Topology-Aware GNNs: Teaching Molecules Their Quantum Features

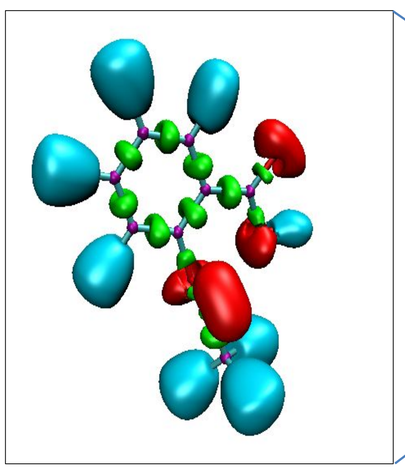
Subashini Kennedy [1], Rodolphe Vuilleumier [1], Marc Bianciotto [2]

1 PASTEUR, Département de Chimie, École Normale Supérieure, PSL University, Sorbonne Université, 75005 Paris, France

2 Molecular Design Sciences – Integrated Drug Discovery, Sanofi R&D, 94400 Vitry-sur-Seine, France

Highlights

- Can incorporating ELF-derived descriptors into GNNs improve model performance?
- ELF descriptors provide quantum-level, quantitative information on atom and bond properties.
- Utilizing a database of 44K WFX files with TopChem2 to extract topology descriptors improves model performance compared to models relying only on atom and bond features from molecular graphs.



Electron Localization function (ELF) graph


```
{
"Critical Points": {
"1": {
        Label: "C(C2),
        Type: "core",
        Atom list: "C2",
        Volume: 0.823,
        Population: 2.0844,
        Charge: 3.9156,
        X: 1.9805547,
        Y: -0.7057066,
        Z: 0.33965203,
        group: 0,
        Value: 0.99999966
}, .....
```

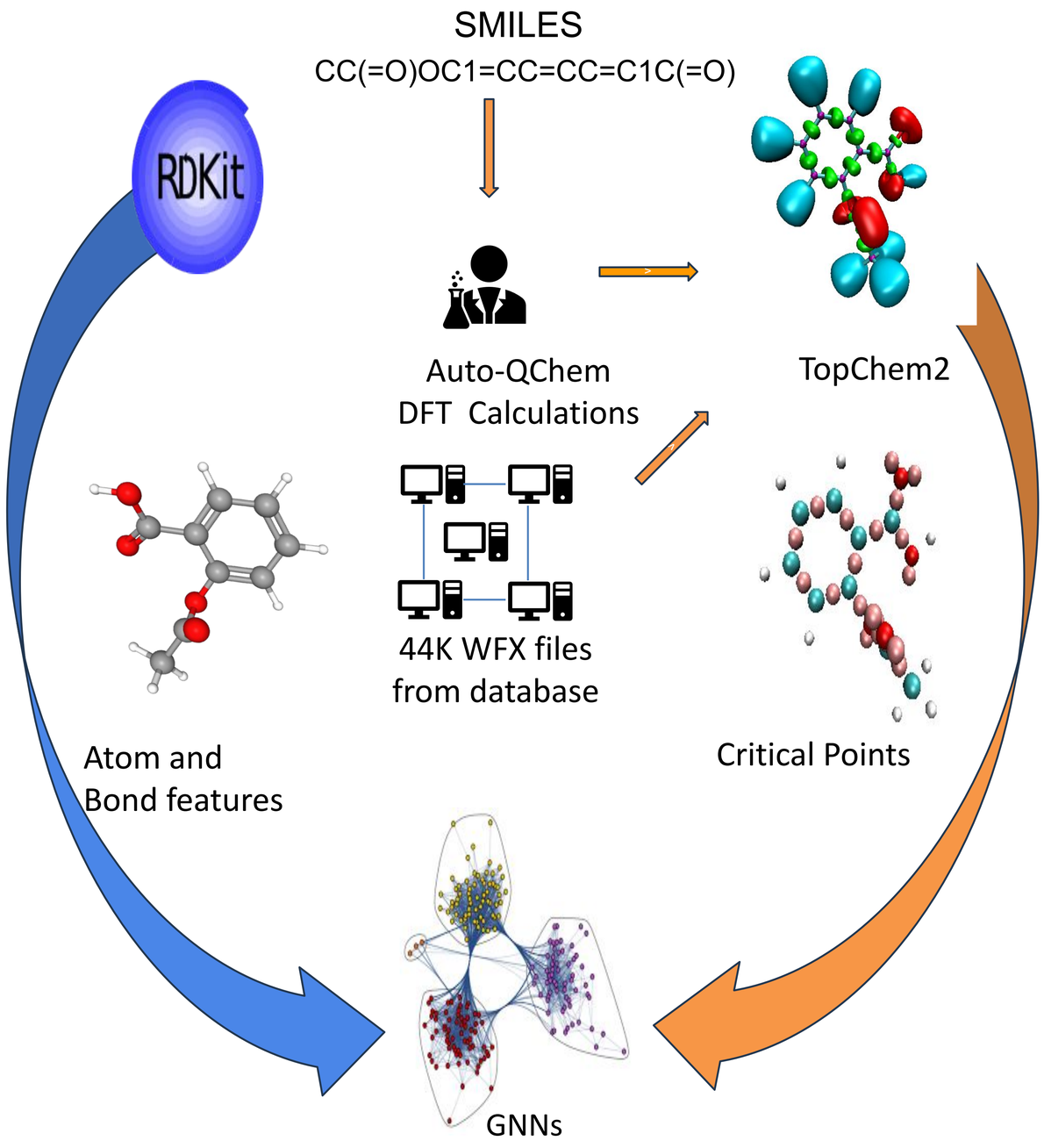
```
"41": {
        Label: "V(C2,O3),
        Type: "valence",
        Atom list: "C2,O3",
        Volume: 38.226,
        Population: 2.4191,
        Charge: -2.4191,
        X: 1.79692437,
        Y: -0.68776169,
        Z: 0.92143002,
        group: 3,
        Value: 0.85703409
}}}
```


Background

Drug discovery is a high-dimensional optimization process, with Quantitative Structure Activity Relationship (QSAR) models playing a key role in linking molecular descriptors to biological activity.

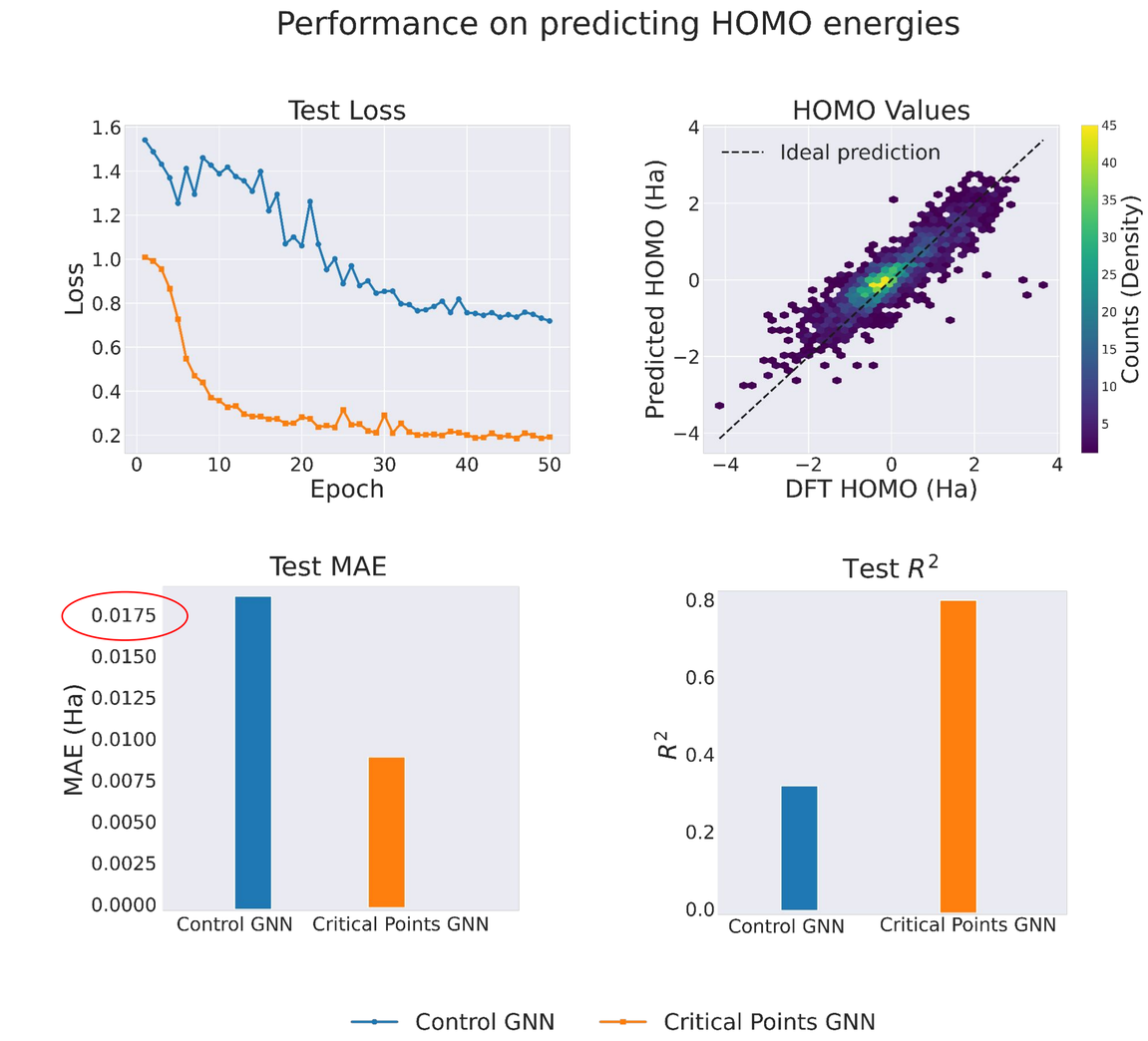
Extending molecular representations used by Graph Neural Networks (GNNs) with quantum level features extracted from ELF (Electron Localization Function) analysis offers the potential to enhance model's predictive performance. These ELF descriptors include critical points (nuclear, bond and ring) as well as lone pairs.

Workflow



- DFT static calculations (B3LYP/6-31G)
- 44K WFX files from 44K randomly selected QM9 molecules (excluding those containing fluorine atoms)

Preliminary Model using 11K graphs



Conclusion and Perspectives

- Preliminary results show that training on 11k graphs was sufficient to achieve better performance compared to models using only atom and bond features.
- Future work will involve utilizing the full 44K molecule dataset, extending the approach to additional datasets, and applying it across diverse machine learning architectures to enhance property predictions

Reference

1. Meza-González, B., Ramírez-Palma, D.I., Carpio-Martínez, P., 2024.. Quantum Topological Atomic Properties of 44K molecules.
2. Żurański, A.M., Wang, J.Y., Shields, B.J., Doyle, A.G., 2022. Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules.
3. Chevreau, H., Pilmé, J., 2023. Promising insights in parallel grid-based algorithms for quantum chemical topology.

Acknowledgement

This project is funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101120466, as part of the AiChemist Doctoral Network within the Marie Skłodowska-Curie Actions (MSCA-DN).