**NovaMol**

GENERATIVE MODEL FOR NOVEL (NEW) CHEMICAL MOLECULES

**WHAT IS YOUR PROJECT ABOUT?**

Our project uses artificial intelligence to generate new chemical molecules and predict their properties, such as dipole moments. By training on existing molecular datasets, it can model molecular structures as graphs, learn patterns in atomic interactions, and suggest both novel molecules and their potential physical or chemical properties. This allows researchers and industry professionals to explore new compounds efficiently, assess their characteristics, and evaluate their novelty compared to known molecules.

**DID YOU LOOK INTO ANY PAPERS OR STUFF FOR THIS?**

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.

QM9 / QM40 datasets:

Standard datasets used for molecular property prediction. Your use of dipole moment is aligned with these benchmarks.

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.

5. Visualization

Py3Dmol / RDKit visualizations: widely used for 3D molecular visualization in interactive notebooks.

Py3Dmol GitHub

RDKit Documentation

**SUMMARISE YOUR PROJECT CODE**

Dataset:

QM40 Molecular Dataset (via Kaggle: nikitamanaenkov/qm40-molecular-qm-dataset)

It contains three CSV files:

main.csv

Contains molecule-level data.

Key columns:

Zinc\_id – unique molecule identifier

smile – SMILES representation of the molecule

dipol\_mom – dipole moment (target property)

Additional quantum properties

xyz.csv

Atom-level coordinates for each molecule.

Columns: Zinc\_id, atom, final\_x, final\_y, final\_z

bond.csv

Bond connectivity for each molecule.

Columns: Zinc\_id, atom1, atom2, bond

Libraries Used:

Molecular Processing: RDKit, SELFIES

Deep Learning: PyTorch, torch\_geometric

Data Handling: pandas, numpy

Visualization: py3Dmol

Utilities: tqdm, requests

Pipeline Overview:

Data Preprocessing:

Convert SMILES to SELFIES for robust molecule encoding.

Convert molecules to graphs (atoms as nodes, bonds as edges).

Compute RDKit descriptors (MW, LogP, TPSA, H-bond donors/acceptors).

Graph Neural Network (GNN):

Model: Simple GIN-based GNN with two GINConv layers and linear readout layers.

Input: Node features (atomic numbers) and bond adjacency.

Target: Predict dipole moment.

Training: Adam optimizer, L1 loss, train/validation split.

SELFIES RNN for Molecular Generation:

Trained on SELFIES sequences from dataset molecules.

Generates novel molecules token-by-token.

Decoded to SMILES and filtered for chemical validity.

Candidate Molecule Evaluation:

Convert generated SMILES to graph format.

Predict dipole moment using trained GNN.

Compute novelty using Tanimoto similarity (Morgan fingerprints) against dataset.

Check existence in PubChem via API.

3D Visualization:

Use RDKit + Py3Dmol to generate 3D structures.

Display stick representation of molecules.

**OKAY JUST TELL US WHAT THE CODE DOES NOW?**

1. Cell 1:Installs required chemistry, ML, and data libraries like RDKit, PyTorch, and PyGeometric.

2. Cell 2: Imports KaggleHub and OS libraries.

3. Cell 3: Downloads the QM40 molecular dataset from Kaggle and lists its files.

4. Cell 4: Loads main.csv, xyz.csv, and bond.csv into Pandas DataFrames and prints column names.

5. Cell 5: Defines a mapping from element symbols to atomic numbers for graph construction.

6. Cell 6: Builds graph representations from the dataset with atom features, edge indices, and target properties.

7. Cell 7: Splits the dataset into train and validation sets and creates PyG DataLoaders.

8. Cell 8: Defines a GNN model (GINConv-based) for predicting molecular properties and sets up optimizer and loss.

9. Cell 9:Defines an evaluation function to compute MAE and trains the GNN for a few epochs, printing train/val performance.

10. Cell 10: Uses py3Dmol to visualize random molecules in 3D with predicted properties.

11. Cell 11: Converts dataset SMILES to SELFIES representation for sequential modeling.

12. Cell 12: Builds a vocabulary from SELFIES tokens and converts them into tensors for training.

13. Cell 13: Defines an LSTM-based RNN for modeling SELFIES sequences.

14. Cell 14: Trains the SELFIES RNN on the dataset for sequence prediction.

15. Cell 15: Samples new SELFIES sequences from the trained RNN and decodes them back to SMILES.

16. Cell 16: Converts generated SMILES into graph representations compatible with the GNN.

17. Cell 17: Uses the trained GNN to predict properties for the generated molecules.

18. Cell 18: Filters generated molecules to keep only chemically valid SMILES.

19. Cell 19: Visualizes valid molecules in 3D using RDKit and py3Dmol with hydrogens and UFF optimization.

20. Cell 20: Checks PubChem for each generated molecule to classify them as existing or novel.

**WAIT WHY DID WE PREDICT DIPOLE MOMENT OF ALREADY PRESENT MOLECULES:**

You use the known dataset as a training ground to teach the model the relationship between molecular structure and a property like dipole moment. Then, when the model sees novel, generated molecules, it can predict their properties even though those molecules are not in the dataset.

It is essentially learning a function:

molecular structure

⟶

property value

molecular structure⟶property value

Once learned, this function can be applied to new, unseen molecules, which is crucial for virtual screening and drug/material discovery.

This way, you do not need to measure or simulate the property experimentally for every new molecule—the model can estimate it quickly.

**OKAY BUT THE PROJECT IS NOT ENOUGH FOR A “10 MAN DAY PROJECT”**

project extension:

Existing Molecule Analysis & Industrial Relevance

For molecules already in PubChem:

Scrape or query known properties (physicochemical, pharmacological, material properties).

Collect industrial applications or biological relevance from databases or literature.

Verify manually or semi-automatically via Google or PubChem references.

Present these in a summary table in your demo.

Prediction for Novel Molecules

Use your trained GNN and SELFIES-RNN generated molecules:

Predict molecular properties (dipole moment, LogP, TPSA, etc.).

Suggest potential applications or property-based uses using ML reasoning or LLMs.

Can generate a ranked list of promising candidates with predicted utility.

Website / Demo Visualization

Interactive frontend similar to AlphaFold:

Allow users to input molecules or select generated molecules.

Display: 3D structure (Py3Dmol), predicted properties, predicted uses, similarity to existing molecules.

Include metrics, statistics, and a dashboard summarizing:

Number of valid molecules generated.

Novelty vs. dataset molecules (Tanimoto similarity).

Prediction confidence or MAE/accuracy of property prediction.

Backend: Fast API or Flask to serve the model and handle queries.

Extra Metrics / Stats:

Property Prediction: MAE, RMSE for dipole moment or other properties.

Novelty: % of molecules below similarity threshold.

Validity: % of chemically valid generated molecules.

Potential Applications: Number of molecules with predicted industrial relevance.

**SO WHAT’S THE BUSINESS MODEL OR USE CASE OF THIS MODEL U MADE?**

1. Drug Discovery & Pharmaceuticals

Predicting bioactive molecules for specific targets (enzymes, receptors).

Screening for molecules with desirable ADMET (absorption, distribution, metabolism, excretion, toxicity) properties.

Generating novel small molecules for lead optimization.

Predicting physicochemical properties to accelerate preclinical testing.

2. Material Science & Industrial Chemicals

Designing polymers, resins, and composites with specific thermal or mechanical properties.

Predicting molecular properties for dyes, pigments, or coatings.

Discovering novel lubricants, solvents, or surfactants.

Optimizing molecules for batteries, solar cells, or other energy materials.

3. Flavor & Fragrance Industry

Generating novel aromatic compounds with target olfactory profiles.

Predicting volatility, solubility, or safety of fragrance molecules.

4. Agrochemicals & Food Additives

Designing pesticides, herbicides, or fungicides with improved efficacy and lower environmental impact.

Predicting solubility and stability of food additives.

5. Biotech & Enzyme Engineering

Designing small molecules that modulate enzyme activity.

Predicting ligand binding affinity or inhibition potential.

6. Environmental Chemistry & Green Chemistry

Identifying molecules with low toxicity or high biodegradability.

Predicting environmental persistence or pollutant capture efficiency.

7. Chemical Safety & Regulatory

Screening compounds for potential hazards before synthesis.

Predicting flammability, reactivity, or toxicity for compliance.

8. Academic & Research Applications

Virtual exploration of chemical space.

Studying structure-property relationships for novel molecules.

Benchmarking predictive models against experimental datasets.

9. AI-Driven Custom Molecule Design Services

Offering clients (pharma, chemicals, cosmetics) on-demand generation of novel compounds with predicted properties.

Visual dashboards showing 3D structures, predicted properties, and novelty scores.