Title

Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature

Authors

He, Tanjin
Huo, Haoyan
Bartel, Christopher J
et al.

Peer reviewed

MATERIALS SCIENCE

Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature

Tanjin He[1,2], Haoyan Huo[1,2], Christopher J. Bartel[1,2,3], Zheren Wang[1,2], Kevin Cruse[1,2], Gerbrand Ceder[1,2]*

Synthesis prediction is a key accelerator for the rapid design of advanced materials. However, determining synthesis variables such as the choice of precursor materials is challenging for inorganic materials because the sequence of reactions during heating is not well understood. In this work, we use a knowledge base of 29,900 solid-state synthesis recipes, text-mined from the scientific literature, to automatically learn which precursors to recommend for the synthesis of a novel target material. The data-driven approach learns chemical similarity of materials and refers the synthesis of a new target to precedent synthesis procedures of similar materials, mimicking human synthesis design. When proposing five precursor sets for each of 2654 unseen test target materials, the recommendation strategy achieves a success rate of at least 82%. Our approach captures decades of heuristic synthesis data in a mathematical form, making it accessible for use in recommendation engines and autonomous laboratories.

INTRODUCTION

Predictive synthesis is a grand challenge that would accelerate the discovery of advanced inorganic materials (1). The complexity of synthesis mainly originates from the interactions of many design variables, including the diversity of precursor candidates for each element in the target material (oxides, hydroxides, carbonates, etc.), the experimental conditions (temperature, atmosphere, etc.), and the chronological organization of operations (mixing, firing, reducing, etc.). Properly selecting the combination of experimental variables is crucial and demanding for successful synthesis (2–4). Here, we focus on the rational design of precursor combinations for solid-state synthesis, a widely used approach to create inorganic materials.

Because of the lack of a general theory for how phases evolve during heating, synthesis design is mostly driven by heuristics and basic chemical insights. Unlike the success of retrosynthesis and automated design for organic materials based on the conservation and transformation of functional groups (5–7), the mechanisms underlying inorganic solid-state synthesis are not well understood (6, 8–10). Here, we define a recipe to be any structured information about a target material, including the precursors, operations, conditions, and other experimental details. Experimental researchers usually approach a new inorganic synthesis by manually looking up similar materials in the literature and repurposing precedent recipes for a novel material. However, deciding what materials are similar and thus where to look is often driven by intuition and limited by individuals' personal experience in specific chemical spaces, hindering the ability to rapidly design syntheses for new chemistries. With the emergence of large-scale materials synthesis datasets from text-mining efforts (11–14), it is becoming possible to statistically learn the similarity of materials and the correlation of their synthesis variables in a more systematic and quantitative fashion, and provide such tools as a guide to scientists when approaching the synthesis of novel compounds.

Several studies have demonstrated the promise of building general models for the predictive synthesis of inorganic materials. Aykol et al. (15) and McDermott et al. (16) proposed heuristic models to rank the favorability of synthesis reactions or pathways based on thermodynamic metrics such as the reaction energy, nucleation barrier, and the number of competing phases. Kim et al. (17) used the stochasticity of a conditional variational autoencoder model to generate various samples of synthesis actions and precursors for the target material. Huo et al. (18) predicted synthesis conditions using large solid-state synthesis datasets text-mined from scientific journal articles. An interesting yet unexplored angle is to machine learn how the precursors of different target materials are shared and varied to enable the recommendation of multiple synthesis recipes with some ranked potential of success. In addition, extending the assessment from specific case studies to a large test set is also valuable for the development and improvement of predictive synthesis models.

We propose a precursor recommendation strategy (Fig. 1) based on machine-learned similarity of materials to automate the literature-based approach used by experimental researchers. Inspired by natural language processing models (19–21), we designed an encoding neural network to learn the vectorized representation of a material based on its corresponding precursors for the quantification of materials similarity. Assuming that the target material can be synthesized using an experimental design adapted from a similar material, synthesis variables such as precursors, operations, and conditions can be proposed and ranked by querying the knowledge base of previously synthesized materials. In this work, we applied the recommendation strategy to predict precursors for 2654 test target materials in a historical validation. Learning from a knowledge base of 29,900 synthesis reactions text-mined from the

[1]Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455, USA.
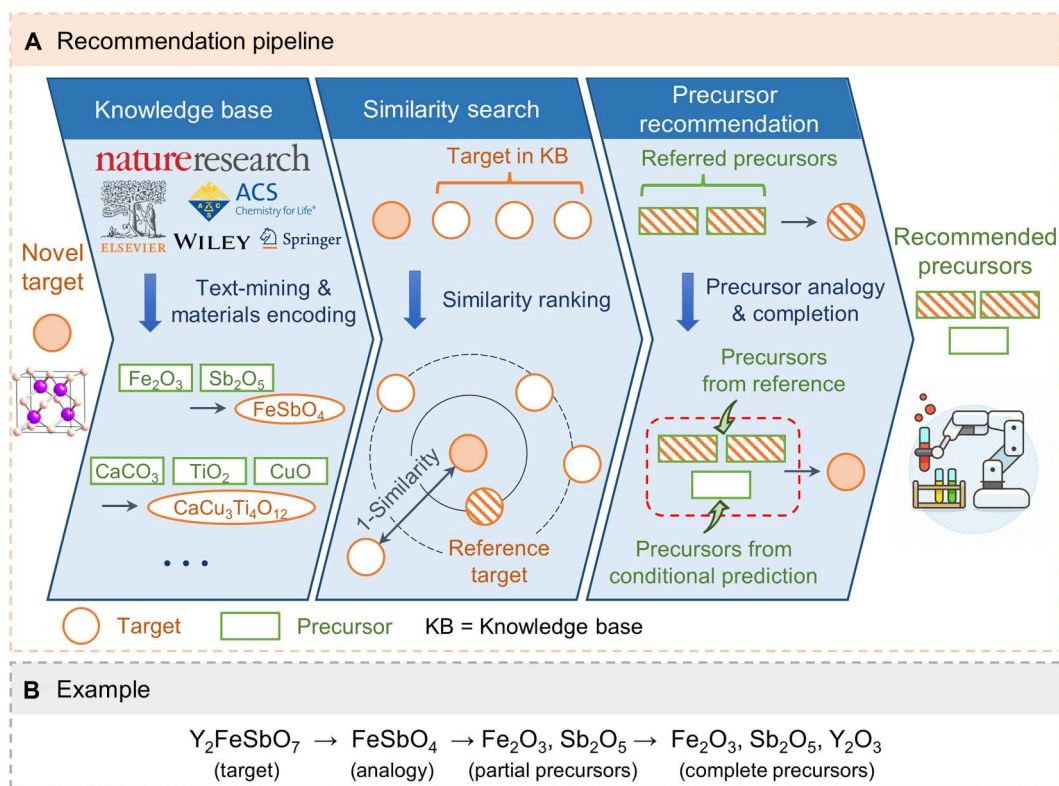*Corresponding author. Email: gceder@berkeley.edu

Fig. 1. Precursor recommendation strategy. (**A**) Pipeline for precursor recommendation consisting of three steps: (1) digitize target materials in the synthesis knowledge base text-mined from scientific literature, (2) rank target materials in the knowledge base according to the similarity to the novel target, and (3) recommend precursors based on analogy to the most similar target. (**B**) An example of precursor recommendation for Y_2FeSbO_7 by referring to the synthesis of $FeSbO_4$.

scientific literature, we demonstrate that the algorithm can acquire chemical knowledge on materials similarity via self-supervised learning and make promising decisions on precursor selection. Our quantitative recommendation pipeline captures how experimental researchers learn synthesis from the literature and enables rational and rapid precursor selection for new inorganic materials. It also provides meaningful initial solutions in the active learning and decision-making process for autonomous synthesis.

RESULTS

We begin with statistical insights from solid-state synthesis experiments reported in 24,304 papers (*11*) to better understand the problem of precursor selection (the "Problem of precursor selection" section). Because a universal model for solid-state synthesis has not yet been established, we use a data-driven method to recommend potential precursor sets for the given target material (Fig. 1). The recommendation pipeline consists of three steps: (i) an encoding model to digitize the target material as well as known materials in the knowledge base (the "Materials encoding for precursor selection" section), (ii) similarity query based on the materials encoding to identify a reference material that is most similar to the target (the "Similarity of target materials" section), and (iii) recipe completion to (a) compile the precursors referred from the reference material and (b) add any possibly missed precursors if element conservation is not achieved using conditional predictions based on referred precursors (the "Recommendation of precursor materials" section).

Problem of precursor selection

In the solid-state synthesis of inorganic materials, precursor selection plays a crucial role in governing the synthesis pathway by yielding intermediates that may lead to the desired material or alternative phases (*2–4*). For each metal/metalloid element, one precursor is often used predominantly over all others, which we denote as the common precursor (*22*). However, in a solid-state synthesis dataset of 33,343 experimental recipes extracted from 24,304 materials science papers (*11*), we find that approximately half of the target materials were synthesized using at least one uncommon precursor. Figure 2A presents the fraction of targets in the text-mined dataset (*11*) that can be achieved as one increases the number of available precursors. The precursors on the *x* axis are ordered by the relative frequency with which they are used to bring a specific element into a synthesis target. Uncommon precursors may be used for a variety of reasons including synthetic constraints (e.g., temperature and time), purity, morphology, and anthropogenic factors (*2, 22, 23*).

In addition, a probability analysis of the text-mined dataset indicates that precursors for different chemical elements are not randomly combined. The joint probability to select a specific precursor pair (A_i, B_i) can be compared to the marginal probability to select A_i for element Ele_a and B_i for Ele_b. If the choices of A_i and B_i are independent, then the joint probability should equal the product of the marginal probabilities, namely, $P(A_i, B_i) = P(A_i)P(B_i)$. However, inspection of 6472 pairs of precursors from our text-mined dataset (Fig. 2B) reveals that many show a strong dependency
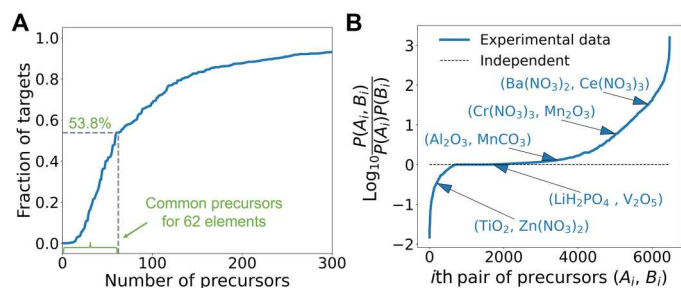
Fig. 2. Usage of precursors in solid-state synthesis. (**A**) Fraction of targets that can be synthesized with limited number of available precursors. The precursors are ordered by relative frequency per metal/metalloid element. Precursors for 62 elements are considered. A target is included if at least one reported reaction for that target was performed with the available precursors. (**B**) Pairwise dependency of precursors A_i and B_i characterized by $\frac{P(A_i, B_i)}{P(A_i)P(B_i)}$. Probability is estimated from the frequency of occurrence in the solid-state synthesis dataset. The value of $\log_{10} \frac{P(A_i, B_i)}{P(A_i)P(B_i)}$ is zero when A_i and B_i are independent, positive when A_i and B_i tends to be used in the same experiment more frequently than $P(A_i)P(B_i)$, negative otherwise.

on each other [i.e., $P(A_i, B_i)$ deviating significantly from $P(A_i)P(B_i)$]. A well-known example is that nitrates such as $Ba(NO_3)_2$ and $Ce(NO_3)_3$ tend to be used together, likely because of their solubility and applicability for solution processing (e.g., slurry preparation). Unfortunately, these decisions regarding dependencies of precursors are usually empirical and hard to standardize. Machine learning is a possible solution to ingest the heuristics that underlie such selections.

Materials encoding for precursor selection

Our precursor recommendation model for the synthesis of a novel target will mimic the human approach of trying to identify similar target materials for which successful synthesis reactions are known. To find similar materials, digital processing requires an encoding model that transforms any arbitrary inorganic material into a numerical vector. For organic synthesis, structural fingerprinting such as Morgan2Feat (*24*) is a good choice (*25*) because it is natural to track the conservation and change of functional groups in organic reactions, but the concept of functional groups is not applicable to inorganic synthesis. Chemical formulas of inorganic solids have been represented using a variety of approaches [e.g., Magpie (*26*, *27*), Roost (*28*), CrabNet (*29*)]. However, these representations are typically used as inputs to predict thermodynamic or electronic properties of materials. Here, we attempt to directly incorporate synthesis information into the representation of a material with arbitrary composition. Local text-based encodings such as Word2Vec (*30*, *31*) and FastText (*17*) are able to capture contextual information from the materials science literature, of which synthesis information is a part; however, they are not applicable to unseen materials when the materials text (sub)strings are not in the vocabulary or when the materials are not in the predefined composition space. For example, Pei *et al.* (*31*) computed the similarity of high-entropy alloys as the average similarity of element strings by assuming that the elements are present in equal proportions in the material (e.g., CoCrFeNiV). However, this approach is not applicable to unseen materials different from such composition template and consequently would not be practical in our work on synthesis of diverse inorganic materials. Substitution modeling can evaluate

similarity of precursors by assessing the viability of substituting one precursor with another while retaining the same target, but it cannot be used to identify analogues for new target materials (*22*). In this work, we propose a synthesis context–based encoding model using the idea that target materials produced with similar synthesis variables are similar.

Analogous to how language models (*19–21*) pretrain word representations by predicting context for each word, we use a self-supervised representation learning model to encode arbitrary materials by predicting precursors for each target material, which we refer to as PrecursorSelector encoding (Fig. 3A). The upstream part is an encoder where properties of the target material are projected into a latent space as the encoded vector representation. In principle, any intrinsic materials property could be included at this step. Here, we use only composition for simplification. The downstream part consists of multiple tasks where the encoded vector is used as the input to predict different variables related to precursor selection. Here, we use a masked precursor completion (MPC) task (Fig. 3B) to capture (i) the correlation between the target and precursors and (ii) the dependency between different precursors in the same experiment. For each target material and corresponding precursors in the training set, we randomly mask part of the precursors and use the remaining precursors as a condition to predict the complete precursor set. We also add a task of reconstructing the chemical composition to conserve the compositional information of the target material. The downstream task part is designed to be extensible; other synthesis variables such as operations and conditions can be incorporated by adding corresponding prediction tasks in a similar fashion. By training the entire neural network, the encoded vectors for target materials with similar precursors are automatically dragged closer to each other in the latent space because that reduces the overall prediction error. This PrecursorSelector encoding thus takes the correlation induced by precursor selection and serves as a useful metric to measure similarity of target materials in syntheses.

To demonstrate that the neural network is able to learn precursor information, we present the results of the MPC task (Fig. 3B) for $LaAlO_3$ as an example (Table 1). $LaAlO_3$ is a ternary material that normally requires two precursors (one to deliver each cation, La and Al). In this test, we masked one precursor and asked the model to predict the complete precursor set. For the same target conditioned with different partial precursors, the predicted probabilities of precursors strongly depend on the given precursor and agree with some rules of thumb for precursor selection. When the partial precursors are oxides such as La_2O_3 or Al_2O_3, the most probable precursors are predicted to be oxides for the other element, i.e., Al_2O_3 for La_2O_3 and La_2O_3 for Al_2O_3 (*32*). When the partial precursors are nitrates such as $La(NO_3)_3$ or $Al(NO_3)_3$, nitrates for the other element are prompted with higher probabilities, i.e., $Al(NO_3)_3$ for $La(NO_3)_3$ and $La(NO_3)_3$ for $Al(NO_3)_3$ (*33*). If both precursors are masked, oxides rank first in the prediction because the common precursors for elements La and Al are La_2O_3 and Al_2O_3, respectively. The simple successful prediction shows that our PrecursorSelector encoding model is able to learn the correlation between the target and precursors in different contexts of synthesis without explicit input of chemical rules about synthesis. In addition, the use of different precursors suggests that various synthetic routes may lead to the same target material. When a practical preference for a particular route exists, the framework we introduce in this work can be
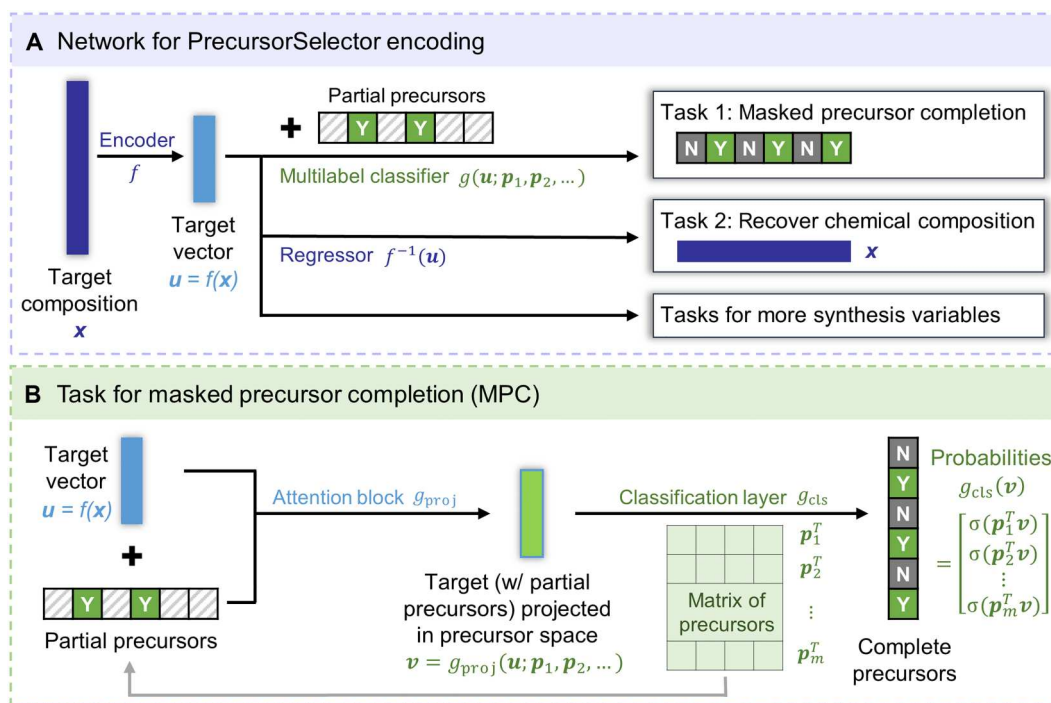
Fig. 3. Representation learning to encode precursor information for target materials. (**A**) Multitask network structure to encode the target material in the upstream and to predict the complete precursor set, chemical composition, and more synthesis variables in the downstream. x and u represent the composition and encoded vector of the target material, respectively. p_i represents the ith precursor in a predefined ordered precursor list. Dense layers are used in each layer unless specified differently. (**B**) Submodel of multilabel classification for the MPC task. Part of the precursors are randomly masked; the remaining precursors (marked as "Y") are used as a condition to predict the probabilities of other precursors for the target material. The probabilities corresponding to the complete precursors (marked as "Y") are expected to be higher than that of unused precursors (marked as "N"). The attention block g_{proj} (61) is used to aggregate the target vector and conditional precursors. The final classification layer g_{cls} and the embedding matrix for conditional precursors share the same weights. σ represents the sigmoid function.

extended to include more constraints, such as synthesis type, temperature, morphology, particle size, and cost of precursors, by learning from pertinent datasets (23, 34, 35).

Similarity of target materials

Similarity establishes a link between a novel material to synthesize and the known materials in the knowledge base because it is reasonable to assume that similar target materials share similar synthesis variables in experiments. Although the understanding of similarity is generally based on heuristics, the PrecursorSelector encoding introduced in the "Materials encoding for precursor selection" section provides a meaningful representation for quantified similarity analysis. Dedicated to precursor prediction in this study, we define the similarity of two target materials as the similarity of the precursors used in their respective syntheses. Although precursors for a new target material are not known in advance, the PrecursorSelector encoding serves as a proxy reflecting the potential precursors to use. In that latent space, we can take the cosine similarity (19, 20, 30) of the PrecursorSelector encoding as a measure of the similarity (Sim) of two target materials x_1 and x_2

$$\text{Sim}(x_1, x_2) \sim \cos[f(x_1), f(x_2)] \quad (1)$$

where f is the encoder part of the PrecursorSelector model transforming the composition of the target material x into the encoded target vector (Fig. 3A).

To demonstrate that the similarity estimated from PrecursorSelector encoding is reasonable, we show typical materials with different levels of similarity to an example target material $NaZr_2(PO_4)_3$ (Table 2). The most similar materials are the ones with the same elements such as Zr-containing phosphates and other sodium super ionic conductor (NASICON) materials. The similarity decreases slightly as additional elements are introduced (e.g., $Na_3Zr_{1.9}Ti_{0.1}Si_2PO_{12}$) or when one element is substituted [e.g., $LiZr_2(PO_4)_3$]. When the phosphate groups are replaced with another anion, the similarity decreases further, with oxides having generally mild similarity to the phosphate $NaZr_2(PO_4)_3$. The similarity decreases even further for compounds with no anion (e.g., intermetallics) and for non-oxygen anions (e.g., chalcogenides). This finding agrees with our experimental experience that when seeking a reference material, researchers will usually refer to compositions in the same chemical system or to cases where some elements are substituted. It is also worth noting that our quantitative similarity is purely a data-driven abstraction from the literature and uses no externally chemical knowledge.

To better understand the similarity, we conducted a relationship analysis (19, 20, 30) by visualizing four groups of target materials synthesized using one shared precursor and one distinct precursor (Fig. 4). For example, the syntheses of $YCuO_2$, $Ba_3Y_4O_9$, and $Ti_3Y_2O_9$ share Y_2O_3 as a precursor and separately use CuO, $BaCO_3$, and TiO_2. The three other groups share the precursors In_2O_3, Al_2O_3, and Fe_2O_3, respectively. To separate the effect of

Table 1. MPC conditioned on different partial precursors for the same target material LaAlO$_3$. The predicted complete precursors are the ones with the highest probabilities (bold). N/A denotes the absence of partial precursors, i.e., all precursors are masked in the MPC task.

Partial precursors (condition)	Probability to use different precursors (output)					
	La$_2$O$_3$	Al$_2$O$_3$	La(NO$_3$)$_3$	Al(NO$_3$)$_3$	La$_2$(CO$_3$)$_3$	Al(OH)$_3$
La$_2$O$_3$	**0.75**	**0.71**	0.58	0.57	0.57	0.57
Al$_2$O$_3$	**0.72**	**0.73**	0.58	0.57	0.58	0.56
La(NO$_3$)$_3$	0.60	0.59	**0.64**	**0.63**	0.61	0.61
Al(NO$_3$)$_3$	0.62	0.58	**0.65**	**0.65**	0.62	0.60
N/A	**0.70**	**0.69**	0.59	0.58	0.59	0.59

the precursor variation, we align the original points of the target vectors by first projecting each target vector to the same vector space as the precursors and then subtracting the vector of the shared precursor, providing a difference vector showing the relationship between the target material and the shared precursor (more details in the "Representation learning for similarity of materials" section). Next, we plot the top two principal components (*36*) of these difference vectors in a two-dimensional plane. The difference vectors are automatically separated into three clusters according to the precursor variate, representing three types of relationships, "react with BaCO$_3$," "react with CuO," and "react with TiO$_2$," respectively. For example, Ba$_3$Y$_4$O$_9$ is to Y$_2$O$_3$ as BaAl$_2$O$_4$ is to Al$_2$O$_3$ (i.e., Ba$_3$Y$_4$O$_9$ − Y$_2$O$_3$ ≈ BaAl$_2$O$_4$ − Al$_2$O$_3$) because both syntheses use BaCO$_3$. The consistency between this automatic clustering and the chemical intuition again affirms the efficacy of using PrecursorSelector encoding as a similarity metric.

Recommendation of precursor materials

With the capability of measuring similarity, a natural solution to precursor selection is to replicate the literature-based approach used by experimental researchers. Given a novel material to synthesize, we initialize our recommendation by first proposing a recipe consisting of common precursors for each metal/metalloid element in the target material because this might be the first attempt in a lab. Then, we encode the novel target material and

known target materials in the knowledge base using PrecursorSelector encoding model from the "Materials encoding for precursor selection" section and calculate the similarity between the novel target and each known material with Eq. 1. We rank known materials based on their similarity to the target such that a reference material can be identified that is the most similar to the novel target. When the precursors used in the synthesis of the reference material cannot cover all elements of the target, we use MPC in Fig. 3B to predict the missing precursors. For example, for Y$_2$FeSbO$_7$ (Fig. 1B), the most similar material in the knowledge base is FeSbO$_4$. It is reasonable to assume that the precursors Fe$_2$O$_3$ and Sb$_2$O$_5$ used in the synthesis of FeSbO$_4$ (*37*) can also be used to synthesize Y$_2$FeSbO$_7$. Because the Y source is missing, MPC finds that Y$_2$O$_3$ is likely to fit with Fe$_2$O$_3$ and Sb$_2$O$_5$ for the synthesis of Y$_2$FeSbO$_7$, ending up as a complete precursor set (Fe$_2$O$_3$, Sb$_2$O$_5$, and Y$_2$O$_3$) (*38*). Multiple attempts of recommendation are feasible by moving down the list of known materials ranked to be most similar to the novel target.

To evaluate our recommendation pipeline, we conduct a validation (Fig. 5) using the 33,343 synthesis recipes text-mined from the scientific literature. Using the knowledge base of 24,034 materials reported by the year 2014, we predict precursors for 2654 test target materials newly reported from 2017 to 2020 (more details in the "Data preparation" section). Because multiple precursors exist for each element, the number of possible precursor combinations increases combinatorially with the number of elements present in the target material. A good precursor prediction algorithm is anticipated to select from hundreds of possible precursor combinations those that have a higher probability of success. For each test material, we attempt to propose five different precursor sets. For each attempt, we calculate the percentage of test materials being successfully synthesized, where success means at least one set of proposed precursors has been observed in previous experiments. The similarity-based reference already increases the success rate to 73% at the second attempt. The first guess is set to default to the most common precursors which leads to 36% success rate. Within five attempts, the success rate of our recommendation pipeline using PrecursorSelector encoding is 82%, comparable to the performance of recommendations for organic synthesis (*25*). We note that as defined here, "success" will be underestimated since some suggested precursor sets may actually lead to successful target synthesis although they may not have been tried (and therefore do not appear in the data).

We also establish a baseline model ("Most frequent" in Fig. 5) that ranks precursor sets based on the product of frequencies with

Table 2. Different levels of similarity between NaZr$_2$(PO$_4$)$_3$ and materials in the knowledge base.

| Target	Similarity	Target	Similarity
Zr$_3$(PO$_4$)$_4$	0.946	Li$_{1.8}$ZrO$_3$	0.701
Na$_3$Zr$_2$Si$_2$PO$_{12}$	0.929	NaNbO$_3$	0.600
Na$_3$Zr$_{1.8}$Ge$_{0.2}$Si$_2$PO$_{12}$	0.921	Li$_2$Mg$_2$(MoO$_4$)$_3$	0.500
Na$_3$Ca$_{0.1}$Zr$_{1.9}$Si$_2$PO$_{11.9}$	0.908	Sr$_2$Ce$_2$Ti$_5$O$_{16}$	0.400
Na$_3$Zr$_{1.9}$Ti$_{0.1}$Si$_2$PO$_{12}$	0.900	Ga$_{0.75}$Al$_{0.25}$FeO$_3$	0.300
LiZr$_2$(PO$_4$)$_3$	0.896	Cu$_2$Te	0.200
NaLa(PO$_3$)$_4$	0.874	Ni$_{60}$Fe$_{30}$Mn$_{10}$	0.100
Sr$_{0.125}$Ca$_{0.375}$Zr$_2$(PO$_4$)$_3$	0.852	AgCrSe$_2$	0.000
Na$_5$Cu$_2$(PO$_4$)$_3$	0.830	Zn$_{0.1}$Cd$_{0.9}$Cr$_2$S$_4$	−0.099
LiGe$_2$(PO$_4$)$_3$	0.796	Cr$_2$AlC	−0.202
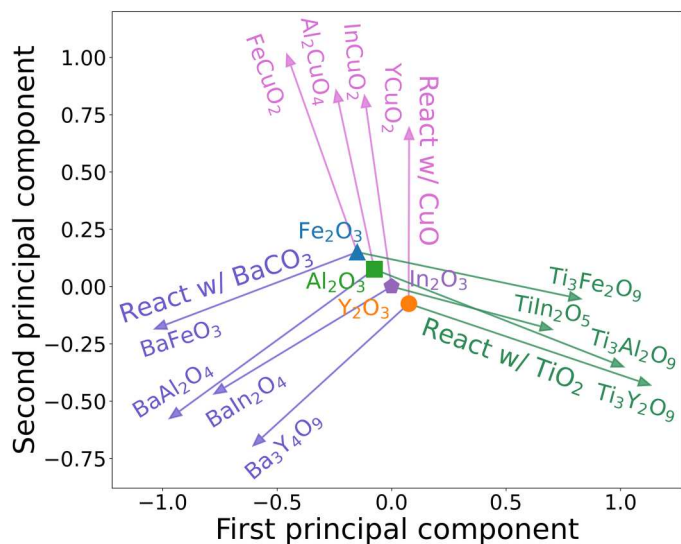
Fig. 4. Relationships between targets and their shared precursors. Four groups of target materials are synthesized each using one shared precursor shown as the original point (Y_2O_3, In_2O_3, Al_2O_3, or Fe_2O_3) and one distinct precursor shown as the edge ($BaCO_3$, CuO, or TiO_2). The relationship of "react with another precursor" is visualized as the first two principal components of the difference vector between the target and the shared precursor $g_{proj}[f(\boldsymbol{x})] - \boldsymbol{p}_i$. The original points corresponding to different precursors $\boldsymbol{p}_i$'s are jittered for clarity.

Fig. 5. Performance of various precursor prediction algorithms. For each of the 2654 test target materials, the algorithm attempts to propose n ($1 \leq n \leq 5$ as the x axis) precursor sets. The y axis shows the success rate that at least one of the n proposed precursor set is observed in previous experimental records. PrecursorSelector encoding: This work. Magpie encoding/FastText encoding/Raw composition: Similar recommendation pipeline to this work but using Magpie representation (26, 27)/FastText representation (17)/the raw material composition. Most frequent: Select precursors by frequency.

which different precursors are used in the literature (more details in the "Baseline models"). This baseline simulates the typical early stage of the trial-and-error process where researchers grid-search different combinations of precursors matching elements present in the target material without the knowledge of dependency of precursors (Fig. 2B). The success rate of this baseline is 58% within five attempts. Our recommendation pipeline performs better because the dependency of precursors is more easily captured when the combination of precursors is sourced from a previously used successful recipe for a similar target. Through in situ diffraction of synthesis (2–4), it is now better understood that some precursor sets do not lead to the target material because they form intermediate phases which have consumed much of the overall reaction energy, thereby leaving a low driving force to form the target. It is likely that our literature-informed precursor prediction approach implicitly captures some of this reactivity and pathway information, resulting in a higher prediction power than random selection or selection based on how common a precursor is.

In addition, we compare with three other baseline models ("Magpie encoding," "FastText encoding," and "Raw composition" in Fig. 5) using the same recommendation strategy but different encoding methods (more details in the "Baseline models"). Magpie encoding (26, 27) is a set of attributes computed using the fraction of elements in a material, including stoichiometric attributes, elemental property statistics, electronic structure attributes, and ionic compound attributes. Precursor recommendation with Magpie encoding achieves a success rate of 68% within five attempts; it performs reasonably well because these properties reflect the material composition, and generally, materials with close compositions tend to be similar. Similarly, precursor recommendation directly with the raw material composition achieves a success rate of 66% within five attempts. FastText encoding (17)
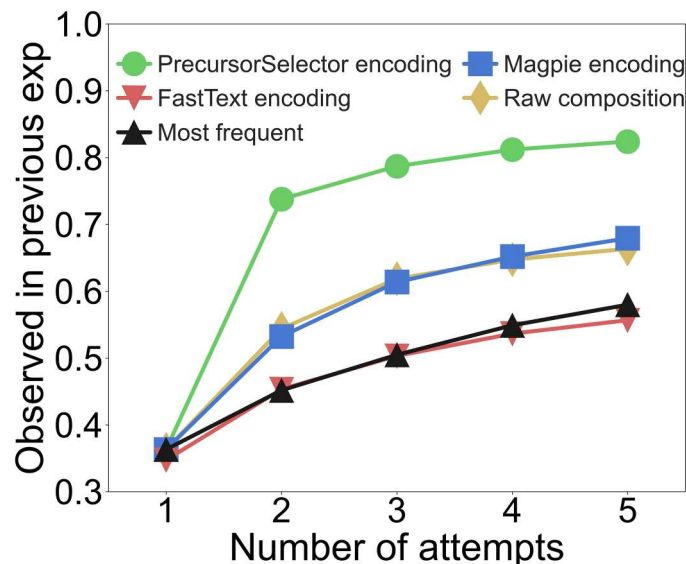
uses the FastText model (39) to capture information about the co-occurrences of context words around material formulas/names in the literature. However, only 1985 test materials can be digitized with FastText encoding due to the conflict between the limited vocabulary of n-grams and the variety of float numbers in material formulas. The success rate using FastText encoding is 56% within five attempts. Overall, the recommendation with PrecursorSelector encoding performs substantially better because Magpie and FastText encodings are more generic but not dedicated to predictive synthesis. The PrecursorSelector encoding and MPC capture the correlation between synthesis variables and known target materials, which better extends to novel materials.

DISCUSSION
Because of its heuristic nature, it is challenging to capture the decades of synthesis knowledge established in the literature. By establishing a materials similarity measure that is a natural handle of chemical knowledge and leveraging a large-scale dataset of precedent synthesis recipes, our similarity-based recommendation strategy mimics human synthesis design and succeeds in precursor selection. The incorporation of precursor information into materials representations (Fig. 3) leads to a quantitative similarity metric that successfully reproduces a known precursor set 82% of the time in five attempts or less (Fig. 5). We discuss the strengths and weaknesses of this recommendation algorithm and its generalizability to broader synthesis prediction problems.

In this work, materials similarity is learned through an automatic feature extraction process mapping a target material to the combination of precursors. While learning the usage of precursors,

Table 3. Representative successful and failed examples for precursor prediction using the similarity-based recommendation pipeline in this study.

Target	Reference target(s)	Expected precursors	Error in recommendation
Successful			
$Li_7La_3Nb_2O_{13}$ (65)	$Li_5La_3Nb_2O_{12}$ (66)	$LiOH$, La_2O_3, Nb_2O_5	N/A
$NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ (40)	$Mn_{0.24}Zn_{1.76}SiO_4$ (67)	$MnCO_3$, Na_2CO_3, Ga_2O_3, SiO_2, ZnO	N/A
$SrZnSO$ (68)	$CaZnSO$ (69)	$SrCO_3$, ZnS	N/A
$Na_3TiV(PO_4)_3$ (42)	$Na_3V_2(PO_4)_3$ (43)	NaH_2PO_4, NH_4VO_3, TiO_2	N/A
$GdLu(MoO_4)_3$ (45)	$Gd_2(MoO_4)_3$ (70)	$(NH_4)_6Mo_7O_{24}$, Lu_2O_3, Gd_2O_3	N/A
$BaYSi_2O_5N$ (47)	$YSiO_2N$ (48)	Si_3N_4, SiO_2, $BaCO_3$, Y_2O_3	N/A
$Cu_3Yb(SeO_3)_2O_2Cl$ (49)	$Cu_4Se_5O_{12}Cl_2$ (50)	CuO, $CuCl_2$, SeO_2, Yb_2O_3	N/A
$LiMn_{0.5}Fe_{0.5}PO_4$ (51, 52)	$LiMn_{0.8}Fe_{0.2}PO_4$ (53), $LiMn_{0.9}Fe_{0.1}PO_4$ (54)	$MnCO_3$, FeC_2O_4, LiH_2PO_4; $Mn(CH_3COO)_2$, FeC_2O_4, LiH_2PO_4	N/A
Failed			
Li_3CoTeO_6 (55)	$LiCoO_2$ (71)	Co, Te, Li_2CO_3	Co_3O_4, TeO_2, $LiOH$
$Sr_4Al_6SO_{16}$ (56)	$SrAl_2O_4$ (72)	$SrCO_3$, $SrSO_4$, $Al(OH)_3$	$SrCO_3$, H_2SO_4, $Al(OH)_3$
$Ca_{7.5}Ba_{1.5}Bi(VO_4)_7$ (57)	$Bi_3Ca_9V_{11}O_{41}$ (73)	$BaCO_3$, NH_4VO_3, $CaCO_3$, Bi_2O_3	BaO, NH_4VO_3, $CaCO_3$, Bi_2O_3

useful chemical knowledge for synthesis practice is accordingly embedded in PrecursorSelector encoding. The first level of knowledge about materials similarity is based on composition. For example, to synthesize $Li_7La_3Nb_2O_{13}$, PrecursorSelector encoding finds $Li_5La_3Nb_2O_{12}$ as a reference target material (Table 3) because their difference in composition is only one Li_2O unit. PrecursorSelector encoding also reflects the consideration of valence in synthesis. Although it is not necessary to keep the valence in the precursor the same as that in the target, a precursor with similar valence states to the target is frequently used in practical synthesis (22). For example, to synthesize $NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ (40), $MnCO_3$ was used as the Mn source because the valence state of Mn is 2+ in both the target and precursor. PrecursorSelector encoding finds $Mn_{0.24}Zn_{1.76}SiO_4$ similar to $NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ because the valence state of Mn is also 2+ in $Mn_{0.24}Zn_{1.76}SiO_4$, despite $NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ containing large fractions of Na and Ga while $Mn_{0.24}Zn_{1.76}SiO_4$ does not. Our algorithm also captures the similarity of syntheses between compounds which have one element substituted. For example, PrecursorSelector encoding refers to CaZnSO for synthesizing SrZnSO because the elements Ca and Sr are regarded as similar. While such knowledge may appear obvious to the trained chemist, our approach enables it to be automatically extracted and convoluted as a vectorized representation (Fig. 3), making it thereby available in a mathematical form, convenient to be used in recommendation engines or automated labs (41).

Because of this customized synthesis similarity of materials and our precursor recommendation pipeline, we are able to not only recommend trivial solutions for target synthesis such as the use of common precursors but also deal with more challenging situations. One typical scenario is the adoption of uncommon precursors. For example, Lalère *et al.* (42) used NaH_2PO_4 as the source of Na and P to synthesize $Na_3TiV(PO_4)_3$, while the common precursors for Na and P are Na_2CO_3 and $NH_4H_2PO_4$, respectively. It is not apparent to

conclude from the composition of $Na_3TiV(PO_4)_3$ that the uncommon precursor NaH_2PO_4 is needed. However, the similarity-based recommendation pipeline successfully predicts the use of NaH_2PO_4 by referring to a similar material $Na_3V_2(PO_4)_3$ (43). A plausible reason for the choice of NaH_2PO_4 for $Na_3TiV(PO_4)_3$ can also be inferred from the synthesis of $Na_3V_2(PO_4)_3$. Feng *et al.* (43) reported that NaH_2PO_4 was used to implement a one-pot solid-state synthesis of $Na_3V_2(PO_4)_3$, while Fang *et al.* (44) reported that a reductive agent and additional complex operations are needed when using Na_2CO_3 and $NH_4H_2PO_4$. Similar outcomes may also apply to the synthesis of $Na_3TiV(PO_4)_3$. A second example is the successful precursor recommendation for the target compound $GdLu(MoO_4)_3$. Instead of the common precursor MoO_3, a less common precursor $(NH_4)_6Mo_7O_{24}$ was adopted as the Mo source (45). The use of $(NH_4)_6Mo_7O_{24}$ may facilitate the mixing of different ions in the synthesis of $GdLu(MoO_4)_3$. The adoption of uncommon precursors also provides clues in underexplored chemical spaces such as mixed-anion compounds (46). Taking the pentanary oxynitride material $BaYSi_2O_5N$ (47) as an example, the five-component system, including multiple anions, implies that many precursor combinations can potentially yield the target phase, including oxides, nitrides, carbonates, etc. Our recommendation pipeline correctly identifies that a combination of SiO_2 and Si_3N_4 facilitates the formation of $BaYSi_2O_5N$ by referring to a quaternary oxynitride material, $YSiO_2N$ (48). Another challenging situation is that multiple precursors may be used for the same element. Usually, only one precursor is used for each metal/metalloid element in the target material, but exceptions do exist. For example, CuO and $CuCl_2$ were used as the Cu source in the synthesis of $Cu_3Yb(SeO_3)_2O_2Cl$ (49). Through analogy to $Cu_4Se_5O_{12}Cl_2$ (50), the recommended precursor set includes both CuO and $CuCl_2$. Moreover, it is possible to predict multiple correct precursor sets by referring to multiple similar target materials. For example, two different sets of precursors for $LiMn_{0.5}Fe_{0.5}PO_4$ were reported by Zhuang *et al.* (51) and

Wang *et al.* (*52*). The recommendation pipeline predicts both by repurposing the precursor sets for $LiMn_{0.8}Fe_{0.2}PO_4$ (*53*) and $LiMn_{0.9}Fe_{0.1}PO_4$ (*54*).

The recommendation of precursors presented here is still imperfect. The engine we present is inherently limited by the knowledge base it is trained on, thereby biasing recommendations toward what has been done previously and lacking creativity for unprecedented combinations of precursors. For example, metals Co and Te were used in the synthesis of Li_3CoTeO_6 (*55*), but no similar materials in the knowledge base use the combination of Co and Te as precursors. Another example is that $SrCO_3$ and $SrSO_4$ were used in the synthesis of $Sr_4Al_6SO_{16}$ (*56*). Although the recommendation pipeline is, in principle, able to predict multiple precursors for the same element, a similar case using both $SrCO_3$ and $SrSO_4$ as the Sr source is not found in the knowledge base. Both examples end up being mispredictions. This situation could be improved when more data from text mining and high-throughput experiments (*41*) are added to the knowledge base. Furthermore, the success rate of the recommendation strategy may be underestimated in some cases. For example, BaO is predicted as the Ba source for synthesizing $Ca_{7.5}Ba_{1.5}Bi(VO_4)_7$, while $BaCO_3$ is used in the reported synthesis (*57*). Given the slight difference between BaO and $BaCO_3$, BaO may actually be suitable.

Besides the prediction of precursors, the similarity-based recommendation framework is a potential step toward general synthesis prediction. The same strategy can be extended to the recommendation of more synthesis variables, such as operations, device setups, and experimental conditions, by adding corresponding prediction tasks to the downstream part of the multitask network (Fig. 3) for similarity measurement. For example, we may infer that reduced atmosphere is necessary for synthesizing $Na_3TiV(PO_4)_3$ (*42*) because it is used in the synthesis of a similar material $Na_3V_2(PO_4)_3$ (*43*). Moreover, synthesis constraints such as the type of synthesis method, temperature, morphology of the target material, particle size, and cost can be added as conditions of synthesis prediction. For example, we may integrate our effort of synthesis temperature prediction to prioritize the predicted precursors within expected temperature regime. Our automated algorithm, mimicking human design process for the synthesis of a new target, provides a practical solution to query decades of heuristic synthesis data in recommendation engines and autonomous laboratories.

MATERIALS AND METHODS
Representation learning for similarity of materials
The neural network consists of an encoder part for encoding target materials and a task part for predicting variables related to precursor selection. The encoder part f is a three-layer fully connected submodel transforming the composition of the target material x into a 32-dimensional target vector $u = f(x)$. The input composition is an array with 83 units showing the fraction of each element. The reduced dimension of the encoded target vector is inspired by the bottleneck architecture of autoencoders (*58*). By limiting the dimension of the encoded vector, the network is forced to learn a more compact and efficient representation of the input data, which is more appropriate for the precursor selection-related downstream tasks (*59*). The task part uses different network architectures for different tasks of prediction, including precursor completion and composition recovery in this work. The MPC task replaces part of the precursors with a placeholder "[MASK]" (*21*) at random and uses the remaining precursors as a condition to predict the complete precursor set for the target material, which is formulated as a multilabel classification problem (*60*). An attention block g_{proj} (*61*) is used to aggregate the target vector and the vectors for conditional precursors as a projected vector $v = g_{proj}(u; p_1, p_2, ...)$ with dimensionality of 32. Then, v is passed to the precursor classification layer represented by a 417×32 matrix P, of which each row is the 32-dimensional vector representation of a potentially used precursor p_i. To avoid having too many neural network weights to learn, the precursor completion task only considers 417 precursors used in at least five reactions in the knowledge base. The probability to use each precursor is indicated by sigmoid($p_i^\top v$), allowing nonexclusive prediction of multiple precursors (*60*). Here, v acts as a probe corresponding to the target material projected in the precursor space and is used to search for p_i's with similar vector representations via a dot product. The conditional precursors input to g_{proj} share the same trainable vector representations as p_i's. Circle loss (*62*) is used because of its benefits in capturing the dependency between different labels in multilabel classification and deep feature learning. The composition recovery task is a two-layer fully connected submodel decoding back to the chemical composition x from the target vector u, similar to the mechanism of autoencoders (*58, 63*). Mean squared error loss is used because it is the most popular for regression. More tasks predicting other synthesis variables such as operations and conditions can be appended in a similar fashion. To combine the loss functions in this multitask neural network, an adaptive loss (*64*) is used to automatically weigh different loss by considering the homoscedastic uncertainty of each task.

Baseline models
Most frequent
This baseline model ranks precursor sets based on an empirical joint probability without considering the dependency of precursors (Fig. 2B). Assuming that the choices of precursors are independent of each other, the joint probability of selecting a specific set of precursors can be estimated as the product of their marginal probabilities. For each metal/metalloid element, different precursors can be used as the source. The marginal probability to use a precursor is estimated as the relative frequency of using that precursor over all precursors contributing the same metal/metalloid element. For example, the precursor set ranked in first place is always the combination of common precursors for each metal/metalloid element in the target material, which is also typically the first attempt in the lab.
Magpie encoding
This baseline model uses the same recommendation strategy as Fig. 1, except that the similarity is calculated using Magpie encoding (*26, 27*). The composition of each target material is converted into a vector consisting of 132 statistical quantities such as the average and standard deviation of various elemental properties. The cosine similarity is used, as shown in Eq. 1. When the precursors from the reference target material cannot cover all elements of the novel target, the common precursors for the missing elements are supplemented because MPC (Fig. 3B) is only trained for PrecursorSelector encoding.

FastText encoding

Similar to the baseline of Magpie encoding, this baseline model uses the same recommendation strategy as Fig. 1, except that the similarity is calculated using FastText encoding (*17*). The formula of each target material is converted into a 100-dimensional vector using the FastText model trained with materials science papers (*17*). The total number of target materials tested in this baseline model is 1985 instead of 2654 because some n-grams such as certain float numbers corresponding to the amount of elements are not in the vocabulary.

Raw composition

Similar to the baseline of Magpie encoding, this baseline model uses the same recommendation strategy as Fig. 1, except that the similarity is calculated using the cosine similarity of raw material composition. The formula of each target material is converted into an 83-dimensional vector corresponding to the fraction of each element.

Data preparation

In total, 33,343 inorganic solid-state synthesis recipes extracted from 24,304 materials science papers (*11*) were used in this work. Because some material strings (e.g., $Ba_{1-x}Sr_xTiO_3$) extracted from the literature contain variables corresponding to different amounts of elements, we substituted these variables with their values from the text to ensure that a material in any reaction only corresponds to one composition, resulting in 49,924 expanded reactions and 28,598 target materials. An ideal test for generalizability and applicability of this method would be to synthesize many entirely new materials using recommended precursors. In the absence of performing extensive new synthesis experiments, we designed a robust test to simulate precursor recommendation for target materials that are new to the trained model. We split the data based on the year of publication, i.e., training set (or knowledge base) for reactions published by 2014, validation set for reactions in 2015 and 2016, and test set for reactions from 2017 to 2020. In addition, to avoid data leakage where the synthesis of the same material can be reported again in a more recent year, we placed reactions for target materials with the same prototype formula in the same dataset as the earliest record. The prototype formula was defined as the formula corresponding to a family of materials including (i) the formula itself, (iii) formulas derived from a small amount (<0.3) of substitution (e.g., $Ca_{0.2}La_{0.8}MnO_3$ for prototype formula $LaMnO_3$), and (iii) formulas able to be coarse-grained by rounding the amount of elements to one decimal place (e.g., $Ba_{1.001}La_{0.004}TiO_3$ for the prototype formula $BaTiO_3$). In the end, the number of reactions in the training/validation/test set was 44,736/2254/2934 from 29,900/1451/1992 original recipes. The number of target materials in the training/validation/test set was 24,304/1910/2654, respectively.

Supplementary Materials

This PDF file includes:
Supplementary Text
Fig. S1
References

REFERENCES AND NOTES

1. J. C. Hemminger, J. Sarrao, G. Crabtree, G. Flemming, M. Ratner, Challenges at the frontiers of matter and energy: Transformative opportunities for discovery science, *Tech. rep.*, USDOE Office of Science (SC) (United States) (2015).
2. A. Miura, C. J. Bartel, Y. Goto, Y. Mizuguchi, C. Moriyoshi, Y. Kuroiwa, Y. Wang, T. Yaguchi, M. Shirai, M. Nagao, N. C. Rosero-Navarro, K. Tadanaga, G. Ceder, W. Sun, Observing and modeling the sequential pairwise reactions that drive solid-state ceramic synthesis. *Adv. Mater.* **33**, 2100312 (2021).
3. M. Bianchini, J. Wang, R. J. Clément, B. Ouyang, P. Xiao, D. Kitchaev, T. Shi, Y. Zhang, Y. Wang, H. Kim, M. Zhang, J. Bai, F. Wang, W. Sun, G. Ceder, The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nat. Mater.* **19**, 1088–1095 (2020).
4. Z. Jiang, A. Ramanathan, D. P. Shoemaker, In situ identification of kinetic factors that expedite inorganic crystal formation and discovery. *J. Mater. Chem. C* **5**, 5709–5717 (2017).
5. E. J. Corey, Robert Robinson lecture. Retrosynthetic thinking—essentials and examples. *Chem. Soc. Rev.* **17**, 111–133 (1988).
6. A. Stein, S. W. Keller, T. E. Mallouk, Turning down the heat: Design and mechanism in solid-state synthesis. *Science* **259**, 1558–1564 (1993).
7. M. H. S. Segler, M. Preuss, M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
8. J. R. Chamorro, T. M. McQueen, Progress toward solid state synthesis by design. *Acc. Chem. Res.* **51**, 2918–2925 (2018).
9. H. Kohlmann, Looking into the black box of solid-state synthesis. *Eur. J. Inorganic Chem.* **2019**, 4174–4180 (2019).
10. H. Schäfer, Preparative solid state chemistry: The present position. *Angew. Chem. Int. Ed. Engl.* **10**, 43–50 (1971).
11. O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).
12. E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
13. M. C. Swain, J. M. Cole, Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
14. A. M. Hiszpanski, B. Gallagher, K. Chellappan, P. Li, S. Liu, H. Kim, J. Han, B. Kailkhura, D. J. Buttler, T. Y.-J. Han, Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *J. Chem. Inf. Model.* **60**, 2876–2887 (2020).
15. M. Aykol, J. H. Montoya, J. Hummelshøj, Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).
16. M. J. McDermott, S. S. Dwaraknath, K. A. Persson, A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat. Commun.* **12**, 3097 (2021).
17. E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, E. Olivetti, Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
18. H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain, G. Ceder, Machine-learning rationalization and prediction of solid-state synthesis conditions. *Chemistry of Materials* **34**, 7323–7336 (2022).
19. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013).
20. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **26**, 3111–3119 (2013).
21. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018).
22. T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, G. Ceder, Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem. Mater.* **32**, 7861–7873 (2020).
23. X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, J. Schrier, Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).
24. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
25. C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).
26. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).

27. L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).

28. R. E. Goodall, A. A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **11**, 6280 (2020).

29. A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, T. D. Sparks, Compositionally restricted attention-based network for materials property predictions. *npj Comput. Mater.* **7**, 77 (2021).

30. V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).

31. Z. Pei, J. Yin, P. K. Liaw, D. Raabe, Toward the design of ultrahigh-entropy alloys via mining six million texts. *Nat. Commun.* **14**, 54 (2023).

32. Z.-y. Mao, Y.-c. Zhu, Q.-n. Fei, D.-j. Wang, Investigation of 515nm green-light emission for full color emission LaAlo$_3$ phosphor with varied valence Eu. *JOL* **131**, 1048–1051 (2011).

33. E. Mendoza-Mendoza, K. P. Padmasree, S. M. Montemayor, A. F. Fuentes, Molten salts synthesis and electrical properties of Sr- and/or mg-doped perovskite-type LaAlo$_3$ powders. *J. Mater. Sci.* **47**, 6076–6085 (2012).

34. Z. Wang, O. Kononova, K. Cruse, T. He, H. Huo, Y. Fei, Y. Zeng, Y. Sun, Z. Cai, W. Sun, G. Ceder, Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Sci. Data* **9**, 231 (2022).

35. K. Cruse, A. Trewartha, S. Lee, Z. Wang, H. Huo, T. He, O. Kononova, A. Jain, G. Ceder, Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. *Sci. Data* **9**, 234 (2022).

36. K. Pearson, LIII. *On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).

37. E. Zvereva, O. Savelieva, Y. D. Titov, M. Evstigneeva, V. Nalbandyan, C. Kao, J.-Y. Lin, I. Presniakov, A. Sobolev, S. Ibragimov, M. Abdel-Hafiez, Y. Krupskaya, C. Jähne, G. Tan, R. Klingeler, B. Büchner, A. N. Vasiliev, A new layered triangular antiferromagnet Li$_4$FeSbO$_6$: Spin order, field-induced transitions and anomalous critical behavior. *Dalton Trans.* **42**, 1550–1566 (2013).

38. J. Luan, L. Zhang, K. Ma, Y. Li, Z. Zou, Preparation and property characterization of new Y$_2$FeSbO$_7$ and In$_2$FeSbO$_7$ photocatalysts. *Solid State Sci.* **13**, 185–194 (2011).

39. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2017).

40. S. Lv, B. Shanmugavelu, Y. Wang, Q. Mao, Y. Zhao, Y. Yu, J. Hao, Q. Zhang, J. Qiu, S. Zhou, Transition metal doped smart glass with pressure and temperature sensitive luminescence. *Adv. Opt. Mater.* **6**, 1800881 (2018).

41. N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, G. Ceder, Toward autonomous design and synthesis of novel inorganic materials. *Mater. Horiz.* **8**, 2169–2198 (2021).

42. F. Lalère, V. Seznec, M. Courty, J. Chotard, C. Masquelier, Coupled X-ray diffraction and electrochemical studies of the mixed Ti/V-containing NASICON: Na$_2$TiV(PO$_4$)$_3$. *J. Mater. Chem. A* **6**, 6654–6659 (2018).

43. P. Feng, W. Wang, K. Wang, S. Cheng, K. Jiang, Na$_3$V$_2$(PO$_4$)$_3$/C synthesized by a facile solid-phase method assisted with agarose as a high-performance cathode for sodium-ion batteries. *J. Mater. Chem. A* **5**, 10261–10268 (2017).

44. Y. Fang, L. Xiao, X. Ai, Y. Cao, H. Yang, Hierarchical carbon framework wrapped Na$_3$V$_2$(PO$_4$)$_3$ as a superior high-rate and extended lifespan cathode for sodium-ion batteries. *Adv. Mater.* **27**, 5895–5900 (2015).

45. B. Wang, X. Li, Q. Zeng, G. Yang, J. Luo, X. He, Y. Chen, Efficiently enhanced photolumi-nescence in Eu^{3+}-doped Lu$_2$(MoO$_4$)$_3$ by Gd^{3+} substituting. *Mater. Res. Bull.* **100**, 97–101 (2018).

46. H. Kageyama, K. Hayashi, K. Maeda, J. P. Attfield, Z. Hiroi, J. M. Rondinelli, K. R. Poeppelmeier, Expanding frontiers in materials chemistry and physics with multiple anions. *Nat. Commun.* **9**, 772 (2018).

47. T. Yasunaga, M. Kobayashi, K. Hongo, K. Fujii, S. Yamamoto, R. Maezono, M. Yashima, M. Mitsuishi, H. Kato, M. Kakihana, Synthesis of Ba$_{1−x}$Sr$_x$YSi$_2$O$_5$N and discussion based on structure analysis and dft calculation. *J. Solid State Chem.* **276**, 266–271 (2019).

48. Y. Kitagawa, J. Ueda, M. G. Brik, S. Tanabe, Intense hypersensitive luminescence of Eu^{3+}-doped YSiO$_2$N oxynitride with near-uv excitation. *Opt. Mater.* **83**, 111–117 (2018).

49. M. Markina, K. Zakharov, E. Ovchenkov, P. Berdonosov, V. Dolgikh, E. Kuznetsova, A. Olenev, S. Klimin, M. Kashchenko, I. Budkin et al., Interplay of rare-earth and transition-metal subsystems in Cu$_3$Yb(SeO$_3$)$_2$O$_2$Cl. *Phys. Rev. B* **96**, 134422 (2017).

50. D. Zhang, H. Berger, R. K. Kremer, D. Wulferding, P. Lemmens, M. Johnsson, Synthesis, crystal structure, and magnetic properties of the copper selenite chloride Cu$_5$(SeO$_3$)$_4$Cl$_2$. *Inorg. Chem.* **49**, 9683–9688 (2010).

51. H. Zhuang, Y. Bao, Y. Nie, Y. Qian, Y. Deng, G. Chen, Synergistic effect of composite carbon source and simple pre-calcining process on significantly enhanced electrochemical per-formance of porous LiFe$_{0.5}$Mn$_{0.5}$PO$_4$/C agglomerations. *Electrochim. Acta* **314**, 102–114 (2019).

52. L. Wang, Y. Li, J. Wu, F. Liang, K. Zhang, R. Xu, H. Wan, Y. Dai, Y. Yao, Synthesis mechanism and characterization of LiMn$_{0.5}$Fe$_{0.5}$PO$_4$/C composite cathode material for lithium-ion batteries. *J. Alloys Compd.* **839**, 155653 (2020).

53. Q.-Q. Zou, G.-N. Zhu, Y.-Y. Xia, Preparation of carbon-coated LiFe$_{0.2}$Mn$_{0.8}$PO$_4$ cathode material and its application in a novel battery with Li$_4$Ti$_5$O$_{12}$ anode. *J. Power Sources* **206**, 222–229 (2012).

54. H. Yi, C. Hu, H. Fang, B. Yang, Y. Yao, W. Ma, Y. Dai, Optimized electrochemical performance of LiMn$_{0.9}$Fe$_{0.1−x}$Mg$_x$PO$_4$/C for lithium ion batteries. *Electrochim. Acta* **56**, 4052–4057 (2011).

55. G. Heymann, E. Selb, M. Kogler, T. Götsch, E.-M. Köck, S. Penner, M. Tribus, O. Janka, Li$_{3−}$Co$_{1.06(1)}$TeO$_6$: Synthesis, single-crystal structure and physical properties of a new tellurate compound with Co^{II}/Co^{III} mixed valence and orthogonally oriented Li-ion channels. *Dalton Trans.* **46**, 12663–12674 (2017).

56. J. S. Ndzila, S. Liu, G. Jing, J. Wu, L. Saruchera, S. Wang, Z. Ye, Regulation of Fe^{3+}-doped Sr$_4$Al$_6$SO$_{16}$ crystalline structure. *J. Solid State Chem.* **288**, 121415 (2020).

57. N. G. Dorbakov, V. V. Titkov, S. Y. Stefanovich, O. V. Baryshnikova, V. A. Morozov, A. A. Belik, B. I. Lazoryak, Barium-induced effects on structure and properties of β-Ca$_3$(PO$_4$)$_2$-type Ca$_9$Bi(VO$_4$)$_7$. *J. Alloys Compd.* **793**, 56–64 (2019).

58. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).

59. M. Tschannen, O. Bachem, M. Lucic, Recent advances in autoencoder-based representation learning. arXiv:1812.05069 (2018).

60. F. Herrera, F. Charte, A. J. Rivera, M. J. del Jesus, *Multilabel Classification* (Springer, 2016), pp. 17–31.

61. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).

62. Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified per-spective of pair similarity optimization. arXiv:2002.10857 (2020).

63. G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural net-works. *Science* **313**, 504–507 (2006).

64. R. Cipolla, Y. Gal, A. Kendall, Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, paper presented at Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, 18 to 23 June 2018, pp. 7482–7491.

65. H. Peng, X. Luan, L. Li, Y. Zhang, Y. Zou, Synthesis and ion conductivity of Li$_7$La$_3$Nb$_2$O$_{13}$ Ceramics with cubic garnet-type structure. *J. Electrochem. Soc.* **164**, A1192–A1194 (2017).

66. L. van Wüllen, T. Echelmeyer, H.-W. Meyer, D. Wilmer, The mechanism of li-ion transport in the garnet Li$_5$La$_3$Nb$_2$O$_{12}$. *Phys. Chem. Chem. Phys.* **9**, 3298–3303 (2007).

67. K. Park, H. Lim, S. Park, G. Deressa, J. Kim, Strong blue absorption of green Zn$_2$SiO$_4$: Mn^{2+} phosphor by doping heavy Mn^{2+} concentrations. *Chem. Phys. Lett.* **636**, 141–145 (2015).

68. C. Chen, Y. Zhuang, D. Tu, X. Wang, C. Pan, R.-J. Xie, Creating visible-to-near-infrared me-chanoluminescence in mixed-anion compounds SrZn$_2$S$_2$O and SrZnSO. *Nano Energy* **68**, 104329 (2020).

69. C. Duan, A. Delsing, H. Hintzen, Photoluminescence properties of novel red-emitting Mn^{2+}-activated MZnOS (M = Ca, Ba) phosphors. *Chem. Mater.* **21**, 1010–1016 (2009).

70. J. Thirumalai, R. Krishnan, I. Shameem Banu, R. Chandramohan, Controlled synthesis, formation mechanism and lumincence properties of novel 3-dimensional Gd$_2$(MoO$_4$)$_3$:Eu^{3+} nanostructures. *J. Mater. Sci. Mater. Electron.* **24**, 253–259 (2013).

71. R. Alcantara, J. Jumas, P. Lavela, J. Olivier-Fourcade, C. Pérez-Vicente, J. Tirado, X-ray diffraction, ^{57}Fe Mössbauer and step potential electrochemical spectroscopy study of LiFe$_y$Co$_{1−y}$O$_2$ compounds. *J. Power Sources* **81–82**, 547–553 (1999).

72. Y. Zhu, J. Zeng, W. Li, L. Xu, Q. Guan, Y. Liu, Encapsulation of strontium aluminate phos-phors to enhance water resistance and luminescence. *Appl. Surf. Sci.* **255**, 7580–7585 (2009).

73. I. Radosavljevic, J. A. Howard, A. W. Sleight, J. S. Evans, Synthesis and structure of Bi$_3$Ca$_9$V$_{11}$O$_{41}$. *J. Mater. Chem.* **10**, 2091–2095 (2000).

74. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019).

75. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

76. L. Prechelt, *Neural Networks: Tricks of the trade* (Springer, 2002), pp. 55–69.