Large Language Models as Materials Science Adapted Learners

Tong Xie[a], **Yuwei Wan**[b], **Yuchen Zeng**[c], **Clara Grazian**[d], **Wenjie Zhang**[e], **Dongzhan Zhou**[f], **Yixuan Liu**[a], **Shaozhou Wang**[a], **Chunyu Kit**[b], **Ouyang Wanli**[f], **Bram Hoex**[e]

[a] *GreenDynamics, Sydney, NSW, Australia* tong@greendynamics.com.au,yuwei@greendynamics.com.au

[b] *City University of Hong Kong, Hong Kong, China*

[c] *University of Wisconsin-Madison, Madison, Wisconsin, United States*

[d] *University of Sydney, Camperdown, NSW, Australia*

[e] *University of New South Wales, Kensington, NSW, Australia* b.hoex@unsw.edu.au

[f] *Shanghai Artificial Intelligence Laboratory, Shanghai, China* zhoudongzhan@pjlab.org.cn

1. Introduction

Novel materials can revolutionize technologies, addressing global challenges like climate change [1, 2], sustainable development [3, 4, 5], and public health [6, 7]. To navigate vast chemical spaces and complex structure-property relationships in material discovery, researchers employ principle calculation through high-throughput simulations [8, 9, 10] and machine learning (ML) [11, 12, 13] to expedite discovery workflows. However, experimental materials' heterogeneity from defects and impurities presents obstacles not captured by basic structural representations, causing ML models to gauge the experimental performance ineffectively. This necessitates a foundational model that bridge human understanding, computational models, and experimental realities while adapting across diverse tasks and material classes.

Large language models (LLMs) [14, 15] can process human-readable descriptions directly and generalize across tasks. Works like [16] demonstrate GPT models' efficacy for chemical tasks with scarce data. However, many state-of-the-art LLMs remain proprietary, requiring fine-tuning on proprietary servers with high costs, limited customizability, and potential privacy concerns. The main contributions of this study are: 1) We propose DARWIN 1.5, an open-source foundational LLM for materials science. 2) The QA-multi (2-stage) strategy proves superior by integrating scientific literature comprehension with multi-task prediction capabilities, outperforming both GPT-3.5 fine-tuning and GPT-4 prompting across diverse materials prediction tasks. 3) Comprehensive ablation studies reveal that exposure to diverse task formats enhances instruction-following capabilities, while multi-task training enables knowledge transfer between properties.

2. Methods

We constructed 22 tasks, comprising 5 classification problems and 17 regression problems, which characterize fundamental physical, chemical, and electrochemical properties using various material representations. To harness the full potential of

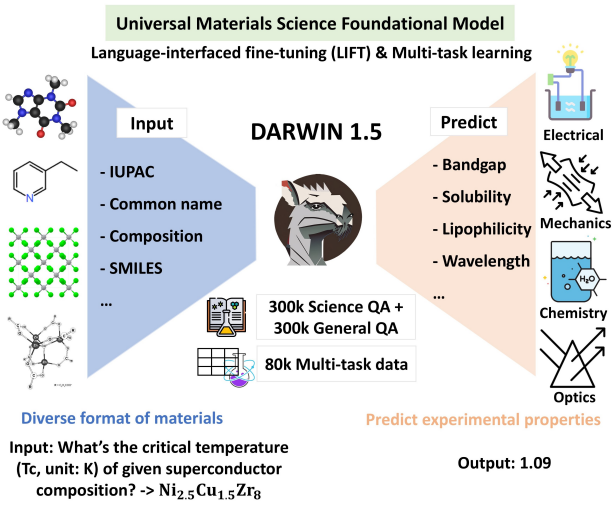

Fig. 1: Overview of DARWIN 1.5, with example input and output.

LLMs for materials science applications, we converted original datasets into language-interfaced format instructions [17] suitable for fine-tuning these models. For example, an instruction may look like: 'What is the band gap of given composition?' with input 'CdCu2SnS4', and our model should give a text output '1.37', which can be converted to a numeric value. Before multi-task fine-tuning, we enhanced LLMs' scientific reasoning by incorporating scientific QA fine-tuning alongside multi-task fine-tuning. We used SciQAG framework [18] to generate 332,997 open-ended scientific QA pairs, preserving essential information from lengthy scientific texts, which are particularly suitable for experimental contexts. To maintain the model's general language capabilities and prevent overfitting to scientific content, we balanced training with general QA pairs from the Tulu dataset [19]. Details of data are available in Appendix B. We explore the impact of QA and multi-task fine-tuning on 22 task performance using the open-source LLaMA-7B [20]. Our experimental setups include 4 specific fine-tuning strategies:

1) Single-task (**Base-ST**): Fine-tuning LLMs on in-

dividual task training sets to establish baselines and create task-specific models (22 separate models).

2) Multi-task (**Base-MT**): Fine-tuning LLMs on a mixture of all 22 task datasets to create a single model capable of handling all tasks.

3) QA-single (**QA-ST**): Two-stage approach where LLMs are first fine-tuned on QA data (creating Base-QA), then further fine-tuned on individual tasks (22 separate models).

4) QA-multi (**QA-MT**): Two-stage approach combining QA and multi-task learning, where the Base-QA model is further fine-tuned on a mixture of all 22 task datasets.

3. Results

3.1 Results of QA and multi-task fine-tuning

As shown in Table 1, we compare different fine-tuning strategies on performance of LLaMA-7B on 22 tasks. Two-stage fine-tuning (QA-MT) provides the best results by combining QA fine-tuning and multi-task learning, showing average improvements of 3.38% for classification tasks and 11.79% for regression tasks compared to the baseline. QA-MT also outperforms GPT-3.5 (fine-tuned), GPT-4 (few-shot), and traditional ML algorithms in most tasks, demonstrating its effectiveness for diverse materials science applications. In a bandgap case study (Appendix, our QA-MT model achieves a performance comparable to the high-quality Heyd–Scuseria–Ernzerhof (HSE) method and significantly outperforms the Perdew–Burke–Ernzerhof (PBE) method.

Table 1: Performance comparison of different fine-tuning strategies (CI and RI refers to performance improvement on classification tasks and regression tasks, respectively)

Compare	Strategy	CI (%)	RI (%)
vs. Base-ST	QA-ST	1.55	2.30
	Base-MT	2.65	10.77
	QA-MT	**3.38**	**11.79**
vs. Random-ST	Random-MT	8.08	24.08
	Base-ST	**11.04**	**33.57**

To evaluate how pre-training impacts fine-tuning performance, we created "Random-ST" models by fine-tuning the untrained model on single tasks, and a "Random-MT" model through multi-task fine-tuning, enabling direct assessment of pre-training's contribution to materials science tasks. Pre-training and multi-task fine-tuning play complementary roles in LLM performance. Pre-training provides significant advantages (11.04% for classification and 33.57% for regression tasks) compared to randomly initialized models, with greater benefits for general material representations (compositions, names) than specialized ones (SMILES, MOFs)(see Appendix). Multi-task fine-tuning can partially compensate for lack of pre-training, suggesting that language-interfaced multi-task learning effectively leverages encoded knowledge to integrate diverse material representations.

3.2 Investigating factors that affect prediction performance

We conducted ablation studies using two recognized regression benchmark datasets: matbench_exp_bandgap and matbench_steel. We designed multiple small-scale multi-task datasets by combining each target dataset with various auxiliary datasets:

$$
\begin{aligned}
D_{\text{matbench}} &= \{(x_i, y_i) \mid x_i \in \mathcal{X}_{\text{comp}}, y_i \in R\} \\
D_{\text{other}} &= \{(x_i, y_i) \mid x_i \in \mathcal{X}_{\text{SMILES}}, y_i \in R\} \\
D_{\text{syn\_1}} &= \{(x_i, \tilde{y}_i) \mid x_i \in \mathcal{X}_{\text{comp}}, \tilde{y}_i \sim U(y_{\min}, y_{\max})\} \\
D_{\text{syn\_2}} &= \{(\tilde{x}_i, \tilde{y}_i) \mid \tilde{x}_i \in \mathcal{X}_{\text{rand}}, \tilde{y}_i \sim U(y_{\min}, y_{\max})\} \\
D_{\text{syn\_3}} &= \{(\tilde{x}_i, y_i) \mid \tilde{x}_i \in \mathcal{X}_{\text{rand}}, y_i \in R\}
\end{aligned}
$$

where $\mathcal{X}_{\text{comp}}$ and $\mathcal{X}_{\text{SMILES}}$ represents composition-based and SMILES material representations, $\tilde{y}_i$ is a random value sampled uniformly within the original dataset's property range, and $\tilde{x}_i \in \mathcal{X}_{\text{rand}}$ represents randomly generated alphabetical codes.

Table 2: Performance comparison using different auxiliary datasets (The metrics here is MAE).

Auxiliary data	Bandgap	Steel
Base-ST	0.386	194.9
+Syn_1	0.659	220.5
+Syn_2	0.371	148.3
+Syn_3	0.371	128.9
+Other	0.368	116.0
+Matbench	0.289	109.9
+QA+Matbench	**0.269**	**93.66**

As shown in Table 2, auxiliary datasets generally improved performance except for +Syn_1 (containing incorrect property data). Interestingly, completely fabricated +Syn_2 data improved model performance by enhancing instruction-following abilities through pattern recognition. +Syn_3, which maintained authentic property distributions with fabricated materials, performed slightly better than +Syn_2.

4. Conclusions

Our study demonstrates that LLMs can effectively internalize materials science knowledge across multiple systems and modalities through multi-stage training. The model develops sophisticated understanding of materials relationships through exposure to scientific literature and multi-task training data. The result is essentially a "materials science GPT" - a foundational model that learns and reasons about materials in ways that parallel human scientific thinking, opening new possibilities for materials discovery and design.

Acknowledgments

This research was supported by generous funding and resources from multiple organizations. We gratefully acknowledge the support of: 1) The National Computational Infrastructure, Paswsey Super Computer Centre and Argonne National Lab, which provided critical computational resources essential to our research infrastructure. 2) Microsoft Research through AFMR, whose technological support and computational resources were instrumental in advancing our project. 3) The Australian Renewable Energy Agency (ARENA), whose financial support and commitment to innovative energy research made this work possible. 4) The Australian Centre for Advanced Photovoltaics (ACAP), whose funding and scientific collaboration significantly contributed to our research outcomes.

We extend our sincere thanks to these organizations for their crucial support in enabling this research endeavor.

References

[1] Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, et al. Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183, 2020.

[2] Xiao-Lei Shi, Jin Zou, and Zhi-Gang Chen. Advanced thermoelectric design: from materials and structures to devices. *Chemical reviews*, 120(15):7399–7515, 2020.

[3] Shasha Zhang, Zonghao Liu, Wenjun Zhang, Zhaoyi Jiang, Weitao Chen, Rui Chen, Yuqian Huang, Zhichun Yang, Yiqiang Zhang, Liyuan Han, et al. Barrier designs in perovskite solar cells for long-term stability. *Advanced Energy Materials*, 10(35):2001610, 2020.

[4] Muge Acik and Seth B Darling. Graphene in perovskite solar cells: device design, characterization and implementation. *Journal of Materials Chemistry A*, 4(17):6185–6235, 2016.

[5] Richard Song, Maxwell Murphy, Chenshuang Li, Kang Ting, Chia Soo, and Zhong Zheng. Current development of biodegradable polymeric materials for biomedical applications. *Drug design, development and therapy*, pages 3117–3145, 2018.

[6] Wenlong Li, Eng San Thian, Miao Wang, Zuyong Wang, and Lei Ren. Surface design for antibacterial materials: from fundamentals to advanced strategies. *Advanced Science*, 8(19):2100368, 2021.

[7] Kara Lavender Law and Ramani Narayan. Reducing environmental plastic pollution by designing polymer materials for managed end-of-life. *Nature Reviews Materials*, 7(2):104–116, 2022.

[8] Lili Xi, Shanshan Pan, Xin Li, Yonglin Xu, Jianyue Ni, Xin Sun, Jiong Yang, Jun Luo, Jinyang Xi, Wenhao Zhu, et al. Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening. *Journal of the American Chemical Society*, 140(34):10785–10793, 2018.

[9] Stefano Curtarolo, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V Chepulskii, Richard H Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

[10] Guillaume Brunin, Francesco Ricci, Viet-Anh Ha, Gian-Marco Rignanese, and Geoffroy Hautier. Transparent conducting materials discovery using high-throughput computing. *npj Computational Materials*, 5(1):63, 2019.

[11] Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

[12] Taoyong Cui, Chenyu Tang, Mao Su, Shufei Zhang, Yuqiang Li, Lei Bai, Yuhan Dong, Xingao Gong, and Wanli Ouyang. Gpip: Geometry-enhanced pre-training on interatomic potentials. *arXiv preprint arXiv:2309.15718*, 2023.

[13] Duo Zhang, Xinzijian Liu, Xiangyu Zhang, Chengqian Zhang, Chun Cai, Hangrui Bi, Yiming Du, Xuejian Qin, Anyang Peng, Jiameng Huang, et al. Dpa-2: a large atomic model as a multi-task learner. *npj Computational Materials*, 10(1):293, 2024.

[14] OpenAI. Gpt-4 technical report, 2023. Preprint at https://arxiv.org/abs/2303.08774.

[15] AI@Meta. Llama 3 model card, 2024. GitHubhttps://github.com/meta-lama/llama3/blob/main/MODEL_CARD.md.

[16] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[17] Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

[18] Yuwei Wan, Aswathy Ajith, Yixuan Liu, Ke Lu, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. Sciqag: A framework for auto-generated scientific question answering dataset with fine-grained evaluation, 2024. Preprint at https://arxiv.org/abs/2405.09939.

[19] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023. Preprint at https://arxiv.org/abs/2311.10702.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[21] Sebastian Ruder. An overview of multi-task learning in deep neural networks, 2017. Preprint at https://arxiv.org/abs/1706.05098.

[22] Soumya Sanyal, Janakiraman Balachandran, Naganand Yadati, Abhishek Kumar, Padmini Rajagopalan, Suchismita Sanyal, and Partha Talukdar. Mt-cgcnn: Integrating crystal graph convolutional neural network with multitask learning for material property prediction, 2018. Preprint at https://arxiv.org/abs/1811.05660.

[23] Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *npj computational materials*, 7(1):83, 2021.

[24] Christopher Kuenneth, Arunkumar Chitteth Rajan, Huan Tran, Lihua Chen, Chiho Kim, and Rampi Ramprasad. Polymer informatics with multi-task learning. *Patterns*, 2(4), 2021.

[25] Ryan Jacobs, Maciej P Polak, Lane E Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. Regression with large language models for materials and molecular property prediction, 2024. Preprint at https://arxiv.org/abs/2409.06080.

[26] Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions, 2023. Preprint at https://arxiv.org/abs/2310.14029.

[27] Tong Xie, Yuwei Wan, Yufei Zhou, Wei Huang, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. Creation of a structured solar cell material dataset and performance prediction using large language models. *Patterns*, 5(5), 2024.

[28] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022. Preprint at https://arxiv.org/abs/2212.10560.

[29] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S Grethe, Jaap Heringa, Peter A C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J Lusher, Maryann E Martone, Albert Mons, Abel L Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.

[30] Matthias Scheffler, Martin Aeschlimann, Martin Albrecht, Tristan Bereau, Hans-Joachim Bungartz, Claudia Felser, Mark Greiner, Axel Groß, Christoph T Koch, Kurt Kremer, et al. Fair data enabling new horizons for materials research. *Nature*, 604(7907):635–642, 2022.

[31] 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. *Nature communications*, 11(1):1–10, 2020.

[32] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

[33] James L McDonagh, Neetika Nath, Luna De Ferrari, Tanja Van Mourik, and John BO Mitchell. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *Journal of chemical information and modeling*, 54(3):844–856, 2014.

[34] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

[35] Yoshiyuki Kawazoe, J-Z Yu, A-P Tsai, and T Masumoto. Nonequilibrium phase diagrams of ternary amorphous alloys, 1997.

[36] Zongrui Pei, Junqi Yin, Jeffrey A Hawk, David E Alman, and Michael C Gao. Machine-learning informed prediction of high-entropy solid solution formation: Beyond the hume-rothery rules. *npj Computational Materials*, 6(1):50, 2020.

[37] Rohit Batra, Carmen Chen, Tania G Evans, Krista S Walton, and Rampi Ramprasad. Prediction of water stability of metal–organic frameworks using machine learning. *Nature Machine Intelligence*, 2(11):704–710, 2020.

[38] Edward J Beard, Ganesh Sivaraman, Álvaro Vázquez-Mayagoitia, Venkatram Vishwanath, and Jacqueline M Cole. Comparative dataset of experimental and computational attributes of uv/vis absorption spectra. *Scientific data*, 6(1):307, 2019.

[39] Shinji Nagasawa, Eman Al-Naamani, and Akinori Saeki. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *The Journal of Physical Chemistry Letters*, 9(10):2639–2646, 2018.

[40] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, and Bissan Al-Lazikani. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[41] YG Chung, E Haldoupis, BJ Bucior, M Haranczyk, S Lee, KD Vogiatzis, S Ling, M Milisavljevic, H Zhang, JS Camp, et al. Computation-ready experimental metal-organic framework (core mof) 2019 dataset, 2019. Zenodo https://doi.org/10.5281/zenodo.

[42] Seyed Mohamad Moosavi, Balázs Álmos Novotny, Daniele Ongari, Elias Moubarak, Mehrdad Asgari, Özge Kadioglu, Charithea Charalambous, Andres Ortega-Guerrero, Amir H Farmahini, Lev Sarkisov, et al. A data-science approach to predict the heat capacity of nanoporous materials. *Nature materials*, 21(12):1419–1425, 2022.

[43] David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720, 2014.

[44] Ryan-Rhys Griffiths, Jake L Greenfield, Aditya R Thawani, Arian R Jamasb, Henry B Moss, Anthony Bourached, Penelope Jones, William McCorkindale, Alexander A Aldrick, Matthew J Fuchter, et al. Data-driven discovery of molecular photoswitches with multioutput gaussian processes. *Chemical Science*, 13(45):13541–13551, 2022.

[45] Valentin Stanev, Corey Oses, A Gilad Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and Ichiro Takeuchi. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4(1):29, 2018.

[46] Materials Information Station National Institute of Materials Science. Supercon, 2011. http://supercon.nims.go.jp/index_en.html.

[47] Michael W Gaultois, Taylor D Sparks, Christopher KH Borg, Ram Seshadri, William D Bonificio, and David R Clarke. Data-driven review of thermoelectric materials: performance and resource considerations. *Chemistry of Materials*, 25(15):2911–2920, 2013.

[48] Gyoung S Na and Hyunju Chang. A public database of thermoelectric materials and system-identified material representation for data-driven discovery. *npj Computational Materials*, 8(1):214, 2022.

[49] Dingyun Huang and Jacqueline M Cole. A database of thermally activated delayed fluorescent molecules auto-generated from scientific literature with chemdataextractor. *Scientific Data*, 11(1):80, 2024.

[50] Jiuyang Zhao and Jacqueline M Cole. A database of refractive indices and dielectric constants auto-generated using chemdataextractor. *Scientific data*, 9(1):192, 2022.

[51] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673, 2018.

[52] Qingyang Dong and Jacqueline M Cole. Auto-generated database of semiconductor band gaps using chemdataextractor. *Scientific Data*, 9(1):193, 2022.

[53] Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022.

[54] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[55] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):77, 2021.

[56] Pierre-Paul De Breuck, Matthew L Evans, and Gian-Marco Rignanese. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *Journal of Physics: Condensed Matter*, 33(40):404002, 2021.

[57] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

[58] Branimir Radisavljevic, Aleksandra Radenovic, Jacopo Brivio, Valentina Giacometti, and Andras Kis. Single-layer mos2 transistors. *Nature nanotechnology*, 6(3):147–150, 2011.

[59] Albert Polman, Mark Knight, Erik C Garnett, Bruno Ehrler, and Wim C Sinke. Photovoltaic materials: Present efficiencies and future challenges. *Science*, 352(6283):aad4424, 2016.

[60] E Fred Schubert and Jong Kyu Kim. Solid-state light sources getting smart. *Science*, 308(5726):1274–1278, 2005.

[61] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[62] Jochen Heyd and Gustavo E Scuseria. Efficient hybrid density functional calculations in solids: Assessment of the heyd–scuseria–ernzerhof screened coulomb hybrid functional. *The Journal of chemical physics*, 121(3):1187–1192, 2004.

[63] Alejandro J Garza and Gustavo E Scuseria. Predicting band gaps with hybrid density functionals. *The journal of physical chemistry letters*, 7(20):4165–4170, 2016.

[64] Matteo Gerosa, Carlo Enrico Bottani, Lucia Caramella, Giovanni Onida, Cristiana Di Valentin, and Gianfranco Pacchioni. Electronic structure and phase stability of oxide semiconductors: Performance of dielectric-dependent hybrid functional dft, benchmarked against gw band structure calculations and experiments. *Physical Review B*, 91(15):155201, 2015.

[65] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):15679, 2017.

Appendix A. Related work

Multi-task learning, as a ML strategy, offers a promising alternative by enabling models to simultaneously learn multiple related tasks through shared representations, thereby improving generalization capabilities [21]. It has demonstrated significant value in materials science by effectively capturing the inherent correlations among various material properties [22]. When applying this strategy in the materials domain, several key factors are usually considered: the careful identification and selection of physically meaningful features derived from descriptors associated with physical, chemical, and geometrical properties [23]; the reliance on standardized representation formats such as SMILES strings for polymeric materials [24] to ensure compatibility; and a focus on interrelated physical properties, such as Formation Energy (ΔE^f), Band Gap (E_g), and Fermi Energy (E^F) [22]. In contrast, LLMs has potential to bypass these constraints as inherently multi-task models. LLMs leverage natural language as a universal carrier of information, allowing them to process a diverse range of tasks and datasets without the need for domain-specific feature selection. Recent works [16, 25, 26, 27] have demonstrated the potential of LLM in materials science; however, these studies focused on fine-tuning separate models for individual property prediction tasks, leaving the potential of a unified predictive framework largely unexplored.

Appendix B. Datasets

To improve the capability of LLMs, we generated scientific QA pairs as training set of QA fine-tuning using SciQAG framework [18]. The main idea is to train a QA generator to convert full-text scientific papers into QA pairs and use an evaluator to filter out those that do not meet quality standards. The task of QA generator is defined as follows: given seed input texts T, for each input text t, the generator should firstly generate 15 keywords k that capture the most important terms and concepts in the text, then generate a set $S = \{(q_i, a_i)\}_{i=1}^{n}$ focusing on the generated keywords k, where $\forall i \in \{1, ..., 10\}$, q_i is the question and a_i is the answer to q_i. To generate S, one should learn a generator $G(S|T;\theta)$ with θ the model parameters. Thus, given a new input text $\hat{t}$, following $G(S|T;\theta)$, one can directly generate a $\hat{S}$ consisting of QA pairs (by firstly generating 15 keywords to guide the QA generation). To fine-tune an open-source LLM as QA generator, we first randomly selected 700 papers from the paper collection as input to produce 7000 seed QA pairs by prompting GPT-4 (see **??**). Then, we fine-tuned `llama3-64k` model [15] on seed data. The data employs the instruction schema [28] composed of three elements: <instruction>, <input>, and <output>. The seed QA generation prompt was converted into the <instruction>. The seed paper filled the <input> field, and the <output> were the generated seed QA pairs. We concatenated the instruction and instance input as a prompt and train the model to generate the instance output in a standard supervised way. Using the trained QA generator, we performed inference on the remaining papers to form a training set. To reduce the occurrence of article-specific information in questions (i.e., non-knowledge-based questions that can only be answered using the given article information), a simple rule-based approach was used to remove all pairs containing “this paper" or “this study". The distribution of scientific QA categories of final 332,997 QA pairs can be found in A2.

FAIR stands for 'Findable, Accessible, Interoperable, and Reusable', which is a set of principles for enhancing the value and accessibility of data [29]. Due to the strong impact of 4V (volume, variety, velocity, and veracity) of big data on materials science, efforts have been made in recent years to collect comprehensive data from research groups worldwide, including unpublished data, and ensure its FAIRness [30]. We collect 21 open-accessed FAIR datasets from highly cited publications in materials science. 5 classification tasks and 17 regression tasks are derived from these datasets according to their property types. It is important to note that there is not a one-to-one correspondence between tasks and datasets. In some cases, multiple tasks are derived from a single dataset. For instance, from the MoosaviDiversity dataset [31], we construct two regression tasks, R5 and R6, which predict the log Henry's Law constant for CH4 and CO2, respectively, based on SMILES representations. Conversely, some datasets are consolidated into a single task to prevent data leakage. For example, both the ESOL [32] and DLS-100 [33] datasets, which focus on solubility prediction, are merged into a single regression task (R17). The visualization can be found in A1. Following task partitioning, we design prompt templates to transform tabular data samples into natural language sentences suitable for each task. These templates follow an instruction-based format to align with the LLaMA fine-tuning paradigm. Detailed specifications of the prompt templates are also provided.

- **Matbench_is_metal** [34]: Dataset from Zhuo et al.'s work, containing 4921 chemical formulas for classifying metallicity from composition. **[Task: C1]**
- **Matbench_glass** [35]: Retrieved from the Landolt–Börnstein collection, containing 5680 chemical formulas for bulk metallic glass formation ability classification. **[Task: C2]**
- **Pei** [36]: Dataset of 1252 observations (625 single-phase, 627 multi-phase alloys) for binary classification of alloy phases. **[Task: C3]**
- **WaterStability** [37]: Dataset of water stability for 200 MOFs, including metal node, organic ligand, and molar ratios. Contains 170 pairs of activated formula units and stability (high/low). **[Task: C4]**
- **UV** [38]: Includes 18,309 experimental UV/vis absorption maxima records. A subset of 5,158 SMILES is used for absorption region classification (ultraviolet or visible). **[Task: C5]**
- **NagasawaOPV** [39]: Dataset of 1203 experimental OPV material parameters. Three regression tasks: SMILES to bandgap, HOMO, and polydispersity index (PDI). **[Tasks: R1, R2]**
- **Matbench_steels** [34]: Retrieved from Citrine informatics, containing steel yield strengths for 312 chemical formulas. **[Task: R3]**
- **ChEMBL** [40]: Curated database of 1899 bioactive molecules, focused on lipophilicity (water-octanol partition coefficient, logD). **[Task: R4]**
- **MoosaviDiversity** [31]: Dataset with 5941 SMILES and log Henry's Law constants for CH4 and CO2 from experimental CoRE-2019 [41]. **[Tasks: R5, R6]**
- **MoosaviCp** [42]: Predicts the heat capacity of materials using density functional theory simulations. **[Task: R7]**
- **FreeSolv** [43]: Experimental and calculated hydration free energies for 641 SMILES. **[Task: R8]**
- **photoswitch** [44]: Contains 405 experimentally determined photoswitch properties, used to predict E-isomer transition wavelength from SMILES. **[Task: R9]**
- **SuperCon_ML** [45]: Over 16,000 compositions from SuperCon [46] database, used for critical temperature (Tc) prediction. **[Task: R10]**
- **UCSB+ESTM** [47, 48]: Combination of UCSB (1,100 thermoelectric materials) and ESTM databases to predict thermoelectric figure of merit (zT) for 5747 compositions with temperature conditions. **[Task: R11]**
- **TADF** [49]: Contains 5,349 TADF molecule records. Three tasks: delayed lifetime (435 samples), emission wavelength (937 samples), and photoluminescence quantum yield (719 samples). **[Tasks: R12, R13, R14]**
- **Refractive** [50]: 49,076 refractive index and 60,804 dielectric constant records on 11,054 unique chemicals. 6,262 pairs of compound and refractive index. **[Task: R15]**
- **Matbench_expt_gap** [51]: Experimental band gaps and DFT zero band gaps for 4604 compounds. **[Task: R16]**
- **Semiconductor** [52]: Auto-generated database of 100,236 semiconductor band gap records. 5,000 samples used for material name and averaged bandgap prediction. **[Task: R16]**
- **QMUG** [53]: Quantum mechanical properties of 665k biologically relevant molecules. 6,592 samples for SMILES to HOMO-LUMO gap prediction. **[Task: R16]**
- **ESOL** [32]: Compilation of aqueous solubility (LogS) values for 927 molecular compounds, relevant for drug discovery. **[Task: R17]**
- **DLS-100** [33]: 100 molecules with intrinsic aqueous solubility measurements, including a 75-25 training-test split. **[Task: R17]**

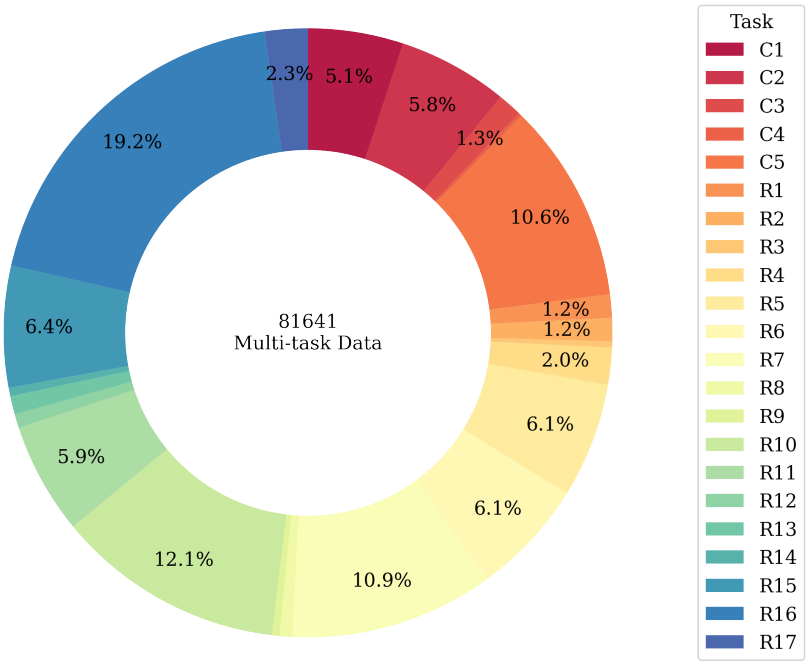

Fig. A1: Visualization of multi-task dataset size. Labels of all percentages below 1% are hidden.

Label definition

<material_type>: The specific format (e.g. composition) used to represent a material's characteristics or structure.

<material_representation>: The actual representation of material (e.g. TiO2), typically showing its elemental composition or structural details.

<yes_no>: A binary response indicating the presence or absence of a specific characteristic. The format typically follows the pattern "Yes/No".

<has_>: A linguistic marker that indicates the presence or absence of a specific material property or characteristic, typically used to explicitly state whether a material possesses a particular attribute. It connects the material representation with a specific property in a grammatically complete statement. The format typically follows the pattern "has/have/does not have".

<property>: A characteristic or attribute (e.g. bandgap) of a material that describes its behavior, performance, or physical characteristics.

<property_value>: The numerical or quantitative measurement of a specific property of the material.

- **Classification templates**

```
Template 1:
Instruction: Tell me if
given <material_type> <has_>
<property>.
Input: <material_representation>
Output: <yes_no>,
<material_representation>
<has_> <property>.

Template 2:
Instruction: Does given
<material_type> <has_>
<property>?
Input: <material_representation>
Output: <yes_no>,
<material_representation>
<has_> <property>.
```

- **Regression templates**

```
Template 1:
Instruction: Given a
<material_type>, write its
<property>.
Input: <material_representation>
Output: <property_value>.

Template 2:
Instruction: Predict the
<property> of this given
<material_type>.
Input: <material_representation>
Output: <property_value>.

Template 3:
Instruction: What is the
```

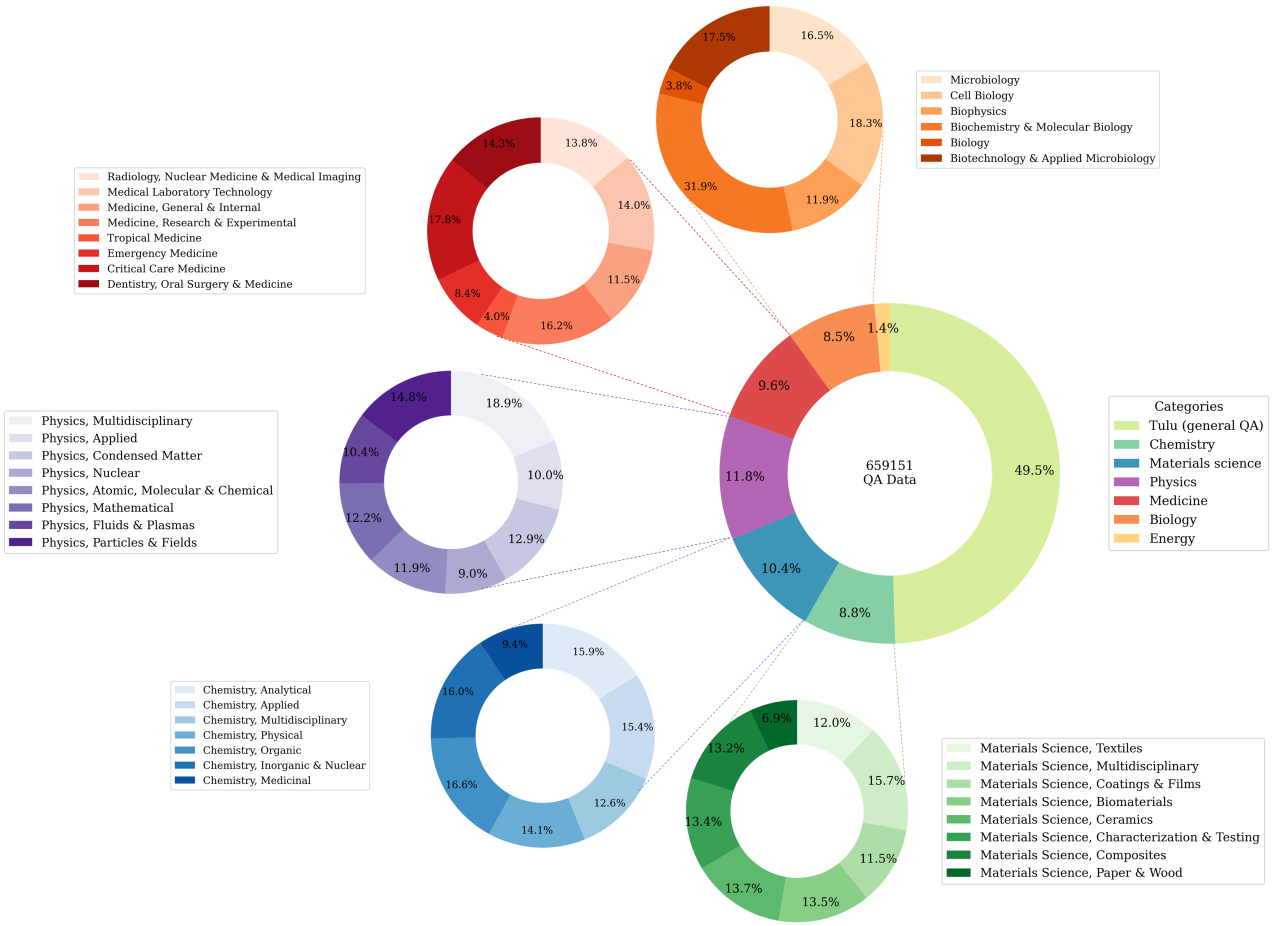

Fig. A2: Visualization of QA data categories.

```
<property> of this given
<material_type>?
Input:  <material_representation>
Output:  <property_value>.
```

Note: The above templates are applicable to data conversion for most tasks, with individual tasks having some adjustments. Please refer to the following examples for specifics.

Appendix C. GPT baselines and machine learning baselines

In Figure A3, we compare the performance of QA-MT with closed-source GPT-series models and competitive ML algorithms ('Machine learning results' section in Methods). For the GPT-series, we conducted single-task fine-tuning experiments with GPT-3.5 and few-shot prompting with GPT-4 [54] (the maximum file upload size per fine-tuning job is limited to 16 MB and our data volume exceeded this size limitation for both the first-stage QA and the second-stage multi-task fine-tuning). The results show that QA-MT consistently outperforms GPT-3.5 (fine-tuned) and GPT-4 (few-shot) across most tasks. Notable exceptions are observed in tasks C4 (MOF, water stability) and R5 (SMILES, log Henry's Law constant for CO2), where the GPT-series models slightly surpass QA-MT. Additionally, QA-MT achieves better performance than the ML baselines in 11 out of 14 tasks where ML baselines are available, highlighting its effectiveness in handling diverse material science applications. We include results from several ML models as references, like CrabNet [55], MODNet (v0.1.1) [56], and AMMExpress v2020 from matbench [34] and ML algorithm Gaussian process regression (GPR) used in GPTchem [16]. For some tasks, we directly use the ML result from its original papers of FAIR dataset. It should be noted that ML result of each task was individually trained on a specific single-task data. Take GPR for R1 and GPR for R2 as an example, they are results from two GPR models trained on R1 and R2 task data, respectively. For each ML algorithm we used in this study, it usually receives only one kind of input format and does not have results for all tasks. And due to different representation formats of the input, not all tasks have ML baselines. For example, R16 task data contains common names of the materials which cannot be converted into input of ML algorithms.

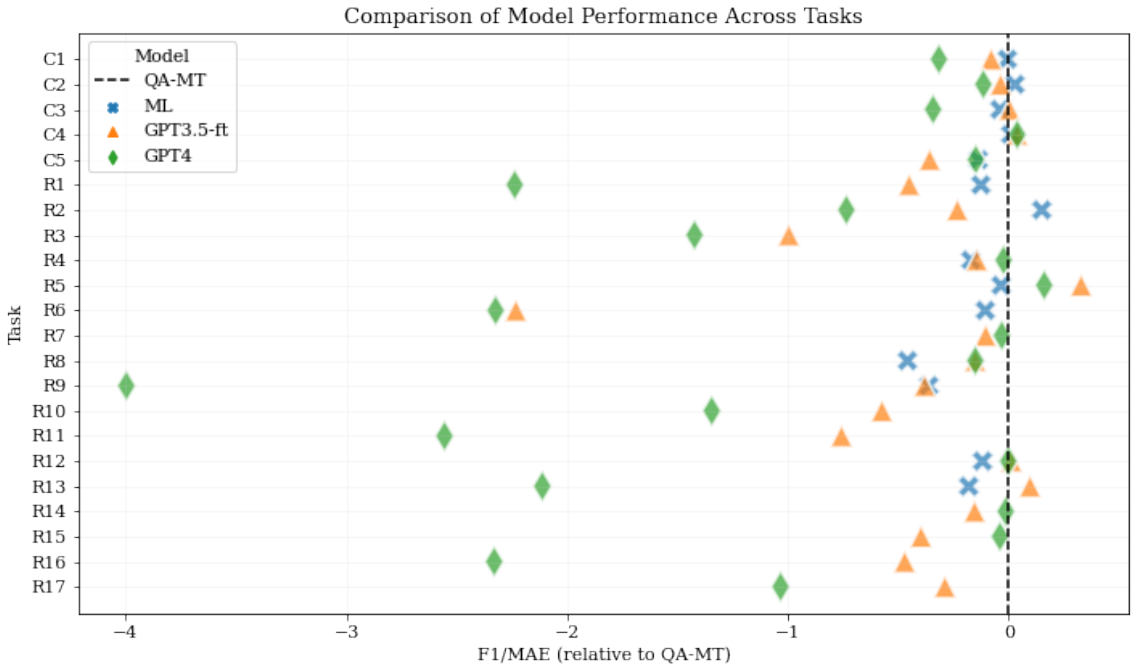

Fig. A3: Comparison of model performance relative to QA-MT across tasks. Points on the left of the vertical line indicate poorer performance compared to QA-MT. The x-axis represents the performance ratio, where a value of -1 indicates performance 100% worse than QA-MT. For tasks with more than one ML results, the best one is used in this visualization.

Appendix D. Fine-tuning strategy

For all QA-generator model, QA-base model and ST/MT models, We fine-tune the LLaMA models following established methods, using a setup of 8xAMD MI250X GPUs and employing the Brain Floating Point 16 (BF16) data format for an optimal balance between precision and computational efficiency. For inference, we use a temperature of 0.6 and top_p of 0.9 for logical and diverse text generation.

Similarly, for QA-base model and MT model, the `LLaMA-7b` model is fine-tuned. DeepSpeed stage 2 [57] is employed with a batch size of 2 per device.During inference, we set the temperature to 0.8 and top_p to 0.75, which makes text generated more logical with rich vocabulary.

In experiments involving the `gpt-series` models, we use `gpt-3.5-turbo-0613` for fine-tuning and `gpt-4-0613` for few-shot learning. For training, we utilize default parameters determined algorithmically by Azure OpenAI based on the size of the training data. During inference, we set the temperature to 0.8 to enhance generation diversity.

Appendix E. The application of QA-MT on bandgap prediction

The bandgap, a fundamental electronic property of materials, is the energy difference between the highest occupied molecular orbital (valence band) and the lowest unoccupied molecular orbital (conduction band). It plays a crucial role in determining a material's electrical and optical properties, making it a key parameter in fields like transistors [58], photovoltaics [59], and light-emitting diodes (LEDs) [60]. For AI models in materials science, predicting bandgaps serves as a robust benchmark to evaluate their effectiveness and adaptability to domain-specific tasks. Accurate bandgap predictions can indicate a model's capability in capturing essential material properties.

We compare several common methods for bandgap prediction with our QA-MT model, including:

1. Perdew–Burke–Ernzerhof (PBE) [61]: A generalized gradient approximation (GGA) functional for density functional theory (DFT) that improves upon local density approximation (LDA) by including electron density gradients.

2. Heyd–Scuseria–Ernzerhof (HSE) [62]: A hybrid functional that combines PBE with a fraction of exact exchange at short ranges, providing better accuracy but requiring more computational resources. Its accuracy for wide band gap materials can be enhanced by increasing the short-range Hartree-Fock exchange component, while maintaining semiconductor accuracy [63].

3. GW Approximation [64]: A many-body perturbation theory approach that calculates the electronic self-energy by expanding it in terms of the single particle Green's function (G) and the screened Coulomb interaction (W).

4. AFLOW [65]: A PBE-trained ML model

which predicts various material properties, including bandgaps, focusing on efficiency and scalability.

We predict bandgap for 7 specific compositions, unseen in QA-MT corresponding training set, that cover a wide energy range. These materials were chosen as test cases because researchers have thoroughly studied them using various theoretical methods, making them valuable benchmarks for evaluating new exchange-correlation functionals. We calculated results from each method, including Mean Absolute Deviation (MAD) and Root Mean Squared Error (RMSE) compared to experimental bandgap values ($E_{\mathrm{g,exp}}$). They both evaluate the error between predicted and observed values, while RMSE penalizes larger errors more significantly. PBE method suffers from DFT systematic error and usually underestimates the bandgap values compared with the experimental $E_{\mathrm{g,exp}}$. Compared with PBE, HSE raises more computational cost but the discrepancy between computation and experiment has undoubtedly been reduced. GW, the most complex and computational expensive one, shows the most accurate results (0.12 MAD and 0.15 RMSE) for test cases. Compared with these first-principle methods, AFLOW has improved based on PBE but still far from HSE, with lower inference costs and faster speed as a ML algorithm. It also has same systematic error as PBE since it trained using PBE values. Our model QA-MT achieves a MAD of 0.51 and an RMSE of 0.65, which are comparable to the HSE results. This shows that our model, while not reaching GW's accuracy, performs significantly better than other methods like PBE and AFLOWE.

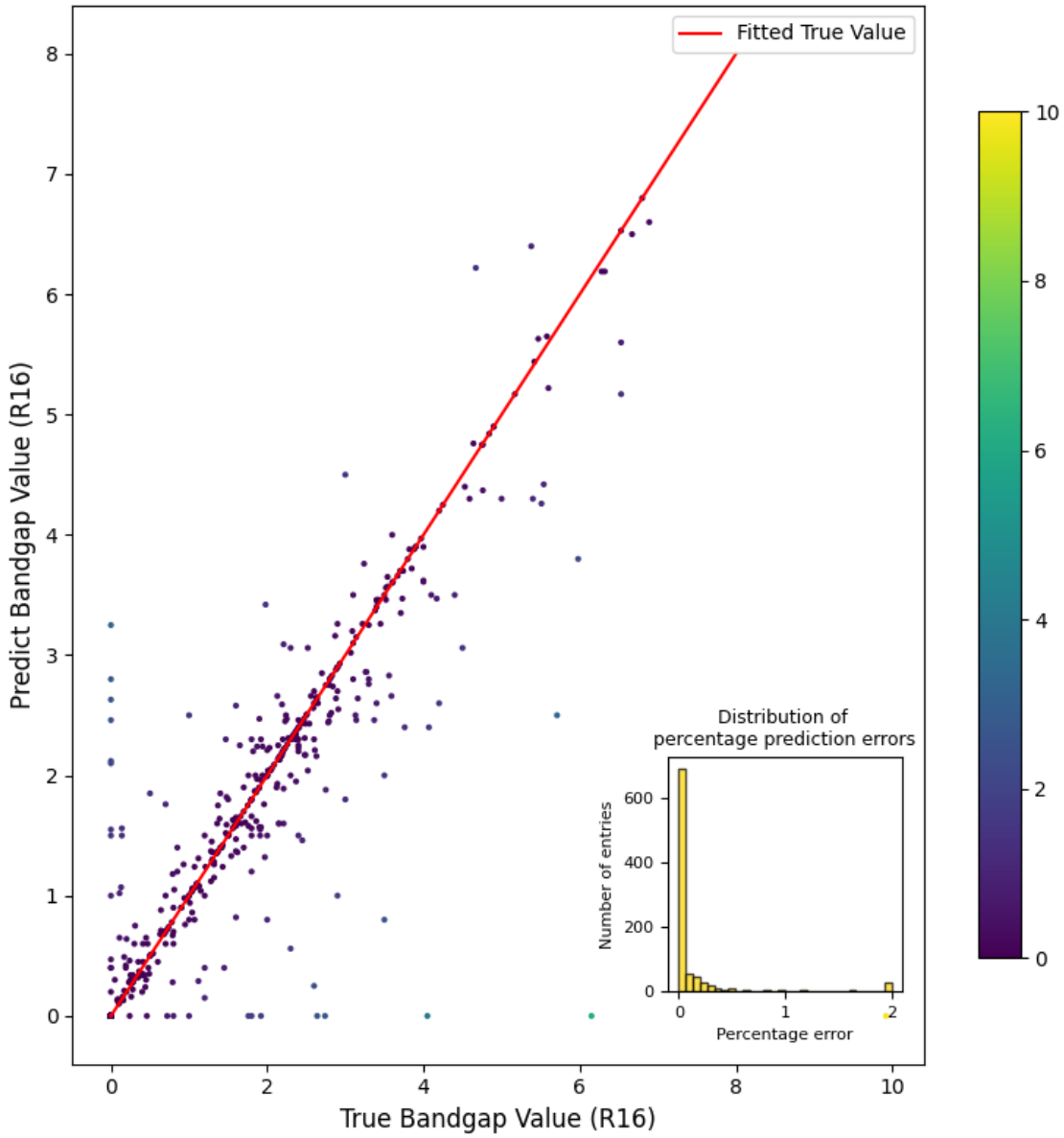

Fig. A4: Comparison between true and predicted bandgap values, with percentage error