AiChemist meeting
25.04.2025, Lausanne

Subashini Kennedy
Supervisors: Prof. Rodolphe Vuilleumier, Dr. Marc Bianciotto

Can quantum approach to QSAR improve the model's performance ?

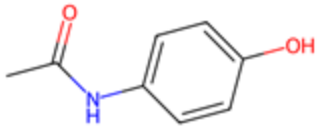
SMILES

CC(=O)OC1=CC=CC=C1C(=O)O

RDKit

Open-Source Cheminformatics and Machine Learning

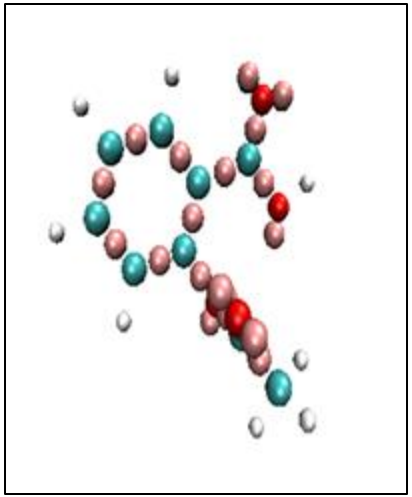
Molecular Graphs → Atom and bond features → Graph Neural Network



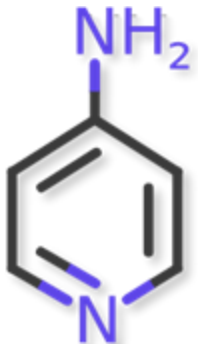
Auto-QChem

String Representations | DFT Calculations | Cloud Database | Web Interface python

TopChem2 →

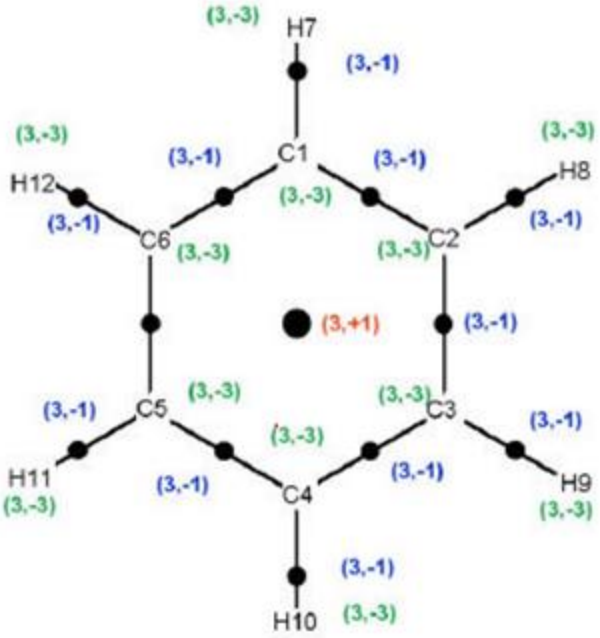


Critical points and including them in ML models



Types of Critical points:

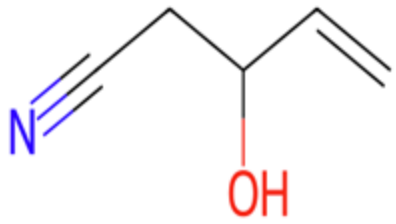
- Nuclear (3, -3)
- Bond (3, -1)
- Ring (3, 1)
- Cage (3, 3)



Baseline model with atom and bond features for comparison

```
In [80]: mols[10]
```

Out[80]:




```
In [81]: g=convert_mol_to_graph(mols[10])
         g.x,g.edge_index,g.edge_attr
```

```
Out[81]: (tensor([[8., 0., 0., 4., 0., 1., 1.],
                  [6., 0., 0., 4., 0., 3., 1.],
                  [6., 0., 0., 4., 0., 2., 2.],
                  [6., 0., 0., 2., 0., 2., 0.],
                  [7., 0., 0., 2., 0., 1., 0.],
                  [6., 0., 0., 3., 0., 2., 1.],
                  [6., 0., 0., 3., 0., 1., 2.]]),
          tensor([[0, 1, 2, 3, 1, 5],
                  [1, 2, 3, 4, 5, 6]]),
          tensor([[1., 0., 0.],
                  [1., 0., 0.],
                  [1., 0., 0.],
                  [3., 0., 0.],
                  [1., 0., 0.],
                  [2., 0., 0.]]))
```

Atom type
Is in ring
Is aromatic
Hybridization
Charge
No. of bonds associated
No. of bonded hydrogen atoms

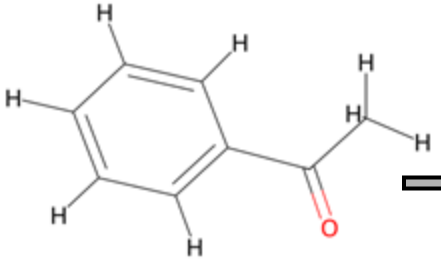
Node tensor

Edge tensor

Bond type
Conjugated
Is in ring

Edge attributes

Automated AutoQchem + topchem2 workflow to extract quantum data



Acetophenone

DFT calculations

.sh,.gjf,.chk,.log
+
.wfn

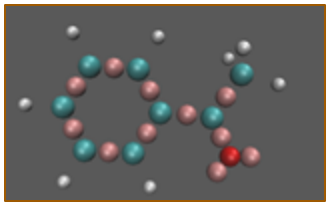
Mongo database

Quantum descriptors

Graph Neural Network

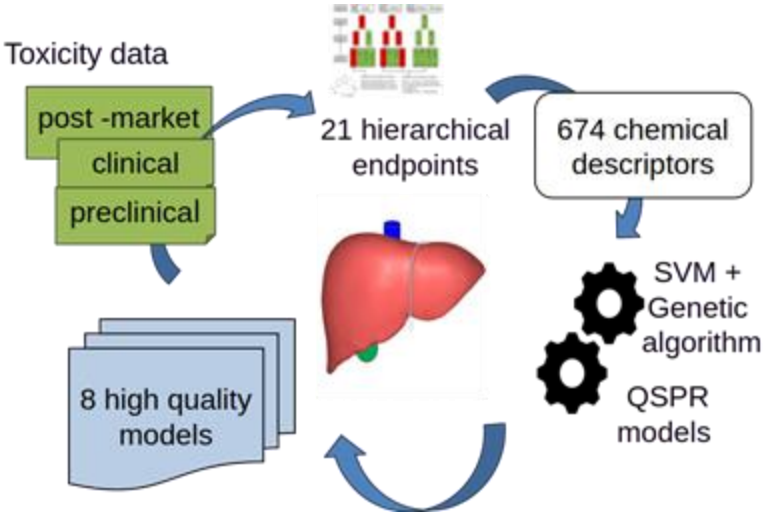
TopChem2

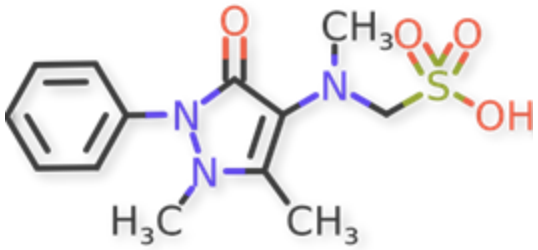
Output.txt
Cpelf.xyz
data.sum
cube files…

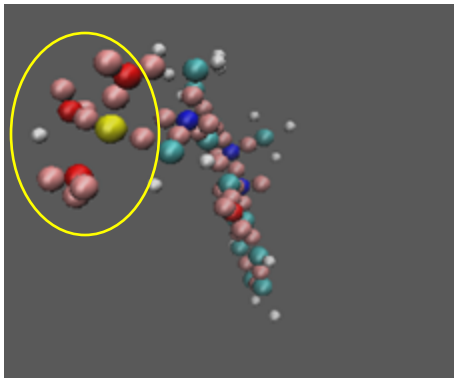


Cpelf.xyz

Sanofi dataset with preclinical and clinical hepatotoxicity







- **Size**: ~3,800 molecules
- **Molecular composition**: Drug molecules (C, H, N, O, S, Cl..)
- **Computed with**: DFT calculation using B3LYP/6-31G(d,p) with PCM
- **QM properties:** 19 including HOMO-LUMO energy, HOMO-LUMO gap

Smaller dataset, Complex molecules, topchem issues Alternatives?

Converting json files to graph tensors

Node tensor :

Atoms	Z value	0	0	0	0
Critical points	group	Vol.	Pop.	Charge	Value

Edge tensor:

Atoms ←→ Critical points

Edge attributes:

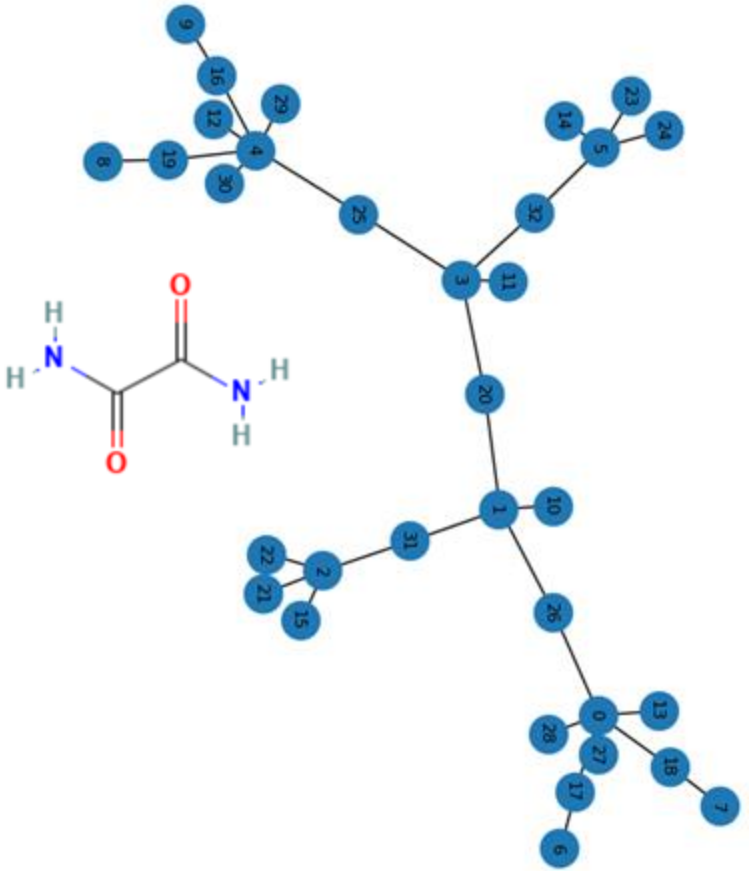
None

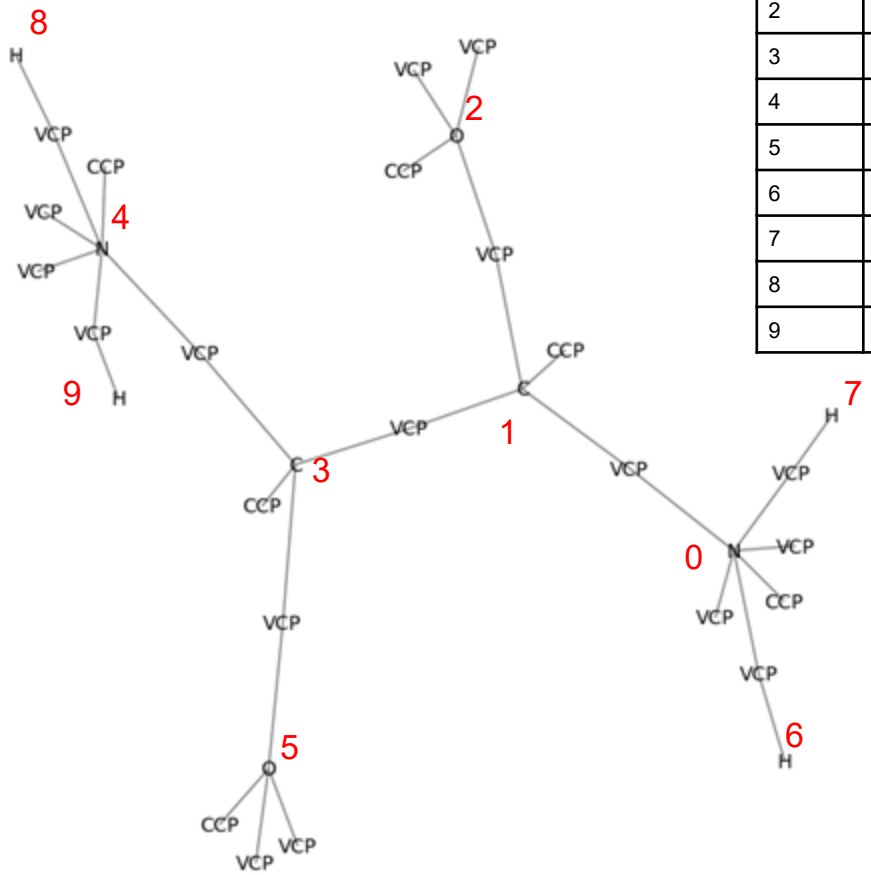
```
Data(x=[33, 5], edge_index=[2, 32], y=[1], z=[33])
Node Features (x): tensor([[  7.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  6.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  8.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  6.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  7.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  8.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  1.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  1.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  1.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  1.0000,   0.0000,   0.0000,   0.0000,   0.0000],
        [  0.0000,   0.8200,   2.0884,   3.9116,   1.0000],
        [  0.0000,   0.8200,   2.0873,   3.9127,   1.0000],
        [  0.0000,   0.4430,   2.1090,   4.8910,   1.0000],
        [  0.0000,   0.4430,   2.1050,   4.8950,   1.0000],
        [  0.0000,   0.2610,   2.1165,   5.8835,   1.0000],
        [  0.0000,   0.2610,   2.1061,   5.8939,   1.0000],
        [ -3.0000, 106.1320,   1.9479,  -0.9479,   0.9994],
        [ -3.0000, 106.1480,   1.9486,  -0.9486,   0.9994],
        [ -3.0000,  82.3420,   1.9583,  -0.9583,   0.9994],
        [ -3.0000,  82.3380,   1.9545,  -0.9545,   0.9994],
        [ -3.0000,  40.7080,   2.3423,  -2.3423,   0.9651],
        [ -3.0000, 103.2180,   2.7167,  -2.7167,   0.9192],
        [ -3.0000,  86.4900,   2.6883,  -2.6883,   0.9192],
        [ -3.0000, 103.2070,   2.7286,  -2.7286,   0.9181],
        [ -3.0000,  86.5030,   2.6806,  -2.6806,   0.9181],
        [ -3.0000,  16.9730,   2.0010,  -2.0010,   0.9156],
        [ -3.0000,  16.9490,   2.0048,  -2.0048,   0.9156],
        [ -3.0000,  55.8650,   0.9551,  -0.9551,   0.8784],
        [ -3.0000,  55.7980,   0.9567,  -0.9567,   0.8784],
        [ -3.0000,  55.8540,   0.9526,  -0.9526,   0.8783],
        [ -3.0000,  55.8020,   0.9591,  -0.9591,   0.8783],
        [ -3.0000,  34.0750,   2.2896,  -2.2896,   0.8626],
        [ -3.0000,  34.0730,   2.2816,  -2.2816,   0.8626]])
Shape: torch.Size([33, 5])
Edge Index: tensor([[10, 11, 12, 13, 14, 15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 20, 21, 22,
         23, 24, 25, 25, 26, 26, 27, 28, 29, 30, 31, 31, 32, 32],
        [ 1,  3,  4,  0,  5,  2,  9,  4,  6,  0,  7,  0,  8,  4,  1,  3,  2,  2,
          5,  5,  3,  4,  1,  0,  0,  0,  4,  4,  1,  2,  3,  5]])
Shape: torch.Size([2, 32])
HOMO Value (y): tensor([-0.2422])
```

Index	atoms
0	7 → Nitrogen
1	6 → Carbon
2	8 → Oxygen
3	6 → Carbon
4	7 → Nitrogen
5	8 → Oxygen
6	1 → Hydrogen
7	1 → Hydrogen
8	1 → Hydrogen
9	1 → Hydrogen

Converting json files to graph tensors

Index	atoms
0	7 → Nitrogen
1	6 → Carbon
2	8 → Oxygen
3	6 → Carbon
4	7 → Nitrogen
5	8 → Oxygen
6	1 → Hydrogen
7	1 → Hydrogen
8	1 → Hydrogen
9	1 → Hydrogen





Database with 44K wfx files

Data Descriptor | Open access | Published: 29 August 2024

Quantum Topological Atomic Properties of 44K molecules

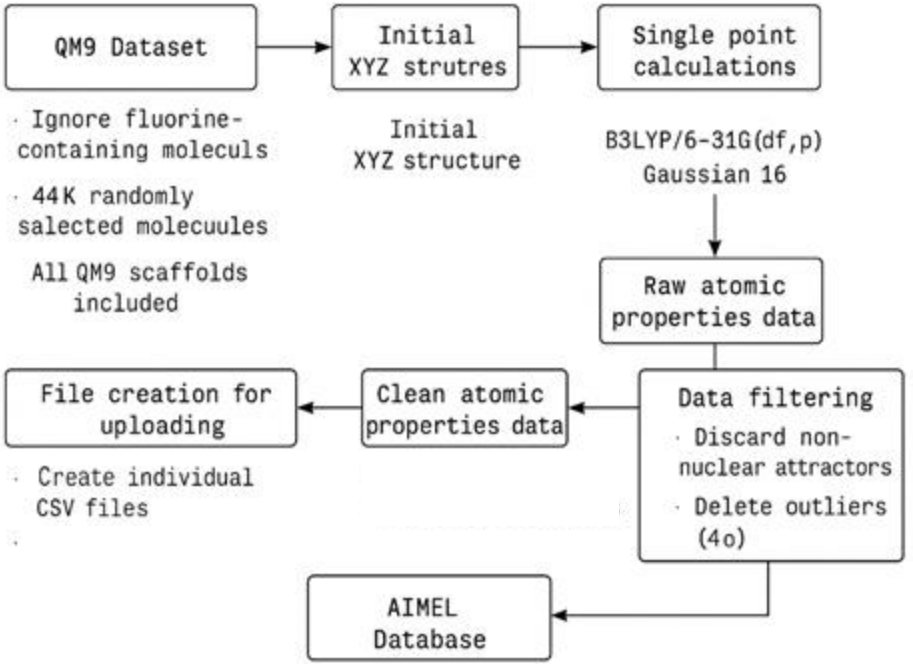
Brandon Meza-González, David I. Ramírez-Palma, Pablo Carpio-Martínez, David Vázquez-Cuevas, Karina Martínez-Mayorga & Fernando Cortés-Guzmán

Scientific Data 11, Article number: 945 (2024) | Cite this article

1531 Accesses | 1 Citations | Metrics

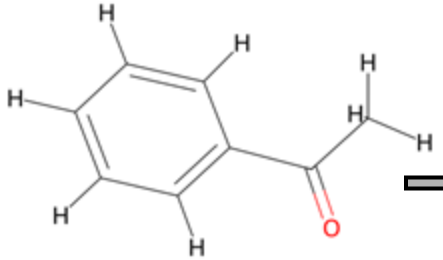
Abstract

We present a data set of quantum topological properties of atoms of 44K randomly selected molecules from the GDB-9 data set. These atomic properties were obtained as defined by the quantum theory of atoms in molecules (QTAIM) within an atomic basin, a region of real space bounded by zero-flux surfaces in the electron density gradient vector field. The wave function files were generated through DFT static calculations (B3LYP/6-31G), and the atomic properties were calculated using QTAIM. The calculated atomic properties include the energy of the atomic basin, the electronic population, the magnitude of the total dipole moment, and the magnitude of the total quadrupole moment. The atomic properties allow one to understand the chemical structure, reactivity, and molecular recognition. They can be incorporated into force fields for molecular dynamics or for predicting reactive sites. We believe that this data set could facilitate new studies in chemical informatics, machine learning applied to chemistry, and computational molecular design.



Gas Phase

~~Automated AutoQchem~~ + topchem2 workflow to extract quantum data



Acetophenone

DFT calculations

.sh,.gjf,.chk,.log
+
.wfn

Mongo database

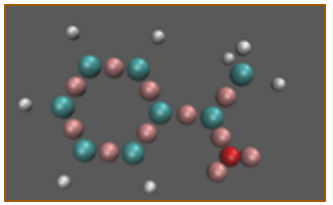
Quantum descriptors

Skipping DFT

Graph Neural Network

TopChem2

Output.txt
Cpelf.xyz
data.sum
cube files…



Cpelf.xyz

Evaluating Critical points GNN Performance in Predicting HOMO Energies

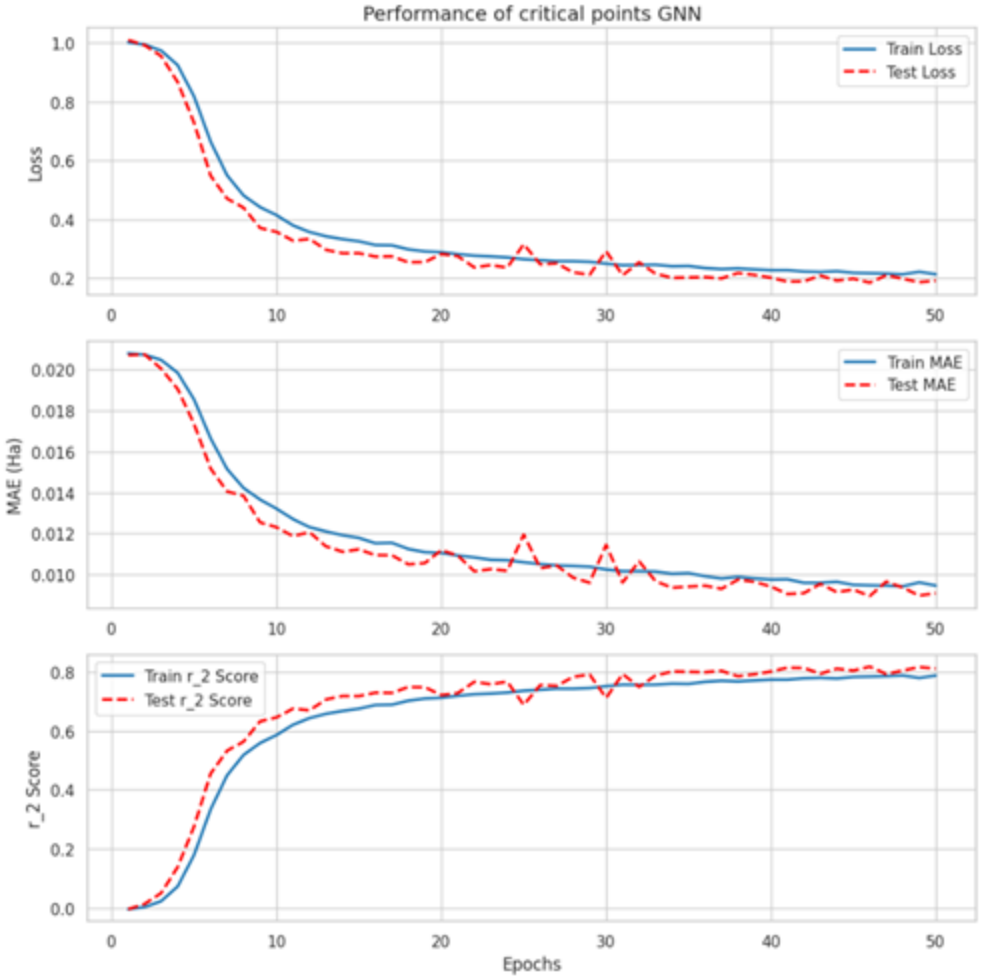
Graphs generated: 11004

Training graphs: 8803 (80%)
Testing graphs: 2201 (20%)

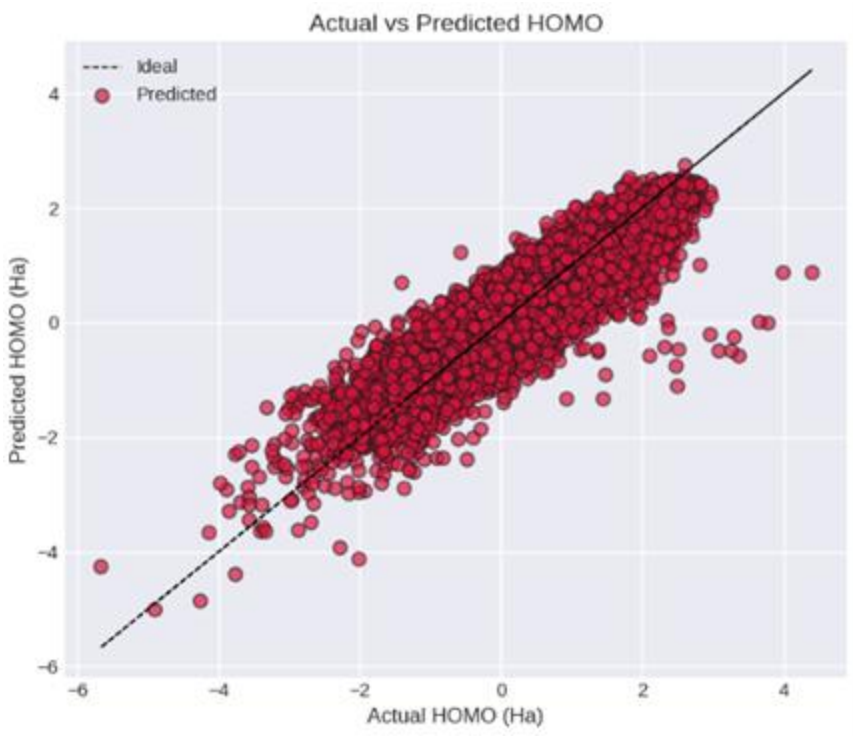
Exciting models:

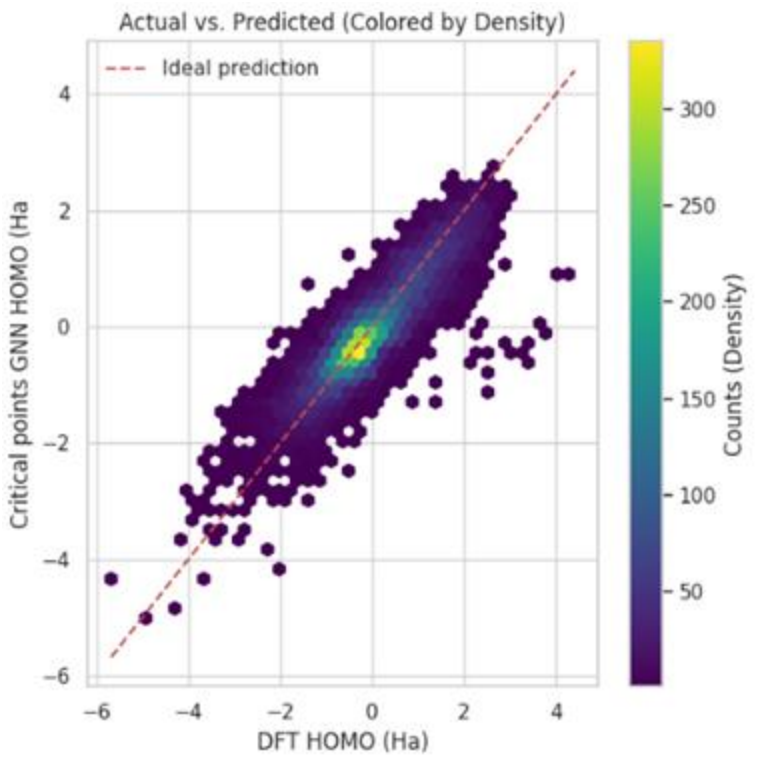
MAE < 0.00184 (Ha) or < 0.05 eV

r_2 Score: 0.75 – 0.90

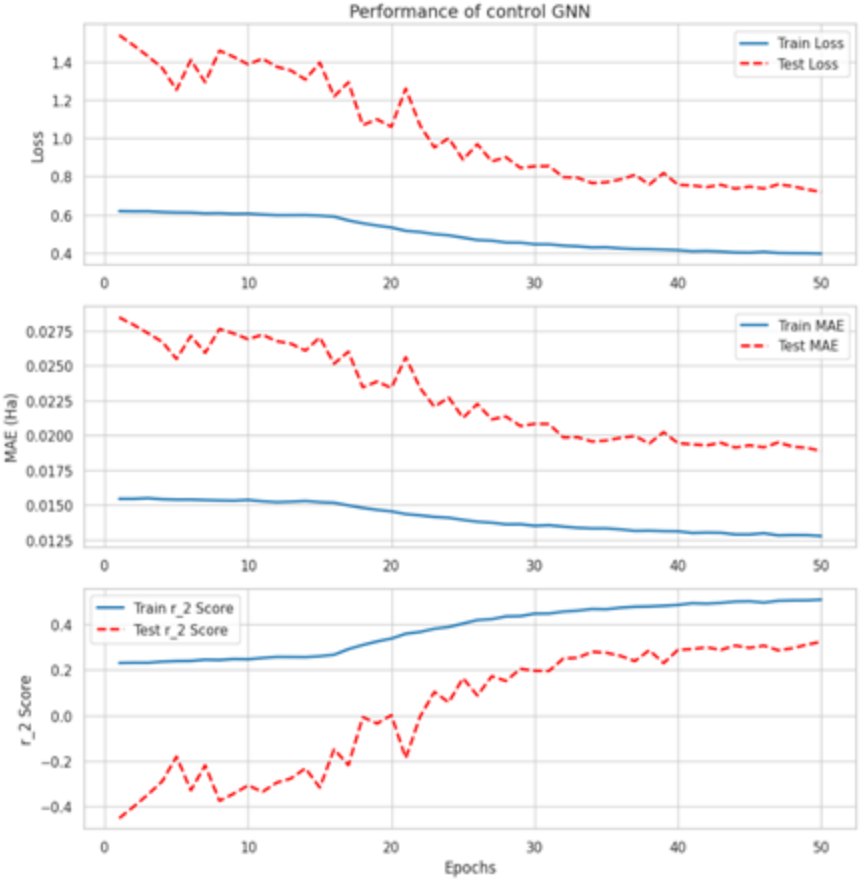


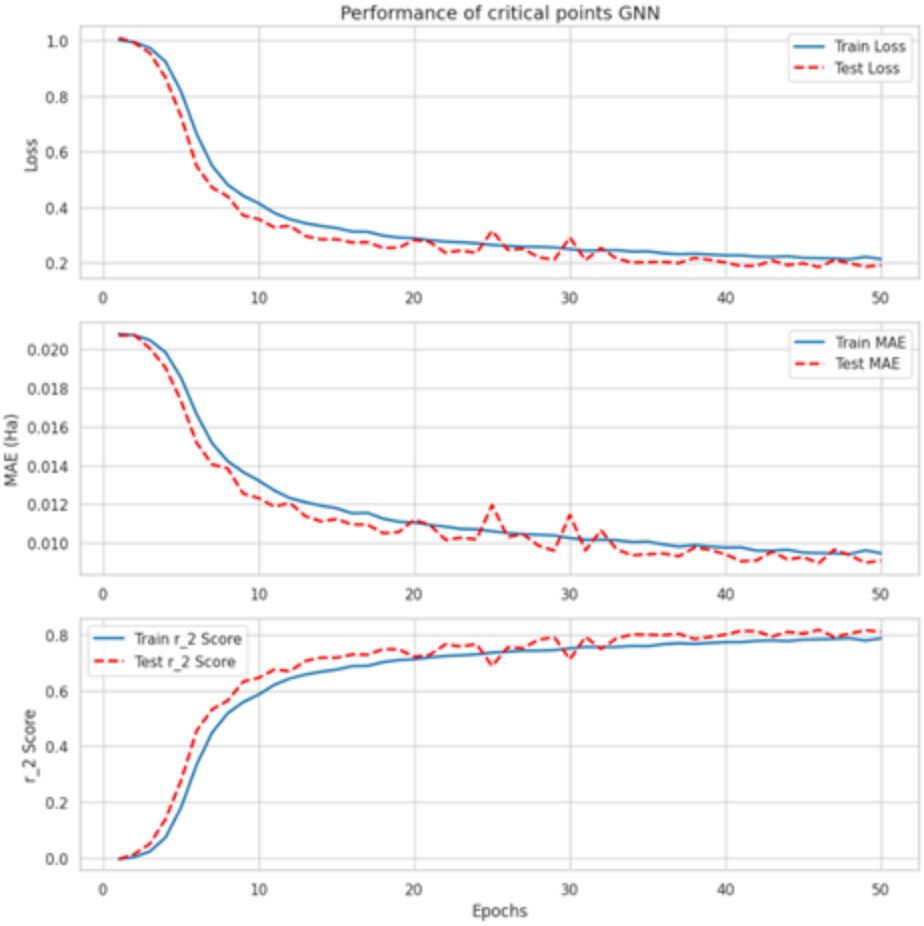
Evaluating Critical points GNN Performance in Predicting HOMO Energies





Control vs Critical points GNN





GCNs trained on critical points shows improved performance :11004 graphs

GNN	TRAIN			TEST		
Property	**MAE**	**RMSE**	**R2 Score**	**MAE**	**RMSE**	**R2 Score**
HOMO (Ha)	0.0091	0.0118	0.8117	0.0088	0.0116	0.8085
LUMO(Ha)	0.0142	0.0184	0.8523	0.0148	0.0191	0.8410
Gap (Ha)	0.0176	0.0227	0.7969	0.0174	0.0228	0.8004
R^2 (a_0^2)	124.252	169.101	0.430	130.58	179.356	0.399
Dipole (D)	0.834	1.202	0.431	0.828	1.126	0.459
Alpha (a_0^3)	3.878	5.094	0.654	3.919	5.040	0.644
ZPVE (Ha)	0.0075	0.0091	0.9094	0.0075	0.0092	0.9071
U0 (Ha)	16.645	22.476	0.748	16.216	21.935	0.753
U298 (Ha)	16.731	22.488	0.746	16.754	22.387	0.749
H298 (Ha)	21.771	26.190	0.654	22.080	26.580	0.654
G298 (Ha)	15.545	21.162	0.776	15.395	20.876	0.780
Cv (cal/mol K)	1.507	2.039	0.728	1.586	2.243	0.676

Control	TRAIN			TEST		
Property	**MAE**	**RMSE**	**R2 Score**	**MAE**	**RMSE**	**R2 Score**
HOMO (Ha)	0.0123	0.0166	0.5296	0.0184	0.0225	0.3502
LUMO (Ha)	0.0230	0.0299	0.6001	0.0287	0.0347	0.3443
Gap (Ha)	0.0220	0.02278	0.6779	0.0205	0.0259	0.1590
R^2 (a_0^2)	126.297	179.767	0.388	265.613	281.036	-4.291
Dipole (D)	1.028	1.319	0.206	1.363	1.749	0.022
Alpha (a_0^3)	4.353	6.116	0.487	7.995	8.766	-0.939
ZPVE (Ha)	0.0111	0.0145	0.7799	0.0124	0.0147	0.5969
U0 (Ha)	19.773	29.953	0.471	43.169	48.312	-1.836
U298 (Ha)	22.256	32.238	0.3888	47.741	52.849	-2.393
H298 (Ha)	24.681	35.185	0.270	37.935	43.533	-1.302
G298 (Ha)	20.682	30.624	0.447	43.410	49.194	-1.940
Cv (cal/mol K)	1.949	2.742	0.553	3.772	4.044	-1.119

Steps forward:

Critical points data → HOMO prediction

- Include all 40K molecules
- Investigate the node embeddings
- Hyperparameter tuning
- Interpretability methods

Potential workflow:

Critical points data + atom,bond features from control GNN → HOMO prediction

Critical points data + other parameters → HOMO prediction (multi-task regression)

Complex architectures ex. GAT
Pretrain and finetune for drug molecules

Thanks for listening!