Data-Driven Electronic Structure Analysis of Metal-Organic Frameworks

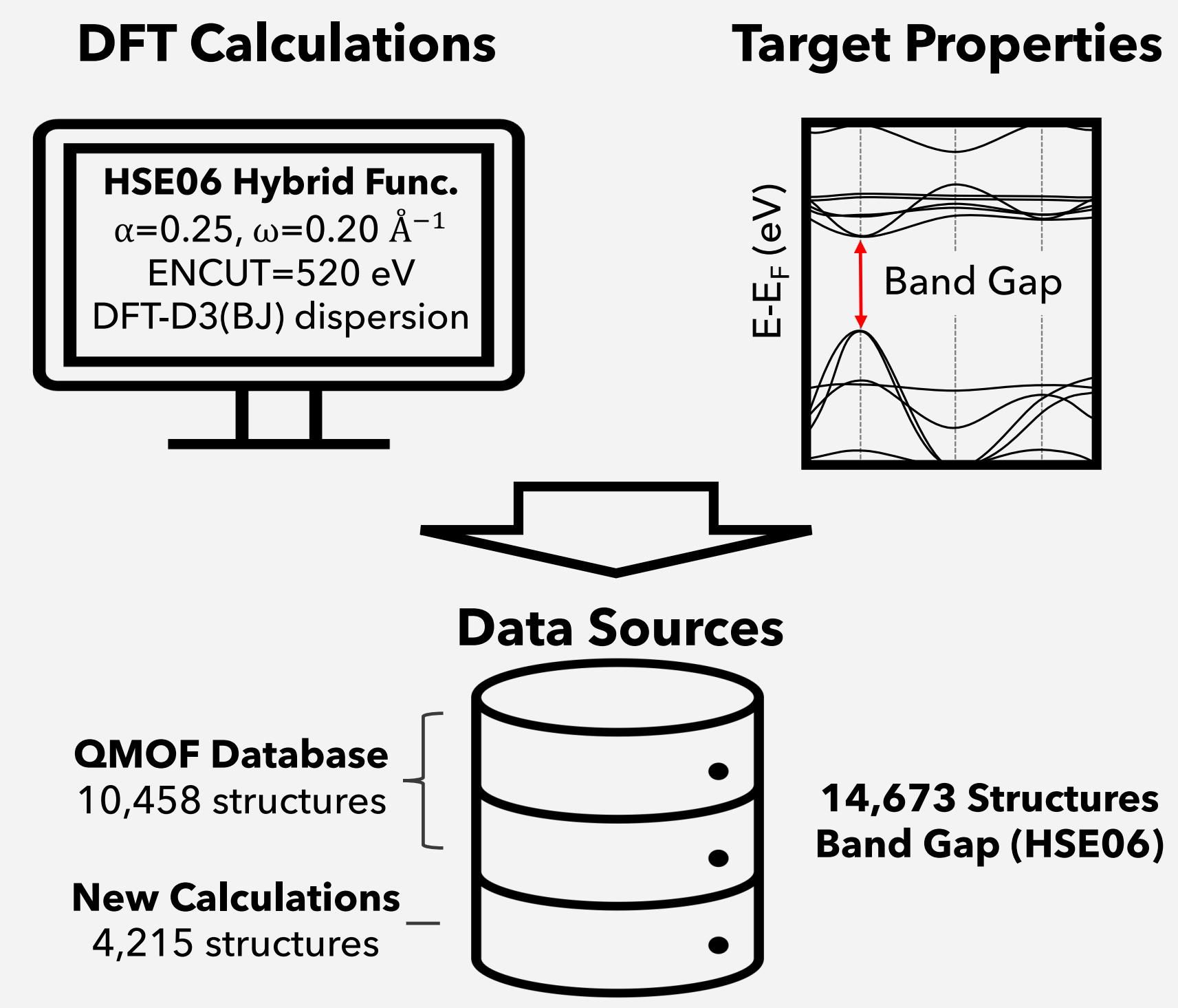
Minhyuk Kang[1], Seung-Jae Shin[2], Tianshu Li[3], Aron Walsh[3]

[1] Department of Chemical Engineering, Imperial College London, London, United Kingdom
[2] School of Energy and Chemical Engineering, UNIST, Ulsan, Republic of Korea
[3] Department of Materials, Imperial College London, London, United Kingdom

Introduction ...

⊟ Motivation

- **Problem:** Prior MOF band gap ML relies on PBE labels, which systematically underestimate E_g and hinder real-world screening
- **Solution:** We assembled a high-fidelity HSE06 dataset of ~15,000 MOFs
- **Composition:** 10,458 entries sourced from QMOF[1] + 4,215 newly calculated structures
- **Impact:** Provides accurate electronic labels to train fairer models, improving identification of semi-conductive MOFs and reducing PBE-induced bias

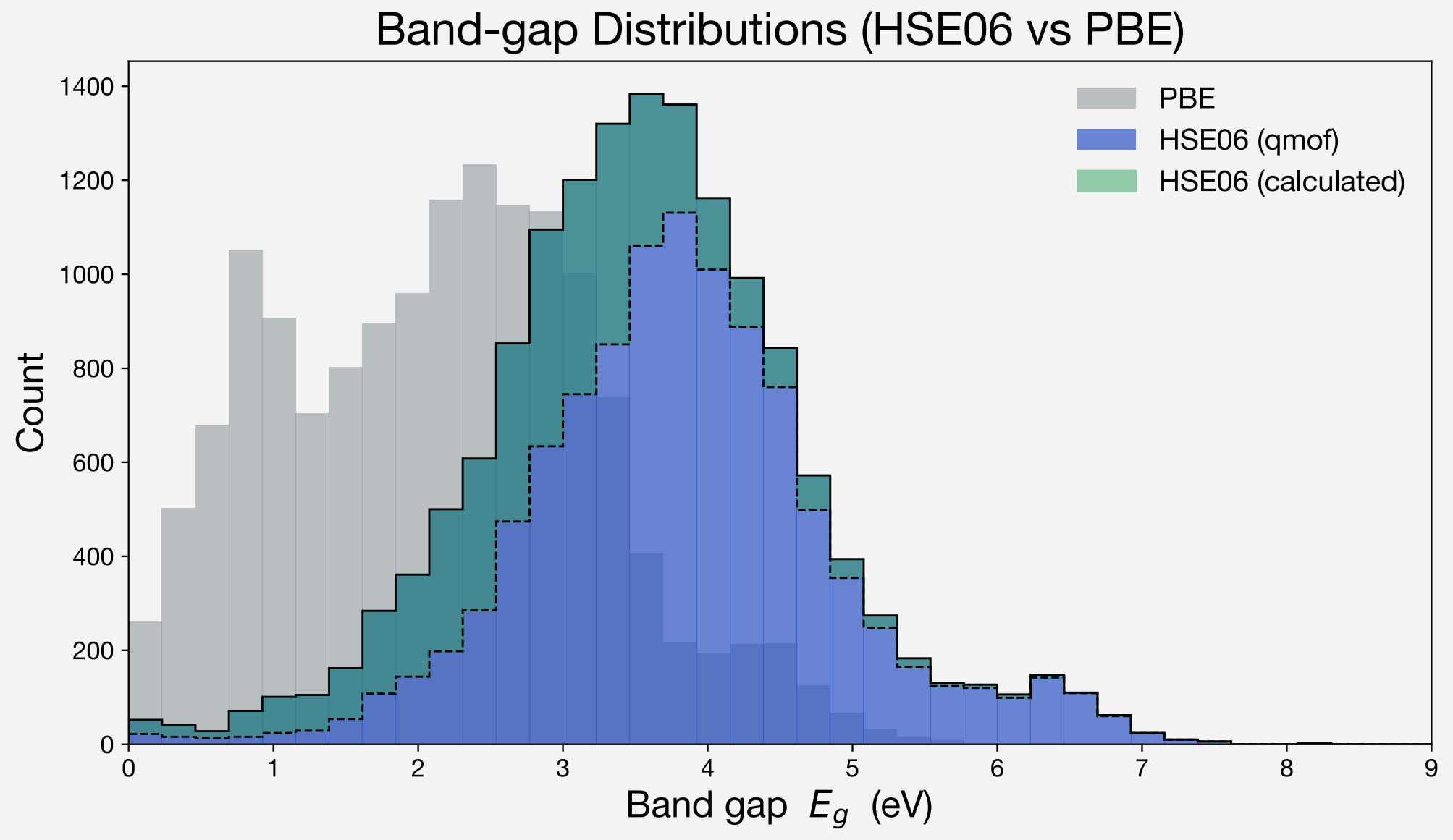


⊟ Dataset Overview

HSE06 vs. PBE Systematic Bias Correction

- **PBE Limitation:** Distribution clearly shows the underestimation of band gaps across entire domain
- **Largest errors:** Low band gap MOFs (E_g < 1.5eV) show 0.5-1.5eV discrepancies
- **HSE06 Solution:** Hybrid functional corrects bias and enables accurate ML training

Band-gap Distributions (HSE06 vs PBE)



Enabling generalizable model training

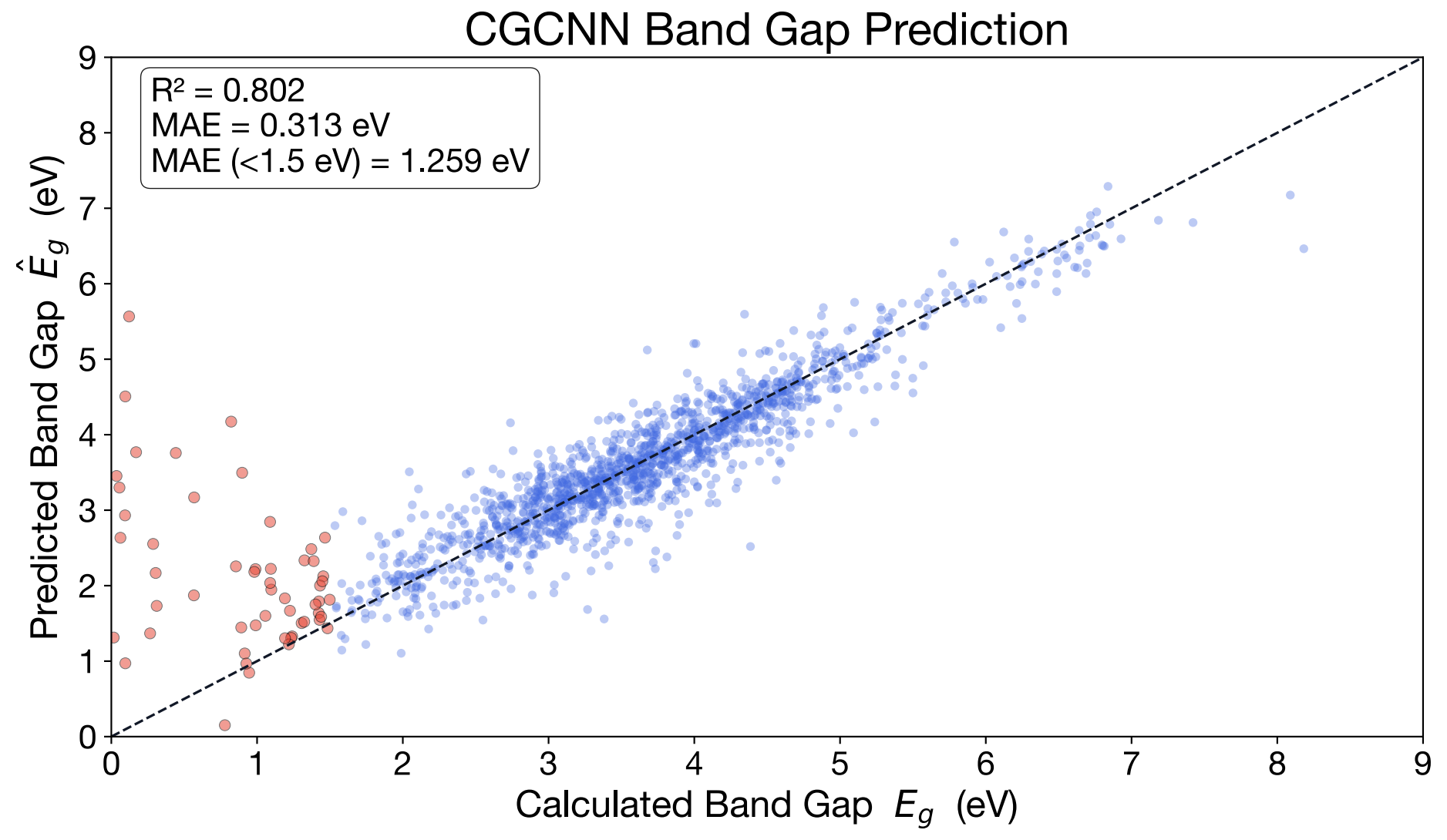
- **Complete spectrum:** 14,673 MOFs spanning 0–8 eV
- **Balanced extremes:**
  - Near-conductive: 472 structures (E_g < 1.5 eV)
  - Insulating: 468 structures (E_g > 6 eV)
- **Peak region:** Majority at 2-4 eV (semiconductors)
- **Why it matters:** Prevents model overfitting to central peak; captures underrepresented low-gap materials

Initial Analysis & Methodology ✕ ...

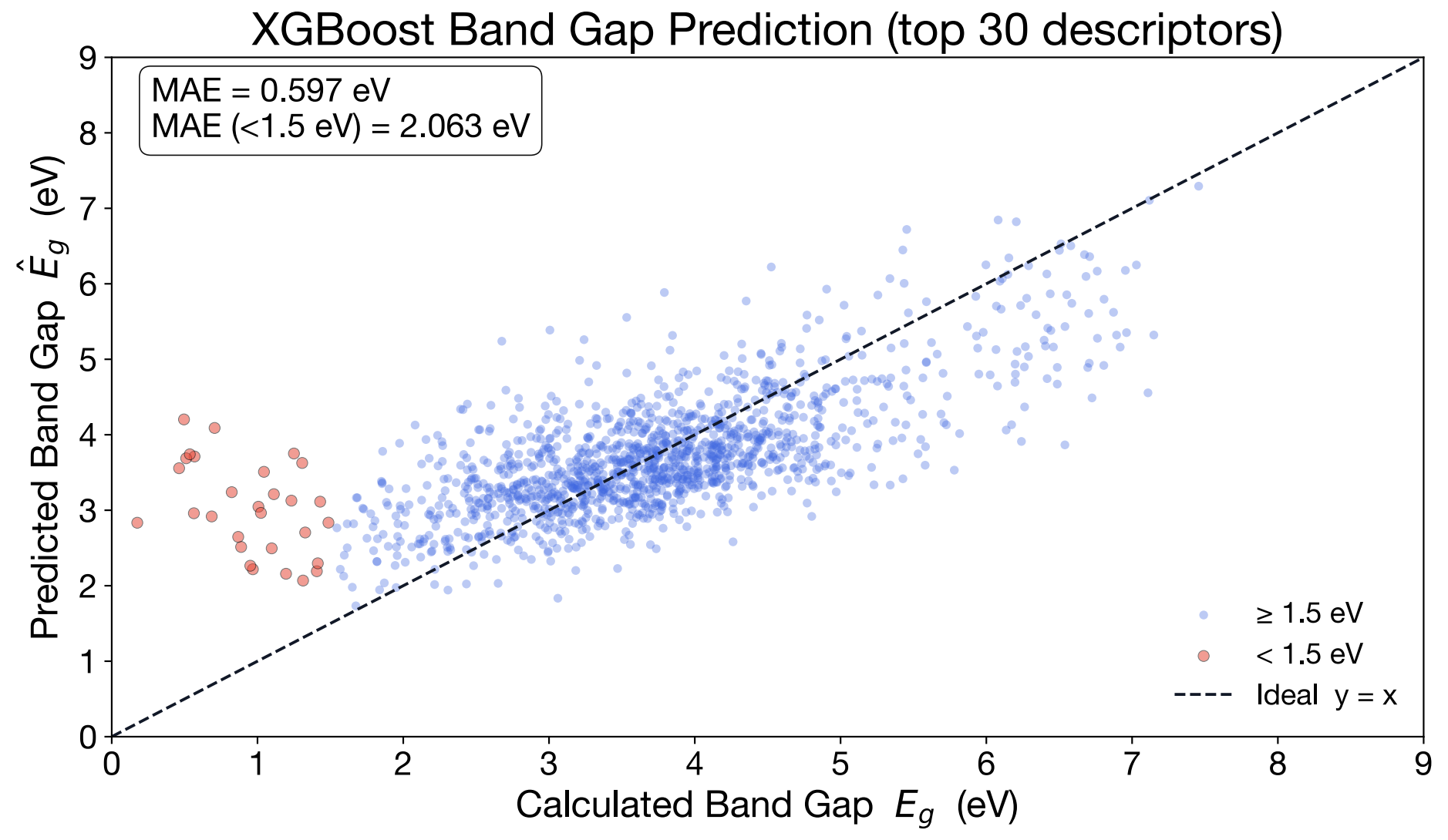
# Initial Analysis of ML Performance

- **GNN baseline:** CGCNN[2], a foundational crystal GNN model, was first evaluated for band gap prediction
- **Systematic limitation:** Despite strong global correlation, CGCNN exhibits consistent overestimation and increased error within the low-band-gap regime (E_g < 1.5 eV)
- **Interpretable approach:** Gradient-boosted decision trees (XGBoost) were employed to elucidate key structural and geometric features
- **Descriptor set:** Feature selection encompassed radial distribution function (RDF) peaks, crystallographic symmetry parameters, pore metrics, bond lengths, and coordination shell statistics
- **Feature contributions:** Short-range RDFs and symmetry descriptors emerged as most important, yet geometry-only features proved insufficient for accurate prediction of low-gap MOFs

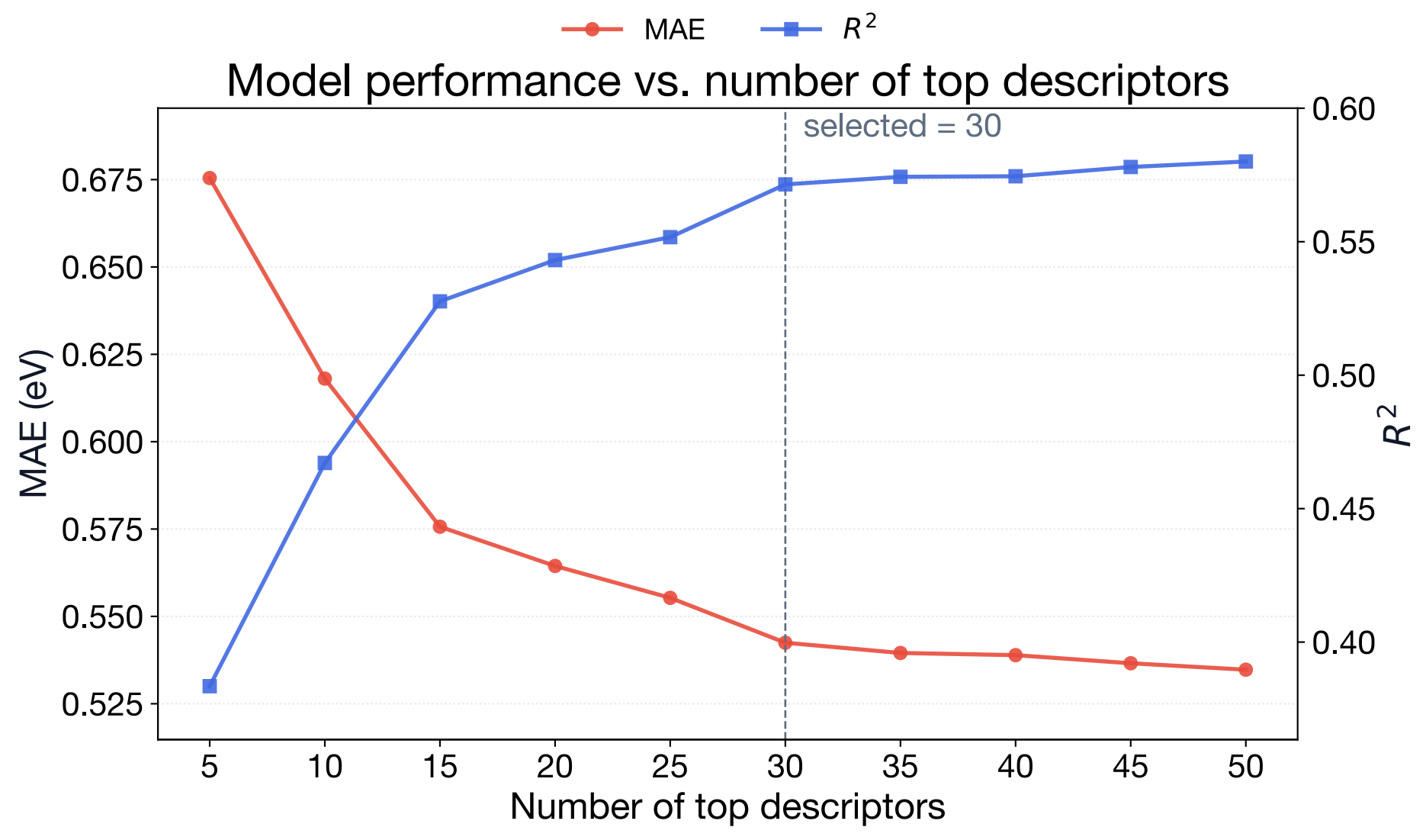
CGCNN Band Gap Prediction



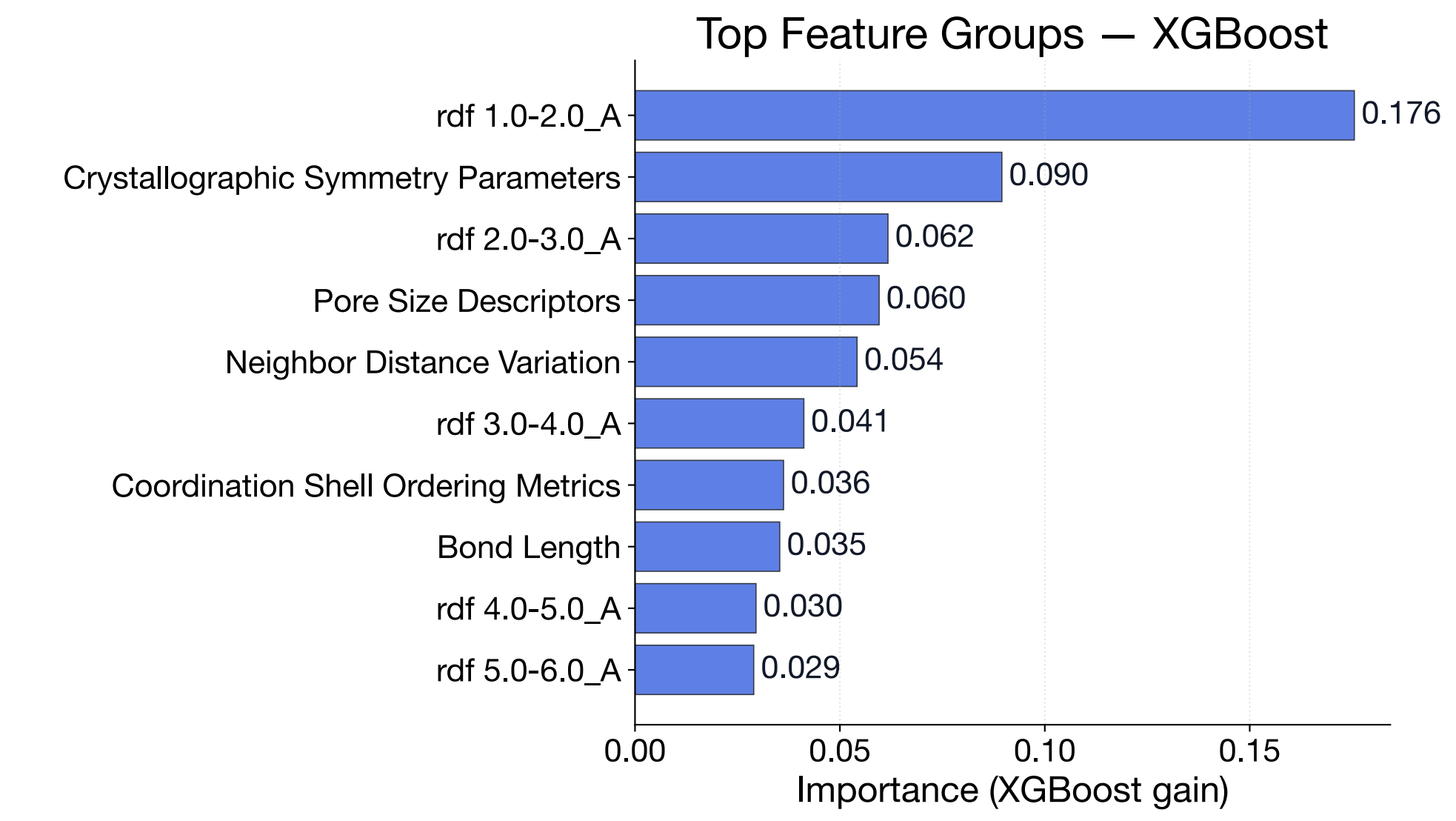
XGBoost Band Gap Prediction (top 30 descriptors)



Model performance vs. number of top descriptors

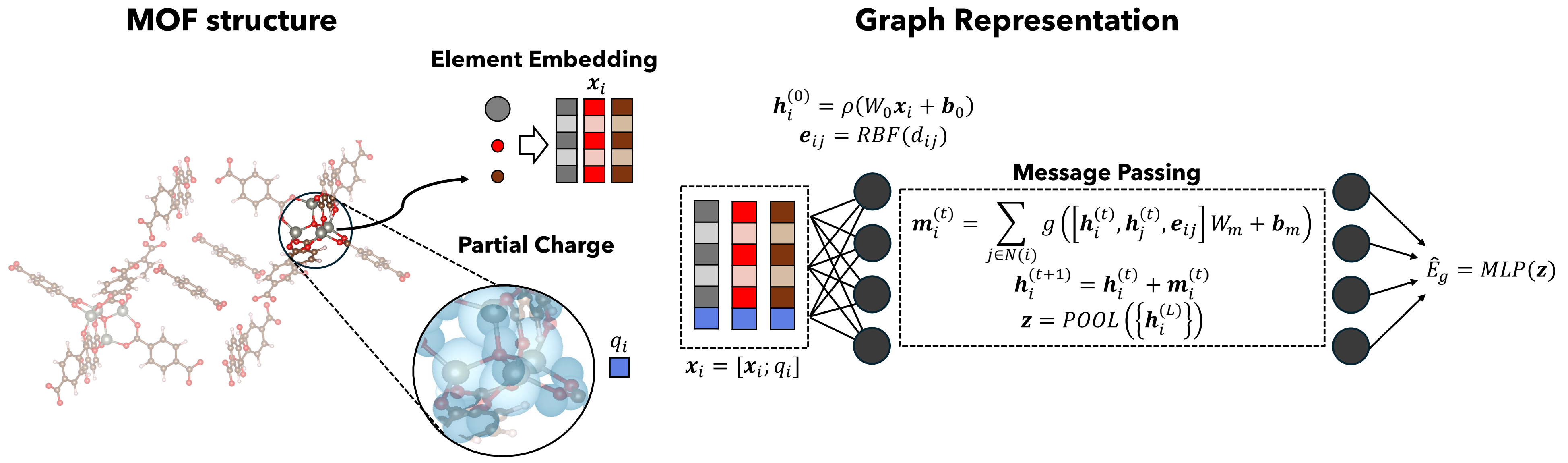


Top Feature Groups — XGBoost



# Model Description

- **Physics-informed node features:** Use per-atom partial charges predicted by PACMAN[3] to encode electrostatics and donor-acceptor character, which can be key for near-conductive MOFs
- **Graph construction:** Each atom's feature vector x_i is extended to $[x_i; q_i]$, directly integrating partial charge information into message passing and global pooling layers
- **Comparative methodology:** Benchmark scalar addition strategies by appending single-valued elemental features in the same fashion as charge, to evaluate the impact of different node augmentations.


MOF structure
Graph Representation
Element Embedding
$\boldsymbol{x}_i$
Partial Charge
q_i
$\boldsymbol{h}_i^{(0)} = \rho(W_0 \boldsymbol{x}_i + \boldsymbol{b}_0)$
$\boldsymbol{e}_{ij} = RBF(d_{ij})$
Message Passing
$\boldsymbol{m}_i^{(t)} = \sum_{j \in N(i)} g\left(\left[\boldsymbol{h}_i^{(t)}, \boldsymbol{h}_j^{(t)}, \boldsymbol{e}_{ij}\right] W_m + \boldsymbol{b}_m\right)$
$\boldsymbol{h}_i^{(t+1)} = \boldsymbol{h}_i^{(t)} + \boldsymbol{m}_i^{(t)}$
$\boldsymbol{z} = POOL\left(\left\{\boldsymbol{h}_i^{(L)}\right\}\right)$
$\boldsymbol{x}_i = [\boldsymbol{x}_i; q_i]$
$\hat{E}_g = MLP(\boldsymbol{z})$

Results & Conclusion ✕ ...

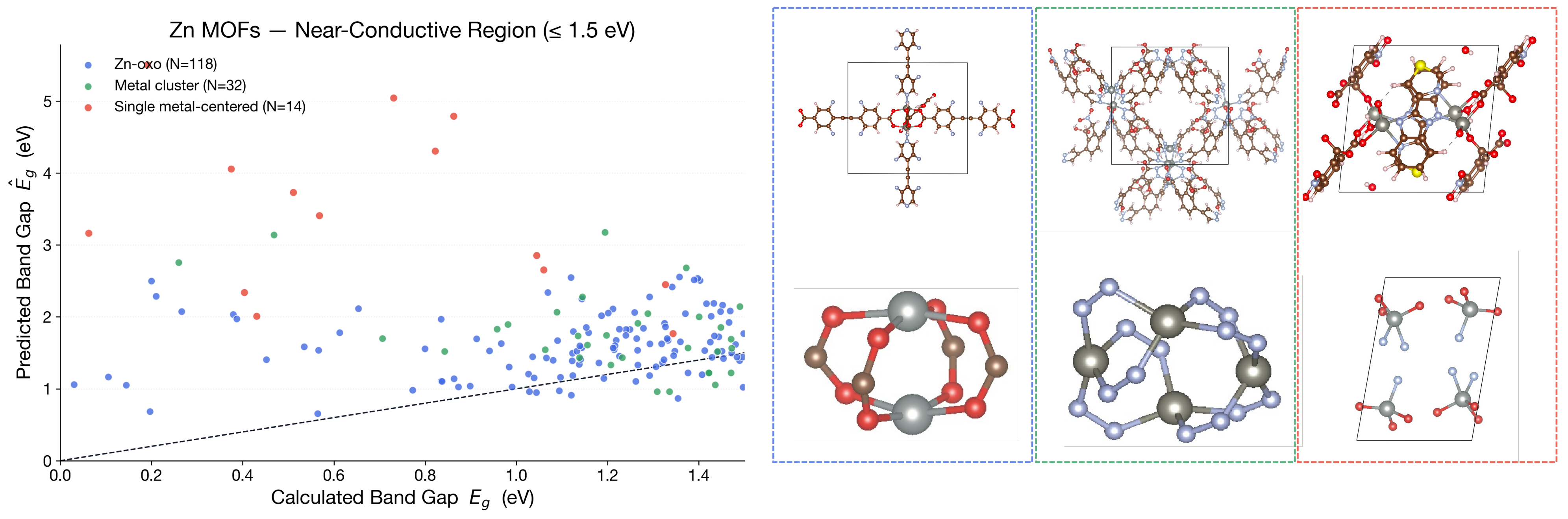
# Results

- **Charge embedding impact:** Incorporation of partial atomic charges produced the highest global accuracy and the lowest mean absolute error (MAE) for low-band-gap MOFs among all tested scalar augmentations, demonstrating that electrostatics-aware representations deliver information beyond geometric descriptors
- **Low-gap challenge remains:** Despite these gains, models continue to systematically overpredict for E_g < 1.5 eV, indicating the need for further integration of physical mechanisms to fully resolve conductive MOF prediction

Embedding Feature	MAE	MAE (E_g<1.5eV)	R^2
Charge	**0.2879 ± 0.0027**	**1.1492 ± 0.1362**	**0.8320 ± 0.0150**
Nf Unfilled	0.3064 ± 0.0088	1.2482 ± 0.0423	0.8161 ± 0.0244
Default	0.3067 ± 0.0058	1.1774 ± 0.1162	0.8198 ± 0.0183
GS Magnetic Moment	0.3077 ± 0.0118	1.2191 ± 0.1284	0.8198 ± 0.0209
GS Band Gap	0.3094 ± 0.0018	1.2319 ± 0.0650	0.8181 ± 0.0114
Np Unfilled	0.3099 ± 0.0073	1.2137 ± 0.0350	0.8189 ± 0.0175
Nd Unfilled	0.3110 ± 0.0077	1.2015 ± 0.0206	0.8178 ± 0.0141
GS Volume per Area	0.3146 ± 0.0110	1.2303 ± 0.1534	0.8086 ± 0.0215
Average	0.3077 ± 0.0099	1.2161 ± 0.0883	0.8179 ± 0.0163

# Case Study: Zn MOFs

- **Dataset focus:** Near-conductive Zn-MOFs (E_g < 1.5 eV) classified into Zn-oxo bridges (~71%), metal clusters (~20%), and single-metal-centered motifs (~9%)
- **Motif-specific results:** Zn-oxo structures yield the tightest residuals, clusters show moderate prediction error, single-metal-centered motifs exhibit the largest dispersion
- **Interpretation:** Predictive difficulty intensifies with increasing frontier orbital localization: Zn-oxo < clusters < single-center, highlighting underrepresented ligand-field and charge transfer effects in isolated centers, motivating further physical model enhancements



# Conclusion

- **Electrostatics-aware GNNs:** Integrating per-atom charge embeddings generally improves band gap prediction for MOFs, outperforming purely geometry-based or elemental scalar augmentation strategies
- **Persistent challenge:** Systematic overestimation remains for highly localized, low-gap cases—especially in isolated metal-center motifs—highlighting the need for further incorporation of ligand-field effects and advanced electronic descriptors
- **Outlook:** Ongoing work will focus on embedding deeper physical priors and environment-dependent features to achieve robust, transferable predictions across diverse MOF chemistries

References

[1] Rosen AS, et al. *npj Comput. Mater.* 8, 112 (2022)
[2] Xie T, Grossman JC. *Phys. Rev. Lett.* 120, 145301 (2018)
[3] Zhao G, Chung YG. *J. Chem. Theory Comput.* 20, 5368-5380 (2024)

Acknowledgements

This project was undertaken as part of the Undergraduate Research Opportunities Programme (UROP) at Imperial College London, Department of Materials, in the Walsh Materials Design (WMD) group under Professor Aron Walsh. I also thank Professor Seung-Jae Shin and the THEMMES Group (UNIST) for valuable discussions.

Contact
minhyuk.kang21@imperial.ac.uk
Imperial College London,
Exhibition Rd, London, UK, SW7 2AZ
LinkedIn

Resources https://github.com/kang-minhyuk/mof-electronic-structure

IMPERIAL