JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

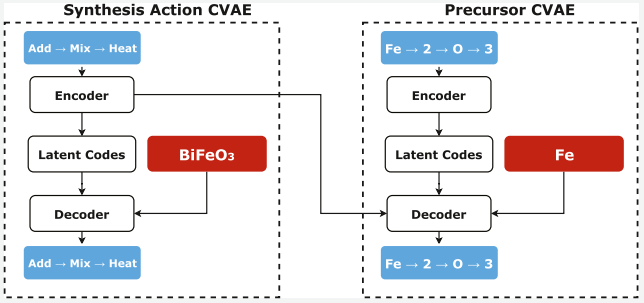
Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks

Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, and Elsa Olivetti*

Cite This: J. Chem. Inf. Model. 2020, 60, 1194−1201 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Leveraging new data sources is a key step in accelerating the pace of materials design and discovery. To complement the strides in synthesis planning driven by historical, experimental, and computed data, we present an automated, unsupervised method for connecting scientific literature to inorganic synthesis insights. Starting from the natural language text, we apply word embeddings from language models, which are fed into a named entity recognition model, upon which a conditional variational autoencoder is trained to generate syntheses for any inorganic materials of interest. We show the potential of this technique by predicting precursors for two perovskite materials, using only training data published over a decade prior to their first reported syntheses. We demonstrate that the model learns representations of materials corresponding to synthesis-related properties and that the model's behavior complements the existing thermodynamic knowledge. Finally, we apply the model to perform synthesizability screening for proposed novel perovskite compounds.



INTRODUCTION

Recent advances in predicting material properties,[1−3] screening synthesizable compounds,[4−7] and organic reaction prediction[8−10] have been driven, in part, by the accessibility of machine-readable datasets[11−13] and consequently, data-driven models. In stark contrast to organic reaction databases,[13] the overwhelming majority of inorganic synthesis knowledge lies locked within the text of journal articles[14−16] and laboratory notebooks.[17] While the latter has been shown as an effective source for guiding successful syntheses of specific materials systems, there is no existing method for automatically and broadly translating literature knowledge into insights for the syntheses of novel inorganic materials.

Scientific literature has previously been used to illuminate patterns in nanoscale morphologies,[14] solid-state reactions,[16] device performances,[18] and apparatus parameters,[19] but each of these efforts have required tailored, material-specific data representations. Recently, Tshitoyan et al.[20] have shown that the Word2Vec embedding algorithm[21] can capture useful correlations in the materials science literature without the use of any supervised algorithms or hand-engineered data representations. The model used by Tshitoyan et al. focuses on predictions regarding materials it has seen before, while the work presented here aims to add chemical insights for undiscovered materials. To provide actionable insights for materials discovery, data-driven models must provide inferences about never-before-seen materials, and these models must consider not only the topical context (e.g., if a material is a thermoelectric) but also synthetic context (e.g., which precursors are often used to synthesize it). Moreover, all of this must be achieved with minimal hand-labeling of training data.[16,22]

In this work, we present an automated method for connecting scientific literature to context-aware insights for inorganic materials synthesis planning. We show that an unsupervised conditional variational autoencoder (CVAE)[23−25] can generate synthesis predictions for a variety of materials, including materials never before seen by the model. This CVAE learns directly from the materials synthesis literature and produces an internal representation of precursors which corresponds to physical and chemical trends without receiving any explicit domain knowledge. We then use the literature knowledge captured by the CVAE to complement first-principles techniques in materials screening tasks.

In contrast to existing generative models for materials synthesis planning,[26] this unsupervised CVAE model requires no fine-tuned feature engineering when performing synthesis planning for different categories of materials. Additionally, it is

Received: October 27, 2019
Published: January 7, 2020

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201

trained on data that is automatically produced from a neural network natural language processing pipeline[14] that uses context-aware, character-based word representations to allow inferences to be made for never-before-seen materials. In short, while previous models have shown the capability to predict if a previously discovered material may be seen in a new context,[14,20,26] our model predicts not only if, but how, new materials can be made, using only data learned autonomously from the literature. Critically, our model not only learns to follow the "rules" of commonly accepted synthesis intuition but also learns to "break" these rules while still obeying thermodynamics (e.g., suggesting nitride precursors where oxides and carbonates are more typical, while maintaining negative reaction enthalpy). Thus, our model provides a literature-driven synthesis planning approach that is complementary to first-principles methods for exploring rare materials, such as inorganic nitrides.[7,27]

To accelerate the efforts of the materials science community, we open-source several key resources used in this work at www.github.com/olivettigroup/materials-synthesis-generative-models: We release context-sensitive embeddings from language models (ELMo) that have been adapted for materials science text[28] along with a pre-trained FastText word embedding model for materials science.[29] Each of these embedding models has been trained on a collection of over 2.5 million materials science journal articles.[14,22] We also provide the full architecture and code for the CVAE model, along with a Python tutorial for running the code. Finally, we provide over two hundred annotated literature synthesis routes for named entity recognition (NER) tasks, such as identifying reaction conditions and materials.

RESULTS AND DISCUSSION

We first describe our automated workflow. After a recurrent neural network[30] identifies synthesis sections of journal articles, context-sensitive ELMo word embeddings are computed and passed into another recurrent neural network which performs NER to identify precursors, synthesis target materials, and synthesis actions. Then, a CVAE model, shown in Figure 1, is trained to learn representations of synthesis routes from the named entities in an unsupervised manner. Intuitively, an autoencoder (AE) learns to compress synthesis parameters into a lower-dimensional representation. AEs can be modified to have a variational component, resulting in a variational autoencoder (VAE) which allows for novel synthesis parameters to be efficiently and accurately generated.[26] By adding inputs to the decoder, the VAE becomes a CVAE, and the model is then able to produce different synthesis parameters depending on the target material of interest. In contrast to previous work (which was not conditional),[26] the CVAE model presented here requires no material-specific feature engineering, respects the order of operations performed in a synthesis, and can perform synthesis planning for arbitrary materials after a single instance of model training (whereas prior models required retraining and additional feature engineering for each materials system). More details on these methods are provided in the Experimental Section.

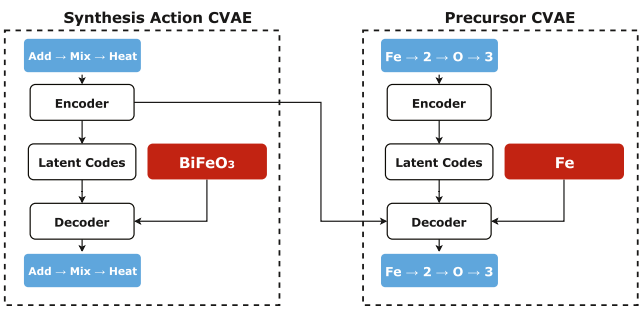


Figure 1. Schematic diagram of the CVAE architecture. The model consists of two joined CVAEs used for learning synthesis actions and precursors. The synthesis action CVAE learns distributions of synthesis action sequences conditioned on target materials. The precursor CVAE learns distributions of precursor formulas conditioned on both a target element and an encoded representation of the jointly observed synthesis action sequence. Target materials are represented by FastText embeddings, and all other inputs to the model are sequences of one-hot vectors.

To maximize the opportunity for transfer learning of synthesis trends, we choose a broad definition for synthesis routes that requires minimal assumptions. For a given target material m, a synthesis route S_m^i is a 2-tuple consisting of a sequence of n synthesis actions $(a_1, a_2, ..., a_n)$ acting on a set of l precursors $\{p_1, ..., p_l\}$

$$S_m^i = ((a_k)^n, \{p_j\}^l) \tag{1}$$

and in general, a single target material m may have $N > 1$ valid synthesis routes and thus S_m^i represents the ith valid synthesis route for m. We also define precursors p as "element sources," such that they are materials sharing an element with m. The CVAE model is then constructed to model the following distributions

$$\mathbb{P}((a_k)^n|\theta_a, m) \tag{2}$$

$$\mathbb{P}(p_j|\theta_p, e_j, (a_k)^n) \tag{3}$$

where θ_a and θ_p are model parameters for the synthesis action and precursor CVAEs, respectively, and e_j is the shared element between a precursor p_j and target material m (e.g., titanium). Because CVAEs are generative models, novel synthesis actions and precursors can be generated by sampling from a Gaussian prior distribution.[23]

Critically, we represent m by a FastText word embedding model trained on the materials science literature, which enables the transfer of synthesis trends between existing and novel materials by leveraging literature-based similarity. Although unsupervised word representations for materials science have been recently explored,[20,22] existing methods cannot draw inferences about materials that were previously unseen by the model. As an example, using a model which has never before encountered the formula $LiNi_{1-x}Co_xO_2$, the word embedding model we fine-tuned enables reasonable inferences via cosine similarity. Similarity of the previously unseen material is ranked to be higher with another battery cathode material compared to a binary metal oxide

$$\text{Similarity}\ (LiCoO_2,\ \underline{LiNi_{1-x}Co_xO_2}) = 61\% \tag{4}$$

$$\text{Similarity}\ (MnO_2,\ \underline{LiNi_{1-x}Co_xO_2}) = 47\% \tag{5}$$

Here, the underlined text denotes a material which was never observed by the model during training. Nonetheless, we are able to represent the material as a real-valued vector as well as compute a reasonable similarity to related materials—both tasks which would not be possible using previous methods.[20,22]

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201

The authors note, however, that FastText embeddings may overemphasize similarity based on morphological likeness (e.g., words containing the same substrings), and so appropriate caution should be exercised.

To demonstrate the applicability of our CVAE method, we construct a dataset of approximately 51,000 synthesis action sequences and 116,000 precursors via a general set of search terms ("perovskite + thermoelectric + multiferroic + photovoltaic + solar + nano + cathode") and apply our neural network pipeline. We investigate the effectiveness of the CVAE model in synthesis planning by performing a publication-year-split experiment, where the model is trained only on syntheses published prior to 2005 (~2800 syntheses). We apply the model in predicting precursors for materials that were unseen during training, are computationally predicted as stable perovskites,[1] and only recently appear in the literature: $InWO_3$ and $PbMoO_3$, first reported in 2016 and 2017, respectively.[31,32] Table 1 shows a report of the data generated

Table 1. Generated Precursors for $InWO_3$ and $PbMoO_3$, Drawn from the CVAE Model[a]

target material	precursors
$InWO_3$	$In_2S_3 + WCl_4$
	$In(NO_3)_3 + WCl_4$
	$In_2O_3 + WO_2$
	$In_2O_3 + WN$
	[b] $InCl_3 + Na_2WO_4$
$PbMoO_3$	$PbCl_2 + MoCl_2$
	$PbSO_4 + MoCl_2$
	[c] $PbO + MoO_2$

[a]The CVAE model was trained on synthesis routes published during or before 2005. [b]Precursors match Kamalakkannan et al. (2016).[31] [c]Precursors match Takatsu et al. (2017).[32]

by sampling from the CVAE's Gaussian prior distributions, where the CVAE suggests the precursors for both materials (see Table S1 for additional details). The CVAE model is thus capable of predicting synthesis precursors while relying only on literature knowledge from more than a decade prior to the literature-reported syntheses of these materials. Trial-and-error (or random) precursor selection is substantially less efficient, as the number of possible precursor sets for each material is in the hundreds. Thus literature-driven models may greatly accelerate future synthesis attempts of novel materials.

During the data generation process, the CVAE model proposes several plausible syntheses beyond the literature-matching samples. To the best of the authors' knowledge, the only reported synthesis of $InWO_3$ is via a solution-phase route. However, the CVAE model suggests that solid-state synthesis of $InWO_3$ may be possible, using In_2O_3 and either WO_2 or WN as precursors. Such syntheses may be feasible, as they are thermodynamically favorable (using data at 0 K and 0 atm from OQMD).[11] Thus, by observing the actions performed in the laboratory, the model has extracted the underlying thermochemical trends and physical reasoning used by the experimenters.

$$In_2O_3 + 2WO_2 \rightarrow 2InWO_3 + 2O_2 \quad (6)$$

$$\Delta H = -158\ \text{kJ/mol} \quad (7)$$

$$In_2O_3 + 2WN \rightarrow 2InWO_3 + N_2 \quad (8)$$

$$\Delta H = -930\ \text{kJ/mol} \quad (9)$$

We do note, however, that these thermodynamic analyses should only be used as rough guidelines. Besides the limitations of estimating an overall thermochemical reaction for the synthesis, along with extrapolation from STP conditions, kinetic effects are not considered here. Indeed, while it is common to mix binary oxide precursors in solid-state syntheses of ternary (or quaternary, etc.) oxides, the use of nitride precursors is less common due to the high bond energies of many nitride compounds.[27] To achieve a clearer understanding of kinetic effects, experimental verification would be required alongside a model which incorporates reaction conditions (e.g., temperatures), and this is an area for future work. Nonetheless, it is interesting to observe that the CVAE model is able to suggest synthetic precursors that are both typical (oxides) and atypical (nitrides) while respecting thermodynamics.

In the suggested recipes for $PbMoO_3$, the CVAE model suggests a solution-phase route using $PbSO_4$ and $MoCl_2$, both of which are soluble under acidic conditions. The CVAE model thus provides chemical insights into new, potentially viable paths toward synthesizing $PbMoO_3$, which has only been realized so far in the laboratory by solid-state synthesis methods.[32] However, we stress that these suggestions still need human evaluation and should not be applied "out of the box."

Despite the fact that chemical knowledge is never given to the CVAE model, we find one example where solubility rules emerge from the model results. To demonstrate this, we generate W-bearing precursors for $InWO_3$ conditioned on two representative action sequences sampled from the Synthesis Action CVAE: a solid-state synthesis (mix, grind, calcine, press, sinter, cool) and a solution-phase synthesis (add, dissolve, stir, heat, wash, dry). Following this, we generate 10,000 CVAE-suggested precursors. The most common W-bearing precursor generated for the solid-state synthesis is the water-insoluble WO_3, while the most common precursor generated for the solution-phase synthesis is the highly soluble Na_2WO_4. The differences of these precursor likelihoods in each case is substantial, with −16 and +21% changes to the likelihoods of the CVAE suggesting WO_3 and Na_2WO_4, respectively, when switching from conditioning on solid-state to solution-phase synthesis actions. This shows that, for this particular case, the model has captured the physical concept of aqueous solubility purely by observing the selective use of certain precursors in solution-phase literature syntheses.

This effect of learning precursor trends from the literature is further demonstrated upon inspecting latent codes learned by the model. Because the CVAE learns conditional distributions, the input precursors are projected into a degenerate latent space, where the degeneracy is split by the conditional input received by the decoder. By investigating several examples (see Figure S4), we find that the CVAE learns to group precursors with similar synthesis-relevant properties, including insoluble binary oxides, water-soluble polyanion compounds, and pure/alloyed metals. Because the CVAE only receives one-hot representations of precursors as inputs, with no prior knowledge encoded, this suggests that the CVAE model is capable of capturing chemical intuition and composition-driven similarity solely by joint observations of precursors, synthesis actions, and target materials. Despite the lack of "negative" data in the literature, the diversity of published

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201

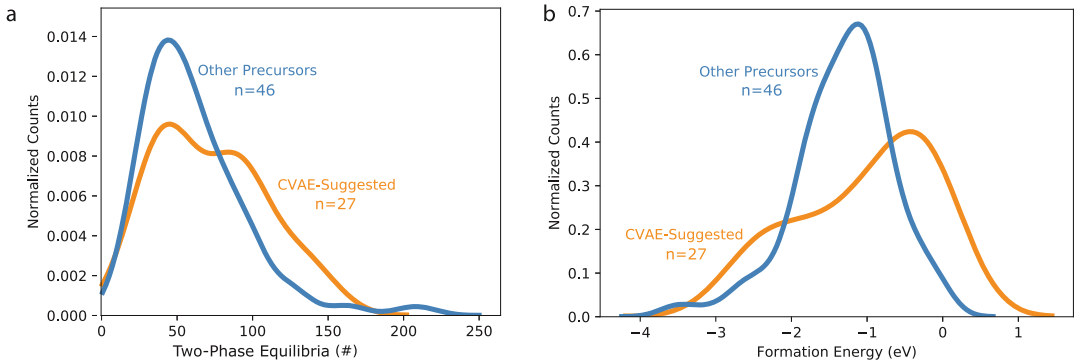


Figure 2. Two-phase equilibria between precursors and associated precursor formation energies for $InWO_3$ precursors,[33] using data from 1000 precursor sets generated by the CVAE. (a) Normalized distributions for number of two-phase equilibria between precursors for CVAE-suggested and non-suggested precursors. (b) Normalized distributions of formation energies for CVAE-suggested and non-suggested precursors.

synthesis literature is sufficient to drive the CVAE model in learning meaningful representations of precursors.

To emphasize the particular nature of synthesis planning via a literature-trained model, we contrast suggested precursors by the CVAE model with thermodynamic stability computations from OQMD,[11] which we have used to compute two-phase equilibria and formation energies for all compounds in the chemical space spanned by all CVAE-suggested precursors.[33] Given that only a subset of the (meta)stable precursor materials are selected by the CVAE, as shown in Figure 2, the CVAE is clearly not suggesting the full set of thermodynamically viable precursors (blue vs orange curves). Indeed, the CVAE has filtered precursors from a set of 73 possible precursors to only 27, thus minimizing the set of candidate precursors to test in the laboratory. Additionally, Figure 2a,b shows that the CVAE is not selecting precursors in correspondence with isolated thermodynamic metrics: the CVAE's suggestions are explained neither by thermodynamic reactivity (i.e., the number of relevant two-phase equilibria with respect to other precursors) nor individual precursor stability (i.e., formation energy). This suggests that there is a meaningful difference between the thermodynamically driven and literature-driven synthesis planning methods. While the former probes the realm of physical possibility, the latter emphasizes practical choices and historical trends. In other words, the CVAE model has uncovered a new physical metric for synthesis planning that is driven by the aggregate reported successes of past experiments and complements existing thermodynamic theory.

We next train the CVAE model on our full dataset, using no publication-year cutoffs. To investigate the capability of the model for suggesting syntheses of a novel, never-before-synthesized material, we consider novel ABO_3 perovskite materials proposed by Balachandran et al.[34] These proposed perovskites have not previously been synthesized and have high thermodynamic stability as measured by energy differences against their convex hulls. We note that ABO_3 perovskites are used here as a representative example because of their chemical variety and diverse range of properties, but the CVAE model does indeed generalize to other categories of materials (see Table S3).

$HgZrO_3$ is one such example of a thermodynamically stable, unsynthesized perovskite material,[34] and we perform synthesis predictions using the CVAE model (see Table S2). We find that the CVAE proposes solid-state syntheses which appear to be thermodynamically reasonable:

$$HgO + ZrC + 2O_2 \rightarrow HgZrO_3 + CO_2 \quad (10)$$

$$\Delta H = -1340\ \text{kJ/mol} \quad (11)$$

$$HgO + ZrO_2 \rightarrow HgZrO_3 \quad (12)$$

$$\Delta H = -0.29\ \text{kJ/mol} \quad (13)$$

Again, we emphasize that thermodynamic analyses are often insufficient to evaluate reaction plausibility. As an additional utility for evaluating generated synthesis parameters, we develop a similarity metric based on the latent codes learned by the CVAE. By measuring nearest-neighbors of latent codes for the recipe using mercuric oxide and zirconium carbide, we find that the two closest literature recipes are for solid-state syntheses of $SrZrO_3$ and $BaAl_2O_4$ (see Figure S5). Besides providing insight into which observed literature examples "inspired" this particular prediction, we are also led to further insights on precursor selections. ZrC is an uncommon choice of precursor, but carbonate precursors are readily used in solid-state syntheses. Indeed, both of the near-neighbor syntheses for $HgZrO_3$ use carbonate precursors rather than carbides.

While similarity methods have previously been produced for materials (e.g., based on crystal structures),[35] the CVAE incorporates synthesis knowledge to produce a distinct measure of similarity. Indeed, from a structural point of view, it would not be expected that $SrZrO_3$ and $BaAl_2O_4$ should have high similarity to $HgZrO_3$ because all three materials form ground-state structures in different crystal systems (orthorhombic, hexagonal, and cubic, respectively).

Moreover, rather than measuring similarity against entire materials, the CVAE-based metric operates at the level of individual reported (or generated) synthesis routes. Because the nearest-neighbor search is computationally inefficient in high dimensional spaces, the dimensionality reduction imposed by the CVAE enables this "latent citation" model to be used as a rapid, data-driven synthesis planning method.

Finally, we present results for synthesis screening using the CVAE model to suggest syntheses for numerous ABO_3 suggested by Balachandran et al.[34] The CVAE was used to generate syntheses with ten data generation attempts per compound, and only compounds which had at least one suggested synthesis route with commercially available precursors were considered to have passed the test, which is the same criterion used by Segler et al.[8] to evaluate retrosynthetic routes for organic molecules.

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201

Figure 3 shows a grid of possible A-site and B-site atoms for ABO_3 perovskite materials, with screened compounds

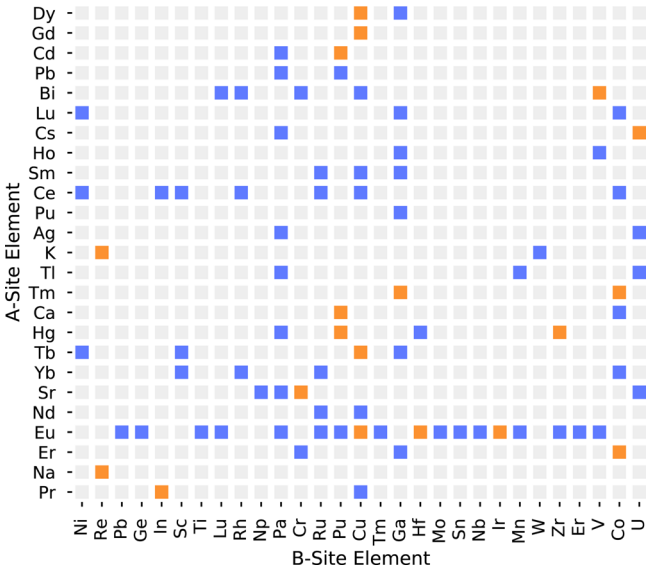


Figure 3. Unsynthesized ABO_3 perovskite compounds, labeled by their A-site and B-site elements. Colored-in squares are perovskites predicted to be stable.[34] Orange and blue colors correspond to compounds which passed or failed the CVAE screening, respectively.

represented by highlighted combinations of A-site and B-site atoms. While a joint machine learning and density functional theory method[34] selects a set of materials which are thermodynamically stable in the perovskite form, the further-imposed synthesis screening selects a subset that is most readily synthesizable based on existing literature knowledge. From a set of 83 proposed ABO_3 perovskite compounds, the CVAE has selected a subset of only 19. In analogy to the results found by Segler et al. for data-driven retrosynthesis of organic molecules,[8] we find here that the CVAE model has derived new chemical selection "rules" based on the results reported by past experiments.

CONCLUSIONS

The CVAE model, combined with the rest of our neural network workflow, enables a new axis of synthesis screening which complements the existing domain knowledge.[5,34] By incorporating and extending patterns in the historical literature, materials which are theoretically synthesizable can be rapidly and automatically filtered by their practical synthesizability. This capability of the model emphasizes the ability to capture and extend physical and chemical insights from the literature itself. Although natural language is not inherently bound by physical principles, the reported steps in successful materials synthesis experiments are ultimately governed by physics and motivated by scientific reasoning. While this latent scientific reasoning exists primarily in the minds of experimenters, we have shown that observations of experimental reports within the literature are sufficient to uncover key aspects of this reasoning, including rules of solubility and thermodynamics. Moreover, the CVAE model encourages synthesis planning beyond the norm of usual synthesis routes, as demonstrated by the suggestion of nitride precursors and precursors that are soluble under specific pH conditions.

While the methods presented in this paper are applicable to various materials systems and synthesis methods, we recognize that our broad representation of synthesis routes omits information such as temperatures, solvents, and morphologies and additionally assumes that there is a one-to-one relation between elements in precursors and targets. We thus believe that a promising future work lies in the direction of generative models with a narrower scope but finer-grained detail. For example, limiting the dataset to solvothermal syntheses may facilitate prediction of solvent choices, solvothermal reaction temperatures, and dwell times. This additional domain knowledge may be incorporated by filtering proposed synthesis parameters[8] or constraining model outputs.[36] Motivated by these possibilities, our open-source NER annotations include the necessary labels (e.g., reaction conditions) to enable these future studies.

EXPERIMENTAL SECTION

The text extraction methodology follows the high-level workflow as reported in the literature;[14] however, all machine learning models have been redesigned from the ground-up. While the prior work used local-window neural networks and logistic regression, in this work, recurrent and convolutional neural networks are used, along with higher performance word embedding models.

All neural network models are implemented in the Keras library using the TensorFlow backend,[37] with the exception of ELMo which is implemented directly in TensorFlow.[38] The Pymatgen library is used for computing all chemical formulas.[39]

A bidirectional-GRU[30] recurrent neural network is trained to classify all paragraphs of a journal article using FastText word embeddings as input features. An annotated collection of 4000 paragraphs were used to train the paragraph classification model. This model has an overall accuracy of 92% across eight classes (abstract, introduction, synthesis recipe, characterization and other methods, results, conclusions, captions, and miscellaneous) and achieves a recall and precision of 96 and 81% for classifying synthesis recipes, respectively. For the purposes of the CVAE model, recall is the most relevant metric, as false-positive synthesis recipes have a minimal effect on downstream NER processes (with the exception of increasing overall computation time). Training the paragraph classification model took less than an hour using an Intel Xeon E5-2620 v4 at 2.10 GHz.

We highlight that static, context-insensitive word embeddings are used for this task because the computation time for character-based embeddings is intractable for a full-text corpus of this size. Training the FastText model took approximately 2 weeks using an Intel Xeon E5-2620 v4 at 2.10 GHz.

A bidirectional-GRU, operating at a sentence-level context, is trained on character-based ELMo word embeddings. ChemDataExtractor is used for tokenizing sentences and words.[40] The macro-averaged categorical accuracy is 93%, and a full confusion matrix is available in Table 2. The ELMo model was fine-tuned on our collection of 2.5M + materials science journal articles, starting from the standard pretrained weights.[28] In practice, the authors found that fine-tuning an ELMo model to materials science literature improved performance on NER tasks by upwards of 10% in some categories. The fine-tuning process took approximately 1 week to complete training on two NVIDIA Titan Xp GPUs. The NER model took less than an hour to train on the same hardware.

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201

Table 2. NER Model Confusion Matrix for Test-Set Predictions[a]

	null (P)	precursor (P)	target (P)	synthesis action (P)
null (T)	98	01	00	01
precursor (T)	06	88	06	00
target (T)	01	09	90	00
synthesis action (T)	03	00	00	97

[a]"P" denotes predicted labels, and "T" denotes true labels. Entries are row-normalized percentages. Total numbers of each ground-truth entity type are as follows: 13,630 null, 140 precursor, 69 target, and 261 operation.

The NER labels are produced from a manual annotation process where 230 journal articles were annotated word-by-word. This annotated dataset includes labels for additional categories which are not used in the current study, such as temperatures, solvents, apparatuses, and brand names.

Table 2 shows the confusion matrix for the NER model used to identify precursors, synthesis actions, and synthesized target materials. We also experimented with training NER models that incorporate additional class labels available in the annotated NER data and found promising accuracies for several types of synthesis-relevant data. On the test set of annotations, using the same NER model architecture (with more output classes), we achieved F1 scores of 92% on atmospheric gases (e.g., argon and nitrogen), 96% on amounts of materials (e.g., molarity and mass), 94% on reaction conditions (e.g., temperatures and reaction times), and 73% on material descriptors (e.g., phase names and morphologies).

Synthesis recipes are treated as a set of precursors, along with a sequence of in-lab synthesis actions and a target-synthesized material. Both the synthesis action CVAE and the precursor CVAE use identical architectures: convolutional encoders feed into a latent parameter space for means and variances of Gaussian variational posteriors, and outputs from a latent sampling function are concatenated with conditional inputs as inputs to a recurrent decoder. In producing the results for this study, 8 latent dimensions were used for both the CVAEs.

Synthesis actions are encoded as sequences of one-hot vectors. Because there are many synonymous actions reported in the literature (e.g., press and compress), cosine similarity between FastText embeddings of synthesis actions are used to automatically cluster and prune the total vocabulary of synthesis actions. The authors found that using a total vocabulary size of 50 actions was adequate to capture the variety of possible synthesis actions. The authors would like to caution that the choice of vocabulary affects the variety and quality of the generated data: in particular, including a larger scope of synthesis methods (e.g., including characterization details) demands a larger vocabulary and consequently increases the difficulty of accurate data generation for the model. While vocabulary sizes in this study were chosen by the authors, automated vocabulary pruning methods have also achieved success in recent work.[15]

The synthesis action CVAE learns to auto-encode sequences of synthesis actions, conditioned on the character-based FastText embedding of the target material. The use of character-based embeddings allows for the representation of target materials unseen at the training time. The reconstruction accuracy varies strongly with the dataset size (e.g., when truncating the dataset by year) and the number of latent dimensions in the "bottleneck" of the model. We found that, using 10% randomly held-out test data, the model achieves categorical accuracies of 51, 56, and 60% on training datasets of 100, 1000, and 10,000 recipes, respectively. Accuracies as high as 70% were achieved by increasing the number of latent dimensions to 64, but this was found to greatly reduce the conditional nature of the model because the model tends to "ignore" the conditional information provided from the target material and instead learn a single, global distribution.

In general, the authors found that high test-set accuracies were not necessary to generate reasonable data, as data generation is computationally cheap. The accuracy for the synthesis action CVAE is also comparable to the accuracies reported in the literature,[8] as there are multiple valid methods for the synthesis of a single material.

Precursors (represented by their chemical formulas) are encoded as sequences of one-hot vectors, where the total vocabulary is a character set consisting of the different elements and the numerical digits. We define a precursor as a chemical formula, used in a synthesis recipe, which shares an element with the target synthesized material (and intuitively acts as an "element source" during synthesis). In our case studies for ABO_3 perovskites, we assume that oxygen is ubiquitous and does not require an associated inorganic precursor. The authors note that information regarding hydrates, non-stoichiometric formulas, fractional formulas, abbreviations, and common names are ignored. This introduces some level of bias into our data, and a more robust representation of precursors is an area for future work. We also validate generated chemical formulas using Pymatgen, as the CVAE also suggests invalid precursors during random sampling attempts, similar to the invalid SMILES noted by Gómez-Bombarelli et al.[25] We found that this added only trivial amounts of computational time (i.e., seconds) to the overall data generation process.

The precursor CVAE model learns to auto-encode chemical formulas, represented as character sequences, conditioned on a one-hot vector of the target element along with the encoded representation of the synthesis action sequence in which the precursor is observed from the literature. Analogous to the CVAE for synthesis action sequences, the reconstruction accuracy is strongly influenced by dataset size and latent dimensions. On 10% random holdouts, the model achieves accuracies of 59, 67, and 91% on training datasets of 100, 1000, and 10,000 precursors, respectively. Increasing the number of latent dimensions to 64 resulted in accuracies as high as 98% but with the same effect on conditioning as observed for the synthesis action CVAE.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00995.

Dataset statistics, model accuracies, learned latent codes, and generated syntheses (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Elsa Olivetti − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge,*

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201

Massachusetts 02139, United States; orcid.org/0000-0002-8043-2385; Email: elsao@mit.edu

Authors

Edward Kim − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-0781-5531

Zach Jensen − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-7635-5711

Alexander van Grootel − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Kevin Huang − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Matthew Staib − *Department of EECS and CSAIL, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Sheshera Mysore − *College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States*

Haw-Shiuan Chang − *College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States*

Emma Strubell − *College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States*

Andrew McCallum − *College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States*

Stefanie Jegelka − *Department of EECS and CSAIL, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b00995

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge funding from the National Science Foundation Award 1534340 and 1534341, DMREF that provided support to make this work possible, support from the Office of Naval Research (ONR) under contract no. N00014-16-1-2432, the MIT Energy Initiative, and National Science Foundation CAREER Award 1553284. Early work was collaborative under the Dept. of Energy's Basic Energy Science Program through the Materials Project under grant no. EDCBEE. This work was also partly funded by the MIT-Sensetime Alliance on Artificial Intelligence. We would also like to acknowledge valuable feedback from Rafael Jaramillo, Gerbrand Ceder, Olga Kononova, Wenhao Sun, Haoyan Huo, and Tanjin He.

■ REFERENCES

(1) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(2) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 014106.

(3) Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chem. Mater.* **2015**, *27*, 735−743.

(4) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J.; Doak, J.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 094104.

(5) Aykol, M.; Hegde, V. I.; Hung, L.; Suram, S.; Herring, P.; Wolverton, C.; Hummelshøj, J. S. Network analysis of synthesizable materials discovery. *Nat. Commun.* **2019**, *10*, 2018.

(6) Kim, K.; Ward, L.; He, J.; Krishna, A.; Agrawal, A.; Wolverton, C. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. *Phys. Rev. Mater.* **2018**, *2*, 123801.

(7) Sun, W.; Bartel, C. J.; Arca, E.; Bauers, S. R.; Matthews, B.; Orvañanos, B.; Chen, B.-R.; Toney, M. F.; Schelhas, L. T.; Tumas, W.; Tate, J.; Zakutayev, A.; Lany, S.; Holder, A. M.; Ceder, G. A map of the inorganic ternary metal nitrides. *Nat. Mater.* **2019**, *18*, 732−739.

(8) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.

(9) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465−1476.

(10) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(11) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *J. Mater.* **2013**, *65*, 1501−1509.

(12) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(13) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49*, 2897.

(14) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436−9444.

(15) Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *J. Chem. Inf. Model.* **2019**, *59*, 3692−3702.

(16) Huo, H.; Rong, Z.; Kononova, O.; Sun, W.; Botari, T.; He, T.; Tshitoyan, V.; Ceder, G. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Comput. Mater.* **2019**, *5*, 62.

(17) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73−76.

(18) Ghadbeigi, L.; Harada, J. K.; Lettiere, B. R.; Sparks, T. D. Performance and resource considerations of Li-ion battery electrode materials. *Energy Environ. Sci.* **2015**, *8*, 1640−1650.

(19) Young, S. R.; Maksov, A.; Ziatdinov, M.; Cao, Y.; Burch, M.; Balachandran, J.; Li, L.; Somnath, S.; Patton, R. M.; Kalinin, S. V.; Vasudevan, R. K. Data mining for better material synthesis: The case of pulsed laser deposition of complex oxides. *J. Appl. Phys.* **2018**, *123*, 115303.

(20) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word

embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95−98.
(21) Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. *Advances in Neural Information Processing Systems 26*; Curran Associates Inc.: Lake Tahoe, Nevada, 2013; pp 3111−3119.
(22) Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **2017**, *4*, 170127.
(23) Kingma, D. P.; Welling, M. *International Conference on Learning Representations*, 2014.
(24) Sohn, K.; Lee, H.; Yan, X. *Advances in Neural Information Processing Systems 28*; Curran Associates, Inc., 2015; pp 3483−3491.
(25) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(26) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **2017**, *3*, 53.
(27) Sun, W.; Holder, A.; Orvañanos, B.; Arca, E.; Zakutayev, A.; Lany, S.; Ceder, G. Thermodynamic Routes to Novel Metastable Nitrogen-Rich Nitrides. *Chem. Mater.* **2017**, *29*, 6936−6946.
(28) Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. *North American Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics, 2018.
(29) Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *TACL* **2017**, *5*, 135−146.
(30) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. *NIPS 2014 Workshop on Deep Learning*, 2014.
(31) Kamalakkannan, J.; Chandraboss, V. L.; Senthilvelan, S. Synthesis and characterization of InWO3-TiO2 nanocomposite material and multi application. *World Sci. News* **2016**, *58*, 97−121.
(32) Takatsu, H.; Hernandez, O.; Yoshimune, W.; Prestipino, C.; Yamamoto, T.; Tassel, C.; Kobayashi, Y.; Batuk, D.; Shibata, Y.; Abakumov, A. M.; Brown, C. M.; Kageyama, H. Cubic lead perovskite PbMoO3 with anomalous metallic behavior. *Phys. Rev. B* **2017**, *95*, 155105.
(33) Hegde, V. I.; Aykol, M.; Kirklin, S.; Wolverton, C. The Phase Diagram of all Inorganic Materials. **2018**, arXiv:1808.10869.
(34) Balachandran, P. V.; Emery, A. A.; Gubernatis, J. E.; Lookman, T.; Wolverton, C.; Zunger, A. Predictions of new ABO3 perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2018**, *2*, 043802.
(35) Yang, L.; Ceder, G. Data-mined similarity function between material compositions. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 224107.
(36) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. *Proceedings of the 34th International Conference on Machine Learning*; PMLR, 2017, pp 1945−1954.
(37) Chollet, F. Keras (accesed December 18, 2018) , 2015.
(38) Yu, Y.; Hawkins, P.; Isard, M.; Kudlur, M.; Monga, R.; Murray, D.; Zheng, X.; Abadi, M.; Barham, P.; Brevdo, E.; Burrows, M.; Davis, A.; Dean, J.; Ghemawat, S.; Harley, T. *EuroSys*; ACM Press, 2018; Vol. *16*, pp 265−283.
(39) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(40) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894−1904.

https://dx.doi.org/10.1021/acs.jcim.9b00995
J. Chem. Inf. Model. 2020, 60, 1194−1201