Loughborough University

Efficiency Prediction for Perovskite Solar Cells Using Automated Curation of Charge Transport Layer Material Properties

Master Thesis

---

by Ariane D. Wilhelm
Student ID: F317561

COPT18: Data Science

Supervisors:
Dr. Lars Nagel
Dr. Fasil Dejene

Collaborators:
Dr. José A. Márquez Prieto,
FAIRmat consortium

24.08.2024

Abstract

With the Perovskite Database, a large collection of perovskite solar cell data is available for analysis. In the database, the materials used as charge transport layers (CTLs)—key components for device performance—are only described by their mostly chemically meaningless common names. This limits their usefulness for machine learning applications, creating the need for a more informative format. This thesis presents three objectives:

Firstly, a pipeline was developed that compiles a materials dictionary for 809 CTL materials, linking them with machine-readable SMILES codes.

Secondly, an XGBoost model, a CrabNet model, and a graph neural network were trained to predict solar cell efficiency, incorporating the structural CTL information and comparing it with baseline models and models with label-encoded CTLs. The XGBoost model with structural CTL information emerged as the best, predicting power conversion efficiency with a mean absolute error of 2.45 %pt.

Thirdly, an application was created that allows users to access this XGBoost model to rank device architectures with varying CTLs.

In summary, the distinguishing features of this work are its consideration of the structure and composition of CTLs, the use of simple data inputs, and the utilisation of the Perovskite Database as a large open data set.

Keywords: perovskite solar cell, efficiency prediction, charge transport materials, graph neural networks, Perovskite Database

Word Count of Thesis Main Text: 7,776

Acknowledgements: My sincerest thanks to Pepe Márquez from FAIRmat, who conceived the idea of analysing CTLs and guided me through the process, along with my first supervisor, Lars Nagel, whom I also thank for overseeing my work.

To Jim and Jonas, who gave me input about graph neural networks, and to Dominik, my official master-theses-proofreader: Thank you for letting me bother you during your holidays!

A big, heartfelt thank you to my parents, Judith & Stephan, for all your support through 7 years of studying! For always showing interest in my studies, for all the house moves, for every call and visit, and—Judith— for the Gemeinwohlbindung. There is not enough space here to list everything you deserve credit for.

Thank you to my dearest friend, Karen, for the hundreds of hours of digital co-working. Without you, I would have felt much more alone in the process. If finishing university means seeing less of you, I would rather not graduate at all.

Finally, to my partner, Daniel: thank you for supporting my decision to study abroad and for never questioning me as a psychologist sticking my nose into photovoltaics. You never held me back, always believed in me, and provided unwavering support every step of the way.

Contents

List of Figures

List of Tables

List of Abbreviations

AI ..... Artificial intelligence

ANOVA ..... Analysis of variance

API ..... Application programming interface

DOI ..... Digital object identifier

ELN ..... Electronic lab notebook

ETL ..... Electron transport layer

GNN ..... Graph neural network

GCN ..... Graph convolutional network

GPT ..... Generative pre-trained transformer

HTL ..... Hole transport layer

InChI ..... International chemical identifier

LLM ..... Large language model

MAE ..... Mean absolute error

MSE ..... Mean squared error

ML ..... Machine learning

NOMAD ..... Novel Materials Discovery

NORTH ..... NOMAD remote tools hub

NREL ..... National Renewable Energy Laboratory

PCE ..... Power conversion efficiency

PSC ..... Perovskite solar cell

ReLU ..... Rectified linear unit

RMSE ..... Root mean square error

SMILES ..... Simplified molecular input line entry system

%pt ..... Percentage points

1 Introduction

Transitioning to renewable energy is crucial for responding to and acting against climate change. Solar energy was listed by the *Intergovernmental Panel on Climate Change* (IPCC) as having the greatest potential contribution to achieving net emission reduction in their report from 2023 (Calvin et al., 2023). Solar irradiation hits the earth with about 1000 W/m^2 in reasonably good weather conditions[1] (ASTM International, 2023). Using the *photovoltaic effect*, this radiation can be converted into electrical energy (Becquerel, 1839). Since the development of the first modern silicon solar cells in 1954, which had a power conversion efficiency (PCE) of 4–6 % (Righini and Enrichi, 2020), PCEs have constantly improved. In 2024, silicon-based cells achieved about 27 % PCE in lab conditions (Green et al., 2024), and the currently most efficient solar cell achieved a PCE of 47.6 % with a highly specialised, four-junction[2] architecture (Schygulla et al., 2022). The US American *National Renewable Energy Laboratory* (NREL) maintains a comprehensive comparison of solar cell efficiencies since 1975 (see Figure 1). In this chart, a steep increase in PCE can be observed for a type of recently invented solar cells: solar cell devices using *perovskites*[3] as their absorber layers have in record-breaking time increased their PCE from an initial 3.8 % in 2009 (Kojima et al., 2009) to 26.1 % in 2023 (as reported by NREL). A better understanding of materials used in perovskite solar cells (PSCs) could advance this promising solar-cell type and thereby contribute to a clean, carbon neutral, and cheap energy future.

1.1 Cheminformatics in Solar Cell Research

Experimental PSC research is a time-consuming, manual process. To identify useful materials and improve PSC performance, researchers typically fabricate many solar cells with various materials, layer stacks, deposition methods, and architectures to then compare their properties. In recent years, many attempts at expediting this time-consuming process using complex data analysis have been made.

[1] The *Standard Tables for Reference Solar Spectral Irradiances* indicate 1001.92 W/m^2 on sun-facing surfaces tilted by 37°, including scattering and atmospheric reflection effects (ASTM International, 2023).

[2] In multi-junction cells, multiple absorber layers are stacked, allowing absorption of a broader spectrum of sunlight.

[3] Perovskites are a class of materials defined by their distinct crystal structure, typically ABX_3.

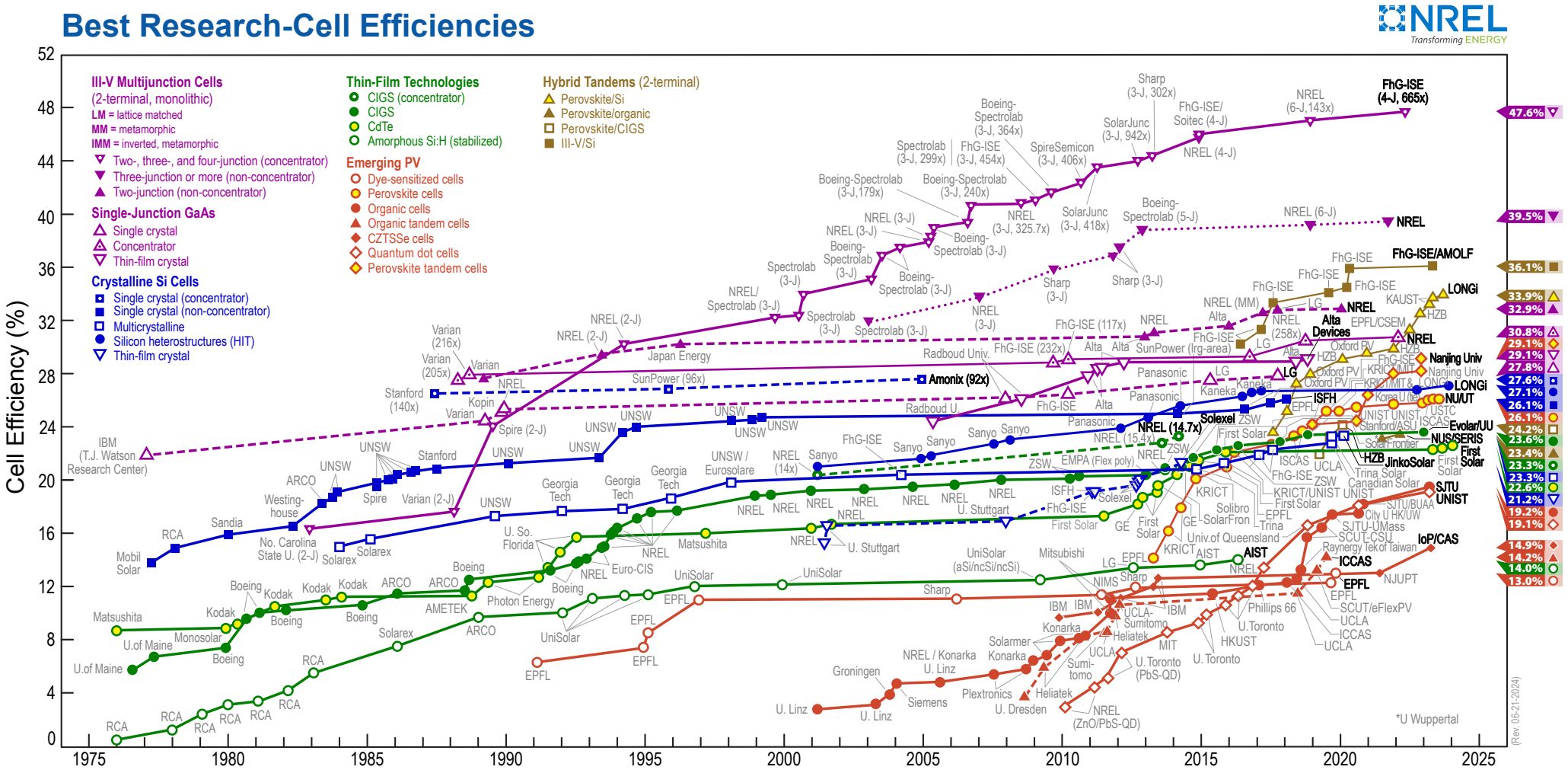


Figure 1: NREL Chart

Note the steep increase in PCE for perovskite solar cells, represented by red circles filled with yellow. Reproduced from the website of the *National Renewable Energy Laboratory (2024)*.

For example, the *Solar Cell Capacitance Simulator* (SCAPS, Niemegeers et al., 2020) is widely used to identify promising material combinations and device architectures, a *Google Scholar* search for "SCAPS solar cell" producing over 22,000 results. Moreover, machine learning (ML) approaches have gained importance in recent years, especially for property prediction and thereby selection of materials (e.g., Yao et al., 2022; Joshi et al., 2023; Padula et al., 2019; Guo et al., 2014). For example, the effects of the band gap have been investigated using neural networks (Li et al., 2019), as have been the effects of morphology using random forest classifiers, gradient boosting, and k-nearest-neighbour algorithms (Weston and Stampfl, 2018) and specific types of coatings (Yan et al., 2012). In 2017, the imperatively-titled paper "Use machine learning to find energy materials" was published in Nature (De Luna et al., 2017). Since then, the use of ML methods for photovoltaics has been tremendously increasing, providing promising, if not yet groundbreaking results (Basit et al., 2023).

1.2 Data Availability and Data Sharing Infrastructures

One central requirement for ML methods is data availability. Mavračić et al. (2021) reason that while ML may hold a prominent place in the future of material sciences, it is met with the problem of acquiring large data sets, as experimentation typically costs time and money. Fortunately, in the past 70 years of photovoltaics research, many cells have been built, measured, and recorded (Fraas, 2014). Unfortunately, however, researchers typically record their experiments locally, creating rich but inaccessible data silos. In data curation terms, these data are not *FAIR*—findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). This problem is being addressed on two pathways. Firstly, to make future data more FAIR, data sharing infrastructures are implemented for documenting experimental data. One such example is the *NOvel MAterials Discovery* (NOMAD) infrastructure (Scheidgen et al., 2023). Maintained by the research consortium *FAIRmat* (FAIRmat, 2024), the NOMAD infrastructure offers a way to store data in a standardised manner, both centrally, in the NOMAD database, and decentrally, in *NOMAD Oases* on institutional levels (FAIRmat, 2023). Secondly, to make past data FAIR, existing experimental records need to be curated and entered into such repositories. An example for this is the *Perovskite Database*, which has been created through manual data collection for 42,400 photovoltaic devices from their respective publications (Jacobsson et al., 2021), and was later integrated into NOMAD.

1.3 Research Objectives

With the Perovskite Database in NOMAD, a large collection of PSC data is openly available. This work aims to contribute to PSC research by enhancing the database and improving efficiency predictions. To do so, three objectives were addressed:

1.3.1 Automated Identification of Material Names

In the NOMAD Perovskite Database, each entry contains information on the charge transport layers (CTLs) used. However, this information consists only of the commonly used name for the CTL materials, which is neither chemically meaningful nor machine-readable and thereby of limited use for data analyses.

Objective 1 Automatically transfer the commonly used names of CTL materials into chemically meaningful and machine-readable form through a database search enhanced by publication text data processing.

1.3.2 PCE Prediction Using CTL Information

With chemically meaningful, machine-readable CTL data available, this information can then be used for predictions of device PCE.

Objective 2 Improve the prediction of PCE by including chemical information on the CTLs.

1.3.3 Construction of CTL Selection Helper Tool

Finally, the results need to be made easily accessible to researchers in order to be useful in practice. A tool that predicts PCE for different material combinations and thereby facilitates selection of CTLs may benefit PSC research.

Objective 3 Combining the results of Objectives 1 and 2, create a tool for helping researchers select appropriate CTL materials to maximise PCE.

1.3.4 Associated GitHub repository

All program code and data downloads can be found in this GitHub repository under an open-source MIT licence: https://github.com/ADWilhelm/psc-ctls-ml

2 Theoretical Background

Selected concepts and previous work that are relevant for this work will be presented in the following.[A]

2.1 Charge Transport in Solar Cells

Solar irradiation can be converted to electrical energy using the photovoltaic effect. Photons that strike the surface of a photoabsorbant semiconductor material excite electrons, creating electron-hole pairs and thus—if the charge carriers are then separated—electric current (Nelson, 2003). In recent years, perovskites have emerged as a promising photoabsorber. In addition to a relatively low fabrication cost, energy-conserving production potential and good optoelectronic properties, perovskites allow tuning of the band gap[4], making them suitable for usage in multi-junction cells. However, perovskite films often contain defects or impurities that can trap and recombine charge carriers, reducing the overall efficiency of the solar cell (Bhattarai et al., 2022). To facilitate charge carriers' movement to the respective electrode, the light-absorbing layer of PSCs is typically sandwiched between an *electron transport layer* (ETL) and a *hole transport layer* (HTL). These *charge transport layers* (CTLs) enable the effective extraction of the generated charge carriers from the absorber layers to the electrodes and mitigate recombination losses, thus playing a pivotal role in achieving high PCEs for PSCs (Foo et al., 2022; Mahmood et al., 2017). A typical PSC composition is shown in Figure 2.

More specific information on CTL materials is presented in Appendix A but is not required for understanding this thesis.

[A] These special footnotes point out where the author of this thesis made use of knowledge learned in the Data Science study programme. The theoretical background presented here relates to the *Research Methods* module, where students were taught how to perform and write a literature review.

[4] The band gap is the energy difference between the valence band and the conduction band in a material. It indicates how much energy is needed to excite an electron to a state where it can conduct electricity. This energy can be provided by an absorbed photon with the corresponding energy. Tuning the band gap modifies the spectrum of light the solar cell can absorb.

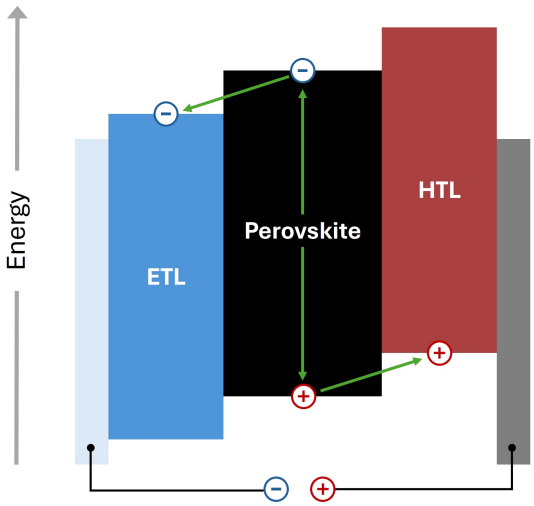


Figure 2: Typical composition of a PSC.
Left to right: transparent front electrode, ETL, perovskite absorber layer, HTL, metal back electrode. Green arrows indicate charge carrier generation and extraction. *Adapted from Ameen et al. (2018), Figure 1(B).*

2.2 ML for PSC research

While ML can serve many purposes in PSC research,[5] predicting PCE—arguably the most important device property—has been the focus of many publications. For example, Liu et al. (2022) present various models capable of predicting PCE. They constructed several models using gradient boosted regressions and random forest algorithms, achieving an RMSE of 1.58 percentage points (%pt[6]) with an ensemble approach. Their data consisted of 814 data points selected from previous publications with regard to measurement methods and materials, resulting in a highly informative data set. Another study achieved $RMSE = 1.28\,\%\text{pt}$ with an gradient boosting algorithm on data also curated from publications (Lu et al.,

[5] Chen et al. (2024) summarise them: Improving the absorber layer's properties such as band gap, stability and crystal structure as well as making predictions on device level. This enables pre-selection of novel materials, reducing the practical experimentation effort required. Also, where PCE can be predicted with explainable models, these may give insight on the relationships between materials, properties and outcome variables (Chen et al., 2023).

[6] PCE is stated as a percentage of solar irradiation. To avoid confusion, the term *percentage points*, abbreviated as %pt, is used for absolute PCE differences or aggregates of PCE.

2023). Similarly, Li et al. (2023) compiled data for 846 chalcogenide perovskite cells and achieved a mean absolute error (MAE) of 2.33 %pt at 71 % explained variance, which they stated was better than results in previous, comparable approaches. For all these examples, the data were compiled from published papers, creating high-quality but relatively small data sets (max. 2,000 devices), especially for ML purposes (Chen et al., 2023). Also, in order to have high-quality data, strict criteria for data exclusions were used, potentially limiting generalisability.

2.3 Data for PCE Predictions

An alternative approach to curating specialised datasets would be to rely on public PSC databases. The largest open collection of PSC device data was curated in the *Perovskite Database Project* (Jacobsson et al., 2021) and originally contained data from over 42,400 devices. To compile these data, the authors manually searched through over 15,000 papers published before February 2020, which presumably covers almost all research published on PSCs until that date, and documented up to 95 attributes per device. This resulting *Perovskite Database* is accessible at the project's website (www.perovskitedatabase.com), but has more recently also been integrated into the more actively maintained NOMAD infrastructure (Scheidgen et al., 2023). There, the data can be accessed via NOMAD's application programming interface (API). While the Perovskite Database Project itself drew attention within the field, only few publications have used the compiled data in the years since: A search in *Google Scholar* for the keywords "machine learning" and "perovskite database project" produced only 45 results, some of which are reviews or use only a small partition of the data. When filtering further for "PCE prediction" or "efficiency prediction", five publications remain, two of which are reviews.[7] The three machine learning papers will be presented in the following, together with another very recent[8] addition to the literature which did not show up in the Google Scholar search (yet) but which also uses the Perovskite Database data for efficiency predictions.

[7]Exact search term: "machine learning" AND "perovskite database project" AND ("PCE prediction" OR "efficiency prediction")

[8]published 05.08.2024

2.4 Related Work on Perovskite Database Project Data

First of these four, Hussain et al. (2023) predicted the PCE from the perovskite composition of 613 unique absorber layer materials within the Perovskite Database. They achieved $RMSE = 2.41\,\%$pt with a gradient boosting algorithm. Unfortunately, the ML methods are not described beyond mentioning that the Python package XGBoost was used (Chen and Guestrin, 2016). Notably, the authors also suggested using such ML results for selecting suitable ETL and HTL materials by aligning their band gap to the absorber layer.

Second, Khan et al. (2023) similarly predicted PCE from 171 unique perovskite compositions, using a *CatBoostRegressor* to achieve a mean absolute error (MAE) of 2.9 %pt and R^2 of 63 %. They implemented the resulting model as a tool accessible on the web, where the user can enter perovskite composition and thickness of the perovskite layer to receive an efficiency prediction.

Third, Hu et al. (2024) presented a PCE prediction with a subset of 16,000 cells from the Perovskite Database. They converted what would be a regression task to a classification task to predict "high PCE" cells which they defined as cells with a PCE over 18 %. Using a *voting classification* algorithm to combine the results of multiple models, they achieved an accuracy of 87.6 %. The authors discussed feature importance, pointing out the prominent relevance of the absorber layer's band gap, which by limiting the absorbable spectrum of light also limits the possibly achievable PCE[9].

Last, Liu et al. (2024) similarly predicted "high PCE" cells from absorber composition and additional device properties, setting the cut-off for high-performing devices at 17 % PCE.[10] They compared the performance of six models and built a combined voting classification model from the better-performing, tree-based models. The combined model predicted high-performing device architectures at around 81 % accuracy on the data it was trained on. However, Liu et al. (2024) only used devices with a specific absorber ($N = 3,526$) and label-encoded all categorical

[9]Solar radiation cannot be fully converted to electrical energy. The maximum possible efficiency (*Shockley-Queisser limit*) of a single-junction cell was calculated to be at 30 %, requiring a band gap of 1.1 eV (Shockley and Queisser, 1961). Despite this, higher band gaps of around 1.6 eV are more common for PSCs.

[10]The different cut-offs for high-performing PSCs (17 % and 18 %, respectively) chosen by Hu et al. (2024) and Liu et al. (2024) is not supported in either paper by any literature or profound reasoning. The difference is particularly astonishing given the fact that both papers were written by almost the exact team of authors and published only weeks apart.

variables, thereby reducing the value of their prediction for unseen materials.

2.5 Data Formats for Cheminformatics

Quantitative analyses require the data to be machine-readable and often at least at the interval level, where numeric distances have meaningful interpretations. While some chemical data naturally meet this criterion, many of them are also categorical or even in the form of molecule graphs. Computational chemistry science, or cheminformatics, has brought forth several data formats that enable detailed machine-readable descriptions of materials. For example, the *International Chemical Identifier* (InChI) and *Simplified Molecular Input Line Entry System* (SMILES) codes both encode molecular structures into strings of characters. Molecular structure data can also be represented in the *Mol* files format, where bonds, atoms and connectivity information are stored numerically.

2.5.1 Featurisation Through Descriptor Vectors

From such standardised representations, ML-ready data sets can be prepared by featurising the molecules further. One such featurisation technique is the creation of descriptor vectors that contain information on the molecule's physical (e.g., molecular weight), structural (e.g., number of rings in the molecule) and electrical (e.g., partial charges) properties. Here, the right choice of descriptors is challenging, as too many can lead to overfitting. Also, the fixed length of these vectors is both an advantage, as many algorithms require fixed input size, and a disadvantage, as it may result in sparse data sets.

2.5.2 Featurisation Through Fingerprinting

Molecular fingerprints, another featurisation technique, are more specialised in capturing structural information. In a fingerprint, the molecular structure is represented as a fixed-length binary vector. Each bit in the vector corresponds to the presence or absence of a specific circular atom environment around each atom in the molecule, determined by a defined radius. There are many types of fingerprinting techniques, prominent examples of which are *Extended Connectivity Fingerprints* and *Morgan Circular Fingerprints*.

2.5.3 Representation as Graphs

Yet another way to represent molecules are molecular graphs. In such graphs, atoms become nodes and bonds between atoms become directed or undirected edges. The nodes themselves are represented with feature vectors describing them. This relatively simple data format can be used to train Graph Neural Networks (GNNs, see respective section below).

2.6 Cheminformatics Toolboxes

Many tools can assist the digital processing of chemical information. Python libraries such as *RDKit* (Landrum et al., 2024) contribute functions to transform molecules' SMILES codes into Mol files, create visual representations, or generate Morgan fingerprints. Additionally, there are toolboxes for processing textual chemistry data. For example, ChemDataExtractor (Swain and Cole, 2016; Mavračić et al., 2021) was already capable of processing paper texts in search of chemical information before the age of large language models (LLMs) with attention-based transformers. Since the introduction of the latter, many efforts have been made in training generative pre-trained transformers (GPTs) to assist with chemistry-specific tasks. A prominent example is *ChemCrow* (M. Bran et al., 2024), a GPT equipped with chemistry-specific tools that allow it to create more accurate responses and reduce hallucinations[11]. However, even general purpose GPTs can help with basic to advanced questions (Polak and Morgan, 2024; Polak et al., 2024).

2.7 ML Methods for Material Property Prediction

While ML publications within the field of material research typically compare different approaches, no single best has emerged, which can likely be attributed to each task requiring different strengths. Here, a selection of very different models will be presented. The models were employed in this work to make Objective 2: PCE Prediction Using CTL Information.

[11] Hallucinations are nonsensical answers or answers that clearly contradict information available to the LLM (Farquhar et al., 2024).

2.7.1 Gradient Boosting Regression[B]

As a conventional benchmark, gradient boosting algorithms have proven to be successful in predicting material properties for perovskites (e.g., Guo and Lin, 2021; Rath et al., 2022). Zhao et al. (2022) compared twelve methods for property prediction in perovskite materials, among others support vector machines, k-nearest neighbours and extreme gradient boosting. The latter, implemented using the Python package *XGBoost* (Chen and Guestrin, 2016), outperformed the other methods. Extreme gradient boosting uses a sequence of decision trees, with each tree correcting the errors of the previous tree. This method can be used for both classification and regression tasks, making it suitable for property prediction. However, a requirement of gradient boosting is that the input data is numerical, meaning that molecule data needs to be featurised (see section Data Formats for Cheminformatics.

2.7.2 CrabNet

Neural networks with transformer architectures cannot only be used for generative purposes, but also for property predictions. The *Compositionally-Restricted Attention-Based Network* (CrabNet) is capable of making property predictions for materials using their chemical formula (Wang et al., 2021). It represents the atoms contained within a formula by their property vectors, aggregating these according to the relative amount of each atom within the formula. These representations are processed using self-attention, meaning that each atom will be influenced by the atoms around it. Further processing with fully connected layers results in the property prediction output. Wang et al. (2021) show CrabNet to outperform a standard *Random Forest* model and, more importantly, the chemistry-specifically trained models *Roost* (Goodall and Lee, 2020) and *ElemNet* (Jha et al., 2018).

2.7.3 Graph Neural Networks[C]

In recent years, graph neural networks (GNNs) have become popular for molecule processing (Reiser et al., 2022). They are particularly suitable for ML in chemistry

[B] In the the *Data Mining* module, Random Forest models like the extreme gradient boosting algorithm used by XGBoost were explained, along with other ML methods mentioned here.

[C] In the module *AI and Applied Machine Learning*, students learned how neural networks function. This helped a lot with understanding GNNs.

or material sciences because molecules can naturally be represented as graphs (see Representation as Graphs). In GNNs, the graph is typically processed by *Message Passing*, where information from one node is passed to its neighbours (Veličković, 2022). The deeper a GNN, the further the information can travel. However, with increasing depth, not only general neural network issues such as overfitting may arise, but also GNN specific problems such as over-smoothing or over-squashing (Ud Din and Qureshi, 2024). Consequently, a balance needs to be struck, keeping the GNN as simple as possible while allowing sufficient complexity to cover the complexity within the data. Overall, GNNs appear to be very promising for materials research, in several instances outperforming other ML methods in property prediction tasks (e.g., Schütt et al., 2018; Gasteiger et al., 2020; Fung et al., 2021).

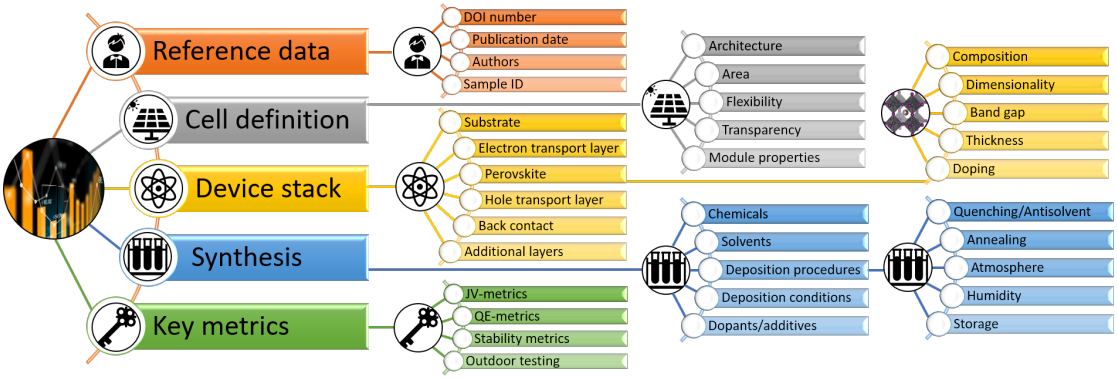


Figure 3: Data structure of solar cell entries within the Perovskite Database. *Taken from Jacobsson et al. (2021), Figure 2.*

3 Objective 1: Automated Identification of Material Names

In its current state within the NOMAD repository, the Perovskite Database contains data on over 43,000 solar cells, covering most PSCs published before 2021 (Jacobsson et al., 2021). The data are publicly accessible under a Creative Commons BY 4.0 licence[D]. The structure of the database is depicted in Figure 3. The present work focuses on the device stack (yellow in Figure 3), i.e., the stack of materials layered sequentially to form the solar cell. While for the perovskite layer, the composition is recorded in the database, the ETL and HTL are only recorded by their commonly used names. Sometimes this is their chemical name (e.g., TiO_2) but often it is an abbreviation or the name under which the material is most commonly sold (e.g., Spiro-MeOTAD, 2PACz). The informational content of these names is limited to being distinguishable. A chemically meaningful representation could be highly beneficial to include more fine-grained information into ML models.

Finding chemically meaningful representations for the CTL materials could be done manually, using chemistry databases and domain knowledge. However, this would be a tedious process as 2,450 unique CTL material names can be found

[D]The meaning of copyright and licences was explained in the module *Data Governance and Ethics*. This licence allows all re-use of the data with the requirement that the source be indicated.

in the Perovskite Database. Also, in the future, more materials may be added. Therefore, an automated solution is needed for associating the common material names with a chemically meaningful identifier such as a SMILES code. This need is addressed by Objective 1.

3.1 Method

To describe the CTL materials, a dictionary is needed that associated each material with its SMILES code. The data used for this were downloaded via the NOMAD API, resulting in a total data set of 43,108 solar cells.[12]

3.1.1 Data Description and Preparation

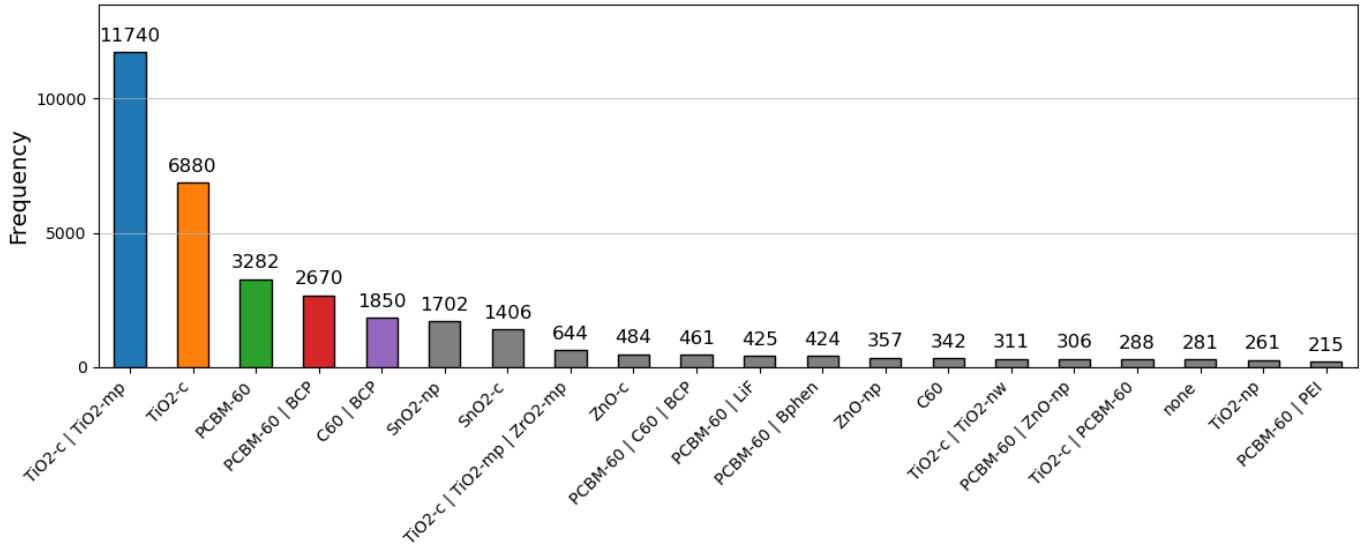
For Objective 1, the relevant features were the ETLs and HTLs consisting of the CTL material name(s), as well as the Digital Object Identifiers (DOIs) of the original publication for the cell. Many ETLs and some HTLs consist of multiple materials, effectively making them sub-stacks of the device stack. Also, the distribution of materials is extremely skewed with very few stacks accounting for a vast majority of cells (see Figure 4).[13]

3.1.2 Pre-processing

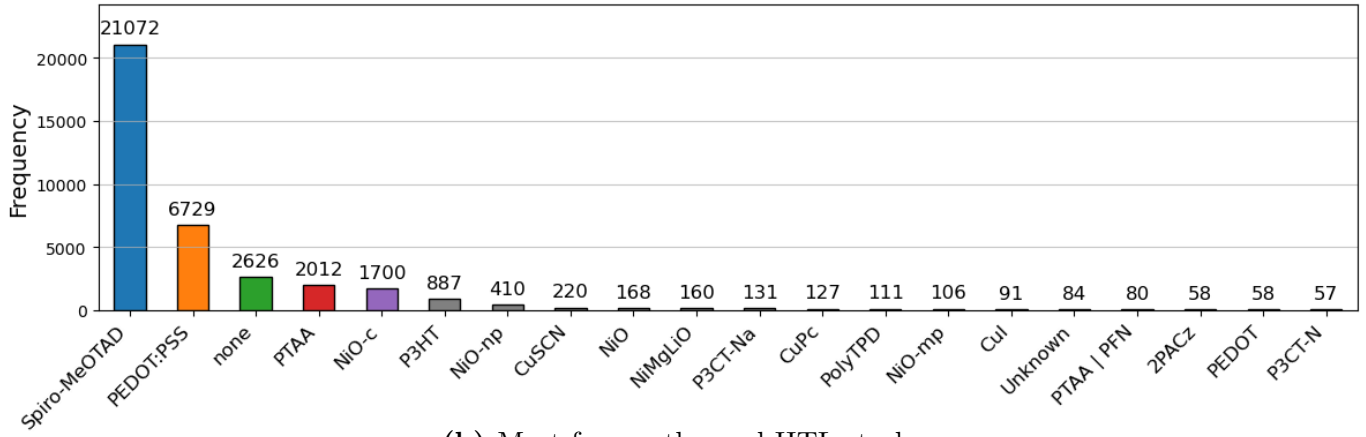
Minimal data cleaning was applied. CTL stacks were transformed into a uniform format, as some multiple-materials stacks had been separated by semicolons while others were separated by commas. Furthermore, the stacks were then broken up into their component materials, which were listed along with the references for each publication that used the respective material. This resulted in a list of 2,450 unique CTL materials.

[12]The difference between this number and the original 42,400 entries curated in the Perovskite Database Project is due to the option for researchers to upload further data. However, this option was hardly used, so most data is from before 2021.

[13]For data description regarding PCE, see Data Description and Preparation for Objective 2.



(a) Most frequently used ETL stacks



(b) Most frequently used HTL stacks

Figure 4: Frequencies of CTL stacks.

Note that of the 1,463 unique ETL and 1,973 unique HTL stacks (not materials, hence the sum exceeds the number of 2,450 materials) only the 20 most common ones are shown here each. As can be seen, some few stacks are used across a vastly disproportionate amount of solar cells.

3.1.3 Material Identification Pipeline

To associate the CTL material names with a SMILES code, an algorithm was designed that sequences calls to the PubChem database to search for the material with calls to an LLM to extract information on the materials' chemical names from the associated publications. The text of the latter was accessed via the *Elsevier* API. The result of this pipeline was a dictionary of material names to PubChem *Compound Identifiers* (CID) that unambiguously point to a PubChem entry from which the SMILES code can eventually be retrieved. See Figure 5 for an overview of the identification pipeline.

Initially, the material name is used to search for an entry within the PubChem Compounds and Substance libraries (blue in Figure 5). When no entry can be found, a search is initiated within the text of the reference publications for cells using that material. First, a list of all papers in which the material has been used is compiled. Then, their DOIs are used to programmatically query the Elsevier's ScienceDirect database. Once a paper text can be retrieved, it is next given to the LLM *Llama 3 70b*. The LLM was accessed through the API of the AI infrastructure company *Groq*. Its *system prompt* (which influences all answers and general answer tendencies) was set to: "You are a solar cell scientist proficient in reading papers. You output only the chemical name of the compound asked for, nothing else." The *user prompt* then given for each paper was: "What is the chemical name pertaining to this abbreviation: *"common material name"*? You can find it in this text:" followed by the text as retrieved from the Elsevier API. After some prompt engineering, the LLM would successfully extract a material name if it was given within the text. This name was then used to again query the PubChem database, continuing with the next paper if no entry could be found. If there was no success after searching and LLM-reading all papers in which the material was used, the search was marked unsuccessful.

The result was a dictionary in which each material name that had been successfully identified was associated with a PubChem CID. With this CID, the SMILES codes could be retrieved through the PubChem API.

3.2 Results

The materials identification pipeline was able to identify 809 (33 %) of the 2,450 unique CTL materials. See Figure 6 for a breakdown of the pipeline's results.

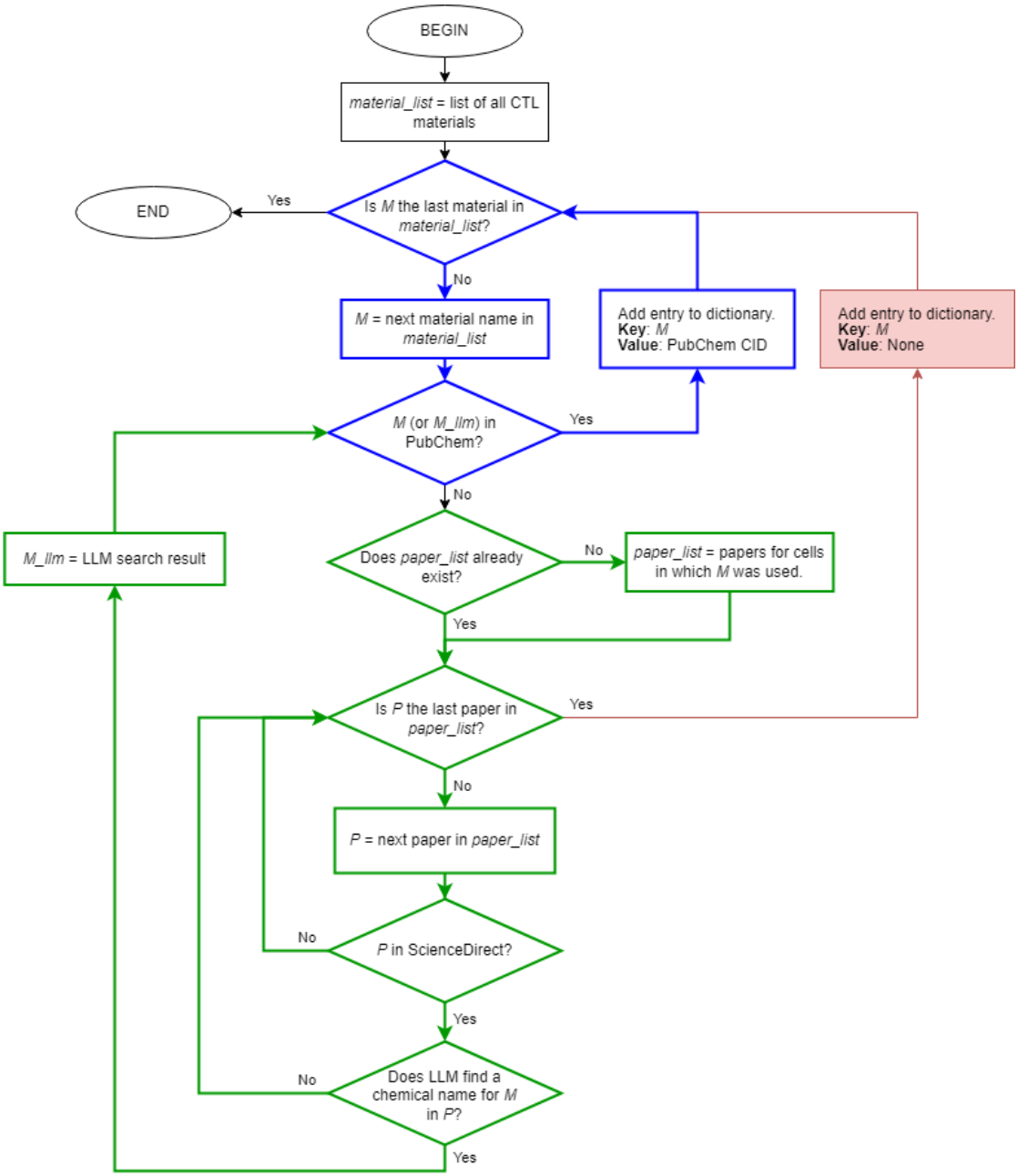


Figure 5: Overview of the identification pipeline.
Blue parts show the loop around calls to the PubChem API. The LLM loop (green) was entered once the initial PubChem call was unsuccessful.

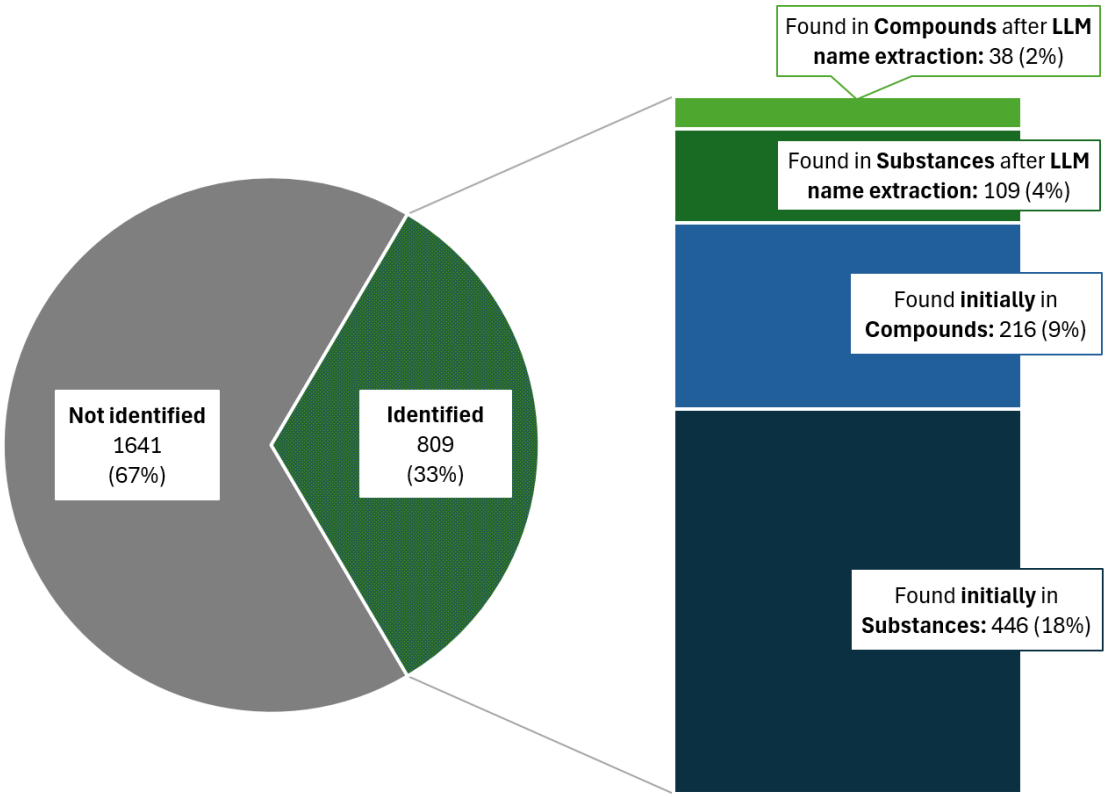


Figure 6: Results of the identification pipeline
Note that *Compounds* and *Substances* mean their respective PubChem library.

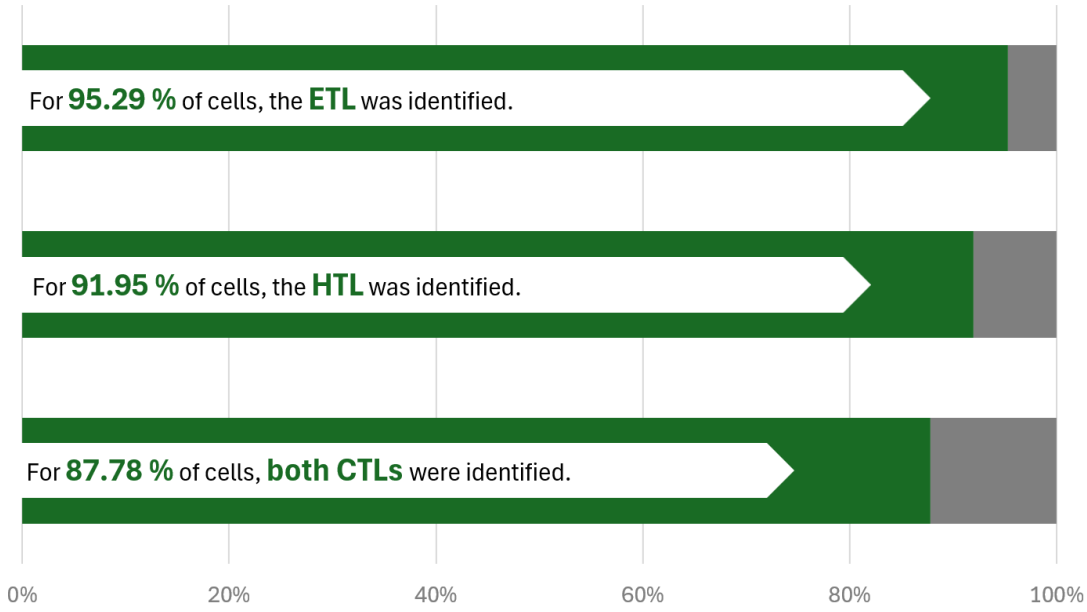


Figure 7: Transfer of identified materials to described PSCs
For 41,090 cells, the ETL stack could be fully associated with SMILES codes. For 39,647 cells, the HTL stack could be fully described. In total, for 37,849 cells, both the ETL and the HTL materials were associated with SMILES codes.

Due to the skewed distribution of the material usage, the identification of 33 % of materials allows the full CTL description of a much larger proportion of the PSCs in the database. When no CTL was used in a cell, this naturally also means that the CTL can be sufficiently described. Altogether, the dictionary allows the description of the vast majority of PSC devices within the NOMAD database (see Figure 7).

3.2.1 Evaluation by PSC Expert

To evaluate the quality of the results, a PSC researcher from the *Karlsruhe Institute of Technology* was approached. They manually identified 30 CTL materials selected at random from the ones identified automatically. The results of manual identification corresponded with the automated identification in 21 (70 %) cases. The expert noted that the mis-identifications happened primarily due to ambiguously named materials. Manual identification took about five minutes per material.

3.3 Discussion

Objective 1, transferring the commonly used names of CTL materials into chemically interpretable and machine-readable form, was achieved through a pipeline of API calls to the PubChem materials database, additionally extracting information from publications using an LLM. With the resulting dictionary, the CTLs of 87.78 % of the solar cells registered in the NOMAD database could be fully described. The results were evaluated by a PSC expert to be moderately accurate.

3.3.1 Limitations

The identification pipeline only producing correct results in 70 % of cases crucially limits the validity of the results, in particular because errors can only be detected manually. Importantly though, while this accuracy value may at face value appear disappointingly low, it must be noted that 100 % accuracy is likely unachievable due to ambiguous or inadequate naming of materials in the data. Besides accuracy, identification yield is a limiting factor: While 87 % are a considerable portion of the PSCs, only 33 % of the materials used could be identified. A higher identification yield could increase the diversity of materials. To increase both accuracy and yield, several further improvements could be considered:

Materials databases Firstly, in addition or alternatively to PubChem, other materials databases could be used. PubChem was chosen due to its open accessibility and programmatic access, its sheer size as well as its popularity within the field.[14] Other databases could be used in conjunction with PubChem to possibly identify more materials, for example the ChemSpider database (Royal Society of Chemistry, 2024) or a database specific to photovoltaics.[15]

Paper retrieval strategies and LLM extraction The LLM's performance depends on the availability of paper texts. In the LLM section of the current pipeline (green in Figure 5), paper texts were accessed through the API of ScienceDirect, a literature database maintained by Elsevier. While ScienceDirect is a sensible choice, containing 21 million articles, 3.3 million of which are openly

[14] A Google Scholar search of "PubChem" yielded over 110 000 results compared to 19,000 for "ChemSpider" or 5,000 for "Reaxys", to name a few other general chemistry databases.

[15] For a list of examples for the latter, see Shang et al. (2024), Table 1.

accessible, additionally searching within other collections warrants finding more articles. In addition to identifying more materials it would reduce the effect that more common materials are more likely to be identified by the LLM due to more papers being available.

Possible additional paper retrieval strategies could be including other scientific collections, such as *arXiv* or the *Royal Society of Chemistry*'s collection. Both of these do offer some level of programmatic access, though they are less accessible than ScienceDirect. Another alternative could be web scraping tools such as *paperscraper*, though this should be done carefully as it may potentially violate publishers' rights.

Moreover, the LLM extraction might be improved by testing alternative GPTs, both chemistry-specific and general purpose. Additionally, results from multiple rounds of paper reading could be combined, for example using a voting algorithm, to reduce the amount of wrong identifications.

Data limitations In addition to methodological considerations, the quality of the data within the NOMAD database is imperfect in many cases. Importantly, materials appear in different forms (e.g., "TiO2-c" and "TiO2-mp", where one is crystalline and the other mesoporous TiO_2). Also, artefacts from manual compilation can be observed. For example, the CTL name "PBCM-60" appears as an unidentified material, which is very likely to be a misspelling of PCBM-60. In some cases, the original paper may lack information, leading the person processing it to note a nondescript name.[16] Moreover, CTL names often contain potentially ambiguous acronyms. Another limitation for the LLM identification is the fact that oftentimes, materials are often more closely described in the supporting information of publications than they are in the main text, meaning they will not be contained within the API accessible main paper text. Addressing these limitations can mainly be done by discovering additional data sources, both for PSC data as well as for materials and publications.

3.3.2 Further research

Besides addressing these limitations separately, further research could attempt a more combined approach: Only recently, the chemistry-aware LLM *ChemCrow*

[16] Suspects are, e.g., "Polymer1", "Polymer2" etc.

has been presented. It can access chemistry-specific tools that allow it to give high-quality responses for chemistry-related questions and tasks (M. Bran et al., 2024). These tools comprise web searching, accessing Web of Science publications and Wikipedia, and transforming chemical names into SMILES codes, potentially making ChemCrow highly suitable for a task that involves processing common material names. Unfortunately, the model is built using *OpenAI*'s *GPT-4*, which currently means that payment per token is required to use it.

3.3.3 Outlook

The identification pipeline presented here has twofold value for PSC research:

Contribution to FAIR data in material science Firstly, the pipeline itself may be useful for further data curation. The approach can easily be transferred to other material applications where common names need to be identified. For the NOMAD database specifically, it could be used to curate further cells in the future or to simplify the process by which researchers enter their experimental data into the database or into NOMAD Oases, FAIRmat's electronic lab notebook (ELN) software. The pipeline could pre-process data which could then be reviewed by the researcher (mitigating the impact of the limited accuracy). This would be particularly useful as ELNs may be tedious to use, discouraging researchers to sacrifice valuable time. In the bigger picture, improving ELNs can contribute to making more data within material science FAIR.

Using the enhanced Perovskite Database data for ML Secondly, the identification pipeline's results can be used for data analysis. Including information on CTLs may benefit property predictions or PSC architecture design efforts. In the following section, an exemplary use is presented with PCE predictions for PSC devices using the newly curated SMILES codes for the CTL materials.

4 Objective 2: PCE Prediction Using CTL Information

As explained in the theory section on PCE prediction, training ML models to be able to predict PCE can reduce the need for manual experimentation when optimising device properties. Previous attempts at PCE prediction using the Perovskite Database were focused on the absorber layer. With the results from Objective 1, it is now possible to include CTL property data into predictions. To demonstrate this possibility and investigate the influence of including CTL information into property prediction, Objective 2 is to improve the prediction of PCE by including chemical information on the CTLs.

4.1 Method

For PCE prediction, the same data source as for Objective 1 was used, now curating more features and applying stronger exclusion criteria than before to ensure a coherent data set for the ML efforts.

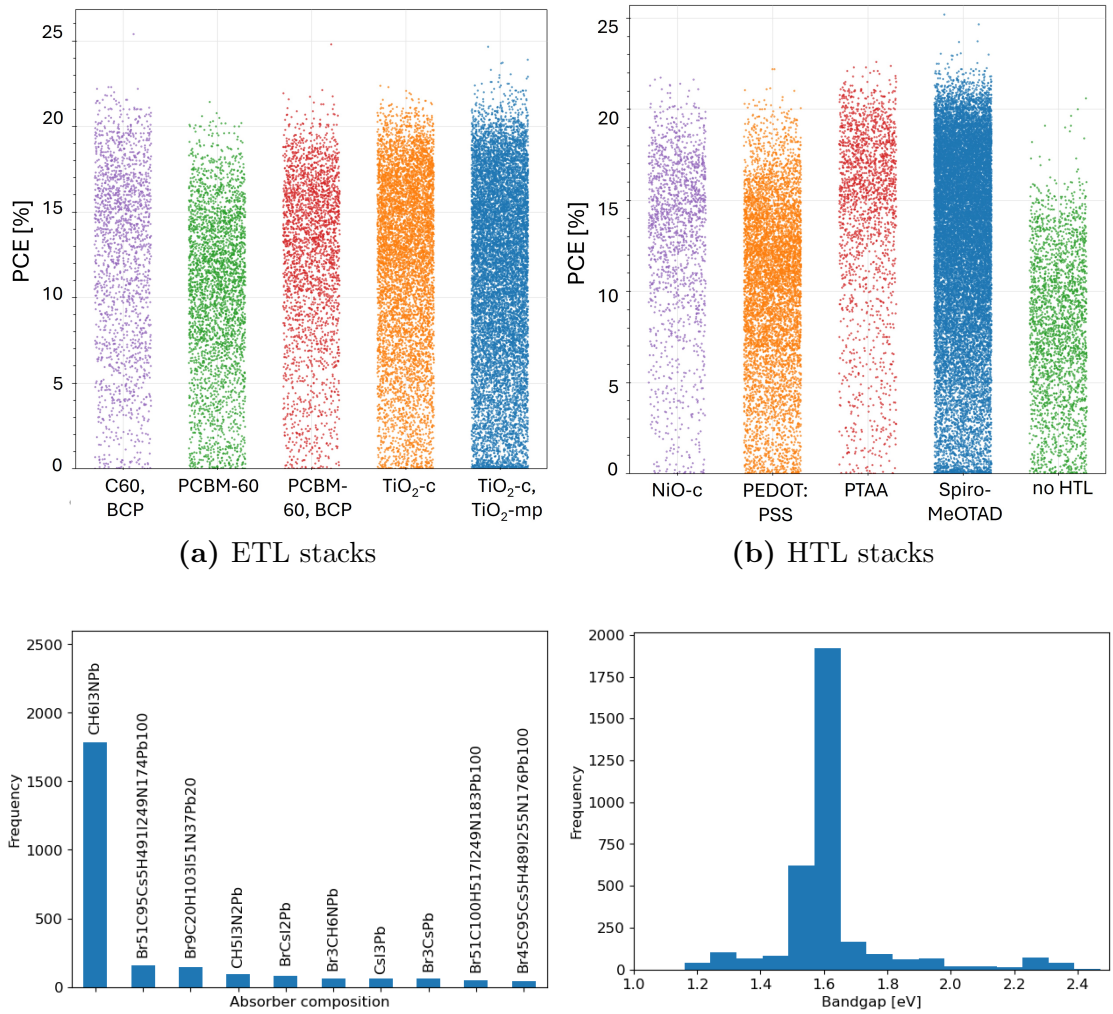
4.1.1 Data description and preparation

Again, 43,108 entries were retrieved from the Perovskite Database in NOMAD. The chemical composition of the absorber layer, its band gap and the CTL names were used for the PCE prediction. Table 1 shows exemplary instances. The prediction target, PCE, ranged between 0 and 25 %.[17]

See Figure 8, a and b, for a PCE depiction by CTLs used and Appendix B for a depiction of PCE development over time.

Data Set Filtering Data cleaning was performed in several steps: First, 5,259 entries for which the ETL and HTL stacks could not be fully transformed into the machine-readable SMILES format with the results from Objective 1 were eliminated. This includes entries with missing values but not solar cells that explicitly use no ETL or HTL. Another 552 entries were eliminated due to missing absorber composition information. Then, 419 solar cells measured at illumination intensities

[17]Except for 25 outliers with unrealistically high PCE values which are possibly incorrectly registered multi-junction cells.



(a) ETL stacks **(b)** HTL stacks

(c) Absorber compositions **(d)** Band gap distribution

Figure 8: Distributions of the predictor features
a) and b) depict cells that used the five most common ETL and HTL stacks (c.f., Figure 4). c) depicts the elemental compositions of the most commonly used absorber layers. One third (1,801) of the 5,468 aggregated entries used methylammonium lead tri-iodide ($MAPbI_3$, Chemical composition: CH_6I_3NPb) as their absorber layer. 1,478 other materials were used as absorber layers. d) depicts the distribution of absorbers' band gaps.

Absorber formula	Band gap	ETL stack	HTL stack	PCE
CH6I3NPb	1.6	SnO2-nanosheets, C60	Spiro-MeOTAD	18.3
BiCs100I300Pb99	0	TiO2-c	CuI	7.23
CH6I3NPb	1.59	PCBM-60, BCP	PolyTPD	13.3
...	...	...	...	...

Table 1: Example entries from the data set

other than $1.000\,\mathrm{W/m^2}$ were excluded.[18] Another 2,677 entries were eliminated that had a device area greater than $24\,\mathrm{mm^2}$.[19] Lastly, 721 entries without PCE were dropped as well as 1,748 solar cells that had $PCE < 2\,\%$.[20] After these exclusions, 31,635 entries remained. These were condensed down to 5,468 by removing duplicate independent variable combinations and averaging their PCE.

Features The absorber composition was given as a chemical formula indicating the elements and their amount within the material. The band gap variable contained many missing values (2,039; 37.4 %), so these cases were not excluded but instead set to zero as it would have meant a grave loss of data. The ETL and HTL stacks were transformed into SMILES codes using the results of Objective 1. For example, the SMILES code for the frequently used ETL material titanium dioxide (TiO_2) is "O=[Ti]=O" and for the frequently used HTL *Spiro-MeOTAD* it is "C OC1=CC=C(C=C1)N(C2=CC=C(C=C2)OC)C3=CC4=C(C=C3)C5=C(C46C 7=C(C=CC(=C7)N(C8=CC=C(C=C8)OC)C9=CC=C(C=C9)OC)C1=C6C=C

[18] These standard test conditions of $1.000\,\mathrm{W/m^2}$ intensity are equivalent to the sun's irradiance hitting the earth's surface in good weather conditions as explained in the introduction.

[19] Such "large-area" devices are relevant for upscaling and commercialising the perovskite technology, but are less relevant for material exploration.

[20] This last exclusion was done for several reasons: It cannot be ruled out that in the compilation of the Perovskite Database, researchers entered "0" as PCE when they could not find the information. Also, very low-performing devices may be due to highly unconventional methods being employed, as even the very first perovskite solar cells performed better than 2 % (Kojima et al., 2009).

(C=C1)N(C1=CC=C(C=C1)OC)C1=CC=C(C=C1)OC)C=C(C=C5)N(C1=C C=C(C=C1)OC)C1=CC=C(C=C1)OC". Where multiple materials were used in a CTL stack, they were concatenated into one string with a period between the SMILES codes.

4.1.2 Machine Learning

The PCE prediction was done using three different approaches: extreme gradient boosting using *XGBoost* (version 2.1.1), CrabNet (version 2.0.8), and a GNN constructed with *Pytorch Geometric* (version 2.5.3) (Fey and Lenssen, 2019). All three have been presented in the theory section. Each model was trained on 80 % of the data, using an additional 10 % for validation, and the remaining 10 % for testing. The splitting was done at random, but the same partitioning was used for all models. Each model type was trained in three configurations:

1. a baseline model with only the absorber layer information (material and band gap),
2. a version with the CTL information passed in label-encoded form,
3. a version including the full CTL information (as fingerprints or graphs).

This enables an evaluation of effectiveness of including CTL information into PCE prediction and the effect of having this additional information in chemically meaningful form instead of the simple label-encoded form.

All training and evaluation computations were performed on the *NOMAD Remote Tools Hub* (NORTH) platform using *Jupyter Notebooks* running Python version 3.11. The underlying infrastructure consisted of a *Linux*-based server (*Debian* 5.10.0-23 kernel) equipped with 504 GB RAM. The *Pytorch* version used as basis for some models and for utilities was 1.12.0+cu102.

XGBoost For the XGBoost modelling, the absorber layer was transformed into numerical format by introducing a column for each possible element into the data set representing the amount of that element in the absorber layer (e.g., "TiO2" would result in a 1 in the Titanium column and a 2 in the Oxygen column). The ETLs and HTLs, for which structural information was available in the form of the SMILES codes, were transformed into Morgan Fingerprints with a bit size of

1024 and a radius of 2 using RDKit.[21] Three models were configured as explained above. They were trained with the objective function set to minimise the squared error and validated using 10-fold cross-validation across three repetitions to ensure robustness and mitigate overfitting. Then, each was trained again on the full training set. Finally, the models were further optimised using bayesian hyperparameter optimisation implemented using the *hyperopt* package. (Bergstra et al., 2013).[22]

CrabNet Unlike with XGBoost, the absorber layer formula did not require featurisation to be passed to CrabNet, as its attention-based transformer architecture is specifically designed for processing chemical formulae. To enhance prediction, CrabNet can be given additional features ("extend_features") which are processed by a hyperoptimised XGBoost regressor, similar to the above model. Label encoding and Morgan Fingerprint generation were therefore done like for the XGBoost models and then passed to CrabNet as additional features. Again, three model configurations were trained.

Graph Neural Network For the GNN, the absorber formulae were featurised by atom occurrences, like for the XGBoost model. Each CTL material was represented by a graph, represented by a nodes tensor and an edges tensor. The nodes represent the elements within the molecule. They were described by a nine-dimensional property tensor for each element.[23] The edges were characterised by the two nodes they connected and they were undirected, meaning neither node was specified as start or end node. One such molecule graph was constructed for each ETL stack and one for each HTL stack. The variable graph size resulted in the graphs not being able to be stacked into batches. Therefore, the data was given to the GNN one by one, updating the model parameters after each solar cell.

[21] The radius is the circular area around each atom in a molecule that is considered when generating the fingerprint. A radius of 2 is a common choice that suffices for capturing most molecules adequately.

[22] The parameters optimised were the number of features selected for each tree ("colsample_bytrees"), the minimum loss reduction required to make a further partition on a leaf node ("gamma"), the learning rate, the maximum depth of each tree, the number of trees built ("n_estimators") and the fraction of solar cells randomly selected for building each tree ("subsample").

[23] The properties describing the nodes were: atomic number, mass, explicit valence, total valence, formal charge, hybridisation, number of radical electrons, whether the atom is part of a ring, whether the atom is part of an aromatic ring.

Model architecture[E] The GNN was built using the *MessagePassing* class[24] from Pytorch Geometric. The ETL and HTL graphs are each processed by two sequential *Graph Convolutional Network* (GCN) layers ("GCNconv"). The first GCN layer projects the 9-dimensional node feature vectors into a 32-dimensional embedding space.[25] Then, a *Rectified Linear Unit* (ReLU) activation function[26] is applied in between the two convolutional layers before the second GCN layer further refines the embeddings within the same 32-dimensional space. The output from the GCN layers is then aggregated across all nodes using global mean pooling, which condenses the graph information into a single vector representation for both ETL and HTL. This is then concatenated with the feature vectors for the absorber material and the band gap. The concatenated vector is fed into a series of three fully connected layers with 64 dimensions, each followed by a Leaky ReLU activation function.[27] Finally, a linear regression layer outputs a single continuous value for the predicted PCE of the solar cell.

Over 30 different model architectures were trained and compared, with varying embedding dimensions for the CGN layers, varying numbers of regression layers, different activation functions and layer architectures including more, fewer or other orders of the layers. The architecture described here belongs to the best performing model.

Training Training parameters were equally varied to optimise performance. The champion model's parameters are described here. This model was optimised using the *Adam optimiser* with an initial learning rate of 0.00005 and weight decay of 0.0004. The weight decay term helps prevent overfitting by penalising large weights. The learning rate was dynamically adjusted using a *OneCycleLR*

[E] The task of building the GNN required a deep dive into neural network architecture and would not have been possible without the knowledge gained from the tutorials and coursework in the module *AI and Applied Machine Learning*, where a convolutional neural network was built from scratch.

[24] Message passing is the basic mechanism for processing graph information by which each node passes information on its own features to its neighbouring nodes.

[25] The embedding is the representation by its parameters within the neural net.

[26] The ReLU activation function outputs the input directly if it is positive, and zero otherwise. Thereby, it introduces non-linearity and allows the model to capture more complex relationships.

[27] Leaky ReLU allows a small, non-zero gradient for negative input values instead of setting them to zero as regular ReLU does. This reduces the risk of neurons ceasing updating their weights effectively during training.

Model	XGBoost		CrabNet		GNN	
	MAE	R^2	MAE	R^2	MAE	R^2
Baseline (no CTLs)	2.75	48.6 %	2.89	39.2 %	2.93	42.2 %
Label-encoded CTLs	2.49	56.8 %	2.92	38.2 %	2.84	39.8 %
Featurised CTLs	2.45	57.8 %	2.90	39.1 %	2.50	56.6 %

Table 2: Model comparison for the PCE predictions
Note that MAE is indicated in percentage points.

scheduler, which allows the learning rate to increase to a peak (set at 0.0001) before gradually decreasing. The scheduler was configured to reach the peak after 100 of the 1,000 epochs that it trained for, decreasing to a final level of 0.0000007. The model was trained to minimise the mean squared error (MSE) loss. After the tests converged on a best performing model, its architecture was tweaked for training two additional models to mimic the comparisons made with XGBoost and CrabNet: a baseline without CTL information and a model using label-encoded CTL information, in addition to the full graph model.

4.2 Results

After training the nine models, predictions were made on the unseen test data. The mean absolute error (MAE) and the explained variance R^2 are depicted in Table 2.

For the XGBoost models and the GNN, both the MAE scores and the explained variance improve from the baseline model to the full model. The CrabNet's performance remains similar across all configurations. On all configuration levels, the XGBoost models appear best, exhibiting the smallest MAE and highest explained variance on each configuration level. These differences in prediction quality were tested with inferential statistics. However, the results were inconclusive, especially considering limitations arising from distributional assumptions. Due to limited space, any explanation given here would need to be inappropriately simplified. Therefore, no statements on significance will be made or interpreted here. However, it is highly recommended to refer to Appendix C for a detailed account

of the statistical evaluation of model performance.[F]

Besides the comparison of MAE and explained variance, a difference in prediction quality becomes apparent when inspecting how well the predictions reflect the true values: the scatter plots in Figure 9 show that some models predict overly many cells to have a PCE around 12 %, ignoring the original distribution which has a far less prominent peak around 12 %.

This 12 % peak is very apparent in all predictions, though it is less prominent in the two non-baseline XGBoost models. Another, smaller local maximum or at least plateau around 15-16 % can be observed in all models. For the full XGBoost model, it is almost equal to the 12 % peak. Overall, the shape of the true values' distribution is best matched by the full XGBoost and full GNN models. Overall, the XGBoost model with CTLs fingerprints emerges as the best model: it combines the smallest MAE and most variance explained and its predictions reflect the distribution of the true values reasonably well. In addition, its computational cost for training was at least 50 times lower than for the GNNs and at least 10 times lower than for the CrabNet models.

4.3 Discussion

Objective 2, improving the prediction of PCE by including chemical information on the CTLs, was addressed by training various ML models. Including CTL information compared to using only a baseline model resulted in a reduced MAE, more explained variance, and better prediction distribution in two out of three cases (XGBoost and GNN, but not CrabNet). Over all criteria, the XGBoost model that uses CTL information in the form of Morgan Fingerprints emerged as the best model with $MAE = 2.45\,\%\text{pt}$ and $R^2 = 57.8\,\%$.

4.3.1 Limitations

Some methodological limitations need to be considered regarding the results.

Data limitations Predictions can only be as good as the data allows. While the data set used was large and data quality overall was good, there were still some limiting factors: Firstly, for many solar cells (33 %), the absorber's band

[F] Here, the module *Statistical Methods and Data Analysis* had provided in introduction. Additionally, transfer from experience in statistics for a psychology M.Sc. was possible.

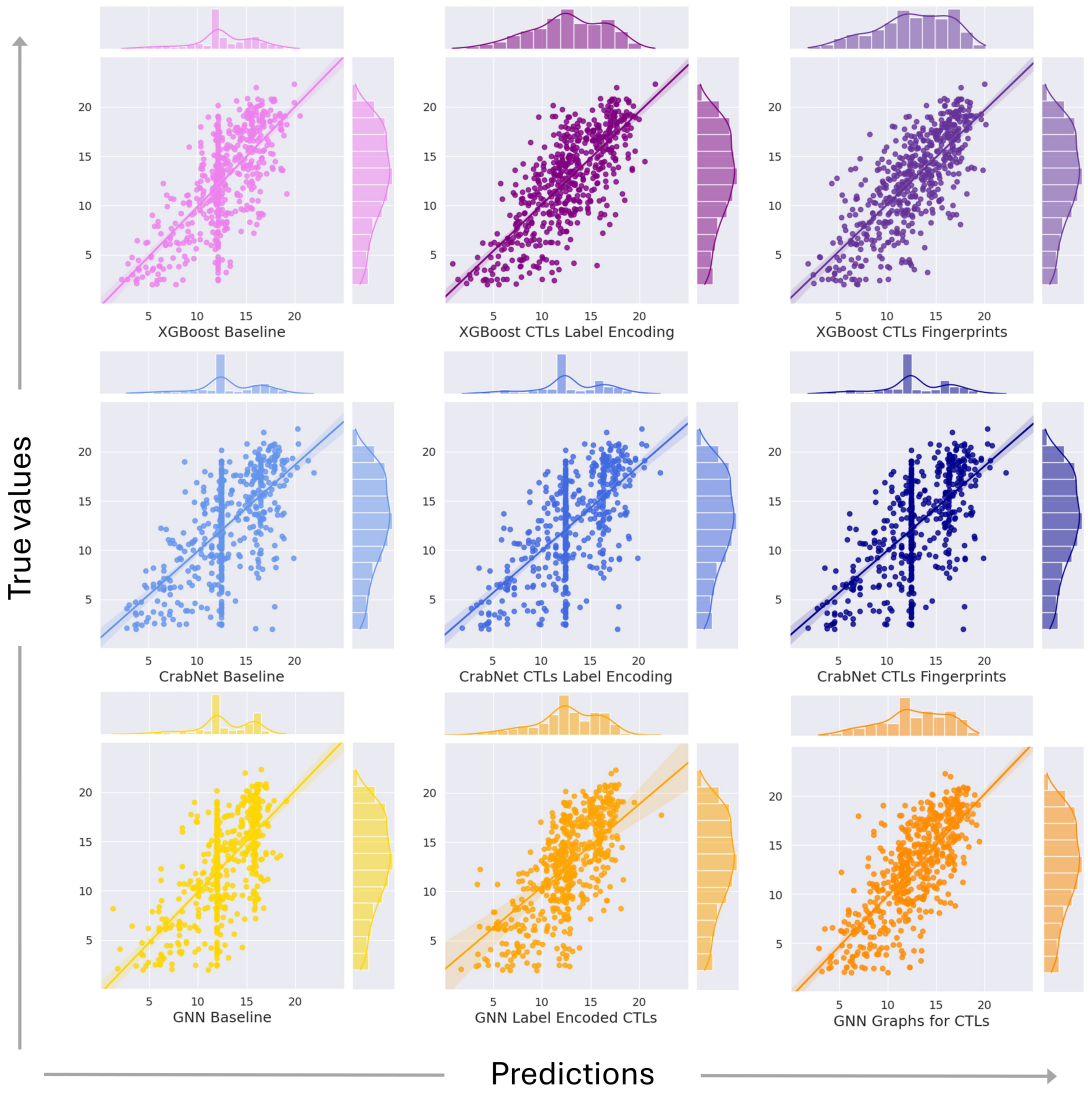


Figure 9: Distributions of predictions and true values by model type and configuration

gap was not available. Secondly, with an identification accuracy of 70 % (see Evaluation by PSC Expert), the results from Objective 1 likely resulted in wrongly identified materials that become noise in the prediction. Thirdly, the absorber materials were only described by their chemical formula, which does not contain structural information. Future research could apply the material identification pipeline presented here to the absorber layers, enhancing the absorber information and potentially improving the predictions.

Featurisation Another limitation may have been caused by the selection of featurisation techniques. The transformation of the material information into atom counts or fingerprints adds features to the data set, potentially invoking the "Curse of Dimensionality" (Altman and Krzywinski, 2018), which means that when including too many variables in a model, prediction quality will deteriorate. Using descriptor vectors from molecule's properties instead of fingerprints would reduce dimensionality, but also poses the challenge of selecting appropriate chemical descriptors. Lastly, a limitation may arise from the way CTL stacks with multiple materials were featurised. These were treated as one material, concatenating the SMILES for the two materials. While it can be assumed that this does not pose a problem for the graph representation, as messages will simply not be passed between multiple materials, it is unclear how it may have affected fingerprint generation. Such model-specific limitations within the featurisation could be mitigated by aggregating the predictions of multiple models.

4.3.2 Further research

Many options can still be explored that may further improve the models. XGBoost could be improved further by widening the hyperoptimisation space, for example. CrabNet has only recently been developed and is continuously worked on, which may in the future also improve its applicability for PCE prediction. The GNN could benefit from further architecture exploration, among which could be the development of message passing layers tailored to perovskites or CTL materials, similar to the chemistry-specific graph convolutions developed by Xie and Grossman (2018) for crystal graphs and Duvenaud et al. (2015) for constructing Morgan fingerprints.[28]

[28]Both (available in Pytorch Geometric as "CGConv" and "MFConv") were tested for the GNN but were found inferior to conventional message passing layers for the present application.

4.3.3 Outlook

When comparing to previous works, some by far outperform the models constructed here: Liu et al. (2022) predict PCE with $RMSE = 1.58\,\%\text{pt}$ and Lu et al. (2023) even achieve $RMSE = 1.28\,\%\text{pt}$.[29] However, the ones that use the Perovskite Database generally perform in a similar range, with Hussain et al. (2023) achieving $RMSE = 2.42\,\%\text{pt}$ and Khan et al. (2023) achieving $MAE = 2.9\,\%\text{pt}$. All of these studies not only used very rich information for their predictions, but also filtered the data rigorously to facilitate ML. In contrast, the main benefit of the present work is its extraordinary simplicity, using only sparse information on the absorber and CTLs. With this, the achieved MAE of 2.45 %pt can be considered remarkable. Adding more variables and increasing model complexity warrants potential for improvement and may in the long run even become a foundation for reaching new levels of prediction quality.

[29] RMSE and MAE are comparable though not exactly the same. In RMSE, larger errors have more impact than in MAE, rendering it the stricter criterion.

5 Objective 3: Construction of CTL Selection Helper Tool

With the results of Objective 1, it is now possible to automatically translate commonly used CTL material names into SMILES codes. With Objective 2, these can be used to predict PCE, thereby allowing a conclusion on suitable CTL materials for PSCs. In order to make both results accessible and usable in practice, Objective 3 was to create a tool that can assist researchers in selecting appropriate CTL materials to maximise device efficiency.

5.1 Method

The *CTL Selection Helper* was constructed using *Jupyter Widgets*[G] (also known as *IPython Widgets*) and consists of two parts (see Figure 10):

The main part is the *CTL finder*, a wrapper for the XGBoost predictions. To use it, the user first enters an absorber material and its band gap. Then, they provide a CTL material that will stay fixed for the comparisons and a list of CTL materials that will be compared. Clicking the prediction button for either ETL or HTL will output a prediction table that ranks the combinations by PCE which is also displayed. The predictions are calculated by the champion model from Objective 2, the XGBoost model with featurisation of CTLs through fingerprinting. The CTL materials need to be entered in SMILES encoded form.

If the user does not have the SMILES code readily available, the second part, the *SMILES Translator* can transform a common CTL name into a SMILES code. The function underlying this feature is derived from the identification pipeline in Objective 1. First, a given material is searched in the dictionary that was the result of the identification pipeline. If the material is not present in the dictionary and cannot be found directly within the PubChem database, the SMILES Translator will ask the user to provide DOIs for papers in which the material was used. These are then given to the LLM pipeline which processes them as described for Objective 1. Finally, a SMILES code will be shown if the search was successful.

[G] The use of widgets to create user interfaces had been taught in the module *Introduction to Programming*, along with many useful skills for data wrangling.

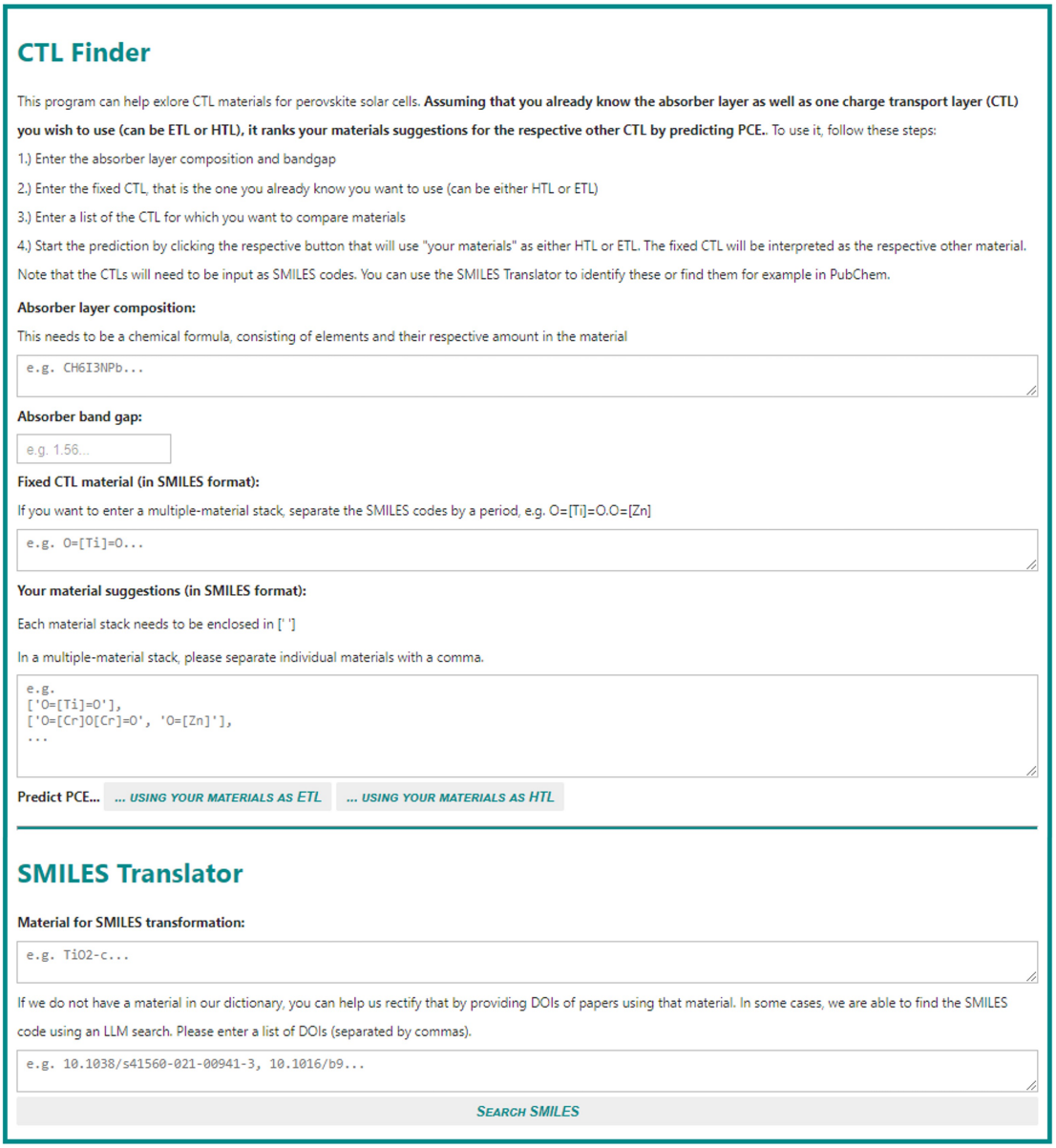


Figure 10: Screenshot of the CTL Selection Helper.

5.2 Discussion

The above described CTL selection helper is able to rank potential CTLs for a given PSC device stack. While it could be considered to be only at alpha level in software release cycle terms, it does perform its core functionality, the prediction of PCE and ranking of CTL materials reliably and without known bugs. Future development could add further features to tailor it more to the needs of PSC researchers.

5.2.1 Usability Testing and Evaluation

In order to explore usability and evaluate the development results, usability tests could be conducted with PSC researchers. A guideline for a simplified usability test can be found in Appendix D. It is to be noted that such tests were not conducted due to the requirement of ethical approval for testing with humans.[H]

5.2.2 Outlook

To make the application publicly available, it could be implemented in the NOMAD-lab, i.e., the web interface of the NOMAD database, where other tools for exploring the data from the Perovskite Database Project are also available. This would enable PSC researchers to make use of it for their research. Additionally, it would be possible to collect user inputs for additional materials and user feedback on wrong identifications to continuously improve the dictionary linking CTL materials with their SMILES codes.

[H]The *Data Governance and Ethics* module taught students to be very careful when dealing with human test subjects or human data in general.

6 General Discussion

The three objectives of this master thesis were successfully completed. For Objective 1, the developed pipeline compiles a materials dictionary for 809 CTL materials linking them with machine-readable SMILES codes. For Objective 2, three ML approaches were compared to identify the impact of incorporating the structural CTL information into PCE prediction models. While statistical testing was inconclusive, the XGBoost model emerged as the best, predicting PCE with $MAE = 2.45\,\%\text{pt}$. Making these results accessible to researches in an interactive application completes Objective 3.

For each objective, the specific limitations, suggestions for further research, and an outlook have been discussed in their respective sections and a broader perspective will be taken here.

6.1 Usefulness of CTL Materials in the Prediction

This work is based on the assumption that CTL materials have a relevant impact on PSC performance (see Charge Transport in Solar Cells). The improvements in the PCE predictions through the addition of CTL information into the models confirm this assumption. Previous publications have also considered CTLs (e.g., Hu et al., 2024; Liu et al., 2024), but usually only include them via label encoding (see related work). With label encoding, models are inevitably limited to materials seen during training. Using structural and compositional information, however, makes the model more generalisable for materials not previously seen. This means that a model such as the presented GNN or XGBoost could be used to investigate more recent data with novel materials. Especially data generated since 2021, which are strongly underrepresented in the Perovskite Database in NOMAD (see Appendix A), could be investigated further.

6.2 The Place of This Work Within the Field

This work ties in with previous publications on ML in the context of PSC research, similarly presenting a model capable of predicting PCE. The distinguishing features of the present work are its consideration of the structure and composition of CTLs, the use of simple data inputs, and the use of a large data set to enhance validity. Additionally, the CTL selection helper tool makes the results of the present

work practically usable for PSC researchers. This is particularly important as such tools may help bridge the gap between the researchers practically experimenting with PSCs, who are typically chemists, (electrical) engineers or physicists, and researchers concerned with ML and simulation approaches, who are typically associated to the fields of computer/data science, mathematics or statistics. Without such tools that make ML results practically usable, such results may miss their true potential.

6.3 The Future of the Perovskite Database

Since the publication of the Perovskite Database Project, researchers were able to add data to the open database. However, this option was hardly used, presumably due to it costing the researchers valuable time. Tools such as the identification pipeline presented here could serve as the foundation for automated pre-processing of PSC data to facilitate entering them into the database.

7 Conclusion

In summary, this thesis presented an approach to PCE prediction in PSCs using the vast open database collected in the Perovskite Database Project. Automated transformation of CTL materials' common names allowed including structural information on CTLs into models that predict PCE, improving their performance. A user interface was built around the champion model, allowing researchers to use it for exploring CTL materials. With the potential to improve the re-usability and accessibility of the currently largest available collection of FAIR PSC data, these results are a small but relevant contribution to the development of perovskite photovoltaics.

References

Altman, N. and Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, 15(6):399–400.

Ameen, S., Akhtar, M. S., Shin, H.-S., and Nazeeruddin, M. K. (2018). *Charge-Transporting Materials for Perovskite Solar Cells*, page 185–246. Elsevier.

ASTM International (2023). Standard Tables for Reference Solar Spectral Irradiances: Direct Normal and Hemispherical on 37° Tilted Surface.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, 160(901):268–282.

Basit, M. A., Aanish Ali, M., and Yasmeen, M. (2023). *Solar Cells and Relevant Machine Learning*, page 1–20. Springer Nature Singapore.

Becquerel, A. E. (1839). *Mémoire sur les effets électriques produits sous l'influence des rayons solaires*, volume 9. Comptes Rendus.

Bergstra, J., Yamins, D., and Cox, D. D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page 115–123. Journal of Machine Learning Research.

Bhattarai, S., Mhamdi, A., Hossain, I., Raoui, Y., Pandey, R., Madan, J., Bouazizi, A., Maiti, M., Gogoi, D., and Sharma, A. (2022). A detailed review of perovskite solar cells: Introduction, working principle, modelling, fabrication techniques, future challenges. *Micro and Nanostructures*, 172:207450.

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., and 79 others (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.*

Chen, C., Maqsood, A., and Jacobsson, T. J. (2023). The role of machine learning in perovskite solar cell research. *Journal of Alloys and Compounds*, 960:170824.

Chen, M., Yin, Z., Shan, Z., Zheng, X., Liu, L., Dai, Z., Zhang, J., Liu, S. F., and Xu, Z. (2024). Application of machine learning in perovskite materials and devices: A review. *Journal of Energy Chemistry*, 94:254–272.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM.

Cheng, M., Zuo, C., Wu, Y., Li, Z., Xu, B., Hua, Y., and Ding, L. (2020). Charge-transport layer engineering in perovskite solar cells. *Science Bulletin*, 65(15):1237–1241.

De Luna, P., Wei, J., Bengio, Y., Aspuru-Guzik, A., and Sargent, E. (2017). Use machine learning to find energy materials. *Nature*, 552(7683):23–27.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2224–2232. MIT Press.

FAIRmat (2023). NOMAD Oasis. https://nomad-lab.eu/nomad-lab/nomad-oasis.html. [Accessed 14-03-2024].

FAIRmat (2024). https://www.fairmat-nfdi.eu/fairmat. [Accessed 21-03-2024].

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Fearn, T. (2024). Testing differences in predictive ability: A tutorial. *Journal of Chemometrics*, 38(8).

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. *arXiv*.

Foo, S., Thambidurai, M., Senthil Kumar, P., Yuvakkumar, R., Huang, Y., and Dang, C. (2022). Recent review on electron transport layers in perovskite solar cells. *International Journal of Energy Research*, 46(15):21441–21451.

Fraas, L. M. (2014). *History of Solar Cell Development*, page 1–12. Springer

International Publishing.

Fung, V., Zhang, J., Juarez, E., and Sumpter, B. G. (2021). Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7(1).

Gasteiger, J., Groß, J., and Günnemann, S. (2020). Directional message passing for molecular graphs. *arXiv*.

Goodall, R. E. A. and Lee, A. A. (2020). Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nature Communications*, 11(1).

Green, M. A., Dunlop, E. D., Yoshita, M., Kopidakis, N., Bothe, K., Siefer, G., Hinken, D., Rauer, M., Hohl-Ebinger, J., and Hao, X. (2024). Solar cell efficiency tables (version 64). *Progress in Photovoltaics: Research and Applications*, 32(7):425–441.

Guo, X., Zhou, H. Y., Guo, S., Luan, X. X., Cui, W. K., Ma, Y. F., and Shi, L. (2014). Design of broadband omnidirectional antireflection coatings using ant colony algorithm. *Optics Express*, 22(S4):A1137.

Guo, Z. and Lin, B. (2021). Machine learning stability and band gap of lead-free halide double perovskite materials for perovskite solar cells. *Solar Energy*, 228:689–699.

Hu, J., Chen, Z., Chen, Y., Liu, H., Li, W., Wang, Y., Peng, L., Liu, X., Lin, J., Chen, X., and Wu, J. (2024). Interpretable machine learning predictions for efficient perovskite solar cell development. *Solar Energy Materials and Solar Cells*, 271:112826.

Hussain, W., Sawar, S., and Sultan, M. (2023). Leveraging machine learning to consolidate the diversity in experimental results of perovskite solar cells. *RSC Advances*, 13(32):22529–22537.

Jacobsson, T. J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., and 88 others (2021). An open-access database and analysis tool for perovskite solar cells based on the fair data principles. *Nature Energy*, 7(1):107–115.

Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C., and

Agrawal, A. (2018). Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific Reports*, 8(1).

Joshi, N., Kushvaha, V., and Madhushri, P., editors (2023). *Machine Learning for Advanced Functional Materials*. Springer Nature Singapore.

Khan, A., Kandel, J., Tayara, H., and Chong, K. T. (2023). Predicting the bandgap and efficiency of perovskite solar cells using machine learning methods. *Molecular Informatics*, 43(2).

Kojima, A., Teshima, K., Shirai, Y., and Miyasaka, T. (2009). Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *Journal of the American Chemical Society*, 131(17):6050–6051.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:89–91.

Landrum, G., Tosco, P., Kelley, B., and 14 others (2024). Rdkit. url = https://zenodo.org/doi/10.5281/zenodo.591637.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292.

Li, F. and Jen, A. K.-Y. (2022). Interface engineering in solution-processed thin-film solar cells. *Accounts of Materials Research*, 3(3):272–282.

Li, F., Peng, X., Wang, Z., Zhou, Y., Wu, Y., Jiang, M., and Xu, M. (2019). Machine learning (ml)-assisted design and fabrication for solar cells. *Energy & Environmental Materials*, 2(4):280–291.

Li, W., Hu, J., Chen, Z., Jiang, H., Wu, J., Meng, X., Fang, X., Lin, J., Ma, X., Yang, T., Cheng, P., and Xie, R. (2023). Performance prediction and optimization of perovskite solar cells based on the bayesian approach. *Solar Energy*, 262:111853.

Liu, H., Chen, Z., Zhang, Y., Wu, J., Peng, L., Wang, Y., Liu, X., Chen, X., and Lin, J. (2024). Bayesian reverse design of high-efficiency perovskite solar cells based on experimental knowledge constraints. *Applied Physics Letters*, 125(6).

Liu, Y., Yan, W., Han, S., Zhu, H., Tu, Y., Guan, L., and Tan, X. (2022). How

machine learning predicts and explains the performance of perovskite solar cells. *Solar RRL*, 6(6).

Lu, Y., Wei, D., Liu, W., Meng, J., Huo, X., Zhang, Y., Liang, Z., Qiao, B., Zhao, S., Song, D., and Xu, Z. (2023). Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results. *Journal of Energy Chemistry*, 77:200–208.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Magomedov, A., Al-Ashouri, A., Kasparavičius, E., Strazdaite, S., Niaura, G., Jošt, M., Malinauskas, T., Albrecht, S., and Getautis, V. (2018). Hole transporting monolayers: Self-assembled hole transporting monolayer for highly efficient perovskite solar cells (adv. energy mater. 32/2018). *Advanced Energy Materials*, 8(32).

Mahmood, K., Sarwar, S., and Mehran, M. T. (2017). Current status of electron transport layers in perovskite solar cells: materials and properties. *RSC Advances*, 7(28):17044–17062.

Mavračić, J., Court, C. J., Isazawa, T., Elliott, S. R., and Cole, J. M. (2021). Chemdataextractor 2.0: Autopopulated ontologies for materials science. *Journal of Chemical Information and Modeling*, 61(9):4280–4289.

National Renewable Energy Laboratory (2024). Best Research-Cell Efficiency Chart. https://www.nrel.gov/pv/cell-efficiency.html. [Accessed 11-07-2024].

Nelson, J. (2003). *The Physics of Solar Cells*. Imperial College Press.

Niemegeers, A., Burgelman, M., Decock, K., Degrave, S., and Verschraegen, J. (2020). Simulation programme SCAPS-1D for thin film solar cells developed at ELIS, University of Gent. https://scaps.elis.ugent.be/. [Accessed 23-08-2024].

Padula, D., Simpson, J. D., and Troisi, A. (2019). Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons*, 6(2):343–349.

Polak, M. P., Modi, S., Latosinska, A., Zhang, J., Wang, C.-W., Wang, S., Hazra, A. D., and Morgan, D. (2024). Flexible, model-agnostic method for materials data extraction from text using general purpose language models. *Digital Discovery*, 3(6):1221–1235.

Polak, M. P. and Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1).

Rath, S., Sudha Priyanga, G., Nagappan, N., and Thomas, T. (2022). Discovery of direct band gap perovskites for light harvesting by using machine learning. *Computational Materials Science*, 210:111476.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., and Friederich, P. (2022). Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1).

Righini, G. C. and Enrichi, F. (2020). *Solar cells' evolution and perspectives: a short review*, page 1–32. Elsevier.

Royal Society of Chemistry (2024). ChemSpider - Search and share chemistry. https://www.chemspider.com/. [Accessed 23-07-2024].

Said, A. A., Xie, J., and Zhang, Q. (2019). Recent progress in organic electron transport materials in inverted perovskite solar cells. *Small*, 15(27).

Scheidgen, M., Himanen, L., Ladines, A. N., Sikter, D., Nakhaee, M., Fekete, , Chang, T., and 22 others (2023). Nomad: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8(90):5388.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. (2018). Schnet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24).

Schygulla, P., Beutel, P., Heckelmann, S., Höhn, O., Klitzke, M., Schön, J., Oliva, E., Predan, F., Schachtner, M., Siefer, G., Helmers, H., Dimroth, F., and Lackner, D. (2022). Quadruple junction solar cell with 47.6% conversion efficiency

under concentration. Presentation held at The 20th international Conference on Metal Organic Vapor Phase Epitaxy.

Shang, Y., Xiong, Z., An, K., Hauch, J. A., Brabec, C. J., and Li, N. (2024). Materials genome engineering accelerates the research and development of organic and perovskite photovoltaics. *Materials Genome Engineering Advances*, 2(1).

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611.

Shockley, W. and Queisser, H. J. (1961). Detailed balance limit of efficiency of p-n junction solar cells. *Journal of Applied Physics*, 32(3):510–519.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281.

Swain, M. C. and Cole, J. M. (2016). Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904.

Ud Din, A. and Qureshi, S. (2024). Limits of depth: Over-smoothing and over-squashing in gnns. *Big Data Mining and Analytics*, 7(1):205–216.

Veličković, P. (2022). Message passing all the way up.

Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J., and Sparks, T. D. (2021). Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 7(1).

Weston, L. and Stampfl, C. (2018). Machine learning the band gap properties of kesterite I2-II-IV-V4 quaternary compounds for photovoltaics applications. *Physical Review Materials*, 2(8).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., and 47 others (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1).

Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301.

Yan, X., Poxson, D. J., Cho, J., Welser, R. E., Sood, A. K., Kim, J. K., and Schubert, E. F. (2012). Enhanced omnidirectional photovoltaic performance of solar cells using multiple-discrete-layer tailored- and low-refractive index anti-reflection coatings. *Advanced Functional Materials*, 23(5):583–590.

Yao, Z., Lum, Y., Johnston, A., Mejia-Mendoza, L. M., Zhou, X., Wen, Y., Aspuru-Guzik, A., Sargent, E. H., and Seh, Z. W. (2022). Machine learning for a sustainable energy future. *Nature Reviews Materials*, 8(3):202–215.

Zhang, T., He, Q., Yu, J., Chen, A., Zhang, Z., and Pan, J. (2022). Recent progress in improving strategies of inorganic electron transport layers for perovskite solar cells. *Nano Energy*, 104:107918.

Zhao, Q., Zhou, B., Luo, L., Duan, Z., Xie, Z., and Hu, Y. (2023). A literature overview of cell layer materials for perovskite solar cells. *MRS Communications*, 13(6):1076–1086.

Zhao, R., Xing, B., Mu, H., Fu, Y., and Zhang, L. (2022). Evaluation of performance of machine learning methods in mining structure–property data of halide perovskite materials. *Chinese Physics B*, 31(5):056302.

A Charge Transport Layer Materials

The choice of appropriate CTL materials is important and many different combinations have been tested in the laboratory as well as in simulation studies. Relevant properties of CTL materials are (Foo et al., 2022; Li and Jen, 2022):

- *Band alignment* with the absorber layer: The conduction band of the ETL should align with that of the absorber to facilitate the transfer of electrons, while the valence band of the HTL should align with that of the absorber to facilitate the transfer of holes
- The absorber layer's *carrier selective mobility*, i.e., the ease with which electrons or holes move through the material.
- *Optical transparency*, which influences how much light reaches the absorber layer.
- *Ion migration*, i.e., the movement of ions through a material and how it affects its electrical conductivity and stability.
- *Trap density*, i.e., the measure of the frequency of defect sites or localised states within a material that can capture and hold charge carriers, which can impact the material's electronic behaviour, such as its conductivity, carrier lifetimes and recombination rates.
- *Long term stability*, which is crucial for practicality and commercialisation.
- *Interfacial interactions* with other device layers.

In addition to these properties, there are also relevant processing conditions when fabricating full solar cell devices. For example, the commonly used ETL material titanium dioxide (TiO_2) may require a high temperature sintering process for optimal functionality which might damage the absorber layer (Cheng et al., 2020). From these considerations a vast field of potential materials has emerged, the most prominent of which are depicted in Figure 11.

Metal oxides (e.g., TiO_2, ZnO) and organic fullerene-based materials (e.g., C60, PCBM[30]) are used for ETLs, while organic molecules (e.g., Spiro-MeOTAD[31]),

[30] phenyl-C61-butyric acid methyl ester
[31] 2,2',7,7'-tetrakis[N,N-di(4-methoxyphenyl)amino]-9-9'-spirobifluorene

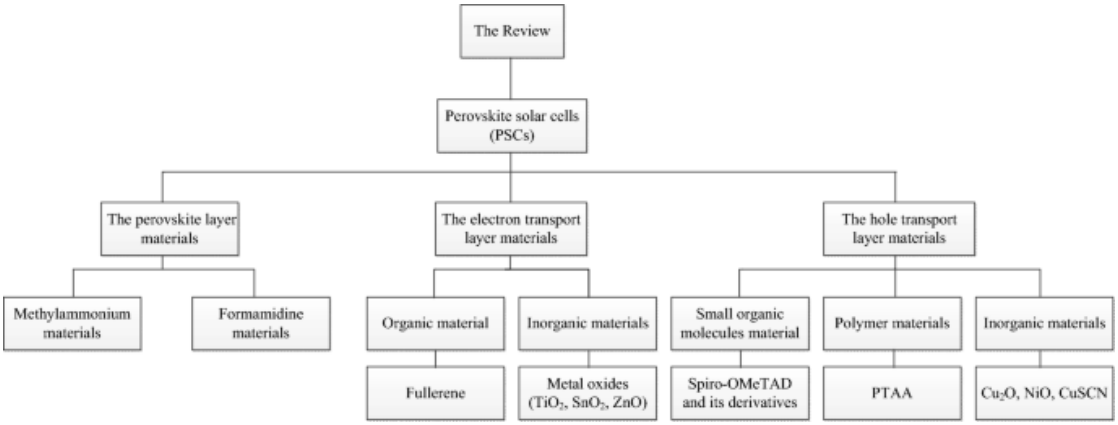


Figure 11: Typical materials used in PSCs
Reproduced from the review by Zhao et al. (2023), Graphical Abstract.

polymers (e.g., PTAA[32], PEDOT:PSS[33]), inorganic materials as well as *self-assembled mono-layers* (e.g., 2PACz[34]) are used for HTLs (Mahmood et al., 2017; Zhang et al., 2022; Said et al., 2019; Magomedov et al., 2018; Li and Jen, 2022).

[32] poly[bis(4-phenyl)(2,4,6-trimethylphenyl)amine]
[33] poly(3,4-ethylenedioxythiophene) polystyrene sulfonate
[34] [2-(9H-carbazol-9-yl)ethyl]phosphonic acid

B PCEs by Publication Date and CTLs

Figure 12: PCEs by publication date and CTLs for the most frequently used ETL stacks.

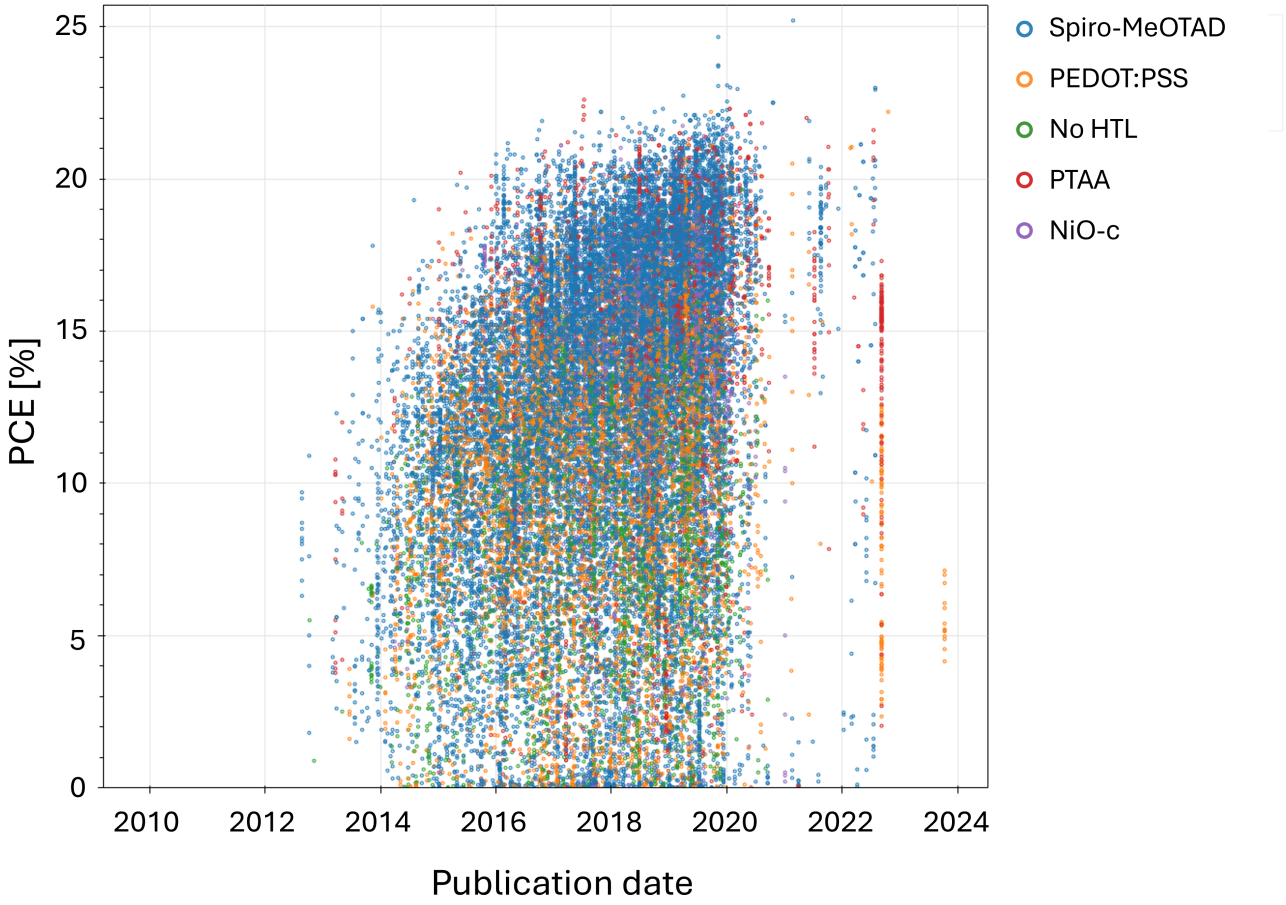


Figure 13: PCEs by publication date and CTLs for the most frequently used HTL stacks.

C Statistical Evaluation of ML Prediction Quality Differences

To analyse whether the differences in prediction quality of different ML models are significant, statistical tests need to be performed. Fearn (2024) recommend using a paired mean comparison test with the differences in error for each individual prediction. $d_j = |e_{1j}| - |e_{2j}|$ calculates the absolute error difference for each sample, while $\overline{d} = \frac{\sum_j^n d_j}{n}$ calculates the mean difference of paired prediction errors and, for a t-test, $t = \frac{\sqrt{n} \cdot \overline{d}}{s_d}$ with the standard deviation in the differences being $s_d = \sqrt{\frac{\sum_j^n (d_j - \overline{d})^2}{n-1}}$ can be evaluated. Since the absolute errors and their differences are typically not normally distributed, a recommended alternative to the t-test is the Wilcoxon Signed-Rank Test (Wilcoxon, 1945). Similar to these comparisons, with which two models can be compared, for multiple comparisons along factors an analysis of variance (ANOVA) can be suitable. If the paired difference evaluation is not possible—e.g., because the test data set is not identical for the compared models—one could also compare the aggregated errors (e.g., $MAE_i = \frac{\sum_j^n |e_{ij}|}{n}$) of each predictor i. Using a statistical test (e.g., t-test) on these has a significantly lower power due to likely unfulfilled distributional requirements.

C.1 Tests Used for Model Comparison

Here, a paired comparison was possible, as the test sets were identical for all models.

C.1.1 ANOVA

Following the methods described by Fearn (2024), the prediction errors were evaluated with a two-way ANOVA on a data set created out of the prediction errors from the models (N = 4,887 predictions, comprising 543 predictions for each of the nine models). One factor represented the model type (XGBoost, CrabNet or GNN) and the other factor represented the model configuration (Baseline, label-encoding for CTLs, fingerprint or graph featurisation for CTLs). Note that all p values from statistical tests were compared against an alpha level of 0.05 to determine significance. In the ANOVA, the main effect for model type was significant, $F(2, 4878) = 9.84$, $p < .001$, as was the effect of model configuration level,

	XGBoost	CrabNet	GNN
MAE difference [%pt]	0.3	0.01	0.43
W statistic	71843	62230	63833
p	.58	0.001	.006

Table 3: T-tests between baseline models and models with full CTL information.

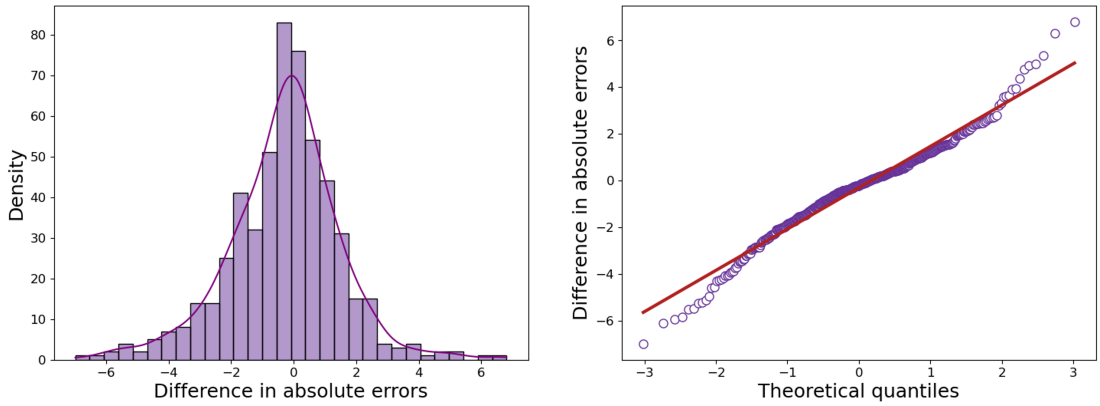
$F(2, 4878) = 4.51$, $p = .01$. The interaction between the two was non-significant, $F(4, 4878) = 1.99$, $p = .09$.

C.1.2 Pairwise Mean Comparisons Using Wilcoxon Signed-Rank Test

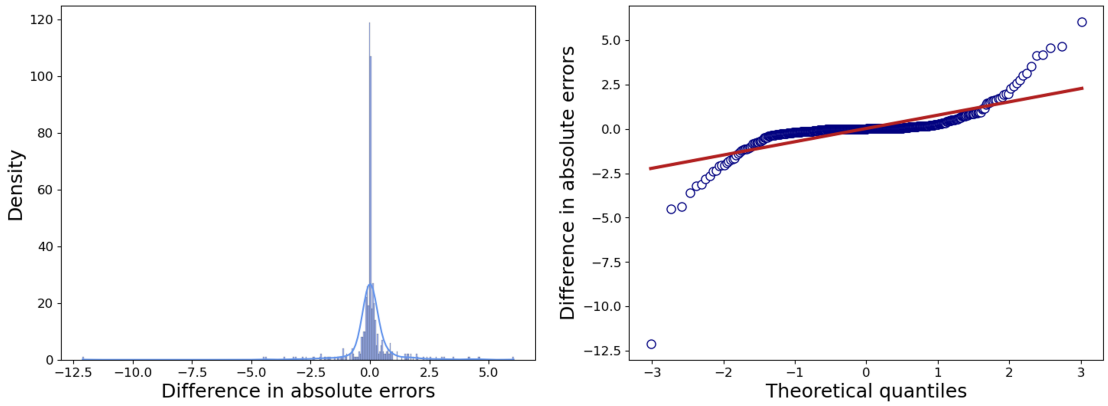
With the ANOVA testing for all effects at once, specific comparisons were conducted additionally. Following the logic that more information should lead to better predictions, the full models were tested against the baseline models. For this, Wilcoxon's Signed-Rank Test (Wilcoxon, 1945) was used to compare the differences in absolute prediction error for two models at a time against zero. Unlike its parametric counterpart, the paired t-test, this non-parametric test does not require the data to follow a normal distribution. More on distribution requirements and problems follows suit. Table 3 shows the results for the comparison of full models and baseline, which was significant for both the GNN and CrabNet, but not for XGBoost. Even conservative family-wise error correction in the form of a Bonferroni-Holm correction does not change these results.

C.2 Distributions of Prediction Errors

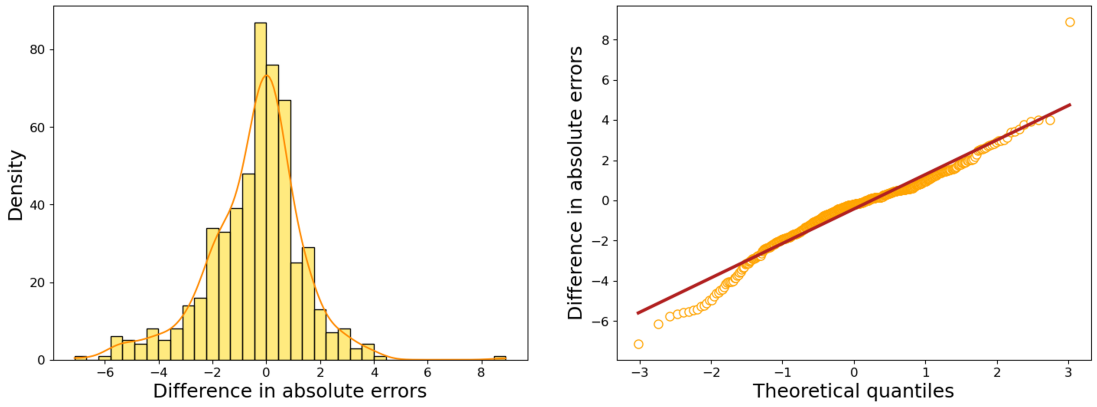
The statistical tests need to be interpreted with caution as their underlying distributions may limit the validity of the tests. Normality was judged through visual inspection, as general use of tests like the the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1948) or Shapiro-Wilk test (Shapiro and Wilk, 1965) can lead to alpha error inflation. As an example for visual inspection, the distributions of the error differences for the three Wilcoxon test comparisons can be seen in Figure 14.



(a) XGBoost: The distribution is approximately normal.



(b) CrabNet: The distribution is not normal. Many error differences amount to zero, meaning that predictions of full and baseline models are similar.



(c) GNN: The distribution appears approximately normal.

Figure 14: Distribution of prediction error differences for the Wilcoxon Signed-Rank Tests between full model and baseline model.
Normality would mean the left graphs approach a bell curve and the QQ-plots (right) approach the diagonal.

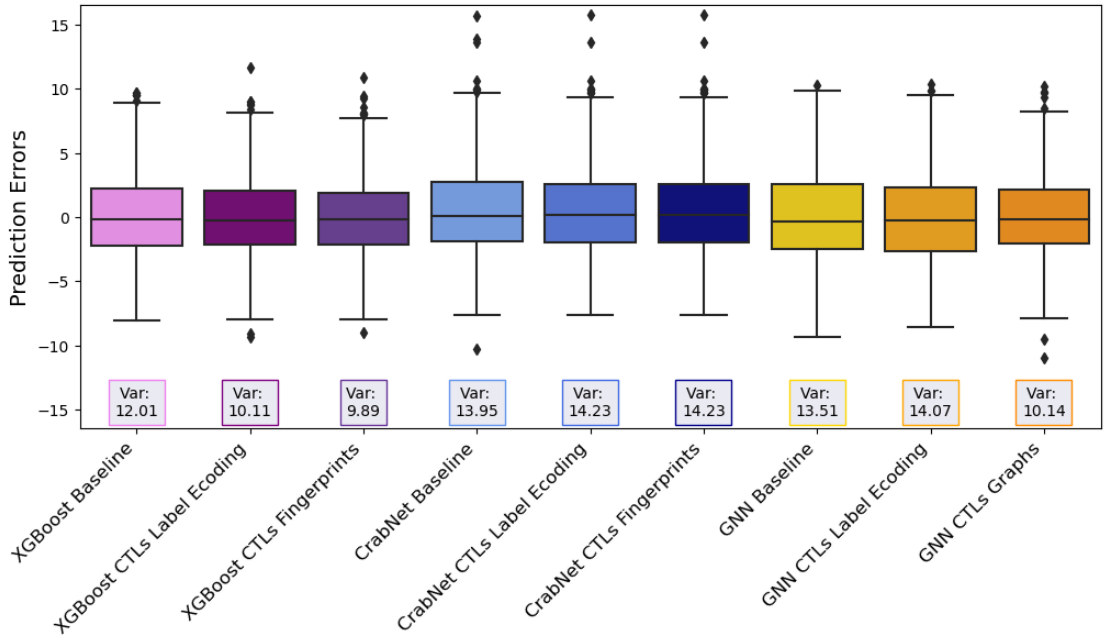


Figure 15: Variances across all predicted sets
Whiskers extend to the lowest and highest value within 1.5 times the interquartile range. Points beyond that can be considered outliers.

For the three full-model-vs-baseline comparisons, normality can be assumed for the XGBoost and GNN error differences, but not for the ones from CrabNet. For ANOVA, additionally, homogeneity of variances needs to be considered. Figure 15 shows that across all models and levels, variances are roughly equal. Again, this could be tested (using Bartlett's test (Bartlett, 1937) or Levene's test (Levene, 1960) if normality is compromised), but the tests would be overpowered and therefore result in significance even with irrelevant deviations from normal distributions.

C.3 Outliers

As apparent in Figure 14 and Figure 15, there is also the issue of outliers. Dealing with outliers in the context of prediction errors is difficult: Is an outlier caused by unusual values in the true values? In the prediction? Some outliers only appear in the error differences between models, some are present in the absolute predictions but not the error differences. All tests were conducted once with very liberal

outlier removal, removing up to eight outliers, but this did not impact any results to become significant. Here, the results without any outlier removal are reported in order to be as conservative as possible.

C.4 Statistical Conclusions

Statistical tests are not trivial and do not always give a clear picture, especially when taking into account distributional properties and assumptions. Here, the conclusion needs to be viewed with respect to the limitations. While there are significant differences, they could not be determined reliably. Due to space constraints, it would not have been possible to present a differentiated assessment in the main text. Therefore, it was decided not to interpret statistical significance there as any conclusion would have needed to be inappropriately simplified.

D Usability Test Guideline for the CTL Selection Helper

In the early stage of development, a rudimentary usability test could help screen for the practical usefulness of the program as well as potential requested features. The test should be conducted as semi-structured interviews, i.e., loosely following a guideline such as the one below. The results yielded from this test would be:

- the "think-aloud" comments of participants, potentially pointing at misconceptions, lacking usability, or errors in the program,
- the answers to questions a, b, and c, which could be averaged and be compared either against a pre-defined threshold to determine program quality or against previous prototype versions to determine development progress,
- the answers to question d, which can point at additional features that could improve the program.

Interview guideline

1. Greeting, express thanks for participation
2. Inform about time (15 minutes) and voluntariness of participation, briefly explain procedure, get **informed consent**
3. **Explain Task**: Imagine you are a researcher interested in trying new materials for their next batch of PSCs. Use the program accordingly. During the task, please comment aloud what you are doing and thinking.
4. Time for task completion
5. After completion, ask the following questions (read aloud literally)
   (a) How easy or hard was it to use the program? Please answer with a number on a scale from 1 to 7, where 1 means "very easy" and 7 means "very hard". 4 means "neither easy nor hard".
   (b) How useful do you find the program for your personal research? Please answer with a number on a scale from 1 to 7, where 1 means "not useful at all" and 7 means "very useful". 4 means "neither".

(c) How useful do you find the program for PSC research in general? Please answer with a number on a scale from 1 to 7, where 1 means "not useful at all" and 7 means "very useful". 4 means "neither".

(d) Which features would you like to have in addition to the ones currently in the program?