Accurate Band Gap Prediction in Porous Materials using $\triangle$-Learning

Ashna Jose
Department of Materials
Imperial College London
London SW7 2AZ, United Kingdom
a.jose@imperial.ac.uk

Aron Walsh
Department of Materials
Imperial College London
London SW7 2AZ, United Kingdom
a.walsh@imperial.ac.uk

Abstract

Metal-organic frameworks (MOFs) are versatile materials with tunable crystal structures, morphologies, and chemistries, offering diverse physical and chemical properties. Although typically electrically insulating, specific combinations of organic and inorganic components can impart electrical conductivity to MOFs. The virtually limitless chemical space of MOFs, however, presents a significant challenge in identifying optimal candidates for electrochemical applications. Although Density Functional Theory (DFT) can probe their electronic structure, its high computational cost hinders the discovery of novel electroactive MOFs using machine learning due to limited data. To tackle these challenges, a semi-empirical extended tight binding approach (GFN1-xTB) is employed to compute electronic properties of a dataset of MOFs, and it is shown that GFN1-xTB approximates MOF band gaps well, as compared to semi-local DFT. This data is used to train an interpretable Δ-learning model that predicts the difference between low and high fidelity band gaps, given by xTB and DFT data at the hybrid level, respectively. This model outperforms direct models trained using only the DFT values. With limited high-quality DFT band gaps, taking advantage of Δ-learning using low-cost GFN1-xTB leads to better predictions as opposed to relying on DFT data alone.

1 Introduction

Metal-organic frameworks (MOFs) [1, 2] are formed through coordination bonds between metal ions and organic ligands. They show promise in a plethora of applications [3] such as gas capture [4, 5], energy storage [6, 7], catalysis [8, 9] and sensing [10]. They have also been exploited for applications in electrochemistry [11], such as for batteries [12], super-capacitors [13], and fuel cells [14, 15]. This versatility of MOFs is due to their highly porous and chemically tunable nature [16]. Further, electronic properties of MOFs can be altered by introducing ions or guest molecules in their porous framework [17, 18], which has fueled a growing interest in investigating these properties [19, 20, 21, 22]. More than 100,000 MOFs have been reported in the Cambridge Structural Database [23, 24] so far, and there exist various publicly available datasets with synthesized and hypothetical MOFs [25, 26, 27, 28]. However, the search for novel MOFs with specific properties becomes challenging due to the vastness of its design space.

Computational techniques such as density-functional theory (DFT) [29, 30, 31, 32] are often used to screen these datasets, although due to the high number of atoms in MOFs, this is computationally intensive, particularly for DFT at the hybrid level of theory, using which is unfeasible to explore a substantial portion of the MOF design space. In recent years, artificial intelligence (AI) has been utilized to accelerate MOF discovery [33, 34, 35, 36, 37, 38, 39, 40, 41]. However, deep machine learning (ML) architectures [42, 43, 44, 45, 46, 47, 48, 49] are well-suited for large datasets, and the lack of such datasets with high quality DFT data is a significant bottleneck in advancing this field.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

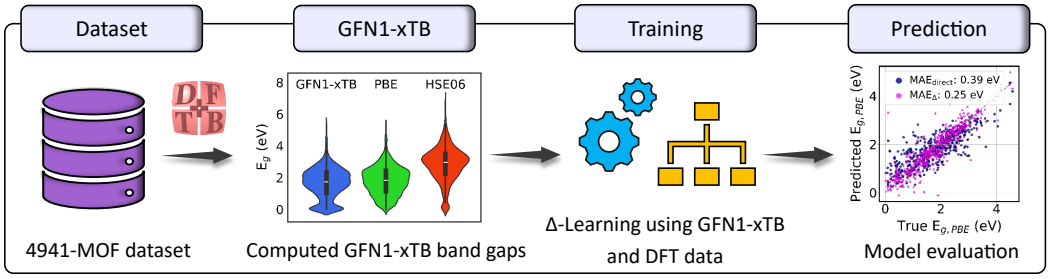

Figure 1: Schematic of the band gap prediction framework developed in this work.

In this context, semi-empirical methods like the self-consistent charge density functional tight binding (SCC-DFTB [50, 51]) balance accuracy and computational cost by approximating DFT through a Taylor expansion of the total energy around a reference electron density, significantly reducing computational cost by 2-3 orders of magnitude. However, its limited element coverage and low transferability from pairwise repulsive potentials restricts its applicability in high-throughput studies. Recently, Grimme et al. developed two extended tight-binding (xTB) methods for calculating structures, vibrational frequencies, and non-covalent interactions (GFNn, n = 1, 2) [52, 53] of large molecular systems. These methods include precomputed parameters for each atomic species, encompassing both global and element-specific parameters for most elements up to atomic number Z = 86. Thus, the total number of parameters are fixed and no pairwise parameters are required, making it attractive for large-scale studies. The method has been used to investigate a wide range of properties in a variety of materials [54, 55, 56, 57, 58, 59]. GFNn-xTB has also shown promise in accurate geometry optimization [60], and in reliably approximating XRD patterns, textural characteristics, as well as electronic properties of 2D and 3D MOFs [61, 62].

In this work, GFN1-xTB is employed to compute electronic band gaps for a dataset of 4941 MOFs from the QMOF dataset [27]. GFN1-xTB band gaps are found to be in agreement with those obtained using DFT (PBE) for the majority of MOFs. Materials descriptors [63] are used to encode the MOFs and dimensionality reduction [64] is used to analyze the data structure in its descriptor space. The analysis reveals that GFN1-xTB predicts band gaps of MOFs with transition metals with low accuracy. Further, Δ-learning [65] using an XGBoost [66] model is utilized to train the difference between band gaps computed using GFN1-xTB and DFT (PBE/HSE06). This model outperforms the XGBoost model trained directly using the DFT data as targets, as well as deep learning models such as MOFTransformer [44] and Crystal Graph Convolutional Neural Networks (CGCNN) [67] that are commonly used to predict band gaps in complex systems. This shows that using classical machine learning models, while exploiting low and high fidelity data, is an effective approach, while also being computationally less intensive and more interpretable, leading to a better understanding of structure-property relationships. A schematic of the band gap prediction framework developed in this work is shown in Figure 1.

2 Computational Methods

GFN1-xTB: The semi-empirical GFN1-xTB [52] approach employs minimal basis sets of atom-centered orbitals and considers only valence electrons in a linear combination of atomic orbitals (LCAO), including terms up to the third order in energy. Contrary to previous semi-empirical methods [51], the repulsive term is not designed to reproduce the original term relative to the DFT contribution, and to compensate for the approximations of other terms in DFT. This eliminates the need to redefine the potential for each element pair, avoiding pair-wise parameters. The D3 method [68] is used to compute the dispersion energy term. Slater-type orbitals are used to describe AOs, approximated through contractions of standard primitive Gaussian functions. Hydrogen bonding is described by a second s-function for hydrogen and d-polarization functions for higher row elements are also included. Using these parameterizations, GFN1-xTB comprises of 16 global and about 1000 element-specific parameters, making it magnitudes faster than DFT, and a computationally inexpensive way to obtain

large labeled datasets. Note that GFN2-xTB [53] was tested preliminarily in this study, but was not pursued further due to convergence failures, consistent with findings in previous works [69, 70, 62].

Δ-Learning: This approach [65] is used in ML to evaluate a correction to the true values of a target, where the correction is calculated with respect to low fidelity approximations of the true target values. In this work, band gaps computed using GFN1-xTB, $E_{g,xTB}$, are used as the low fidelity data, and DFT band gaps, $E_{g,DFT}$, are considered as true target values. An XGBoost model [66] is used to train Δ using each MOF, i, in the training set, and Δ_i is given by:

$$\Delta_i = E_{g,DFT_i}(x_i) - E_{g,xTB_i}(x_i). \tag{1}$$

Here, x_i denotes the set of descriptors used to encode the structure. Learning this correction can be highly valuable in situations where access to high-quality data is limited. In such cases, Δ-learning from cheaper approximations of the targets can be leveraged to train ML models with higher accuracy, as was previously shown by Zhang et al. [71]. XGBoost models are employed in this work due to their strong predictive performance and interpretability. It also enables analysis of structure-property relationships and provides insights into feature importance, helping to identify the key factors influencing the model's predictions. A concatenation of three descriptors is considered: Stoichiometric-135 (ST-135) [72], Revised Auto-Correlations (RACs) [37, 73] and Atomic Property Weighted Radial Distribution Functions (APRDFs) [74]. These are commonly used to encode MOFs and ML models trained using these predict DFT band gaps accurately [39]. ST-135 is composed of elemental fractions and statistical attributes of elemental properties such as atomic number, radii etc. RACs are products and differences of heuristic atomic properties on graphs. Metal-centered, linker and functional-group descriptors are generated, weighted by atomic properties, and averaging over all atoms for each MOF produces 384 features. APRDFs use the weighted probability distribution of finding an atom pair in a given spherical volume inside the unit cell to encode a structure. Atomic properties are also used to weigh the RDFs, leading to a feature length of 648.

3 Results

A subset of 4941 MOFs from the QMOF dataset [27] curated in Ref. [75] was used in this work. DFT band gaps at two levels of theory were obtained, PBE values from the QMOF dataset, and high-fidelity HSE06 values from Ref. [75]. To compute band gaps using GFN1-xTB, PBE optimized structures from the QMOF dataset were used. xTB calculations were performed using DFTB+ [76, 77] (automated using a custom Python code available publicly at https://github.com/AshnaJose/MOF-xTB, details in Appendix A.1). These calculations were successful for 4922 MOFs.

Figure 2 (a) shows the parity plot using hex-bins comparing reference PBE band gaps to computed GFN1-xTB band gaps. The semi-empirical method largely succeeds to capture the electronic behavior of this dataset, with an MAE of 0.37 eV, along with the added advantage of being computationally inexpensive compared to DFT. Figure 2 (b) shows the comparison between the reference HSE06 band gaps and computed GFN1-xTB values. GFN1-xTB, in general, underestimates the HSE06 band gaps, with an MAE of 1.26 eV. However, a clear correlation is observed between the two, resembling the trend seen between PBE and HSE06 band gaps (MAE = 1.16 eV, see Appendix Figure 3). These results demonstrate that GFN1-xTB serves as an effective method for approximating electronic band gaps in MOFs. The error distributions of the GFN1-xTB band gaps relative to PBE and HSE06 are shown in Appendix Figure 4.

It is interesting to note that in both Figures 2 (a) and (b), a subset of MOFs exhibits high DFT band gap values while showing comparatively low band gaps when calculated using GFN1-xTB. Dimensionality reduction using t-SNE [64] was used to visualize the dataset in the ST-135 stoichiometric feature space. Figure 2 (c) shows that t-SNE separates the dataset into distinct clusters. This clustering is primarily based on one of the features from ST-135, the range of atomic number in each MOF (see Appendix Figure 5). This feature, obtained from Matminer [78], is defined as the difference between the highest and lowest atomic numbers of the elements present in a given structure. It effectively reflects the highest atomic number element within the material, which is typically the metal in the case of MOFs. The t-SNE visualization also reveals that the 5% MOFs with the highest $\Delta_{HSE\text{-}xTB}$ values (shown as $\Delta_{HSE\text{-}xTB}$ > 2.6 eV), which are primarily the MOFs incorrectly predicted as highly conductive are clustered in a small region of this descriptor space. Element specific features from

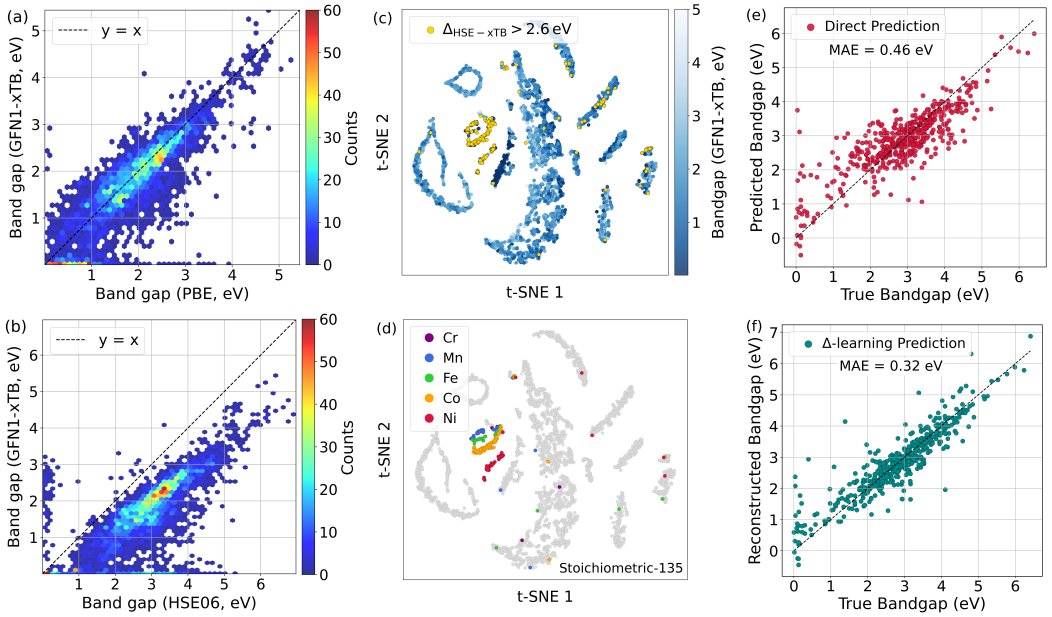

Figure 2: (a) Parity plots using hex-bins comparing computed GFN1-xTB band gaps for the 4922-dataset with reference DFT band gaps at the PBE and (b) HSE06 level. Colors indicate point density; $y = x$ line is shown for reference. (c) t-SNE of the dataset in the ST-135 feature space. Blue shades represent GFN1-xTB band gap values for each MOF. MOFs with high band gap error ($\Delta_{\text{HSE-xTB}}$ above the threshold) are shown in yellow. (d) t-SNE highlighting the clusters that correspond to MOFs with Cr, Mn, Fe, Co and Ni. (e) Parity plot comparing true HSE06 values of the test set with XGBoost-predicted band gaps trained on HSE06 data; (f) and with reconstructed band gap values obtained by training a Δ-learning XGBoost model using the GFN1-xTB and HSE06 band gaps.

ST-135 reveal that these clusters correspond to MOFs that contain the transition metals Cr, Mn, Fe, Co and Ni (Figure 2 (d)). This implies that GFN1-xTB was unable to correctly approximate the electronic behavior of MOFs with transition metals. This is likely because of the complex d-electron behavior of transition metals, which is difficult to capture using semi-empirical methods like GFN1-xTB due to the simplification of the self-consistency in DFT using parameterization. Furthermore, GFN1-xTB is inadequate in describing strong correlation effects, which are prominent in materials with transition metals, and are captured by hybrid functionals and DFT+U.

This data was then used to train ML models (XGBoost, see Appendix A.5 for details on hyper-parameters) that predict PBE and HSE06 band gaps. ST-135 (computed using matminer [78]), APRDFs and RACs features (obtained using `mofdscribe` [63]) were used to encode the structures. 4714 MOFs were featurizable using RACs, thus this set of MOFs is used hereafter. The training, validation and test sets were constructed in the ratio 8:1:1 using random splits (5-fold cross validation; same split was used to compare various methods). Two different approaches were used: first, an XGBoost model was trained using PBE band gap values as the target property (namely, direct model). The average MAE on the test set using this method was found to be 0.382 eV (obtained after 5-fold cross validation). Secondly, a Δ-learning model was trained i.e. a model that uses Δ as the target quantity, where Δ is the difference between the band gaps given by xTB and DFT in this work. After training, it is used to predict the Δ values for the test set. The PBE band gaps for the test set were then reconstructed using this correction to the GFN1-xTB band gaps. This process led to an average MAE of 0.237 eV, i.e. a 38% decrease in comparison with the direct model. Similar models were trained using HSE06 band gaps as the target property, and the same trend was observed: the average MAE using the direct model was found to be 0.486 eV, and that with the Δ-learning model was found to be 0.304 eV, about 37% lower. This significant improvement obtained using Δ-learning shows how low-fidelity low-cost data can be exploited to train ML models with high accuracy for porous materials. Our models were also compared to two methods employing deep learning architectures, MOFTransformer [44] and CGCNN [67], fine-tuned on our custom dataset. The results, presented in Table 1, show that our descriptor-based XGBoost model performs comparably to these methods,

Table 1: Comparison of different methods for predicting band gaps using PBE and HSE06 functionals. Metrics shown are Mean Absolute Error (MAE) in eV and coefficient of determination (R^2), evaluated on the test set and averaged over five different cross-validation runs. The best performer has been highlighted in bold, * denotes models developed in this work.

Method	$E_{g,PBE}$		$E_{g,HSE06}$	
	MAE (eV)	R^2	MAE (eV)	R^2
XGBoost*	0.382	0.651	0.486	0.598
xTB-$\Delta_{XGBoost}$*	**0.237**	**0.823**	**0.304**	**0.829**
MOFTransformer [44]	0.406	0.591	0.531	0.501
CGCNN [67]	0.356	0.540	0.458	0.442

while the Δ-learning approach using xTB data outperforms both CGCNN and MOFTransformer, reinforcing the impact of our contribution.

4 Discussion and Conclusions

We have demonstrated that Δ-learning using low-cost GFN1-xTB band gaps enables accurate prediction of high-fidelity (HSE06) band gaps for MOFs. GFN1-xTB approximates PBE band gaps well (MAE = 0.37 eV), and the trend relative to HSE06 is comparable to that of PBE itself. However, its performance degrades for MOFs with 3d transition metals, due to limitations in treating open-shell electronic states. By training XGBoost models on the difference between GFN1-xTB and reference DFT values, Δ-learning outperforms direct models, as well as deep-learning benchmarks. This approach allows efficient use of large, inexpensive datasets for training predictive models and is useful in low-data regimes. For MOFs without strongly correlated metals, GFN1-xTB provides a reliable baseline, enabling scalable screening and potential integration with generative design workflows.

Acknowledgments and Disclosure of Funding

We thank Tianshu Li for the HSE06 calculations for the MOF dataset and Hyunsoo Park for useful discussions. This work was supported by EPSRC project EP/X037754/1. Via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/X035859/1), this work used the ARCHER2 UK National Supercomputing Service (www.archer2.ac.uk). We acknowledge computational resources and support provided by the Imperial College Research Computing Service (doi.org/10.14469/hpc/2232). The project also benefited from membership of the AI for Chemistry: AIchemy hub (EPSRC grant EP/Y028775/1 and EP/Y028759/1).

References

[1] Hong-Cai Zhou, Jeffrey R. Long, and Omar M. Yaghi. Introduction to metal–organic frameworks. *Chemical Reviews*, 112(2):673–674, January 2012.

[2] Hao Lyu, Zhe Ji, Stefan Wuttke, and Omar M. Yaghi. Digital reticular chemistry. *Chem*, 6(9):2219–2241, 2020.

[3] Ashley M. Wright, Matthew T. Kapelewski, Stefan Marx, Omar K. Farha, and William Morris. Transitioning metal–organic frameworks from the laboratory to market through applied research. *Nature Materials*, 24(2):178–187, August 2024.

[4] Hailian Li, Mohamed Eddaoudi, M O'Keeffe, and O M Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402(6759):276–279, November 1999.

[5] Meili Ding, Robinson W. Flaig, Hai-Long Jiang, and Omar M. Yaghi. Carbon capture and conversion using metal–organic frameworks and mof-based materials. *Chem. Soc. Rev.*, 48:2783–2828, 2019.

[6] Yang Zhao, Zhongxin Song, Xia Li, Qian Sun, Niancai Cheng, Stephen Lawes, and Xueliang Sun. Metal organic frameworks for energy storage and conversion. *Energy Storage Mater.*, 2:35–62, January 2016.

[7] Avery E Baumann, David A Burns, Bingqian Liu, and V Sara Thoi. Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices. *Commun. Chem.*, 2(1), July 2019.

[8] JeongYong Lee, Omar K. Farha, John Roberts, Karl A. Scheidt, SonBinh T. Nguyen, and Joseph T. Hupp. Metal–organic framework materials as catalysts. *Chem. Soc. Rev.*, 38:1450–1459, 2009.

[9] Yuan-Biao Huang, Jun Liang, Xu-Sheng Wang, and Rong Cao. Multifunctional metal–organic framework catalysts: synergistic catalysis and tandem reactions. *Chem. Soc. Rev.*, 46:126–157, 2017.

[10] Arturo Gamonal, Chen Sun, A Lorenzo Mariano, Estefania Fernandez-Bartolome, Elena Guerrero-SanVicente, Bess Vlaisavljevich, Javier Castells-Gil, Carlos Marti-Gastaldo, Roberta Poloni, Reinhold Wannemacher, Juan Cabanillas-Gonzalez, and Jose Sanchez Costa. Divergent adsorption-dependent luminescence of amino-functionalized lanthanide metal-organic frameworks for highly sensitive NO2 sensors. *J. Phys. Chem. Lett.*, 11(9):3362–3368, May 2020.

[11] Mark D. Allendorf, Renhao Dong, Xinliang Feng, Stefan Kaskel, Dariusz Matoga, and Vitalie Stavila. Electronic devices using open framework materials. *Chemical Reviews*, 120(16):8581–8640, 2020. PMID: 32692163.

[12] Sarah Foley, Hugh Geaney, Gerard Bree, Killian Stokes, Sinead Connolly, Michael J. Zaworotko, and Kevin M. Ryan. Copper sulfide (cuxs) nanowire-in-carbon composites formed from direct sulfurization of the metal-organic framework hkust-1 and their use as li-ion battery cathodes. *Advanced Functional Materials*, 28(19):1800587, 2018.

[13] Guiyin Xu, Ping Nie, Hui Dou, Bing Ding, Laiyang Li, and Xiaogang Zhang. Exploring metal organic frameworks for energy storage in batteries and supercapacitors. *Materials Today*, 20(4):191–209, 2017.

[14] Yuqian Ren, Guo Hui Chia, and Zhiqiang Gao. Metal–organic frameworks in fuel cell technologies. *Nano Today*, 8(6):577–597, 2013.

[15] Lilia S. Xie, Grigorii Skorupskii, and Mircea Dincă. Electrically conductive metal–organic frameworks. *Chemical Reviews*, 120(16):8536–8580, 2020. PMID: 32275412.

[16] Cheng Wang, Demin Liu, and Wenbin Lin. Metal–organic frameworks as a tunable platform for designing functional molecular materials. *J. Am. Chem. Soc.*, 135(36):13222–13234, 2013.

[17] Nivedita Sikdar, Kolleboyina Jayaramulu, Venkayala Kiran, K. Venkata Rao, Srinivasan Sampath, Subi J. George, and Tapas Kumar Maji. Redox-active metal–organic frameworks: Highly stable charge-separated states through strut/guest-to-strut electron transfer. *Chemistry – A European Journal*, 21(33):11701–11706, 2015.

[18] Xiao-Ting Liu, Bin-Bin Qian, Da-Shuai Zhang, Mei-Hui Yu, Ze Chang, and Xian-He Bu. Recent progress in host–guest metal–organic frameworks: Construction and emergent properties. *Coordination Chemistry Reviews*, 476:214921, 2023.

[19] Huabin Zhang, Jianwei Nai, Le Yu, and Xiong Wen (David) Lou. Metal-organic-framework-based materials as platforms for renewable energy and environmental applications. *Joule*, 1(1):77–107, 2017.

[20] Eric M Johnson, Stefan Ilic, and Amanda J Morris. Design strategies for enhanced conductivity in metal-organic frameworks. *ACS Cent. Sci.*, 7(3):445–453, March 2021.

[21] Seung-Jae Shin, Jamie W. Gittins, Matthias J. Golomb, Alexander C. Forse, and Aron Walsh. Microscopic origin of electrochemical capacitance in metal–organic frameworks. *Journal of the American Chemical Society*, 145(26):14529–14538, 2023. PMID: 37341453.

[22] A. L. Mariano, A. Fernández-Blanco, and R. Poloni. Perspective from a Hubbard U-density corrected scheme towards a spin crossover-mediated change in gas affinity. *J. Chem. Phys.*, 159(15):154108, 10 2023.

[23] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a cambridge structural database subset: A collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials*, 29(7):2618–2625, 2017.

[24] Peyman Z. Moghadam, Aurelia Li, Xiao-Wei Liu, Rocio Bueno-Perez, Shu-Dong Wang, Seth B. Wiggin, Peter A. Wood, and David Fairen-Jimenez. Targeted classification of metal–organic frameworks in the cambridge structural database (csd). *Chem. Sci.*, 11:8373–8387, 2020.

[25] Guobin Zhao, Logan M. Brabson, Saumil Chheda, Ju Huang, Haewon Kim, Kunhuan Liu, Kenji Mochida, Thang D. Pham, Prerna, Gianmarco G. Terrones, Sunghyun Yoon, Lionel Zoubritzky, François-Xavier Coudert, Maciej Haranczyk, Heather J. Kulik, Seyed Mohamad Moosavi, David S. Sholl, J. Ilja Siepmann, Randall.Q. Snurr, and Yongchul G. Chung. Core mof db: A curated experimental metal-organic framework database with machine-learned properties for integrated material-process screening. *Matter*, 8(6):102140, 2025.

[26] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.*, 4(2):83–89, November 2011.

[27] Andrew S Rosen, Victor Fung, Patrick Huck, Cody T O'Donnell, Matthew K Horton, Donald G Truhlar, Kristin A Persson, Justin M Notestein, and Randall Q Snurr. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *Npj Comput. Mater.*, 8(1), May 2022.

[28] Jake Burner, Jun Luo, Andrew White, Adam Mirmiran, Ohmin Kwon, Peter G. Boyd, Stephen Maley, Marco Gibaldi, Scott Simrod, Victoria Ogden, and Tom K. Woo. Arc–mof: A diverse database of metal-organic frameworks with dft-derived partial atomic charges and descriptors for machine learning. *Chem. Mater.*, 35(3):900–916, 2023.

[29] Dalar Nazarian, P Ganesh, and David S Sholl. Benchmarking density functional theory predictions of framework structures and properties in a chemically diverse test set of metal–organic frameworks. *J. Mater. Chem. A Mater. Energy Sustain.*, 3(44):22432–22440, 2015.

[30] Andrew S Rosen, Justin M Notestein, and Randall Q Snurr. Identifying promising metal-organic frameworks for heterogeneous catalysis via high-throughput periodic density functional theory. *J. Comput. Chem.*, 40(12):1305–1318, May 2019.

[31] Emmanuel Ren, Philippe Guilbaud, and François-Xavier Coudert. High-throughput computational screening of nanoporous materials in targeted applications. *Digital Discovery*, 1:355–374, 2022.

[32] Chenru Duan, Aditya Nandy, Ralf Meyer, Naveen Arunachalam, and Heather J Kulik. A transferable recommender approach for selecting the best density functional approximations in chemical discovery. *Nat. Comput. Sci.*, 3(1):38–47, December 2022.

[33] Cigdem Altintas, Omer Faruk Altundal, Seda Keskin, and Ramazan Yildirim. Machine learning meets with metal organic frameworks for gas storage and separation. *J. Chem. Inf. Model.*, 61(5):2131–2146, 2021.

[34] Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, and Jihan Kim. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Appl. Mater. Interfaces*, 13(20):23647–23654, 2021.

[35] Aditya Nandy, Chenru Duan, and Heather J. Kulik. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal–organic frameworks. *J. Am. Chem. Soc.*, 143(42):17535–17547, 2021.

[36] Hakan Demir, Hilal Daglar, Hasan Can Gulbalkan, Gokhan Onder Aksu, and Seda Keskin. Recent advances in computational modeling of mofs: From molecular simulations to machine learning. *Coordin. Chem. Rev.*, 484:215112, 2023.

[37] Jon Paul Janet and Heather J. Kulik. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

[38] Emmanuel Ren and François-Xavier Coudert. Enhancing gas separation selectivity prediction through geometrical and chemical descriptors. *Chem. Mater.*, 35(17):6771–6781, 2023.

[39] Ashna Jose, Emilie Devijver, Noel Jakse, and Roberta Poloni. Informative training data for efficient property prediction in metal–organic frameworks by active learning. *Journal of the American Chemical Society*, 146(9):6134–6144, 2024. PMID: 38404041.

[40] Evan Xie, Xijun Wang, J. Ilja Siepmann, Haoyuan Chen, and Randall Q. Snurr. Generative ai for design of nanoporous materials: review and future prospects. *Digital Discovery*, pages –, 2025.

[41] Junkil Park, Honghui Kim, Yeonghun Kang, Yunsung Lim, and Jihan Kim. From data to discovery: Recent trends of machine learning in metal–organic frameworks. *JACS Au*, 4(10):3727–3743, 2024.

[42] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[43] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: Self-supervised transformer model for metal–organic framework property prediction. *J. Am. Chem. Soc.*, 145(5):2958–2967, 2023.

[44] Yeonghun Kang, Hyunsoo Park, Berend Smit, and Jihan Kim. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nat. Mach. Intell.*, 5(3):309–318, March 2023.

[45] Hyunsoo Park, Yeonghun Kang, and Jihan Kim. Enhancing structure–property relationships in porous materials through transfer learning and cross-material few-shot learning. *ACS Appl. Mater. & Inter.*, 15(48):56375–56385, 2023. PMID: 37983088.

[46] Ashna Jose, Emilie Devijver, Roberta Poloni, Valérie Monbet, and Noël Jakse. Tree-based quantile active learning for automated discovery of MOFs. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023.

[47] Hyunsoo Park, Sauradeep Majumdar, Xiaoqi Zhang, Jihan Kim, and Berend Smit. Inverse design of metal–organic frameworks for direct air capture of co2via deep reinforcement learning. *Digital Discovery*, 3:728–741, 2024.

[48] Mahyar Rajabi-Kochi, Negareh Mahboubi, Aseem Partap Singh Gill, and Seyed Mohamad Moosavi. Adaptive representation of molecules and materials in bayesian optimization. *Chem. Sci.*, 16:5464–5474, 2025.

[49] Aditya Nandy. From pages to patterns: Towards extracting catalytic knowledge from structure and text for transition-metal complexes and metal-organic frameworks. *Journal of Catalysis*, 448:116174, 2025.

[50] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, Th. Frauenheim, S. Suhai, and G. Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B*, 58:7260–7268, Sep 1998.

[51] Marcus Elstner and Gotthard Seifert. Density functional tight binding. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2011):20120483, March 2014.

[52] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, 2017. PMID: 28418654.

[53] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019. PMID: 30741547.

[54] Sebastian Spicher, Markus Bursch, and Stefan Grimme. Efficient calculation of small molecule binding in metal–organic frameworks and porous organic cages. *The Journal of Physical Chemistry C*, 124(50):27529–27541, 2020.

[55] José Manuel Vicent-Luna, Sofia Apergi, and Shuxia Tao. Efficient computation of structural and electronic properties of halide perovskites using density functional tight binding: Gfn1-xtb method. *Journal of Chemical Information and Modeling*, 61(9):4415–4424, 2021. PMID: 34414764.

[56] Felix R. S. Purtscher, Leo Christanell, Moritz Schulte, Stefan Seiwald, Markus Rödl, Isabell Ober, Leah K. Maruschka, Hassan Khoder, Heidi A. Schwartz, El-Eulmi Bendeif, and Thomas S. Hofer. Structural properties of metal–organic frameworks at elevated thermal conditions via a combined density functional tight binding molecular dynamics (dftb md) approach. *The Journal of Physical Chemistry C*, 127(3):1560–1575, 2023.

[57] Thanh-Hiep Thi Le, Pablo Gómez-Orellana, and Manuel Angel Ortuño. Evaluation of semiempirical quantum mechanical methods for zr-based metal–organic framework catalysts. *ChemPhysChem*, 26(8), March 2025.

[58] Karoline Schjelde, Oscar B. Obel, Andreas Erbs Hillers-Bendtsen, and Kurt V. Mikkelsen. Semi-automated screening of azobezenes for solar energy storage using extended tight binding methods. *Scientific Reports*, 15(1), July 2025.

[59] Alin Marin Elena, Prathami Divakar Kamath, Théo Jaffrelot Inizan, Andrew S. Rosen, Federica Zanca, and Kristin A. Persson. Machine learned potential for high-throughput phonon calculations of metal—organic frameworks. *npj Computational Materials*, 11(1), May 2025.

[60] Maryam Nurhuda, Carole C. Perry, and Matthew A. Addicoat. Performance of gfn1-xtb for periodic optimization of metal organic frameworks. *Phys. Chem. Chem. Phys.*, 24:10906–10914, 2022.

[61] Masoumeh Mahmoudi Gahrouei, Nikiphoros Vlastos, Ransell D'Souza, Emmanuel C. Odogwu, and Laura de Sousa Oliveira. Benchmark investigation of scc-dftb against standard and hybrid dft to model electronic properties in two-dimensional mofs for thermoelectric applications. *Journal of Chemical Theory and Computation*, 20(9):3976–3992, 2024. PMID: 38708963.

[62] Mateusz Pokora, Jakub Goclon, Johannes Margraf, Chiara Panosetti, Artem Samtsevych, and Piotr Paneth. Low-cost periodic calculations of metal-organic frameworks: A gfn1-xtb perspective. *ChemPhysChem*, 26(14), June 2025.

[63] Kevin Maik Jablonka, Andrew S. Rosen, Aditi S. Krishnapriyan, and Berend Smit. An ecosystem for digital reticular chemistry. *ACS Central Science*, 9(4):563–581, 2023.

[64] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[65] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. PMID: 26574412.

[66] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[67] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, 04 2018.

[68] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurateab initioparametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of Chemical Physics*, 132(15), April 2010.

[69] Vivek Sinha, Jochem J. Laan, and Evgeny A. Pidko. Accurate and rapid prediction of pka of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach. *Physical Chemistry Chemical Physics*, 23(4):2557–2567, 2021.

[70] Jarosław J. Panek. The structural stability of enzymatic proteins in the gas phase: A comparison of semiempirical hamiltonians and the gfn-ff. *Molecules*, 30(10):2131, May 2025.

[71] Lingyao Zhang, Tianhao Su, Musen Li, Fanhao Jia, Shuobo Hu, Peihong Zhang, and Wei Ren. Accurate band gap prediction based on an interpretable δ-machine learning. *Materials Today Communications*, 33:104630, 2022.

[72] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B*, 89:094104, 03 2014.

[73] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather J Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.*, 11(1):4068, August 2020.

[74] Michael Fernandez, Nicholas R. Trefiak, and Tom K. Woo. Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity. *J. Phys. Chem. C*, 117(27):14095–14105, 2013.

[75] Hyunsoo Park, Tianshu Li, Ashna Jose, Seung-Jae Shin, and Aron Walsh. Data-driven design of metal-organic frameworks for photoelectrochemical reactions. In preparation (2025).

[76] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim. Dftb+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of Chemical Physics*, 152(12):124101, 03 2020.

[77] B. Hourahine, M. Berdakin, J. A. Bich, F. P. Bonafé, C. Camacho, Q. Cui, M. Y. Deshaye, G. Díaz Mirón, S. Ehlert, M. Elstner, T. Frauenheim, N. Goldman, R. A. González León, T. van der Heide, S. Irle, T. Kowalczyk, T. Kubař, I. S. Lee, C. R. Lien-Medrano, A. Maryewski, T. Melson, S. K. Min, T. Niehaus, A. M. N. Niklasson, A. Pecchia, K. Reuter, C. G. Sánchez, C. Scheurer, M. A. Sentef, P. V. Stishenko, V. Q. Vuong, and B. Aradi. Recent developments in dftb+, a software package for efficient atomistic quantum mechanical simulations. *The Journal of Physical Chemistry A*, 129(24):5373–5390, 2025. PMID: 40479742.

[78] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E.R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.*, 152:60–69, September 2018.

A Appendix

A.1 Technical details of GFN1-xTB calculations

DFTB+ was used to compute band gaps using GFN1-xTB method, starting from the optimized geometries obtained from the QMOF dataset. The structures were used to obtain the inputs of the DFTB+ code: `cif` files to extract the atomic positions and lattice parameters. The maximum angular momentum was set based on the elements in the structure, using `MaxAngularMomentum`. The K-point mesh was generated automatically using the `automatic_density` function from the `pymatgen`, where the target density was set to 2000. The calculations were run, and the band gaps were extracted from the outputs obtained from DFTB+. This set of calculations and analysis was automated using a custom Python code, available at github.com/AshnaJose/MOF-xTB.

A.2 Reference data

Figure 3 compares the reference band gaps at the HSE06 and PBE levels for the 4941-dataset. The trend shown here is similar to that between HSE06 and GFN1-xTB band gaps, showing that GFN1-xTB predicts electronic properties of MOFs reliably. Here, the MAE is 1.16 eV.

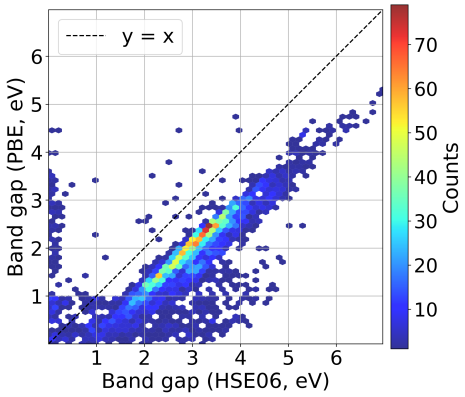

Figure 3: Parity plot using hex-bins comparing the reference PBE and HSE06 band gaps for the 4941-dataset. The colors represent the density of points in each bin. $y = x$ line is shown for reference.

A.3 Error estimation for GFN1-xTB computed band gaps

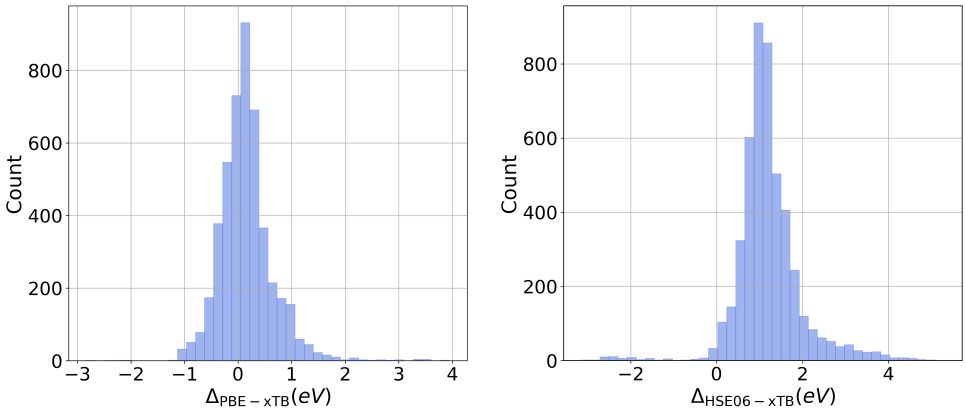

Figure 4: Histograms showing the distribution of $\Delta_{PBE-xTB}$ (left) and $\Delta_{HSE06-xTB}$ (right).

Figure 4 shows the distribution of the errors between the computed GFN1-xTB band gaps and the reference DFT band gaps (PBE and HSE06). It can be seen that GFN1-xTB approximates PBE values well, with an MAE of 0.37 eV. It underestimates HSE06 in general, with an MAE of 1.26 eV.

A.4 Dimensionality reduction analysis

Figure 5 shows a representation of the 4941-dataset using the dimensionality reduction method, t-distributed Stochastic Neighbor Embedding (t-SNE), in the ST-135 feature space. The ST-135 feature set separates the dataset into distinct clusters, which are based on the range of atomic number feature, which broadly indicates the metal center type in each MOF. Thus, this clustering is based on the local metal-center composition of the MOF.

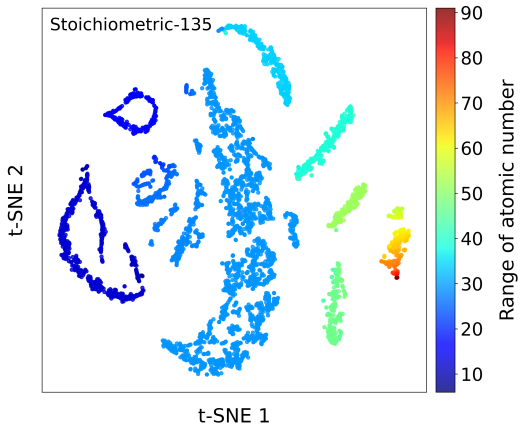

Figure 5: t-SNE of the 4941-dataset in the ST-135 feature space. The colors represent the value of the range of atomic number in each structure, which is one of the features of ST-135.

A.5 Feature selection and model training

XGBoost models with 200 estimators were used to train on the training data. The learning rate was set to 0.1, and the maximum depth of the trees was set as 6. As the dimension of the feature space is large, feature selection was used to select the important features, using the XGBoost feature selection function. This selection was performed using the training set. 3 sets of feature selection were performed, using different target properties, i.e., band gaps obtained using GFN1-xTB, PBE and HSE06. The top (important) features primarily comprise of features from the RACs descriptor, along with a few from the ST-135 and the APRDF descriptor sets. This implies that RACs describe electronic properties, such as band gap in this work, accurately.

To obtain the optimal number of features to be used, different values of the number of top features were chosen and models were trained based on each. The results are shown in the left panel of Figure 6, for GFN1-xTB, PBE and HSE06 band gap values as the target property. The MAE vs number of top features curves plateau after approximately 400 features in each case. This was thus chosen as the optimal number of features for model training.

These features were then used to train XGBoost models using the band gap values as the target property (direct model). Five fold cross-validation was performed to compute statistics (using seeds 42, 78, 14, 115 and 173). Parity plots comparing the true and predicted values of GFN1-xTB, PBE and HSE06 band gaps (for seed = 42) are shown in the right panel of Figure 6. The MAE for these individual models, trained directly on the target property, on the test set are 0.384 eV, 0.388 eV and 0.456 eV for GFN1-xTB, PBE and HSE06 band gaps, respectively.

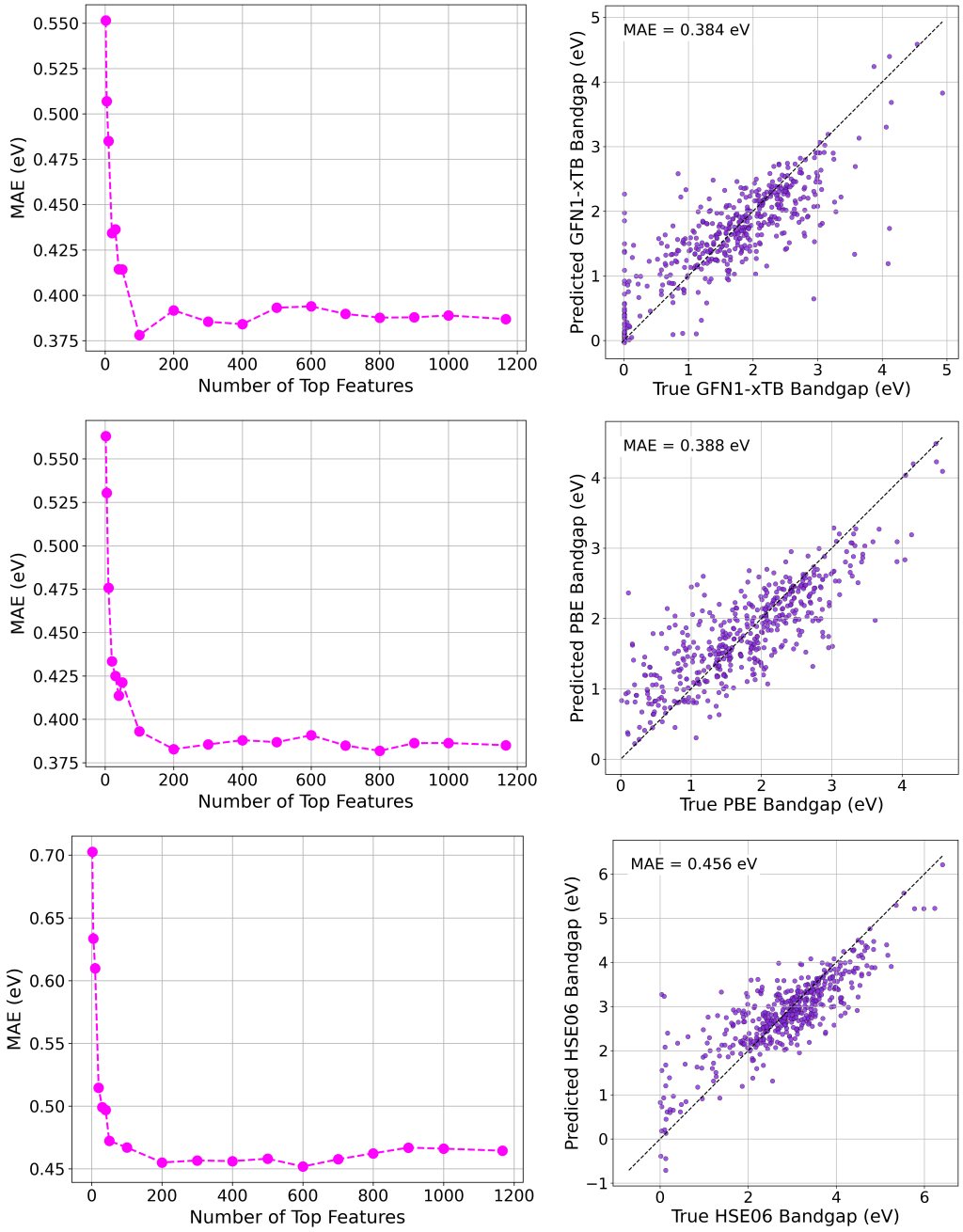

Figure 6: Left panel: Plots showing the variation of MAE with the number of features to predict GFN1-xTB, PBE and HSE06 band gaps as target values. Features are added in their decreasing order of importance. Right panel: Parity plot comparing the true computed GFN1-xTB, PBE and HSE06 band gaps for the test set with predicted band gaps using an XGBoost model trained with the respective band gaps as the targets (direct prediction). The model was trained using the top 400 features. $y = x$ line is shown for reference. All the plots in this figure correspond to one of the five train-test splits (seed = 42) used in this work.

A.6 Δ-Learning

Parity plots comparing true PBE values of the test set with predicted band gaps obtained using the direct and Δ-Learning (using GFN1-xTB band gaps) approaches are shown in Figure 7 for the train-test split using seed = 42. The MAE on the test for the direct prediction was found to be 0.39 eV, while using the Δ-Learning model reduced the MAE to 0.25 eV.

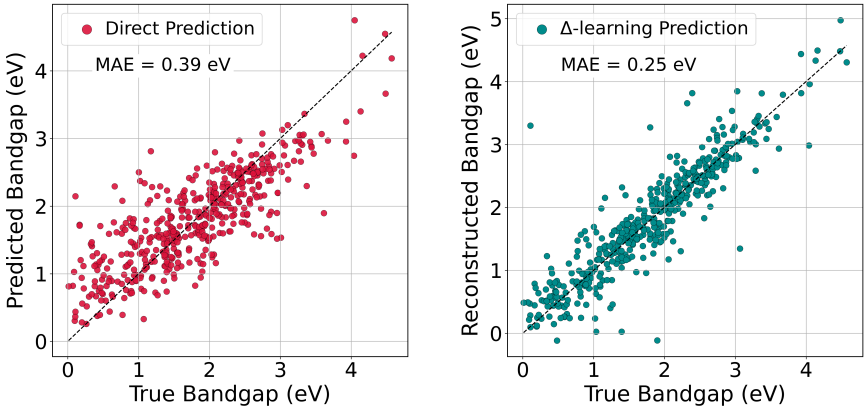

Figure 7: Parity plots comparing true PBE values of the test set (for seed = 42) with predicted band gaps obtained by training an XGBoost model using the PBE values (left) and with reconstructed band gap values obtained by training a Δ-learning XGBoost model using the GFN1-xTB and PBE band gaps (right).