Supporting Information: Graph-Based Networks for Accurate Prediction of Ground and Excited State Molecular Properties from Minimal Features

Denish Trivedi[1,†], Kalyani Patrikar[2,†], and Anirban Mondal[2,*]

[1]Department of Physics, Indian Institute of Technology Gandhinagar, Gujarat, 382355, India

[2]Department of Chemistry, Indian Institute of Technology Gandhinagar, Gujarat, 382355, India

[†]These authors contributed equally to this work.

September 17, 2024

email: amondal@iitgn.ac.in

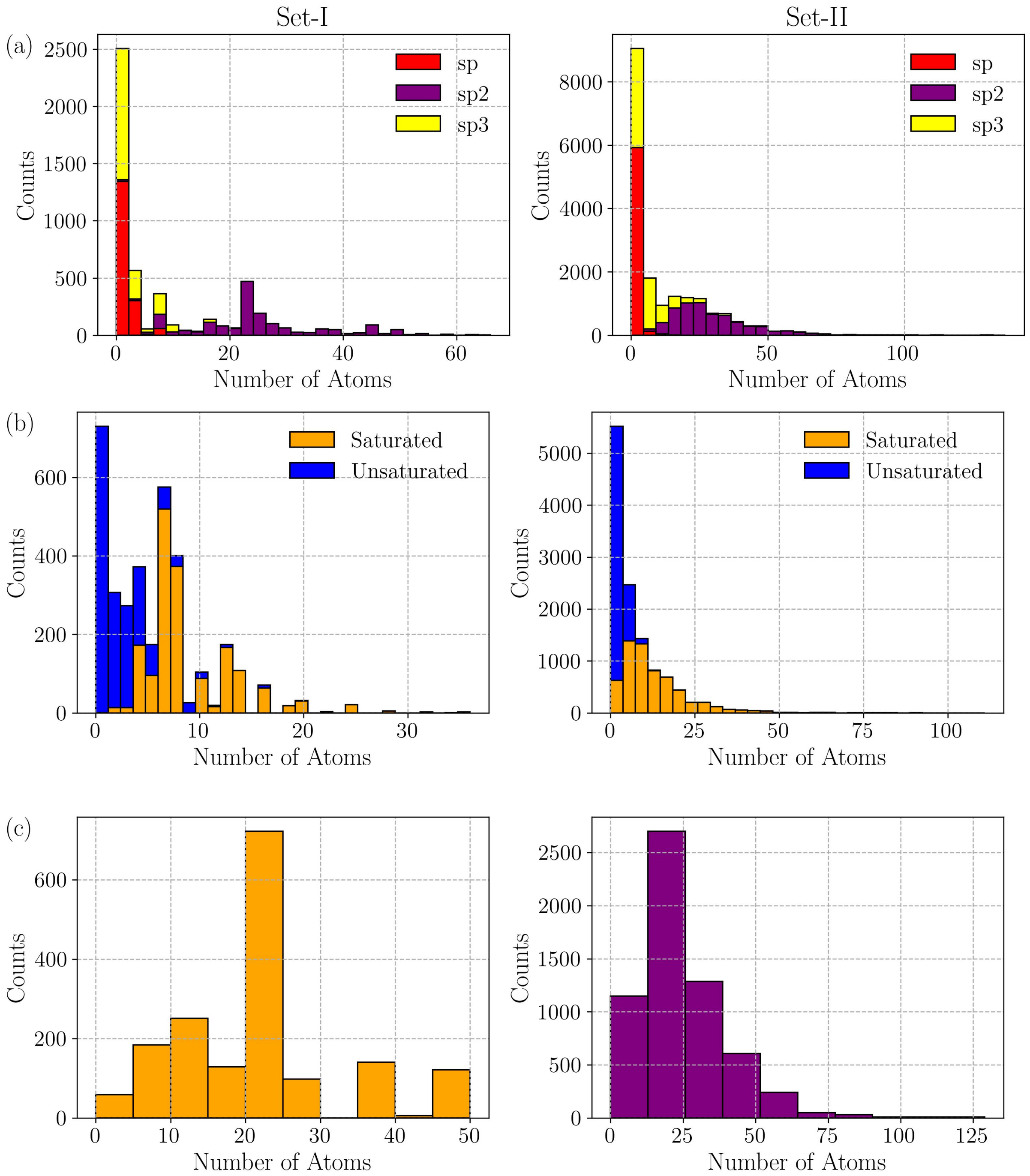

Figure S1: Input features of molecules in Set-I (left column) and Set-II (right column): (a) number of atoms in hybridisation state *sp*, sp^2, or sp^3; (b) number of saturated or unsaturated bonds; and (c) number of atoms in aromatic structures.

S1 General Applicability of GAT Model

To demonstrate the broad applicability of the GAT-based model, we tested it on a separate dataset of organic donor-acceptor molecules. These molecules play a crucial role in enabling high-performance organic devices due to their favorable optoelectronic properties. Various methods for designing such high-performing molecules are actively being explored. One of the key properties for these applications is the difference between the HOMO and LUMO energy levels, which governs the semiconducting behavior of organic molecules. In this study, we used our model to predict the HOMO and LUMO energy levels of molecules from the "Database of Organic Donor-Acceptor Molecules," curated by the Computational Materials Repository (available at https://cmr.fysik.dtu.dk/solar/solar.html). This database contains properties of molecules computed using first-principles methods.

Our results show that the GAT model's predicted values are in close agreement with the computed ones (see Figure S2). For the HOMO energy prediction, the model achieved a mean absolute percentage error (MAPE) of 1.83%, a coefficient of determination (r^2) of 0.84, a mean absolute error (MAE) of 0.09 eV, and a root mean square error (RMSE) of 0.12 eV. For the LUMO energy, the model yielded a MAPE of 3.03%, an impressive r^2 of 0.94, an MAE of 0.08 eV, and an RMSE of 0.10 eV, as summarized in Table S1. These metrics highlight the accuracy and reliability of the GAT model in predicting key optoelectronic properties, demonstrating its effectiveness as a powerful tool for molecular design.

Table S1: Metrics for prediction of HOMO and LUMO energy by GAT model for a separate database.

	MAPE	r^2	MAE	RMSE
E_{HOMO}	1.83	0.84	0.09	0.12
E_{LUMO}	3.03	0.94	0.08	0.10

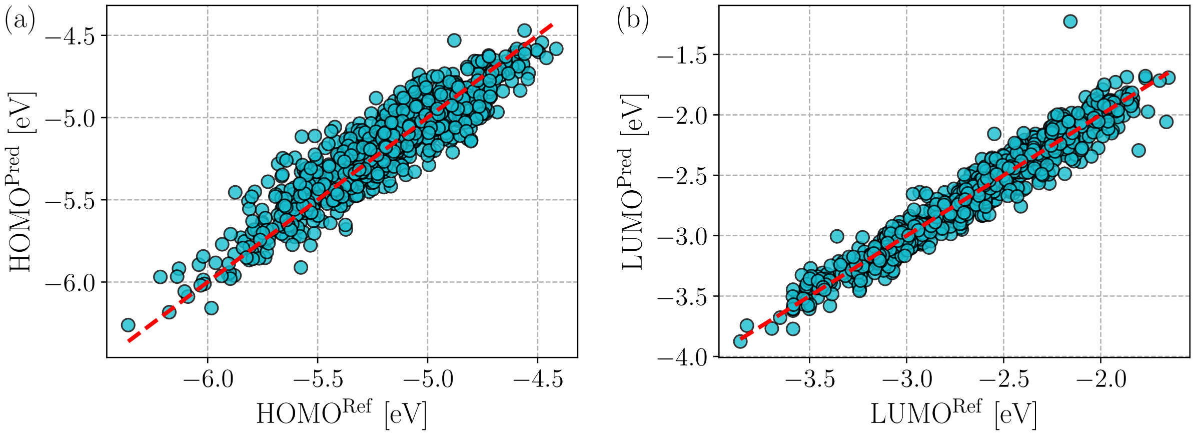

Figure S2: Predicted versus computed values for an additional database for the properties (a) HOMO and (b) LUMO.

S2 Data Invariance of GAT based Model

Maxima of absorption wavelengths of molecules in Set-I (λ_{max}) are also considered for comparison with the prediction of λ_{abs} obtained from Set-II. These values have been computationally derived from first principles, as is the case for all properties reported for Set-I molecules, as opposed to experimentally measured λ_{abs} for the molecules in Set-II. Figure S3(a) shows the spread in properties for λ_{max} of Set-I molecules. While there is a skew in the distribution, it is lower than that for the values of λ_{abs} for molecules of Set-II, as shown in Figure 1. The values predicted by GAT and actual values are shown in Figure S3. The prediction accuracy attains an r^2 of 0.79, similar to that obtained for λ_{abs} of Set-II molecules. Further, the MAE is obtained as 41.10 nm, leading to a MAPE of 8.06%, which is also similar to those for λ_{abs} of Set-II. This demonstrates the robustness as well as invariance of the graph-based model in predicting properties with high accuracy, regardless of the origin of the data. While the method of obtaining absorption wavelength values and dataset size are markedly different for Set-I and Set-II, the model accurately predicts properties for a given molecular structure.

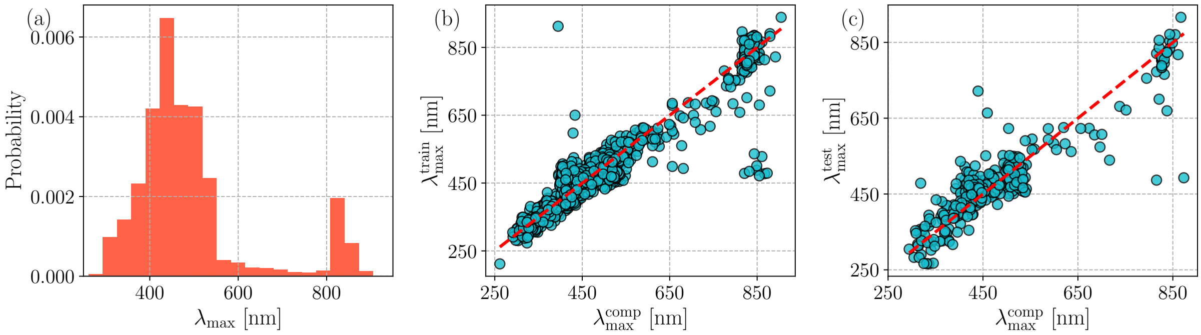

Figure S3: (a) Distribution in the absolute values of λ_{max} trained (tested) in this study, (b) & (c) predicted versus actual values for λ_{max} for training and testing steps.