Predicting HOMO and LUMO Energy Levels Using a Graph Convolutional Network-Based Deep Learning Model

Mayuri Patil, Sanchita Hati, and Sudeep Bhattacharyya,
Department of Chemistry and Biochemistry, University of Wisconsin-Eau Claire

ABSTRACT

HOMO (Highest Occupied Molecular Orbital) and LUMO (Lowest Unoccupied Molecular Orbital) are two key properties that play a pivotal role in determining a molecule's chemical reactivity and electronic properties. Our research leverages Graph Convolutional Networks (GCNs) to accurately predict HOMO and LUMO values across a diverse chemical dataset. By representing molecules as graphs, we extract key substructures contributing most to orbital properties. This data-driven approach accelerates prediction while offering insights into molecular electronic behavior, potentially aiding in drug discovery and materials design.

BACKGROUND

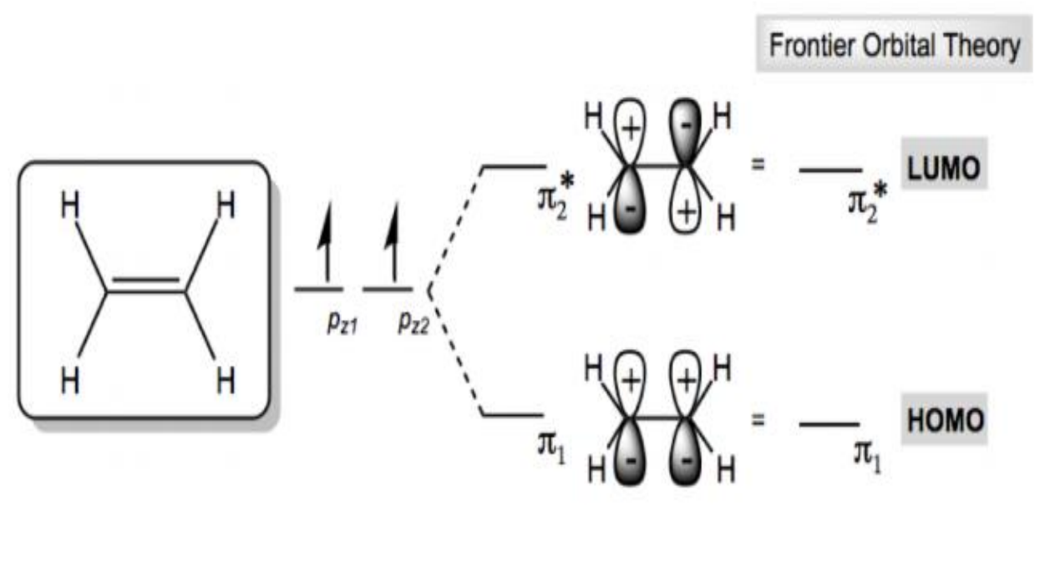

Figure 1: *Visual Representation of HOMO and LUMO Levels*

- •HOMO (Highest Occupied Molecular Orbital) is the highest energy level occupied by electrons in a molecule.
- •LUMO (Lowest Unoccupied Molecular Orbital) is the lowest energy level available to accept electrons.
- •The HOMO-LUMO gap is a key indicator of a molecule's stability, reactivity, and electronic behavior.
- •The HOMO-LUMO gap correlates with molecular stability, reactivity, and optical/electronic behavior.
- •Traditional quantum chemistry methods (e.g., Density Functional Theory, DFT) are accurate but computationally expensive.
- •Graph theory is a branch of mathematics that studies networks of nodes and edges, a perfect fit for representing molecules, where atoms are nodes and bonds are edges.
- •Graph Convolutional Networks (GCNs) apply deep learning to these graph structures, learning atom-level features by aggregating chemical context across the molecule.
- •An Artificial Neural Network (ANN) is a machine learning model inspired by the human brain which receives the molecular fingerprint created by the GCN and it maps that fingerprint to a prediction.

IMPORTANCE

In drug discovery, speed and accuracy are critical. Rather than running slow quantum simulations for every new molecule, we use deep learning to make fast, reliable predictions of HOMO and LUMO values. Just as importantly, the model highlights the molecular substructures that drive these properties.

The HOMO-LUMO gap, the energy difference between the Highest Occupied and Lowest Unoccupied Molecular Orbitals is a key indicator of a molecule's reactivity and stability. A small gap suggests high reactivity; a large gap, greater stability. This enables faster identification of drug candidates and helps chemists design better molecules, making molecular modeling more efficient and insightful.

METHODS

Dataset

We used the QM9 dataset, which contains approximately 133,885 small organic molecules. Each molecule includes quantum-chemically calculated properties, including HOMO and LUMO energy levels.

SMILES to GCN based Fingerprint

Molecules were represented using SMILES (Simplified Molecular Input Line Entry System), a compact text format that describes molecular structure.

These were converted into molecular graphs, where:

- Atoms :-Nodes with features (atomic number, hybridization, aromaticity, etc.)
- Bonds :- Edges representing connectivity and bond type

This representation forms a graph-based fingerprint, capturing the molecule's structure and local chemical environment.

Unlike traditional bit-vector fingerprints, these are learned directly by the model through graph neural networks, allowing it to identify complex substructures and patterns relevant to molecular properties like HOMO & LUMO values.

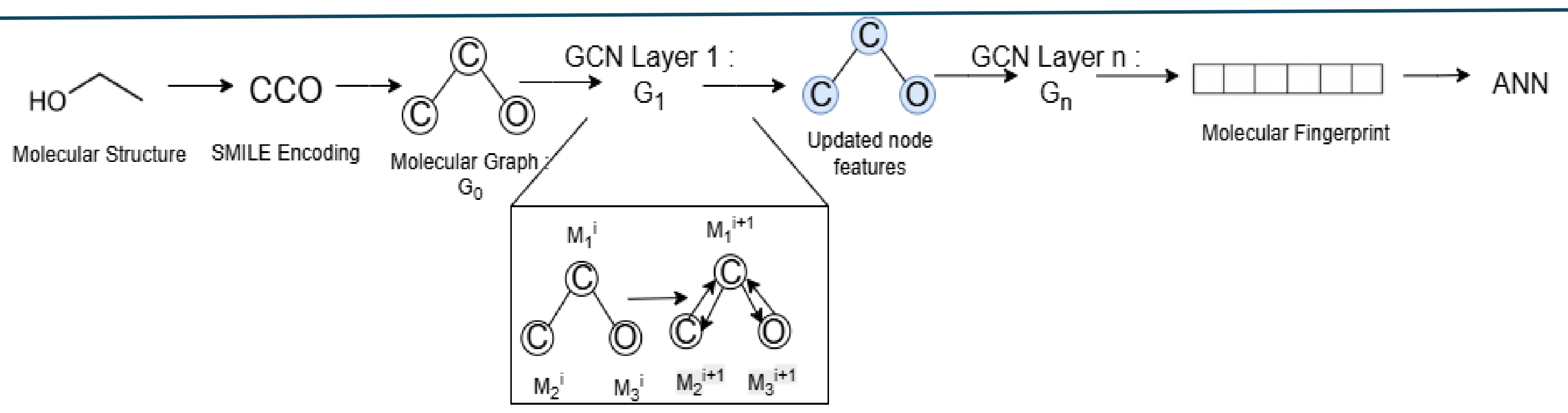

Figure 2: *Graph Convolutional Network (GCN) Architecture for Molecular Representation Learning*

Model Architecture

Each model combined:

- Graph Convolutional Network (GCN) layers to translate the molecular structure to graph-based fingerprint.
- A fully connected Artificial Neural Network (ANN) for regression and to predict HOMO & LUMO values based on the graph-based fingerprint as input

We created 10 unique model architectures with varied hyperparameters to enhance robustness.

Training Pipeline

- •Creating model configurations: Generated model configurations
- •Training individual models: Trained each model for 250 epochs using an 70/15/15 split of the data. Stratified binning was used to ensure HOMO and LUMO values were evenly distributed across training and validation sets.
- •Building ensemble predictions: Combined outputs from all 10 models into a single ensemble prediction.

Evaluation : Assessed performance using:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 Score
- Output included scatterplots and CSVs of predicted vs. actual values

Substructure Interpretation & Prediction :

- Predicts HOMO or LUMO values using a trained model on a specified directory (e.g., homo/ 3, i.e : directory homo, model 3)
- Highlights important substructures contributing to the prediction via gradient-based attention
- Outputs include both the numerical predictions and annotated PNG images showing key atom contributions

RESULTS

Quantitative Performance

- The ensemble model achieved high predictive accuracy on the QM9 dataset for both HOMO and LUMO energy values.
- •The model was trained to predict frontier orbital energies (HOMO and LUMO) from molecular graph representations derived from SMILES strings.
- •Evaluation metrics were computed on a withheld test set using a script that calculates standard regression metrics by comparing predicted and true values..
- Representative values included:
  - MAE (Mean Absolute Error): *≤ 0.04 eV*
  - RMSE (Root Mean Squared Error): *≤ 0.06 eV*
  - R^2 Score: *> 0.95*

METRICS	HOMO: - Ensemble	LUMO: - Ensemble
MAE (eV)	0.0692	0.0819
RMSE (eV)	0.1011	0.1154
R^2	0.9721	0.9917
SMAPE	1.0829	18.5089

Table 1:- *Evaluation Metrics*

Substructure & Prediction Insights

- We used Model 7 (HOMO) and Model 8 (LUMO) to generate:
  - Orbital predictions (HOMO/LUMO values)
  - Visualized substructures via gradient-based attention
- Outputs included:
  - Annotated PNG images highlighting key atoms influencing predictions
  - CSV logs of fingerprint indices, SMILES, and predicted values
- Substructure activations provided interpretability of predictions
- All 10 models were assessed individually; the ensemble improved generalization and reduced error

Molecule	Calculated HOMO (eV)	Predicted HOMO (eV)	Calculated LUMO (eV)	Predicted LUMO (eV)
Amoxapine	-5.86	-6.10	-1.73	-1.80
Milnacipran	-6.23	-6.17	-0.62	-0.23

Table 2:- *HOMO-LUMO Prediction Accuracy*

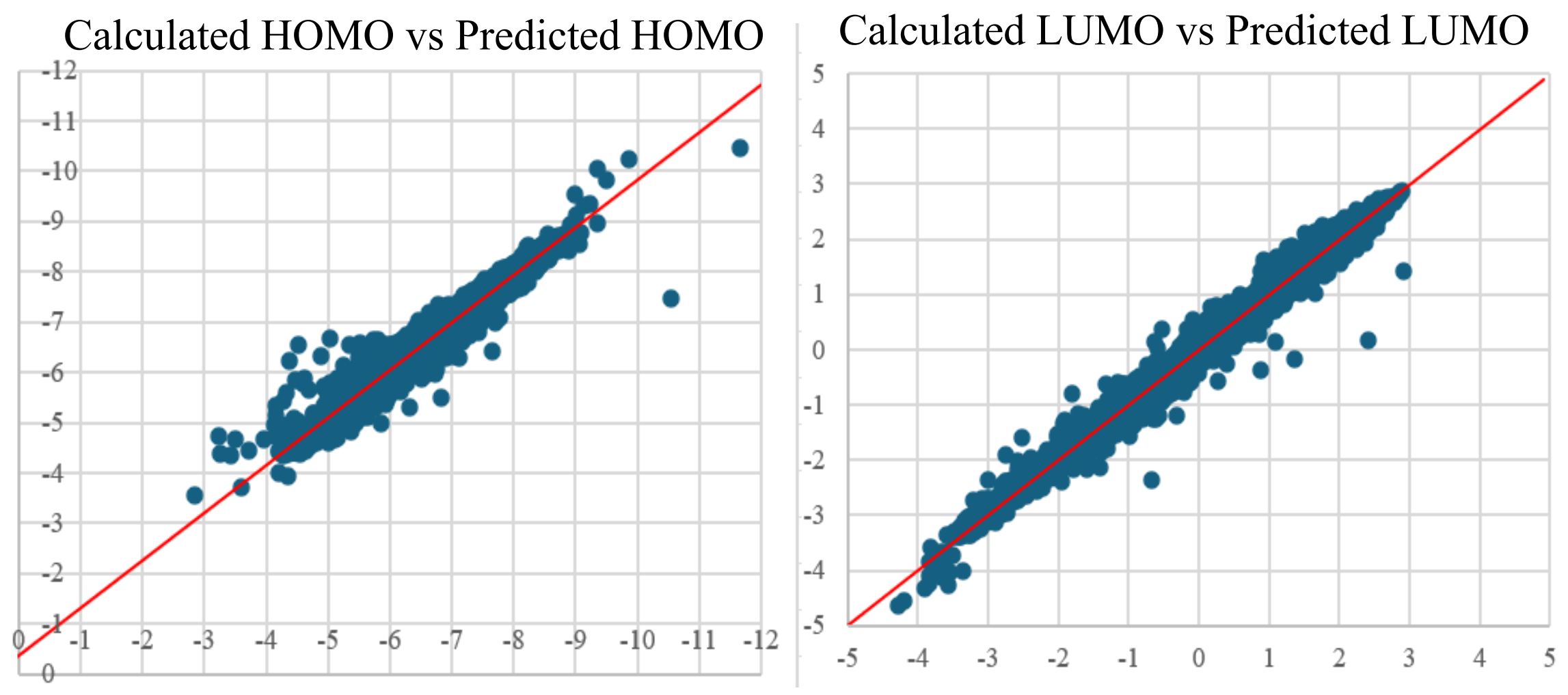

Figure 3: *Scatter plots showing correlation between calculated and predicted HOMO (left) and LUMO (right) energy values*

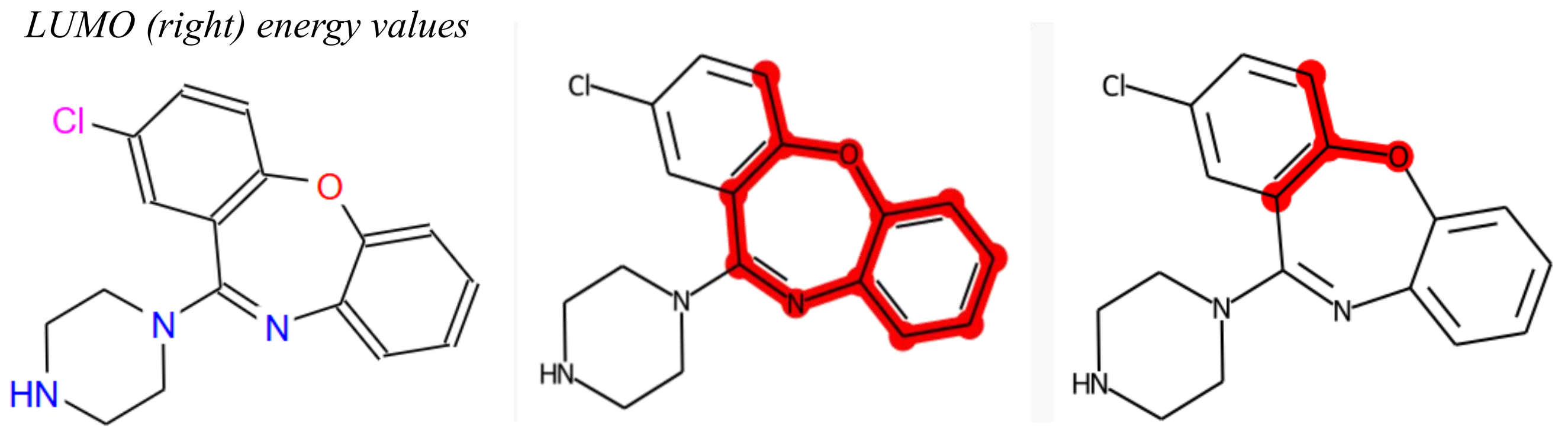

Figure 4: *Amoxapine structure (left) with predicted HOMO (middle) and LUMO (right) regions highlighted in red.*

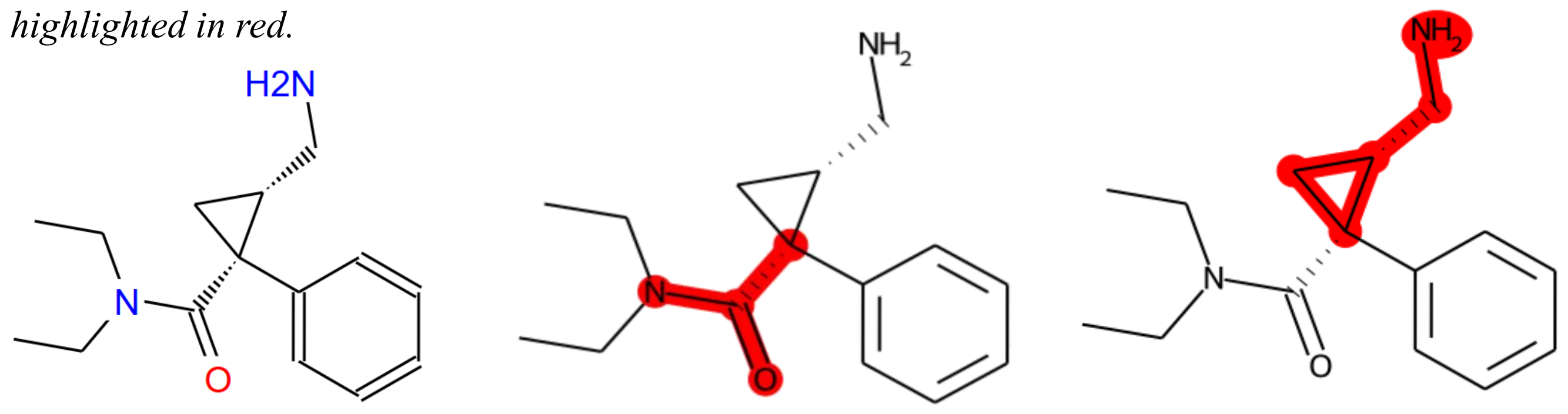

Figure 5: *Milnacipran structure (left) with predicted HOMO (middle) and LUMO (right) regions highlighted in red.*

CONCLUSION & FUTURE WORK

- •Demonstrated the effectiveness of Graph Convolutional Networks (GCNs) in predicting HOMO and LUMO energies from SMILES-derived molecular graphs.
- •Developed a GCN + ANN ensemble model for improved accuracy and robustness
- •Enabled rapid, DFT free electronic property predictions, scalable to molecular libraries
- •Achieved high performance on the QM9 dataset, a benchmark for quantum molecular property prediction.
- •Implemented substructure-level visualization to interpret atom-wise contributions.
- •Supports drug discovery, materials science, and high-throughput screening through fast, explainable predictions
- •Outperformed traditional molecular fingerprints (eg: ECFP) by learning task-specific chemical representations.
- •Implementing Harris Hawk Optimization to improve model accuracy.
- •Using Coulomb Matrix technique to improve the fingerprint which considers the 3-D geometry of the molecule while creating a fingerprint.

ACKNOWLEDGEMENT

We acknowledge that UW-Eau Claire occupies the sacred and ancestral lands of the Ojibwe and Dakota Nations. This research was funded by the National Science Foundation (NSF) Research Experience for Undergraduates (REU) grant OAC- 2447779. The computational resources of this study were provided by the Blugold Center for High-Performance Computing under NSF grant CNS-1920220. We sincerely thank, WiSys SPARK Symposium Organizers, University Printing Services , Hati-Bhattacharyay group members

REFERENCES

- Gilmer, Justin, et al. "Neural Message Passing for Quantum Chemistry.", June 2017, https://arxiv.org/abs/1704.01212
- Duvenaud, David, et al. *Convolutional Networks on Graphs for Learning Molecular Fingerprints*. Jan 2015.
- "HOMO LUMO - Video Tutorials & Practice Problems | Channels for Pearson+." *Www.pearson.com*, www.pearson.com/channels/organic-chemistry/learn/johnny/conjugated-systems/homo-lumo.
- Kearnes, Steven, et al. "Molecular Graph Convolutions: Moving beyond Fingerprints." *Journal of Computer-Aided Molecular Design*, vol. 30, no. 8, Aug. 2016, pp. 595–608, https://doi.org/10.1007/s10822-016-9938-8.
- Ramakrishnan, Raghunathan, et al. "Quantum Chemistry Structures and Properties of 134 Kilo Molecules." *Scientific Data*, vol. 1, no. 1, Aug. 2014, p. 140022, https://doi.org/10.1038/sdata.2014.22.