NOVAMOL

GENERATIVE MODEL FOR NOVEL CHEMICAL MOLECULES

**Group No: 17**

**Group Members:**

Ahmed Abdul Wasae Annan - A0328301R

Hari Krishna Dhamodaran - A0333806E

Raja Vijayakumar - A0327096X

John Joseph Peter - A0328845R

# **Introduction**

Discovering new molecules and understanding their properties is a critical step in developing advanced drugs, sustainable materials, and innovative chemicals. Traditional discovery methods are often slow, costly, and resource intensive. This project leverages artificial intelligence to identify patterns in existing molecular structures and properties, enabling the generation of novel candidate molecules and the prediction of their potential functionalities. By combining data-driven techniques with chemical knowledge, the project aims to accelerate molecular discovery while reducing experimental costs.

# **Project Background and Market Context**

* Problem Today: Discovering new molecules is slow, costly, and trial-and-error based. Traditional lab experiments and simulations take months or years.
* Why AI Helps: AI can predict molecular properties without running every experiment. This saves time, money, and resources in discovery.
* Market Needs
  1. Pharma: Faster drug discovery and lead optimization.
  2. Materials Science: New polymers, batteries, solar materials.
  3. Food & Fragrance: Novel safe flavour and aroma compounds.
  4. Agrochemicals: Eco-friendly pesticides and additives.
* Opportunity: AI-driven molecule generation can shorten R&D cycles. Growing demand for faster, cheaper, and sustainable innovation across industries.

# **Literature Review**

**1. Molecular Graph Neural Networks**

**Paper:** “Neural Message Passing for Quantum Chemistry” – Gilmer et al., ICML 2017

Introduced message-passing GNNs for predicting molecular properties.

**Paper:** “Graph Attention Networks” – Veličković et al., ICLR 2018

Attention-based GNN for molecular and general graph data.

**2. Molecular Generation with RNN / SELFIES**

**Paper:** “SELFIES: A Robust Representation of Semantically Constrained Graphs” – Krenn et al., 2020

SELFIES ensure chemically valid sequences for generative models.

**Paper:** “Molecular Generation with Recurrent Neural Networks” – Segler et al., 2018

Demonstrates SMILES-based RNNs for de novo molecule generation.

**3. Graph-based Property Prediction**

**Paper:** “Benchmarking Graph Neural Networks for Molecule Property Prediction” – Wu et al., 2018

Extensive comparison of GNN architectures for quantum chemistry datasets.

**4. Integration of Generation + Property Prediction**

**Paper:** “Junction Tree Variational Autoencoder for Molecular Graph Generation” – Jin et al., ICML 2018

Generates molecules as graphs while ensuring chemical validity.

**Paper:** “GraphAF: A Flow-based Autoregressive Model for Molecular Graph Generation” – Shi et al., NeurIPS 2020

Uses graph-based autoregressive models to generate molecules with target properties.

# **Project Scope**

The project aims to use Artificial Intelligence to automate the task of finding new molecules and predicting their properties by training on existing molecular datasets. This allows researchers and industry professionals to explore new compounds efficiently, assess their properties, and evaluate their novelty.

**Domain Knowledge Requirements**

To succeed in this project, expertise is needed across multiple disciplines:

* Artificial Intelligence / Machine Learning:
  + Graph Neural Networks (GNNs)
  + Variational Autoencoders (VAEs)
  + Recurrent Neural Networks (RNNs) and Transformers for sequence-based molecular generation
  + Generative models (flow-based, autoregressive, diffusion models)
* Chemistry / Molecular Science:
  + Chemical representations (SMILES, SELFIES, molecular graphs)
  + Quantum chemical properties (energies, charges, vibrational spectra)
  + Understanding molecular stability and reactivity principles
* Data Science / Engineering:
  + Handling large-scale datasets (cleaning, preprocessing molecular datasets)
  + Efficient data storage and GPU-optimized training pipelines
  + Integrating external APIs (PubChem, RDKit) for molecule validation

# **Data Collection**

## QM40 Molecular Dataset (Kaggle)

* Contains ~163,000 neutral molecules from the ZINC database (C, N, O, S, F, Cl)
* Provides quantum mechanical data: geometries, energies, Mulliken charges, vibrational properties
* Suitable for training and evaluating ML models on structure–property relationships
* Manageable size compared to larger QM datasets

## Why QM40?

* Initial plan: OMAT24 dataset → too computationally intensive
* QM40 = efficient alternative for experimentation
* Covers diverse molecular properties without overwhelming resources

**Initial Challenges**

1. **Dataset Scalability:** The original plan to use the OMAT24 dataset (20 GB) proved infeasible due to storage and memory limitations, making preprocessing and training impractical. This forced the team to pivot to a smaller yet comprehensive dataset (QM40).
2. **Computational Resources:** Training deep generative models and graph neural networks for molecular data requires powerful GPUs. Due to limited GPU capacity, the training and property prediction pipelines run slower than expected, limiting rapid experimentation and hyperparameter tuning.
3. **Chemical Validity of Generated Molecules:** While SELFIES and SMILES improve validity, ensuring generated molecules are chemically feasible and interpretable remains a practical challenge.

# **Appendix**

## References

* Gilmer et al. (2017) – Neural Message Passing for Quantum Chemistry
* Veličković et al. (2018) – Graph Attention Networks
* Krenn et al. (2020) – SELFIES robust molecular representations
* Jin et al. (2018) – Junction Tree VAE for molecular graph generation
* Shi et al. (2020) – GraphAF autoregressive flow models

## AI Tools Assistance

ChatGPT (GPT-5) – Assisted in drafting diagrams, flowcharts, code explanations, and slide content.

Gemini – assistance in code and debugging