MOLECULAR PROPERTY CLIFF DETECTION

DEV VASANI, SARTHAK GOEL, YESHA RAVANI

MLPR END-TERM REPORT

PROBLEM STATEMENT

Machine learning models assume smooth structure-property relationships, but molecular property cliffs violate this assumption; small structural changes can cause large shifts in HOMO-LUMO gap. Our goal is to formulate property cliff detection as a supervised classification problem: given a pair of structurally similar molecules, predict whether they form a property cliff or not.

What is the problem?

- Most ML models assume:
  Small structural change → small property change.
  Hence, they often overlook small structural changes that cause cliffs.
- But in reality, we observe molecular property cliffs.
- Two molecules can be structurally very similar yet show large differences in HOMO–LUMO gaps.

Why this is a problem?

- Standard regression models average out these sharp transitions.
- Predictions become unreliable near these cliff regions.
- This makes ML unreliable for molecular design, screening, and optimization pipelines.

POTENTIAL APPLICATIONS

- Faster Materials Discovery
- Reducing Failed Laboratory Experiments
- Enhancing Explainability in Molecular ML Systems

POTENTIAL IMPACT

- Safer molecular optimization in drug discovery
- Better design of organic electronic materials
- More reliable AI-assisted chemical screening

LITERATURE SURVEY

Study / Approach	Author	What They Developed	Limitations	Our Extension / Improvement
Exploring Molecular Machine Learning Models for Activity-Cliff Prediction (2023)	Sheridan, R. P. (2023)	• Treated cliff detection as a separate binary classification • showed Graph Neural Networks (GNNs) outperform traditional fingerprints (ECFP) for cliff detection	• Focused on biological activity (potency) • Not tested on electronic properties	Apply graph-based models to HOMO-LUMO cliff detection
Exposing the Limitations of Molecular Machine Learning with Activity Cliffs (MoleculeACE) (2022)	van Tilborg, D., Alenicheva, A., & Grisoni, F. (2022)	• Defined formal cliff criteria (Similarity > 0.9, Property difference > 1 log unit)	• Designed for drug discovery • Bioactivity-specific thresholds	Adapt cliff definition to HOMO–LUMO gap thresholds and extend MoleculeACE benchmarking to materials science
Activity Cliff-Informed Contrastive Learning (ACA) (2024)	Wanxiang Shen (2024)	• Introduced Triplet Loss (ACA-loss) to distinguish similar molecules with large property differences	• Computationally intensive • Bioactivity-focused	Incorporate learning to improve separation of structurally similar but energetically different molecules
A Semi-supervised Molecular Learning Framework for Activity Cliff (SemiMol) (2024)	Ai, H., Yu, J., Li, L., & Zhao, Q. (2024)	• Addressed class imbalance (rare cliffs) using semi-supervised training	• Requires labeled cliff examples • Not tailored to electronic datasets	Use supervised learning to improve detection of HOMO–LUMO cliffs
Selected Machine Learning for HOMO–LUMO Gaps Prediction (2021)	Magar, R., Doebler, J. E., & Farimani, A. B. (2021)	• Identified electronic properties as data-inefficient • partitioned molecules into chemical classes (aromatic, unsaturated, saturated) to improve learning efficiency	• Primarily regression-based • No explicit cliff handling	Propose specialized prediction which is to integrate chemical-class structural indicators

THE DATASET

About the dataset

QM9 is a publicly available quantum chemistry dataset containing ~134k small organic molecules and their computed molecular properties. Each molecule has a molecular structure, SMILES representation, and target values like HOMO, LUMO, and the HOMO–LUMO gap.

Why was it chosen?

- Molecules are naturally represented as graphs (atoms as nodes, bonds as edges), which aligns well with Graph Neural Networks.
- The dataset is large enough to generate many structurally similar molecule pairs for cliff detection.
- It provides the HOMO–LUMO gap, which is the property used in our cliff definition.
- QM9 molecules contain up to 9 heavy atoms (C, N, O, F), enabling efficient GNN training on available hardware.

THE DATASET

How was the data collected?

QM9 was introduced by Raghunathan Ramakrishnan et al. (2014).
The molecules were:

- Systematically enumerated from chemical space.
- Geometry-optimized using Density Functional Theory (DFT).
- Evaluated using the B3LYP functional and 6-31G basis set.

Therefore, the molecular properties are obtained from quantum chemistry calculations rather than experimental measurements.

Ethical Considerations

- The dataset does not contain any personal, medical, or sensitive information.
- The molecules included are small organic compounds used for scientific benchmarking.
- Therefore, there are no privacy or direct ethical concerns associated with its use.

THE DATASET

Dataset Size

Our dataset consists of 50,000 molecule pairs, chosen randomly from the QM9 dataset.

Features

Raw Data: For each atom, we have

- Atomic symbol
- 3D position

Derived atom/molecule features:

- Atomic number
- Degree (how many bonds it has)
- Formal charge
- Aromatic or not aromatic
- Hybridization type (sp, sp^2, sp^3)
- Number of attached hydrogens

Derived bond features:

- Bond type (single, double, triple)
- Conjugated or not conjugated
- Is it part of a ring or not

Target

Binary label (0 or 1) indicating whether a pair of structurally similar molecules forms a property cliff or not.

THE DATASET

Pairwise Dataset Construction - How were molecule pairs created?

Step 1: Compute structural similarity for all the molecules.

- Morgan fingerprints are generated for each molecule using RDKit.
- Tanimoto similarity is computed to measure structural similarity between molecules.

Step 2: Generate candidate pairs

- For each molecule, the top-K most similar molecules are selected to create candidate pairs.
- This avoids comparing every molecule with every other molecule.

Step 3: Compute property difference

- For each pair of molecules, the HOMO–LUMO gap difference is calculated:

$$\Delta gap = | gapA - gapB |$$

FEATURE PREPROCESSING

Feature Construction

- Raw molecular structures were provided in .xyz format.
- We used RDKit to convert each molecule into a graph representation.
- From this, we extracted:
  1. Atom-level features (atomic number, degree, formal charge, aromaticity, hybridization, number of hydrogens attached).
  2. Bond-level features (bond type, conjugation, part of a ring).
- HOMO-LUMO gap values were standardized using z-score normalization.
- Cliff labels were generated using raw HOMO-LUMO gap differences.

FEATURE PREPROCESSING

Missing Data Handling

- The dataset contained no missing values.
- Engineered features data produced no missing values.
- Therefore, no interpolation or imputation methods were required.

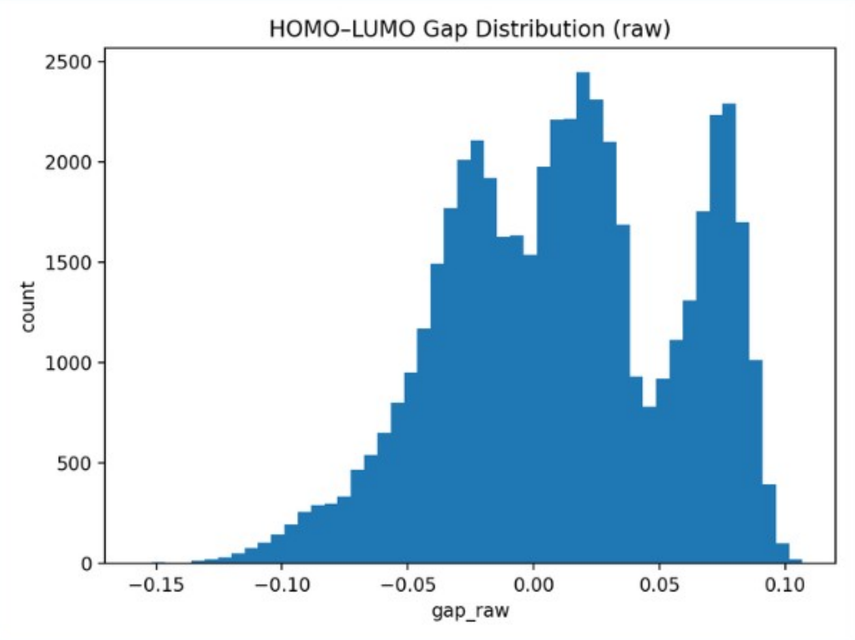
Dimensionality Reduction

- Dimensionality reduction techniques like PCA, LDA were not applied.
- Since the graph feature space was already compact, dimensionality reduction techniques were not required.

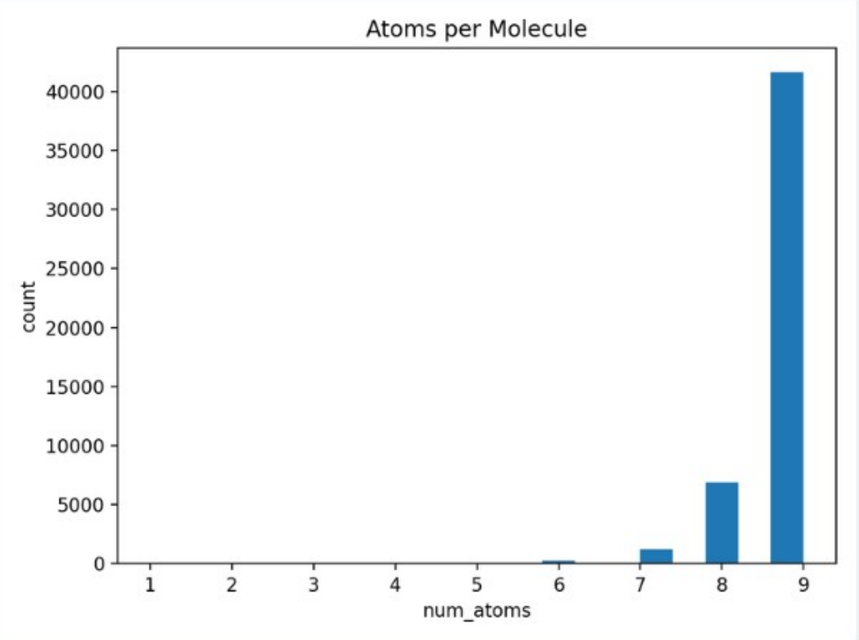
Feature Importance Analysis

- We computed Pearson correlation coefficients between molecular features and the HOMO–LUMO gap to analyze linear relationships between structural descriptors and target property.
- Features related to sp2 atoms, double bonds, and conjugation showed negative correlation with the gap and those related to sp3 atoms and single bonds showed positive correlation.

THE DATA

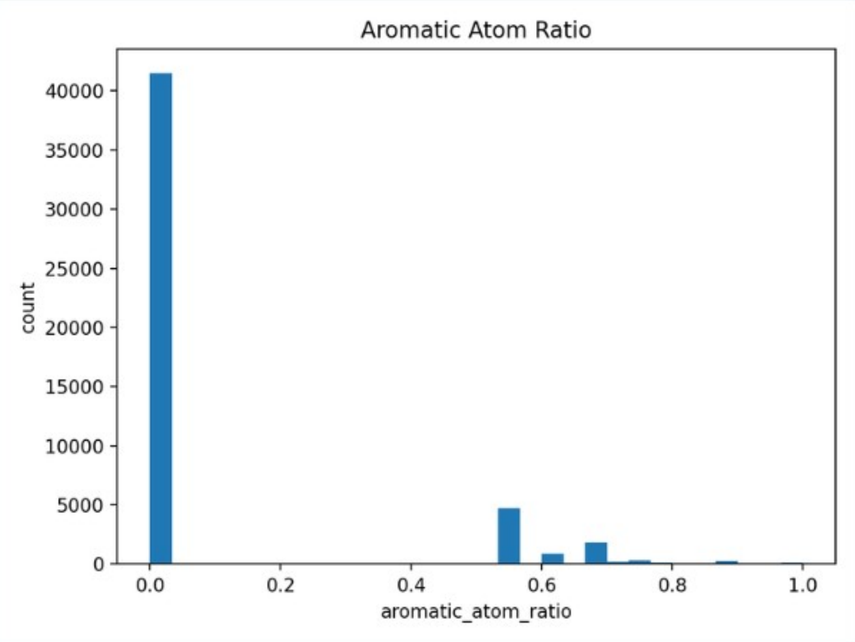


- This graph shows the distribution of the target variable across 50,000 molecule pairs.
- Mean ≈ 0.0119, Std ≈ 0.0467, Range ≈ [−0.157, 0.107]
- It helps understand the range and variability of the target variable and check for outliers before we standardize the data.



- This graph shows:
  - Most molecules contain 7–9 atoms.
  - The dataset is restricted to small organic molecules (≤9 heavy atoms).
- It demonstrates the structural complexity of the dataset.

THE DATA



	Correlation coefficient
no. of sp3 atoms	0.688599929117741
avg. no. of hydrogens per atom	0.667649894196232
no. of double bonds in molecule	-0.660501584017812
no. of sp2 atoms	-0.629764309332082
no. of single bonds in molecule	0.562341251593809
ratio of bonds in conjugation	-0.485618794902218
no. of conjugated bonds	-0.479962867613172
avg. atomic no. of molecule	-0.418808656126541
avg. degree of molecule	0.3514510396368

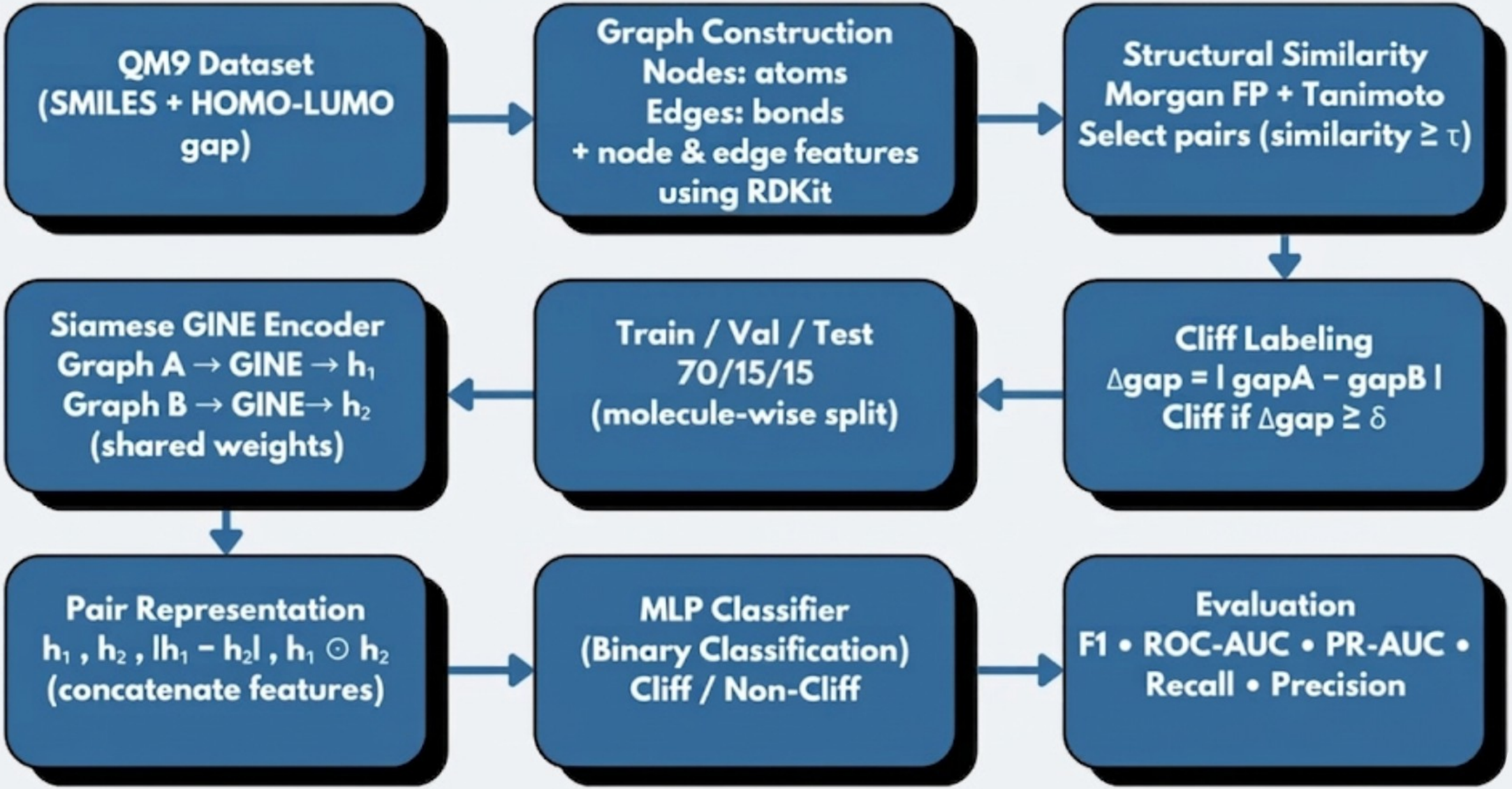
- Most molecules are non-aromatic, while a smaller subset contains aromatic systems.
- This helps analyze the influence of conjugation on HOMO-LUMO behavior.

Pearson correlation analysis was used to study relationships between structural descriptors and HOMO-LUMO gap values.

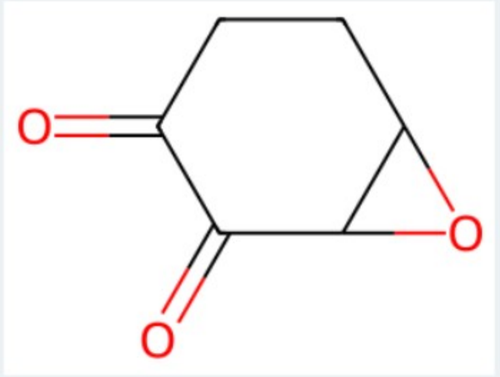
ML METHODS

Model	Role in Project	Why It Was Used	How It Works
Shared-Encoder Graph Neural Network	Final model	Learns molecular structure directly from atom-bond graphs and compares both molecules in the same embedding space. Chosen because property cliffs depend on subtle structural differences.	Converts each molecule into a graph of atoms and bonds. Both molecules pass through the same GNN encoder to generate embeddings. The embeddings are compared using concatenation, absolute difference, and element-wise product and then passed to a classifier.
Two-Encoder Graph Neural Network	Benchmark 1	Uses separate GNN encoders for each molecule. Used to test whether independent molecular representations improve performance over shared encoding.	Converts both molecules into graph representations. Each molecule is processed by a separate GNN encoder. The two embeddings are then combined and classified as property cliff or non-cliff.
TF-IDF SMILES + Logistic Regression	Benchmark 2	Treats SMILES strings as text and uses character n-gram features. Used as a simple classical baseline.	Treats the SMILES pair as a text sequence. Character n-gram TF-IDF features are extracted, and logistic regression predicts whether the pair is a property cliff.
Morgan Fingerprint + Random Forest	Benchmark 3	Uses standard molecular fingerprints with a tree-based classifier. Used as a strong classical cheminformatics benchmark.	Converts each molecule into a Morgan fingerprint, which encodes circular substructures. Pairwise fingerprint features are passed into a Random Forest classifier made of multiple decision trees.
Morgan Fingerprint + MLP	Benchmark 4	Uses Morgan fingerprints with a neural network classifier. Used to compare graph-learned representations against fixed molecular descriptors.	Converts molecules into Morgan fingerprints. The fingerprint pair features are given to a feed-forward neural network, which learns nonlinear patterns for cliff classification.

METHODOLOGY PIPELINE

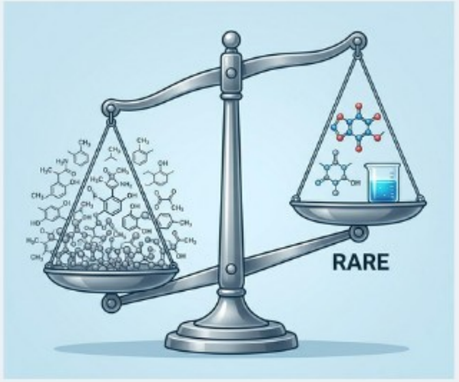

QM9 Dataset
(SMILES + HOMO-LUMO gap)
Graph Construction
Nodes: atoms
Edges: bonds
+ node & edge features using RDKit
Structural Similarity
Morgan FP + Tanimoto
Select pairs (similarity ≥ τ)
Cliff Labeling
Δgap = | gapA − gapB |
Cliff if Δgap ≥ δ
Train / Val / Test
70/15/15
(molecule-wise split)
Siamese GINE Encoder
Graph A → GINE → h1
Graph B → GINE→ h2
(shared weights)
Pair Representation
h1 , h2 , |h1 − h2| , h1 ⊙ h2
(concatenate features)
MLP Classifier
(Binary Classification)
Cliff / Non-Cliff
Evaluation
F1 • ROC-AUC • PR-AUC •
Recall • Precision

CHALLENGES



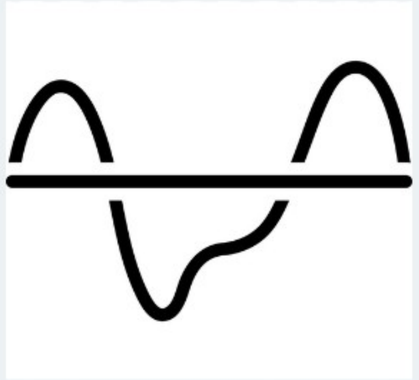
Molecular graph representation

Molecules had to be converted from SMILES strings into graph structures with atom and bond features. RDKit and PyTorch Geometric were used to build graph inputs for the GNN models.



Class imbalance

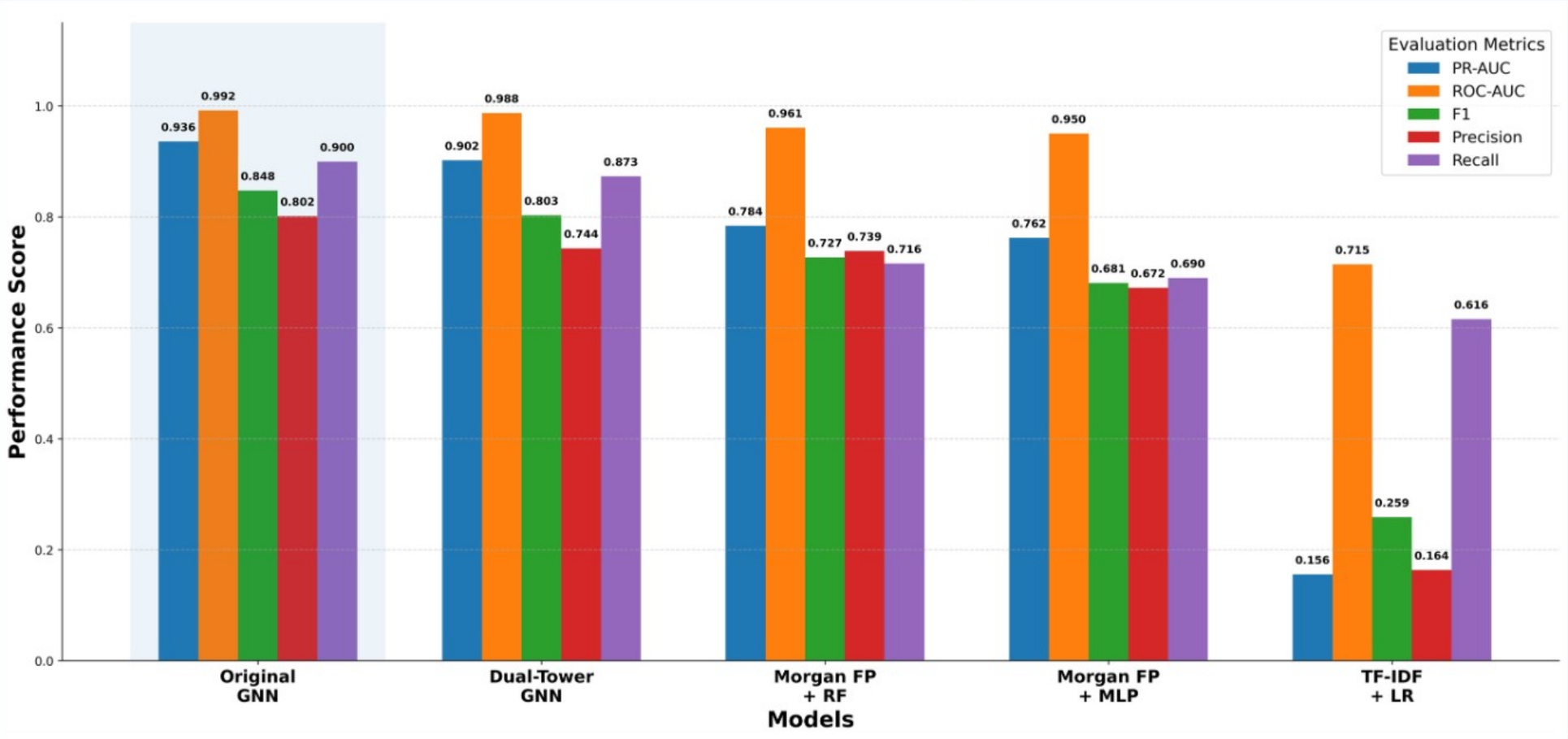
Property cliff pairs were much fewer than non-cliff pairs. This was handled using weighted binary cross-entropy loss and evaluation metrics such as PR-AUC, precision, recall, and F1 rather than relying only on accuracy.



Threshold selection

The classification threshold strongly affected precision and recall. Threshold tuning was performed on the validation set to choose an operating point that balanced F1, precision, and recall.

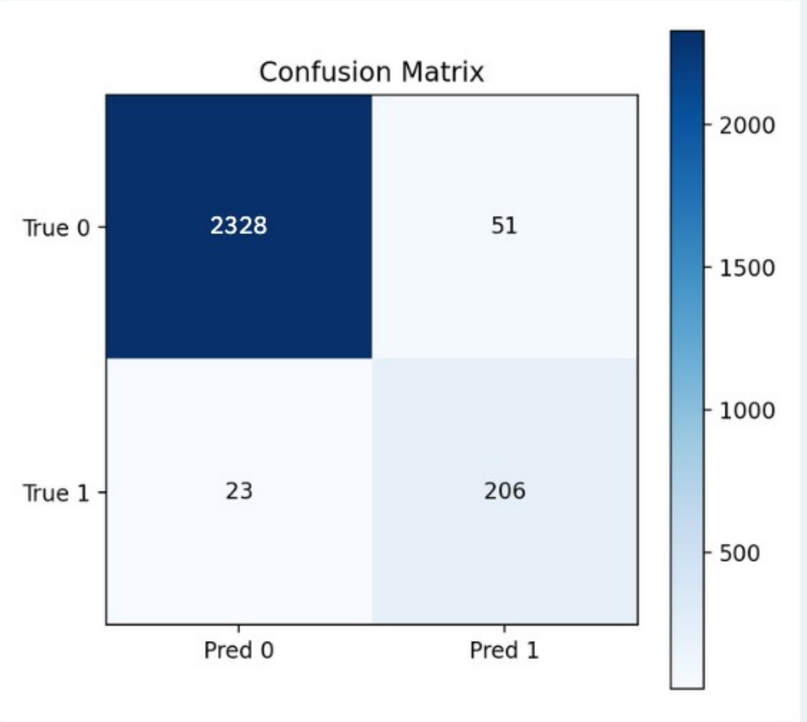
MODEL PERFORMANCE COMPARISON



KEY INSIGHT: ORIGINAL GNN ACHIEVES THE BEST PERFORMANCE ACROSS ALL THE METRICS

PERFORMANCE METRICS

Metric	Value	Interpretation
PR-AUC	**0.9361**	Excellent balance between precision and recall across thresholds
ROC-AUC	**0.9922**	Near-perfect ability to distinguish positive and negative classes
F1 Score	**0.8477**	Strong overall classification performance
Precision	**0.8016**	~80% of predicted positives were correct
Recall	**0.8996**	~90% of actual positives were successfully detected



- **HIGH PR-AUC AND ROC-AUC INDICATE STRONG SEPARATION BETWEEN CLIFF AND NON-CLIFF PAIRS.**
- **BALANCED PRECISION AND RECALL INDICATE STABLE CLASSIFICATION PERFORMANCE.**

- **2328 NON-CLIFF PAIRS AND 206 CLIFF PAIRS WERE CLASSIFIED CORRECTLY**
- **FALSE POSITIVE AND FALSE NEGATIVE COUNTS REMAINED RELATIVELY LOW**
- **THE MODEL MAINTAINED HIGH RECALL WHILE PRESERVING REASONABLE PRECISION**

DEPLOYMENT SCALABILITY CHALLENGES

Although this solution cannot be deployed within the current Plaksha project environment, it has strong potential for deployment in real-world industrial settings, particularly in pharmaceutical research, molecular screening, and lead optimization.

What problems appear at scale?

- **Computational Scalability:** Large molecular libraries may require GPU-based inference pipelines.
- **Large-Scale Data Handling:** Industrial datasets may contain millions of compounds and molecule pairs.
- **Real-World Data Quality:** Real-world datasets may include noisy labels, duplicates, and chemically invalid molecular structures.
- **Workflow Integration:** The solution would need to integrate with existing cheminformatics databases and industrial workflows.

THANK YOU