be redundant or irrelevant. Reducing these descriptors is essential for improving model performance, interpretability, and computational efficiency. This study compares two widely used reduction strategies: SHAP-based feature selection and autoencoder-based compression, within the context of Quantitative Structure-Activity Relationship (QSAR) classification. LightGBM is used as a consistent modeling framework to evaluate models trained on all descriptors, the top 50 and 100 SHAP-ranked descriptors, and a 64-dimensional autoencoder embedding. The results show that SHAP-based selection produces interpretable and stable models with minimal performance loss, particularly when using the top 100 descriptors. In contrast, the autoencoder achieves the highest test performance by capturing nonlinear patterns in a compact, low-dimensional representation, although this comes at the cost of interpretability and consistency across data splits. These findings reflect the differing inductive biases of each method. SHAP prioritizes sparsity and attribution, while autoencoders focus on reconstruction and continuity. The analysis emphasizes that descriptor reduction strategies are not interchangeable. SHAP-based selection is suitable for applications where interpretability and reliability are essential, such as in hypothesis-driven or regulatory settings. Autoencoders are more appropriate for performance-driven tasks, including virtual screening. The choice of reduction strategy should be guided not only by performance metrics but also by the specific modeling requirements and assumptions relevant to cheminformatics workflows.



## 1. Introduction

Machine learning models are widely used in drug discovery to predict properties such as biological activity from molecular structure [1–3]. These models often rely on molecular descriptors, which are numerical features that summarize the chemical properties of compounds [4]. However, these descriptors can be high-dimensional, noisy, and include redundant or irrelevant information [5]. Using all available descriptors can make models more difficult to interpret and sometimes reduce their performance [6]. A key challenge is deciding how to reduce the number of descriptors while keeping the most important information for accurate predictions.

Several descriptor reduction techniques exist, each embedding different assumptions about what information is relevant. Explainable machine learning methods, such as SHAP (SHapley Additive exPlanations), assess post-hoc feature importance, enabling the selection of descriptors most influential to model predictions [7, 8]. Alternatively, autoencoders, a type of neural network, learn a compressed, latent representation of the input by minimizing reconstruction loss, capturing global patterns in an unsupervised manner [9]. These methods represent fundamentally different inductive biases: SHAP emphasizes sparsity and attribution, while autoencoders favor continuity and reconstruction of the feature space.

Despite the widespread use of descriptor reduction techniques in cheminformatics, few studies have systematically compared these methods through the lens of their inductive biases. Existing work often treats feature selection or dimensionality reduction as a preprocessing step, without considering how the underlying assumptions of each approach affect generalization, stability, and interpretability, critical factors in real-world drug discovery workflows [10, 11].

This study presents a comparative analysis of SHAP-based feature selection and autoencoder-based dimensionality reduction in Quantitative Structure-Activity Relationship (QSAR) modeling. The analysis evaluates how the inductive bias of each method impacts classification performance, using LightGBM as a consistent model framework. Three configurations are examined: using all descriptors without any reduction, selecting the top 50 and 100 descriptors based on SHAP values, and compressing the descriptors into a 64-dimensional latent space using an autoencoder.

Although interpretability and compression techniques such as SHAP and autoencoders are increasingly applied in cheminformatics, few studies provide a rigorous, side-by-side empirical evaluation of their inductive biases in QSAR modeling. This study offers a systematic, application-driven comparison that is directly relevant to drug discovery practice. By employing a unified modeling framework, a standardized dataset, and consistent evaluation metrics, the analysis illustrates how different descriptor reduction strategies affect model performance, generalization, stability, and interpretability. The results contribute to a clearer understanding of the conditions under which sparse, interpretable representations are preferable to dense, learned embeddings in bioactivity classification tasks. These findings are intended to support the selection of appropriate feature reduction techniques based on specific modeling objectives and practical constraints.

The comparison was guided by key expectations grounded in the inductive biases of each method. SHAP-based feature selection, with its emphasis on sparsity and attribution, was expected to produce models that are more interpretable and stable across different data splits, while maintaining competitive predictive performance. In contrast, autoencoder-based compression was anticipated to better capture nonlinear patterns in the descriptor space, potentially enhancing generalization to unseen data at the expense of interpretability. These expectations informed the design of the comparative analysis and provide a framework for evaluating the strengths and trade-offs of each approach within a realistic QSAR modeling context.

The rest of this paper is organized as follows: Section 2 describes the methods. Section 3 presents experimental results, discusses findings, and their implications. Section 4 concludes with recommendations and directions for future work.

## 2. Materials and Methods

An overview of the experimental workflow is shown in Figure 1. This includes data preprocessing, application of different descriptor reduction strategies, model training using LightGBM, and performance evaluation via cross-validation (CV) and a held-out test set.

### *2.1. Dataset*

The dataset was constructed from ChEMBL bioactivity data for the tyrosine-protein kinase receptor (Target ID CHEMBL4895) [12]. This receptor is a well-studied target in drug development. Compound-target pairs were retrieved where the reported bioactivity measurement was $IC_{50}$, expressed in nanomolar (nM) units. To ensure data consistency, only standard $IC_{50}$ values were retained, with erroneous entries excluded.

Each compound was labeled as either active or inactive based on a predefined threshold: compounds with $IC_{50}$ values less than 1000 nM were labeled as active, while those with $IC_{50}$ values equal to or greater than 1000 nM were labeled as inactive [13]. This binary classification approach is commonly used in virtual screening pipelines to distinguish between potentially potent and weak inhibitors [14].

Molecular structures were represented using canonical SMILES strings, and numerical descriptors were computed using the Mordred descriptor calculator. Mordred generates a comprehensive set of descriptors that capture a wide range of physicochemical, topological, and structural properties [15]. Compounds with missing or invalid descriptor values were excluded

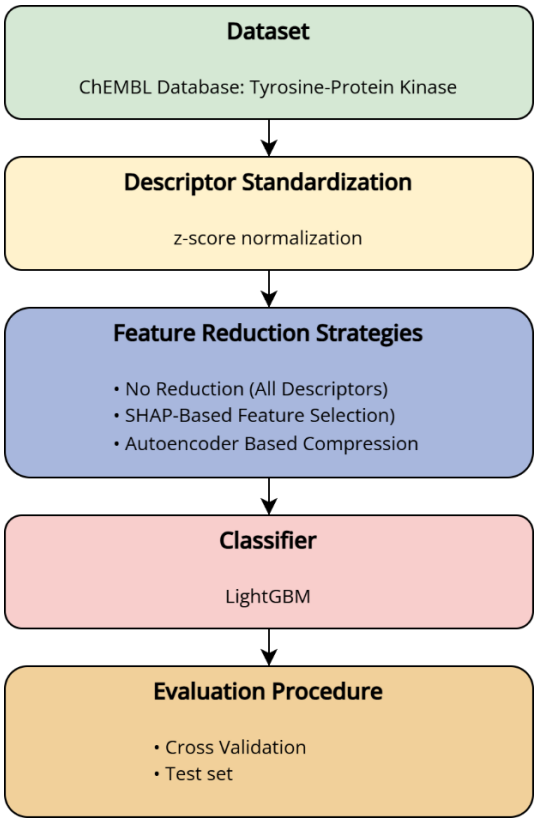


**Figure 1.** Workflow of this study.

to ensure a clean and fully usable dataset. After filtering and preprocessing, the final dataset consisted of 944 compounds described by 1,073 molecular descriptors.

The dataset was randomly split into training (80%) and test (20%) sets using stratified sampling to preserve the proportion of active and inactive compounds in both subsets [16]. The training set was used for all model development and CV procedures, while the test set was held out and used exclusively for final performance evaluation.

*2.2. Descriptor Standardization*

Before model training, all molecular descriptors were standardized to ensure comparability across features and to mitigate biases introduced by differing scales. Descriptor values were transformed using z-score normalization, where each feature was rescaled to have a mean of zero and a variance of one, based on statistics computed from the training set [17]. This standardization step is essential for algorithms that are sensitive to feature magnitude, such as gradient-boosted trees and neural networks [18].

The same transformation parameters computed from the training set were applied to the test set to prevent information leakage. Standardization was performed independently for each fold during CV to maintain the integrity of the validation procedure [19]. This preprocessing step ensured that each descriptor contributed proportionally during model learning and that models remained robust to irrelevant variation in feature scales.

*2.3. Feature Reduction Strategies*

To investigate how different assumptions about feature importance influence model performance, three descriptor reduction strategies were implemented: using all descriptors without reduction, selecting the top 50 and 100 descriptors based on SHAP values, and compressing the descriptors into a 64-dimensional latent space using an autoencoder. Each approach introduces a distinct inductive bias, representing a set of prior assumptions about which types of information are most relevant for learning from data.

In this study, SHAP-based feature selection introduces a sparsity and attribution bias, prioritizing features that have the most consistent and high-magnitude influence on model output. This aligns with the assumption that only a small subset of features carries most of the predictive signal, favoring transparency and interpretability. In contrast, the autoencoder approach assumes a continuity and reconstruction bias. It operates under the premise that useful information about molecular structure lies on a lower-dimensional, continuous manifold that can be uncovered through reconstruction. This assumption supports dense, abstract representations that may better capture complex, nonlinear feature interactions, but at the expense of interpretability. These inductive biases influence not only how the descriptor space is reduced but also how well the resulting models generalize to unseen data, how interpretable their outputs are, and how stable their performance is across different data splits. The following subsections describe the specific implementation of each strategy.

*2.3.1. No Reduction (All Descriptors)*

As a baseline, models were trained using the complete set of standardized molecular descriptors, without applying any feature selection or dimensionality reduction. This strategy assumes that the learning algorithm (in this case, LightGBM) can internally identify and downweight irrelevant or redundant features. While this provides access to the full descriptor space, it may also increase the risk of overfitting, reduce interpretability, and slow down model training.

*2.3.2. SHAP-Based Feature Selection*

To introduce a sparsity-driven inductive bias, SHAP values were employed for feature attribution. After training an

initial LightGBM model on the full set of descriptors, the TreeExplainer implementation of SHAP was used to compute feature importance scores, quantified as the mean absolute SHAP value for each descriptor across the training set, as shown in Equation 1:

$$\phi_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_{ij}| \quad (1)$$

where $\phi_{ij}$ is the SHAP value of descriptor $j$ for compound $i$, and $\phi_j$ represents the average contribution of descriptor $j$ to model predictions across all samples [7].

Descriptors were then ranked by importance, and two reduced sets were created: one containing the top 50 descriptors and another containing the top 100. This method assumes that the most predictive information resides in a small number of features with high attribution scores, and that removing low-impact descriptors improves model generalization without sacrificing accuracy.

This strategy favors interpretability and sparsity. Each selected descriptor retains its original chemical meaning, enabling downstream analysis and supporting explainable AI applications in drug discovery.

*2.3.3. Autoencoder-Based Dimensionality Compression*

As an alternative to feature selection, the Autoencoder (64) model was also evaluated. This unsupervised neural network learns a compressed representation of the descriptors by reducing the full descriptor space into a fixed-size latent embedding, based on the assumption that underlying chemical variation can be captured in fewer dimensions.

The autoencoder architecture consisted of fully connected layers with ReLU activations and was trained to minimize the reconstruction error between the input and its output. The loss function used was mean squared error (MSE), calculated as shown in Equation 2:

$$LAE = \frac{1}{N}\sum i = 1^{N}|x_i - \hat{x_i}|^2 \quad (2)$$

where $x_i$ is the input descriptor vector of the $i$ –th compound, and $\hat{x}_i$ is the corresponding reconstructed vector produced by the autoencoder. The objective is to learn a compact representation that retains as much information as possible from the original data.

A bottleneck layer with 64 neurons served as the compressed representation. After training on the standardized descriptor matrix, only the 64-dimensional encoded vectors were used as input to the downstream LightGBM classifier.

This method introduces a continuity and reconstruction bias, favoring descriptors that preserve structure across a non-linear transformation, rather than selecting features based on importance to prediction. While the learned embeddings are not directly interpretable, they can capture complex interactions and redundant information in a more compact form, potentially improving generalization, especially in high-dimensional, noisy datasets.

*2.4. Machine Learning Classifier*

All classification tasks in this study were performed using the Light Gradient Boosting Machine (LightGBM), a decision tree–based ensemble algorithm known for its efficiency and performance in high-dimensional, structured datasets such as molecular descriptor matrices [20, 21]. LightGBM constructs boosted decision trees using a histogram-based learning approach, which makes it highly scalable and robust [22].

LightGBM was selected due to its widespread success in cheminformatics and QSAR modeling, especially in contexts where interpretability and feature importance estimation are advantageous [23]. Additionally, it supports integration with SHAP, enabling consistent use of the same model architecture for both classification and feature attribution.

Models were trained using default hyperparameters provided by the LightGBM library unless otherwise noted. A fixed random seed was set to ensure reproducibility across experiments. During CV, model training was confined to the training folds, and no information from the validation or test sets was used during the learning process. For final evaluation, each model was retrained on the full training set and tested on the held-out test set using the same preprocessing and feature reduction pipeline.

*2.5. Evaluation Procedure*

Model performance and the impact of different feature reduction strategies were assessed using a two-stage evaluation framework consisting of cross-validation and held-out test set assessment. This design ensured robustness to sampling variability and provided a fair estimation of generalization performance.

During model development, 5-fold stratified cross-validation was conducted on the training set. Stratification was applied to preserve the proportion of active and inactive compounds within each fold [24, 25]. For each split, models were trained on four folds and

**Table 1.** Classification performance of different descriptor reduction strategies using LightGBM

| Model | CV Accuracy (%) | CV F1 Score (%) | Test Accuracy (%) | Test F1 Score (%) |
|---|---|---|---|---|
| All Features | 0.8172 | 0.8161 | 0.7672 | 0.7634 |
| Top 50 SHAP | 0.8185 | 0.8174 | 0.7725 | 0.7693 |
| Top 100 SHAP | 0.8172 | 0.8162 | 0.7778 | 0.7758 |
| Autoencoder (64) | 0.8079 | 0.8069 | 0.7884 | 0.7872 |

evaluated on the remaining fold. This process was repeated across all five folds, and the mean and standard deviation of key classification metrics were reported. Importantly, data preprocessing steps, including descriptor standardization and feature reduction, were performed independently within each fold to prevent information leakage from the validation data.

Following CV, each model was retrained on the entire training set using the same feature reduction strategy, and final performance was measured on the held-out test set. This test set was not used at any point during model selection, training, or validation, thereby providing an unbiased estimate of the model's real-world predictive performance.

Performance was quantified using accuracy and weighted F1 score. Accuracy captures the overall correctness of predictions, while the weighted F1 score accounts for class imbalance by combining precision and recall, weighted by class support [26]. These metrics were chosen to strike a balance between interpretability and relevance in drug discovery, where both precision (reducing false positives) and recall (capturing actives) are crucial.

### *2.6. Reproducibility and Implementation Details*

All experiments were conducted in Python, utilizing open-source libraries to ensure transparency and reproducibility. Molecular descriptors were calculated using the Mordred descriptor calculator (version 1.2.0) and standardized with StandardScaler from scikit-learn (version 1.5.2) [27].

SHAP-based feature selection was conducted using the SHAP library (v0.47.2). Autoencoders were implemented using PyTorch (version 2.6.0), and classification tasks were performed with LightGBM (version 4.1.0) using the default hyperparameters. A fixed random seed (42) was used throughout to ensure reproducibility.

## 3. Results and Discussion

### *3.1. Overall Performance*

Table 1 summarizes the performance of the four descriptor configurations evaluated in this study: using all descriptors, SHAP-based selection (Top 50 and Top 100), and autoencoder-based compression to 64 dimensions. Results are reported for a 5-fold stratified CV and an independent held-out test set.

All models achieved competitive results, with the baseline model, which utilized all descriptors, performing well across both the CV and test sets. SHAP-based feature selection maintained or slightly improved performance, particularly with the Top 100 feature set, which achieved the highest test F1 score (77.58%). This suggests that removing less informative descriptors can benefit generalization while simplifying the model input.

Interestingly, the model using the Top 50 SHAP features performed nearly as well as the full feature set, while using only a small fraction of the total descriptors. This supports the utility of SHAP as an effective feature selection tool, offering both predictive strength and interpretability.

The Autoencoder (64) model achieved the highest test accuracy (78.84%) and F1 score (78.72%), although its CV performance was slightly lower than that of the SHAP-based models. This indicates that the compressed representation learned by the autoencoder may generalize well, though it lacks the direct interpretability of SHAP-based selection.

To better understand the distribution of CV scores, Figures 2 and 3 show boxplots of accuracy and F1 across folds for each descriptor configuration. The accuracy comparison (Figure 2) shows similar median values across all models. The Top 100 SHAP model displays a slightly higher upper range, while the Top 50 and all-feature models show more tightly grouped scores. The autoencoder (64) model's distribution is broader, suggesting more variation in fold performance.

The F1 score distribution (Figure 3) follows a similar pattern. SHAP-based models performed consistently, with the Top 100 SHAP variant again showing slightly higher central performance. The autoencoder (64) model achieved comparable median F1 but showed more variability across folds.

Overall, these results demonstrate the effectiveness of SHAP for descriptor reduction, with minimal

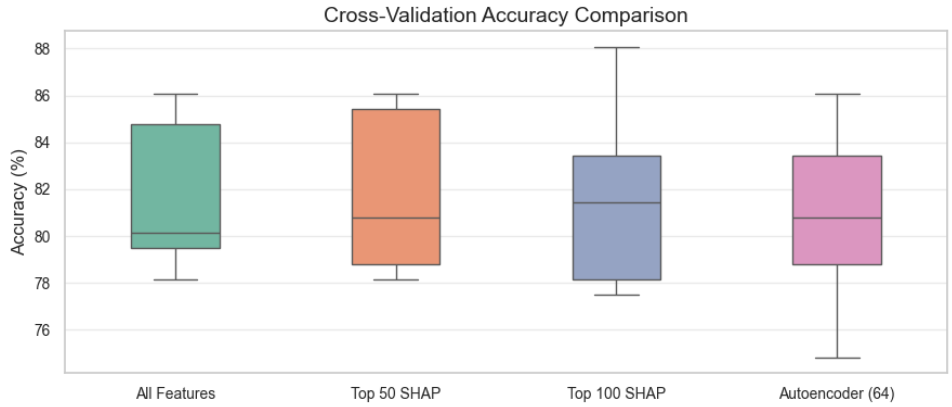


**Figure 2.** Boxplot of cross-validation accuracy for all descriptor reduction strategies.

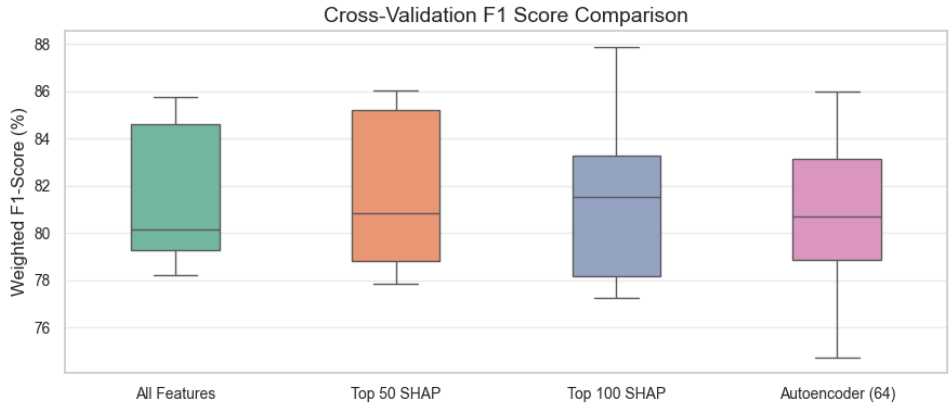


**Figure 3.** Boxplot of cross-validation weighted F1 scores for all descriptor reduction strategies.

performance loss and, in some cases, improvement. The autoencoder offers strong generalization potential but may be more sensitive to training dynamics. These trends suggest that SHAP-based selection is well-suited for tasks where model simplicity and interpretability are crucial, while autoencoders may be more suitable when compact representations are required.

*3.2. SHAP-Based Feature Selection*

Two SHAP-based feature selection strategies were compared: selecting the top 50 and top 100 descriptors based on their mean absolute SHAP values computed from a LightGBM model trained on the full feature set. As shown in Table 1, both approaches performed well and remained competitive with the baseline model using all descriptors.

The model using the Top 100 SHAP features achieved the highest test set scores among the SHAP-based configurations, with a test accuracy of 77.78% and a weighted F1 score of 77.58%. The Top 50 SHAP model also performed strongly, with slightly lower test accuracy and F1 (77.25% and 76.93%, respectively), while using half the number of features. These results suggest that including more features up to 100 helps capture additional predictive signal, leading to marginal but meaningful performance improvements.

CV performance follows a similar trend, with both SHAP-based models exhibiting comparable accuracy and F1 scores across folds. The median scores are nearly identical, though the Top 100 SHAP model exhibits a slightly broader upper performance range, indicating better generalization capacity in some splits.

The SHAP importance decay curve shown in Figure 4 provides further insight into these results. It illustrates how SHAP values decrease sharply across the top-ranked features, with the most influential descriptors contributing significantly more than those ranked lower. This steep drop supports the effectiveness of selecting the top 50 descriptors, where most predictive signal resides, while also explaining the slight performance boost gained by expanding to the top 100.

This highlights a trade-off between dimensionality and performance. The Top 50 SHAP model provides a compact and efficient representation with minimal loss in accuracy, making it advantageous when simplicity, speed, or memory usage are critical. The Top 100 model, while s

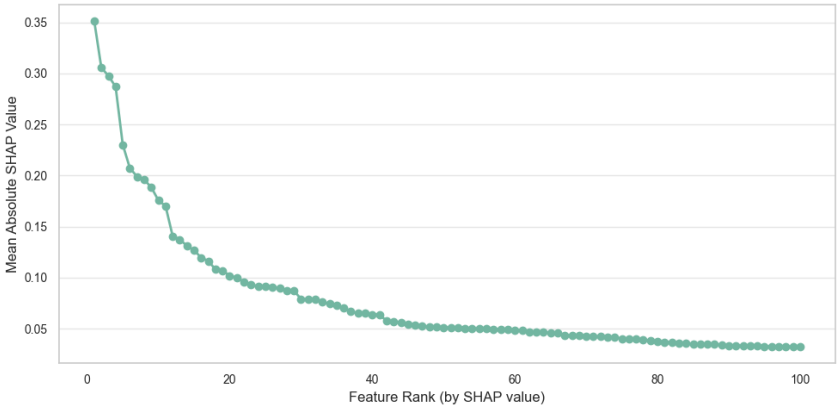


**Figure 4.** SHAP importance decay across the top 100 molecular descriptors.

lightly larger, delivers the best overall generalization, making it preferable in performance-sensitive applications.

A key advantage of SHAP-based feature selection is its interpretability. Unlike unsupervised methods such as autoencoders, each retained descriptor in the SHAP-ranked set maintains its original chemical meaning. This facilitates downstream analysis and supports explainability, which is particularly important in cheminformatics, where understanding which molecular features drive activity is often as important as making accurate predictions.

*3.3. Autoencoder-Based Compression*

The autoencoder-based approach involved compressing the full descriptor set into a 64-dimensional latent space using an unsupervised neural network. Unlike SHAP-based feature selection, which relies on feature importance derived from model predictions, this method learns a compact latent representation by minimizing the reconstruction error of the input descriptors. In doing so, it identifies a lower-dimensional embedding that preserves structural patterns in the data, without supervision from the classification labels.

As shown in Table 1, the Autoencoder (64) model achieved the highest overall test set performance, with a test accuracy of 78.84% and a weighted F1 score of 78.72%. This result is particularly noteworthy given that its average CV accuracy and F1 score (80.79% and 80.69%, respectively) were slightly lower than those of the SHAP-based models. This suggests that the autoencoder learned a representation that generalizes especially well to unseen data, even though it was not explicitly trained for the classification task.

Figures 2 and 3 further illustrate that the Autoencoder (64) model exhibits greater variability in CV performance compared to the SHAP-based models. While the median values are comparable, the broader spread across folds reflects the autoencoder's sensitivity to training dynamics, such as initialization, learning rate, and fold composition. This is expected behavior for unsupervised models that do not directly optimize for task-specific performance.

The training process of the autoencoder is visualized in Figure 5, which shows the reconstruction loss (mean squared error) decreasing steadily over 100 epochs. This indicates that the model effectively learned to encode and reconstruct the descriptor space, reinforcing that a meaningful lower-dimensional structure was captured during training.

One major advantage of the autoencoder is its ability to capture non-linear relationships and interactions among descriptors, patterns that SHAP-based selection might overlook due to its reliance on model-derived attributions. The resulting 64-dimensional embeddings are dense, learned representations that can be especially valuable in deep learning pipelines or for integrating heterogeneous data sources.

However, this benefit comes at the cost of interpretability. Unlike SHAP-based methods, where each retained descriptor has a clear chemical meaning, the autoencoder's compressed features are not human-interpretable. This can limit their usefulness in applications where explainability, regulatory compliance, or hypothesis generation is important. Moreover, training autoencoders introduces additional pipeline complexity and may require careful tuning to avoid underfitting or overfitting.

*3.4. Generalization and Stability*

Beyond average performance, it is essential to evaluate how consistently each model performs across various data splits. Generalization refers to how well a model trained on a subset of data performs on unseen

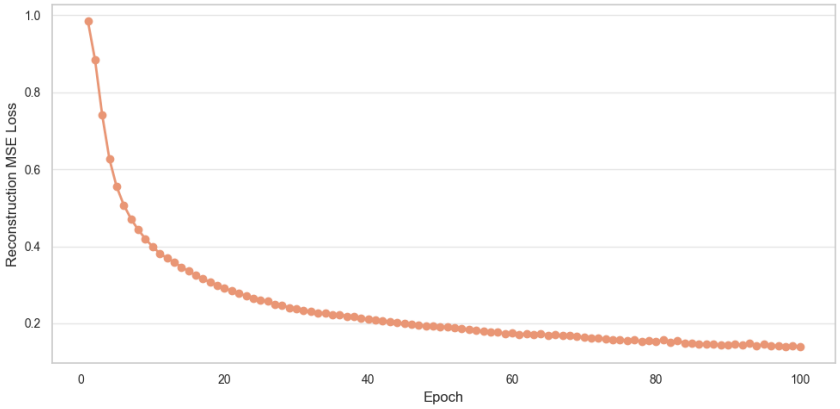


**Figure 5.** Autoencoder training loss over 100 epochs.

examples. At the same time, stability reflects how sensitive the model is to variations in the training data.

The CV results in Table 1, supported visually by Figures 2 and 3, provide insight into this aspect. The models using Top 50 and Top 100 SHAP features demonstrated stable performance across folds, with relatively narrow distributions in both accuracy and F1 scores. This suggests that SHAP-based feature selection, in addition to offering interpretability and compactness, yields models that generalize reliably with minimal performance fluctuation.

The model trained on all features also performed consistently, though with slightly lower test set scores than the SHAP-based configurations. This result indicates that while the full descriptor set contains sufficient information for good predictive performance, it may also include noise or redundancy that slightly limits generalization.

In contrast, the autoencoder (64) model, despite achieving the highest test accuracy and F1 score, showed greater variability across CV folds. As seen in the boxplots, its performance fluctuated more between splits, suggesting sensitivity to initialization, training conditions, or the data partitioning itself. This is expected, given that the autoencoder compresses input features without supervision and relies on reconstruction error, rather than task-specific feedback.

Interestingly, the autoencoder's stronger performance on the held-out test set may indicate that its compressed representation generalizes well beyond the training distribution, even if its validation performance is less stable. This suggests a potential trade-off: autoencoder embeddings may offer better generalization on average, but with less predictable fold-to-fold behavior, which could be significant in high-stakes or regulatory settings.

These patterns reflect the inductive biases of each method. The stability and consistent CV performance of SHAP-based models align with its sparsity and attribution bias, selecting a small, stable set of influential features reduces variance across splits. Conversely, the variability seen in the autoencoder model is consistent with its reconstruction bias. By learning a global, unsupervised embedding of the descriptor space, it captures complex structure at the expense of fold-to-fold consistency.

### *3.5. Implication for Drug Discovery*

The findings from this study provide preliminary insights into how descriptor reduction strategies may influence QSAR model behavior for a specific bioactivity classification task. Although the results are based on a single, well-curated target (CHEMBL4895), they suggest that different dimensionality reduction methods may be more appropriate depending on the specific modeling objective.

SHAP-based feature selection preserved strong classification performance while reducing input dimensionality, making it a practical choice for tasks that benefit from model interpretability and simplicity, such as early-stage virtual screening or exploratory analysis. Since the selected descriptors retain chemical meaning, this approach may also support hypothesis generation and model transparency.

The autoencoder-based method achieved the highest test set performance by learning dense, nonlinear embeddings. Although less interpretable, this method could be useful in situations where predictive performance outweighs explainability, such as large-scale screening or integration into neural network-based workflows.

These observations should be interpreted with caution. The analysis was limited to a single dataset, and while it provides a controlled comparison, broader validation

across diverse targets and endpoints is necessary to confirm generalizability. The findings are intended as context-sensitive guidance rather than definitive recommendations. Different descriptor reduction strategies embed distinct assumptions, and aligning those assumptions with the goals of a modeling task is essential for effective QSAR application in drug discovery.

## 4. Conclusions

This study compared three descriptor reduction strategies in QSAR modeling: using all molecular descriptors, selecting features based on SHAP importance values, and applying autoencoder-based dimensionality compression. SHAP-based selection provided interpretable and compact feature sets, with the Top 100 descriptors achieving strong predictive performance while reducing dimensionality. The autoencoder approach, although less interpretable, achieved the highest test performance by learning dense, low-dimensional representations of molecular structure.

These findings demonstrate that descriptor reduction methods introduce distinct inductive biases that significantly impact model behavior. SHAP's sparsity and attribution bias supports more stable and interpretable models, making it well-suited for settings where transparency and reproducibility are essential. In contrast, the autoencoder's reconstruction bias enables it to capture complex, nonlinear relationships, leading to better generalization to unseen data, especially in performance-driven applications.

Importantly, the results demonstrate that inductive biases are not merely theoretical but manifest in stability, interpretability, and performance trade-offs. Therefore, selecting a descriptor reduction method should involve not only evaluating predictive metrics but also understanding the assumptions each method imposes on the data and their alignment with specific modeling goals in cheminformatics.

**Author Contributions:** Conceptualization, T.R.N. and R.I.; methodology, T.R.N. and R.I.; software, T.R.N., G.M.I. and A.L.; validation, Z.H. and R.I.; formal analysis, T.R.N.; investigation, T.R.N. and G.M.I.; resources, A.L. and R.I.; data curation, G.M.I. and R.I.; writing—original draft preparation, T.R.N., G.M.I. and A.L.; writing—review and editing, Z.H. and R.I.; visualization, T.R.N.; supervision, R.I.; project administration, R.I.; funding acquisition, R.I. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study does not receive external funding.

**Ethical Clearance:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Raw data were retrieved from the ChEMBL database and processed for analysis. The processed datasets supporting the findings of this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** All the authors declare no conflicts of interest.

## References

1. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. (2021). Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery, *Molecular Diversity*, Vol. 25, No. 3, 1315–1360. doi:10.1007/s11030-021-10217-3.
2. Khan, S., Sarfraz, A., Prakash, O., and Khan, F. (2024). Machine Learning-Based QSAR Modeling, Molecular Docking, Dynamics Simulation Studies for Cytotoxicity Prediction in MDA-MB231 Triple-Negative Breast Cancer Cell Line, *Journal of Molecular Structure*, Vol. 1315, 138807. doi:10.1016/j.molstruc.2024.138807.
3. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.
4. Wigh, D. S., Goodman, J. M., and Lapkin, A. A. (2022). A Review of Molecular Representation in the Age of Machine Learning, *WIREs Computational Molecular Science*, Vol. 12, No. 5. doi:10.1002/wcms.1603.
5. Li, J., Luo, D., Wen, T., Liu, Q., and Mo, Z. (2021). Representative Feature Selection of Molecular Descriptors in QSAR Modeling, *Journal of Molecular Structure*, Vol. 1244, 131249. doi:10.1016/j.molstruc.2021.131249.
6. Goodarzi, M., Dejaegher, B., and Heyden, Y. Vander. (2012). Feature Selection Methods in QSAR Studies, *Journal of AOAC INTERNATIONAL*, Vol. 95, No. 3, 636–651. doi:10.5740/jaoacint.SGE_Goodarzi.
7. Lundberg, S. M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, Vol. 30.
8. Noviandy, T. R., Idroes, G. M., Syukri, M., and Idroes, R. (2024). Interpretable Machine Learning for Chronic Kidney Disease Diagnosis: A Gaussian Processes Approach, *Indonesian Journal of Case Reports*, Vol. 2, No. 1, 24–32. doi:10.60084/ijcr.v2i1.204.
9. Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., and Xu, Y. (2024). Autoencoders and Their Applications in Machine Learning: A Survey, *Artificial Intelligence Review*, Vol. 57, No. 2, 28. doi:10.1007/s10462-023-10662-6.
10. Azizah, M., Yanuar, A., and Firdayani, F. (2022). Dimensional Reduction of QSAR Features Using a Machine Learning Approach on the SARS-Cov-2 Inhibitor Database, *Jurnal Penelitian Pendidikan IPA*, Vol. 8, No. 6, 3095–3101. doi:10.29303/jppipa.v8i6.2432.
11. Khan, P. M., and Roy, K. (2018). Current Approaches for Choosing Feature Selection and Learning Algorithms in Quantitative Structure–Activity Relationships (QSAR), *Expert Opinion on Drug Discovery*, Vol. 13, No. 12, 1075–1089. doi:10.1080/17460441.2018.1542428.
12. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012). ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery, *Nucleic Acids Research*, Vol. 40, No. D1, D1100–D1107. doi:10.1093/nar/gkr777.
13. Yu, T., Nantasenamat, C., Kachenton, S., Anuwongcharoen, N., and Piacham, T. (2023). Cheminformatic Analysis and Machine Learning Modeling to Investigate Androgen Receptor

Antagonists to Combat Prostate Cancer, *ACS Omega*, Vol. 8, No. 7, 6729–6742. doi:10.1021/acsomega.2c07346.

14. Toropov, A. A., and Toropova, A. P. (2020). QSPR/QSAR: State-of-Art, Weirdness, the Future, *Molecules*, Vol. 25, No. 6, 1292. doi:10.3390/molecules25061292.

15. Moriwaki, H., Tian, Y. S., Kawashita, N., and Takagi, T. (2018). Mordred: A Molecular Descriptor Calculator, *Journal of Cheminformatics*, Vol. 10, No. 1, 1–14. doi:10.1186/s13321-018-0258-y.

16. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., and Idroes, R. (2024). An Explainable Multi-Model Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates, *Sci*, Vol. 6, No. 4, 81. doi:10.3390/sci6040081.

17. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). Machine Learning Approach to Predict AXL Kinase Inhibitor Activity for Cancer Drug Discovery Using XGBoost and Bayesian Optimization, *Journal of Soft Computing and Data Mining*, Vol. 5, No. 1, 46–56.

18. Ahsan, M., Mahmud, M., Saha, P., Gupta, K., and Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance, *Technologies*, Vol. 9, No. 3, 52. doi:10.3390/technologies9030052.

19. Baron, G., and Stańczyk, U. (2021). Standard vs. Non-Standard Cross-Validation: Evaluation of Performance in a Space with Structured Distribution of Datapoints, *Procedia Computer Science*, Vol. 192, 1245–1254. doi:10.1016/j.procs.2021.08.128.

20. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

21. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.

22. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, Vol. 30.

23. Noviandy, T. R., Maulana, A., Irvanizam, I., Idroes, G. M., Maulydia, N. B., Tallei, T. E., Subianto, M., and Idroes, R. (2025). Interpretable Machine Learning Approach to Predict Hepatitis C Virus NS5B Inhibitor Activity Using Voting-Based LightGBM and SHAP, *Intelligent Systems with Applications*, Vol. 25, 200481. doi:10.1016/j.iswa.2025.200481.

24. Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation, *Molecular Informatics*, Vol. 29, Nos. 6–7, 476–488. doi:10.1002/minf.201000061.

25. Danishuddin, Madhukar, G., Malik, M. Z., and Subbarao, N. (2019). Development and Rigorous Validation of Antimalarial Predictive Models Using Machine Learning Approaches, *SAR and QSAR in Environmental Research*, Vol. 30, No. 8, 543–560. doi:10.1080/1062936X.2019.1635526.

26. Noviandy, T. R., Imelda, E., Idroes, G. M., Suhendra, R., and Idroes, R. (2025). Evaluation of Machine Learning Methods for Identifying Carbonic Anhydrase-II Inhibitors as Drug Candidates for Glaucoma, *Malacca Pharmaceutics*, Vol. 3, No. 1, 32–41. doi:10.60084/mp.v3i1.271.

27. Kramer, O. (2016). Scikit-Learn, 45–53. doi:10.1007/978-3-319-33383-0_5.