Inorganic Synthesis Reaction Condition Prediction with Generative Machine Learning

Christopher Karpovich, Zach Jensen, Vineeth Venugopal, Elsa Olivetti
Dept. of Materials Science and Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139
{ckarp, zjensen, vineethv, elsao}@mit.edu

Abstract

Data-driven synthesis planning with machine learning is a key step in the design and discovery of novel inorganic compounds with desirable properties. Inorganic materials synthesis is often guided by chemists' prior knowledge and experience, built upon experimental trial-and-error that is both time and resource consuming. Recent developments in natural language processing (NLP) have enabled large-scale text mining of scientific literature, providing open source databases of synthesis information of synthesized compounds, material precursors, and reaction conditions (temperatures, times). In this work, we employ a conditional variational autoencoder (CVAE) to predict suitable inorganic reaction conditions for the crucial inorganic synthesis steps of calcination and sintering. We find that the CVAE model is capable of learning subtle differences in target material composition, precursor compound identities, and choice of synthesis route (solid-state, sol-gel) that are present in the inorganic synthesis space. Moreover, the CVAE can generalize well to unseen chemical entities and shows promise for predicting reaction conditions for previously unsynthesized compounds of interest.

1 Introduction

Virtual materials screening and physics-based simulations have in recent years greatly accelerated the design and discovery of novel inorganic compounds with applications in chemical catalysis [1], thermoelectrics [2], and metal-organic frameworks [3]. However, while existing tools have largely focused on the inverse design of inorganic materials, one major remaining challenge is the development of inverse synthesis planning, where, given a target material, appropriate synthesis parameters are suggested as a means to synthesize the compound. In comparison to widely available organic chemistry reaction databases, the vast majority of openly accessible inorganic synthesis information is contained within the text of scientific journal articles [4]. Recent efforts have leveraged advances in Natural Language Processing (NLP) to extract and convert inorganic synthesis information in unstructured scientific text into machine readable databases [5, 6, 7]. In the organic chemistry space, previous works have investigated synthesis temperature prediction using feedforward neural network based models [8]; however, they utilize a dataset on the order of 10^7 points with a small range of reported temperatures (-100 to 300 °C), while inorganic synthesis datasets typically consist of 10^4 to 10^5 points and report temperatures ranging from 200 to 2000 °C. The dearth of available inorganic synthesis data and wide range of reported reaction conditions makes the inorganic prediction problem challenging. In the inorganic space, other works have developed methods for condition generation for specific materials families such as TiO_2, MnO_2, and $SrTiO_3$ [9], inverse prediction of precursor materials and synthesis operation sequences [10], forward prediction of target compositions [11], and precursor selection based on kinetic factors [12]; however, to the best of our knowledge, no studies have explored the prediction of synthesis conditions for novel inorganic compounds.

Generative machine learning models have already proven to be powerful tools in the chemical and materials spaces for materials discovery. Autoencoder models have been leveraged to optimize molecules with desirable properties over a learned latent space [13] and predict crystal structures for new inorganic compounds [14]. Generative adversarial networks have also been explored for drug discovery [15] and to screen inorganic material compositions [16]. In the case of experimental synthesis, reaction conditions present several important variables in the validation of computer-aided synthesis planning. Broadly, in solid-state synthesis, two or more non-volatile solid precursor materials are ground and heated in multiple consecutive steps to temperatures below their melting points to react and form the desired product [17, 18]. In sol-gel synthesis, a "sol" (a colloidal solution of particles in a solvent) is first heated to form a "gel", which is then typically heated in multiple consecutive steps to form the desired product [19, 20]. Common heating steps in both synthesis methods include calcination, where a mixture of compounds is heated to a high temperature to remove impurities and unwanted volatile substances, often through thermal decomposition [17], and sintering, where a compound is heated at a temperature (often higher than that reached by calcination) to induce nucleation and grain growth [17].

2   Methods

Our dataset consists of two publicly released materials synthesis databases (in JSON format) text-mined from scientific literature using a combination of NLP and rule-based extraction techniques [6]. The first is a solid-state synthesis database [6] containing 31,782 inorganic solid-state chemical reactions, while the second is a sol-gel synthesis database [6] containing 9,518 inorganic sol-gel chemical reactions. Each entry in the synthesis database contains a *target material:* the stoichiometric formula of the target compound synthesized in the reaction, *precursor materials:* the starting materials reacted together to form the target material, where a precursor is defined as a compound which shares one or more elements with the target material, excluding abundant elements that can be found in air (e.g. oxygen, hydrogen), and *processing actions and synthsis conditions:* the sequence of synthesis actions (e.g. mix, grind, calcine, sinter, dry) that were performed on the precursor materials to transform them into the target material. Relevant synthsis conditions for these processing actions include temperatures and times (when reported in the synthesis). The datasets report an overall extraction accuracy of 93% [6].

To make predictions for synthesis conditions we used (see Appendix) a conditional variational autoencoder (CVAE) with a convolutional encoder and recurrent decoder. A depiction of the model architecture is shown in Fig. 1. Temperatures and times of the four heating steps of interest (calcination, sintering, annealing, and drying) were represented as an 8-dimensional vector and standardized. Targets and precursors (represented by their chemical formulas, see Appendix) were used as conditions and encoded and concatenated with the latent space representation using convolutional layers over sequences of one-hot vectors, where the total vocabulary is a character set consisting of the different elements and the numerical digits. Two dataset splits were investigated: a random split, and a compositional-based split based on [21] where the train, validation, and test set do not contain materials involving the same set of elements. For instance, if $LiFePO_4$ is in the test set, then no other materials in the Li-Fe-P-O phase system (such as $LiFeP_2O_7$) would be allowed in the training or validation sets.

As shown in Fig. 2, we plot the overall distribution of calcination, sintering, annealing, and drying temperatures and times for both solid-state and sol-gel synthesis methods. The dataset is mainly composed of unique compounds, with the majority of the entries reporting novel target compositions. Notably, calcination temperatures follow a common trend such that nitride > oxide > carbonate > nitrate > acetate, alkoxide, acetylacetonate, oxalate which is correlated to bonding strength between cations and anions [22]. However, these trends are not absolute, as other factors such as decomposition reactions, reactivity of precursors, and identity of the target material play a role in the chosen calcination temperature.

3   Results and Discussion

In Fig. 3, we plot the predicted calcination and sintering temperatures for a selection of samples in the held-out test set. The CVAE model indeed learns a meaningful relationship between the composition of the target material, the identities of the precursors in the synthesis, and a range within
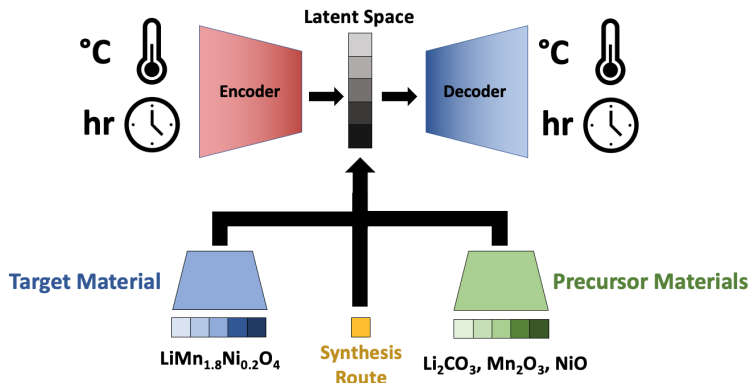
Figure 1: Architecture of CVAE used for temperature and time generation. The encoder embeds the synthesis conditions x to a latent vector z, and the decoder reconstructs the synthesis conditions from the latent vector. The latent vector is concatenated with learned representations of the target and precursor materials as well as a binary indicator of the synthesis route.
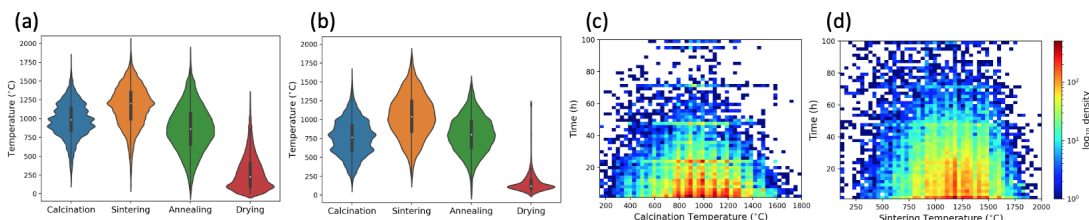

Figure 2: Temperature distributions for extracted operations for (a) solid-state and (b) sol-gel synthesis methods. (c) Calcination and (d) sintering time-temperature distributions for solid-state synthesis.

which the appropriate temperature for these heating operations should occur. For instance, in Fig. 3 (a)-(b) it is evident that when a compound consisting primarily of barium is doped with cesium, yttrium, and zirconium, it should require higher calcination and sintering temperatures than if doped with europium and copper. Moreover, the CVAE learns the common synthesis trend that sintering temperatures tend to be higher than calcination temperatures by the order of 100-300°C. In Fig. 3 (c)-(d), we show the parity plots for predicted vs. true mean calcination and sintering temperatures in the test set.

The quantitative performance metrics of the CVAE model are presented in Table 1, including mean absolute error (MAE), root-mean-square error (RMSE), mean relative error (MRE), and coefficient of determination (R^2). In experimental synthesis, suggesting approximate initial temperatures would be sufficiently helpful to inform experiments and accelerate synthesis of novel materials. On a held-out test set, the CVAE predicts the mean calcination temperature with a mean absolute error of 132.4 °C and mean sintering temperature with a MAE of 129.9 °C. To evaluate model performance against physically meaningful baselines, we leverage two common heuristics in the materials synthesis field as predictors on the dataset. Tamman's rule [23] is a synthesis heuristic which approximates reaction temperature as two-thirds the melting point of the lowest melting temperature reactant. For sintering temperature, we use a heuristic which approximates the value as 200 °C above the calcination temperature predicted by Tamman's rule, which is the average difference between sintering and calcination temperatures in the dataset. From Table 1, it is clear that the CVAE outperforms both heuristics by a factor of three to four.

Another factor to consider in evaluating model performance is the structure of the dataset itself. In scientific literature, the majority of publications report the unique synthesis of a single compound or family of compounds, meaning the reported synthesis temperature may not be representative of either the optimal temperature or the range of temperature within which it is feasible to synthesize the compound. Thus, since the dataset is comprised of single experimental data points per material, comparing the means of generated distributions of temperatures to single experimental points likely underestimates model performance. For instance, in Fig. 4 (a)-(b), we plot the MAE and R^2 metrics
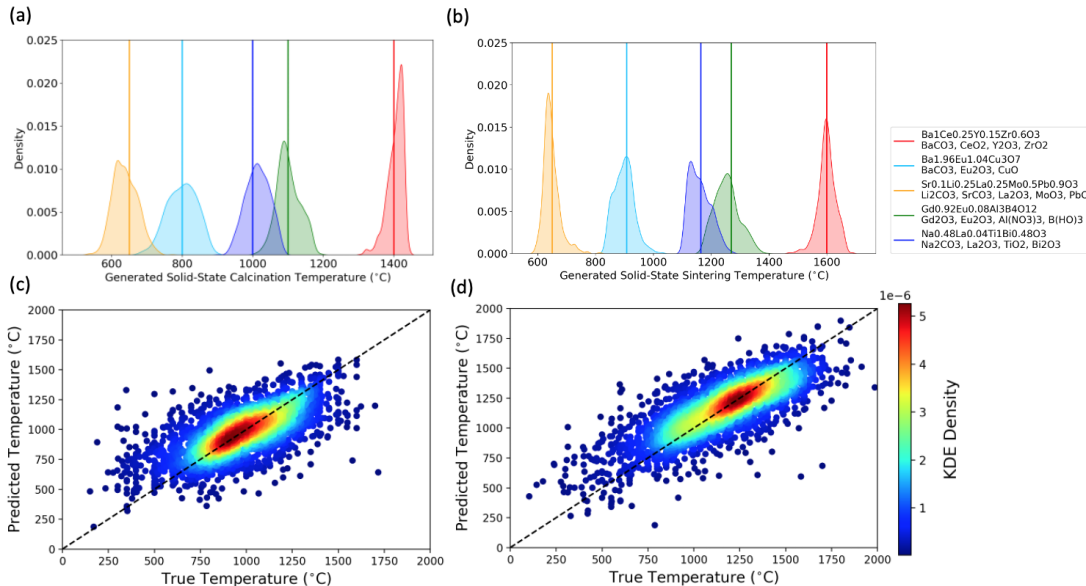
Figure 3: Generated solid-state (a) calcination and (b) sintering temperature distributions for a selection of samples in the held-out test set. Parity plots for predicted vs. true means of test set distributions of (c) calcination and (d) sintering temperatures. In the key, the target material is presented first (above) and precursor materials are presented second (below) for each entry.

Table 1: Dataset-wide performance of CVAE model compared to baseline heuristics

Model	Prediction Task	MAE (°C)	RMSE (°C)	MRE (%)	R^2
CVAE (random split)	Mean Calcination Temp.	132.4	180.9	16	0.40
CVAE (comp. split)	Mean Calcination Temp	147.0	190.9	18	0.32
Tamman's Rule	Mean Calcination Temp.	679.2	596.8	60	-8.68
CVAE (random split)	Mean Sintering Temp.	129.9	173.4	13	0.60
CVAE (comp. split)	Mean Sintering Temp	147.4	191.8	15	0.43
Sintering Heuristic	Mean Sintering Temp.	529.0	613.9	45	-5.03

as a function of the minimum number of literature data points for each example in the test set, showing marked improvement as the minimum number of points increases from one to five. With the experimental mean comprised of at least five points, the MAE is 57.0 °C and R^2 is 0.90 for sintering and 75.3 °C and 0.59 for calcination, respectively. We note that the MAE for calcination increases slightly as the minimum number of literature data points increases from 4 to 5 and the R^2 for calcination decreases slightly as the minimum number of literature data points increases from 3 to 5, which can be attributed to anthropomorphic factors in data reporting and systematic error discussed later.

To show that the CVAE also learns the appropriate trends in synthesis route and precursor substitutions, we plot in Fig. 4(c) the predicted calcination temperature distributions for the synthesis of Sr_2FeO_4 from $Sr(NO_3)_2$ and $Fe(NO_3)_3$, conditioning on either sol-gel or solid-state synthesis. Evidently, the CVAE learns the relationship that if we change our choice of synthesis route from solid-state to sol-gel, the calcination and sintering temperatures should decrease as well, reflective of the calcination temperature distributions in both datasets. We also plot in 5(a)-(b) the predicted calcination and sintering temperatures for the synthesis of $Li_4Fe_7O_{12}$ from Li_2CO_3 while varying the identity of the iron-based precursor. The CVAE model recognizes that when we alter our choice of precursor from high bonding strength candidates such as Fe_3O_4 and Fe_2O_3 to lower bonding strength candidates such as $Fe(NO_3)_3$, $FePO_4$, and FeC_2O_4, the calcination and sintering temperatures should appropriately decrease. These trends are reflective of the overall shift in calcination and sintering temperature across the entire dataset portrayed in Fig. 5 (c)-(d). However, these trends are not absolute, as for
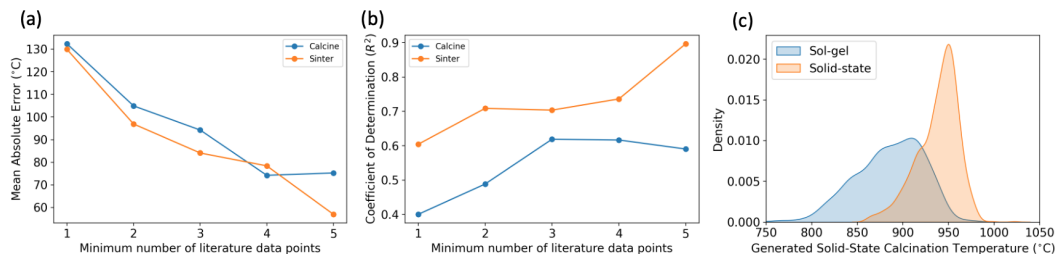
4

Figure 4: (a) MAE and (b) R^2 metrics as a function of minimum number of literature data points in the test set. (c) Generated calcination temperature distributions conditioned on either sol-gel or solid-state synthesis routes.

example the generated calcination temperature distribution using $FePO_4$ is higher than what would be expected simply by comparing with the literature-wide trends. While reported trends in synthesis temperatures and average bonding strength do correlate to an extent, other factors such as reactivity of precursors, intermediate reactions such as decomposition, the identity of the target compound, errors in automated extraction, and anthropomorphic factors [22, 24] such as experimentalist bias and past reported literature success all affect reported reaction conditions and influence the learned distributions.
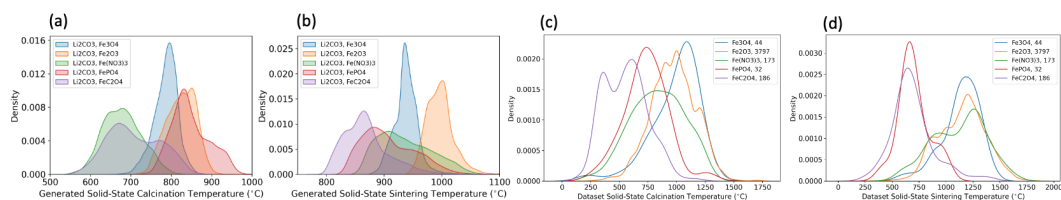

Figure 5: Generated solid-state (a) calcination and (b) sintering temperatures for the synthesis of $Li_4Fe_7O_{12}$ from Li_2CO_3 and various iron-containing precursors. (c) Calcination and (d) sintering temperature trends as a function of iron-containing precursors across the dataset. The inset number is the frequency of the reported precursor in the dataset.

4 Conclusions

We propose a CVAE model which suggests appropriate distributions for calcination and sintering temperatures in inorganic synthesis based on synthesis route, precursor identity, and the desired target compound. The model captures physics-based trends in the doping of target materials as well as choice of precursor and significantly outperforms common heuristics in the field in suggesting predictions of synthesis temperatures. We envision this model as a stepping-stone to high-throughput inorganic synthesis, where laboratory experiments will ultimately be informed by machine learning to hasten the synthesis and optimization of inorganic materials with desirable properties.

5 Broader Impact

Inorganic materials play a crucial role in the advancement of fields such as energy storage, chemical catalysis, thermoelectrics, and microelectronics. Therefore, new computational tools leveraging machine learning and scientific data need to be developed in order to accelerate the design and realization of novel materials with desirable properties. Our approach guides experimentalists with initial suggestions to aid in synthesis planning for the discovery and design of new materials. Outside of inorganic materials science, our approach can potentially be applied to a wide range of other fields which are currently adapting machine learning-based tools to unify theoretical predictions and experimental validation, such as drug discovery and electronic device manufacturing.

5

6 Acknowledgements

References

[1] Takashi Toyao, Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catalysis*, 10(3):2260–2297, feb 2019.

[2] Yuma Iwasaki, Ryohto Sawada, Valentin Stanev, Masahiko Ishida, Akihiro Kirihara, Yasutomo Omori, Hiroko Someya, Ichiro Takeuchi, Eiji Saitoh, and Shinichi Yorozu. Identification of advanced spin-driven thermoelectric materials via interpretable machine learning. *npj Computational Materials 2019 5:1*, 5(1):1–6, oct 2019.

[3] Michael Fernandez, Peter G Boyd, Thomas D Daff, Mohammad Zein Aghaji, and Tom K Woo. Rapid and accurate machine learning recognition of high performing metal organic frameworks for co2 capture. *The journal of physical chemistry letters*, 5(17):3056–3060, 2014.

[4] Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317, dec 2020.

[5] Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum, and Elsa Olivetti. Machine-learned and codified synthesis parameters of oxide materials. *Scientific Data 2017 4:1*, 4(1):1–9, sep 2017.

[6] Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):203, oct 2019.

[7] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702, sep 2019.

[8] Hanyu Gao, Thomas J. Struble, Connor W. Coley, Yuran Wang, William H. Green, and Klavs F. Jensen. Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Central Science*, 4(11):1465–1476, nov 2018.

[9] Edward Kim, Kevin Huang, Stefanie Jegelka, and Elsa Olivetti. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational Materials*, 3(1):53, dec 2017.

[10] 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. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *Journal of chemical information and modeling*, 60(3):1194–1201, mar 2020.

[11] Shreshth A. Malik, Rhys E. A. Goodall, and Alpha A. Lee. Predicting the Outcomes of Material Syntheses with Deep Learning. *Chemistry of Materials*, 16:acs.chemmater.0c03885, jan 2021.

[12] Muratahan Aykol, Joseph H. Montoya, and Jens Hummelshøj. Rational Solid-State Synthesis Routes for Inorganic Materials. *Journal of the American Chemical Society*, page jacs.1c04888, jun 2021.

[13] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, feb 2018.

[14] Callum J. Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M. Cole. 3-D inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, oct 2020.

[15] Gabriel Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, may 2017.

[16] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6(1):1–7, dec 2020.

[17] Anthony R West. *Solid state chemistry and its applications*. John Wiley & Sons, 2014.

[18] Juan R. Chamorro and Tyrel M. McQueen. Progress toward Solid State Synthesis by Design. *Accounts of Chemical Research*, 51(11):2918–2925, nov 2018.

[19] A. Vioux. Nonhydrolic sol-gel routes to oxides. *Chemistry of Materials*, 9(11):2292–2299, 1997.

[20] A. E. Danks, S. R. Hall, and Z. Schnepp. The evolution of 'sol–gel' chemistry as a technique for materials synthesis. *Materials Horizons*, 3(2):91–112, feb 2016.

[21] Christopher J. Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials 2020 6:1*, 6(1):1–11, jul 2020.

[22] Tanjin He, Wenhao Sun, Haoyan Huo, Olga Kononova, Ziqin Rong, Vahe Tshitoyan, Tiago Botari, and Gerbrand Ceder. Similarity of Precursors in Solid-State Synthesis as Text-Mined from Scientific Literature. *Chemistry of Materials*, 32(18):7861–7873, sep 2020.

[23] Rotraut Merkle and Joachim Maier. On the Tammann–Rule. *Zeitschrift für anorganische und allgemeine Chemie*, 631(6-7):1163–1166, may 2005.

[24] Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron E. Ruby, Hao Wang, Sorelle A. Friedler, Alexander J. Norquist, and Joshua Schrier. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature 2019 573:7773*, 573(7773):251–255, sep 2019.

7 Appendix

7.1 Theory

In a variational autoencoder (VAE), the loss function consists of the variational lower bound, also known as the evidence lower bound (ELBO):

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p_\theta(z)) \tag{1}$$

where $\mathbb{E}$ is the expectation value, p and q are probability distributions, D_{KL} is the Kullback–Leibler divergence, and x and z are the data and latent spaces, respectively. The first and second terms are often called the reconstruction loss and the KL loss, respectively. The reconstruction loss encourages the decoder to learn to reconstruct the data, while the KL loss is a regularization term which measures how similar the variational distribution encoded from the input data $q_\phi(z|x)$ and the latent space distribution $p_\theta(z)$ are. We take $p_\theta(z)$ to be a standard normal distribution with zero mean and unit variance, such that $p_\theta(z) = \mathcal{N}(0, 1)$. In a VAE, $q_\phi(z|x)$ and $p_\theta(x|z)$ are approximated by an encoder and a decoder, respectively. For a conditional variational autoencoder (CVAE), we embed

the conditional information as a vector (denoted by c) in the objective function of the VAE, leading to the revised loss function as follows:

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z,c)] - D_{KL}(q_\phi(z|x,c)\|p_\theta(z|c)) \tag{2}$$

In our model architecture, the data x consists of standardized temperatures and times of the four heating steps of interest (calcination, sintering, annealing, and drying) represented as an 8-dimensional vector. The condition c consists of targets and precursors (represented by their chemical formulas), encoded and concatenated with the latent space representation using convolutional layers over sequences of one-hot vectors, where the total vocabulary is a character set consisting of the different elements and the numerical digits. An additional condition (if desired) is the synthesis route, which is a binary condition (solid-state or sol-gel). Using this framework, the CVAE model can generate temperature and time distributions conditioned on the target and precursor materials of interest.

7.2 Model architecture

All neural network models were implemented in the Keras library using the TensorFlow backend. For the conditional variational autoencoder (CVAE) model, convolutional layers encode the temperature-time vector 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. The encoder is comprised of three convolutional layers and the decoder comprised of three gated recurrent unit (GRU) layers. In producing the results for this study, 3 latent dimensions were used. Targets and precursors (represented by their chemical formulas) were used as conditions and encoded and concatenated with the latent space representation using convolutional layers over sequences of one-hot vectors, where the total vocabulary is a character set consisting of the different elements and the numerical digits. A period was included in the character set for targets to represent non-stoichiometric target formulas. Training was conducted on two NVIDIA Titan Xp GPUs with a batch size of 128 and the Adam optimizer with default hyperparameters. Hyperparameter selection was performed by grid searches, where the latent layer dimension was varied from 2 to 5 dimensions and the standard deviation of the Gaussian prior was varied between 0.001 and 10.0. The data was split in a 75/15/10 train/validation/test ratio using either random or compositional splits, and hyperparameters were selected based on minimizing validation loss.

7.3 Data post-processing

Our post-processing of the data was completed as follows. First, reactions with organic precursors and targets, non-stoichiometric precursors, unsubstituted target stoichiometries, less than two or greater than five precursors, or not containing at least one relevant heating step with a reported temperature were removed. Hydrate precursors were truncated to their base chemical formula. Temperatures and times were converted into units of Celsius and Hours and limited to between 100 °C and 2000 °C and less than 100 hours, and if an operation step was reported with more than one temperature or time, the highest value was taken. If a relevant heating step occurred more than once in a recipe, the last value was taken. Because not every synthesis recipe employed all four heating steps, data imputation was conducted using the `IterativeImputer` module in `scikit-learn` with the BayesianRidge estimator, default hyperparameters, and minimum and maximum imputation values set to those in the dataset. To aid in accurate imputation, precursors were one-hot encoded and used as additional features. Temperatures and times were standardized by removing the mean and scaling to unit variance per feature.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work focuses on leveraging scientific data from literature to advance materials discovery, so there is no potential negative societal impact to discuss.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [N/A]
    (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The data is open-source and freely available. The code and project are still a work-in-progress and code will be released upon full publication of this work at a later date.
    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [Yes]
    (b) Did you mention the license of the assets? [Yes]
    (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]