

Faraday Discussions

Accepted Manuscript



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

This article can be cited before page numbers have been issued, to do this please use: D. A. Garzn, L. Himanen, L. Andrade, S. Sadewasser and J. A. Márquez, *Faraday Discuss.*, 2026, DOI: 10.1039/D6FD00030D.

ML-guided screening of chalcogenide perovskites as solar energy materials

Diego A. Garzón^{1,2,✉}, Lauri Himanen³, Luisa Andrade², Sascha Sadewasser¹, and José A. Márquez³

¹International Iberian Nanotechnology Laboratory, Av. Mte. José Veiga s/n, 4715-330 Braga, Portugal

²LEPABE, ALiCE, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

³Physics Department and CSMB, Humboldt-Universität zu Berlin, Zum Großen Windkanal 2, D-12489 Berlin, Germany
✉ diego.garzon@inl.int

April 8, 2026

Abstract

Chalcogenide perovskites have emerged as promising absorber materials for next-generation photovoltaic devices, yet their experimental realization remains limited by competing phases, structural polymorphism, and synthetic challenges. Here, we present a fully data-driven and experimentally grounded screening and ranking framework to assess the stability and experimental feasibility of chalcogenide perovskites, integrating interpretable analytical descriptors, machine-learning models, and sustainability metrics. Using a curated experimental dataset of halide and chalcogenide compounds, we derive a new tolerance factor via the SISO (sure independence screening and sparsifying operator) algorithm that more accurately distinguishes perovskite-forming compositions than established tolerance-factor-based screening criteria. This descriptor is combined with generative crystal structure prediction, composition-based bandgap estimation, and machine-learning-based feasibility assessment to systematically explore a wide chemical space of hypothetical chalcogenide perovskites. The resulting candidates are further evaluated using sustainability indicators, enabling multi-objective ranking tailored to both single-junction and tandem photovoltaic architectures. Beyond identifying several promising and previously unexplored chalcogenide perovskites, this work demonstrates a transferable screening strategy for chemically constrained materials spaces that balances optoelectronic performance, experimental viability, and long-term sustainability.

Introduction

Chalcogenide perovskites have recently emerged as a promising class of inorganic materials for optoelectronic and energy conversion applications, particularly photovoltaics. In these compounds, the anions occupying the X-site in the ABX_3 perovskite-type structure are chalcogenides ($X = S^{2-}, Se^{2-}, Te^{2-}$), excluding oxygen-based systems. Despite less than a decade of experimental development, chalcogenide perovskites have attracted significant attention due to their intrinsic structural robustness, tunable optoelectronic properties, and absence of toxic elements [1].

In the context of thin film solar cell technology, hybrid inorganic-organic halide perovskites currently dominate the research field, having achieved a record laboratory efficiency of 27% [2]. However, concerns regarding their long-term stability, ion migration, and toxicity primarily due to the presence of lead, continue to hinder large-scale deployment [3]. These limitations have motivated the search for alternative perovskite-inspired absorbers that combine high performance with enhanced chemical and thermal stability. In this regard, chalcogenide perovskites emerge as a compelling candidate, although whether they can ultimately deliver comparable photovoltaic performance remains an open question.

Among chalcogenide perovskites, $BaZrS_3$ is the most extensively studied compound, with a reported bandgap in

the range of 1.7–1.95 eV. $BaZrS_3$ exhibits a high optical absorption coefficient above the bandgap, excellent stability under exposure to air, moisture, and heat, and is composed of earth-abundant, non-toxic elements [4]. These characteristics make it particularly attractive as an absorber material for top cells in tandem photovoltaic configurations. Nevertheless, the prominence of $BaZrS_3$ also underscores the limited experimental diversity currently available within this materials family.

Although theoretical studies predict that many chalcogenide perovskites possess favorable optoelectronic properties [5–8], experimental realization of the perovskite-type phase remains scarce beyond a few compositions, including $BaHfS_3$ [9], $Sr(Hf,Zr)S_3$ [10], and $EuZrS_3$ [11]. Moreover, the synthesis of these materials typically requires high-temperature processes, raising concerns for device integration and scalability. As a result, significant experimental effort has been directed toward lowering synthesis temperatures and identifying viable processing routes [12, 13]. Many of these challenges arise from the high thermodynamic stability of binary and polysulfide phases, as well as the presence of competing non-perovskite structures, which can hinder the formation of the perovskite-type phase. This disconnect between theoretical predictions and experimental feasibility represents a major bottleneck in the discovery of new chalcogenide perovskites.

The perovskite-type structure consists of a corner-sharing BX_6 octahedral network with A cations occupying



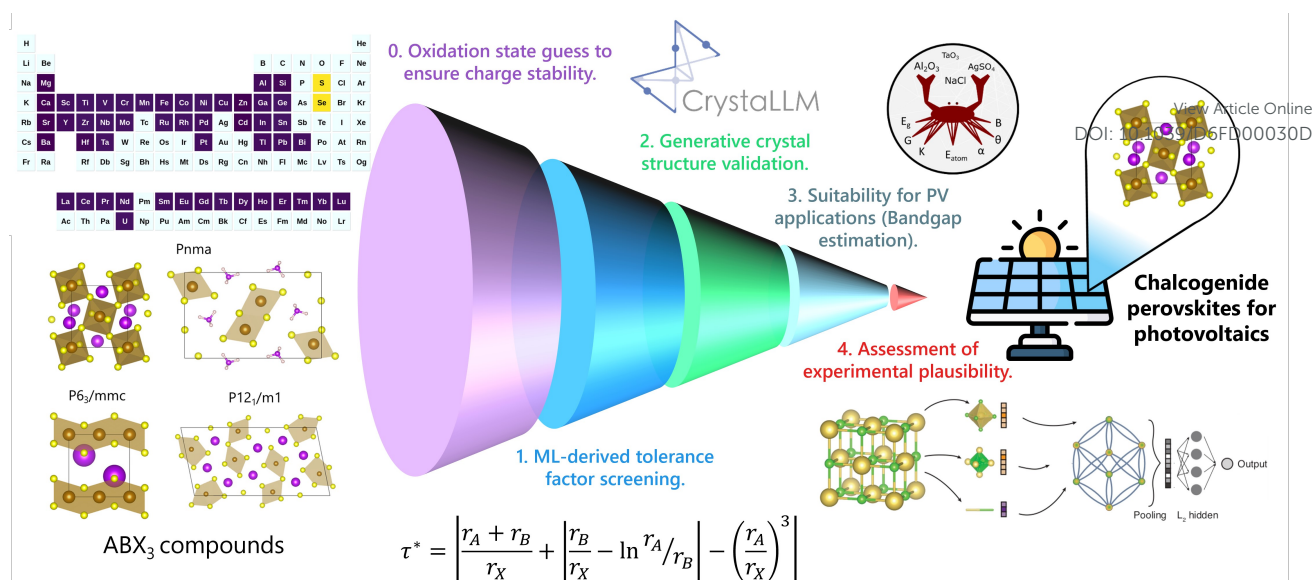


Figure 1: Overview of the ML-guided screening pipeline employed to assess chalcogenide perovskites, combining experimentally motivated descriptors and machine-learning models to evaluate structural stability, crystal structure, experimental plausibility, and photovoltaic suitability.

the dodecahedral sites [14, 15]. To rationalize perovskite phase stability, simple geometrical descriptors such as the Goldschmidt tolerance factor ($t_{\text{Goldschmidt}}$) and the octahedral factor (μ) have long been employed based on ionic radii (r_A , r_B , and r_X) [16]. These dimensionless parameters predict perovskite formation when their values fall within the approximate ranges of 0.8-1 and 0.414-0.732, respectively:

$$t_{\text{Goldschmidt}} = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \quad \mu = \frac{r_B}{r_X} \quad (1)$$

However, these classical descriptors fail to reliably predict phase stability in chalcogenide perovskites, as several non-perovskite chalcogenide phases also fall within these nominal stability ranges [17, 18]. This limitation has been attributed to the increased covalent character of metal-chalcogen bonds compared to oxide or halide analogues. To address this issue, modified tolerance factors incorporating electronegativity differences have been proposed, including the Jess *et al.* tolerance factor (t_{Jess}) and electronegativity-based descriptors like χ_{diff} [17, 18].

$$t_{\text{Jess}} = \frac{\frac{\Delta\chi_{A-X}}{\Delta\chi_{A-O}}(r_A + r_X)}{\sqrt{2} \frac{\Delta\chi_{B-X}}{\Delta\chi_{B-O}}(r_B + r_X)} \quad (2)$$

While these descriptors improve discrimination between perovskite and non-perovskite chalcogenide phases, their generality and predictive power across unexplored compositional spaces remain uncertain, motivating the development of more systematic screening approaches.

Alternative screening methodologies rely on high-throughput first-principles calculations, where crystal structures are relaxed and materials properties such as thermodynamic stability, electronic band structure, and optical response are computed [19–21]. The rise of these computational approaches has significantly impacted the identification of lead-free perovskites, particularly in data-scarce regimes where experimental exploration is costly and

time-consuming. When coupled with machine-learning techniques, these methods enable rapid exploration of large chemical spaces. However, experimental validation of computational predictions remains limited, and structures predicted to be stable by density functional theory (DFT) do not always correspond to experimentally realizable phases [22].

In parallel with high-throughput first-principles approaches, materials informatics has evolved rapidly over the past decade, enabling property prediction directly from composition and/or crystal structure. Early efforts focused on descriptor engineering and sparse regression frameworks such as SISSO (sure independence screening and sparsifying operator), which identify compact, physically interpretable analytical expressions from large feature spaces [23]. More recently, deep learning models such as CrabNet and MODNet have demonstrated accurate composition-based property prediction without explicit structural input [24, 25]. In addition, generative models for crystal structure prediction have emerged [26], ranging from variational autoencoders and diffusion-based approaches to large language models capable of generating complete crystallographic information files [27]. These advances increasingly enable not only forward property prediction but also constraint-aware candidate generation and screening, expanding the scope of data-driven materials discovery beyond conventional paradigms.

Here, we present a machine-learning-guided screening pipeline aimed at identifying potentially synthesizable chalcogenide perovskites, with an emphasis on experimentally grounded descriptors (Figure 1). A new analytical tolerance factor derived using the SISSO framework is introduced to assess structural stability. Sequentially, crystal structure, experimental plausibility, and bandgap are inferred using machine-learning models, including CrystaLLM, CrabNet, GCNNs [24, 27, 28]. The resulting ranked list of candidate materials is intended to provide a focused



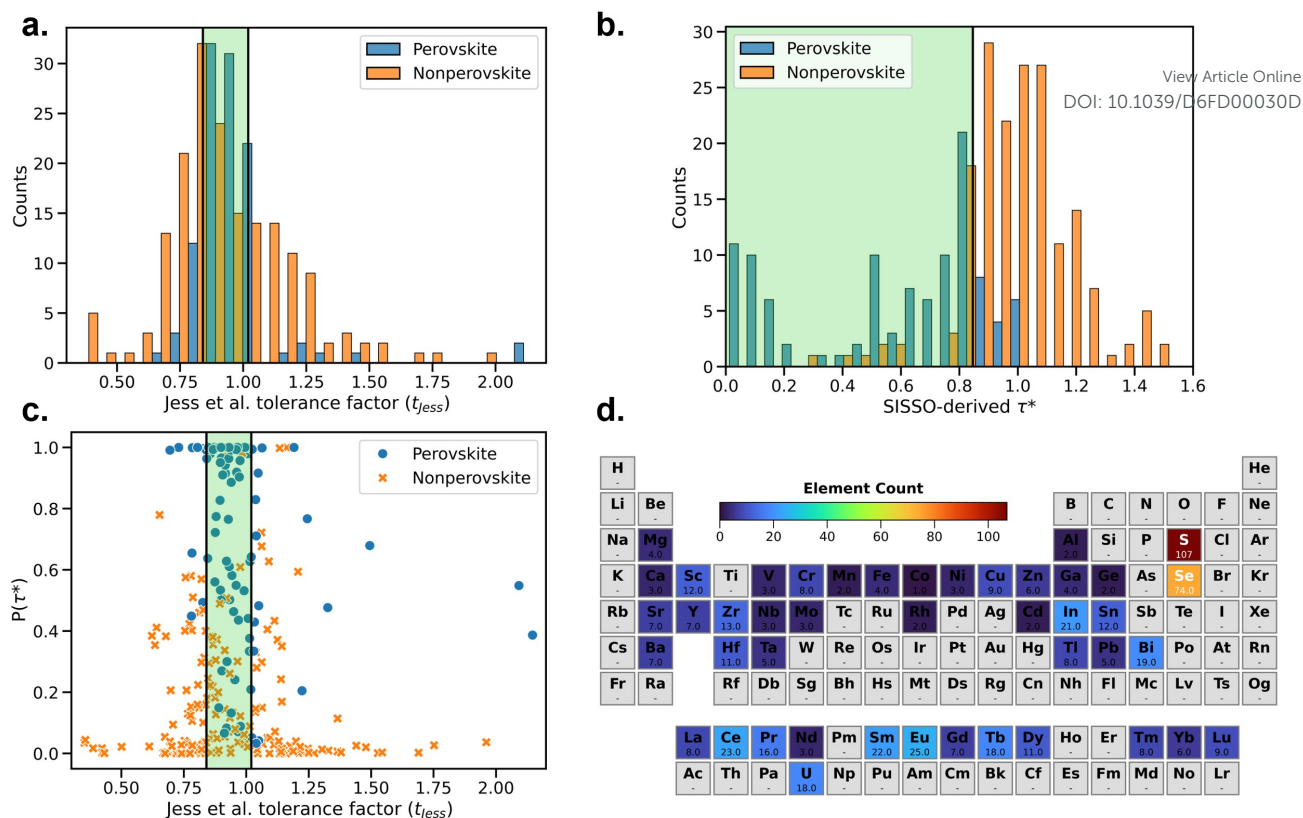


Figure 2: a. Jess *et al.* tolerance factor (t_{JESS}) distribution on the experimental ABX_3 data used for training; b. SISSO-derived τ^* tolerance factor distribution on the experimental ABX_3 data; c. Logistic-calibrated probability of perovskite-type stability based on τ^* as a function of the Jess *et al.* tolerance factor for the experimental ABX_3 data; in each plot the stability region is delimited with a green background; d. Elemental distribution of the predicted perovskite-type phases with $\tau^* < 0.846$.

platform for experimental validation and to stimulate discussion on the strengths, limitations, and future directions of ML-driven materials identification in chalcogenide perovskites. Notably, each stage of the pipeline introduces its own uncertainty; the ensemble therefore does not remove uncertainty but distributes decision-making across partially independent screening criteria.

Results and Discussion

ML-derived tolerance factor

The tolerance factor τ^* was derived using the SISSO algorithm based on experimentally reported halide and chalcogenide perovskites [23]. The resulting expression for τ^* represents an analytically compact descriptor that balances model sparsity with predictive accuracy, while retaining direct physical interpretability in terms of geometric packing and size mismatch. τ^* (Eq. 3) is a function of the ionic radii, where r_A , r_B , and r_X denote the ionic radii of the A-site cation, B-site cation, and anion, respectively. Perovskite stability, including distorted configurations, is defined by the criterion $\tau^* < 0.846$, which was learned from the experimental data with a decision tree classifier (see [Data & Code Availability](#)).

$$\tau^* = \left| \frac{r_A + r_B}{r_X} + \left| \frac{r_B}{r_X} - \ln\left(\frac{r_A}{r_B}\right) \right| - \left(\frac{r_A}{r_X}\right)^3 \right|, \quad (3)$$

Following Turnley *et al.*, we used their chalcogenide-tailored ionic radii for larger and more polarizable anions (S^{2-} , Se^{2-} , Br^- , I^-), while Shannon ionic radii were retained for the more ionic F^- and Cl^- systems [18]. This choice reflects the increased covalent character of metal–chalcogen bonds compared to oxide and fluoride analogues. We note that the selection of ionic radii constitutes an important modeling assumption, and further refinement of radii datasets for specific ABX_3 chemistries may influence the quantitative form of the derived descriptor.

Table 1: Classification performance metrics for the Goldschmidt ($t_{\text{Goldschmidt}}$), Jess *et al.* (t_{JESS}), and Bartel *et al.* (t_{Bartel}) tolerance factors together with the SISSO-derived τ^* for perovskite stability prediction.

Metric	$t_{\text{Goldschmidt}}$	t_{JESS}	t_{Bartel}	τ^*
Accuracy (%)	59.6	70.2	70.2	91.2
Precision (%)	50.0	59.4	59.4	95.0
Recall (%)	60.9	82.6	82.6	82.6
Specificity (%)	58.8	61.8	61.8	97.1
F1-score (%)	55.1	69.1	69.1	88.4



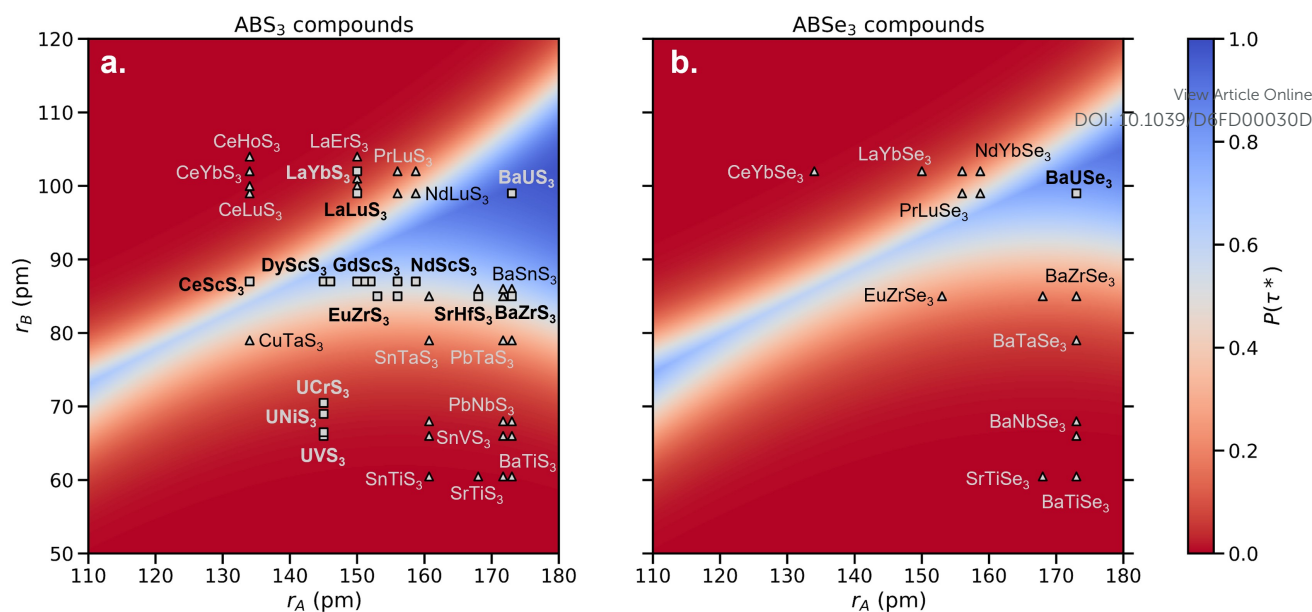


Figure 3: The effects of ionic radii on the stability of chalcogenide perovskites for a. $AB\text{S}_3$ and b. $AB\text{Se}_3$ compounds. The scattered symbols correspond to experimentally synthesized materials, materials with a perovskite-type structure are represented with squares and a bold label, while non-perovskite structures are triangles.

Figure 2 shows the distribution of the experimental dataset for the reported Jess *et al.* tolerance factor (t_{JESS}) [17] (a) and the SISSO-derived τ^* introduced in this work (b). In addition, inspired by Bartel *et al.*, we fitted a logistic calibration that maps τ^* values to a probability of perovskite-type stability derived from the experimental phase labels [29]. Figure 2c shows the calibrated probability as a function of t_{JESS} . Notably, most experimentally observed perovskites—both within and outside the nominal t_{JESS} stability window—exhibit calibrated probabilities above 50%.

Table 1 compares the derived τ^* with already established descriptors like t_{JESS} [17], Goldschmidt tolerance factor ($t_{\text{Goldschmidt}}$) and Bartel tolerance factor (t_{Bartel}) [29], while Fig. S1 shows the respective confusion matrices. The four formulae exhibit distinct predictive performance in terms of accuracy and F1-score for the experimental test dataset. t_{JESS} and t_{Bartel} achieve identical results in terms of accuracy and F1-score, while the baseline model $t_{\text{Goldschmidt}}$ performs worst, reaching an F1-score of 55.1%. In contrast, τ^* clearly outperforms all other approaches, attaining the highest accuracy and F1-score.

t_{JESS} , t_{Bartel} , and τ^* exhibit comparable recall values (82.6%), indicating a similar ability to correctly identify experimentally stable perovskites. However, t_{JESS} and t_{Bartel} produce a substantial number of false positives, corresponding to unstable perovskite-type compositions incorrectly classified as stable; such errors are particularly undesirable, as they can lead to unnecessary synthesis and characterization efforts.

τ^* significantly mitigates this limitation by reducing the number of false positives to a single case, while preserving the same recall as t_{JESS} and t_{Bartel} . This improvement results in a marked increase in precision (95.0%) and specificity (97.1%), indicating a much higher reliability of stabil-

ity predictions. Figure 2d shows the elemental distribution of the 181 predicted perovskites from the full chemical space of 1392 possible $AB\text{X}_3$ chalcogenide compositions. This low fraction is consistent with experimental observations, as only a limited number of chalcogenide compounds have been reported to crystallize in the perovskite-type structure, with many candidate compositions instead favoring non-perovskite phases. As expected, more sulfur compounds are predicted to be stable than selenium-based materials, consistent with the larger ionic radius of Se. Predictions for $AB\text{Se}_3$ compositions should therefore be interpreted with caution, as experimentally reported selenide perovskites are underrepresented in the training dataset, limiting the statistical confidence of the learned stability boundary for this subset. In addition, common observed elements in the chalcogenide perovskite literature are well represented including Hf, Zr, Ba, Eu, and U.

Unlike the classical Goldschmidt tolerance factor, τ^* incorporates higher-order and mismatch-dependent terms. A defining feature of τ^* is its intrinsic asymmetry with respect to the A- and B-site cations. The A-site radius r_A enters both linearly and through a dominant cubic term $(r_A/r_X)^3$, while the B-site radius r_B enters linearly and logarithmically via a size-mismatch term. This asymmetry reflects the distinct structural roles of the two cation sites: the A-site cation controls the size and rigidity of the anion framework, whereas the B-site cation primarily determines the dimensions of the corner-sharing BX_6 octahedra.

When evaluated over the entire chemically accessible range of ionic radii, the τ^* expression admits two analytic branches, separated by the condition $\frac{r_B}{r_X} = \ln\left(\frac{r_A}{r_B}\right)$. On one branch, τ^* retains an explicit linear dependence on r_B ; on the other, the linear B-site contributions cancel, leaving only a logarithmic dependence on the B-site radius. Figure 3 shows color maps of τ^* across the full (r_A, r_B) chem-



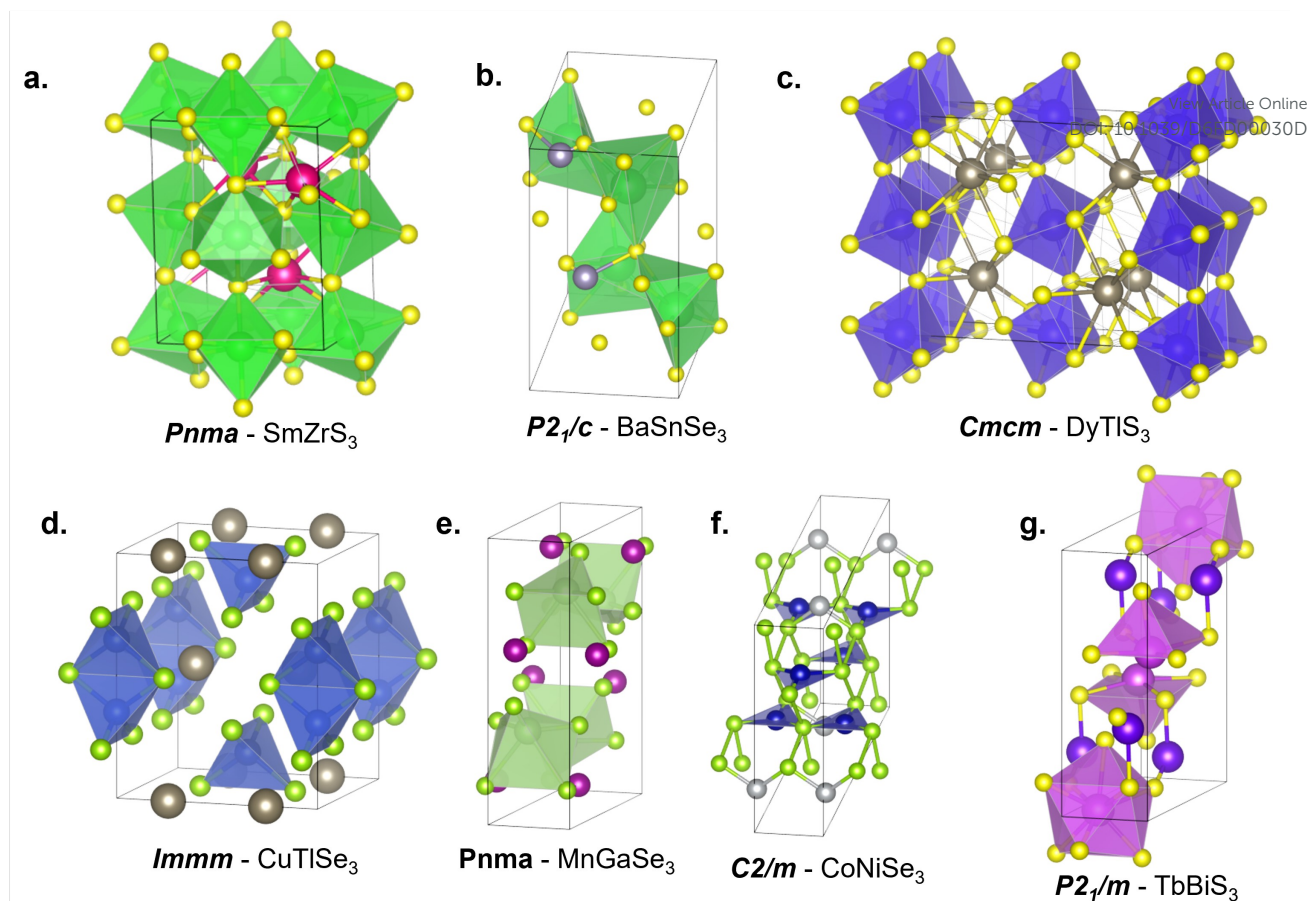


Figure 4: Some of the generated crystal structures with *CrystaLLM*, and their respective space group, based on the formulae that are predicted stable with τ^* including some specific compounds as examples: a. SmZrS_3 , b. BaSnSe_3 , c. DyTlS_3 , d. CuTlSe_3 , e. MnGaSe_3 , f. CoNiSe_3 , g. TbBiS_3 . Yellow circles correspond to the chalcogenide anions, while the other colors are for the A and B cations.

ical space for both anions. In our screened radius range, the stability region occurs primarily for $r_A > r_B$, which lies within the latter branch. Consequently, although linear B-site contributions are present in the general mathematical form of τ^* , they largely cancel within the stability-allowed region.

Within the stability-allowed region of the training and screening data, τ^* exhibits stronger sensitivity to the A-site ionic radius, the cubic term $(r_A/r_X)^3$ introduces a sharp upper bound on r_A in the descriptor, beyond which τ^* increases rapidly. In contrast, variations in the B-site radius contribute more gradually through the logarithmic mismatch term. As a result, r_B acts as a fine-tuning parameter, while r_A largely governs whether a perovskite-type structure is geometrically feasible within this model. The resulting stability region defined by τ^* therefore takes the form of a narrow wedge in the (r_A, r_B) space. Fig. S2 shows an elemental matrix color-coded with the logistic-calibrated stability probability derived from τ^* .

Structural validation

To further improve the precision of our screening, we employ an ensemble strategy that combines the τ^* criterion with explicit structure generation using *CrystaLLM*. While τ^* identifies compositions that are geometrically and chem-

ically plausible perovskites, *CrystaLLM* provides an independent assessment of whether these compositions can realistically adopt a perovskite-type structure based solely on their chemical formula, without resorting to computationally expensive first-principles calculations [27].

Crystal structures were generated for each of the 181 chalcogenide compositions satisfying $\tau^* < 0.846$. At this stage, density functional theory (DFT) verification was intentionally not performed, as *CrystaLLM* is trained on DFT-derived structures [27]. Consequently, the generated structures are not guaranteed to correspond to fully relaxed energetic minima at the DFT level of theory. These generated structures (see [Data & Code Availability](#)) were further analyzed to determine whether they adopted the true perovskite-type topology characterized by a corner-sharing BX_6 octahedral network. Non-perovskite structures were filtered by analyzing the a–b–c lattice parameters of the generated crystal structures, where clusters corresponding to edge-sharing and corner-sharing motifs within the *Pnma* space group were observed.

To further assess the physical plausibility of the generated *Pnma* structures, we performed structure relaxation using the UMA (Universal Models for Atoms, checkpoint uma-s-1p2) machine-learned interatomic potential (MLIP) [30], which has been trained on DFT data and represents an approximation to DFT-level calculations. Fig. S3 shows



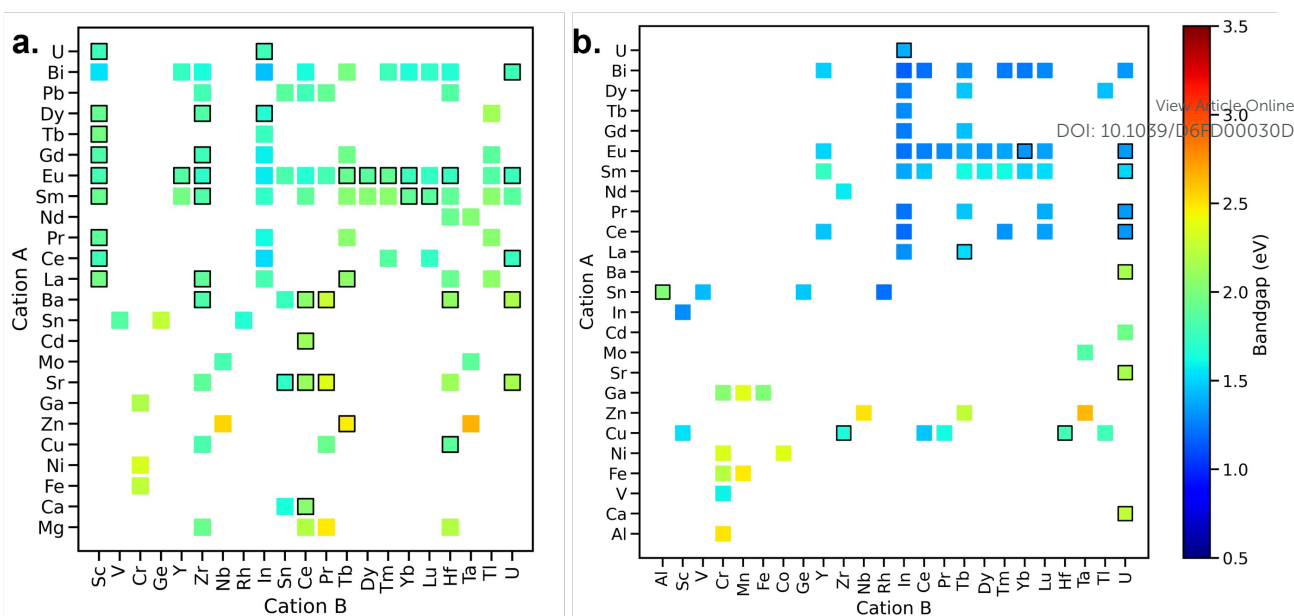


Figure 5: CrabNet-estimated bandgaps for predicted chalcogenide perovskites displayed on an element–element matrix for a. ABS_3 and b. $ABSe_3$ compositions. Color indicates the predicted bandgap value, while squares outlined in black correspond to compounds predicted to adopt a corner-sharing perovskite-type structure according to *CrystaLLM*.

that most structures are already close to relaxed configurations, as evidenced by their proximity to the parity line (with a mean energy difference of 0.104 eV/atom). Atomic positions were relaxed using the FIRE optimizer until the maximum residual force on any atom fell below 0.05 eV/Å, while keeping the lattice parameters fixed. This supports the overall robustness of the generated *CrystaLLM* structures. Moreover, there are some outliers such as $CuZrSe_3$, $CuHfS_3$, $BiUS_3$, $EuYbSe_3$, and $EuYbS_3$, which may reflect underlying chemical instabilities.

Requiring agreement between both τ^* and *CrystaLLM* models reduces the candidate set to 54 compositions, corresponding to approximately 30% of the τ^* -stable subset and about 4% of the original chemical space of 1392 possible compounds. This reduction reflects an intentional trade-off between recall and precision, because our ensemble prioritizes candidates supported by both geometric and structural hypotheses, giving a more experimentally actionable set of materials. Higher precision, or confidence in the predicted candidates, is favorable for screening tasks, especially for this type of materials as synthesis efforts have reported challenges related to byproducts and formation of non-perovskite phases [1, 31]. In addition, Fig. S4 presents parity comparisons of the lattice parameters (a , b , c) and unit-cell volume between the generated structures and ICSD-reported experimental data, showing close quantitative agreement.

The distributions of normalized ionic radii for the 54 *CrystaLLM*-validated perovskites reveal a significantly narrower stability window than the one defined by the tolerance factor alone. The A-site ratio r_A/r_X is strongly concentrated toward large values, with a mean of 0.84 ± 0.09 ($r_A = 155 \pm 17$ pm for ABS_3 compounds), indicating that successful formation of chalcogenide perovskites requires near-optimal filling of the anion framework by the A-site cation. Only a small number of compounds exhibit sub-

stantially smaller A-site cations, and the upper bound of the distribution approaches the geometric cutoff imposed by the cubic A-site term in τ^* . In contrast, the B-site ratio r_B/r_X is confined to a narrow interval centered around 0.48 ± 0.04 ($r_B = 88 \pm 7$ pm for ABS_3 compounds). Both smaller and larger B-site cations are systematically eliminated during structure generation, despite being permitted by the tolerance-factor criterion, indicating that only a limited range of octahedral sizes is compatible with stable corner-sharing connectivity in chalcogenide systems.

Further insight into this refinement is obtained by examining the space-group symmetries of the 181 τ^* -stable compounds. Figure 4 shows some of the generated structures, which span several low-symmetry space groups, including 116 in $Pnma$, 46 in $Cmcm$, 11 in $C2/m$, 4 in $P2_1/c$, 3 in $P2_1/m$, and 1 in $Immm$. The orthorhombic $Pnma$ space group dominates this set, consistent with its well-established role as a common distorted derivative of the ideal cubic perovskite-type structure, either as corner- or edge-sharing octahedral networks.

These results demonstrate that the τ^* tolerance factor efficiently narrows the chemical space to geometrically plausible candidates, while *CrystaLLM* structure generation resolves topological competition and isolates the small subset of compounds that reliably realize the perovskite-type structure. These 54 candidates include already reported compounds like $A\text{ScS}_3$ ($A = \text{Ce, Dy, Gd, La, Pr, Sm, Tb}$), BaBS_3 ($B = \text{Hf, U, Zr}$), EuBS_3 ($B = \text{Hf, Zr}$) and SrSnS_3 , which is reported as a non-perovskite phase [32]. Interestingly, there are computational studies that describe the potential of this material in the perovskite-type structure as an absorber layer in solar cells [8, 33]. Moreover, new chalcogenide perovskites are predicted to form stable perovskite-type phases including several Uranium-based compounds, AZrS_3 ($A = \text{Dy, Gd, La, Sm}$), EuBS_3 ($B^{3+} = \text{Dy, Tb, Tm, Yb}$), BaCeS_3 , EuScS_3 , DyInS_3 , among others.



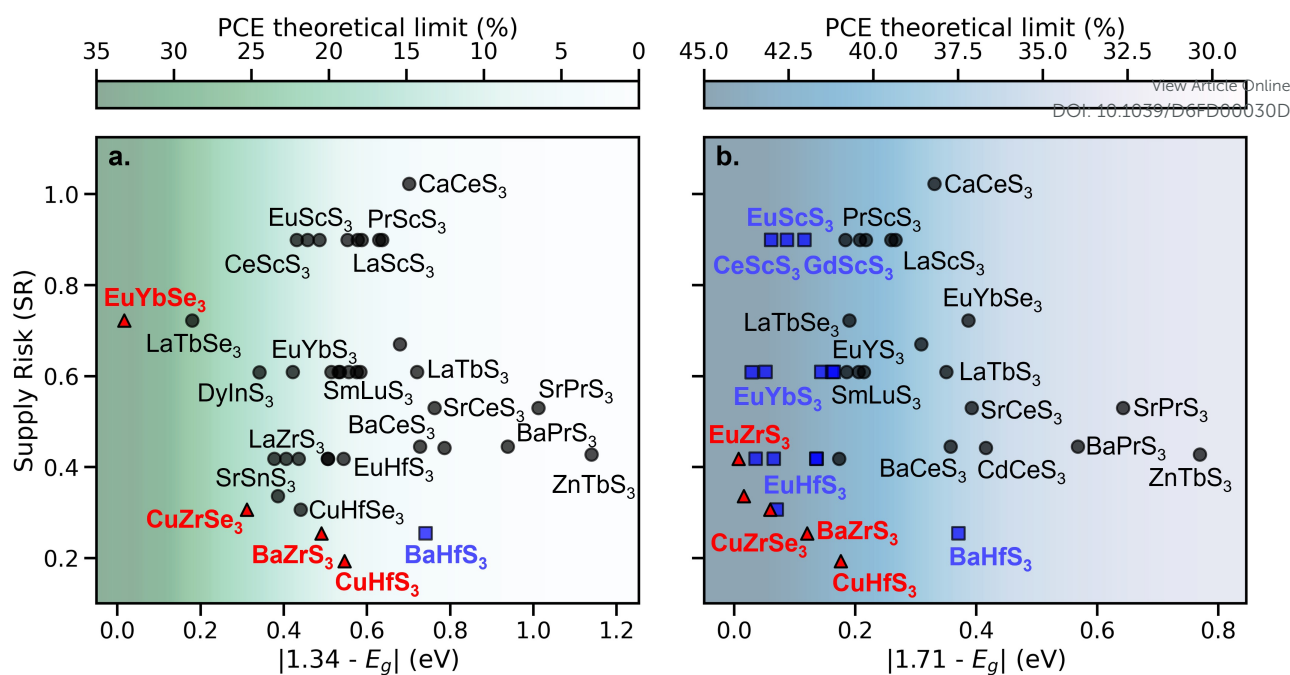


Figure 6: Evaluation of material sustainability using the supply risk (SR), derived from the Herfindahl-Hirschman Index (HHI) and ESG scores, as a function of the deviation of the estimated bandgap from the optimal values for (a) single-junction photovoltaics ($E_g^{opt} = 1.34$ eV) and (b) the top cell in a tandem photovoltaic configuration ($E_g^{opt} = 1.71$ eV). Red triangles indicate Pareto-optimal materials that simultaneously minimize the bandgap deviation $|E_g - E_g^{opt}|$ and SR. Blue squares correspond to materials lying within 10% of the minimum values of both optimization objectives. The background color indicates the theoretical limit in power conversion efficiency (PCE) based on the Shockley-Queisser limit [34].

Suitability for PV applications

To assess the suitability of the 54 candidates for PV applications, two indicators were studied. Firstly, the bandgap of each material was estimated using a composition-based regression algorithm [24]. Secondly, a sustainability factor was estimated with indices based on the elemental world production (Herfindahl-Hirschman Index, HHI) and environmental, social and governance (ESG) scores. Inspired by Nominé *et al.* a supply risk was determined for each composition [35].

Bandgap estimation was possible by training a CrabNet model based on reported experimental data from halide perovskites, chalcogenide semiconductors, and chalcogenide perovskites. This dataset was chosen to have an equilibrium between the perovskite-type structure knowledge and the chalcogenide effect. In addition, training solely with experimental chalcogenide perovskites was not feasible as the reported data is limited. The influence of training dataset size on model performance is shown in Fig. S5, illustrating convergence of the mean absolute error as the dataset grows.

Evaluation on the held-out test set yielded a mean absolute error (MAE) of 248 meV for the bandgap. The corresponding parity plot is shown in Fig. S6. This is consistent with the validation performance and demonstrates stable generalization to unseen compositions drawn from the same chemical distribution. Notably, the test MAE did not show a significant change when using different

element encoders, including Magpie [36], Mat2Vec [37], one-hot, Pettifor [38] descriptors and even randomly initialized embeddings (Fig. S7). The near-invariance of the final performance to the choice of encoder indicates that CrabNet's attention mechanism learns chemically meaningful element representations directly from composition during training, rapidly overriding the specific priors encoded in the initial embeddings [39]. In this regime, the dominant limitation is not feature expressiveness but rather the absence of structural information and the inherent noise in experimental bandgap measurements, which depend on synthesis conditions, measurement techniques, temperature, defects, and sample morphology.

Importantly, the achieved MAE falls within the range typically reported for existing benchmarks for composition-only prediction of experimental bandgaps. Prior studies on standardized benchmarks typically report MAEs in the range of 0.3–0.5 eV for the Matbench v0.1 experimental dataset when crystal structure information is not considered [40–42]. We emphasize that the absolute MAE obtained for property estimation is extremely dataset dependent, reflecting differences in chemical space, data curation, and experimental uncertainty. As such, the comparison to prior benchmarks is not intended as a strict performance evaluation, but rather to contextualize the present results within the typical accuracy range reported for composition-only models trained on experimental bandgap data.

Figure 5 shows elemental matrices with the predicted bandgap color coded for all τ^* stable compounds, with a



black edge for the *CrystaLLM* predicted perovskite-type structures. Narrow bandgaps (blue squares in Fig. 5) are largely filtered out by the *CrystaLLM* step. As a result, the subset of compounds classified as perovskite-type structures exhibits a bandgap distribution with a mean of 1.86 eV. This trend is consistent with literature reports, as many chalcogenide perovskites exhibit relatively wide bandgaps. For example, the well-known chalcogenide perovskites BaZrS₃ and BaHfS₃ are predicted to have bandgaps of 1.8 ± 0.4 eV and 2.1 ± 0.5 eV, respectively, in agreement with reported experimental values [1, 4, 9, 43]. In contrast, the predicted bandgap of LaScS₃ (2.0 ± 0.5 eV) underestimates a recently reported experimental value of 2.9 eV [44], even though it was included in the training dataset. While a single discrepancy does not allow for a quantitative assessment of model accuracy, these results support the use of the model as a rapid screening tool for identifying photovoltaic-relevant candidates rather than as a substitute for high-level theoretical or experimental methods. Notably, this experimental study does not report an error associated with the measurement, and strong absorption is observed in the data at photon energies approaching 2 eV [44]. Thus further corroboration of this value may be required.

To further illustrate the chemical origin of the CrabNet bandgap predictions, we performed a principal component analysis (PCA) shown in Fig. S8. The first two principal components, which together account for more than 60% of the bandgap variance, are primarily governed by descriptors associated with the B-site cation and the chalcogenide anion, emphasizing the dominant role of the BX₆ octahedral framework in defining the electronic structure of chalcogenide perovskites. This trend is consistent with established electronic-structure models of transition metal-based perovskites, in which the valence and conduction band edges are primarily derived from X-anion p states and B-cation d states, respectively, while the A-site cation mainly influences the bandgap indirectly through structural distortions and lattice effects [1].

Sustainability analysis

Figure 6 illustrates the combined influence of optoelectronic performance and sustainability by plotting the supply risk (SR) against the deviation of the predicted bandgap from optimal photovoltaic values. Two device architectures were considered. For single-junction photovoltaics, the optimal bandgap was taken as $E_g^{opt} \approx 1.34$ eV, corresponding to the maximum theoretical efficiency predicted by the Shockley-Queisser limit [34]. In contrast, the typically wider bandgaps of chalcogenide perovskites make them more suitable as absorbers for top cells in tandem photovoltaic architectures. In this configuration, the top cell absorbs high-energy photons while remaining transparent to lower-energy photons transmitted to the bottom cell, thereby surpassing the single-junction efficiency limit. For tandem devices paired with silicon bottom cells, the optimal bandgap of the top absorber was taken as $E_g^{opt} = 1.71$ eV [45]. Uranium-based compounds were excluded from this analysis due to their inherent toxicity

and radioactivity, which prevent practical photovoltaic applications.

Pareto-optimal candidates (red triangles) identify materials that achieve the best compromise between electronic suitability and sustainability. Notably, the well-studied compound BaZrS₃ remains optimal for tandem photovoltaic configurations, consistent with its frequent identification as a benchmark absorber in prior experimental studies. In addition, several previously unexplored compositions, including CuHfS₃, CuZrSe₃, and EuYbSe₃, emerge as promising photovoltaic absorbers. Materials highlighted in blue (squares) delineate a practically attractive region of the design space, lying within 10% of the minimum values for both optimization objectives. This subset exhibits increased chemical diversity, including compounds containing rare-earth elements, consistent with their large ionic radii and favorable geometric compatibility with the perovskite-type structure. While materials such as EuYbS₃, CeScS₃, and EuScS₃ may be of fundamental interest, their relatively high supply risk may limit their feasibility for large-scale photovoltaic deployment.

Assessment of experimental plausibility

Although the 54 *CrystaLLM* candidates appear promising, their experimental realization may prove challenging. To address this, a final screening step was introduced to evaluate their experimental feasibility. High-throughput materials screening workflows often rely on thermodynamic descriptors, such as formation energy or energy above the convex hull, which are typically derived from DFT calculations and are therefore computationally demanding [46]. As an initial attempt, available thermodynamic data were collected from open databases including the Materials Project (MP) [47], following the approach of Chen *et al.* [48]. However, for the majority of the predicted and previously unexplored compositions, thermodynamic data were unavailable, underscoring a common limitation in evaluating virtual materials.

Recent work by Jung *et al.* has demonstrated that relative synthesizability likelihood can be effectively assessed using positive-unlabeled (PU) machine-learning approaches [28, 46, 49]. PU learning trains on known synthesized compounds as positive examples while treating unlabeled materials as potentially negative. This is particularly useful for synthesizability, since we know which compounds have been synthesized but do not have confirmed data on compounds that cannot be made. Compared to purely thermodynamic models, PU-based approaches exhibit fewer false positives, reflecting the influence of kinetic, chemical, and experimental factors beyond thermodynamic stability [50].

The most recent model proposed by Jung *et al.* converts crystallographic information files (CIFs) into textual representations and employs a fine-tuned GPT-4o-mini model to predict a crystal-likeness score (CLS), which provides a data-driven proxy for the relative synthesizability. This model reported strong performance on general inorganic datasets [50]. However, for perovskite-specific materials, a graph convolutional neural network (GCNN) has been shown to outperform language-based approaches, likely



Table 2: Data derived for *CrystaLLM*-predicted stable, true perovskite-type compositions, excluding U-containing compounds. The compounds are sorted by CLS and the bold compounds have already been seen experimentally. HHI corresponds to the Herfindahl-Hirschman Index, SR is the Supply Risk. The bandgap and its standard deviation were predicted using a trained CrabNet model, while CLS is the crystal-likeness score, which is a data-driven proxy for relative synthesizability derived from a pre-trained GCNN model.

Formula	τ^*	Bandgap (eV)	HHI	SR	CL score
BaZrS₃	0.820	1.830 ± 0.463	0.485	0.254	0.991 ± 0.024
EuScS ₃	0.817	1.797 ± 0.491	1.586	0.899	0.988 ± 0.019
BaHfS₃	0.820	2.081 ± 0.511	0.485	0.254	0.987 ± 0.025
DyScS₃	0.809	1.917 ± 0.377	1.586	0.899	0.985 ± 0.017
CeScS₃	0.821	1.771 ± 0.442	1.586	0.899	0.982 ± 0.031
EuZrS₃	0.840	1.717 ± 0.417	0.778	0.418	0.981 ± 0.030
SrSnS ₃	0.822	1.726 ± 0.385	0.581	0.336	0.981 ± 0.036
GdScS₃	0.820	1.826 ± 0.347	1.586	0.899	0.981 ± 0.043
LaScS₃	0.818	1.969 ± 0.505	1.586	0.899	0.975 ± 0.082
EuHfS₃	0.840	1.745 ± 0.429	0.778	0.418	0.973 ± 0.067
GdZrS ₃	0.844	1.775 ± 0.325	0.778	0.418	0.971 ± 0.091
EuTmS ₃	0.775	1.916 ± 0.343	1.076	0.609	0.959 ± 0.111
EuTbS ₃	0.757	1.924 ± 0.302	1.076	0.609	0.957 ± 0.106
TbScS₃	0.819	1.976 ± 0.349	1.586	0.899	0.956 ± 0.068
LaZrS ₃	0.841	1.883 ± 0.462	0.778	0.418	0.956 ± 0.114
LaTbS ₃	0.793	2.060 ± 0.225	1.076	0.609	0.955 ± 0.100
DyZrS ₃	0.833	1.845 ± 0.350	0.778	0.418	0.950 ± 0.125
EuYS ₃	0.796	1.854 ± 0.301	1.076	0.609	0.944 ± 0.109
LaTbSe ₃	0.807	1.520 ± 0.214	1.281	0.723	0.934 ± 0.135
EuDyS ₃	0.838	1.872 ± 0.319	1.076	0.609	0.932 ± 0.118
PrScS₃	0.822	1.893 ± 0.511	1.586	0.899	0.927 ± 0.180
BaPrS ₃	0.820	2.278 ± 0.298	0.783	0.445	0.922 ± 0.166
SmZrS ₃	0.840	1.846 ± 0.314	0.778	0.418	0.915 ± 0.182
SmLuS ₃	0.755	1.875 ± 0.232	1.076	0.609	0.914 ± 0.184
CdCeS ₃	0.825	2.126 ± 0.391	0.792	0.442	0.901 ± 0.208
BaCeS ₃	0.799	2.067 ± 0.518	0.783	0.445	0.894 ± 0.200
SmScS₃	0.817	1.927 ± 0.353	1.586	0.899	0.880 ± 0.230
SmYbS ₃	0.817	1.896 ± 0.334	1.076	0.609	0.827 ± 0.261
DyInS ₃	0.844	1.681 ± 0.205	1.083	0.609	0.816 ± 0.251
SrPrS ₃	0.833	2.352 ± 0.385	0.925	0.530	0.760 ± 0.309
EuYbS ₃	0.817	1.762 ± 0.467	1.076	0.609	0.748 ± 0.276
SnAlSe ₃	0.829	2.019 ± 0.292	1.187	0.670	0.717 ± 0.326
CaCeS ₃	0.825	2.041 ± 0.508	1.805	1.023	0.703 ± 0.305
SrCeS ₃	0.812	2.102 ± 0.612	0.925	0.530	0.665 ± 0.324
EuYbSe ₃	0.815	1.323 ± 0.257	1.281	0.723	0.607 ± 0.371
CuHfS ₃	0.811	1.886 ± 0.613	0.383	0.193	0.585 ± 0.344
CuHfSe ₃	0.822	1.780 ± 0.596	0.589	0.307	0.570 ± 0.360
ZnTbS ₃	0.835	2.479 ± 0.295	0.759	0.428	0.560 ± 0.358
CuZrSe ₃	0.822	1.651 ± 0.588	0.589	0.307	0.518 ± 0.337

due to the relatively simple and repetitive connectivity of the perovskite-type structure [28]. Accordingly, the GCNN model developed by Gu *et al.* was employed in this work.

The GCNN was pre-trained on experimental structures from the ICSD and virtual perovskite-type crystals from MP, encompassing a broad range of chemistries including oxides, halides, hydrides, nitrides, and chalcogenides. The model CLS output can be seen as a statistical synthesizability indicator, where higher values indicate a greater likelihood of successful synthesis; a threshold of CLS = 0.5 has been proposed to identify synthetically plausible perovskites [28]. No additional fine-tuning was performed,

as the model is explicitly designed to generalize across perovskite-type chemistries. Crystal structures of the candidate set generated by *CrystaLLM* were therefore used directly as input to the GCNN.

Chalcogenide candidates were ranked by CLS, most of them with values exceeding 0.5, which indicates greater statistical similarity to known synthesized crystals. Importantly, this indicator is not used to decide whether a perovskite structure exists, but to rank how the generated structures resemble experimentally realized perovskites, providing a complementary signal to geometric and topological screening. Moreover, all experimentally reported



chalcogenide perovskites exhibit CLS values above 0.88, providing additional validation of the model for this class of materials and suggesting an empirical threshold for high synthesizability likelihood. For example, well-known Ba-based sulfide perovskites with B-site cations such as Zr, U, and Hf display CLS values of 0.99 ± 0.02 .

Applying this stricter criterion reduces the candidate pool to approximately 30 compounds exhibiting high similarity to known experimentally reported materials. In terms of compositional trends, higher CLS values are predominantly observed for sulfides, Ba- and La-based A-site cations, and Sc- and Tm-based B-site cations, whereas lower CLS values are associated with selenides, Cu- and Ce-based A-site compositions, and Yb- and Tb-based B-site cations.

Figure 7 summarizes the combined assessment of the statistical synthesizability indicator, sustainability, and optoelectronic suitability for the most promising candidates. The figure reports CLS and SR values, with the predicted bandgap encoded by color and Pareto-optimal materials highlighted for single-junction (triangles) and tandem (squares) configurations. BaZrS₃ emerges as the most promising candidate for tandem applications. While this compound was included in the training dataset and thus cannot serve as an independent validation, its high ranking, together with the statistical analysis presented above, suggests that the individual models capture largely non-redundant aspects of material suitability. Furthermore, a Spearman rank correlation analysis over the 54 candidates confirms that pairwise correlations among the four screening metrics are uniformly weak ($|\rho| \leq 0.27$); only the E_g -CLS pair reaches marginal significance ($\rho = 0.27$, $p < 0.05$), while all remaining pairs are statistically non-significant (Fig. S9). This near-orthogonality indicates that each stage of the pipeline contributes independent discriminating information, reducing the likelihood that the ranking is governed by a latent common variable. Instead geometric stability, statistical synthesizability, sustainability, and optoelectronic suitability appear to capture fundamentally different chemical drivers rather than recapitulating the same compositional trend.

Table 2 summarizes the candidates along with their predicted bandgaps, CLS values, and estimated HHI and supply risks. These candidates should be interpreted as prioritized targets for experimental validation rather than definitive predictions of optoelectronic properties or thermodynamic stability. For instance, several yet-to-be-synthesized interesting compositions include EuScS₃, LaTbS₃, EuYbS₃, EuYbSe₃, CuHfS₃, CuHfSe₃, CuZrSe₃ (ranked by CLS). Nonetheless, we would like to acknowledge that from a chemical perspective, some of these compositions may face additional challenges related to oxidation-state stability and charge balance, as pymatgen oxidation-state assignments may not fully reflect redox thermodynamics. For instance, Cu is frequently stabilized in the +1 oxidation state (soft cation) in sulfides and selenides (soft bases), rendering Cu²⁺-based chalcogenide perovskites potentially susceptible to reduction or mixed-valence behavior based on the hard and soft acids and bases (HSAB) theory [51]. Similarly, Ce⁴⁺ is strongly stabilized in oxides such as CeO₂ but is rarely observed in sulfide environments, where Ce³⁺

phases dominate [52]. While the present screening workflow does not explicitly model redox stability, these considerations highlight the importance of chemical feasibility constraints and underscore the need for experimental verification of the most promising candidates.

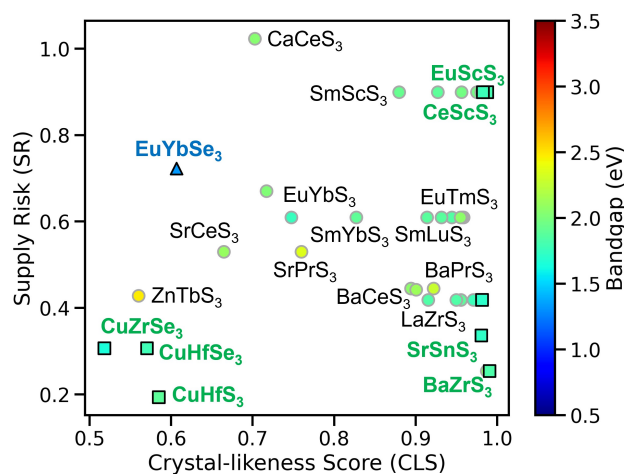


Figure 7: Supply risk (SR) as a function of the Crystal-likeness Score (CLS). The latter is a data-driven proxy for the relative synthesizability of the compound based on a pre-trained GCNN model [28]. The color code corresponds to the bandgap predicted by CrabNet [39]. Squares and triangles indicate Pareto-optimal materials that simultaneously minimize SR, 1-CLS, and the bandgap deviation for tandem configurations ($E_g^{opt} = 1.71\text{eV}$) and single junctions ($E_g^{opt} = 1.34\text{eV}$), respectively.

Beyond the specific materials identified here, the proposed workflow illustrates a general and transferable strategy for data-driven materials screening in chemically complex and experimentally underexplored spaces. By combining interpretable analytical descriptors derived from experimental data, generative crystal structure prediction, composition-based property estimation, and experimental plausibility assessment, the pipeline enables systematic navigation of large compositional spaces while minimizing reliance on computationally intensive first-principles calculations. Importantly, each stage of the workflow produces physically interpretable outputs, allowing chemical intuition to be retained alongside machine-learning predictions.

The integration of sustainability metrics further extends the scope of the screening beyond traditional performance-driven criteria. These results highlight the importance of incorporating sustainability metrics at early stages of materials screening to better assess long-term technological viability, particularly for energy applications where material availability and supply-chain considerations can be as limiting as intrinsic performance. More broadly, this multi-objective framework illustrates how experimental knowledge, machine-learning models, and external socio-economic datasets can be combined to guide the identification of viable functional materials prior to experimental synthesis.



Conclusions

We present a data-driven framework to prioritize chalcogenide perovskite candidates by combining an analytically derived tolerance factor, generative crystal-structure prediction, composition-based bandgap estimation, and sustainability-oriented metrics. Rather than relying on a single screening criterion, the proposed workflow distributes decision-making across partially independent hypotheses, enabling structured reduction of chemical space while retaining physical interpretability at each stage.

The SISO-derived tolerance factor τ^* significantly reduces false positives compared to classical geometrical descriptors while preserving recall for experimentally reported perovskites. Its intrinsic asymmetry between the A and B-site radii provides chemically meaningful insight into the geometric constraints governing chalcogenide perovskites formation. Subsequent structure generation with *CrystalLM* resolves topological ambiguities within the tolerance-factor-stable region, refining the candidate pool to compounds capable of forming true corner-sharing BX_6 networks. Finally, integration of bandgap estimation, supply risk metrics, and crystal-likeness scoring enables multi-objective ranking tailored to photovoltaic applications.

The resulting candidates should be interpreted as prioritized targets for experimental validation rather than definitive predictions of thermodynamic stability or optoelectronic performance. In particular, the limited representation of selenide perovskites in the experimental dataset introduces additional uncertainty for $ABSe_3$ predictions. Nevertheless, the convergence of independent screening criteria and the weak pairwise correlations among them indicate that each stage contributes complementary information rather than reiterating the same chemical trends.

Beyond the specific compositions identified, this study illustrates how experimentally grounded descriptors, generative models, and socio-economic indicators can be integrated into a transparent and transferable screening strategy for chemically constrained materials spaces. Such approaches may help bridge the gap between theoretical prediction and experimental realization, particularly in underexplored material families where data scarcity and competing phases complicate discovery.

Methods

Data acquisition

Experimental data were compiled from multiple literature and database sources to construct datasets for structural stability, electronic properties, experimental feasibility, and sustainability analysis (See the documentation in [Data & Code Availability](#)). For the determination of the tolerance factor τ^* , the materials dataset consisted of experimentally reported ABX_3 compounds with X anions being halide or chalcogenides. Halide perovskite data were taken from Bartel *et al.* [29], while experimentally reported chalcogenide perovskites were collected from the work of Sopiha *et al.* [1]. In addition to compounds adopting the perovskite-

type structure, experimentally reported ABX_3 compositions crystallizing in non-perovskite structure types were included as negative examples, enabling discrimination between phases within the same stoichiometry. Only oxygen-free compounds were considered throughout this study in order to focus on non-oxide perovskite-type systems. Halide compounds were included to increase the dataset size, while accepting the risk of bias toward ionic bonding characteristics.

Ionic radii were obtained from a human-curated dataset reported by Turnley *et al.*, based on Shannon radii with extensions tailored to metal sulfides [18]. This dataset includes extrapolated values for missing entries and provides improved descriptions of metal-chalcogen bond distances, which are not adequately captured by conventional oxide-based radii tables. The use of this dataset reflects the increased covalent character of metal-chalcogen bonds relative to oxide and halide analogues.

Experimental bandgap data were collected from three complementary sources: hybrid organic-inorganic halide perovskites from the NOMAD API [53, 54], chalcogenide perovskites from the dataset of Sopiha *et al.* [1], and a large set of chalcogenide semiconductors extracted using ChemDataExtractor [55]. Combining these sources enabled the construction of a broad, experimentally grounded dataset spanning both perovskite-type and chalcogenide materials, which is particularly valuable given the limited availability of chalcogenide perovskite data.

Sustainability indicators were derived from publicly available datasets describing global commodity production and socio-environmental risk. Commodity production shares were taken from the US Geological Survey 2025 database [56], environmental, social, and governance (ESG) indicators from the World Bank [57]. These datasets were combined to provide comparative estimates of material supply risk.

Tolerance factor determination

A total of 283 experimentally reported ABX_3 compositions were used to develop and evaluate the tolerance factor τ^* . The dataset was divided into an 80% training set and a 20% test set, with stratification to preserve similar distributions of anion types and experimentally stable perovskite-type phases across both subsets. Chalcogenide perovskites constitute approximately 21% and 5% of the training and test datasets for sulfides and selenides, respectively, corresponding to 27 experimentally reported compounds (26 ABS_3 and one $ABSe_3$). As a result, predictions for selenide perovskites are expected to carry a large intrinsic uncertainty.

Primary features were derived from elemental properties of the A, B, and X species, including atomic radii (r_A , r_B , r_X), electronegativity (χ_A , χ_B , χ_X), electron affinity (EA_A , EA_B , EA_X), ionization potential (IP_A , IP_B , IP_X), nuclear charge (Z_A , Z_B , Z_X), and oxidation state (n_A , n_B , n_X). In addition, computed energies of the highest occupied and lowest unoccupied atomic orbitals (E^{HOMO} and E^{LUMO}) and orbital-resolved radii corresponding to the maxima of the radial distribution functions of s and p orbitals were included. Consistent with established perovskite screening approaches, additional derived features



based on ratios of atomic radii and electronegativity differences were constructed, such as the relative change in cation-anion electronegativity difference $\Delta\chi$ relative to oxygen introduced by Jess *et al.* [17, 29].

An initial feature selection was performed using a random forest classifier with Gini importance as the ranking criterion [58]. Given the limited dataset size, the train-test split was randomized, and a grid search combined with cross-validation was used to optimize model hyperparameters with respect to the F_1 -score. This procedure was repeated multiple times to assess robustness, and features were ranked according to their average normalized importance. The five most predictive features identified through this process were:

$$\Phi = \left[\log \left(\frac{r_A}{r_B} \right), \frac{\Delta\chi_{B-X}}{\Delta\chi_{B-O}}, \frac{r_B}{r_X}, \frac{r_A}{r_X}, EA_A \right].$$

These features were subsequently used as inputs to the sure independence screening and sparsifying operator (SISSO) algorithm [23]. The implementation followed the approach of Bartel *et al.*, employing an operator space consisting of 11 mathematical operations: $[+, -, | - |, \times, \nabla \cdot, \exp, \square^{-1}, \square^2, \square^3, \sqrt{\square}, \log]$ [29]. Four iterations of sure independence screening were performed, retaining 200 residuals per iteration, yielding a pool of approximately 3000 candidate descriptors.

A decision tree classifier with a maximum depth of one was employed as the sparsifying operator to select a single optimal descriptor from the candidate pool. This shallow-tree approach was chosen to preserve descriptor interpretability and reduce sensitivity to individual data points, particularly given the limited and heterogeneous nature of the experimental dataset. Classifiers were trained using the entropy criterion, with class weights adjusted according to class frequency to account for the underrepresentation of chalcogenide perovskites. This strategy acknowledges that experimental classification of perovskite versus non-perovskite is often ambiguous due to variations in synthesis conditions.

Finally, a compositional screening of candidate materials was performed by enumerating 1392 charge-balanced ABX_3 compounds formed from combinations of metals and chalcogenides satisfying the perovskite-type stoichiometry [59]. Oxidation states were assigned using the prediction scheme implemented in *pymatgen*, resulting in a final list of 181 compounds predicted to be structurally stable according to the derived τ^* criterion.

Crystal structure generation

The crystal structures of the 181 compounds predicted to be stable based on the τ^* criterion were generated from their chemical formulae using *CrystaLLM* [27]. The large pre-trained model was employed [60], and the number of formula units in the generated periodic cell was fixed to four, such that each ABX_3 composition was represented by a single generated $A_4B_4X_{12}$ structure. *CrystaLLM* is based on an autoregressive large language model trained to generate Crystallographic Information Files (CIFs) and has been trained on over three million compounds drawn from open-access density functional theory databases, including

the Materials Project [47], NOMAD[53], and OQMD [27, 61].

From the generated structures, 116 compounds were predicted to adopt the orthorhombic *Bma* space group, which is commonly associated with distorted perovskite-type structures. However, within this space group, the BX_6 octahedra may be arranged in either edge-sharing or corner-sharing configurations, with only the latter corresponding to the true perovskite-type structure. To distinguish between these motifs, the connectivity of the octahedral network was analyzed for each predicted structure. Only structures exhibiting a corner-sharing BX_6 framework were retained, resulting in a final set of 54 compounds with a generated perovskite-type crystal structure.

It is important to note that this classification relies on generative structural predictions rather than experimentally relaxed geometries, and therefore introduces an additional source of uncertainty. In particular, subtle distortions and alternative low-energy phases may not be fully captured at this stage. Consequently, the predicted crystal structures are intended to serve as screening-level indicators of perovskite-type structure rather than definitive structural assignments.

Bandgap estimation

Bandgaps were estimated using a composition-based machine-learning model trained on experimentally reported data. The initial dataset comprised 50 662 reported materials, including hybrid organic-inorganic halide perovskites, chalcogenide perovskites, and chalcogenide semiconductors [1, 53, 55, 62]. The latter ensures diverse data that encourages generalizable composition-property relationships rather than narrow domain fitting. Materials with reported bandgaps below 1 eV were excluded to focus on semiconductors relevant for photovoltaic applications. In addition, for each composition, reported values more than 0.3 eV below the median bandgap were discarded to suppress anomalously low values likely related to DFT-derived quantities. Compounds were subsequently grouped by chemical formula, resulting in a curated dataset of 3 628 unique compositions, each assigned a representative bandgap given by the median of the reported values. This dataset was randomly divided into training, validation, and test subsets.

A CrabNet model [24] was trained using several compositional encoding schemes including *magpie*, *mat2vec*, *onehot*, *random_200* and *Pettifor* descriptors [36–39, 63]. The model architecture consisted of three attention layers with an embedding dimension of 512 and four attention heads, resulting in approximately 12 million trainable parameters. Training was performed on a CPU using a batch size of 128 and a cyclical learning rate schedule, with early stopping enabled based on validation mean absolute error (MAE). Finally, model performance was evaluated on the held-out test set, yielding a mean absolute error (MAE) of 248 meV for a combined embedding representation with *mat2vec* and *Pettifor*. This combination was chosen as training solely with *mat2vec* embedding resulted in a MAE of 256 meV. For comparison, other models were trained only with the halide perovskites or chalcogenide semiconduc-



tors datasets and baseline models predicting the median or mean bandgap for all compositions were assessed. Table 3 shows the respective test-set MAE and R^2 for these models.

As conventional random train/test split for this case yields zero or one chalcogenide perovskite in the test set, any error estimate on the target chemical space is statistically unreliable. Thus to obtain a more rigorous estimate of the model's ability to extrapolate to unseen chalcogenide perovskites, we employ Leave-One-Out Cross-Validation (LOOCV): in each of the nine folds, one chalcogenide perovskite composition is withheld entirely from training, and CrabNet is retrained from scratch on the remaining corpus of ~ 3600 compounds (halide perovskites and chalcogenide semiconductors), then used to predict the bandgap of the held-out compound. The median experimental bandgap per formula, together with its inter-measurement standard deviation as a proxy for experimental uncertainty, is used as the reference value. This procedure yields a LOOCV MAE of 281 meV across the nine compositions, providing a more reliable estimate of the model's performance on the target class of materials.

Given the compositional nature of the model and the structural diversity of the training data, the predicted bandgaps are intended as approximate screening-level indicators rather than quantitative optoelectronic predictions. Nevertheless, this approach enables rapid identification of candidate materials with bandgaps compatible with photovoltaic applications within the broader screening workflow.

Table 3: Performance comparison of composition-based bandgap models and trivial baselines. Models were trained on different subsets of the experimental dataset. All metrics are reported for the held-out test set. MAE corresponds to the Mean absolute error, while R^2 is the coefficient of determination.

Model	n_{train}	MAE (meV)	R^2
Full dataset	3628	248	0.423
Halide-only	1263	103	0.406
Chalcogenides-only	1630	430	0.290
Median baseline	–	443	–
Mean baseline	–	472	–

Sustainability analysis

To assess the sustainability of the screened chalcogenide perovskites, supply risk indicators were calculated following the methodology proposed by Nominé *et al.* [35]. Two complementary metrics were evaluated using publicly available datasets describing global commodity production and socio-environmental risk.

The Herfindahl-Hirschman Index (HHI) was used to quantify market concentration for each elemental commodity and was calculated as the sum of the squared country-level production shares. Production data were taken from the most recent US Geological Survey dataset (USGS 2025), corresponding to 2023 global production [56]. Commodity-level HHI values were subsequently mapped to elemental species within each candidate compound.

Environmental, social, and governance (ESG) risk was estimated using 18 normalized country-level indicators obtained from the World Bank [57], grouped into environmental, social, and governance categories with six indicators per category. These indicators serve as proxies for ESG conditions associated with raw material production and reflect aggregate national-level trends rather than site-specific mining practices. The selected indicators are listed below.

- **Social score:** average of indicators including internet usage, labor force participation rate, life expectancy at birth, proportion of seats held by women in national parliaments, school enrollment, and unemployment.
- **Environmental score:** average of indicators including value added by agriculture, forestry, and fishing; electricity production from coal; net energy imports; per capita energy use; fossil fuel energy consumption; and extent of terrestrial and marine protected areas.
- **Governance score:** average of World Bank governance estimates for control of corruption, government effectiveness, political stability and absence of violence or terrorism, regulatory quality, rule of law, and voice and accountability.

The average ESG score was combined with country-level production shares as weights in the HHI formulation to estimate an overall supply risk for each elemental commodity. For a given compound, the total supply risk was calculated as the sum of the elemental supply risks. These sustainability metrics provide a comparative framework for prioritizing candidate materials rather than absolute assessments of environmental impact.

Experimental plausibility assessment

Experimental plausibility was evaluated using a machine-learning approach based on positive-unlabeled (PU) learning, which has been shown to effectively capture experimental realizability by learning from known synthesized materials without requiring explicit negative examples. In this work, we employed a pre-trained graph convolutional neural network (GCNN) model developed by Gu *et al.* [28], specifically designed for perovskite-type crystal structures. GCNNs are particularly well suited for this task, as they encode local coordination environments and connectivity patterns that are central to perovskite-type chemistry, allowing structurally meaningful comparison across diverse compositions.

The model outputs a crystal-likeness score (CLS), which quantifies the likelihood that a given crystal structure resembles experimentally observed perovskite-type compounds. Higher CLS values correspond to greater experimental feasibility. Importantly, CLS is not a thermodynamic quantity and does not directly account for kinetic barriers, synthesis pathways, or processing conditions; rather, it serves as a comparative metric to prioritize candidate materials for experimental investigation.



Crystal structures generated by *CrystaLLM* were used directly as input to the GCNN. No geometry optimization or energy minimization was performed prior to evaluation, as the purpose of the synthesizability assessment is to analyze coordination features rather than energetic stability. No additional fine-tuning of the model was performed. A CLS value of 0.5 has previously been proposed as a threshold indicating synthetically plausible perovskite-type structures [28]; however, in this work CLS values are interpreted comparatively rather than as strict pass–fail criteria. Validation against experimentally reported chalcogenide perovskites indicates that known compounds consistently exhibit CLS values above 0.88, providing an empirical reference for high synthesizability likelihood within this class of materials. Accordingly, CLS was used as a ranking and filtering metric to guide candidate prioritization.

Author contributions

DAG: data curation, formal analysis, investigation, methodology, software, visualization and writing - original draft. LH: writing - review & editing. LA: supervision, writing - review & editing. SaS: supervision, writing - review & editing. JAM: conceptualization, software, visualization, methodology, supervision, writing - review & editing.

Conflicts of interest

There are no conflicts to declare.

Data & Code Availability

The curated datasets and code developed in this work are openly available at the GitHub repository [64]: <https://github.com/GarzonDiegoFEUP/chalcogenide-perovskite-screening>. An interactive version of the figures is available in the documentation at the GitHub repository. A citable archived version (v1.0.1) is deposited on Zenodo at <https://doi.org/10.5281/zenodo.18743650>.

Acknowledgements

DAG acknowledges the support by FCT - Fundação para a Ciência e Tecnologia, I.P. with the project reference 2023.00258.BD, and DOI identifier <https://doi.org/10.54499/2023.00258.BD>. Authors acknowledge the COST Action Research and International Networking project "Emerging Inorganic Chalcogenides for Photovoltaics (RENEW-PV)," CA21148, supported by COST (European Cooperation in Science and Technology).

References

- Sopiha, K. V., Comparotto, C., Márquez, J. A. & Scragg, J. J. S. Chalcogenide Perovskites: Tantalizing Prospects, Challenging Materials. *Advanced Optical Materials* **10**, 2101704 (2022).
- Green, M. *et al.* Solar Cell Efficiency Tables (Version 66). *Progress in Photovoltaics: Research and Applications* **33**, 795–810 (2025). View Article Online
DOI: 10.1039/D5PP00077D
- Chowdhury, T. A. *et al.* Stability of perovskite solar cells: issues and prospects. *RSC Advances* **13**, 1787–1810 (2023).
- Agarwal, S., Vincent, K. C. & Agrawal, R. From synthesis to application: a review of BaZrS₃ chalcogenide perovskites. *Nanoscale* **17**, 4250–4300 (2025).
- Basera, P. & Bhattacharya, S. Chalcogenide Perovskites (ABS₃; A = Ba, Ca, Sr; B = Hf, Sn): An Emerging Class of Semiconductors for Optoelectronics. *J. Phys. Chem. Lett.* **13**, 6439–6446 (2022).
- Chakravorty, A., Adhikari, S. & Johari, P. Unlocking the optoelectronic potential of AGeX₃ (A = Ca, Sr, Ba; X = S, Se): A sustainable alternative in chalcogenide perovskites. *Journal of Chemical Physics* **163**, 234708 (2025).
- Adhikari, S. & Johari, P. Optimizing lead-free chalcogenide perovskites for high-efficiency photovoltaics via alloying. *Phys. Rev. B* **112**, 085206 (2025).
- Adhikari, S., Das, S. & Johari, P. Post-transition metal Sn-based chalcogenide perovskites: a promising lead-free and transition metal alternative for stable, high-performance photovoltaics. *J. Mater. Chem. C* **13**, 7792–7805 (2025).
- Nishigaki, Y. *et al.* Extraordinary Strong Band-Edge Absorption in Distorted Chalcogenide Perovskites. *Solar RRL* **4**, 1900555 (2020).
- Liang, Y. *et al.* Parametric Study on Controllable Growth of SrZrS₃ Thin Films with Good Conductivity for Photodetectors. *Nano Research* **16**, 7867–7873 (2023).
- Lelieveld, R. & Ijdo, D. J. W. Sulphides with the GdFeO₃ Structure. *Acta Crystallographica Section B* **36**, 2223–2226 (1980).
- Comparotto, C., Ström, P., Donzel-Gargand, O., Kubart, T. & Scragg, J. J. S. Synthesis of BaZrS₃ Perovskite Thin Films at a Moderate Temperature on Conductive Substrates. *ACS Applied Energy Materials* **5**, 6335–6343 (2022).
- Comparotto, C. *et al.* Thermodynamic insights into the Ba–S system for the formation of BaZrS₃ perovskites and other Ba sulfides. *Journal of Materials Chemistry A* **13**, 9983–9991 (2025).
- Breternitz, J. & Schorr, S. What Defines a Perovskite? *Advanced Energy Materials* **8**, 1802366 (2018).
- Rose, G. Ueber einige neue Mineralien des Urals. *Journal für Praktische Chemie* **19**, 459–468 (1840).
- Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
- Jess, A., Yang, R. & Hages, C. J. On the Phase Stability of Chalcogenide Perovskites. *Chemistry of Materials* **34**, 6894–6901 (2022).
- Turnley, J. W., Agarwal, S. & Agrawal, R. Rethinking tolerance factor analysis for chalcogenide perovskites. *Materials Horizons* **11**, 4802–4808 (2024).



19. Huo, K. Z., Wei, S.-H. & Yin, W.-J. High-throughput screening of chalcogenide single perovskites by first-principles calculations for photovoltaics. *J. Phys. D: Appl. Phys.* **51**, 474003 (2018).
20. Cao, Y., Dai, S., Wang, X. & et al. High-throughput screening of potentially ductile and low thermal conductivity ABX_3 ($X = S, Se, Te$) thermoelectric perovskites. *Appl. Phys. Lett.* **124**, 092101 (2024).
21. Singh, N. *et al.* High-throughput and data-driven search for stable optoelectronic $AMSe_3$ materials. *J. Mater. Chem. A* **13**, 9192–9210 (2025).
22. Fatima, S. K. *et al.* Lead-free perovskites for next-generation applications: a comprehensive computational and data-driven review. *Materials Advances* **6**, 7634–7661 (2025).
23. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Materials* **2**, 083802 (2018).
24. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials* **7**, 77 (2021).
25. De Breuck, P.-P., Hautier, G. & Rignanese, G.-M. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj Computational Materials* **7**, 83 (2021).
26. De Breuck, P.-P., Wang, H.-C., Rignanese, G.-M., Botti, S. & Marques, M. A. Generative AI for crystal structures: a review. *npj Computational Materials* (2025).
27. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nature Communications* **15**, 10570 (2024).
28. Gu, G. H., Jang, J., Noh, J., Walsh, A. & Jung, Y. Perovskite synthesizability using graph neural networks. *npj Computational Materials* **8**, 71 (2022).
29. Bartel, C. J. *et al.* New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances* **5**, eaav0693 (2019).
30. Wood, B. M. *et al.* UMA: A Family of Universal Models for Atoms. *arXiv preprint arXiv:2506.23971* (2025).
31. Carr, A., Glinberg, T., Stull, N., Neilson, J. R. & Bartel, C. J. Origins of chalcogenide perovskite instability. *J. Mater. Chem. C* **13**, 19183–19195 (2025).
32. Yamaoka, S., Fukunaga, O. & Saito, S. Preparations of $BaSnS_3$, $SrSnS_3$ and $PbSnS_3$ at high pressure. *Mater. Res. Bull.* **5**, 789–794 (1970).
33. Ju, M.-G., Dai, J., Ma, L. & Zeng, X. C. Perovskite Chalcogenides with Optimal Bandgap and Desired Optical Absorption for Photovoltaic Devices. *Adv. Energy Mater.* **7**, 1700216 (2017).
34. Shockley, W. & Queisser, H. J. Detailed Balance Limit of Efficiency of p–n Junction Solar Cells. *Journal of Applied Physics* **32**, 510–519 (1961).
35. Nomine, A. *et al.* *The most sustainable high entropy alloys for the future* Preprint, Research Square. 2023.
36. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 16028 (2016).
37. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
38. Cerqueira, T. F. T., Wang, H., Botti, S. & Marques, M. A. L. A non-orthogonal representation of the chemical space. *arXiv preprint arXiv:2406.19761*. Accessed: 2025-10-15 (2025).
39. Wang, A. Y. T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. CrabNet for explainable deep learning in materials science. *Journal of Materials Research and Technology* **10**, 415–428 (2021).
40. Dunn, A., Wang, Q., Ganose, A., Dopp, D., Jain, A., *et al.* Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **6** (2020).
41. Breuck, P.-P. D., Heymans, G. & Rignanese, G.-M. Accurate experimental band gap predictions with multi-fidelity correction learning. *Journal of Materials Informatics* **2**, 10 (2022).
42. Vasylenko, A. *et al.* Digital features of chemical elements extracted from local geometries in crystal structures. *Digital Discovery* **4**, 477–485 (2025).
43. Vincent, K. C., Agarwal, S., Turnley, J. W. & Agrawal, R. Liquid Flux-Assisted Mechanism for Modest Temperature Synthesis of Large-Grain $BaZrS_3$ and $BaHfS_3$ Chalcogenide Perovskites. *Advanced Energy and Sustainability Research* **4** (2023).
44. Zhang, H. *et al.* P-Type Transparent Conducting Material Realized by Composite Thin Film of Chalcogenide Perovskite $LaScS_3$ and Graphene. *Advanced Functional Materials* **35**, 2542382 (2025).
45. Fell, A. *et al.* Elucidating the efficiency limit of silicon-based monolithic tandem cells through the combination of Auger and Shockley–Queisser limits. *EES Solar* **1**, 1030–1039 (2025).
46. Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning. *Journal of the American Chemical Society* **142**, 18836–18843 (2020).
47. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
48. Chen, J. *et al.* Navigating phase diagram complexity to guide robotic inorganic materials synthesis. *Nature Synthesis* **3**. Open access article, 606–614 (2024).
49. Kim, S., Jung, Y. & Schrier, J. Large Language Models for Inorganic Synthesis Predictions. *J. Am. Chem. Soc.* **146**, 19654–19659 (2024).



50. Kim, S., Schrier, J. & Jung, Y. Explainable Synthesizability Prediction of Inorganic Crystal Polymorphs Using Large Language Models. *Angew. Chem. Int. Ed.* **64**, e202423950 (2025).
51. Miessler, G. L., Fischer, P. J. & Tarr, D. A. *Inorganic Chemistry* 5th (Pearson, Boston, 2014).
52. Greenwood, N. N. & Earnshaw, A. *Chemistry of the Elements* 2nd (Butterworth-Heinemann, Oxford, 1997).
53. Scheidgen, M. *et al.* NOMAD: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software* **8**, 5388 (2023).
54. Márquez, J. A. & Scheidgen, M. *Perovskite Solar Cell Database Project* 2024.
55. Dong, Q. & Cole, J. M. Auto-generated database of semiconductor band gaps using ChemDataExtractor. *Scientific Data* **9** (2022).
56. U.S. Geological Survey. *Data release for Mineral Commodity Summaries 2025* U.S. Geological Survey data release. 2025.
57. World Bank. *Environment, Social and Governance Data* World Bank Data Catalog. Accessed: 2025-10-16. 2023.
58. Jiang, X., Liu, G., Xie, J. & Hu, Z. Boosting SISSO performance on small sample datasets by using Random Forests prescreening for complex feature selection. *en. Front. Phys.* **20**, 14209 (2025). View Article Online DOI: 10.1039/D6FD00030D
59. Schilling-Wilhelmi, M. *et al.* From text to insight: large language models for chemical data extraction. *Chemical Society Reviews* (2025).
60. Antunes, L., Butler, K. & Grau-Crespo, R. *Supporting data for: Crystal Structure Generation with Autoregressive Large Language Modeling* version v1. 2024.
61. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).
62. Shabih, S. *et al.* An autonomous living database for perovskite photovoltaics. *arXiv preprint arXiv:2601.17807* (2026).
63. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
64. Garzón, D. A., Himanen, L., Andrade, L., Sadewasser, S. & Márquez, J. A. *chalcogenide-perovskite-screening* version 1.0.1. 2026.



ML-guided screening of chalcogenide perovskites as solar energy materials

View Article Online

DOI: 10.1039/C6FD00000D

Diego A. Garzón^{1,2,✉}, Lauri Himanen³, Luisa Andrade², Sascha Sadewasser¹, and José A. Márquez³

¹International Iberian Nanotechnology Laboratory, Av. Mte. José Veiga s/n, 4715-330 Braga, Portugal

²LEPABE, ALiCE, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

³Physics Department and CSMB, Humboldt-Universität zu Berlin, Zum Großen Windkanal 2, D-12489 Berlin, Germany
✉ diego.garzon@inl.int

February 23, 2026

Data availability statement

The curated datasets and code developed in this work are openly available at the GitHub repository: <https://github.com/GarzonDiegoFEUP/chalcogenide-perovskite-screening>. An interactive version of the figures is available in the documentation at the GitHub repository. A citable archived version (v1.0.1) is deposited on Zenodo at <https://doi.org/10.5281/zenodo.18743650>.

