Faraday Discussions
Cite this: *Faraday Discuss.*, 2025, **256**, 614

PAPER

A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes

Wenhao Sun †* and Nicholas David †

Received 25th May 2024, Accepted 8th August 2024
DOI: 10.1039/d4fd00112e

Synthesis of predicted materials is the key and final step needed to realize a vision of computationally accelerated materials discovery. Because so many materials have been previously synthesized, one would anticipate that text-mining synthesis recipes from the literature would yield a valuable dataset to train machine-learning models that can predict synthesis recipes for new materials. Between 2016 and 2019, the corresponding author (Wenhao Sun) participated in efforts to text-mine 31 782 solid-state synthesis recipes and 35 675 solution-based synthesis recipes from the literature. Here, we characterize these datasets and show that they do not satisfy the "4 Vs" of data-science—that is: volume, variety, veracity and velocity. For this reason, we believe that machine-learned regression or classification models built from these datasets will have limited utility in guiding the predictive synthesis of novel materials. On the other hand, these large datasets provided an opportunity to identify anomalous synthesis recipes—which in fact did inspire new hypotheses on how materials form, which we later validated by experiment. Our case study here urges a re-evaluation on how to extract the most value from large historical materials-science datasets.

1 Introduction

High-throughput computational materials discovery and design methods are reaching maturity, both in predicting new materials for exploratory synthesis,[1–3] and in designing new compounds for diverse functional applications.[4–6] Synthesizability is a major consideration in computational materials search efforts, and is typically evaluated using convex-hull stability—indicated by whether a material lies upon the convex hull and is therefore stable,[7] or if it is metastable, with an 'energy above the hull' commensurate in magnitude with known metastable materials.[8,9] If a computationally designed material is deemed synthesizable and has interesting predicted properties, the material is then passed on to

Department of Materials Science and Engineering, University of Michigan, Ann Arbor, MI, USA. E-mail: whsun@umich.edu
† Equal contribution.

experimental chemists for real-world validation. However, convex-hull stability does not provide any guidance on how to actually synthesize a predicted material—such as which precursors to use, or what reaction temperatures and times are optimal. In this sense, predictive synthesis has become an urgent new bottleneck in the computational materials discovery pipeline.[10–12]

Predictive retrosynthesis has been a long-standing goal of organic chemistry,[13–16] with notable recent advances made using deep neural networks.[17,18] These machine-learning successes were enabled by the availability of large commercial databases of organic reactions, such as SciFinder[19] and Reaxys.[20] Similar commercial databases for inorganic materials synthesis reactions do not currently exist. However, because there have been thousands of successful materials synthesis reports in the literature, text-mining synthesis recipes from published papers could provide a vast source of expert knowledge to train machine-learning models for predictive inorganic materials synthesis.

Between 2016 and 2019, I‡ was a postdoctoral fellow in Gerbrand Ceder's research group at Lawrence Berkeley National Laboratory and participated in the text-mining of 31 782 solid-state synthesis recipes[21] and 35 675 solution-based synthesis recipes[22] from the literature. Here, I offer a retrospective account on attempts to build machine-learning (ML) models for predictive materials synthesis from this dataset. Incidentally, this story follows the Gartner 'hype cycle',[23] which proceeds *via* (1) technology trigger, (2) peak of inflated expectations, (3) valley of disillusionment, (4) slope of enlightenment, and (5) plateau of productivity. The perspectives here are my own, and are not necessarily shared by my co-authors from the text-mining publications.

Here, we begin by reviewing the natural language processing strategies used to build the text-mined recipe database. Then, we evaluate the dataset against the "4 Vs" of data science, and show that the dataset suffers limitations in volume, variety, veracity, and velocity. While some of these limitations stem from technical issues in text-mining, we argue that these limitations primarily arise from the social, cultural, and anthropogenic biases in how chemists have explored and synthesized materials in the past.[24] We show that machine-learning models trained on this text-mined dataset are successful in capturing how chemists think about materials synthesis, but do not offer substantially new guiding insights on how to best synthesize a novel material.

On the other hand, we found that the most interesting recipes in this dataset are actually the anomalous recipes—the ones that defy conventional intuition in solid-state synthesis. These anomalous recipes are also relatively rare, meaning they would not significantly influence regression or classification models. By manually examining some anomalous recipes, we arrived at a new mechanistic hypothesis on how solid-state reactions proceed, and how to select precursors that enhance the reaction kinetics and selectivity of target materials. This hypothesis drove a series of high-visibility follow-up studies,[25–28] which experimentally validated our hypothesized mechanism, gleaned from the text-mined literature dataset.

‡ In this manuscript, the first-person references Wenhao Sun. The "4 Vs" analysis of the text-mining dataset was performed by Nicholas David, a current PhD candidate in the Sun research group, who was not involved in the original text-mining works.

As natural language processing algorithms continue to improve, many other text-mining studies are emerging in materials science, with domains ranging from nanomaterials[29] to alloys,[30,31] catalysts,[32,33] and more.[34–36] Although our retrospective is focused on materials synthesis, we believe that our case study here offers broad and general lessons on how to best leverage large historical datasets for data-driven chemical discovery and design.

2 Natural language processing of materials synthesis paragraphs

To text-mine solid-state materials synthesis recipes from the literature, numerous technical challenges needed to be overcome. Here, we briefly review the natural language processing methods used to build these datasets, with full details available in the original publications.[21,37,38] Our text-mining pipeline can be broken down into five steps: (1) procure full-text literature, (2) identify synthesis paragraphs, (3) extract relevant precursor and target materials, (4) build a list of synthesis operations, and (5) compile data into a common 'recipe' format with balanced stoichiometric reactions. We note that these text-mining studies were performed before the widespread availability of large language models like ChatGPT,[39,40] whose potential impact we will evaluate later in our discussion.

2.1 Full-text literature procurement

First, full-text permissions were obtained from scientific publishers including Springer, Wiley, Elsevier, the Royal Society of Chemistry, the Electrochemical Society, the American Physical Society and the American Chemical Society, enabling large-scale downloads of publication texts. Only papers with HTML/XML formats published after the year 2000 were chosen for extraction, as older publications in only scanned PDF format were difficult to parse. To identify which paragraph in a paper corresponds to a synthesis procedure (which appears at different locations of a manuscript depending on publisher), we made a probabilistic assignment based on which paragraph contained keywords most commonly associated with inorganic materials synthesis.

2.2 Extracting recipe targets and precursors

The target and precursors of a recipe are difficult to assign using rule-based approaches. In different contexts, the same material can play different roles—for example, TiO_2 is sometimes a target material in nanoparticle synthesis, and sometimes TiO_2 is a precursor for ternary oxides like $Li_4Ti_5O_{12}$. Likewise, ZrO_2 is sometimes a precursor, but it can also be used as the grinding medium in ball-milling. There are also many peculiarities on how materials are represented. Solid-solutions are often written as $A_xB_{1-x}C_{2-\delta}$; some materials like $Pb(Zr_{0.5}Ti_{0.5})O_3$ are abbreviated as PZT, and dopants are sometimes represented as $Zn_3Ga_2Ge_{2-x}Si_xO_{10}$: 2.5 mol% Cr^{3+}.

Instead of enumerating all possible representations of target and precursor materials, in He *et al.*[38] we replaced all chemical compounds with <MAT>, and used sentence context clues to label target, precursors, or other (such as atmospheres, reaction media, *etc.*). For example, from the sentence "a spinel-type cathode material <MAT> was prepared from high-purity precursors <MAT>,

<MAT> and <MAT>, at 700 °C for 24 h in <MAT>", it is immediately apparent that the first <MAT> represents a target, the next three are precursors, and the final one corresponds to reaction media. We used a bi-direction long short-term memory neural network with a conditional random field layer (BiLSTM-CRF) to identify these sentence context clues. To train the BiLSTM-CRF, we manually annotated targets, precursors and other reaction media in 834 solid-state synthesis paragraphs.

2.3 Constructing synthesis operations

Identifying materials synthesis operations poses another challenge to text classification, as chemists often use a variety of synonyms to describe the same process—for example, the words 'calcined', 'fired', 'heated', 'baked', all correspond to the same oven heating procedure in solid-state synthesis. To cluster keywords into topics corresponding to specific materials synthesis operations, in Huo *et al.*[37] we used latent Dirichlet allocation (LDA), which builds topic-word distributions for similar processes over tens of thousands of paragraphs. Auxiliary words/tokens are then associated with these topics—for example, keywords for heating include [°C, h, min, air, annealed, samples, atmosphere, films, heat, treatment, annealing, furnace, treated, temperatures, temperature].

We classified sentence tokens into 6 categories: mixing, heating, drying, shaping, quenching, or not operation, corresponding to the main operations in solid-state synthesis. We manually assigned token labels for an annotated set consisting of 100 solid-state synthesis paragraphs (664 sentences). After doing so, for each type of operation, we were able to associate and extract the relevant parameter values (or range of values); for example, the times, temperatures, and atmospheres associated with heating. A Markov chain representation of the experimental operations then was able to reconstruct a flowchart of the synthesis procedures.

2.4 Compiling synthesis recipes and reactions

Finally, in Kononova *et al.*,[21] all the text-mined precursors, targets, and operations were combined into a single JSON database of recipes. We also attempted to build balanced chemical reactions for the identified precursors and target materials, such that we could later compute their reaction energetics using DFT-calculated bulk energies from the Materials Project.[41] Reactions often required including volatile atmospheric gasses, such as O_2, N_2, CO_2, *etc.*, to be balanced.

Altogether, we scraped a total of 4 204 170 papers, which contained 6 218 136 paragraphs in the experimental sections. After classification, 188 198 paragraphs were found to describe inorganic synthesis, such as solid-state, hydrothermal, sol–gel, and co-precipitation syntheses, with 53 538 corresponding to solid-state synthesis. The overall extraction yield of the pipeline is 28%, meaning that out of 53 538 solid-state paragraphs, only 15 144 of them produce a balanced chemical reaction. As a test of the full extraction pipeline, 100 paragraphs were randomly pulled from the set of paragraphs classified as solid-state synthesis and checked for completeness of the extracted data. Out of the 100 paragraphs, we found 30 that did not contain a complete set of starting materials and final products, meaning that even a human expert would not be able to reconstruct a reaction from these paragraphs. To build the final database, we prioritized

providing accurate recipes, so if a paragraph fails the extraction pipeline, we elected to exclude it from the dataset rather than include an incomplete recipe. The final dataset contained 19 488 paragraphs, with 13 009 unique targets, 1845 unique precursors and 16 290 unique reactions. In 2020, the dataset was expanded to 31 782 chemical reactions retrieved from 95 283 solid-state synthesis paragraphs.

2.5 Text-mining solution-based synthesis recipes

A similar text-mining process was executed for solution-based synthesis recipes,[22] which include hydrothermal synthesis and solvothermal synthesis (where synthesis occurs in organic solvents or mixed aqueous/organic solvents). One major difference with solution synthesis recipes is that the precursors also include the volume and molarity of chemical species (*e.g.*, 60 mL of 0.2 M HCl). In all, there were 20 037 hydrothermal synthesis reactions and 15 638 precipitation synthesis reactions. To place the aqueous reactions onto Pourbaix diagrams, we also needed to process these precursor ratios into ion concentrations, solution pH, and effective redox potentials.[28] Because redox reactions can be complicated and are not unique, we did not attempt to build balanced reactions for the solutions dataset.

3 Evaluating the 4 Vs of the text-mined recipe dataset

Developing machine-learning models to predict materials synthesis recipes relies on a robust and diverse training dataset. Here, we characterize the synthesis recipe datasets through the 4 Vs of big data—volume, variety, velocity, and veracity. For each of the Vs, we discuss limitations in the text-mined synthesis recipe dataset, as outlined below:

Volume refers to the number of datapoints in a dataset. Although each dataset seemingly contains over 30 000 entries, many of these entries are redundant in their target chemical systems, meaning that there are fewer unique targets than anticipated.

Variety refers to the diversity of data. Here, we evaluate variety in two ways: first over chemical space, where we show that there is often limited coverage of text-mined recipes, even in materials systems that are experimentally known and well-represented in the Materials Project. Second, we show that even for target materials with numerous recipes, there is little variety in the reported reaction parameters or precursors.

Veracity is a measure of the quality and reliability of the data. We show that many synthesis paragraphs are missing essential information, even for a human chemist to synthesize the target material. Although there are some technical issues, broadly speaking there are many 'unknown unknowns' in materials synthesis that, if unpublished, can confound machine-learned synthesis models.

Velocity describes the speed at which data can be generated, collected and processed. Text-mining scientific literature requires data labeling and curation efforts by domain experts, which is a laborious and time-consuming process. Recently developed large language models like GPT-4 may accelerate this process, but human intervention is still necessary to prevent processing errors or potential hallucinations.

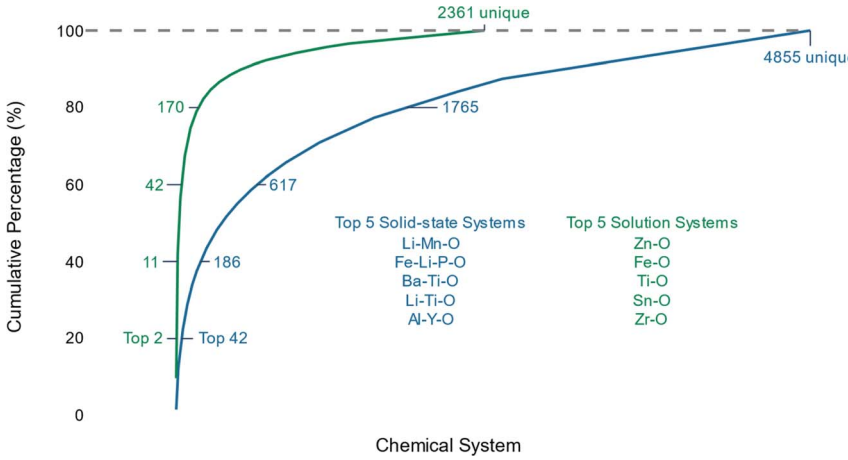

Fig. 1 A Pareto-principle distribution of materials recipes *versus* chemical systems appears in both the solid-state and solution-based text-mined synthesis recipes.

3.1 Volume

Although materials science typically operates in the 'small data' regime,[42,43] both datasets here contain >30 000 entries, which should be sufficient for classical machine learning.[44] However, further interrogation of the dataset shows that there are only 4855 unique chemical systems in the solid-state synthesis dataset, and 2361 unique chemical systems in the solution-based synthesis dataset. This means that many of the 30 000+ recipes are redundant with respect to target chemical systems.

In Fig. 1, we plot the cumulative distribution of the number of recipes *versus* unique chemical systems, revealing a shape reminiscent of the '80/20' Pareto principle—which states that 20% of the population represents 80% of a property distribution.[45] In the solution-based synthesis dataset, the first 7% (170 systems) of most common systems for solution-based synthesis accounts for 80% of the entire dataset, whereas the first 36% (1765 systems) of most common systems for solid-state synthesis composes 80% of that dataset. In fact, the first 2 and first 42 most common chemical systems make up 20% of the solution-based and solid-state datasets, respectively. These highly represented systems, listed beneath the curves in Fig. 1, are all oxides, specifically battery and catalyst materials. For the purposes of chemical discovery, redundant recipes in the same chemical space correspond to reduced coverage of diverse chemical systems to train generalizable machine-learning algorithms.

3.2 Variety across chemical space

Developing a machine-learned synthesis predictor that is effective across broad and novel chemical spaces requires diverse training data that covers various targets, precursors, and synthesis routes. For the 4855 unique target systems in the solid-state synthesis database, we find that the coverage of targets in chemical space is also quite limited.

In characterizing the chemical variety of the dataset, we find that oxides comprise 91.5% of the text-mined recipes. For ternary metal oxides, there are 785 M_1–M_2–O systems with known ternary oxides, of which text-mined synthesis recipes exist for 58% of these systems. Notable gaps in the text-mined recipe (TMR) dataset include ternary oxides containing precious metals, such as Au, Pd, Pt, Ir, Os, Ru and Re, as well as alkali metal oxides with Cs or Rb cations. Precious metal delafossites, such as $PtCoO_2$, are emerging as an exciting materials design space as their ultrahigh electrical conductivity exceeds even that of metallic Au.[46] However, synthesis recipes for these compounds may not be reliable if predicted from machine-learning algorithms trained on this TMR dataset. We note that one of the more reliable single-crystal growth recipes for $PtCoO_2$ proceeds *via* the metathesis reaction $LiCoO_2 + PtCl_2 \rightarrow PtCoO_2 + LiCl$,[47] where the LiCl byproduct is later removed with water. Such metathesis reactions are very effective when thermodynamic driving forces from standard oxide precursors are small.[48,49] However, metathesis reactions do not appear anywhere in the text-mined dataset. Other clever reactions that are facilitated by extrinsic chemical species[50] may also be unaccounted for by the reaction balancer.

Ternary sulfides are well-studied, with 532 experimentally known M_1–M_2–S systems, but they are underrepresented in the text-mined dataset, with only 72 spaces (13%) containing recipes, as visualized on the heatmap in Fig. 2 and summarized in Table 1. It is not immediately clear why the coverage of sulfides in the text-mined dataset is so poor. This is especially unfortunate as sulfides are an important design space for Li-ion-battery solid-state electrolytes,[54,55] thermoelectrics,[56,57] and solar cells.[58] Moreover, experimental validation of DFT-predicted sulfides can often be problematic. For example, Narayan *et al.* attempted to synthesize 24 new ternary sulfides and selenides, where 14 of them were predicted to be convex-hull stable, yet none of these ternary sulfides formed in experiment.[59] A machine-learned synthesis predictor in spaces where convex-hull stability predictors are weak would have been especially beneficial to have, but unfortunately the ternary-sulfide training set is not well-sampled by the text-mined recipe dataset.

A similar lack of text-mined recipes exists for the ternary metal nitrides, likely because much of the chemical exploration in this space was conducted prior to the year 2000,[2,60] where there is no coverage in the text-mined recipe database. Of course, one could supplement the TMR dataset with more recipes of ternary sulfides and nitrides, and other missing chemical spaces. Our point is to highlight that if one did not scrutinize the dataset and used predicted recipes for chemical systems where there are gaps in this dataset, the predictions would likely be unreliable.

3.3 Variety of recipes within a target materials system

Even for specific target materials with many published synthesis recipes, the variety of recipes is limited. We illustrate this in the Y–Ba–Cu–O system, which hosts the famous $YBa_2Cu_3O_{7-x}$ (YBCO) superconductor. The typical recipe for YBCO calls for Y_2O_3/$BaCO_3$/CuO precursors, which are ground in a mortar, compacted, pelletized, and baked in air. Even after many hours of baking, the synthesis reaction is often incomplete, so the pellets must be re-ground, re-pelletized, and re-baked until phase-pure YBCO is obtained.[61]

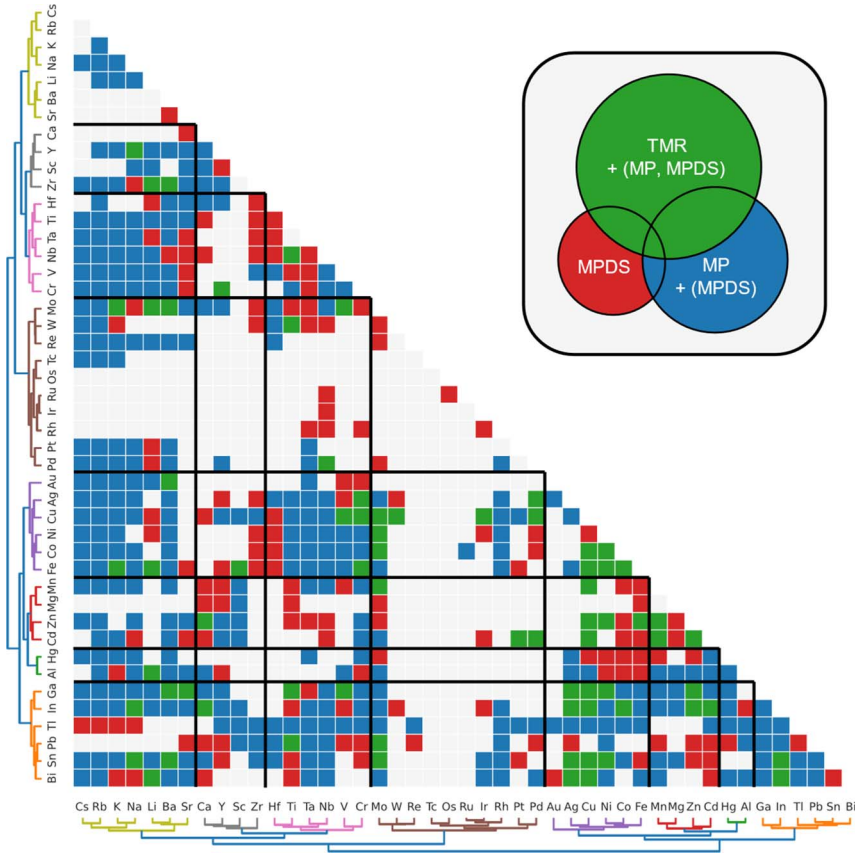

Fig. 2 Heatmap showing the coverage of text-mined sulfides compared to all known ternary metal sulfides. Green squares indicate known materials with text-mined recipes, blue squares indicate systems that are absent from the text-mined recipes dataset but have entries on the Materials Project[51] (and therefore have calculated energetics), and red squares indicate systems missing in the text-mined dataset and the Materials Project but present in the Pauling Files, hosted digitally by the Materials Platform for Data Science (MPDS).[52] Ternary spaces with no entries in any database are colored white. Chemical spaces are clustered hierarchically[53] to elucidate chemical trends in this space.

In the text-mined recipe dataset, there are 237 entries in the Y–Ba–Cu–O-* system. Fig. 3a plots the temperature–time distributions of these recipes, where the average reaction temperature is 950 ± 140 °C, and the average reaction time is

Table 1 Summary statistics for metal 1 (M_1) and metal 2 (M_2) dataset membership grids. The fraction of green, blue, and red entries is taken from the total number of explored (colored) systems

Material system	Total coverage (% possible)	In TMR, and MP or MPDS (green)	Not in TMR, but in MP (blue)	Known but not in TMR or MP (red)
M_1–M_2–O	785 (76%)	454 (58%)	198 (25%)	133 (17%)
M_1–M_2–S	532 (51%)	72 (13%)	317 (60%)	143 (27%)
M_1–M_2–N	296 (29%)	19 (6%)	174 (59%)	103 (35%)

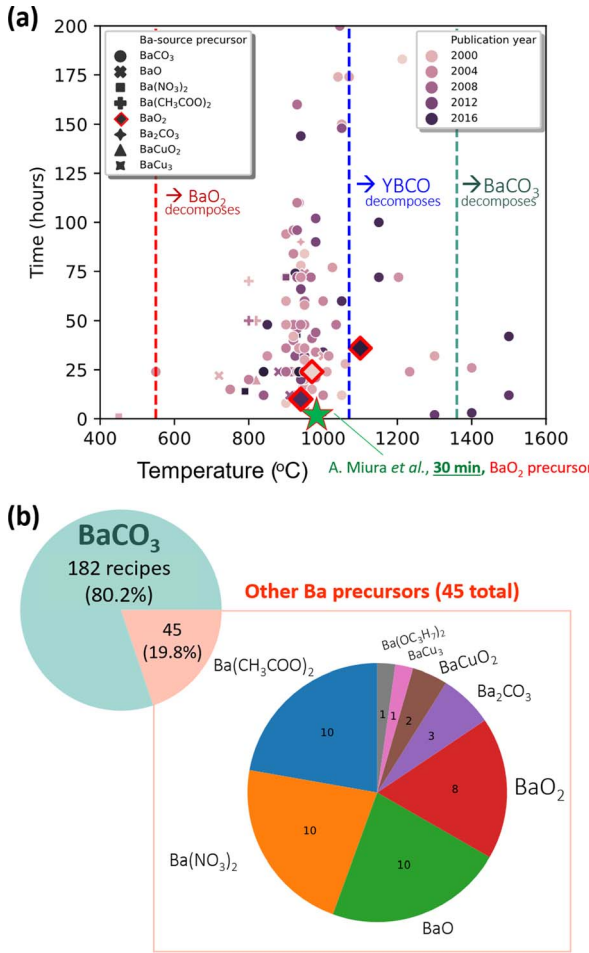

Fig. 3 Distribution of synthesis recipes in the Y–Ba–Cu–O-* chemical systems (a) for reaction temperatures and times and (b) for precursor selection. Emphasis is placed on BaO_2, which is a highly effective precursor in synthesizing $YBa_2Cu_3O_{6+x}$.

54 ± 81 h. In publications since the year 2000 (where our text-mined recipe dataset begins), the scatter in reaction temperatures and times is relatively uniform—there has not been a convergence on optimal reaction times or temperatures. Fig. 3b plots the most common Ba source and shows that 80% of recipes use $BaCO_3$ as the barium precursor.

In 2001, a short 1-page publication in the *Journal of Chemical Education*, 'Superconductor synthesis—an improvement', claimed that replacement of $BaCO_3$ with a BaO_2 precursor could reduce an undergraduate YBCO synthesis lab from 12 hours with regrinding and reannealing to 4 hours in one step.[62] (This recipe was not in the text-mined dataset). Inspired by this article, in Miura *et al.*[26] we used *in situ* synchrotron X-ray diffraction and transmission electron microscopy to observe that a BaO_2 precursor in fact yields YBCO in 25 minutes, and that the reaction occurs as fast as the sample can heat (in our case 30 °C min^{-1}). The YBCO product was indeed found to be superconducting. In separate work, this

fast YBCO reaction was attributed to the formation of a transient liquid phase,[63] and is now being exploited to accelerate the manufacturing of YBCO superconductors.[64]

As illustrated on Fig. 3a, $BaCO_3$ decomposes around 1360 °C,[65] whereas BaO_2 decomposes at 550 °C,[66] and the target compound YBCO decomposes above 1100 °C.[67] It is therefore quite surprising that 80% of literature recipes choose to start from $BaCO_3$ instead of BaO_2, which requires long reaction times and laborious regrinds and reanneals. In discussions with solid-state chemists, we learned that BaO_2 is not very air-stable, so a small fraction (~5%) will transform to $BaCO_3$ over time, making it difficult to weigh samples accurately and properly dose the Ba-stoichiometry. However, the trade-off for using BaO_2 instead of a $BaCO_3$ precursor is the opportunity for a far more efficient solid-state synthesis reaction. In personal communications,[68] we learned that the Cava group at Princeton has long-used $Ba(NO_3)_2$ as a precursor to cuprate superconductors instead of $BaCO_3$, even though this synthesis 'trick' does not seem to be adopted by the solid-state chemistry community at large. This YBCO example illustrates how chemists rely primarily on published recipes by previous chemists, rather than exploring synthesis parameter space for more optimal recipes.

3.4 Veracity

Veracity is a measure of the quality and reliability of data. In other words, veracity is a measure of how well the text-mining algorithm accurately extracted the recipe, and furthermore, how the recipes represent the true experimental procedures executed by the chemists. In general, veracity is inversely correlated with velocity and volume, as ensuring veracity becomes increasingly challenging as the extraction algorithms become more automated and the dataset grows.

From a technical perspective, both solid-state and solution text-mined datasets contain missing values. 1090 recipes from the solid-state dataset are either missing operations or only contain one uninformative operation type "StartingSynthesis" with no additional information. While the quality of the recipes at the chemistry level (only considering compositions) is high, with a reported 93% F1-score, further investigation of chemical formulae reveals some missing data. Many recipes represent 'condensed recipes', where chemical substitutions (indicated by a placeholder, *e.g.*, 'M' for metal) and variable stoichiometries (indicated by a subscript, *e.g.*, '*x*', '*y*', or '*z*') are left unspecified. Roughly 1 in 4 chemical compositions with variable stoichiometries are left unspecified in the solid-state dataset. When considering all attributes of a synthesis recipe, the F1-score drops to 51%.

Reproducibility is also a broader issue in inorganic materials synthesis.[69] Even human chemists often encounter difficulty when reproducing the synthesis of published compounds. Anecdotally, many important aspects of a reaction are often not reported or explained in a published paper. In informal conversations with chemists, we have learned that oxynitrides can be easier to synthesize in the winter than in the summer;[70] that metal *vs.* plastic reaction containers can change the polymorph of $CaCO_3$ precipitated from solution,[71] or that the grinding patterns in a mortar and pestle can influence the performance of synthesized battery materials.[72] From a machine-learning perspective, these unreported aspects of materials synthesis represent 'unknown unknowns,' which cannot even

be expected to be reproduced by human chemists, let alone be captured by text-mining or machine-learning models.[73]

It would benefit chemistry if there were a cultural shift in how synthesis protocols are reported in publications. Instead of only publishing the final successful synthesis recipe, it should also be encouraged to briefly discuss attempted reactions that were sensible but unsuccessful. This could help other chemists avoid pitfalls associated with a tricky reaction. Additionally, if a chemist is aware that environmental considerations or reaction setups were important for successful synthesis of their reported material—such as laboratory humidity and oxygen partial pressure, crucible material, precursors and their associated impurities,[74] mortar-and-pestle grinding technique, *etc*.—these details should definitely be mentioned.

3.5 Velocity

The velocity of a dataset is a measure of how quickly and easily synthesis recipes can be text-mined. Of course, once natural language processing tools are developed, generating new recipes from synthesis paragraphs is fast and scalable. However, labeling the training datasets for effective text-mining required, in our case, a tremendous annotation effort by domain experts. We had several afternoons where the ~40 members of the Ceder Group would label sentences and paragraphs together. Later, the first authors of the text-mining papers still had to manually examine and annotate hundreds more paragraphs each. This time-consuming process cannot be easily outsourced, since it relies on domain expertise. Because the Ceder Research Group had substantial human resources, such an annotation process was viable—in smaller academic groups, similar annotation tasks may be prohibitively labor-intensive.

Our text-mining work was performed before the advent of large language models (LLMs) like GPT, which now dominate natural language processing methodology.[39,40] With LLMs, it should be possible to parse and process published synthesis recipes with fewer manual annotation efforts. The latest OpenAI model available for fine-tuning 'gpt-3.5-turbo' costs $8.00 per 1 M input tokens. Considering the 53 538 synthesis paragraphs and their abstracts (roughly 400 tokens) it would cost roughly $170 per epoch to train – a small and manageable cost for research groups. However, acquiring the relevant papers to text-mine still requires either agreements with publishers or some upfront time to download and prepare inputs for LLMs. Domain expertise is also needed to confirm the fidelity of LLM-extracted recipes, and to check against hallucinations. However, while LLMs could improve velocity and dataset volume, it still would not enhance the variety of published materials or recipes, which again, are confined to narrow domains of chemical and synthesis parameter space due to anthropogenic biases.

4 Machine-learned synthesis prediction based on text-mined literature recipes

To predict materials synthesis recipes, a regression model should be used to predict reaction temperatures and times, and a classification model should be used to predict precursors. The training dataset should be featurized using physically relevant descriptors for the precursors and target materials, such as

elemental and compositional descriptors,[75] as well as thermochemical properties (formation energies, melting temperatures) either from DFT or experiment.[76–78] As a baseline, a machine-learned synthesis prediction should be more sophisticated than how a typical chemist would approach solid-state synthesis—which for a multicomponent oxide would be to start from the constituent binary oxide or carbonate precursors, grind, and fire at typical reaction temperatures for oxide synthesis (between 700–1100 °C).

Using XGBoost algorithms, Huo *et al.*[79] predicted reaction temperatures and times for carbonate and non-carbonate reactions. Predicted reaction temperatures fell within ±100 °C of the reported synthesis temperature in half of the predictions, and ±400 °C within 1.5× the interquartile distribution. The models achieved $R^2 \sim 0.5$–0.6 for heating temperature predictions and $R^2 \sim 0.3$ for $\log_{10}$(reaction times). Karpovich *et al.* trained a conditional variational autoencoder (CVAE) to predict reaction temperatures and times,[80] and obtained similar R^2 values to Huo *et al.* During feature analysis, the average precursor melting point was found to contribute most to reaction temperature predictions. Huo *et al.* noted that this fact is reminiscent of Tamman's Rule, which is a common empirical heuristic that solid-state reactions should be conducted above 1/3 to 1/2 of the precursor melting points.

To predict starting precursors, He *et al.*[81] developed "PrecursorSelector", an algorithm trained to predict the solid-state precursors of quaternary oxides using an encoding scheme for chemical similarity to previously synthesized compounds in the text-mined recipes. For a diversity of high-component oxide materials, including mixed-anion materials, they predict starting precursors that match literature reports with 82% accuracy when up to 5 precursor sets are included. Importantly, the encoding obtains experimentally reported precursors in fewer predictions than a baseline of simply choosing the most common oxide precursors. Using a different literature dataset, E. Kim *et al.*[82] predicted precursor materials using a conditional variational autoencoder (CVAE) trained directly from paragraphs, without any explicit domain knowledge. Given the target material $InWO_3$, which was not included in the training set, their model predicts the following precursor sets for both solution and solid-state reactions: (1) In_2S_3 + WCl_4, (2) $In(NO_3)_3$ + WCl_4, (3) In_2O_3 + WO_2, (4) In_2O_3 + WN, and (5) $InCl_3$ + Na_2WO_4. Reactions 3 and 4 are indeed thermodynamically spontaneous, and the fifth set of precursors was previously used for solution-based synthesis.[83]

In many ways, the machine-learning models outperform the baseline synthesis prediction, and are successful at predicting solid-state reaction temperatures, times, and precursors in alignment with how experimental chemists have previously synthesized inorganic materials.

However, because the models largely capture how chemists think about materials synthesis, it is arguable how much additional value these predictions bring to the experimental chemist.[84,85] The prediction of reaction temperatures in alignment with Tamman's Rule might be because Tamman's Rule drove the chemist to try those reaction temperatures in the first place—and not that these reaction temperatures are necessarily the most optimal. Likewise, reaction times may not be so meaningful. Chemists often leave reactions in the oven for 24 hours or overnight, meaning a regression on reaction times may be capturing this factor of human convenience, rather than some fundamental mechanism of materials synthesis. Finally, it is not necessarily erroneous to predict precursors that lack an

entry in the text-mined dataset, and a precursor that matches the literature dataset does not guarantee that it is an optimal precursor. As exemplified in the YBCO example (Section 3.3), a machine-learned synthesis predictor is unlikely to suggest the BaO_2 precursor given its rarity in the training dataset, even though it is a far superior precursor than $BaCO_3$ for YBCO synthesis.

Overall, our analysis here supports the claims of Jia *et al.*,[24] who argued that anthropogenic biases have narrowly confined exploration of reaction conditions to those similar to previous conditions. Moreover, in the hydrothermal synthesis of amine-templated metal oxides, Jia *et al.* found that neither the popularity of reactants, nor the choices of reaction conditions, were correlated to the success of the reaction, and that machine-learning models trained on randomized reaction datasets outperformed models trained on larger human-selected reaction datasets. Because anthropogenic biases are largely reflected in the 4 Vs of the text-mined dataset, these historical biases strongly limit the creativity and sophistication of downstream machine-learned synthesis recipe predictions.

5 Data-driven insights of fundamental solid-state reaction mechanisms

Since machine-learning models tend to reproduce the existing intuition of chemists, it may be more interesting to examine the anomalous recipes—which defy traditional intuition. How might an anomalous recipe come to be? In speaking with experimentalists, I learned that it often takes 10 to 30 trial-and-error experiments to optimize a publishable synthesis protocol. The initial exploratory trials typically involve grinding the common oxide precursors and firing them in an oven at a temperature commensurate with Tamman's Rule. If an XRD signal of the target compound is detected, the synthesis recipe is adjusted by trial-and-error until the target is produced near 100% yield. Once a phase-pure target is made, the experimentalist will move on to characterizing the functional properties of the material, which is usually their main interest. As soon as the target material is made, the recipe is set in stone—there are rarely any efforts to optimize the efficiency of a synthesis reaction further.

Therefore, if a published recipe reports common precursors and round values for reaction times and temperatures (6, 12 or 24 hours; at 700, 800, 900 or 1000 °C), this means that a chemist probably tried some simple initial experiments and the target phase formed easily. On the other hand, if a final reported recipe is complicated, for example using unusual precursors, laborious precursor mixing steps, or precise reaction temperatures (such as 835 °C) or times (3.75 hours), the chemist probably had to refine the synthesis parameters through a laborious trial-and-error optimization process.

Following this hunch, I classified the text-mined dataset by 'simple' or 'complex' recipes, as illustrated in Fig. 4. Most recipes for quaternary oxides report the three simple binary oxides as precursors. However, a recurrent observation emerged, where ternary oxides were being reported as precursors. Because we had the DOIs for each recipe, I could study the original papers. The following discussion on the synthesis of Sr_2FeMoO_6 provided fascinating and illuminating insights:

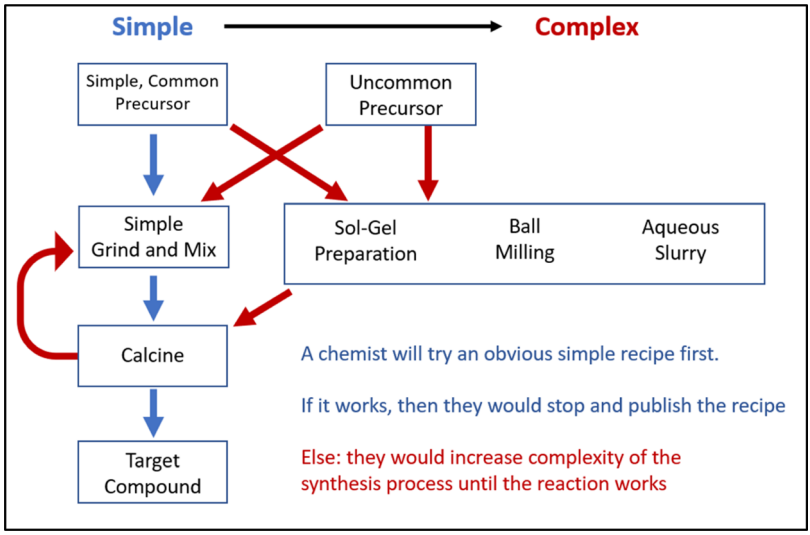

Fig. 4 A flow chart for identifying 'simple' text-mined synthesis reactions and inferring the more interesting 'complex' reactions from them.

"Sr_2FeMoO_6 samples were prepared by solid-state reaction. Two elaboration processes have been used in order to improve the purity of the samples. For the first one, which is close to the protocol used by most of groups, stoichiometric amounts of $SrCO_3$, Fe_2O_3 and MoO_3 were mixed, ground and calcined at 900 °C for 2 h in an Ar atmosphere. The calcined mixtures were reground, pressed and reduced for 1 h under current flow of 5% H_2/95% Ar at 700 °C. Afterwards the mixtures were sintered at 1200 °C under argon flow during 10 h.

Unfortunately, the last protocol does not allow one to obtain a pure Sr_2FeMoO_6 compound. Instead, $SrMoO_4$ is thermodynamically favored. Therefore, a segregation occurs which makes it impossible to obtain a pure phase.

To get rid of this difficulty, we have developed a sintering process in which only one reaction is performed at each step in order to avoid the formation of $SrMoO_4$. Therefore, in the first step, stoichiometric amounts of $SrCO_3$, Fe_2O_3 were mixed, ground and calcined at 1000 °C during 5 h under an Ar flow giving rise to $Sr_2FeO_{3.5}$ compound. Then stoichiometric amounts of $Sr_2FeO_{3.5}$, MoO_2 and MoO_3 were mixed, ground, pressed and sintered at 1200 °C during 2 h under N_2/H_2 flow." [86]

Although this report was purely phenomenological, we could use Materials Project energies to interpret the observation. First, we found that the SrO–MoO_3–Fe_2O_3 pseudo-ternary convex hull (Fig. 5a) is skewed—the $SrMoO_4$ phase is much deeper along the SrO–MoO_3 binary hull than Sr_2FeO_4 is along the SrO–Fe_2O_3 binary hull. This is captured by the following reaction energies, which are easily assessed using the MaterialsProject reaction calculator:

$$SrO + MoO_3 \rightarrow SrMoO_4, \quad \Delta H = -212 \text{ kJ per Sr}$$

$$2SrO + \tfrac{1}{2}Fe_2O_3 + \tfrac{1}{4}O_2 \rightarrow Sr_2FeO_{4-\delta}, \quad \Delta H = -53 \text{ kJ per Sr}$$

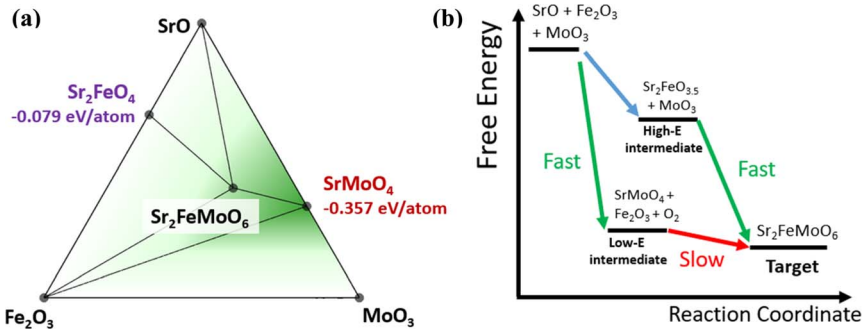

Fig. 5 (a) A ternary compound convex hull for the precursors used to synthesize Sr_2FeMoO_6. (b) A reaction diagram for Sr_2FeMoO_6, illustrating how the choice of a high-energy ternary oxide precursor will save more reaction energy for forming the target in the final step.

Three precursors can only meet in space at a single point. Therefore, reactions probably do not proceed between all three precursors reacting at once. Instead, it is more likely that solid-state reactions initiate at the interfaces between only two precursors at a time. The reaction SrO + MoO_3 is much more favorable than SrO + Fe_2O_3, meaning the large reaction driving force can promote fast reaction kinetics to form the low-energy $SrMoO_4$. Formation of $SrMoO_4$ consumes 93% of the total reaction energy, leaving only $\Delta H = -16$ kJ per Sr for $SrMoO_4$ to react with Fe_2O_3 to form the target phase, Sr_2FeMoO_6. On the other hand, by separately synthesizing $Sr_2FeO_{3.5}$, only 25% of the total reaction energy is consumed. This retains 75% of the reaction driving force for the second step of the reaction, where MoO_x is added, which enables phase-pure formation of Sr_2FeMoO_6 in only 2 hours.

Based on this insight, a synthesis strategy for quaternary oxides should be to first synthesize a high-energy ternary oxide intermediate, like $Sr_2FeO_{3.5}$, and then add the final metal oxide. We examined the text-mined recipe database and found 20 more examples where chemists used a ternary oxide precursor that was not initially favored by the initial pairwise reaction between three precursors, listed in Table 2. This suggests that other chemists may have independently come up with a similar synthesis strategy during their reaction optimization. This discussion associated with reaction design is usually buried deep in the 'results' section of a paper, which was not extracted in our text-mining algorithms. However, once we knew what patterns to look for, our text-mined dataset provided a valuable data source to find more historical examples of this strategy.

These examples led to a series of 'Eureka!' insights. First, we realized that we should reconsider the thermodynamic boundary conditions when analyzing solid-state reactions. While the overall reaction vessel has a stoichiometry fixed at the composition of the dosed precursors, the initial interfacial reactions between precursor powders should be compositionally unconstrained. In other words, the reactants do not 'know' the overall stoichiometry of the reaction vessel, they simply undergo whatever reaction is most favorable at their physical interface with another powder precursor. This led us to distinguish between 'non-equilibrium' and 'metastable' compounds from a convex-hull perspective. For example, $SrMoO_4$ is a hull-stable compound, but in the reaction to Sr_2FeMoO_6,

Table 2 Recipes where chemists synthesized a high-energy intermediate ternary oxide precursor when preparing a target quaternary oxide

Target	Reported precursors	Unusual ternary oxide precursor	Deepest hull ternary oxide	Reference
Sr_2FeMoO_6	$SrCO_3$, Fe_2O_3, MoO_2, $Sr_2FeO_{3.5}$	$Sr_2FeO_{3.5}$	$SrMoO_4$	86
$LiCr(MoO_4)_2$	Li_2MoO_4, $Cr(NO_3)_3 \cdot 9H_2O$, MoO_3	Li_2MoO_4	Li_2CrO_4	87
$Li_3Cr(MoO_4)_3$	Li_2MoO_4, $Cr(NO_3)_3 \cdot 9H_2O$, MoO_3	Li_2MoO_4	Li_2CrO_4	87
Li_2MnSiO_4	$Mn(CH_3COO)_2 \cdot 4H_2O$, Li_2SiO_3	Li_2SiO_3	Li_2MnO_3	88
$TlNd(MoO_4)_2$	Tl_2MoO_4, $Nd_2(MoO_4)_3$	Tl_2MoO_4	$Nd_2(MoO_4)_3$	89
$TlPr(MoO_4)_2$	Pr_6O_{11}, MoO_3, Tl_2MoO_3	Tl_2MoO_3	$Pr_2Mo_4O_{15}$	89
Sr_2CrTaO_6	$SrCO_3$, $CrTaO_4$	$CrTaO_4$	$SrCrO_4$	90
Ca_2CrTaO_6	$CaCO_3$, $CrTaO_4$	$CrTaO_4$	CaCrO4	90
$Ca_3Al_2(SiO_4)_3$	SiO_2, $Ca_3Al_2O_6$	$Ca_3Al_2O_6$	Ca_2SiO_4	91
$Pb(Zr_{0.52}Ti_{0.48})O_3$	TiO_2, $PbZrO_3$, PbO	$PbZrO_3$	Ca_2SiO_4	92
$Ba_3NiSb_2O_9$	$BaCO_3$, $NiSb_2O_6$	$NiSb_2O_6$	$Ba(SbO_3)_2$	93
$Pb(Fe_{0.5}Nb_{0.5})O_3$	$FeNbO_4$, PbO	$FeNbO_4$	$Ba_3V_2O_8$	94
$Pb(Ni_{0.33}Nb_{0.67})O_3$	Nb_2O_5, $Ni_4Nb_2O_9$, PbO	$Ni_4Nb_2O_9$	Nb_2NiO_6	95
$Pb(Zr_{0.5}Ti_{0.5})O_3$	$ZrTiO_4$, PbO	$ZrTiO_4$	Ti_3PbO_7	96
$Pb(Co_{0.33}Nb_{0.67})O_3$	$CoNb_2O_6$, PbO	$CoNb_2O_6$	$TiPbO_3$	96
$Bi(MgTi)_{0.5}O_3$	$MgTiO_3$, Bi_2O_3, TiO_2	$MgTiO_3$	$Mg(BiO_3)_2$	97
$Tl_2Pu(MoO_4)_3$	$Pu(MoO_4)_2$, Tl_2MoO_4	$Pu(MoO_4)_2$	$MgTiO_3$	98
$Tl_4Pu(MoO_4)_4$	$Pu(MoO_4)_2$, Tl_2MoO_4	$Pu(MoO_4)_2$	$MgTiO_3$	98
$CaSnSiO_5$	SiO_2, $CaSnO_3$	$CaSnO_3$	Ca_2SiO_4	99
$CuInGaO_4$	In_2O_3, CuO, $InGaO_3$, Ga_2O_3	$InGaO_3$	$GaCuO_2$	100

$SrMoO_4$ is a non-equilibrium intermediate, which persists kinetically due to small driving forces to complete the reaction to Sr_2FeMoO_6. This example illustrates how convex-hull stability is a limited and incomplete feature in the context of solid-state reactions. Instead, the topology and 'skew' of the convex hull towards low-energy thermodynamic 'traps' is a more relevant descriptor.[75]

Second, three precursors cannot react together at once, as three precursors only meet in space at a single point. It is much more probable that solid-state reactions proceed *via* interfacial reactions between two precursors at a time. Third, we realized that $T = 0$ K DFT convex hulls could effectively evaluate the competition between interfacial reactions, as the energy scale of solid-state reactions ($\sim$0.5 eV per atom) is much larger than the energy scale of DFT errors ($\sim$0.02 eV per atom),[101] $T\Delta S$ ($\sim$15 meV per atom), or the energy scale of kinetics (k_BT)—meaning we do not need time-consuming nudged elastic band calculations for diffusivity rates, or surface energy calculations to evaluate nucleation barriers.[102,103]

6 Hypothesis-driven synthesis science inspired by our literature-derived mechanism

These insights spurred a series of hypothesis-driven experiments into the fundamental synthesis science of oxides. In Bianchini *et al.*,[25] we tested the

hypothesis that interfacial reactions are compositionally unconstrained. For a target reaction of 0.66 Na_2O_2 + CoO → $Na_{0.66}CoO_2$, *in situ* synchrotron XRD found that the first phase to form had $NaCoO_2$ stoichiometry, as this product had the most exothermic compositionally unconstrained reaction. $NaCoO_2$ formed in the undesired O_3 polytype in only 6 minutes, and then required 4 hours of annealing to react with excess CoO precursors to transform to the desired ground-state P_2–$Na_{0.66}CoO_2$ phase.

Next, in Miura *et al.*[26] we tested if reactions between three precursors indeed occur two at a time. While synthesizing $YBa_2Cu_3O_{6+x}$, we saw that the reaction starting from the common $BaCO_3$ precursor initiated at the Y_2O_3|CuO interface, and that the reaction progression was slow because $BaCO_3$ does not decompose until 1100 °C. However, when starting with BaO_2, which decomposes at 550 °C, the BaO_2|CuO interface reacts first. Using *in situ* TEM, we directly observed the sequence of pairwise reactions, which finally resulted in superconducting $YBa_2Cu_3O_{6+x}$ in only 25 minutes.

Finally, we tried to design other analogues to the '$Sr_2FeO_{3.5}$' phase in the Sr_2FeMoO_6 reaction. In Chen *et al.*,[27] we used a robotic inorganic materials synthesis laboratory to synthesize 35 known quaternary oxides from two sets of precursors: (1) the three common binary oxide precursors *versus* (2) a high-energy ternary oxide phase + the remaining oxide. We found that the precursor set with a high-energy ternary oxide (*i.e.*, the $Sr_2FeO_{3.5}$ analogue) frequently outperformed the standard set of three binary oxide precursors. The robotic laboratory enabled us to validate this principle over a broad chemical space spanning 27 elements and 28 unique precursors. Examples of high-energy ternary oxide phases include $LiPO_3$, $LiBO_2$, and $LiNbO_3$, which do not appear in the text-mined recipe dataset. It is therefore very unlikely that a machine-learning model trained on the text-mined dataset would ever predict these highly effective precursors.

Our proposed mechanism regarding the selectivity of pairwise interfacial reactions—colloquially referred to as the 'ΔG_{max}' hypothesis—later drove the active learning reaction optimization algorithm ARROWS[3],[104] which was the decision-making engine behind the A-Lab, the self-driving autonomous robotic synthesis laboratory at Berkeley.[105] The ΔG_{max} hypothesis also was found to be important in hydrothermal synthesis, where a maximized thermodynamic driving force for precipitation also minimizes undesired kinetic byproducts.[28] The ΔG_{max} hypothesis does fail in some systems; in particular it does not reliably reproduce the first interfacial reaction in sulfides[106,107] or for intermetallic reactions.[108] Of course, there remains more work to be done in understanding fundamental synthesis science. However, the success of the ΔG_{max} hypothesis thus far offers an optimistic case study for how fundamental mechanisms can be inferred from large historical datasets of materials data.

7 Reflections (and lessons learned)

Are materials synthesis recipes from the past able to guide chemists on how to synthesize novel materials in the future? Yes, but probably not through machine-learning. This is not to say that machine learning is not useful—certainly sophisticated natural language processing methods were needed to build the text-mined dataset of 31 782 recipes. Specifically, we believe that the typical machine-learning pipeline of featurizing a data set, and then building a classification or

regression model for synthesis prediction, is unlikely to yield insightful predictions of synthesis recipes to novel materials.

As we argued in this retrospective, one critical issue in this machine-learning pipeline arises from limitations in the volume, variety, velocity and veracity of the text-mined recipe dataset. These limitations in the 4 Vs are not failures of natural language processing; in fact, this dataset is probably the best possible product given the practical challenges of text-mining the scientific literature. Rather, many limitations in the volume and variety of the dataset have anthropogenic origins.

From an experimental chemist's perspective, exploratory synthesis of new functional materials is risky, with many potential failure modes.[109] For chemists without access to *ab initio* predictions, it is much easier to arrive at publishable results by characterizing and tweaking known materials, instead of exploring new chemistries and structures. We believe that this human tendency to 'exploit' known materials rather than 'explore' new materials underlies the limited diversity of unique materials explored in the scientific literature—which of course is reflected in the text-mined recipe database. Furthermore, because synthesis is usually seen as a 'means to an end' (the end being materials properties and functionality), published synthesis recipes often do not represent the most efficient or optimized synthesis procedure. Follow-up studies on a given material also tend not to modify or revise published recipes. All in all, this risk aversion leads to a narrow variety of explored materials chemistries, as well as homogeneity in recipe design, which ultimately limits the sophistication of machine-learned synthesis prediction models.

A second critical issue in the typical machine-learning pipeline is the difficulty of featurizing a dataset when the operative physical mechanism is unknown. The current paradigm of machine learning in materials science generally proceeds by collecting a variety of materials descriptors, usually elemental properties and other features that can be conveniently pulled from databases. It is common to include as many features as possible, and hope that the machine-learning algorithm learns something new and interesting—something too complicated for humans to have anticipated. However, if the chosen features are unrelated to the essential underlying physics, the resulting machine-learning algorithms will likely make spurious associations.

In our retrospective here, we did derive new mechanistic insights into how pairwise reactions drive solid-state synthesis, and how clever precursor selection could facilitate more efficient synthesis pathways to high-component oxides. However, arriving at these insights required domain knowledge in materials thermodynamics and kinetics, and a visual abstraction of the microstructure of powder precursors in interfacial reactions. Arriving at the 'Eureka!' moment also required us to recognize how the convex hull was skewed towards certain low-energy competing phases. The skew and topology of the convex hull is not a feature one would likely include *a priori*. If, instead, the dataset was only featurized using DFT bulk formation energies, a machine-learning algorithm is unlikely to have picked up on this essential geometric aspect of the convex hull.

It seems unlikely that any existing machine-learning architecture could derive these physical insights in an unsupervised manner. Even large language models like GPT, which operate by predicting the next word in a sentence, do not have a visual representation of the microstructure of a reaction, nor access to the

thermochemical data needed to assess competing reaction pathways. Of course, once we have a physical principle, we can featurize later machine-learning models properly. As shown in the ARROWS[3] study,[104] synthesis optimization using the ΔG_{max} mechanism results in more efficient synthesis optimization than physics-agnostic statistical approaches. Our point is that it is premature to apply machine-learning techniques before hypothesizing a physical model, and that for the foreseeable future, physical models will probably need to be built by human experts, not machines.

This retrospective is not a criticism of data-driven approaches, as 'bigger data' will always add value and enable more analyses. Our text-mined recipe dataset is orders-of-magnitude larger than any similar dataset. This large database enabled the rapid testing of synthesis hypotheses against prior experimental observations, quantitative surveys of statistical distributions, and visualizations of broad trends across chemical space. Here, this dataset helped classify 'simple' and 'complex' recipes, from which we found the unusual $Sr_2FeO_{3.5}$ precursor for Sr_2FeMoO_6. One cannot search Google Scholar for 'unusual synthesis reaction', nor would this key insight be readily apparent from the title of the original paper, "elaboration and characterization of the Sr_2FeMoO_6 double perovskite".[86]

Going forward, we should not be dismayed by the anthropogenic limitations of the 4 Vs in historical datasets. Rather, we should be encouraged by the fact that scientific data is being generated at an exponentially increasing rate. In particular, the advent of robotic materials synthesis laboratories[27,105,110,111] offers the opportunity to digitize and store all the synthesis metadata that is usually unreported, including 'dark reactions'[112] that do not lead to successful synthesis. Moreover, robotic labs are more likely to maintain high synthesis precision and reproducibility, meaning we will likely produce more high-quality single-source experimental synthesis data in a few years than all the historical synthesis recipes that have been published thus far. We should be optimistic about this incoming deluge of data, which, if interpreted thoughtfully and physically by human experts (and assisted by machine-learning methods), will surely lead to transformative new science on how to synthesize and manufacture novel materials.

Data availability

The code and data needed to reproduce the figures presented are available at **https://github.com/nrdavid/faraday4Vs**.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

W. S. and N. D. acknowledge support through the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering. W. S. thanks leadership by Gerbrand Ceder for the text-mining synthesis project, along with the text-mining team of Tanjin He, Haoyan Huo, Olga Kononona, Amalie Trewartha, Chris Bartel,

Tiago Botari, Zheren Wang, Ziqin Rong, and Vahe Tshitoyan for valuable discussions on these topics.

References

1 G. Hautier, *et al.*, Finding nature's missing ternary oxide compounds using machine learning and density functional theory, *Chem. Mater.*, 2010, **22**(12), 3762–3767.
2 W. Sun, *et al.*, A map of the inorganic ternary metal nitrides, *Nat. Mater.*, 2019, **18**(7), 732–739.
3 A. Merchant, *et al.*, Scaling deep learning for materials discovery, *Nature*, 2023, **624**(7990), 80–85.
4 A. Jain, Y. Shin and K. Persson, Computational predictions of energy materials using density functional theory, *Nat. Rev. Mater.*, 2016, **1**, 15004.
5 S. Curtarolo, G. Hart, M. Nardelli, *et al.*, The high-throughput highway to computational materials design, *Nat. Mater.*, 2013, **12**, 191–201.
6 A. Zunger, Inverse design in search of materials with target functionalities, *Nat. Rev. Chem.*, 2018, **2**, 0121.
7 S. P. Ong, *et al.*, Li−Fe−P−O2 phase diagram from first principles calculations, *Chem. Mater.*, 2008, **20**(5), 1798–1807.
8 W. Sun, *et al.*, The thermodynamic scale of inorganic crystalline metastability, *Sci. Adv.*, 2016, **2**, e1600225.
9 M. Aykol, *et al.*, Thermodynamic limit for synthesis of metastable inorganic materials, *Sci. Adv.*, 2018, **4**(4), eaaq0148.
10 K. Kovnir, Predictive synthesis, *Chem. Mater.*, 2021, **33**(13), 4835–4841.
11 A. K. Cheetham, Ram Seshadri, and Fred Wudl. "Chemical synthesis and materials discovery, *Nat. Synth.*, 2022, **1**(7), 514–520.
12 J. R. Neilson, M. J. McDermott and K. A. Persson, Modernist materials synthesis: Finding thermodynamic shortcuts with hyperdimensional chemistry, *J. Mater. Res.*, 2023, **38**(11), 2885–2893.
13 E. J. Corey, Robert Robinson Lecture. retrosynthetic thinking—essentials and examples, *Chem. Soc. Rev.*, 1988, **17**, 111–133.
14 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, Prediction of organic reaction outcomes using machine learning, *ACS Cent. Sci.*, 2017, **3**, 434–443.
15 C. W. Coley, W. H. Green and K. F. Jensen, Machine learning in computer-aided synthesis planning, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.
16 S. Hong, H. H. Zhuo, K. Jin, *et al.*, Retrosynthetic planning with experience-guided Monte Carlo tree search, *Commun. Chem.*, 2023, **6**, 120.
17 M. Segler, M. Preuss and M. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature*, 2018, **555**, 604–610.
18 C. W. Coley, W. H. Green and K. F. Jensen, Machine learning in computer-aided synthesis planning, *Acc. Chem. Res.*, 2018, **51**(5), 1281–1289.
19 SciFinder; Chemical Abstracts Service, available at **https://scifinder.cas.org**, 2010.
20 Elsevier, Reaxys, available at **https://www.elsevier.com/products/reaxys**, 2018.
21 O. Kononova, H. Huo, T. He, *et al.*, Text-mined dataset of inorganic materials synthesis recipes, *Sci. Data*, 2019, **6**, 203.

22 Z. Wang, *et al.*, Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature, *Sci. Data*, 2022, **9**(1), 231.
23 J. Fenn and M. Raskino, *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*, Harvard Business Press, 2008.
24 X. Jia, *et al.*, Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis, *Nature*, 2019, **573**(7773), 251–255.
25 M. Bianchini, J. Wang, R. J. Clément, *et al.*, The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides, *Nat. Mater.*, 2020, **19**, 1088–1095.
26 A. Miura, *et al.*, Observing and modeling the sequential pairwise reactions that drive solid-state ceramic synthesis, *Adv.Mater.*, 2021, **33**, 2100312.
27 J. Chen, S. R. Cross, L. J. Miara, *et al.*, Navigating phase diagram complexity to guide robotic inorganic materials synthesis, *Nat. Synth.*, 2024, **3**, 606–614.
28 Z. Wang, Y. Sun, K. Cruse, *et al.*, Optimal thermodynamic conditions to minimize kinetic by-products in aqueous materials synthesis, *Nat. Synth.*, 2024, **3**, 527–536.
29 K. Cruse, A. Trewartha, S. Lee, *et al.*, Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities, *Sci. Data*, 2022, **9**, 234.
30 Y. Liu, *et al.*, Text mining of hypereutectic Al-Si alloys literature based on active learning, *Mater. Today Commun.*, 2021, **26**, 102032.
31 W. Wang, X. Jiang, S. Tian, *et al.*, Automated pipeline for superalloy data by text mining, *npj Comput. Mater.*, 2022, **8**, 9.
32 Y. Zhang, *et al.*, Unleashing the power of knowledge extraction from scientific literature in catalysis, *J. Chem. Inf. Model.*, 2022, **62**(14), 3316–3330.
33 L. Bandeira, H. Ferreira, J. M. Almeida, A. Jardim de Paula and G. M. Dalpian, CO_2 reduction beyond copper-based catalysts: A Natural Language Processing Review from the scientific literature, *ACS Sustainable Chem. Eng.*, 2024, **12**, 4411–4422.
34 S. Huang and J. M. Cole, A database of battery materials auto-generated using ChemDataExtractor, *Sci. Data*, 2020, **7**(1), 260.
35 P. Shetty, *et al.*, A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing, *npj Comput. Mater.*, 2023, **9**(1), 52.
36 P. Kumar, S. Kabra and J. M. Cole, Auto-generating databases of yield strength and grain size using chemdataextractor, *Sci. Data*, 2022, **9**(1), 292.
37 H. Huo, Z. Rong, O. Kononova, *et al.*, Semi-supervised machine-learning classification of materials synthesis procedures, *npj Comput. Mater.*, 2019, **5**, 62.
38 T. He, *et al.*, Similarity of precursors in solid-state synthesis as text-mined from scientific literature, *Chem. Mater.*, 2020, **32**, 7861–7873.
39 J. Dagdelen, *et al.*, Structured information extraction from scientific text with large language models, *Nat. Commun.*, 2024, **15**(1), 1418.
40 S. Kim, Y. Jung and J. Schrier, Large Language Models for Inorganic Synthesis Predictions, *ChemRxiv*, 2024, preprint, DOI: **10.26434/chemrxiv-2024-9bmfj-v2**.
41 A. Jain, *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**, 1.
42 P. Xu, X. Ji, M. Li, *et al.*, Small data machine learning in materials science, *npj Comput. Mater.*, 2023, **9**, 42.

43 Y. Zhang and C. Ling, A strategy to apply machine learning to small datasets in materials science, *npj Comput. Mater.*, 2018, **4**, 25.
44 F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
45 L. Amoroso, Vilfredo Pareto, *Econometrica*, 1938, **6**, 1.
46 P. Kushwaha, *et al.*, Nearly free electrons in a 5 d delafossite oxide metal, *Sci. Adv.*, 2015, **1**(9), e1500692.
47 M. Tanaka, M. Hasegawa and H. Takei, Crystal growth of PdCoO2, PtCoO2 and their solid-solution with delafossite structure, *J. Cryst. Growth*, 1997, **173**(3–4), 440–445.
48 A. Miura, *et al.*, Selective metathesis synthesis of MgCr 2 S 4 by control of thermodynamic driving forces, *Mater. Horiz.*, 2020, **7**(5), 1310–1316.
49 A. J. Martinolich and R. N. James, Toward reaction-by-design: achieving kinetic control of solid state chemistry with metathesis, *Chem. Mater.*, 2017, **29**(2), 479–489.
50 Y. Zhang, *et al.*, A novel chemical pathway for energy efficient production of Ti metal from upgraded titanium slag, *Chem. Eng. J.*, 2016, **286**, 517–527.
51 A. Jain, *et al.*, Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**, 011002.
52 E. Blokhin, *Materials Platform for Data Science: from Big Data towards Materials Genome*, 2019.
53 W. Sun, C. J. Bartel, E. Arca, *et al.*, A map of the inorganic ternary metal nitrides, *Nat. Mater.*, 2019, **18**, 732–739.
54 Y. Wang, W. Richards, S. Ong, *et al.*, Design principles for solid-state lithium superionic conductors, *Nat. Mater.*, 2015, **14**, 1026–1031.
55 Q. Zhang, *et al.*, Sulfide-based solid-state electrolytes: Synthesis, stability, and potential for all-solid-state batteries, *Adv. Mater.*, 2019, **31**, 1901131.
56 Z.-H. Ge, *et al.*, Low-cost, abundant binary sulfides as promising thermoelectric materials, *Mater. Today*, 2016, **19**, 227–239.
57 A. V. Powell, Recent developments in Earth-abundant copper-sulfide thermoelectric materials, *J. Appl. Phys.*, 2019, **126**, 100901.
58 M. P. Suryawanshi, *et al.*, CZTS based thin film solar cells: A status review, *Mater. Technol.*, 2013, **28**, 98–109.
59 A. Narayan, *et al.*, Computational and experimental investigation for new transition metal selenides and sulfides: The importance of experimental verification for stability, *Phys. Rev. B*, 2016, **94**, 045105.
60 F. J. DiSalvo and S. J. Clarke, Ternary nitrides: A rapidly growing class of new materials, *Curr. Opin. Solid State Mater. Sci.*, 1996, **1**, 241–249.
61 P. Grant, Do-it-yourself superconductors, *New Scientist*, 1987, vol. 115, pp. 36–39.
62 B. D. Fahlman, Superconductor synthesis—an improvement, *J. Chem. Educ.*, 2001, **78**(9), 1182.
63 L. Soler, *et al.*, Ultrafast transient liquid assisted growth of high current density superconducting films, *Nat. Commun.*, 2020, **11**(1), 344.
64 T. Puig, J. Gutierrez and X. Obradors, Impact of high growth rates on the microstructure and vortex pinning of high-temperature superconducting coated conductors, *Nat. Rev. Phys.*, 2024, **6**(2), 132–148.
65 I. Arvanitidis, D. Siche and S. Seetharaman, A study of the thermal decomposition of BaCO3, *Metall. Mater. Trans. B*, 1996, **27**, 409–416.

66 M. J. Tribelhorn and M. E. Brown, Thermal decomposition of barium and strontium peroxides, *Thermochim. Acta*, 1995, **255**, 143–154.
67 W. Wong-Ng and L. P. Cook, Liquidus diagram of the Ba-Y-Cu-O system in the vicinity of the Ba2YCu3O6+ x phase field, *J. Res. Natl. Inst. Stand. Technol.*, 1998, **103**(4), 379.
68 Private communications with Leslie Schoop.
69 R. B. Canty and K. F. Jensen, Sharing reproducible synthesis recipes, *Nat. Synth.*, 2024, **3**, 428–429.
70 Private communication with Akira Miura.
71 Private communication with Jim Deyoreo.
72 Private communication with Byungwoo Kang.
73 N. David, W. Sun and C. W. Coley, The promise and pitfalls of AI for molecular and materials synthesis, *Nat. Comput. Sci.*, 2023, **3**, 362–364.
74 L. M. Liz-Marzán, C. R. Kagan and J. E. Millstone, Reproducibility in nanocrystal synthesis? watch out for impurities, *ACS Nano*, 2020, **14**, 6359–6361.
75 L. M. Ghiringhelli, *et al.*, Big data of materials science: critical role of the descriptor, *Phys. Rev. Lett.*, 2015, **114**(10), 105503.
76 L. Ward, *et al.*, Matminer: An open source toolkit for materials data mining, *Comput. Mater. Sci.*, 2018, **152**, 60–69.
77 K. T. Butler, *et al.*, Machine learning for molecular and materials science, *Nature*, 2018, **559**(7715), 547–555.
78 L. Ward, *et al.*, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Comput. Mater.*, 2016, **2**(1), 1–7.
79 H. Huo, *et al.*, Machine-learning rationalization and prediction of solid-state synthesis conditions, *Chem. Mater.*, 2022, **34**, 7323–7336.
80 C. Karpovich, E. Pan, Z. Jensen and E. Olivetti, Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction, *Chem. Mater.*, 2023, **35**, 1062–1079.
81 T. He, *et al.*, Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature, *Sci. Adv.*, 2023, **9**, eadg8180.
82 E. Kim, *et al.*, Inorganic Materials Synthesis Planning with literature-trained Neural Networks, *J. Chem. Inf. Model.*, 2020, **60**, 1194–1201.
83 J. Kamalakkannan, V. L. Chandraboss and S. Senthilvelan, Synthesis and characterization of $InWO_3$–TiO_2 nanocomposite material and multi application, *World Sci. News*, 2016, **58**, 97–121.
84 S. K. Kauwe, *et al.*, Can machine learning find extraordinary materials?, *Comput. Mater. Sci.*, 2020, **174**, 109498.
85 N. Wagner and J. M. Rondinelli, Theory-guided machine learning in materials science, *Front. Mater.*, 2016, **3**, 28.
86 A. Dinia, *et al.*, Elaboration and characterization of the Sr_2FeMoO_6 double perovskite, *Catal. Today*, 2004, **89**, 297–302.
87 A. Sarapulova, *et al.*, Crystal structure and magnetic properties of Li,Cr-containing molybdates $Li_3Cr(MoO_4)_3$, $LiCr(MoO_4)_2$ and $Li_{1.8}Cr_{1.2}(MoO_4)_3$, *J. Solid State Chem.*, 2009, **182**, 3262–3268.
88 W. Liu, Y. Xu and R. Yang, Synthesis, characterization and electrochemical performance of Li_2MnSiO_4/C cathode material by solid-state reaction, *J. Alloys Compd.*, 2009, **480**, L1–L4.

89 V. V. Atuchin, *et al.*, Structural and vibrational properties of microcrystalline $TlM(MoO_4)_2$ (M=Nd, Pr) molybdates, *Opt. Mater.*, 2012, **34**, 812–816.
90 M. C. L. Cheah, *et al.*, Synthesis and structures of chromium double perovskites A_2CrTaO_6 (A=Sr, Ca), *Phys. B*, 2006, **385–386**, 184–186.
91 J. M. Rivas Mercury, *et al.*, Solid-state ^{27}Al and ^{29}Si NMR investigations on Si-substituted hydrogarnets, *Acta Mater.*, 2007, **55**, 1183–1191.
92 W. Chaisan, *et al.*, Dielectric properties of solid solutions in the lead zirconate titanate–barium titanate system prepared by a modified mixed-oxide method, *Mater. Lett.*, 2005, **59**, 3732–3737.
93 C. Darie, *et al.*, A new high pressure form of $Ba_3NiSb_2O_9$, *J. Solid State Chem.*, 2016, **237**, 166–173.
94 Y.-C. Liou, Effect of heating rate on properties of $Pb(Fe_{1/2}Nb_{1/2})O_3$ ceramics produced by simplified wolframite route, *Ceram. Int.*, 2004, **30**, 567–569.
95 O. Khamman, R. Yimnirun and S. Ananta, Effect of calcination conditions on phase formation and particle size of lead nickel niobate powders synthesized by using $Ni_4Nb_2O_9$ precursor, *Mater. Lett.*, 2007, **61**, 4466–4470.
96 A. Prasatkhetragarn, *et al.*, Phase formation, microstructure, and dielectric properties of (1-x)PZT-(x)PCN ceramics, *Mater. Lett.*, 2009, **63**, 1281–1284.
97 M. D. Snel, W. A. Groen and G. de With, Investigation of the new piezoelectric system $(1-x)Bi(MgTi)_{0.5}O_3$ –$xPbTiO_3$, *J. Eur. Ceram. Soc.*, 2005, **25**, 3229–3233.
98 N. D. Dahale, M. Keskar, S. K. Sali and V. Venugopal, Preparation and characterisation of $Tl_2Pu(MoO_4)_3$ and $Tl_4Pu(MoO_4)_4$ in Tl–Pu–Mo–O system by X-ray and thermal methods, *J. Nucl. Mater.*, 2008, **376**, 129–132.
99 M. Mouyane, *et al.*, Original electrochemical mechanisms of $CaSnO_3$ and $CaSnSiO_5$ as anode materials for Li-ion batteries, *J. Solid State Chem.*, 2011, **184**, 2877–2886.
100 D. P. Cann, *et al.*, Conductivity anomaly in $CuInGaO_4$ and $CuIn_2Ga_2O_7$ ceramics, *Mater. Lett.*, 2004, **58**, 2147–2151.
101 G. Hautier, S. P. Ong, A. Jain, C. J. Moore and G. Ceder, Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2012, **85**.
102 W. Sun, *et al.*, Nucleation of metastable aragonite CaCO3 in seawater, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 3199–3204.
103 W. Sun, D. A. Kitchaev, D. Kramer, *et al.*, Non-equilibrium crystallization pathways of manganese oxides in aqueous solution, *Nat. Commun.*, 2019, **10**, 573.
104 N. J. Szymanski, P. Nevatia, C. J. Bartel, *et al.*, Autonomous and dynamic precursor selection for solid-state materials synthesis, *Nat. Commun.*, 2023, **14**, 6956.
105 N. J. Szymanski, B. Rendy, Y. Fei, *et al.*, An autonomous laboratory for the accelerated synthesis of novel materials, *Nature*, 2023, **624**, 86–91.
106 R. McClain, *et al.*, Mechanistic insight of $KBiQ_2$ (Q = S, SE) using panoramic synthesis towards synthesis-by-design, *Chem. Sci.*, 2021, **12**, 1378–1391.
107 Z. Jiang, A. Ramanathan and D. P. Shoemaker, In situ identification of kinetic factors that expedite inorganic crystal formation and discovery, *J. Mater. Chem. C*, 2017, **5**, 5709–5717.
108 R. Pretorius, A. M. Vredenberg, F. W. Saris and R. de Reus, Prediction of phase formation sequence and phase stability in binary metal-aluminum

thin-film systems using the effective heat of formation rule, *J. Appl. Phys.*, 1991, **70**, 3636–3646.

109 R. Woods-Robinson, *et al.*, From design to device: challenges and opportunities in computational discovery of p-type transparent conductors, *arXiv*, 2024, preprint arXiv:2402.19378DOI: **10.48550/arXiv.2402.19378**.

110 B. P. MacLeod, *et al.*, Self-driving laboratory for accelerated discovery of thin-film materials, *Sci. Adv.*, 2020, **6**.

111 B. Burger, P. M. Maffettone, V. V. Gusev, *et al.*, A mobile robotic chemist, *Nature*, 2020, **583**, 237–241.

112 P. Raccuglia, K. Elbert, P. Adler, *et al.*, Machine-learning-assisted materials discovery using failed experiments, *Nature*, 2016, **533**, 73–76.