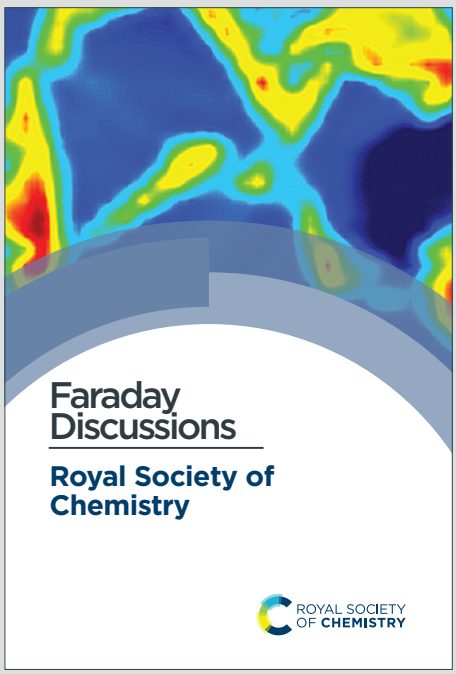
Faraday Discussions

Accepted Manuscript



This article can be cited before page numbers have been issued, to do this please use: B. Das, K. Ji, F. SHENG, K. McCall and T. Buonassisi, *Faraday Discuss.*, 2024, DOI: 10.1039/D4FD00120F.

View Article Online
DOI: 10.1039/D4FD00120F

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Embedding human knowledge in material screening pipeline as filters to identify novel synthesizable inorganic materials

Basita Das,[a], Kangyu Ji,[a] Fang Sheng,[a] Kyle M. McCall,[b] and Tonio Buonassisi[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

How might one embed a chemist's knowledge into an automated materials-discovery pipeline? In generative design for inorganic crystalline materials, generating candidate compounds is no longer a bottleneck — there are now synthetic datasets of millions of compounds. However, weeding out unsynthesizable or difficult to synthesize compounds remains an outstanding challenge. Post-generation "filters" have been proposed as a means of embedding human domain knowledge, either in the form of scientific laws or rules of thumb. Examples include charge neutrality, electronegativity balance, and energy above hull. Some filters are "hard" and some are "soft" — for example, it is difficult to envision creating a stable compound while violating the rule of charge neutrality; however, several compounds break the Hume-Rothery rules. It is therefore natural to wonder: Can one compile a comprehensive list of "filters" that embed domain knowledge, adopt a principled approach to classifying them as either non- conditional or conditional "filters," and envision a software environment to implement combinations of these in a systematic manner? In this commentary we explore such questions, "filters" for screening of novel inorganic compounds for synthesizability.

1 Introduction

"Inverse design" has long been a goal of computational materials science, whereby an algorithm proposes a set of candidate materials satisfying a set of user-defined target properties [1]. Since 2020, several instances of conditional generative-design algorithm have been proposed [2–10]. These are typically based on genetic algorithms or deep neural networks trained using density functional theory (DFT) databases [11,12]. These DFT databases contain well-organized composition, structure, and property relationships for hundreds of thousands of simulated compounds, providing a rich set of training data. Once trained, the algorithms can be directed to generate new, hypothetical compounds, resulting in synthetic databases of up to millions of hypothesized materials.

The ultimate validation of a generative-design workflow is the experimental synthesis and characterization of hypothesized materials. This is not trivial; several teams discuss the difficulties of synthesizing proposed materials in synthetic databases [4,13,14]. There are errors originating from the gap between DFT and experiment [4,15], class imbalances when training the model [16], and errors reconstructing new materials from latent spaces [4], among others. Thus, to date, successes have been relatively modest compared to expectations [13].

Given this, a downselection (or screening) step has often been proposed after the creation of a synthetic materials database. The most obvious approach is to downselect on the basis of properties linked to synthesizability and stability. These can include calculating the convex hull and estimating the value of a new compound using DFT or an ML surrogate model or identifying structural patterns of synthesizable materials [17–19]. Another approach is to perform DFT energy relaxation on proposed compounds, either directly [4,20] or more recently via a machine learning (ML) surrogate model [21]. Lastly, one can apply a set of downselection "filters" [22–25] to a synthetic database to identify candidate compounds that satisfy certain chemical rules embedded in the filter. The latter approach does not use DFT, but rather, aims to encode human domain expertise from synthetic chemistry. This approach can be used not only to downselect candidates within synthetic databases, but also to brute-force screen candidates within ternary phase diagrams.

This "filtering" approach has a rich history. Davies *et al.* [22] introduced the concept of encoding chemical rules for use in high-throughput searches. Davies *et al.* developed a probabilistic framework to assign confidence in the formation of hypothetical compounds, given the proposed oxidation states of their constituent species and was later adopted by Thway *et al.* [25]. Pal *et al.* [23] proposed a series of experimentally-accessible ternary phase diagrams, then applied "filters" based on charge-neutrality rules;

[a] *Dept. of Mechanical Engineering, Massachusetts Institute of Technology, 77 Mass Ave., Cambridge, USA.; E-mail: dasb@mit.edu, buonassi@mit.edu*
[b] *Department of Materials Science and Engineering, University of Texas at Dallas, Richardson, United States*

View Article Online
DOI: 10.1039/D4FD00120F

the team identified 628 thermodynamically stable quaternary chalcogenides, using high-throughput density functional theory (DFT) calculations that also satisfy the charge neutrality principle. More recently, Thway *et al.*[25] expanded upon this filter to include an electronegativity balance filter, which suggests that the most electronegative ion in a compound also has the most negative charge[24]. By applying this to the Cu-In-Te ternary phase diagram, they identified $CuIn_3Te_5$, which was not previously known to the authors, nor in a materials-property database. Further filters have been proposed by Park et al[24] to scan the binary, ternary and quaternary phase diagrams of inorganic materials.

Our current study is directed at three questions: (1) can patterns in chemically similar (adjacent) ternary phase diagrams help identify promising new compounds (*e.g.*, via isovalent substitution), and can this be implemented as a filter? (2) can this approach scale beyond a single ternary phase diagram? (3) in future work, what other forms of human domain knowledge could be embedded in filters, and how best to employ them to minimize false negatives?

2 Overview of our 6-filter Pipeline

Drawing inspiration from Davies *et al.*[22], Pal *et al.*[23], Park *et al.*[24], and Thway *et al.*[25], we expand upon a framework of multiple human-intuition driven filters, aiming to refine the search for synthesizable novel inorganic materials. Building upon prior open-source code we developed a pipeline of six filters to screen ternary phase diagrams (i) charge neutrality filter, (ii) electronegativity balance filter, (iii) unique oxidation state filter, (iv) oxidation state frequency filter (v) intra-phase diagram stoichiometric variation filter, and (vi) cross-phase diagram stoichiometry filter. These six filters are shown graphically in Figure 1. The first four filters were adopted from the work of Park *et al.*[24] and Thway *et al.*[25] and aim to assess viability of each element combination in a proposed compound. The last two of the six filters are developed in this study and are based on the stoichiometric ratios of elements in proposed materials. The words “intra” and “cross” refer to the chemical phase diagrams in consideration. Intra-phase diagram stoichiomtry filter concerns other compounds within the same ternary phase diagram under consideration, whereas cross-phase diagram stoichiometry filter compares stoichiometries of known compounds in adjacent chemical phase diagrams, *e.g.*, by isovalent substitution.

We focus our study on ternary phase diagrams containing known or suspected metal-halide compounds, often known as “perovskite-inspired" materials[26]. These materials include compounds that have compositional ($A_iB_jX_k$) or structural similarity (*e.g.* double perovskites) to lead-halide perovskites. We propose that experimental validation of these compounds is facilitated by a tendency to form stoichiometric compounds (and not compounds with large vacancy concentrations), near room-temperature synthesis via high-throughput liquid approaches, and electronic structures that tend to be more defect tolerant, enabling property measurements even on early-stage, defect-rich materials. We scaled our filter pipeline to 60 different “perovskite-inspired” inorganic ternary phase diagrams ($A_iB_jX_k$) involving elements from group 1 as the *A*-site cation, elements from groups 14 and 15 as the *B*-site cation, and elements from group 17 as anions occupying the *X*-site. To keep code runtimes manageable to a laptop, we screened compounds with up to 20 atoms. We ensured that all compounds satisfy the first “electronegativity balance” filter. We iterated through all the oxidation states of each element reported in literature[27]. We used the Materials Project Dataset[11] to identify existing compounds in the phase diagrams under study and pymatgen[28] for analysis.

A list of more than 50,200 charge-neutral hypothetical “novel compounds" resulted from this process. In this study, we define “novel compound” as one reported neither in the Materials Project database[11,12] nor in the Inorganic Crystal Structure Database (ICSD). While this definition serves our purposes of demonstrating the potential of filters to identify compounds not in our original set, we acknowledge that a more restrictive definition of “novelty” is appropriate when making claims of materials discovery (*e.g.*, credible literature reports but absence in databases may still disqualify a compound as “novel”). After applying all filters (encoding human intuition into the screening process), we generate a downselected list of 27 “novel” hypothetical compounds. The following sections provide details of the design and implementation of each filter, to this case study of 60 ternary phase diagrams. A possible future step of validating the filters using experimental databases, and/or experimentally validating proposed compounds, while out of scope of the current study, is discussed at the end of this paper.

3 Human knowledge driven intuition as "Filters"

The principles of charge neutrality and electronegativity balance are foundational for predicting the stability of chemical compounds. In their groundbreaking study, Park *et al.* (2024)[24] explored an extensive dataset comprising 16,980,849,551 unique binary, ternary, and quaternary compounds, while excluding combinations of equivalent stoichiometry. By applying the charge neutrality principle and electronegativity balance rule, they categorized the compounds as "Allowed" or "Forbidden," based on their compliance with these chemical guidelines. Additionally, they identified compositions as "Known" or "Missing" by cross-referencing with the Materials Project database. This method offers a comprehensive inventory of potential materials for synthesis and property verification. However, this strategy does not account for the actual synthesizability of the reported compounds, indicating that empirical synthesis and testing are necessary to confirm their feasibility as materials.

The challenges for empirical testing of synthesizability of are manifold. The ability to synthesize a compound extends beyond the principle of charge neutrality, encompassing a spectrum of chemical and practical considerations. For a compound to be synthesizable, it must first be thermodynamically stable, *i.e.* exist in it’s lowest energy state or in chemical equilibrium with its environment. This may be a dynamic equilibrium in which individual atoms or molecules are moving but the overall structure is conserved. This type of chemical thermodynamic equilibrium will persist indefinitely unless the system is changed[29]. Furthermore, synthesizing a compound requires identifying a feasible pathway from available starting materials, taking into account the reac-

View Article Online
DOI: 10.1039/D4FD00120F

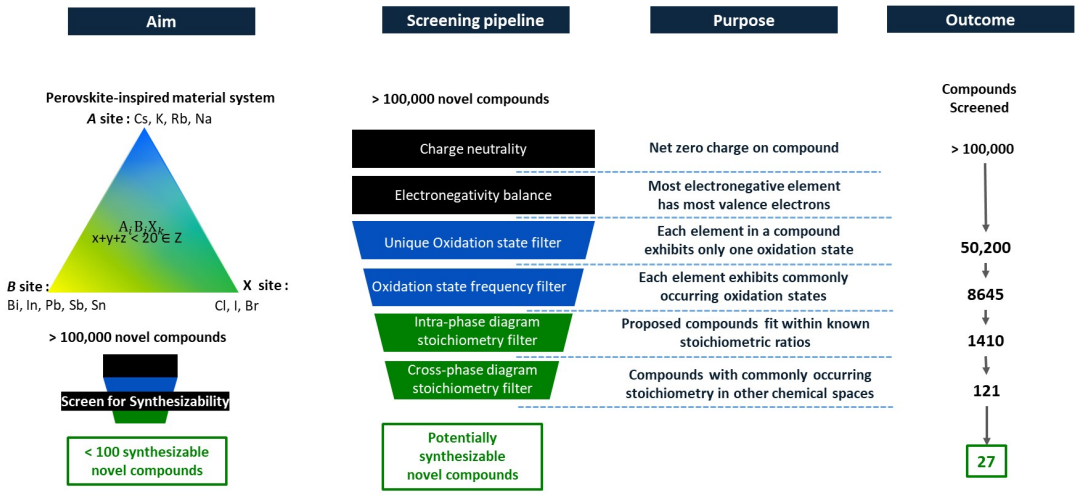


Fig. 1 Screening pipeline formed from stitching together chemical rules and human intuition-driven filters to explore the ternary phase diagrams of "perovskite-inspired" materials. We have searched the chemical phase diagrams comprising of Cesium, Potassium, Sodium, and Rubidium as *A*-site cations; Bismuth, Indium, Lead, Tin, and Antimony for the *B*-site; and Iodine, Chlorine, and Bromine as *X*-site anions. We started with more than > 100,000 novel compounds formed by iterating through various combinations of the elements mentioned above in ternary chemical systems. Over 50,000 charge-neutral compounds satisfying electronegativity balance were initially identified. Applying an oxidation state filter reduced this pool by 80% by excluding compounds with multiple oxidation states per element. Further refinement eliminated compounds with uncommon oxidation states, narrowing the list to approximately 1400. Two additional filters, focusing on stoichiometric ratios, were applied hereafter: the intra-phase diagram stoichiometry filter, which compares new compounds to existing ones within the same chemical phase diagrams, and the cross-phase diagram stoichiometry filter, which assesses common stoichiometries across the broader "perovskite-inspired" phase diagrams. These steps further reduced the pool by 90%, with 27 ultimately meeting all criteria outlined in our study.

tion mechanisms, intermediates and it might simply be that there is no thermodynamically favored reaction path for experimental synthesis. Steric effects, which refer to the physical hindrances caused by the three-dimensional arrangement of atoms, can make certain structures particularly challenging to synthesize. This effect might be pronounced in mixed-cation perovskite structures, where organic and inorganic molecules are used in combination for improving stability. Lastly, the inherent complexity and size of a compound can dictate its synthesizability, with large, intricate molecules often requiring difficult multi-step synthesis with potentially low yields. Thus, the journey from a theoretical compound to a tangible substance is navigated through a landscape shaped by stability, accessibility, energetics, sterics, and practical feasibility. Given the complex landscape of material synthesizability, even with the advent of autonomous labs, synthesizing and validating novel materials in high-throughput remains challenging.

In this context, how can we embed human knowledge and chemical intuition into filters, so we may someday pinpoint compounds that are not only theoretically synthesizable but also practically viable? Our study is broken into three sub-questions: (1) can patterns in chemically similar (adjacent) ternary phase spaces help identify promising new compounds, and can this be implemented as a filter? (2) can this approach scale beyond a single ternary phase space? (3) in future work, what other forms of human domain knowledge could be embedded in filters, and how best to employ them to minimize false negatives? In the following subsections we elaborate upon the four human-intuition driven filters that we designed to condense human knowledge and intuition for synthesizability prediction.

To test our filters we have used "perovskite-inspired" ternary phase diagrams of the stoichiometry $A_iB_jX_k$ where the sum of the stoichiometric fractions is maximum 20, i.e. each compound has maximum of 20 atoms. We have considered Cesium(Cs), Potassium(K), Sodium(Na) and Rubidium(Rb) as the *A*-site cation, Indium (In), Tin (Sn), Antimony (Sb), Lead (Pb) and Bismuth (Bi) as the *B*-site cation and Chlorine (Cl), Bromine (Br) and Iodine (I) as anions occupying the *X*-site. We have used the oxidation states of the elements listed in Ref. 2[27] to form novel compounds. Using the 4 [Cs, K, Na, Rb] , 5 [In, Sb, Sb, Pb, Rb] and 3 [Cl, Br, I] elements as *A*, *B* and *X* - sites elements, respectively, we formed a list of 60 distinct ternary phase diagrams. We generated new compounds by iterating through combinations of their respective oxidation states such that the total number of atoms per compound is less than or equal to 20. By repeating this method for all 60 phase diagrams we generated a compound list of >100,000 novel compounds. To this list of >100,000 compounds, we applied the first two filters in our pipeline (i. charge neutrality filter, and (ii) electronegativity balance filter) to narrow down 50,200 charge neutral and electronegatively balanced compounds. We applied the rest of the four filters in our pipeline on this 50,200 compounds in a successive order to obtain the final list of 27 compounds. In the following four subsections we discuss the four filters in the order they were applied on this list of compounds. To test our fil-

View Article Online
DOI: 10.1039/D4FD00120F

ters we have used "perovskite-inspired" ternary phase diagrams of the stoichiometry where the sum of the stoichiometric fractions is maximum 20, i.e. each compound has maximum of 20 atoms. We have considered Cesium(Cs), Potassium(K), Sodium(Na) and Rubidium(Rb) as the *A*-site cation, Indium (In), Tin (Sn), Antimony (Sb), Lead (Pb) and Bismuth (Bi) as the *B*-site cation and Chlorine (Cl), Bromine (Br) and Iodine (I) as anions occupying the *X*-site. We have used the oxidation states of the elements listed in [27] to form novel compounds. Using the 4 [Cs, K, Na, Rb] , 5 [In, Sb, Sb, Pb, Rb] and 3 [Cl, Br, I] elements as *A*, *B* and *X* - sites elements, respectively, we formed a list of 60 distinct ternary phase diagrams. We generated new compounds by iterating through combinations of their respective oxidation states such that the total number of atoms per compound is less than or equal to 20. By repeating this method for all 60 phase diagrams we generated a compound list of >100,000 novel compounds. To this list of >100,000 compounds, we applied the first two filters in our pipeline (i. charge neutrality filter, and (ii) electronegativity balance filter) to narrow down 50,200 charge neutral and electronegatively balanced compounds. We applied the rest of the four filters in our pipeline on this 50,200 compounds in a successive order to obtain the final list of 27 compounds. In the following four subsections we discuss the four filters in the order they were applied on this list of compounds.

3.1 Unique oxidation states

This filter is adopted from the work of Thway *et al.* [25]. The principle of charge neutrality ensures that the total charge of a newly discovered compound is zero. Nonetheless, this principle does not constrain the possible oxidation states of the elements within the compound. Consequently, when using materials discovery tools or generative design algorithms, we might encounter charge-neutral compounds that exhibit multiple oxidation states for the same element. For instance, in compounds like $CsPb_2Br_7$, cesium is present as Cs^{1+} , lead in both Pb^{2+} and Pb^{4+}, and bromine in Br^{1-}. Although the compound as a whole maintains charge neutrality, the presence of lead in mixed oxidation states casts doubt on its stability as a real, synthesizable material.

Compounds with mixed cation valency (i.e., mixed oxidation states) are known to occur. For example, Pb_3O_4 contains both Pb^{2+} and Pb^{4+}. However, synthesizing such compounds is challenging in practice because it requires precise control of the oxidation potential. To simplify the search for novel compounds, we propose a filter that removes compounds with cations in mixed oxidation states, as they are likely to require significant experimental resources per candidate compound. Therefore, implementing a specialized filter for oxidation states helped us significantly streamline our screening process, efficiently narrowing down the list of candidates by removing compounds with mixed oxidation states. This approach reduces the number of candidates by more than 80%, from 50,200 candidates to only 8645 compounds with single oxidation states for every element

However, synthesizing 8645 compounds is still a huge challenge, even with high-throughput experimentation and hence we need to be more stringent with our screening criteria. Also, even though the unique oxidation state filter removed compounds with mixed oxidation state, it did not remove compounds that have obscure oxidation states of elements. To solve this problem we implemented the oxidation state frequency filter as discussed in the following section.

3.2 Oxidation state frequency filter

Ref. 25 provides a comprehensive overview of all potential oxidation states for elements, including their most commonly observed oxidation states. Thway *et al.* [25], adopted a frequency-based rule for determining the oxidation states of elements, grounded in their prevalence within the Materials Project database. This approach incorporates a variable cutoff value *v*, allowing for adjustable specificity. For the findings presented in this paper, we selected a cutoff of 0.2, meaning that any oxidation state of an element reported in less than 20% of instances across the database is omitted from our considerations. Applying this criterion to the 8645 charge-neutral novel compounds identified in our preliminary screening yielded a refined list of 1410 promising candidates, effectively filtering out over 80% of the initial set.

However, this filtering strategy comes with a potential limitation: it may inadvertently dismiss certain novel compounds that manifest in rare or less conventional oxidation states. Such an exclusion risks overlooking compounds with unique properties or applications, underscoring a trade-off between efficiency and the breadth of discovery in our screening methodology. A search targeting exceptional materials may purposely prioritize candidates with rare oxidation states, following the recommendations of Schrier *et al.* [30]

The two filters we have explored focus exclusively on selecting materials by examining their oxidation states. Yet, the potential for synthesizing novel compounds, even those with unique oxidation states, can be significantly influenced by their stoichiometric ratios. In the upcoming sections, we will delve into methodologies for filtering compounds based on their stoichiometries, addressing how these numerical relationships impact the feasibility of synthesizing new materials.

3.3 Intra-phase diagram stoichiometry filter

The intra-phase diagram stoichiometry filter is designed to eliminate compounds that may not form given the stoichiometric imbalances between the constituent elements, i.e. to eliminate the extreme cases of stoichiometric combinations. For example, we predicted a novel compound in the chemical phase diagram of cesium lead bromide of the formula $CsPb_4I_9$ where the cesium is in Cs^{1+} state, Pb is in Pb^{2+} and I is in I^{1-}, and hence is charge neutral. However, the synthesizing $CsPb_4I_9$ raises challenges in coordination chemistry, primarily due to the potential mismatch between the stoichiometry and the preferred coordination geometries of the lead (II) ions. Typically favoring an octahedral or cubic coordination with four or six points of contact with surrounding halide ions, the lead ions in $CsPb_4I_9$ may struggle to adopt a stable configuration given the compound's unusual lead-to-iodine ratio. Furthermore, it might be difficult to attain an optimal spatial arrangement which would be able to fit the sizable iodine atoms in

View Article Online
DOI: 10.1039/D4FD00120F

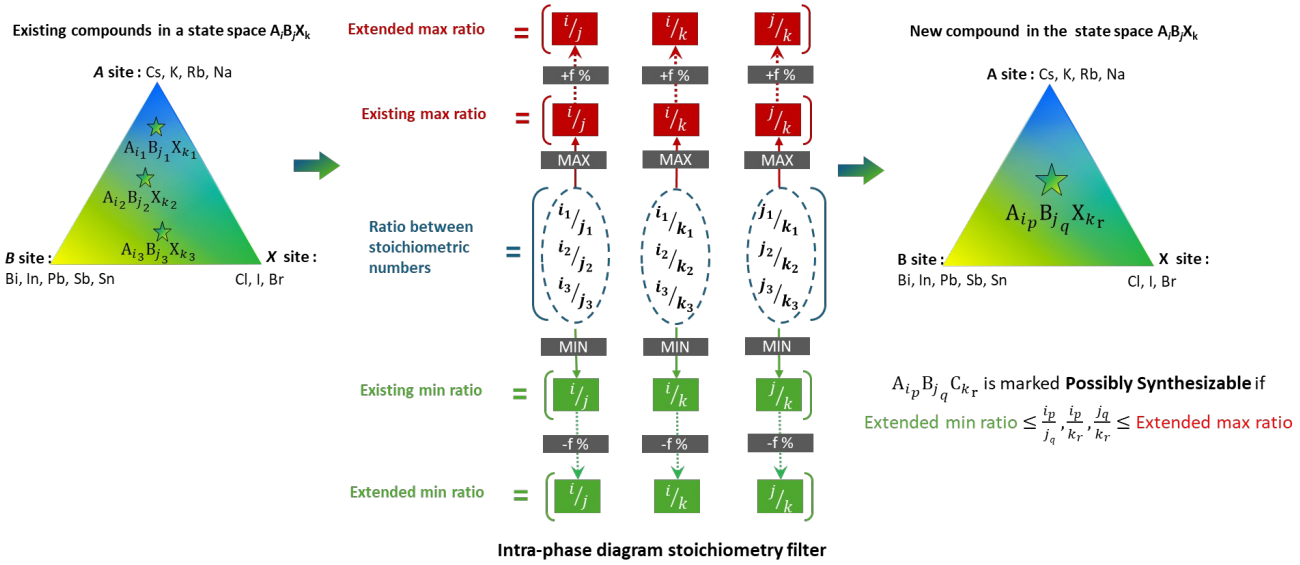


Fig. 2 Intra-phase diagram stoichiometry filter encodes knowledge of existing stoichiometries in a chemical phase diagrams into our material screening protocol. It involves scanning all the existing stoichiometries reported in a chemical phase diagram, and calculating the maximum and minimum ratio between the number of atoms in each positions. These ratios quantifies the structural stability exhibited by the combinations of the elements in the chemical phase diagram. To find new compositions which are likely to be structurally stable, we screen for materials for which the ratio between the number of the atoms in each position are in between the maximum and minimum ratio exhibited by the existing compounds in the particular phase diagram. The factor f% has been included to allow materials for which the ratio between the number of the atoms in each position is very close but outside the range of maximum and minimum ratio of the existing compounds.

a stable geometry.

The intra-phase diagram stoichiometry filter is an approach for discerning the likelihood of formation of a novel compound, based on the stoichiometric ratios of the constituent elements. By establishing a historical range of stoichiometric ratios derived from known stable compounds in the same chemical phase diagram, the filter can predict the structural feasibility of new compounds. To find the range of stoichiometries to consider for the ternary phase diagram of elements *A*, *B*, and *X*, every known compound of the form $A_iB_jX_k$ reported in Materials project database [11,12] is analyzed. The ratio between the stoichiometric fractions $i/j, i/k,$ and j/k are calculated for every reported compound and maximum and minimum value for each ratio is obtained as shown in the Fig. 2. For a compound like $CsPb_4I_9$, the filter would assess its stoichiometric ratios against existing data of $Cs_iPb_jI_k$ phase diagram to determine if such a structure has been previously successful. With the capacity to extend these ratios by a user-defined margin $f\%$, the filter allows for the consideration of slightly unconventional compounds, ensuring that innovative yet stable stoichiometries are not overlooked. For the purpose of this communication we assumed $f\% = 20\%$. It accounts for the preferred coordination geometries and packing efficiencies within a given chemical phase diagram.

Upon applying the intra-phase diagram stoichiometry filter to the pool of 1410 compounds, previously refined through the oxidation state frequency filter, we distilled the selection down to 121 novel compounds. This represents a substantial refinement, effectively excluding over 90% of the initially identified novel compounds. Applying this filter eliminates compounds with unusually imbalanced stoichiometries (like $CsPb_4I_9$), and passes compounds with more balanced stoichiometries (like $Cs_3In_2I_9$). Whether atoms are actually likely form such compounds (like octahedrally coordinated In(3+) in $Cs_3In_2I_9$) is a matter to be addressed by the next filter, which implicitly considers isovalent substitution.

Although a practical tool for preliminary screening, we can also discuss several drawbacks of the intra-phase diagram stoichiometry filter, its reliance on historical data may lead to a conservative approach that overlooks novel compounds with unconventional stoichiometries, which, although rare, might possess unique and desirable properties. This might lead to scenarios where very common stoichiometric ratios which exist in other adjacent chemical phase diagrams are not identified because the $f\%$ was too stringent to encompass those. This historical data dependency also implies that the filter's effectiveness is only as robust as the databases it references; incomplete or biased data sets can result in inaccurate stoichiometric boundaries, leading to potential mis-classification of compounds as unsynthesizable. Additionally, even compounds that fall within the defined stoichiometric ranges are not guaranteed to be stable, as the filter cannot account for kinetic barriers. This filter implicitly quantifies the effect of many different factors using only stoichiometric ratios, and hence is not nuanced to consider the impact of the different factors individually. The introduction of a user-defined tolerance for expanding the range of acceptable stoichiometric ratios injects heuristics into the filter's operation. Such heuristics can skew the filter's objectivity, leading it to be perceived as either overly stringent or excessively relaxed.

View Article Online
DOI: 10.1039/D4FD00120F

As a measure to overcome the drawback where our filtering method might overlook some of the most common stoichiometries in the "perovskite-inspired" phase diagrams, we implemented the filter discussed in the following section.

3.4 Cross-phase diagram stoichiometry filter

This filter is designed to initially aggregate a comprehensive list of stoichiometries observed across various chemical phase diagrams within the Materials Project database. We identified a list of 10 unique stoichiometric ratio [3, 1, 6],[3, 2, 9],[1, 1, 4],[2, 1, 6],[1, 1, 3],[1, 2, 5],[4, 1, 6],[1, 2, 7],[2, 1, 5], and [3, 1, 5] reported among the 60 unique chemical phase diagrams we studied by combining group 1, 14, 15 and 17 elements (perovskite inspired ternary phase diagrams). After compiling this list, we subjected the 1410 charge-neutral and electronegatively balanced compounds which also satisfied the two oxidation state filters to this filter, resulting in the identification of 162 novel compounds as synthesizable.

The filter's proficiency is evaluated based on its capability to accurately identify new compounds with stoichiometric ratios that not only commonly occur within the "perovskite-inspired" chemical phase diagrams but also meet the criteria of the intra-phase diagram stoichiometry variation filter. Among the extensive list of 1410 charge-neutral and electronegatively balanced compounds, we then successfully isolated 27 compounds that conformed to both the intra-phase diagram stoichiometry and cross-phase diagram stoichiometery filters, as listed in Table 1. The results can be found in the online repository[31] of our code.

The materials reported in Table 1 encompasses all the 60 ternary phase diagrams we studied in this communication. To obtain the results presented in Table 1 we configured our pipeline as given below :
(i) Charge neutrality - TRUE,
(ii) electronegativity balance - TRUE,
(iii) Unique oxidation state - TRUE (not allowing for mixed oxidation states of the same elements),
(iv) Oxidation state frequency - 20% (allowing only those oxidation states of an element which occur at least 20% of the all the times that element occurred in the reference database),
(v) Intra-phase diagram filter with $f\% = 20\%$ margin - TRUE , and
(vi) Cross-phase diagram filter - TRUE
However, this is only one of the many configurations one can configure this pipeline to. In the next section we demonstrate the adaptability of this framework.

4 Adaptability of the pipeline

One of the key features of the pipeline is its adaptability to fit the ever changing needs of material screening problems. The overarching goal of this ongoing work is to condense different chemical rules, practical design considerations as well as human intuitions (which often only exists as tacit knowledge) into filters. These individual filters can then be combined in different configurations to make different screening pipeline as needed by the problem in hand. Also, until we study different filter configurations it is also not obvious which set of filters will give us highest success rate in identifying synthesizable materials, hence the adaptable design.

To demonstrate the adaptability of our filter pipeline we showcase the results obtained with Cs-Pb-Br system in Fig. 3 as an example case. To generate Fig. 3 we used the following "Filter" configuration :
(i) Charge neutrality - TRUE,
(ii) electronegativity balance - TRUE,
(iii) Unique oxidation state - FALSE (allowing for mixed oxidation states of the same elements),
(iv) Oxidation state frequency - 20% (allowing only those oxidation states of an element which occur at least 20% of the all the times that element occurred in the reference database),
(v) Intra-phase diagram filter with a $f\% = 20\%$ margin - TRUE,
(vi) Cross-phase diagram filter - TRUE
This particular configuration was selected to demonstrate how the pipeline might be tuned to the needs of a screening problem at hand. When we screened the ternary phase diagram space of Cs-Pb-Br with all six filters in the filter configuration mentioned above, we obtained materials that satisfied the first two chemical rules filter, the combination of the two oxidation state filters, and at least one of the stoichiometric filters. It discarded all materials which satisfied the combination of the two oxidation state filters but did not qualify the screening criteria of either of the two stoichiometric filters. These materials are marked in black as shown in Fig. 3 for the example case cesium Lead Bromide (Cs-Pb-Br) phase diagram. The compounds already existing in the Materials Project database are marked in green. The novel compounds that were identified as "Synthesizable" by the cross-phase diagram stoichiometric filter are marked in "blue" and those by the intra-phase diagram stoichiometric filter are marked in red as shown in Fig. 3. The material marked in blue Cs_3PbBr_5, was deemed synthesizable by both the stoichiometric filters and hence made it to the list of 27 compounds presented in Table 1.

The adjustable percentage values of the oxidation state frequency filter and the margin values $f\%$ of the intra-phase diagram filter gives further flexibility to the user to tune the screening pipeline. By reducing the percentage value of the oxidation state filter we can screen for compounds which might exhibit more obscure oxidation states. Similarly, by increasing the margin value of the intra-phase diagram filter, we go beyond the bounds of the known stoichiometric ratios. Hence, these tunable values limit the influence of the bias in the known material libraries on our novel material discovery pipeline.

5 Conclusion

Our current study focused on the question of whether patterns in chemically similar (adjacent) ternary phase diagrams help identify promising new compounds, and can this be implemented as a filter? In conclusion to this study, the screening framework discussed here is based on human knowledge driven intuition. It incorporated human intuition as well as chemical rules in the form of filters to screen out compounds with mixed oxidation states as well as obscure oxidation states. It also considers historical stoichiometric data in the same chemical phase diagrams and common stoichiometric ratio in analogous chemical phase di-

View Article Online
DOI: 10.1039/D4FD00120F

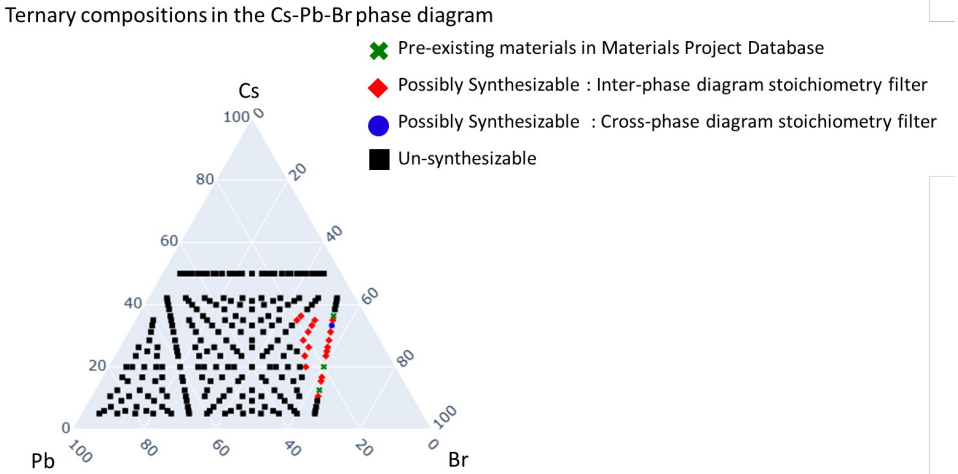


Fig. 3 Ternary phase diagram of the cesium lead bromide. A total of 235 charge neutral compounds were identified in this phase diagram, out of which 3 compounds - $CsPbBr_3$, $CsPb_2Br_5$ and Cs_4PbBr_6 are marked in green cross, were previously reported in Materials Project database. We identified one compound, Cs_3PbBr5 marked in blue using the cross-phase diagram stoichiometry filter and 18 compounds, marked in red using the intra-phase diagram stoichiometry filter. The remaining 214 compounds marked in black were deemed un-synthesizable by the oxidation state based filters and the two stoichiometric filters. Out of all 235 compounds, only one novel compound, Cs_3PbBr_5 (marked in blue) was identified as by the complete pipeline of filters.

Table 1 27 novel charge neutral potential compounds identified using the 6-filter pipeline

Composition	Stoichiometry	Oxidation states
$RbBiBr_4$	[1, 1, 4]	[[1], [3], [-1]]
Rb_2BiBr_5	[2, 1, 5]	[[1], [3], [-1]]
Rb_3PbBr_5	[3, 1, 5]	[[1], [2], [-1]]
Rb_2SbBr_5	[2, 1, 5]	[[1], [3], [-1]]
Rb_2SbI_5	[2, 1, 5]	[[1], [3], [-1]]
Rb_2InI_5	[2, 1, 5]	[[1], [3], [-1]]
$Rb_3In_2I_9$	[3, 2, 9]	[[1], [3], [-1]]
Na_2InBr_5	[2, 1, 5]	[[1], [3], [-1]]
$Na_3In_2Br_9$	[3, 2, 9]	[[1], [3], [-1]]
K_2BiI_5	[2, 1, 5]	[[1], [3], [-1]]
K_2InBr_5	[3, 1, 5]	[[1], [3], [-1]]
$K_3In_2Br_9$	[3, 2, 9]	[[1], [3], [-1]]
K_2InI_5	[2, 1, 5]	[[1], [3], [-1]]
$K_3In_2I_9$	[3, 2, 9]	[[1], [3], [-1]]
$CsBiBr_4$	[1, 1, 4]	[[1], [3], [-1]]
Cs_2BiBr_5	[2, 1, 5]	[[1], [3], [-1]]
Cs_2BiI_5	[2, 1, 5]	[[1], [3], [-1]]
Cs_2BiCl_5	[2, 1, 5]	[[1], [3], [-1]]
Cs_3PbBr_5	[3, 1, 5]	[[1], [2], [-1]]
Cs_3PbI_5	[3, 1, 5]	[[1], [2], [-1]]
Cs_3PbCl_5	[3, 1, 5]	[[1], [2], [-1]]
Cs_3SbBr_5	[3, 1, 5]	[[1], [3], [-1]]
Cs_2SbI_5	[2, 1, 5]	[[1], [3], [-1]]
Cs_2SbCl_5	[2, 1, 5]	[[1], [3], [-1]]
Cs_2InI_5	[2, 1, 5]	[[1], [3], [-1]]
$Cs_3In_2I_9$	[3, 2, 9]	[[1], [3], [-1]]
Cs_2InCl_5	[2, 1, 5]	[[1], [3], [-1]]

View Article Online
DOI: 10.1039/D4FD00120F

agrams to screen for synthesizable materials. This framework is to show how incorporation of such human knowledge driven intuition into our screening pipeline would help us narrow down on synthesizable materials after a huge volume of novel compounds have been generated using material generation algorithm. The screening framework, aimed at identifying synthesizable, charge-neutral compounds, narrows an initial set of 50,200 candidates to 27.

We also want to highlight the scalability of this method beyond ternary phase diagrams. Even though, the results presented in this paper deals with only ternary phase diagram, we have applied the same set of filters to quarternary phase diagrams. Similar chemical rules were applied by Park *et al.*[24] to also screen quaternary phase diagrams.

What additional knowledge or rules of thumb would prove useful to embed in filters? We posit that ionic radii could enable screening materials based on parameters such as Goldschmidt's tolerance factor[32,33] and octahedral factor, providing greater insight into the structural viability of each compound. Also the consideration of the exposed orbital of an element in a particular oxidation state in determining the stoichiometries might lead better predictability of synthesizable stoichiometries. Another candidate filter is "manufacturability," although this would be a multi-factor descriptor, possibly embedding domain knowledge about precursor solubility, chemical reaction kinetics, synthesis tool-specific constraints, thermal budget, and materials availability, among others. Ideally, compounds could be ranked based on ease of synthesis, yield, production speed, and supply-chain resilience.

An open question remains, concerning experimental validation. At this point, we do not know which combination(s) of filters yields the most effective discovery of novel compounds. If the filters are too permissive, filters lose their utility; too selective, they may focus experimental effort on unfruitful compounds (or result in a null set). It is possible that the specific combination of filters must be tailored for different materials diagrams, depending on the relative constraints of each filter, and the amount of background information (training data) for each. Ultimately, this approach of discrete filters may even merge with first-principles or surrogate-model-based screening of candidate compounds, as computational speed increases.

Author Contributions

BD and TB conceived of the work. BD developed the intra-phase diagram stoichiometry filter and cross-phase diagram stoichiometry filters, and built the pipeline by integrating it with existing filters. KJ, FS, KM, and TB provided scientific advice. BD wrote and all authors contributed to the manuscript.

Code availability

All codes used to generate the results in this communication is available here : https://github.com/PV-Lab/SynthesizabilityFilter.git. The code along with the dataset needed is published here : https://doi.org/10.5281/zenodo.12686093.

Dataset availability

The data used to calculate the stoichiometric ratios and identify common stoichiometric combinations in the 60 phase diagrams studied in this paper was downloaded from the Materials Project database on June 2022, and is made available here: https://doi.org/10.5281/zenodo.12686093

Conflicts of interest

The authors have no conflict of interest.

Acknowledgements

BD, KJ, FS and TB acknowledges First Solar for supporting this research. KMM also acknowledges funding from the University of Texas at Dallas. BD also acknowledges Alexander E. Siemenn for valuable inputs on the figures.

Technology use disclosure

The writing of this manuscript was assisted by ChatGPT (specifically for language improvement and literature review). All authors have read, corrected and verified all information presented in this manuscript and Supplementary Information

Notes and references

1 A. Zunger, *Nature Reviews Chemistry*, **2**, 1–16.
2 J. Noh, G. Ho Gu, S. Kim and Y. Jung, *Chemical Science*, 2020, **11**, 4871–4881.
3 H. Choubisa, P. Todorović, J. M. Pina, D. H. Parmar, Z. Li, O. Voznyy, I. Tamblyn and E. H. Sargent, *npj Computational Materials*, 2023, **9**, 117.
4 Z. Ren, S. I. P. Tian, J. Noh, F. Oviedo, G. Xing, J. Li, Q. Liang, R. Zhu, A. G. Aberle, S. Sun, X. Wang, Y. Liu, Q. Li, S. Jayavelu, K. Hippalgaonkar, Y. Jung and T. Buonassisi, **5**, 314–335.
5 T. Xie and J. C. Grossman, *Physical Review Letters*, **120**, 145301.
6 S. I. P. Tian, A. Walsh, Z. Ren, Q. Li and T. Buonassisi, 2022.
7 A. Vasylenko, D. Antypov, V. V. Gusev, M. W. Gaultois, M. S. Dyer and M. J. Rosseinsky, *npj Computational Materials*, 2023, **9**, 1–10.
8 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80–85.
9 M. Alverson, S. G. Baird, R. Murdock, E. Sin-Hang Ho, J. Johnson and T. D. Sparks, *Digital Discovery*, 2024, **3**, 62–80.
10 C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, R. Tomioka and T. Xie, *arXiv*, 2024.
11 *Materials Project*, `https://next-gen.materialsproject.org/`.
12 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, *APL materials*, 2013, **1**, year.
13 J. Leeman, Y. Liu, J. Stiles, S. B. Lee, P. Bhatt, L. M. Schoop and R. G. Palgrave, *PRX Energy*, **3**, 011002.

View Article Online
DOI: 10.1039/D4FD00120F

14 A. K. Cheetham and R. Seshadri, *Chemistry of Materials*, **36**, 3490–3495.

15 K. Lejaeghere, V. Van Speybroeck, G. Van Oost and S. Cottenier, *Critical Reviews in Solid State and Materials Sciences*, 2014, **39**, 1–24.

16 K. Li, D. Persaud, K. Choudhary, B. DeCost, M. Greenwood and J. Hattrick-Simpers, *Nature Communications*, 2023, **14**, 7283.

17 J. Jang, G. H. Gu, J. Noh, J. Kim and Y. Jung, *Journal of the American Chemical Society*, **142**, 18836–18843.

18 G. H. Gu, J. Jang, J. Noh, A. Walsh and Y. Jung, *npj Computational Materials*, **8**, 1–8.

19 F. T. Szczypiński, S. Bennett and K. E. Jelfs, *Chemical Science*, **12**, 830–840.

20 S. Kim, J. Noh, T. Jin, J. Lee and Y. Jung, *npj Computational Materials*, 2023, **9**, 1–9.

21 C. Chen and S. P. Ong, *A Universal Graph Deep Learning Interatomic Potential for the Periodic Table*, `https://arxiv.org/abs/2202.02450v2`.

22 D. W. Davies, K. T. Butler, O. Isayev and A. Walsh, *Faraday Discussions*, 2018, **211**, 553–568.

23 K. Pal, Y. Xia, J. Shen, J. He, Y. Luo, M. G. Kanatzidis and C. Wolverton, *npj Computational Materials*, 2021, **7**, 1–13.

24 H. Park, A. Onwuli, K. T. Butler and A. Walsh, *Faraday Discussions*.

25 M. Thway, A. P. Chen, H. Dai, J. Recatala-Gomez, S. I. P. Tian, R. Zhu, W. Zhai, F. Wei, D. V. M. Repaka, T. Buonassisi, P. Canepa and K. Hippalgaonkar, 2024.

26 R. E. Brandt, J. R. Poindexter, P. Gorai, R. C. Kurchin, R. L. Hoye, L. Nienhaus, M. W. Wilson, J. A. Polizzotti, R. Sereika, R. Žaltauskas, L. C. Lee, J. L. Macmanus-Driscoll, M. Bawendi, V. Stevanović and T. Buonassisi, *Chemistry of Materials*, 2017, **29**, 4667–4674.

27 *List of oxidation states of the elements*, `https://web.archive.org/web/20240708141854/https://en.wikipedia.org/w/index.php?title=Template:List_of_oxidation_states_of_the_elements&oldid=1192756587`, Page Version ID: 1192756587.

28 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Computational Materials Science*, 2013, **68**, 314–319.

29 *Chemical stability*, `https://web.archive.org/web/20240708142428/https://en.wikipedia.org/w/index.php?title=Chemical_stability&oldid=1181112425`, Page Version ID: 1181112425.

30 J. Schrier, A. J. Norquist, T. Buonassisi and J. Brgoch, *Journal of the American Chemical Society*, 2023, **145**, 21699–21716.

31 B. Das, *PV-Lab/Synthesizability-Filter*, 2024, `https://github.com/PV-Lab/Synthesizability-Filter`, original-date: 2024-03-04T22:38:18Z.

32 M. R. Filip and F. Giustino, *Proceedings of the National Academy of Sciences*, 2018, **115**, 5397–5402.

33 C. Li, X. Lu, W. Ding, L. Feng, Y. Gao and Z. Guo, *Acta Crystallographica Section B: Structural Science*, 2008, **64**, 702–707.

View Article Online
DOI: 10.1039/D4FD00120F

Code availability

All codes used to generate the results in this communication is available here : https://github.com/PV-Lab/Synthesizability-Filter.git.

Data availability

All data presented in the paper as well as extra data are available in the results folder of the online repository

Dataset availability

The data used to calculate the stoichiometric ratios and identify common stoichiometric combinations in the 60 phase diagrams studied in this paper was downloaded from the Materials Project database on June 2022, and is made available here:

https://www.dropbox.com/scl/fi/yc7pgthfllij9ne4icw1d/all_materials_9June2022.h5?rlkey=72wl4m6a61dtvsvtqqxpfykm5dl=0

The code as well as the data is also available here: https://doi.org/10.5281/zenodo.12686093