SEMANTIC DEVICE GRAPHS FOR PEROVSKITE SOLAR CELL DESIGN

Anagha Aneesh
FSU Jena
anagha.aneesh@uni-jena.de
Nawaf Alampara
FSU Jena
nawaf.alampara@uni-jena.de

José A. Márquez
HU Berlin
jose.marquez@physik.hu-berlin.de

Kevin Maik Jablonka
FSU Jena, CEEC Jena, JCSM Jena, HIPOLE Jena
mail@kjablonka.com

ABSTRACT

Materials science faces two persistent challenges: the multiscale nature of functional devices, where performance emerges from the complex interplay of components across different length scales, and the prevalence of incomplete characterization data that precludes conventional featurization approaches. These challenges are exemplified in perovskite solar cells, where device optimization requires consideration of multiple interacting layers while much of the materials data exists only as text descriptions. While machine learning has accelerated the discovery of isolated material properties, translating promising materials into functional devices remains a significant bottleneck. Here, we introduce *semantic device graphs*: a physics-inspired representation that captures the multi-scale architecture of perovskite solar cells while leveraging large language models to generate meaningful embeddings for incomplete material descriptions. Our approach achieves a 10% improvement in performance prediction compared to state-of-the-art methods (CrabNet), enabling holistic device optimization rather than isolated material screening. The framework generates physically meaningful device fingerprints that reveal patterns in high-performing architectures, providing insights for future device optimization. This work demonstrates how combining physics-informed architectural choices with language models can address fundamental materials science challenges of multiscale modeling and incomplete information, serving as a stepping stone toward more holistic materials discovery approaches.

1 INTRODUCTION

The accelerated discovery of new materials through machine learning has primarily focused on predicting isolated properties, such as bandgaps or formation energies (Dunn et al., 2020; Alampara et al., 2024; Choudhary et al., 2024; Lee et al., 2023; Jablonka et al., 2020). However, real-world device optimization requires consideration of multiple interacting components and their collective performance (Moosavi et al., 2020). This multiscale challenge is particularly evident in perovskite solar cells, where device efficiency emerges from the complex interplay of multiple material layers and their interfaces (Stolterfoht et al., 2019).

Despite possessing exceptional optoelectronic properties,(Guo et al., 2023) the commercialization of perovskite technologies continues to span decades (Unold, 2022; Dale & Scarpulla, 2023). A fundamental barrier to acceleration lies in the nearly infinite combinations of materials and device architectures required to construct a complete solar cell. While historical design decisions have been guided by physics-based intuition and lessons from previous technologies, the rapidly expanding materials landscape and need for acceleration have rendered this traditional approach a significant bottleneck.

The conventional sequential approach—screening individual materials before considering their integration—can be misleading, potentially overlooking promising combinations that only reveal their potential when considered holistically. This challenge mirrors other complex material systems where performance depends on the interaction of multiple components across different scales where information is available at different qualities and resolutions for different scales (Charalambous et al., 2024).

In this work, we introduce *semantic device graphs*: a physics-inspired graph-based representation of perovskite solar cells that bridges multiple scales of device architecture (see Figure 1). Our approach uniquely leverages large language models (LLMs) to generate embeddings for incomplete material descriptions, enabling meaningful representation even when detailed characterization data is unavailable. This allows us to capture relevant context and prior knowledge for materials where conventional materials informatics tools cannot be applied due to incomplete information.

This work demonstrates how combining physics-informed architectural choices with language models can address fundamental materials science challenges of multiscale modeling and incomplete information. Our framework serves as a stepping stone toward more holistic materials discovery approaches that can better translate promising materials into practical devices.

Concretely, our main contributions are:

- **Semantic device graphs:** We introduce semantic device graphs, a novel physics-inspired graph representation that captures the multi-scale architecture of perovskite solar cells while preserving the hierarchical relationships between different device components.
- **LLM embeddings for incomplete data:** We demonstrate how large language model embeddings can be effectively used to represent materials in device stacks where conventional featurization approaches fail due to incomplete characterization data, achieving a 10% improvement in performance prediction compared to state-of-the-art methods (Wang et al., 2021).
- **GNN for device property prediction:** We develop a heterogeneous graph neural network architecture that combines material-level and layer-level information through specialized node types and edge connections, enabling holistic device optimization rather than isolated material screening.
- **Cartography of perovskite solar cells:** We provide insights into perovskite solar cell design through learned device fingerprints, revealing patterns in high-performing architectures that can guide future device optimization.

2 Related Work

Prediction of perovskite material and solar cell properties Various ML algorithms, including Gradient Boosting Regression (GBR), Kernel Ridge Regression (KRR), and Support Vector Machines (SVM), have been employed to predict bandgaps of metal halide perovskites (Parikh et al., 2022; Im et al., 2019; Li et al., 2019).

Some works also considered the prediction of device properties such as the photoconversion efficiency (PCE) based on device properties but leveraged expensive and time-consuming featurization such as computing the physics-inspired (Godovsky, 2011) difference in highest occupied molecular orbital (HOMO) and lowest unoccupied orbital (LUMO) energies between transport layers and absorbers (Li et al., 2019).

Importantly, this information is not only expensive to compute but also impossible to compute for the largest dataset of perovskite device properties: The perovskite database (Jacobsson et al., 2021), which was created by manually extracting device and performance properties from more than 15,000 papers. Many parts of the device stack of a perovskite solar cell are only encoded as a text string (often an abbreviation).

Multiscale graph representations Graph representations have emerged as a powerful tool for modeling materials across different scales, particularly at the mesoscale and microstructural levels. For example, grains in polycrystalline materials can be represented as nodes in a graph, with

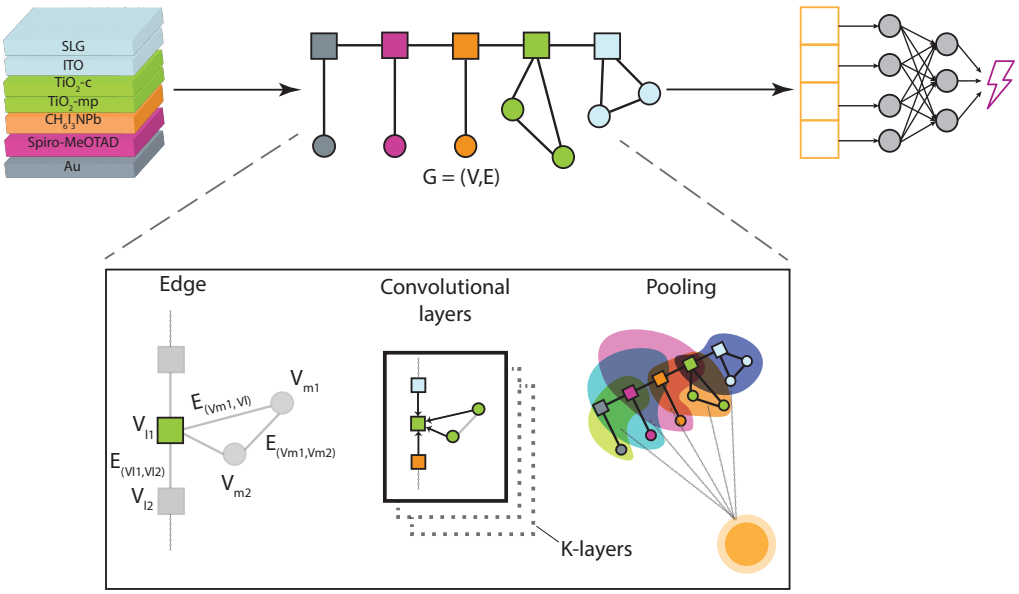

Figure 1: **Overview of the modeling approach.** We convert the device stack of perovskite solar cells in which materials are often only described with a string, such as abbreviations, into semantic device graphs. We use LLM-derived embeddings as the node feature vectors and use GNNs to update the representations. Following mean pooling, the representations are used to predict device performance metrics.

edges representing the physical interfaces between neighboring grains. The nodes typically contain features such as Euler angles, grain volume, and number of neighbors (Dai et al., 2021).

Recent work has shown the effectiveness of heterogeneous graph representations for complex nanomaterials. Sivonxay et al. (2024) demonstrated this for upconverting nanoparticles by representing both dopant species and their energy transfer interactions as distinct node types. This approach allows capturing both spatial relationships and physical interactions while maintaining differentiability for inverse design.

3 METHODS

Data preprocessing The experimental data was obtained from the NOMAD (Scheidgen et al., 2023)repository, which hosts an updated version of the perovskite database.(Márquez & Scheidgen, 2024; Jacobsson et al., 2021) During preprocessing, duplicate entries were eliminated based on device performance metrics and stack configurations to ensure that only unique device configurations are in the dataset.

Crossvalidation To measure performance on unseen data points, we performed a random train/valid/test split, keeping $80\,\%$ for training and $10\,\%$ for testing and validation, respectively.

Embeddings For the material name strings, we obtained embeddings using OpenAI's `Text-embedding-3-small` model.

Graph representation We represent a perovskite solar cell as a heterogeneous graph $G = (V, E)$, where:

- $V = V_m \cup V_l$ represents the set of vertices consisting of:
  - Material nodes V_m representing individual layer materials
  - Meta-nodes V_l representing layer types

The edge set E contains three types of connections:

1. Material-to-meta edges (v_m, v_l) connecting materials to their layer type
2. Material-to-material edges (v_{m1}, v_{m2}) connecting materials within the same layer
3. Meta-to-meta edges (v_{l1}, v_{l2}) connecting adjacent layer types

Each edge $e_{ij} \in E$ is represented by a one-hot encoded vector $\mathbf{t}_{ij} \in \{0, 1\}^3$ indicating the edge type.

Node features are initialized as:

- For material nodes v_m: LLM embeddings of material names $\mathbf{h}_m \in \mathbb{R}^d$
- For meta-nodes v_l: average of the material node embeddings v_m linked to this meta-node

Multiple instances of the same layer type (e.g., in multi-junction cells) are represented by distinct meta-nodes while maintaining the same connectivity pattern.

Graph convolutional neural network We employ a graph-convolutional neural network (Gilmer et al., 2017) to learn device properties. After K graph convolutional layers, the device properties are predicted based on a mean-pooling of the node embedding:

$$\mathbf{h}_G = \frac{1}{|V|} \sum_{i \in V} \mathbf{h}_i^{(K)} \tag{1}$$

Those are passed to a fully-connected neural network (FCNN):

$$\hat{y} = \text{FCNN}(\mathbf{h}_G) \tag{2}$$

The model is optimized using mean squared error loss between predicted and reported device properties.

Architectural details and hyperparameters are provided in Appendix A.1.

4 RESULTS

4.1 PREDICTIVE PERFORMANCE

In Figure 2, we show the performance of our semantic device graph-based model in predicting photoconversion efficiencies. We evaluate model performance using R^2 and mean average error. In Figure 2A. we provide a comparison between our model and those similar to current approaches for modeling solar cells. A baseline model using a one-hot encoding of the absorber material achieves a R^2 of 0.22, while CrabNet (Wang et al., 2021; 2022), a state-of-the-art model, improves to 0.35.

In addition, we also trained a random forest model on the layer-averaged LLM embeddings that we used in our GNN. We find that the additional semantic context the LLM embeddings provide leads to markedly improved predictive performance. We also find that the LLM embeddings outperform those from MatBERT (MAE of $3.38\,\%$), a BERT-based language model specifically trained on materials science literature (Walker et al., 2021).

We observe the best performance (MAE of almost $2.75\,\%$ for PCE) using our GNN architecture, which companies the semantic context provided by the LLM embeddings with physically meaningful inductive biases, such as the connectivity of the graphs, which reflects the different hierarchies in a device and what materials are in physical contact with each other.

In Figure 2B. we plot reported photoconversion efficiencies against the ones predicted by our GNN operating on the semantic device graphs. We can observe that many of the erroneous predictions can be traced to devices with low performance, which might be addressed with improved data preprocessing.

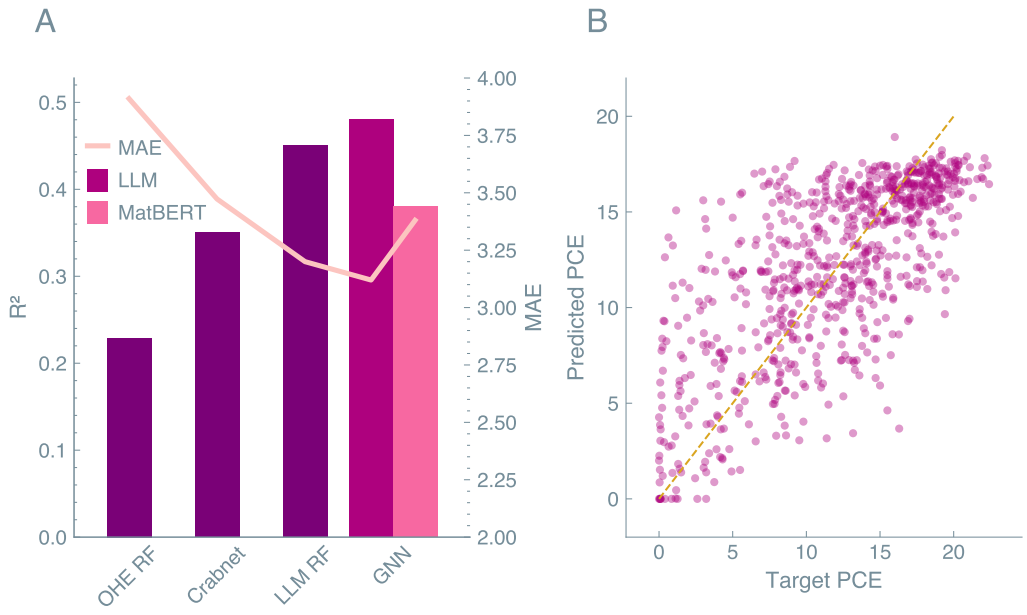

Figure 2: **Analysis of predictive performance of our semantic device-graph-based GNNs and baselines.** A. As the simplest baseline, we consider one-hot-encoded (OHE) materials as input for random forest models. This model is outperformed by CrabNet, a transformer-based model that is optimized for compositions. We can further improve performance by using LLM-derived embeddings instead of OHE to describe materials as input to our random forest. We observe the best performance with our semantic-device graph-based GNN using LLM embeddings compared to MatBERT. B. The parity plot for the predictions of our semantic device graph-based GNN on the test set shows that some errors can still be observed for devices with low performance. This might be addressed with further data preprocessing.

4.2 Cartography of perovskite solar cell devices

Our approach cannot only predict device performances but also device data-driven device fingerprints. That is, we can now convert device stacks into vectors and compare device stacks by their proximity in this vector space. Figure 3 shows how the devices organize in the latent space according to their PCE. For this figure, we embed device stacks using our trained GNNs and reduce the dimensionality using t-SNE (Van der Maaten & Hinton, 2008). We find a color gradient in the image, indicating that the space is meaningfully organized and can be used to distinguish high- from low-performing materials. A similar clustering relationship is observed using PCA dimensionality reduction as shown in Appendix A.3.

Importantly, our architecture allows us to not only compute embeddings for the entire device stack for any layer type and material — enabling continuous similarity measures across the most relevant length scales in the device.

5 Limitations and Future Work

Several important limitations should be considered when interpreting our results. First, our model currently does not account for processing conditions, which significantly impact device performance. These conditions, including deposition methods, annealing temperatures, and environmental factors during fabrication, can dramatically affect device efficiency even for identical material stacks. Detailed error analysis is provided in Appendix A.2. Second, the underlying dataset contains inherent noise from variations in reporting standards and characterization methods across different laboratories. Third, while effective, our graph architecture remains relatively simple compared to

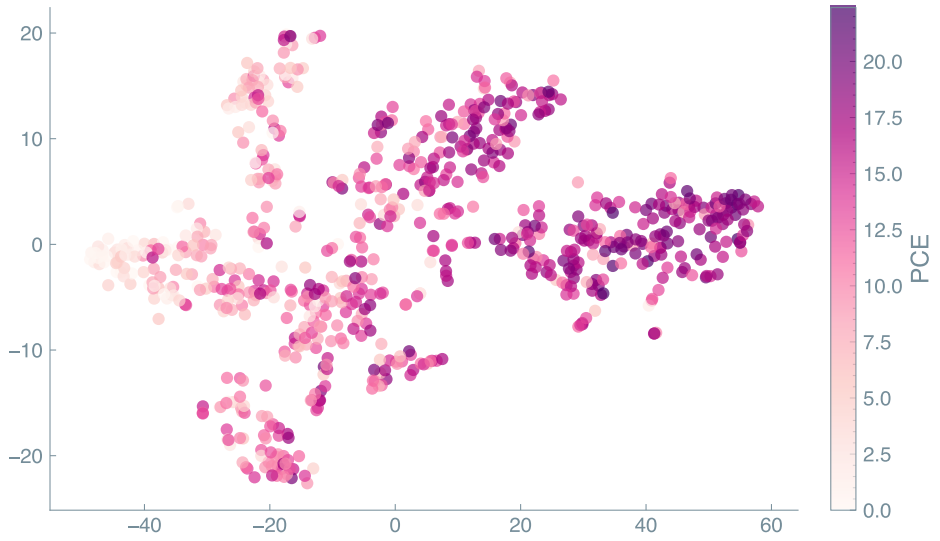

Figure 3: **Two-dimensional representation of device stacks colored by photoconversion efficiencies.** For this figure, we embed the device stacks using our trained GNNs. We then reduce the dimensionality using t-SNE. We can observe that device stacks tend to cluster but also organize in the latent space according to PCE (shown in color).

recent advances in graph neural networks. More sophisticated architectures could potentially extract additional insights from the device structure. Finally, our current representation treats the absorber layer as a single node, whereas a more granular approach treating individual ions as separate nodes could capture additional chemical insights.

6 Conclusions

The prediction and optimization of solar cell architectures represents one of the most impactful applications of machine learning in materials science. However, progress has been bottlenecked by three fundamental challenges: the multiscale nature of device optimization, incomplete materials characterization data, and the vast combinatorial space of possible device configurations. Traditional approaches relying on sequential screening of individual materials have proven inadequate, potentially overlooking promising combinations that only reveal their potential when considered holistically. Previous attempts to address these challenges have been limited by their reliance on expensive computational screening or complete materials characterization data. Here, we have demonstrated that semantic device graphs, combined with language model embeddings, can effectively bridge these gaps by capturing both the physical structure of devices and the semantic relationships between materials. Our approach not only improves predictive performance but also provides interpretable device fingerprints that can guide future optimization efforts. This work highlights the potential of combining physics-inspired architectures with modern machine-learning techniques to address complex materials engineering challenges. The framework we present could accelerate the development of next-generation solar cells while providing a template for similar multiscale optimization problems across materials science.

7 Acknowledgment

This work was supported by the Carl Zeiss Foundation and Intel and Merck via the AWASES programme. A.A. acknowledges financial support for this research by the Fulbright U.S. Student Pro-

gram, which is sponsored by the U.S. Department of State and the German-American Fulbright Commission. Its contents are solely the responsibility of the author and do not necessarily represent the official views of the Fulbright Program, the Government of the United States, or the German-American Fulbright Commission. K.M.J. is part of the NFDI consortium FAIRmat funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 460197019.

REFERENCES

Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? *arXiv preprint arXiv: 2406.17295*, 2024.

Charithea Charalambous, Elias Moubarak, Johannes Schilling, Eva Sanchez Fernandez, Jin-Yu Wang, Laura Herraiz, Fergus Mcilwaine, Shing Bo Peh, Matthew Garvin, Kevin Maik Jablonka, Seyed Mohamad Moosavi, Joren Van Herck, Aysu Yurdusen Ozturk, Alireza Pourghaderi, Ah-Young Song, Georges Mouchaham, Christian Serre, Jeffrey A. Reimer, André Bardow, Berend Smit, and Susana Garcia. A holistic platform for accelerating sorbent-based carbon capture. *Nature*, 632(8023):89–94, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07683-8. URL http://dx.doi.org/10.1038/s41586-024-07683-8.

Kamal Choudhary, Daniel Wines, Kangming Li, Kevin F Garrity, Vishu Gupta, Aldo H Romero, Jaron T Krogel, Kayahan Saritas, Addis Fuhr, Panchapakesan Ganesh, et al. Jarvis-leaderboard: a large scale benchmark of materials design methods. *npj Computational Materials*, 10(1):93, 2024.

Minyi Dai, Mehmet F. Demirel, Yingyu Liang, and Jia-Mian Hu. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. *npj Computational Materials*, 7(1), July 2021. ISSN 2057-3960. doi: 10.1038/s41524-021-00574-w. URL http://dx.doi.org/10.1038/s41524-021-00574-w.

Phillip J. Dale and Michael A. Scarpulla. Efficiency versus effort: A better way to compare best photovoltaic research cell efficiencies? *Solar Energy Materials and Solar Cells*, 251:112097, March 2023. ISSN 0927-0248. doi: 10.1016/j.solmat.2022.112097. URL http://dx.doi.org/10.1016/j.solmat.2022.112097.

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.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv: 1704.01212*, 2017.

Dmitry Godovsky. Modeling the ultimate efficiency of polymer solar cell using marcus theory of electron transfer. *Organic Electronics*, 12(1):190–194, 2011.

Jiahao Guo, Bingzhe Wang, Di Lu, Ting Wang, Tingting Liu, Rui Wang, Xiyue Dong, Tong Zhou, Nan Zheng, Qiang Fu, Zengqi Xie, Xiangjian Wan, Guichuan Xing, Yongsheng Chen, and Yongsheng Liu. Ultralong Carrier Lifetime Exceeding 20 µs in Lead Halide Perovskite Film Enable Efficient Solar Cells. *Advanced Materials*, 35(28):2212126, 2023. ISSN 1521-4095. doi: 10.1002/adma.202212126. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.202212126. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202212126.

Jino Im, Seongwon Lee, Tae-Wook Ko, Hyun Woo Kim, YunKyong Hyon, and Hyunju Chang. Identifying pb-free perovskites for solar cells by machine learning. *npj Computational Materials*, 5(1), March 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0177-0. URL http://dx.doi.org/10.1038/s41524-019-0177-0.

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: Materials genomics and machine learning. *Chemical Reviews*, 120(16):8066–8129, June 2020. ISSN 1520-6890. doi: 10.1021/acs.chemrev.0c00004. URL http://dx.doi.org/10.1021/acs.chemrev.0c00004.

T. Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, Aman Anand, Amran Al-Ashouri, Anders Hagfeldt, Andrea Crovetto, Antonio Abate, Antonio Gaetano Ricciardulli, Anuja Vijayan, Ashish Kulkarni, Assaf Y. Anderson, Barbara Primera Darwich, Bowen Yang, Brendan L. Coles, Carlo A. R. Perini, Carolin Rehermann, Daniel Ramirez, David Fairen-Jimenez, Diego Di Girolamo, Donglin Jia, Elena Avila, Emilio J. Juarez-Perez, Fanny Baumann, Florian Mathies, G. S. Anaya González, Gerrit Boschloo, Giuseppe Nasti, Gopinath Paramasivam, Guillermo Martínez-Denegri, Hampus Näsström, Hannes Michaels, Hans Köbler, Hua Wu, Iacopo Benesperi, M. Ibrahim Dar, Ilknur Bayrak Pehlivan, Isaac E. Gould, Jacob N. Vagott, Janardan Dagar, Jeff Kettle, Jie Yang, Jinzhao Li, Joel A. Smith, Jorge Pascual, Jose J. Jerónimo-Rendón, Juan Felipe Montoya, Juan-Pablo Correa-Baena, Junming Qiu, Junxin Wang, Kári Sveinbjörnsson, Katrin Hirselandt, Krishanu Dey, Kyle Frohna, Lena Mathies, Luigi A. Castriotta, Mahmoud. H. Aldamasy, Manuel Vasquez-Montoya, Marco A. Ruiz-Preciado, Marion A. Flatken, Mark V. Khenkin, Max Grischek, Mayank Kedia, Michael Saliba, Miguel Anaya, Misha Veldhoen, Neha Arora, Oleksandra Shargaieva, Oliver Maus, Onkar S. Game, Ori Yudilevich, Paul Fassl, Qisen Zhou, Rafael Betancur, Rahim Munir, Rahul Patidar, Samuel D. Stranks, Shahidul Alam, Shaoni Kar, Thomas Unold, Tobias Abzieher, Tomas Edvinsson, Tudur Wyn David, Ulrich W. Paetzold, Waqas Zia, Weifei Fu, Weiwei Zuo, Vincent R. F. Schröder, Wolfgang Tress, Xiaoliang Zhang, Yu-Hsien Chiang, Zafar Iqbal, Zhiqiang Xie, and Eva Unger. An open-access database and analysis tool for perovskite solar cells based on the fair data principles. *Nature Energy*, 7(1):107–115, December 2021. ISSN 2058-7546. doi: 10.1038/s41560-021-00941-3. URL http://dx.doi.org/10.1038/s41560-021-00941-3.

Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv: 2309.05934*, 2023.

Jinxin Li, Basudev Pradhan, Surya Gaur, and Jayan Thomas. Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. *Advanced Energy Materials*, 9(46), October 2019. ISSN 1614-6840. doi: 10.1002/aenm.201901891. URL http://dx.doi.org/10.1002/aenm.201901891.

Seyed Mohamad Moosavi, Kevin Maik Jablonka, and Berend Smit. The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society*, 142 (48):20273–20287, November 2020. ISSN 1520-5126. doi: 10.1021/jacs.0c09105. URL http://dx.doi.org/10.1021/jacs.0c09105.

José A. Márquez and Markus Scheidgen. Perovskite Solar Cell Database Project, 2024. URL https://nomad-lab.eu/prod/v1/gui/dataset/doi/10.17172/NOMAD/2024.09.24-1.

Nishi Parikh, Meera Karamta, Neha Yadav, Mohammad Mahdi Tavakoli, Daniel Prochowicz, Seckin Akin, Abul Kalam, Soumitra Satapathi, and Pankaj Yadav. Is machine learning redefining the perovskite solar cells? *Journal of Energy Chemistry*, 66:74–90, March 2022. ISSN 2095-4956. doi: 10.1016/j.jechem.2021.07.020. URL http://dx.doi.org/10.1016/j.jechem.2021.07.020.

Markus Scheidgen, Lauri Himanen, Alvin Noe Ladines, David Sikter, Mohammad Nakhaee, Ádám Fekete, Theodore Chang, Amir Golparvar, José A. Márquez, Sandor Brockhauser, Sebastian Brückner, Luca M. Ghiringhelli, Felix Dietrich, Daniel Lehmberg, Thea Denell, Andrea Albino, Hampus Näsström, Sherjeel Shabih, Florian Dobener, Markus Kühbach, Rubel Mozumder, Joseph F. Rudzinski, Nathan Daelman, José M. Pizarro, Martin Kuban, Cuauhtemoc Salazar, Pavel Ondračka, Hans-Joachim Bungartz, and Claudia Draxl. Nomad: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8 (90):5388, 2023. doi: 10.21105/joss.05388. URL https://doi.org/10.21105/joss.05388.

Eric Sivonxay, Lucas Attia, Evan Walter Clark Spotte-Smith, Benjamin Lengeling, Xiaojing Xia, Daniel Barter, Emory Chan, and Samuel Blau. Inverse design of complex nanoparticle heterostructures via deep learning on heterogeneous graphs. December 2024. doi: 10.26434/chemrxiv-2024-1dw4q. URL http://dx.doi.org/10.26434/chemrxiv-2024-1dw4q.

Martin Stolterfoht, Pietro Caprioglio, Christian M. Wolff, José A. Márquez, Joleik Nordmann, Shanshan Zhang, Daniel Rothhardt, Ulrich Hörmann, Yohai Amir, Alex Redinger, Lukas Kegelmann, Fengshuo Zu, Steve Albrecht, Norbert Koch, Thomas Kirchartz, Michael Saliba, Thomas Unold, and Dieter Neher. The impact of energy alignment and interfacial recombination on the internal and external open-circuit voltage of perovskite solar cells. *Energy &; Environmental Science*, 12(9):2778–2788, 2019. ISSN 1754-5706. doi: 10.1039/c9ee02020a. URL http://dx.doi.org/10.1039/C9EE02020A.

Thomas Unold. Accelerating research on novel photovoltaic materials. *Faraday Discussions*, 239: 235–249, 2022. ISSN 1364-5498. doi: 10.1039/d2fd00085g. URL http://dx.doi.org/10.1039/D2FD00085G.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*, 2021.

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. doi: 10.1038/s41524-021-00545-1.

Anthony Yu-Tung Wang, Mahamad Salah Mahmoud, Mathias Czasny, and Aleksander Gurlo. Crabnet for explainable deep learning in materials science: Bridging the gap between academia and industry. *Integrating Materials and Manufacturing Innovation*, 11(1):41–56, 2022. doi: 10.1007/s40192-021-00247-y. URL https://doi.org/10.1007/s40192-021-00247-y.

Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

A Appendix

A.1 Architectural details

Table 1: **Hyperparameters used in the model.**

parameter	value
GNN	
convolution dimension	64
number of convolutional layers	6
edge MLP dimension	32
dimension message passing layer	64
dropout ratio	0.138
pooling	mean
MLP	
layer sizes	1538, 768, 1
Optimization	
learning rate	1.162×10^{-3}
batch size	256

Hyperparameter optimization was performed using a sweep with Hydra (Yadan, 2019). This approach systematically explored different combinations of hyperparameters, specifically varying the number of graph layers, the number of neurons in message passing, and the number of neurons in the MLP upsampling one-hot encoded edge features. The final hyperparameters are listed in Table 1.

A.2 IMPACT FROM UNACCOUNTED DIFFERENCE IN DEVICE PARAMETERS

Table 2: **Duplicate device error analysis.**

dataset	R^2	MAE
all devices	0.44	3.05
unique devices	0.51	2.85
duplicate devices	0.43	2.91

The experimental dataset used to train the semantic device graph-based model consists solely of unique device configurations. To evaluate the impact of duplicate configurations, we performed a comparative error analysis between the predicted performance of duplicate and unique devices as shown in Table 2. The model achieves an overall R^2 of 0.44. However, when evaluating only duplicate devices, the performance drops slightly to 0.43. In contrast, model performance increases on unique devices, achieving an R^2 of 0.51 and the lowest MAE of 2.91. This analysis suggests that the presence of duplicate configurations negatively affects model performance. The current model architecture does not account for differences in device construction, like material processing and layer thickness, which would distinguish these duplicate devices.

A.3 VARIANCE-PRESERVED DIMENSIONALITY REDUCTION OF DEVICE STACKS VIA PCA

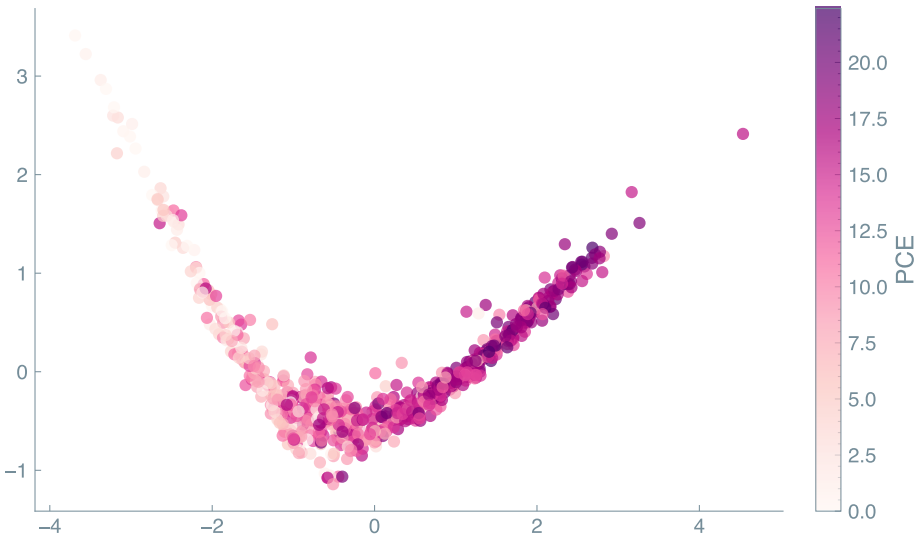

Figure 4: **PCA-based two-dimensional representation of device stacks colored by photoconversion efficiencies.** In this figure, dimensionality reduction reveals clustering patterns that correlate with device performance and confirm the efficiency-based grouping observed in the t-SNE analysis.

The consistency we observe between PCA and t-SNE representations provides clear validation of the clustering patterns because these methods fundamentally differ in their mathematical approach to dimensionality reduction. t-SNE optimizes for local structure preservation, while PCA is a linear reduction technique that maximizes global variance. The replication of efficiency-based clustering across both methods indicates that the observed groupings reflect underlying patterns in device stack embeddings rather than an artifact of t-SNE's nonlinear reduction process.