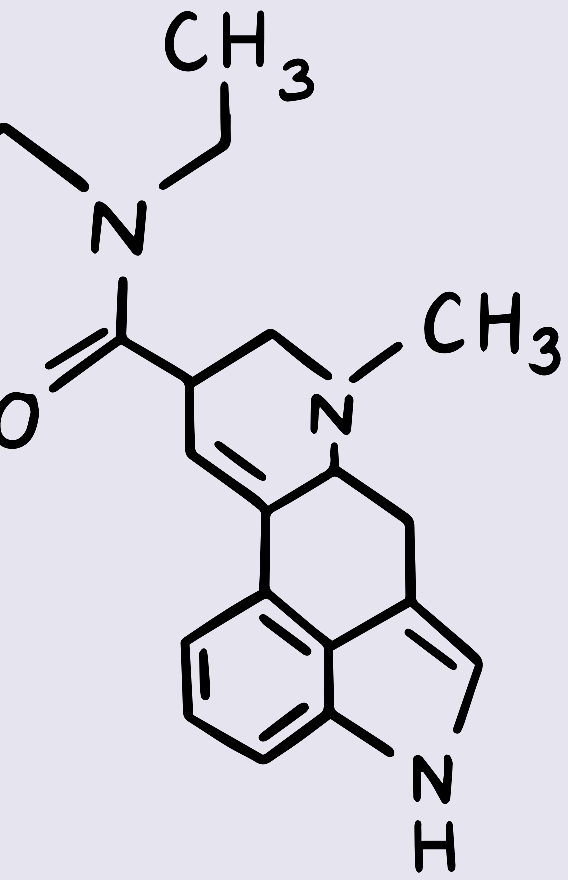
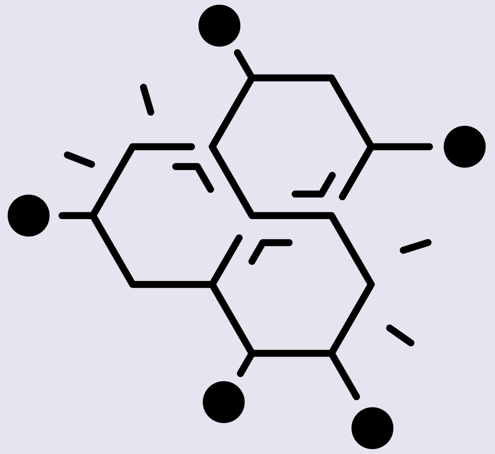




B.TECH PROJECT 2024-25

COMPARING VAE AND GAN MODELS FOR MOLECULAR SMILES GENERATION AND PROPERTY PREDICTION USING GNNS

Name of SIT Mentor : Dr Shruti Patil
Dr Prachi Kadam

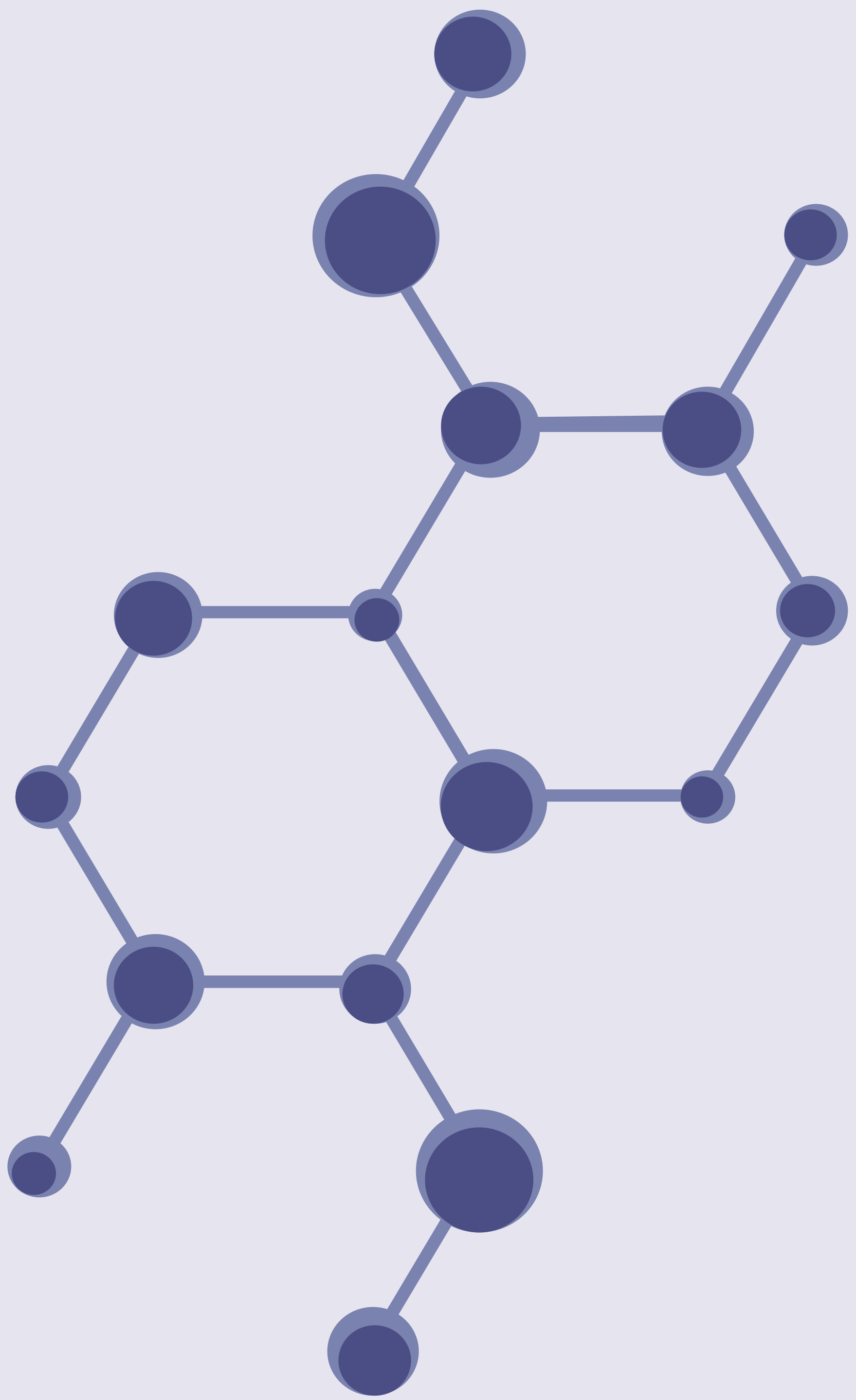
Name of DRDO Mentor : Dr Sunil Jaiswal

PREPARED BY

Arjun Tyagi(21070126020)
Kaustubh Raykar(21070126048)
Nachiket Ropia(21070126056)

OVERVIEW

- Introduction
- Problem Statement and Objectives
- Dataset
- Literature Review
- Methodology
- Results and Discussion
- Comparative Analysis
- Conclusion and Future Scope
- Acknowledgments
- References



INTRODUCTION

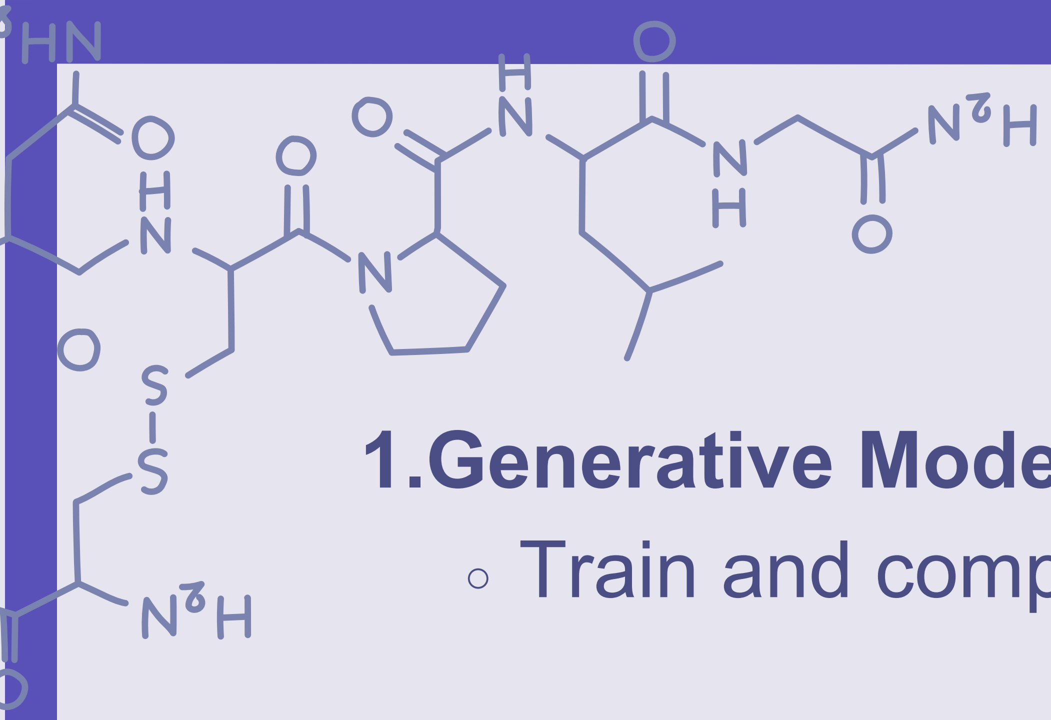
- **Foundation in Molecular Innovation:** The project underpins advancements in pharmaceuticals and materials science by developing novel molecules with specific properties.

- **Shift to Computational Methods:** Traditional experimental approaches in molecular design are being replaced by efficient computational techniques, significantly reducing time and costs.

- **Deep Learning Revolution:** The introduction of deep learning technologies like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) has dramatically changed how molecular structures are generated and properties predicted.

- **Graph Neural Networks (GNNs):** These networks advance property predictions by processing molecular graphs directly, enabling more precise predictions across chemical behaviors.

- **Study Goals:** To integrate and compare the effectiveness of VAEs, GANs, and GNNs to accelerate molecular discovery and streamline the design process, reducing the reliance on trial-and-error methods.



OBJECTIVES

1.Generative Model Development:

- Train and compare VAEs and GANs for generating molecular SMILES strings.

2.Property Prediction Accuracy:

- Evaluate GCN and GIN performance on predicting molecular properties for generated molecules.

3.Integration Framework:

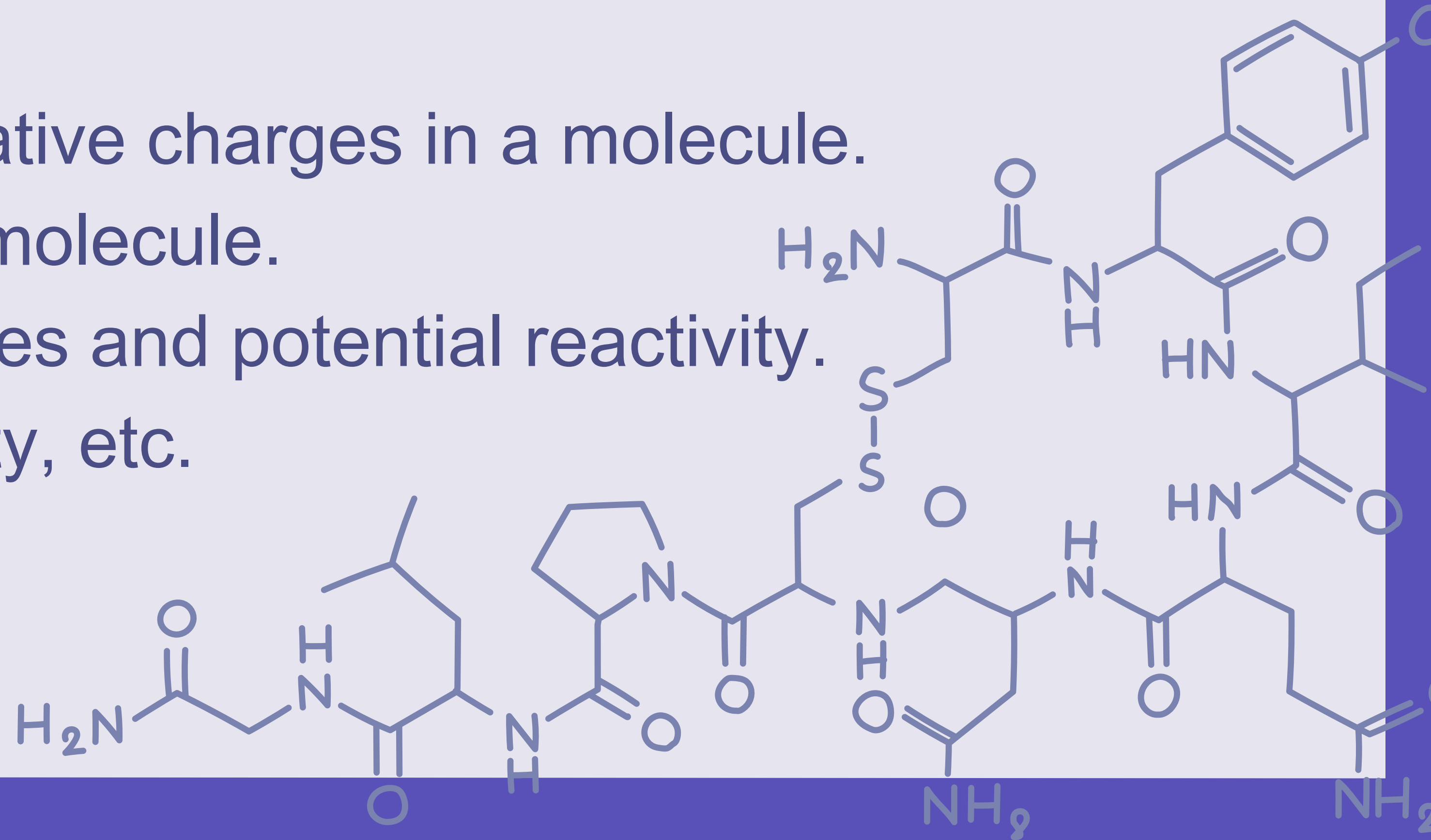
- Develop a robust workflow to combine generative models with GNNs for end-to-end molecular design.

4.Comprehensive Evaluation:

- Compare VAEs and GANs on metrics such as validity, uniqueness, novelty, Fréchet ChemNet Distance (FCD), and internal diversity.

DATASET

- **Size:** Contains approximately 134,000 small organic molecules.
- **Features**
  - **Representation:** Encodes molecular structures into linear text strings, facilitating easy input into machine learning models.
  - **Usage:** Serves as the primary input for generative models like VAEs and GANs to create new molecular structures.
- **19 Quantum Chemical Properties:**
  - **Dipole Moment:** Measures the separation of positive and negative charges in a molecule.
  - **Molecular Energy:** Indicates the stability and reactivity of the molecule.
  - **HOMO-LUMO Gap:** Reflects the molecule's electronic properties and potential reactivity.
  - **Other Properties:** Includes enthalpy, free energy, heat capacity, etc.



LITERATURE REVIEW

Literature review table

METHODOLOGY

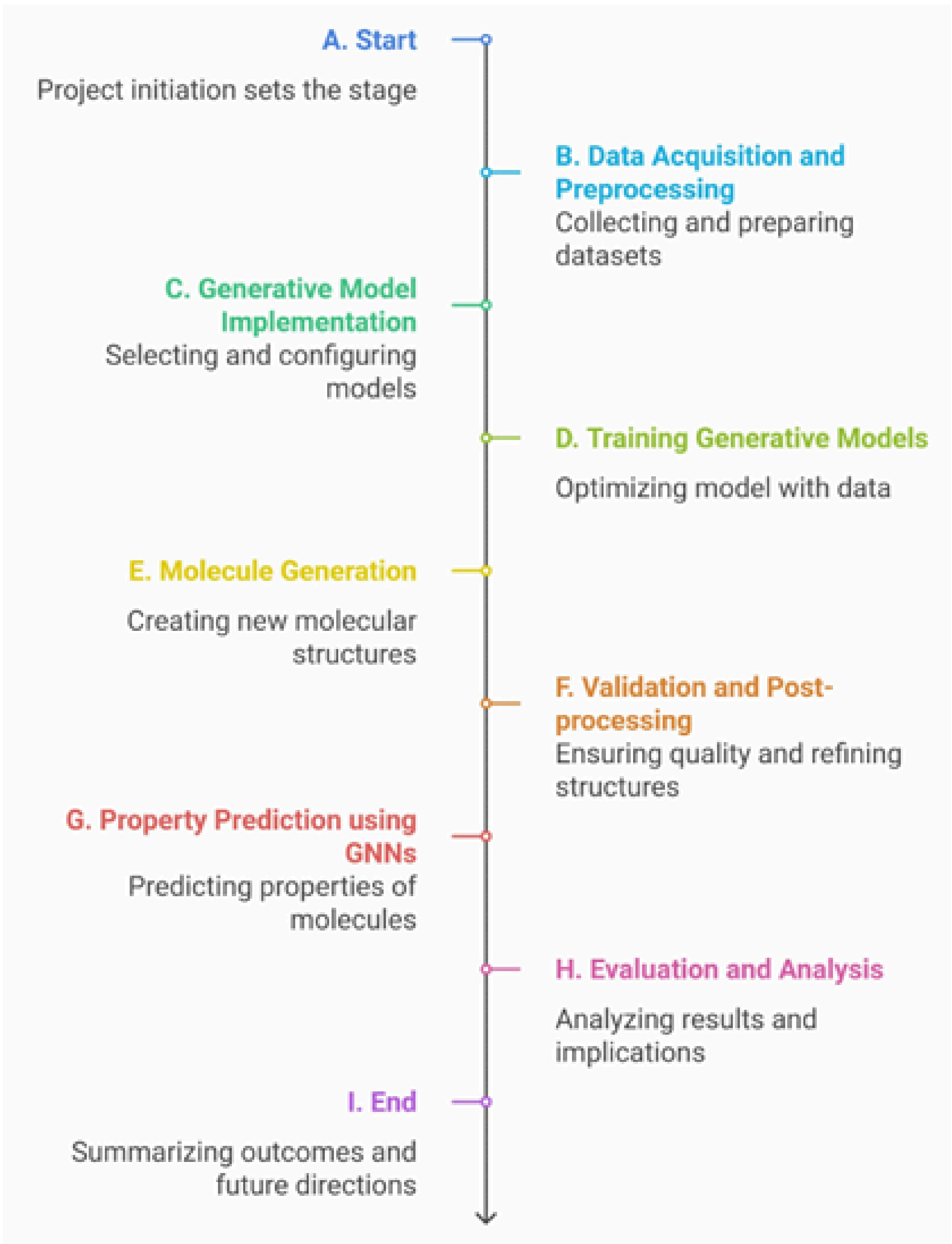
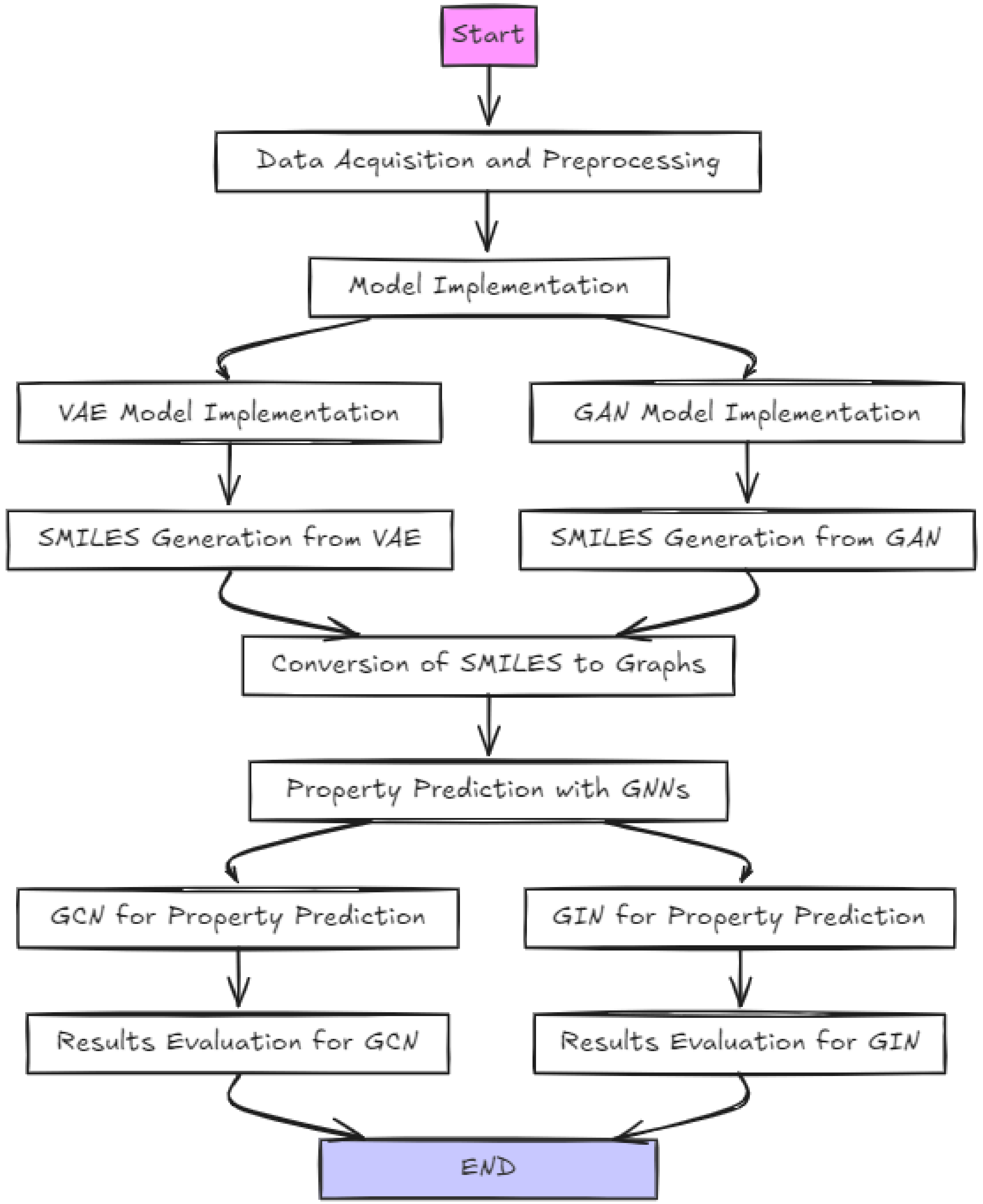
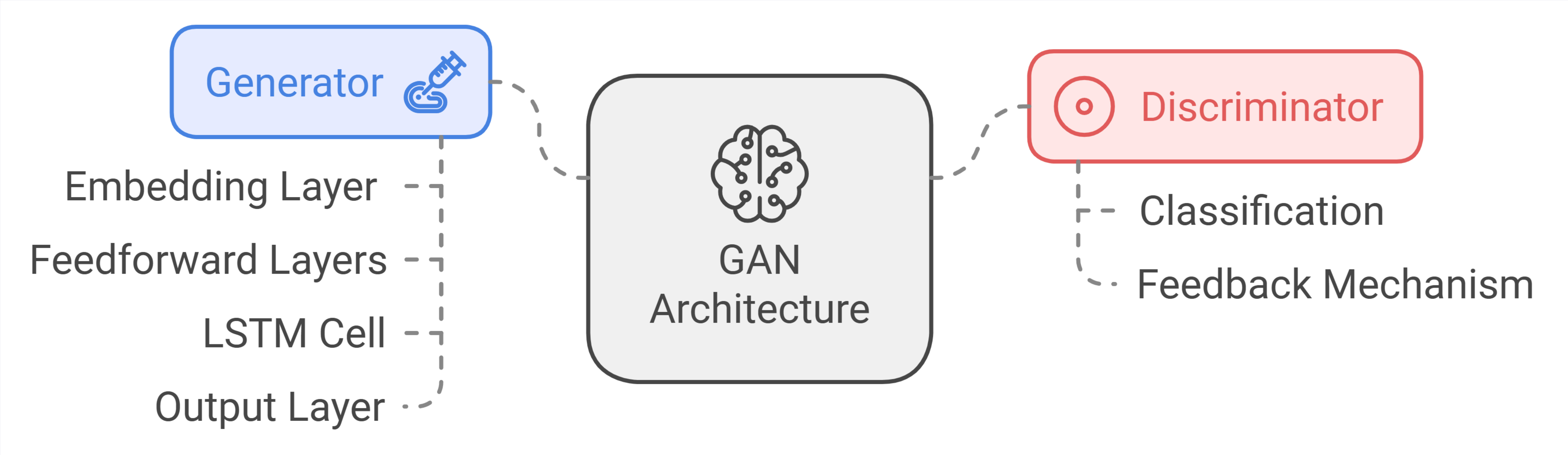


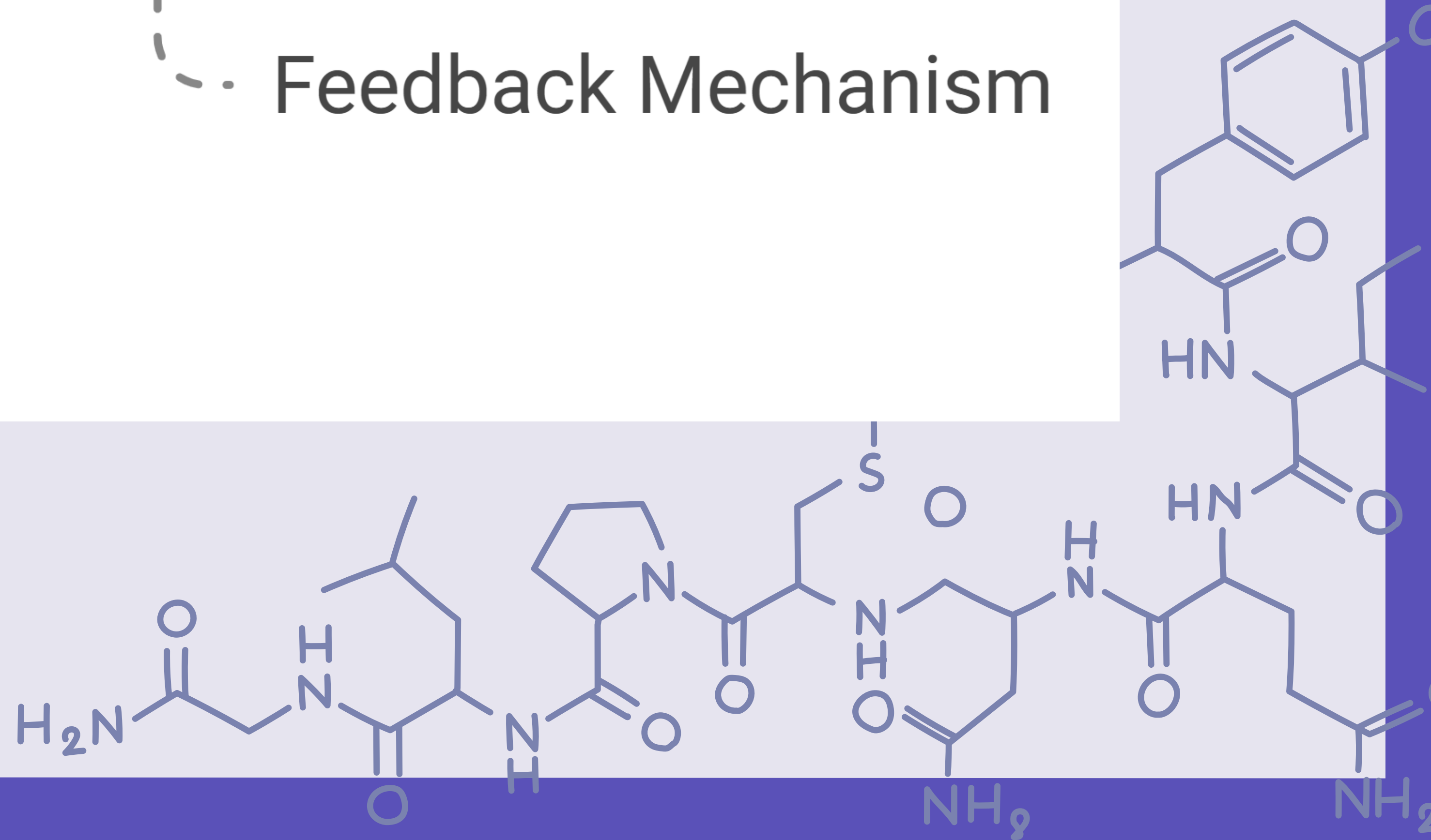
Fig1.2 Project Pipeline



GANS AND VAE FOR GENERATING NOVEL SMILES

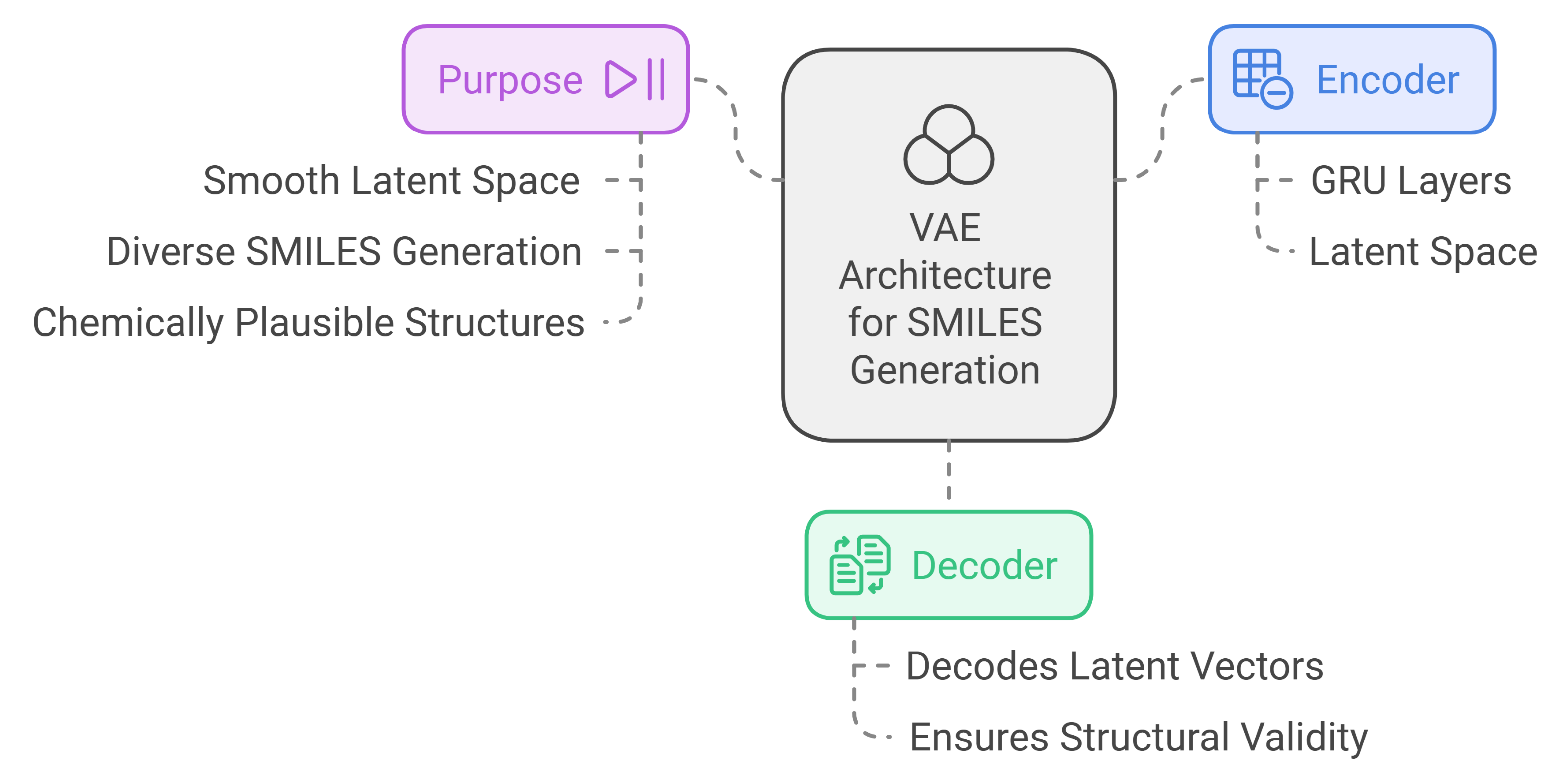
GANS Architecture for SMILES Generation

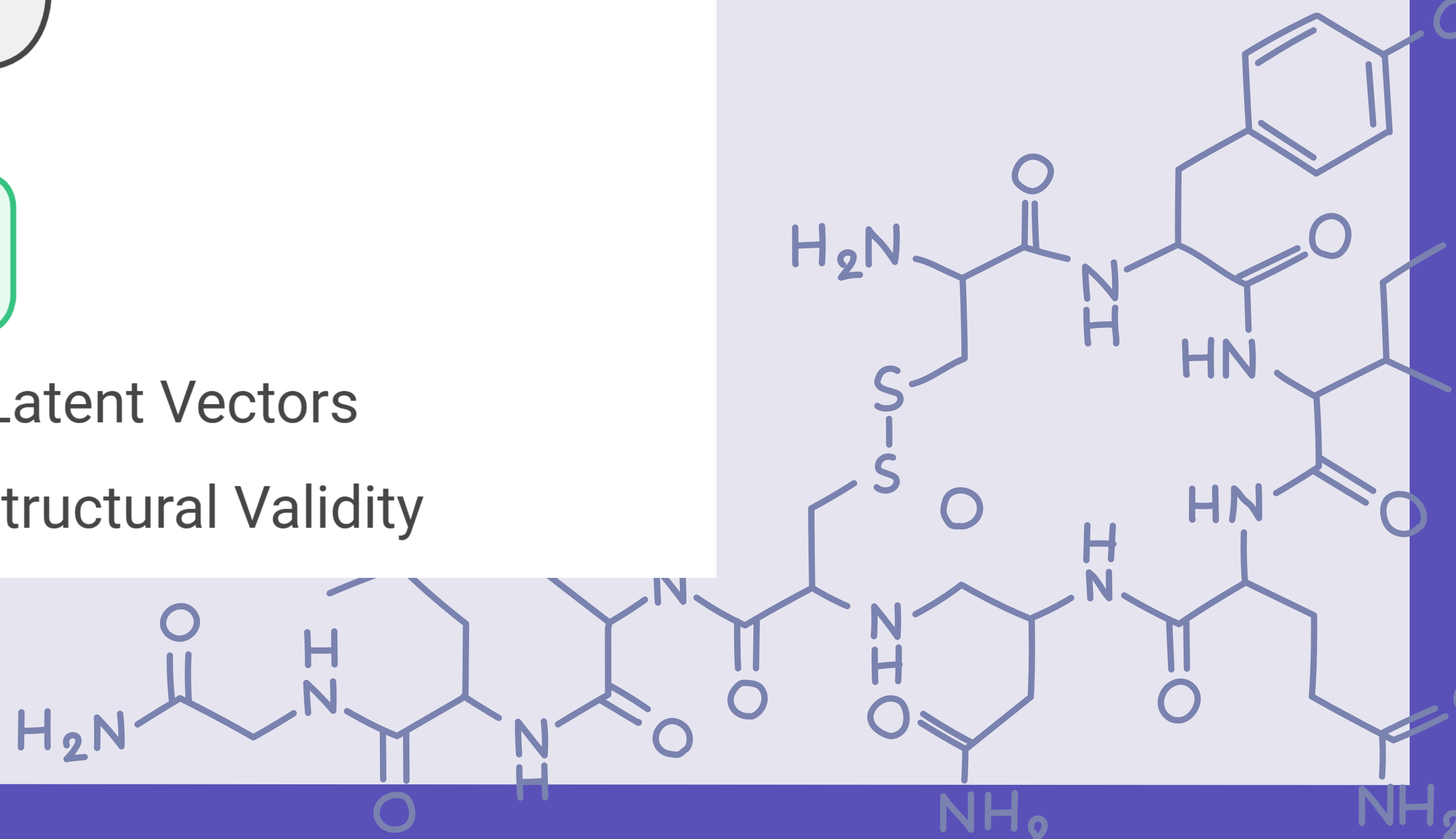




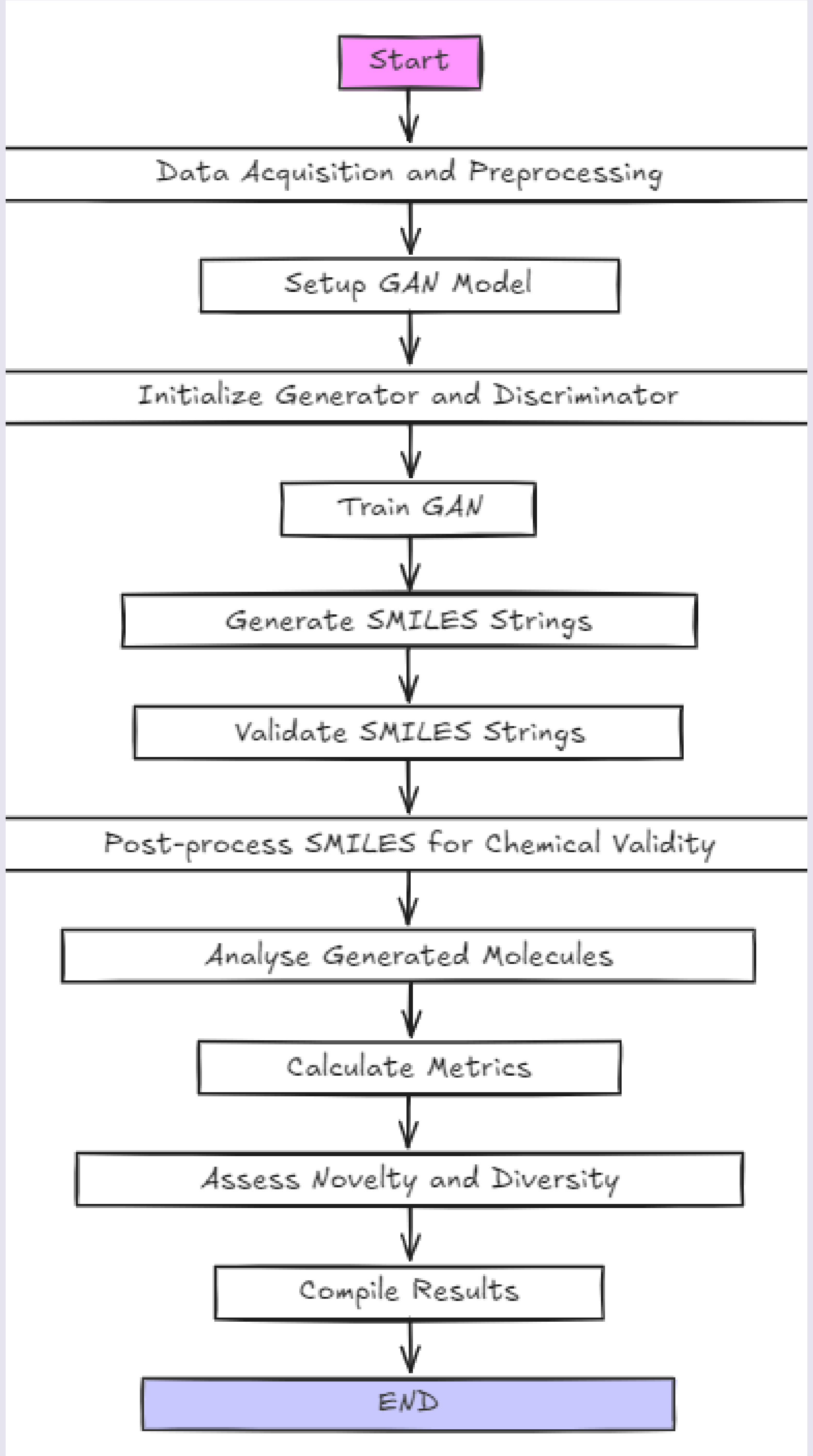
GANS AND VAE FOR GENERATING NOVEL SMILES

VAE Architecture for SMILES Generation

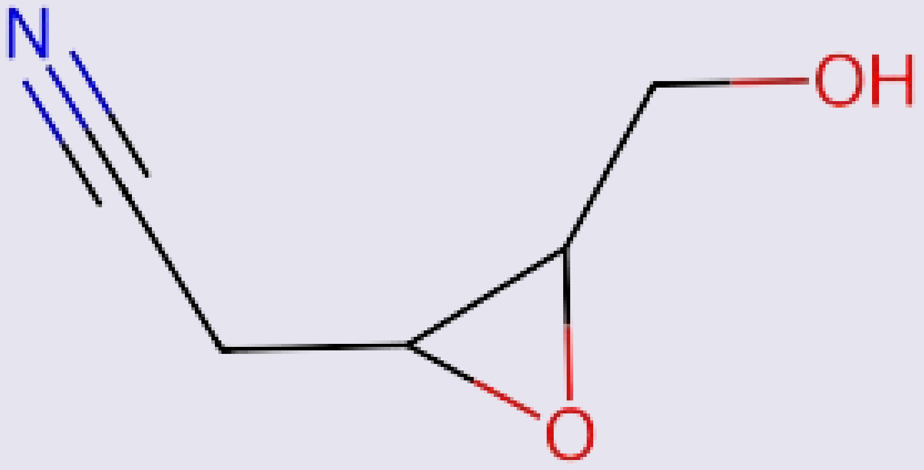


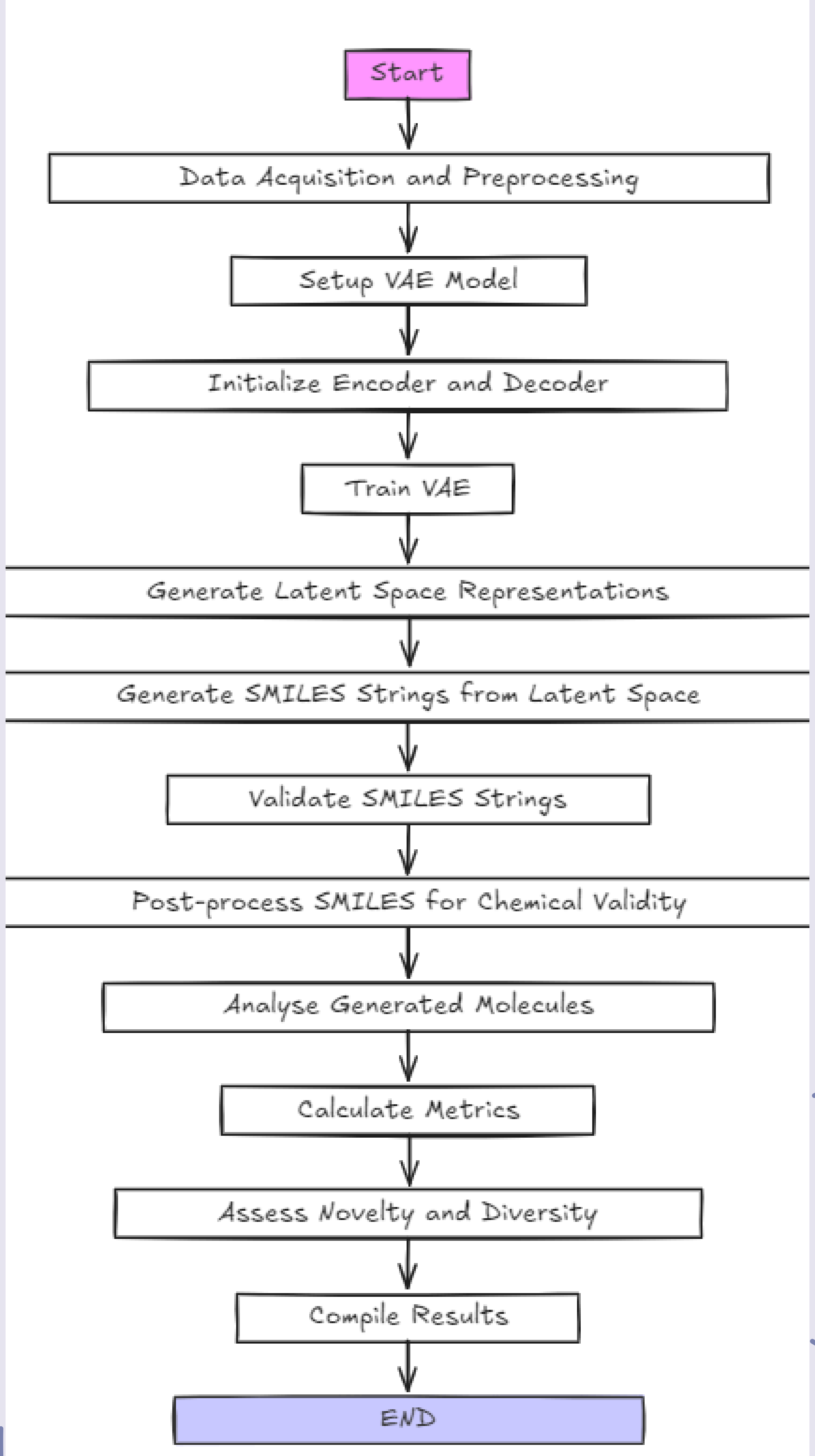


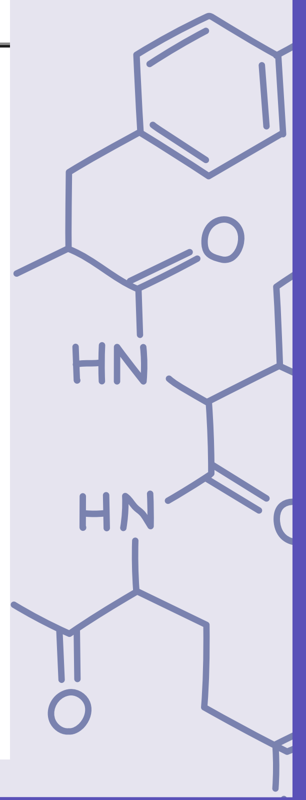
METHODOLOGY



GAN VS VAE

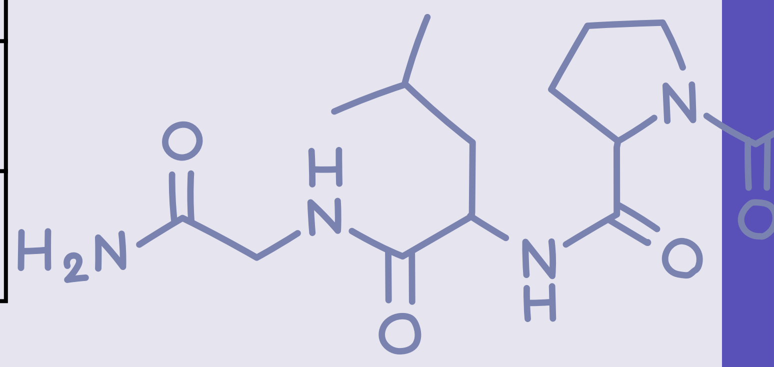




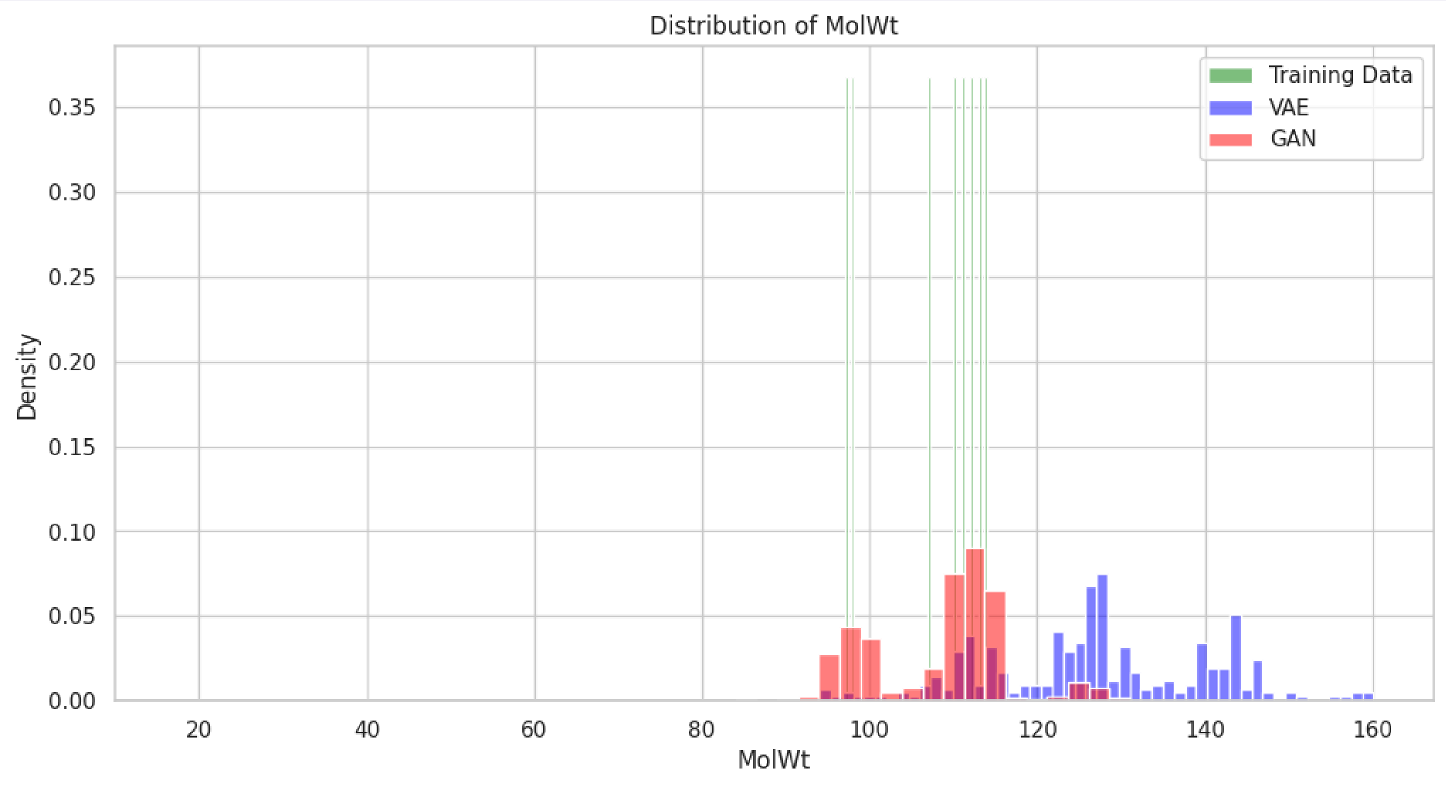


RESULTS GANS VS VAE

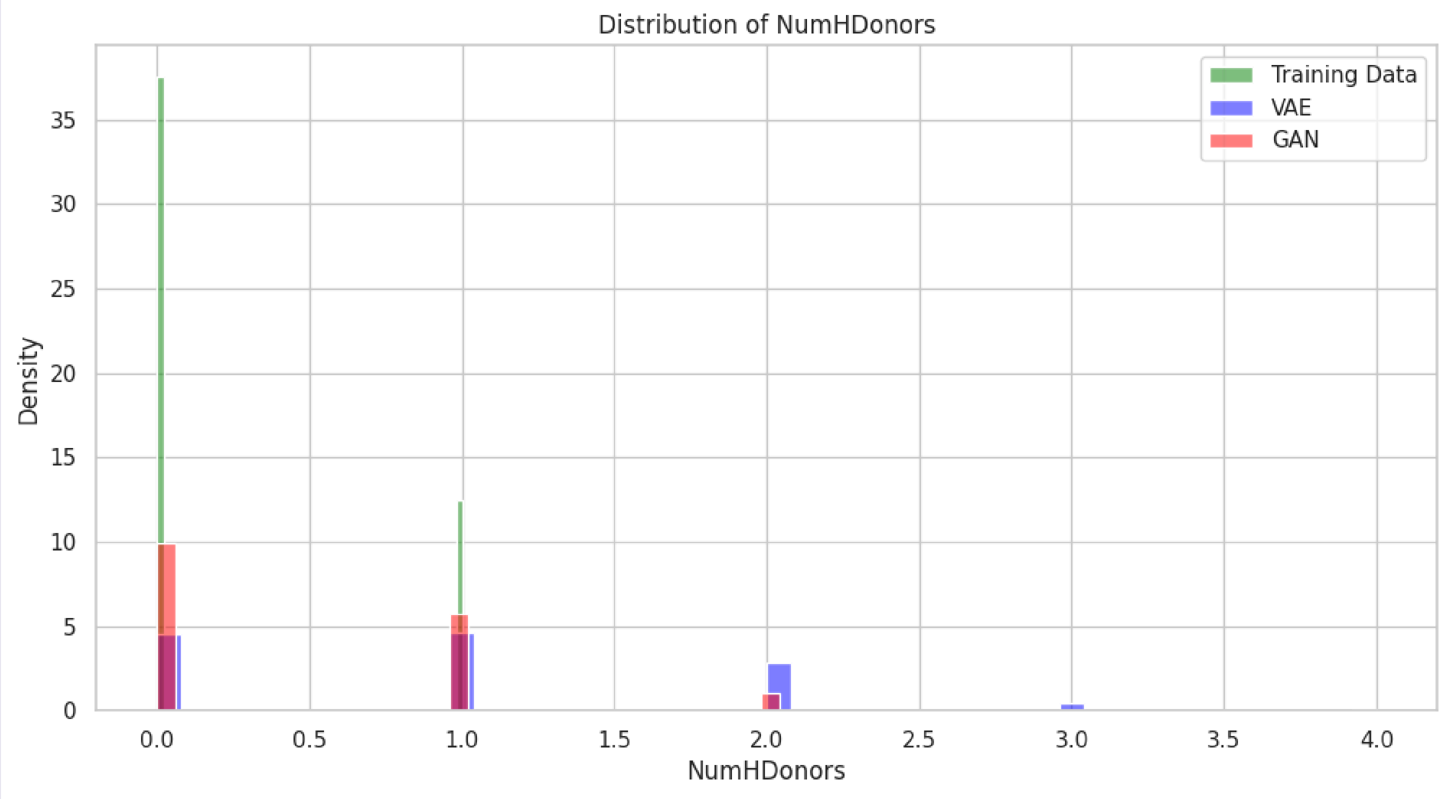
Metric	VAE	GAN
Fréchet ChemNet Distance(similarity between the distributions of generated molecules)	47.773	118.981
Average Tanimoto Similarity(structural similarity between molecules)	0.335	0.123
Internal Diversity(structural variety within a set of generated molecules)	0.937	0.325
Validity Rate (%)	100.0	100.0
Uniqueness Rate (%)	99.678	75.327
Novelty Rate (%)	99.678	74.924
Average MolWt	126.964	108.390
Average LogP	0.412	0.088
Average NumHDonors	0.967	0.473
Average NumHAcceptors	2.260	2.061



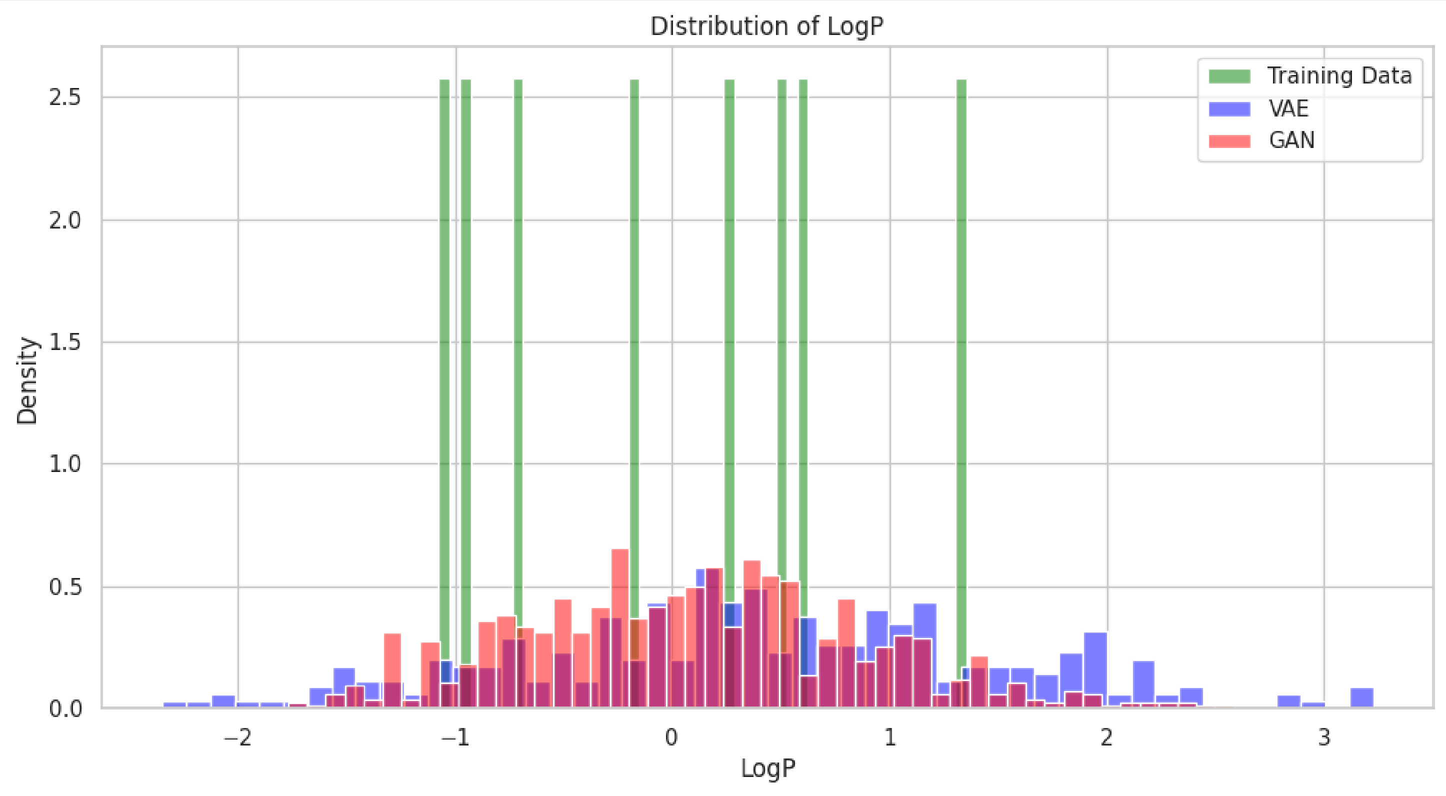
RESULTS GANS VS VAE



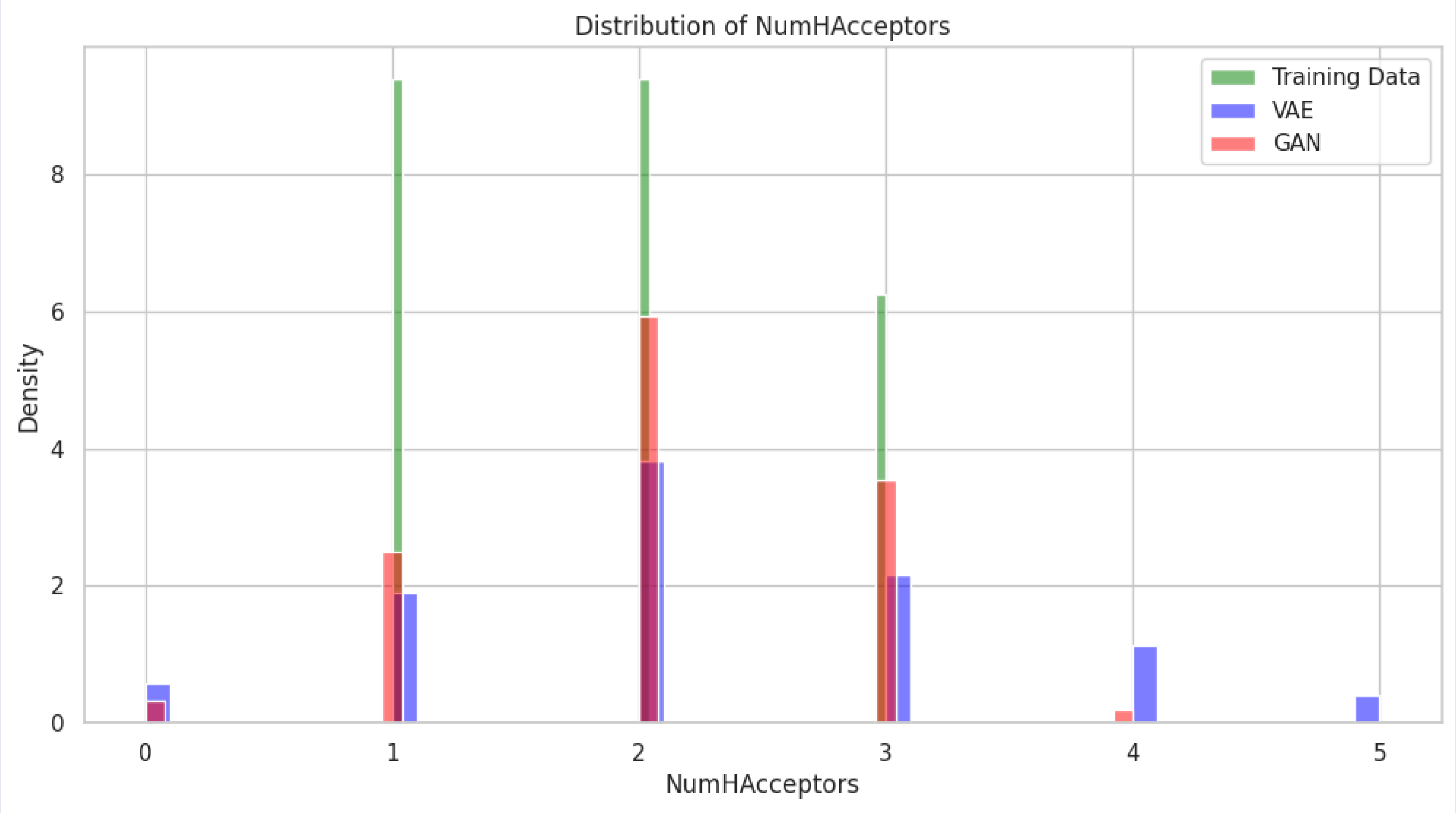
Molecular weight of the generated smile structures



Ability to donate hydrogen in formation of Hydrogen bonds



Representation of Hydrophilicity

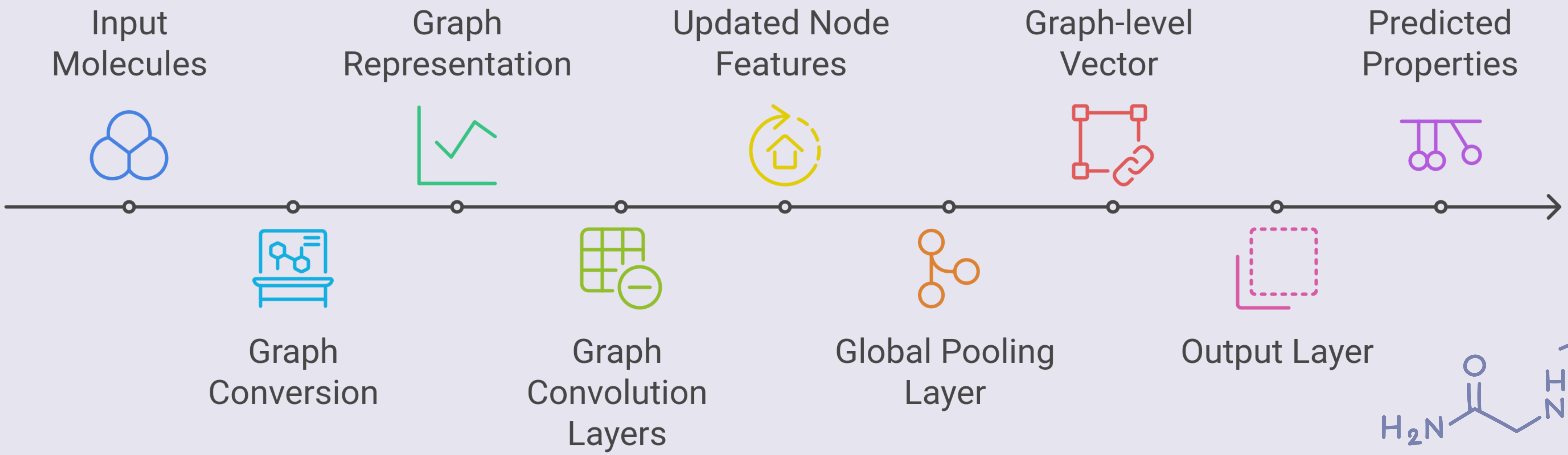


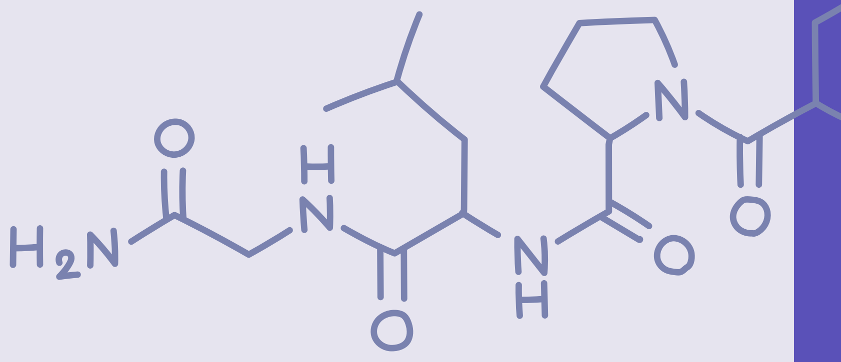
Acceptance in formation of Hydrogen bonds

GCN AND GIN FOR PROPERTY PREDICTION

GCN Architecture for Property Prediction

- **Input Representation:** Molecules are converted into graphs (atoms as nodes, bonds as edges).
- **Graph Convolution Layers:** Aggregate neighboring node features to update each node's representation.
- **Global Pooling Layer:** Combines node features into a single graph-level vector.
- **Output Layer:** Fully connected layers predict properties like dipole moment, energy gap, etc.

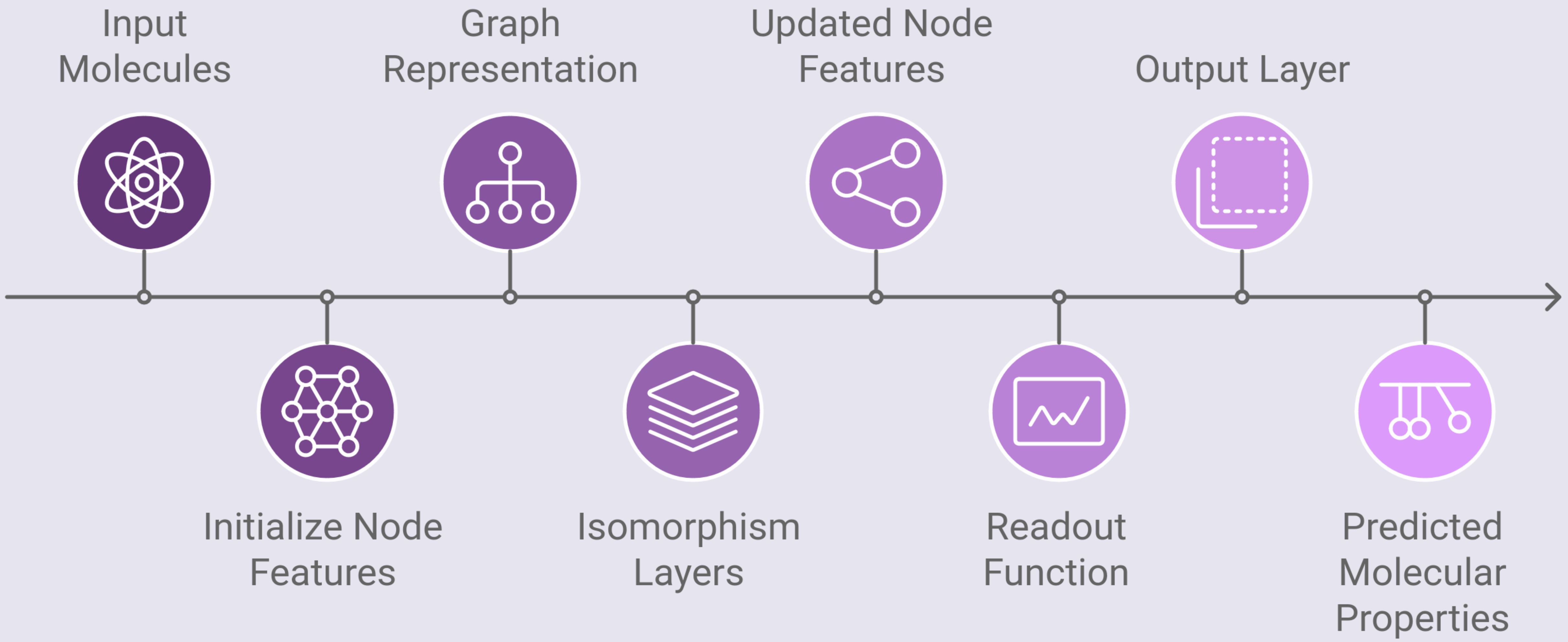


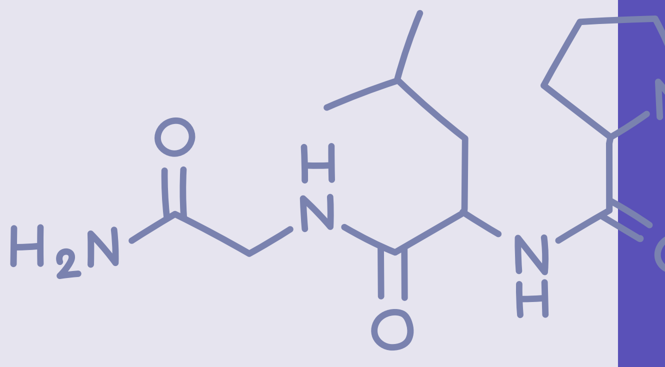


GCN AND GIN FOR PROPERTY PREDICTION

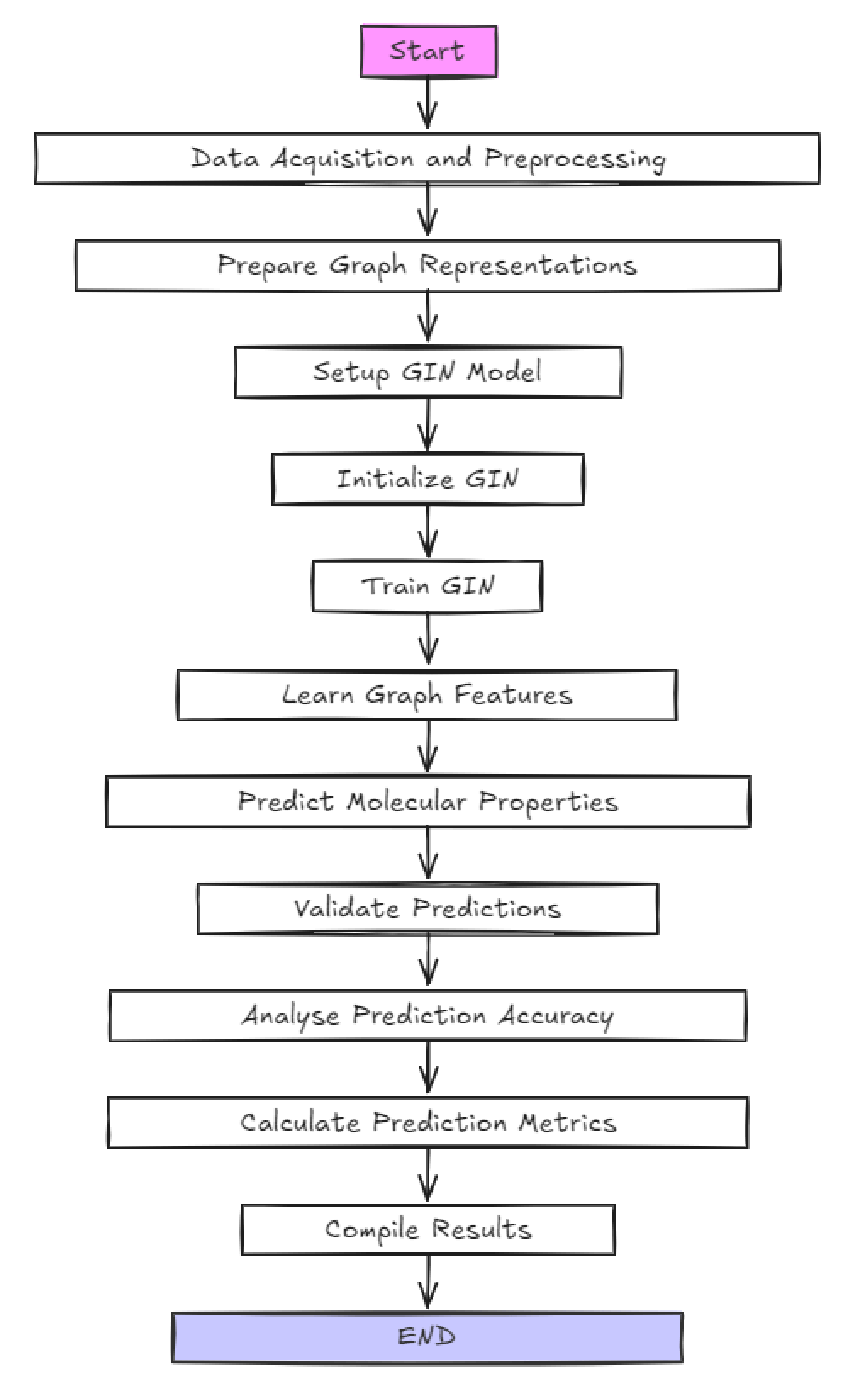
GIN Architecture for Property Prediction

- **Node Features:** Initialized with atomic properties.
- **Isomorphism Layers:** Use a learnable aggregation for feature updates.
- **Readout Function:** Combines updated node features into a graph-level representation.
- **Output Layer:** Predicts molecular properties.

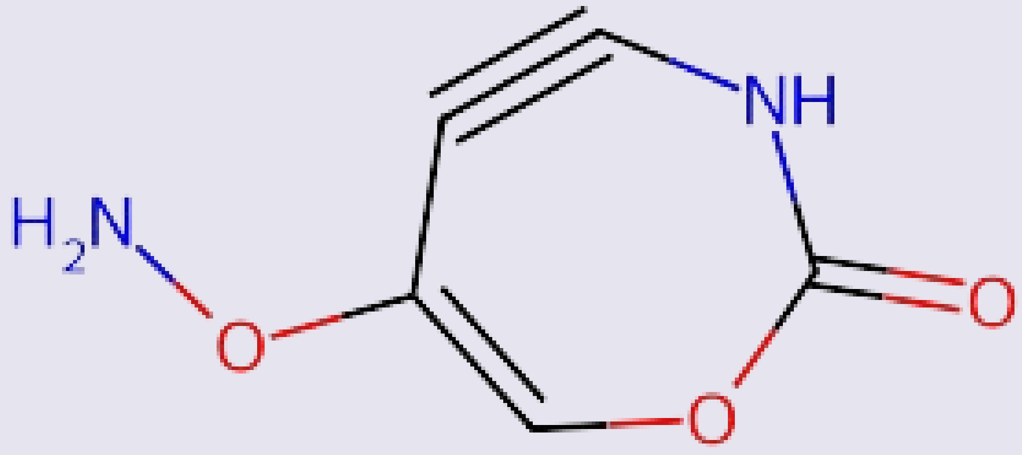


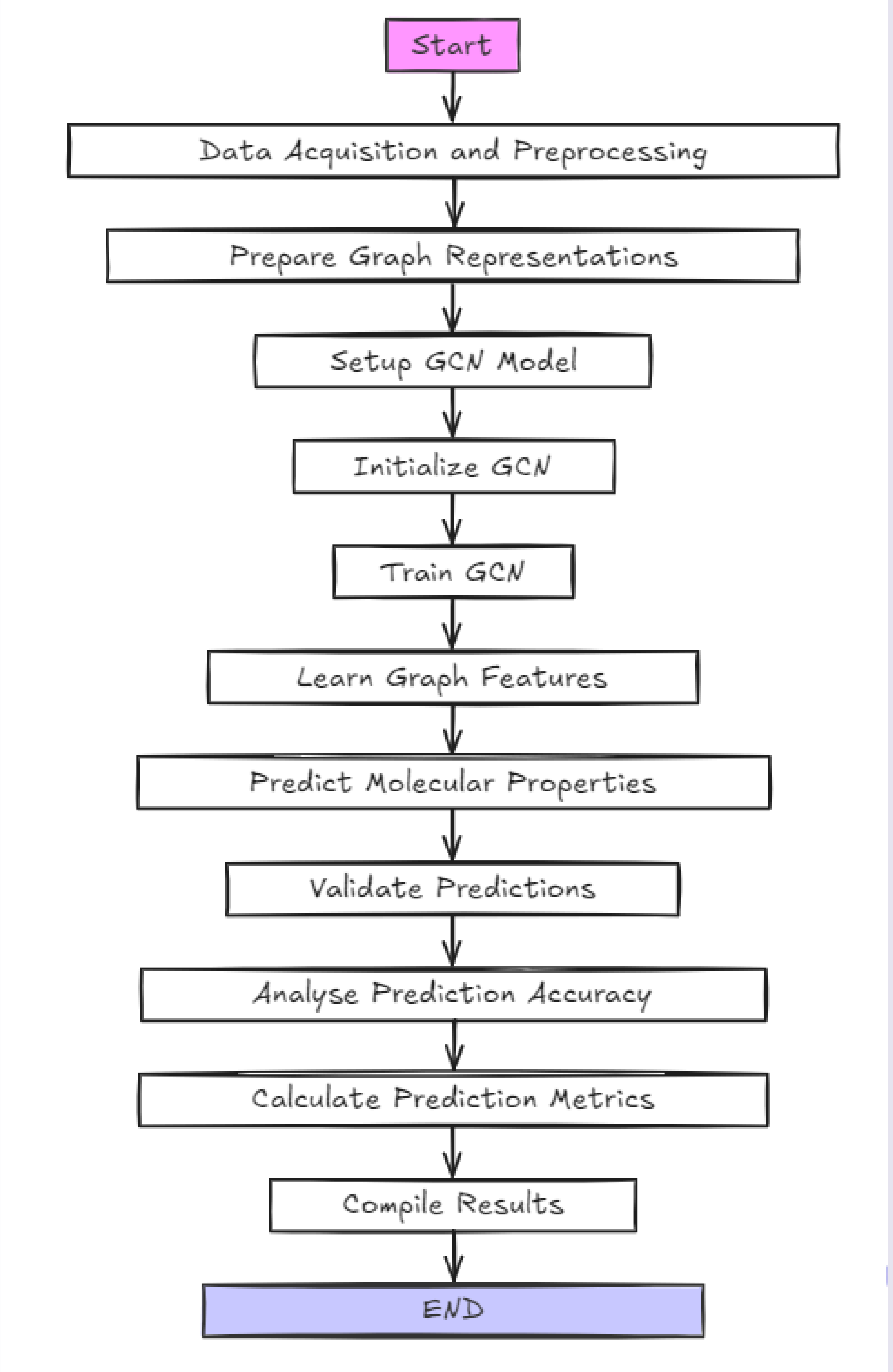


METHODOLOGY

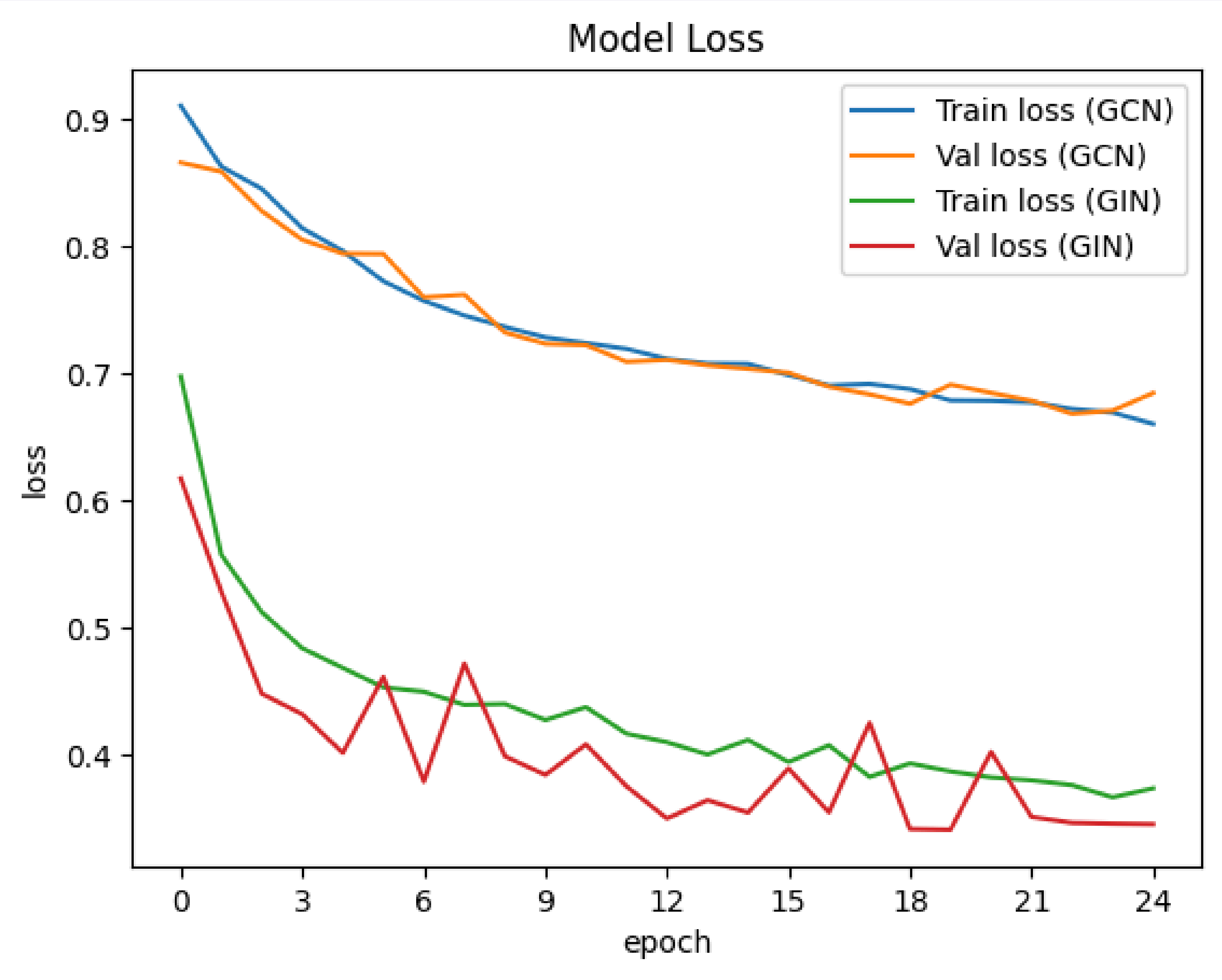


GIN VS GCN

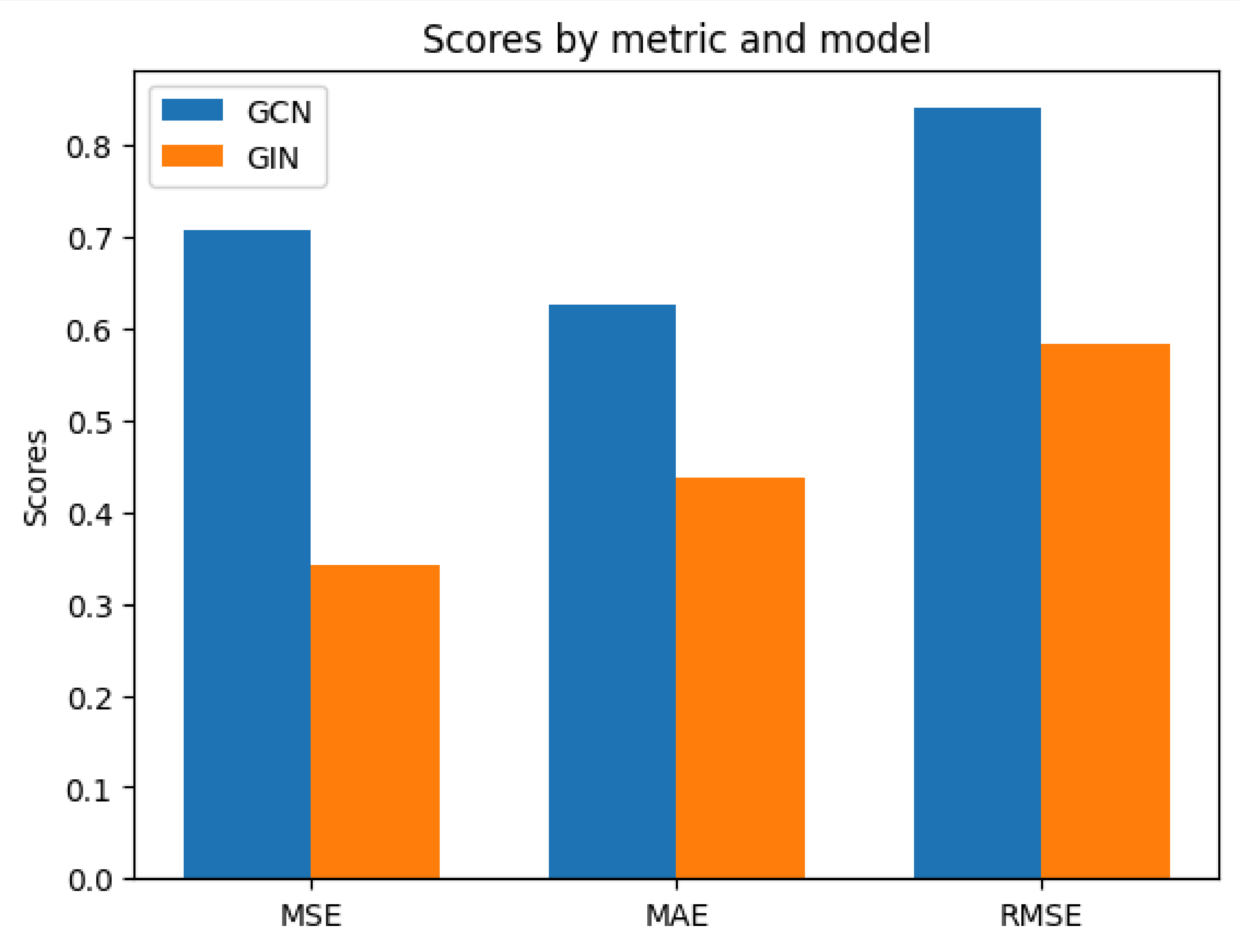




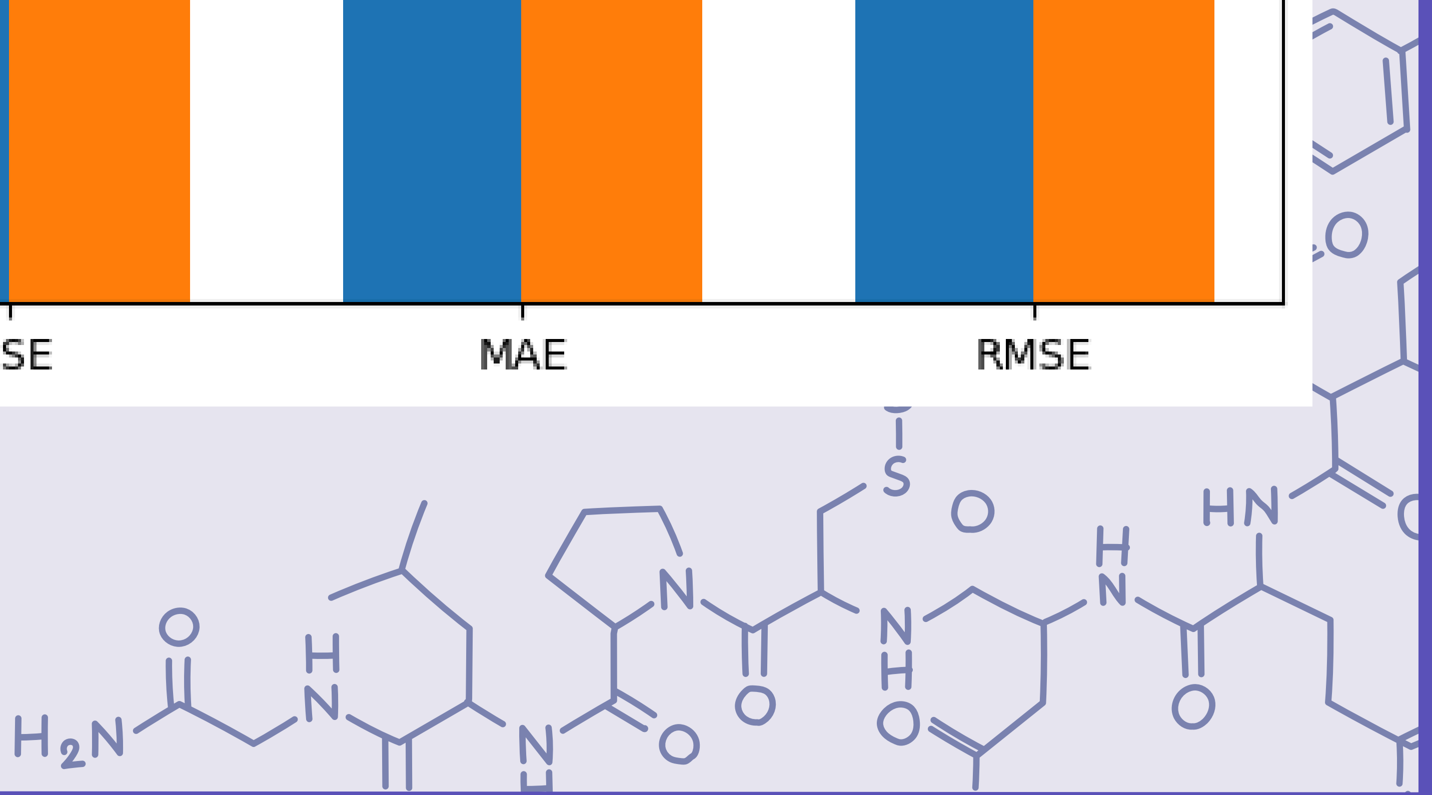
RESULTS GNN



Model Loss
Train loss (GCN)
Val loss (GCN)
Train loss (GIN)
Val loss (GIN)
loss
0.9
0.8
0.7
0.6
0.5
0.4
0
3
6
9
12
15
18
21
24
epoch

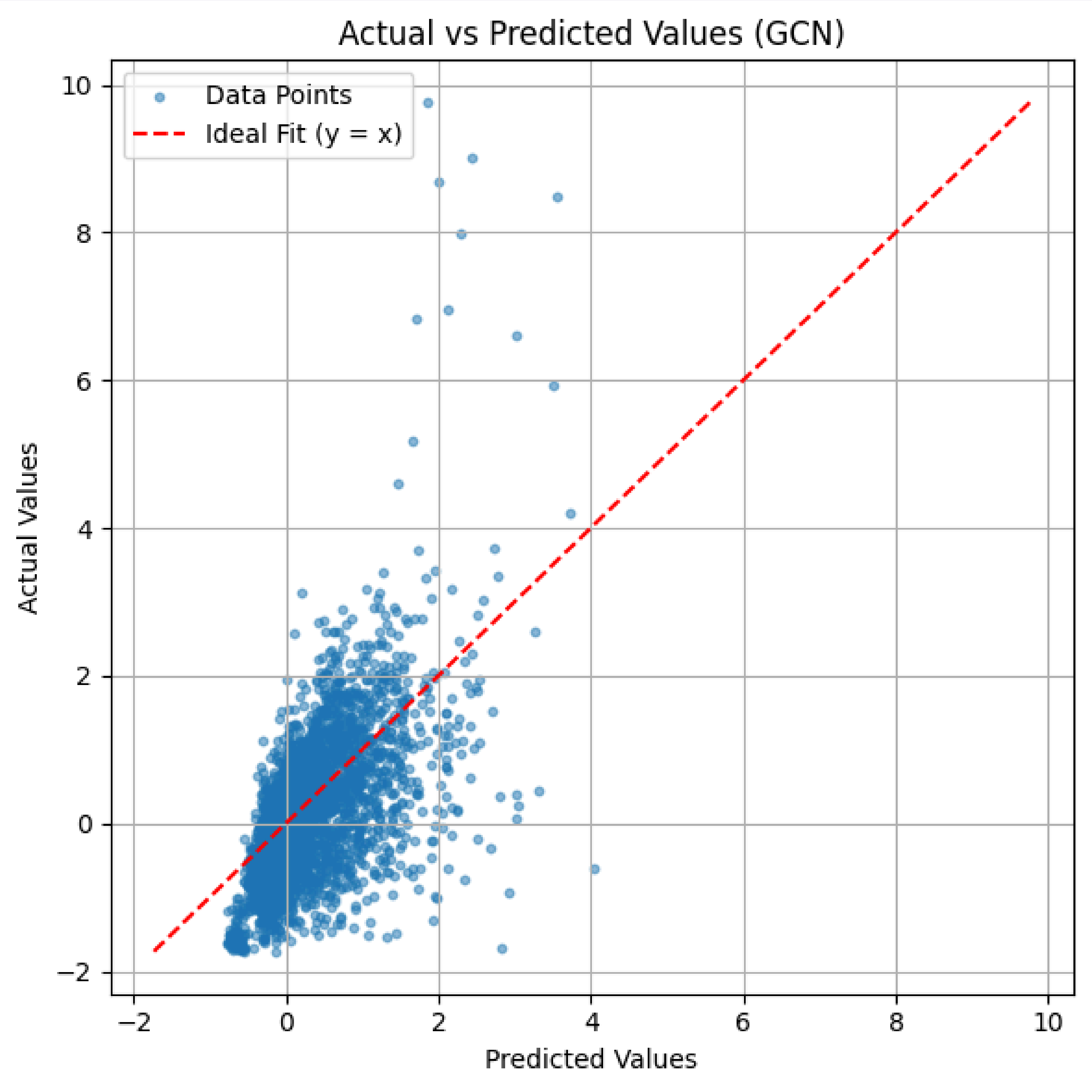


Scores by metric and model
GCN
GIN
Scores
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
MSE
MAE
RMSE

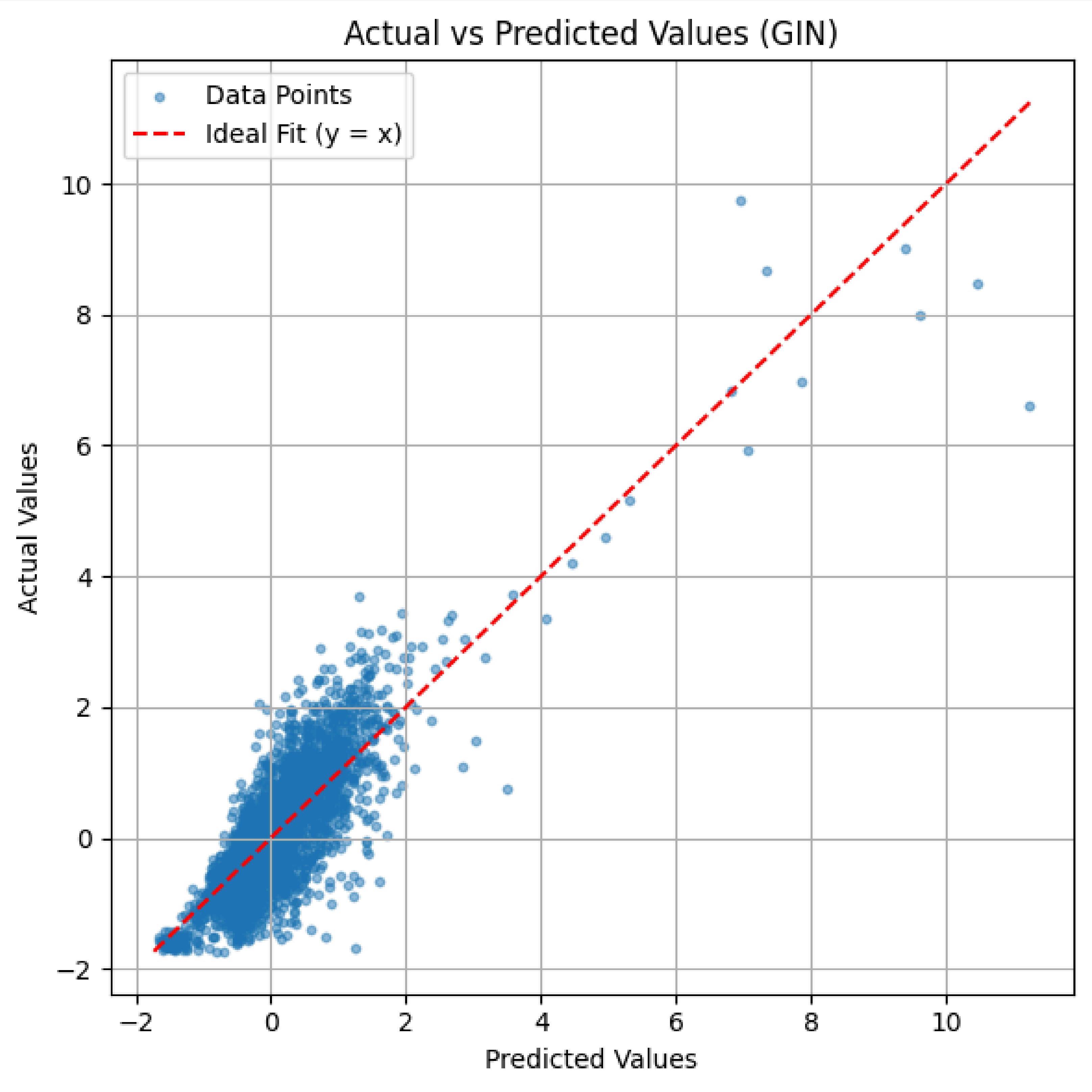


H_2N
HN
S

RESULTS GNN



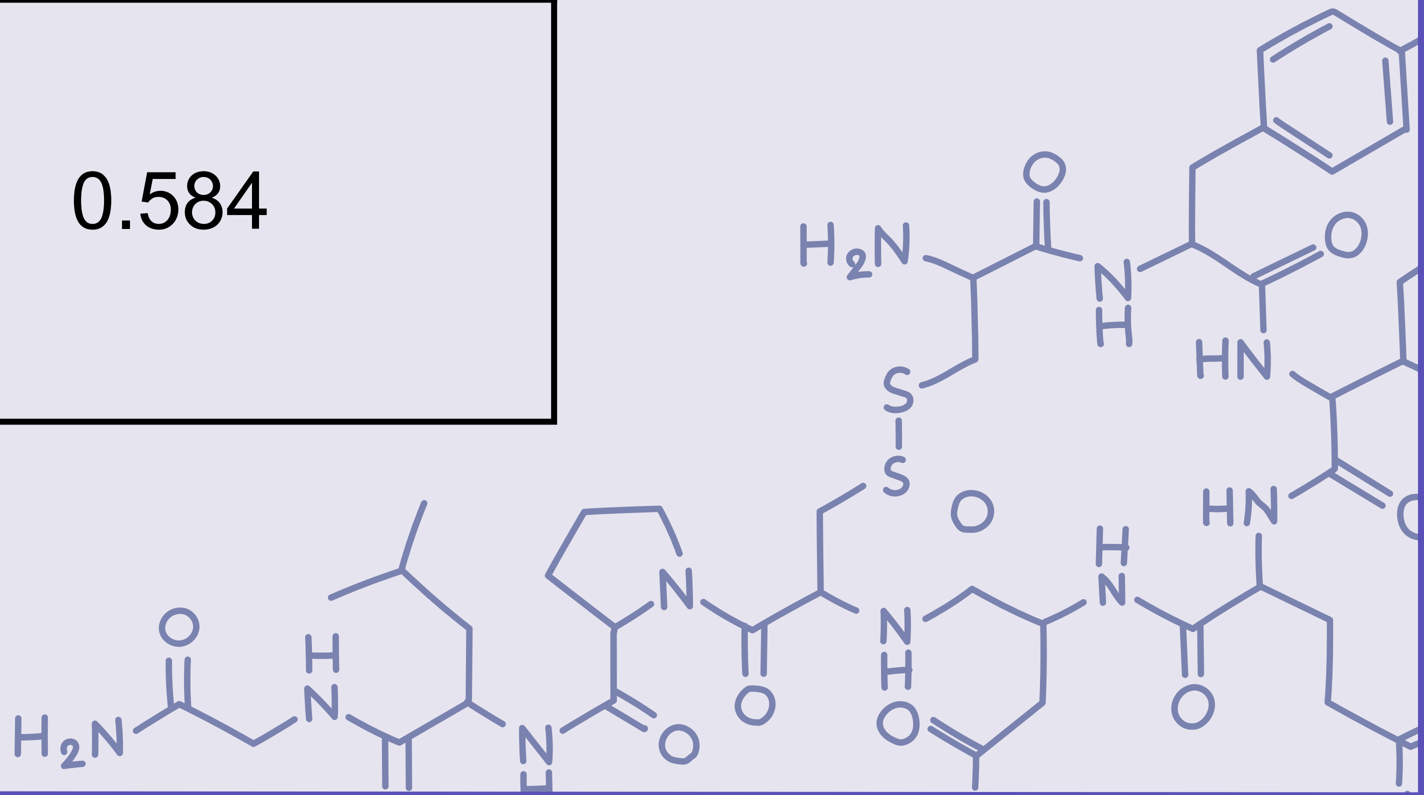
Actual vs Predicted Values (GCN)
Data Points
Ideal Fit (y = x)
Actual Values
Predicted Values



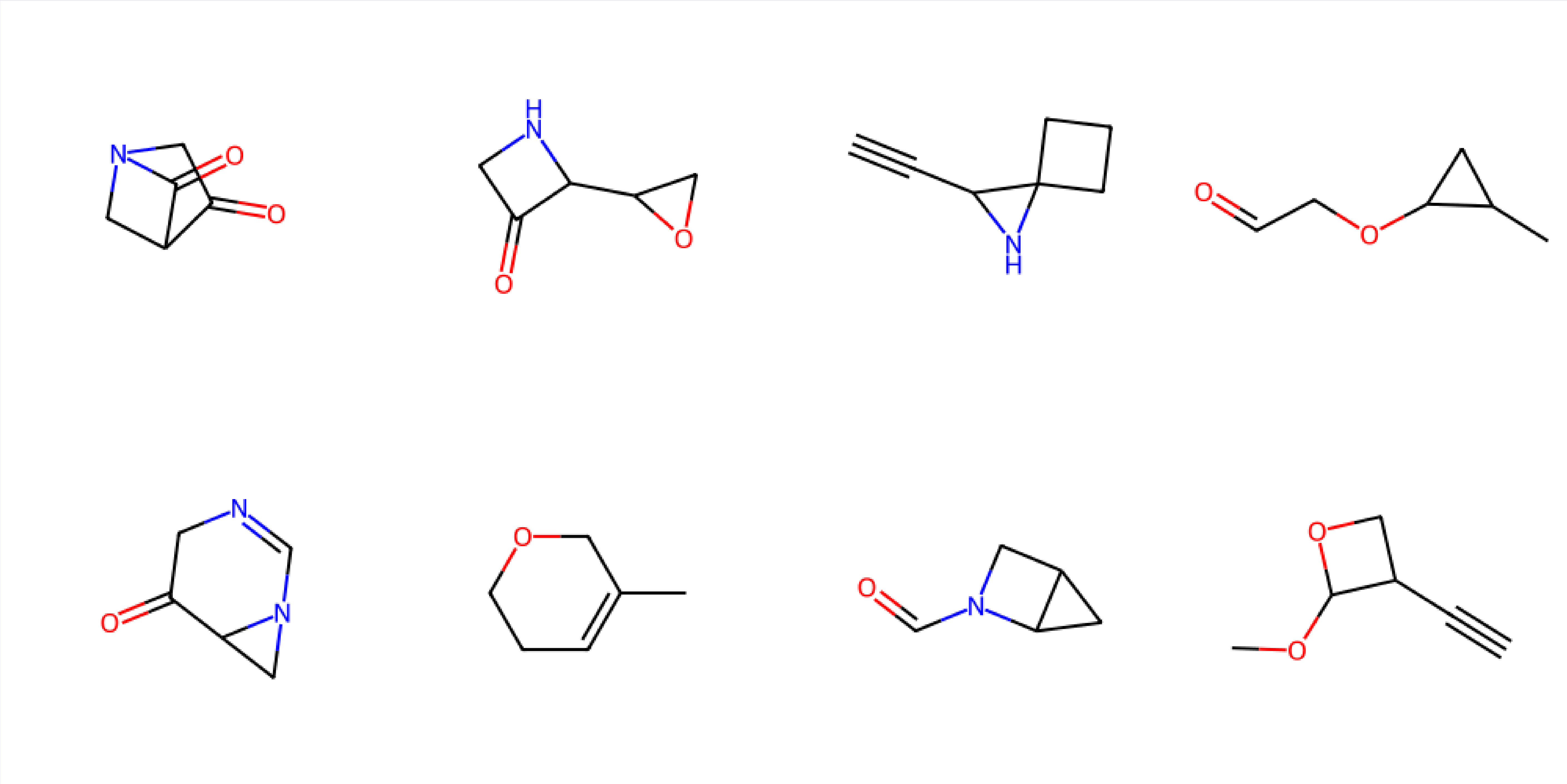
Actual vs Predicted Values (GIN)
Data Points
Ideal Fit (y = x)
Actual Values
Predicted Values

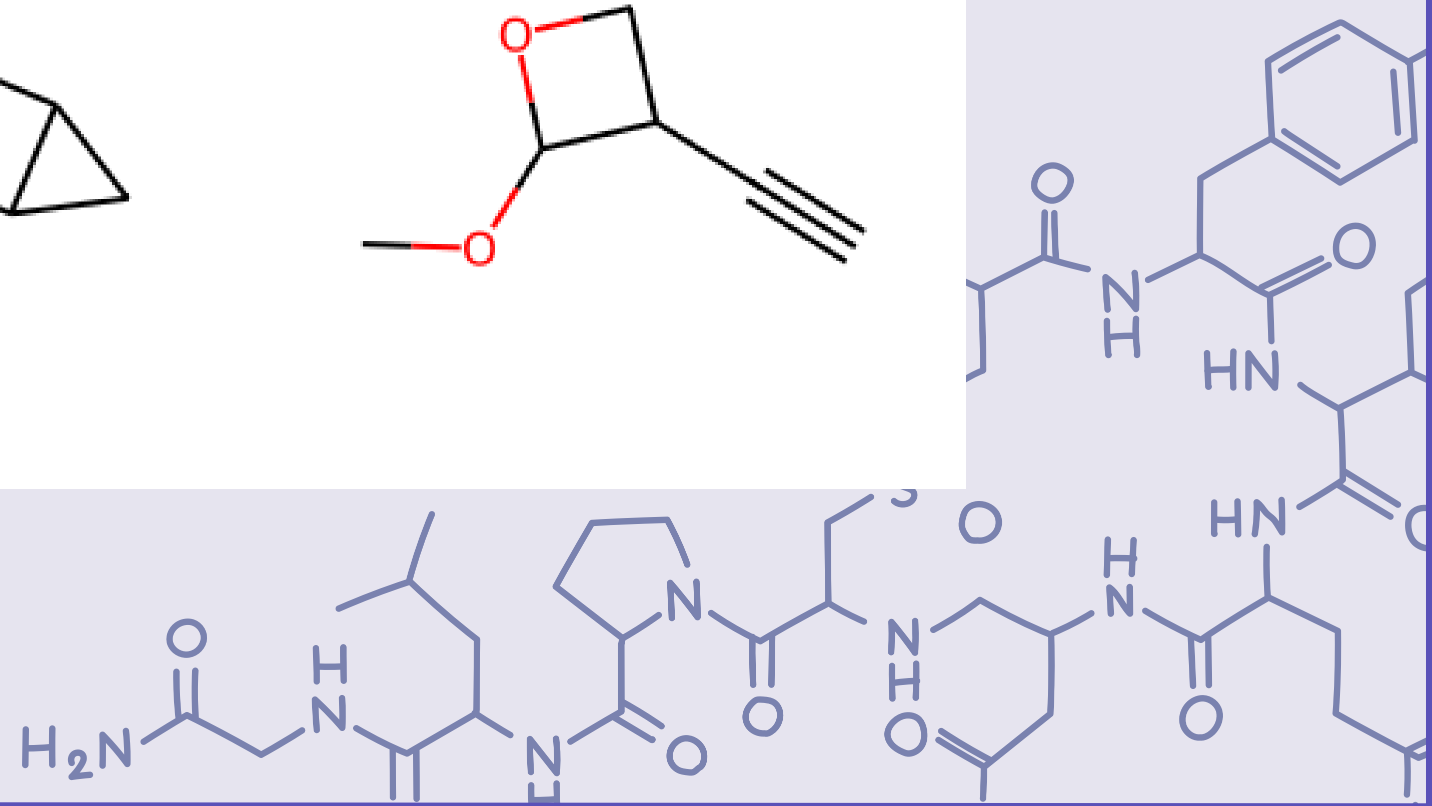
RESULTS GNN

Metric	GCN Model	GIN Model
Mean Squared Error (MSE)	0.707	0.342
Mean Absolute Error (MAE)	0.626	0.439
Root Mean Squared Error (RMSE)	0.841	0.584



DISCUSSION





Thank you very much!

We Appreciate You !