Element-Wise Formulation of Inorganic Retrosynthesis

Seongmin Kim
KAIST
Daejeon 34141, South Korea

Juhwan Noh
KAIST
Daejeon 34141, South Korea

Geun Ho Gu
KENTECH
Naju 58330, South Korea

Shuan Chen
KAIST
Daejeon 34141, South Korea

Yousung Jung[*]
KAIST
Daejeon 34141, South Korea

Abstract

Synthesizing new inorganic functional materials is a practical goal of materials science. While the advances in computational techniques accelerated the virtual design, the actual synthesis of predicted candidate materials still remain as an expensive and slow process. While a few initial studies attempted to predict the synthesis routes for inorganic crystals, the existing models do not yield the uncertainty of the predictions and could produce thermodynamically unrealistic precursor chemicals. Here, we propose an element-wise graph neural network to predict the inorganic synthesis recipes. The trained model outperforms the popularity-based statistical baseline model for *top-k* exact match accuracy test, showing the validity of our approach for inorganic solid-state synthesis. We further validate our model by the publication-year-split test, where the model trained based on the materials data until the year 2016 is shown to successfully predict the synthetic precursors for the materials synthesized after 2016. The high correlation between the classification score and prediction accuracy suggests that the prediction score can be interpreted as a measure of uncertainty.

1 Introduction

Synthesizing new inorganic functional materials is a practical goal of materials science in various fields such as batteries, [1, 2, 3] (photo-)electrochemical catalyst,[4, 5] and solar cell[6] to name a few. While the advances in computational power and electronic structure calculation methods helped to design new materials in a pace much faster than before,[7, 8, 9, 10] the actual synthesis of predicted candidate materials still remain as a slow process due to an empirical nature of synthesis. Thus, to reduce the time and cost associated with failed syntheses, efforts to understand the chemistry of materials synthesizability have been attempted in literature. Fedorovskiy et al.[11] and Ouyang et al.[12] suggested the use of heuristic rules to predict the materials synthesizability, e.g., Goldschmidt's tolerance factor for double halide perovskites or the stability rules for NASICON-structured materials. Since these heuristic rules for synthetic accessibility are usually domain specific, several thermodynamic quantities (the energy above the convex hull and the decomposition enthalpies) obtained from electronic structure calculations have been widely used as a guideline to estimate synthesizability.[13, 14, 15, 16] More recently, data-driven machine learning (ML) models have been proposed to calculate the materials thermodynamics[17, 18, 19] or the synthesizability of materials based on the structural similarity.[20, 21]

[*] Correspondence to ysjn@kaist.ac.kr

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

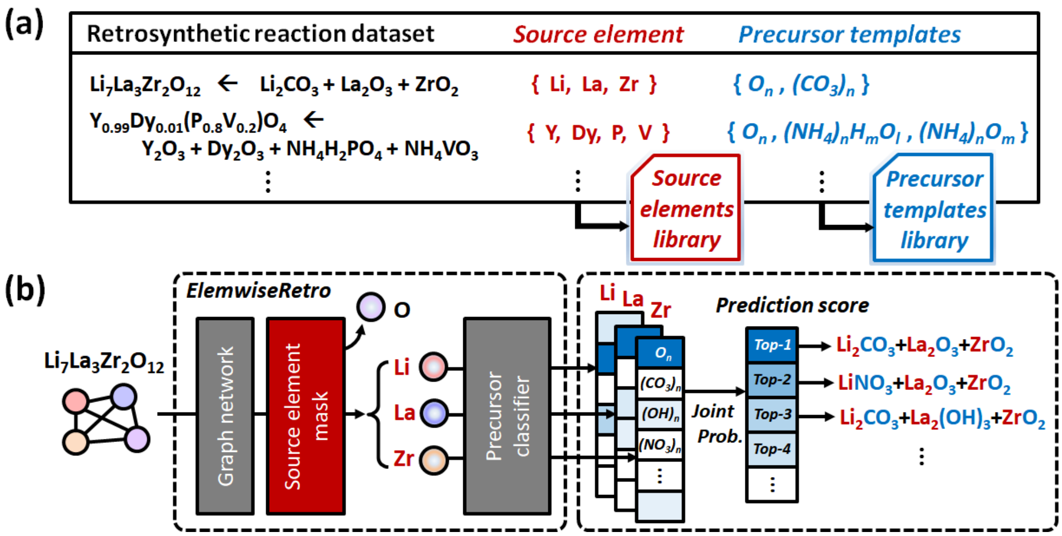

Figure 1: The overview of (a) the formulation of source elements and precursor templates libraries and (b) the inorganic retrosynthetic model architecture (see also Figure 4 for more details on the *ElemwiseRetro*).

Beyond the synthetic feasibility predictions as briefly described above, a few studies attempted to further suggest synthesis routes for inorganic materials. For example, based on thermodynamic parameters and some kinetic heuristics, favorable pathways for inorganic materials could be constructed.[22, 23] Notably, several data-driven models have also been proposed to generate the precursors and synthetic conditions (e.g. heating temperature and time) to synthesize the target materials using the text-mined meta-datasets.[24, 25, 26, 27] However, the outcomes from generative models used in the latter studies[27] can contain thermodynamically unstable precursors and do not generally inform the measure of uncertainty for predicted reactions, requiring an additional process by domain experts to screen or rank the generated reaction recipes.

This early stage of inorganic retrosynthetic reaction prediction can be contrasted with the organic synthesis planning where there are a number of template-free[28, 29, 30] and template-based[31, 32] models with promising prediction accuracy. This difference suggests that the concepts used in the successful organic retrosynthesis models may be borrowed and adapted to address the inorganic retrosynthesis problem. In particular, by noting that most solid-state inorganic syntheses are performed using a finite list of commercial precursors, we envision that the set of popular inorganic precursors used in literature can be seen as a “template" for inorganic solid-state synthesis, and a similar probability-based template selection model used in the organic retrosynthesis can be used in the inorganic synthesis planning. This template-based recommendation would remove the possibility of yielding unrealistic precursor chemicals in some of the existing generative model-based inorganic retrosynthesis predictions.[27]

In this work, we introduce a template-based graph neural network for inorganic synthesis recipe prediction. The model is trained to predict a set of precursors for inorganic crystals by ranking the sets of precursors as the probability scores. Temperature for the solid-state reaction is another important parameter in actual experimental synthesis, that is affected by both the target crystals and detailed precursors chosen. Thus, we additionally constructed a temperature prediction model that is sequentially-connected to the precursor set prediction model. These two models combined then generate a set of precursors and temperature to produce a target solid compound. Due to a high correlation between the prediction score and the prediction accuracy, the proposed model has a key advantage of quantifying uncertainty of the predictions.

2 Methods

Element-wise formulation of inorganic retrosynthesis We formulate the retrosynthetic problems of inorganic materials by first dividing the chemical elements in the target product into two types:

Table 1: The $top\text{-}k$ exact match accuracy for the prediction of inorganic synthesis precursors by *ElemwiseRetro* and the popularity-based baseline model.

	Model		
$Top\text{-}k$ accuracy (%)	*ElemwiseRetro*	*ElemwiseRetro*-TimeSplit	Baseline
$k = 1$	**83.06 ±0.05**	**83.26 ±0.14**	54.00 ±0.46
$k = 2$	**91.14 ±0.04**	**91.28 ±0.17**	74.92 ±0.24
$k = 3$	**94.14 ±0.07**	**94.62 ±0.21**	79.62 ±0.19
$k = 4$	**95.98 ±0.15**	**96.10 ±0.14**	81.56 ±0.20
$k = 5$	**97.60 ±0.15**	**96.92 ±0.13**	82.64 ±0.15

elements that have to be provided as reaction precursor (denoted as "source element") and elements that can come from the reaction environments (denoted as "environmental element"). After selecting the source elements from the given target inorganic compositions, proper anionic frameworks (denoted as "precursors templates") have to be attached to each source element to complete the actual precursor compounds. This formulation of the problem is summarized in Figure 1a.

To categorize the source and environmental elements, we examined the text-mined inorganic reaction database[33]. To that end, we assigned the metal groups (alkali, alkaline, transition, lanthanide, actinide, post transition), metalloid, phosphorus, selenium, and sulfur as the source elements, and the others as environmental elements from the inorganic retrosynthetic point of view. Based on these definitions, we constructed the total 39 precursors templates (tabulated in Table 2) from the 11,122 curated inorganic retrosynthetic datasets. The detailed procedures for the dataset selection and curation, and the precursor template extraction are described in the supplementary section.

Retrosynthetic model Based on these definitions, for a given target composition, the compound is encoded as $Roost$[34] representaiton, which is a 2D graph whose node features are obtained from a separate pretrained representation of inorganic compounds. Once the representation is fed into the model, the inorganic retrosynthetic model predicts the precursors that can provide all source elements contained in the given target composition using the source element mask, as shown in Figure 1b. The formulated source element mask enables the model to discriminate the source elements (Li, La, and Zr) information from the given compositions ($Li_7La_3Zr_2O_{12}$). Each source element is separately used in the following precursor classifier which predicts the precursor in the formulated templates library. By calculating the joint probability of a set of precursors determined for each source element, the precursor-sets (synthesis "recipe") are finally predicted as a probability score which can be ranked. The brief and detailed architecture of the proposed model, *ElemwiseRetro*, is described in Fig. 1 and the supplementary section, respectively.

3 Experiments

Precursor set prediction To demonstrate the *ElemwiseRetro* model performance, we calculated the $top\text{-}k$ exact match accuracy for the test dataset. Since the model might capture merely the popularity trend of literature-reported examples, as recently discussed for some organic retrosynthesis predictions,[35] we constructed the template-popularity-based model as a baseline comparison. In this baseline, the prediction is made statistically based on the number of examples in which a particular template appears in the dataset. The results for the $top\text{-}k$ exact match tests are shown in Table 1. The error bars correspond to the standard deviation of the 5 trained models. The proposed *ElemwiseRetro* shows the promising 83 % top-1, and 97 % top-5 accuracy, as compared to the popularity baseline model whose top-1 and top-5 accuracies are 54 % and 83 %, respectively.

Uncertainty estimation The reliability of the prediction is important to measure in order to prioritize and manage the cost of experimental environments. We split the top-1 exact match accuracy of precursors set prediction depending on their prediction scores. As shown in Figure 2a, a positive correlation between the prediction score and the accuracy is clear. This means that the predictions with higher classification scores can be considered as more reliable predictions.

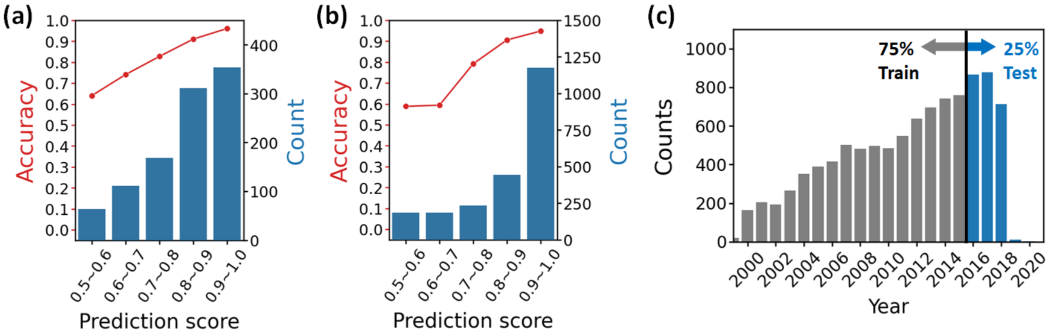

Figure 2: The prediction accuracy of precursors sets as a function of prediction scores for (a) the randomly split and (b) the publication-year-split test dataset. (c) Inorganic reaction dataset counted by published years. After training the model only using the dataset before 2016, the dataset after 2016 were tested to validate the model transferability up to the afterward time space.

Publication-year-split validation To further validate our model, we performed the publication-year-split test, for which we mined the published years of each data using the DOIs tagged in the inorganic reaction database (Figure 2c). In this benchmark, instead of splitting the entire dataset between the training and test sets randomly as in the original *ElemwiseRetro*, we used the time sequence to split the dataset, the data before year 2016 for training ($\sim 75\%$), and the data after 2016 for testing ($\sim 25\%$). As the accuracy results for this time split case is summarized in Table 1, both original *ElemwiseRetro* and *ElemwiseRetro*-TimeSplit yield the consistent model performance. Furthermore, we split the top-1 exact match accuracy depending on their classification scores (Figure 2b), and a positive correlation is still clear even though the test dataset was derived from the out of time domain. This result clearly suggests that our model can be used to discover undiscovered inorganic materials in the future.

Synthetic temperature prediction The synthetic temperature prediction model performance was investigated by the 2D parity heat map as shown in Figure 5. The predicted temperatures from the model reproduces the real temperatures qualitatively with the mean absolute error (MAE) of $117.9°C$, which outperforms the MAE ($\sim 140°C$) of previous results.[25, 26] Nevertheless, a wide range of temperatures ($300 \sim 1600°C$) used to synthesize a target crystal with a limited number of data points is potentially contributing to a relatively large MAE observed here.

4 Conclusion

We proposed an element-wise template-based retrosynthetic model that enables a probabilistic prediction of precursors set for inorganic crystals and the corresponding synthetic temperatures. Based on the concept of "source element" and "environmental element", we derived a set of precursor templates of inorganic crystal compounds. We demonstrated a promising model performance by the top-k exact match accuracy test. The observed positive correlation between the classification score and the prediction accuracy also allows us to estimate the uncertainty of the predictions. We further validated our model by the publication-year-split test, which suggests that our model has a possibility of covering up to the afterward time space where novel inorganic materials will be discovered. While the current approach is the first and initial effort to use a probabilistic modeling for inorganic solid-state retrosynthesis predictions with uncertainty estimation, we expect that the concept of templates, source-element decomposition, and element-wise prediction proposed here can be a promising direction to further develop inorganic retrosynthetic models with improved performance.

Acknowledgments and Disclosure of Funding

We acknowledge the support from the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government under grant No. 2021-0-02068 (Artificial Intelligence Innovation Hub) and the National Science Foundation of Korea (NRF-2019M3D3A1A01069099 and NRF-2019M3D1A1079303).

References

[1] Jan N Reimers and JR Dahn. Electrochemical and in situ x-ray diffraction studies of lithium intercalation in li x coo2. *Journal of the Electrochemical Society*, 139(8):2091, 1992.

[2] Arumugam Manthiram, James C Knight, Seung-Taek Myung, Seung-Min Oh, and Yang-Kook Sun. Nickel-rich and lithium-rich layered oxide cathodes: progress and perspectives. *Advanced Energy Materials*, 6(1):1501010, 2016.

[3] KJPC Mizushima, PC Jones, PJ Wiseman, and John B Goodenough. Lixcoo2 (0< x<-1): A new cathode material for batteries of high energy density. *Materials Research Bulletin*, 15(6):783–789, 1980.

[4] Akihiko Kudo and Yugo Miseki. Heterogeneous photocatalyst materials for water splitting. *Chemical Society Reviews*, 38(1):253–278, 2009.

[5] Yujie Sun, Chong Liu, David C Grauer, Junko Yano, Jeffrey R Long, Peidong Yang, and Christopher J Chang. Electrodeposited cobalt-sulfide catalyst for electrochemical and photoelectrochemical hydrogen generation from water. *Journal of the American Chemical Society*, 135(47):17699–17702, 2013.

[6] Juan-Pablo Correa-Baena, Michael Saliba, Tonio Buonassisi, Michael Grätzel, Antonio Abate, Wolfgang Tress, and Anders Hagfeldt. Promises and challenges of perovskite solar cells. *Science*, 358(6364):739–744, 2017.

[7] Jens Kehlet Nørskov, Thomas Bligaard, Jan Rossmeisl, and Claus Hviid Christensen. Towards the computational design of solid catalysts. *Nature chemistry*, 1(1):37–46, 2009.

[8] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127, 2016.

[9] Tim Mueller, Geoffroy Hautier, Anubhav Jain, and Gerbrand Ceder. Evaluation of tavorite-structured cathode materials for lithium-ion batteries using high-throughput computing. *Chemistry of materials*, 23(17):3854–3862, 2011.

[10] Geoffroy Hautier, Anubhav Jain, Shyue Ping Ong, Byoungwoo Kang, Charles Moore, Robert Doe, and Gerbrand Ceder. Phosphates as lithium-ion battery cathodes: an evaluation based on high-throughput ab initio calculations. *Chemistry of Materials*, 23(15):3495–3508, 2011.

[11] Alexander E Fedorovskiy, Nikita A Drigo, and Mohammad Khaja Nazeeruddin. The role of goldschmidt's tolerance factor in the formation of a2bx6 double halide perovskites and its optimal range. *Small Methods*, 4(5):1900426, 2020.

[12] Bin Ouyang, Jingyang Wang, Tanjin He, Christopher J Bartel, Haoyan Huo, Yan Wang, Valentina Lacivita, Haegyeom Kim, and Gerbrand Ceder. Synthetic accessibility and stability rules of nasicons. *Nature communications*, 12(1):1–11, 2021.

[13] Christopher J Bartel. Review of computational approaches to predict the thermodynamic stability of inorganic solids. *Journal of Materials Science*, pages 1–24, 2022.

[14] Arunima K Singh, Joseph H Montoya, John M Gregoire, and Kristin A Persson. Robust and synthesizable photocatalysts for co2 reduction: a data-driven materials discovery. *Nature communications*, 10(1):1–9, 2019.

[15] Muratahan Aykol, Shyam S Dwaraknath, Wenhao Sun, and Kristin A Persson. Thermodynamic limit for synthesis of metastable inorganic materials. *Science advances*, 4(4):eaaq0148, 2018.

[16] Christopher J Bartel, Alan W Weimer, Stephan Lany, Charles B Musgrave, and Aaron M Holder. The role of decomposition reactions in assessing first-principles predictions of solid stability. *npj Computational Materials*, 5(1):1–9, 2019.

[17] Wei Li, Ryan Jacobs, and Dane Morgan. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Computational Materials Science*, 150:454–463, 2018.

[18] 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*, 6(1):1–11, 2020.

[19] Gordon GC Peterson and Jakoah Brgoch. Materials discovery through machine learning formation energy. *Journal of Physics: Energy*, 3(2):022002, 2021.

[20] Jidon Jang, Geun Ho Gu, Juhwan Noh, Juhwan Kim, and Yousung Jung. Structure-based synthesizability prediction of crystals using partially supervised learning. *Journal of the American Chemical Society*, 142(44):18836–18843, 2020.

[21] Ali Davariashtiyani, Zahra Kadkhodaie, and Sara Kadkhodaei. Predicting synthesizability of crystalline materials via deep learning. *Communications Materials*, 2(1):1–11, 2021.

[22] Matthew J McDermott, Shyam S Dwaraknath, and Kristin A Persson. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nature communications*, 12(1):1–12, 2021.

[23] Muratahan Aykol, Joseph H Montoya, and Jens Hummelshøj. Rational solid-state synthesis routes for inorganic materials. *Journal of the American Chemical Society*, 143(24):9244–9259, 2021.

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

[25] Christopher Karpovich, Zach Jensen, Vineeth Venugopal, and Elsa Olivetti. Inorganic synthesis reaction condition prediction with generative machine learning. *arXiv preprint arXiv:2112.09612*, 2021.

[26] Haoyan Huo, Christopher J Bartel, Tanjin He, Amalie Trewartha, Alexander Dunn, Bin Ouyang, Anubhav Jain, and Gerbrand Ceder. Machine-learning rationalization and prediction of solid-state synthesis conditions. *arXiv preprint arXiv:2204.08151*, 2022.

[27] Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, et al. Inorganic materials synthesis planning with literature-trained neural networks. *Journal of chemical information and modeling*, 60(3):1194–1201, 2020.

[28] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[29] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113, 2017.

[30] Kangjie Lin, Youjun Xu, Jianfeng Pei, and Luhua Lai. Automatic retrosynthetic route planning using template-free models. *Chemical science*, 11(12):3355–3364, 2020.

[31] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

[33] 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):1–11, 2019.

[34] Rhys EA Goodall and Alpha A Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature communications*, 11(1):1–9, 2020.

[35] Wiktor Beker, Rafał Roszak, Agnieszka Wołos, Nicholas H Angello, Vandana Rathore, Martin D Burke, and Bartosz A Grzybowski. Machine learning may sometimes simply capture literature popularity trends: A case study of heterocyclic suzuki–miyaura coupling. *Journal of the American Chemical Society*, 144(11):4819–4827, 2022.

[36] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports*, 8(1):1–13, 2018.

[37] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[38] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section A.1 about the reaction coverage issue.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (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? [Yes] See Section 2 and A.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3.
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)? [No] This work is unpublished and ongoing process, yet.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section C.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 1.
   (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] See Section C.1.
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] We cited [33] for the data curation and [34, 36, 37, 38, 39] for the model construction in References section
   (b) Did you mention the license of the assets? [N/A]
   (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? [Yes]
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]

A Method details

A.1 Dataset preparation

For preparing the dataset for train and test, we started with the inorganic synthesis-related dataset[33] which was text-mined from the literatures published after the year 2000. The raw text-mined data contain some incomplete entries thus we further refined the data. We removed the data with missing

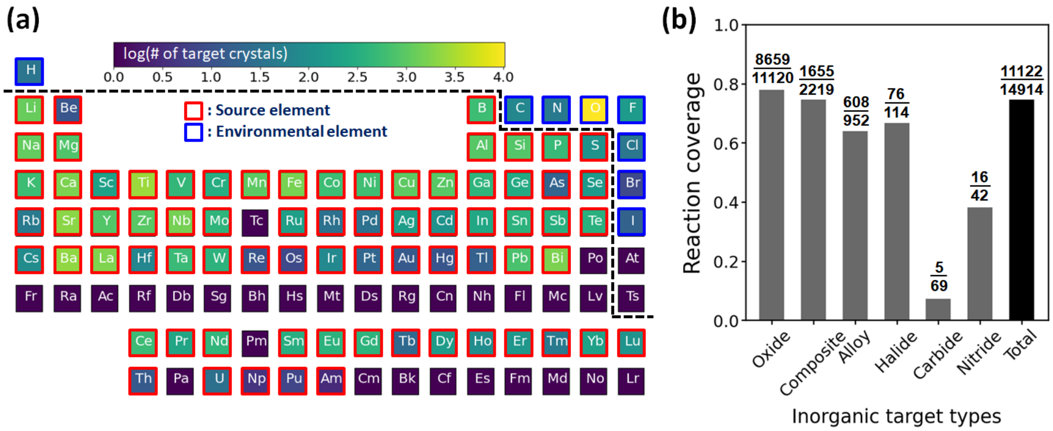

Figure 3: (a) By the definition of the formulated source elements and environmental elements, the element coverage in the periodic table for making up target material compositions are shown as the colored map counted by target crystal compositions. The source elements and the environmental elements are boxed with yellow and blue lines, respectively. (b) Based on the formulated source elements and precursor templates, reaction coverage ratios depending on the most frequent inorganic target types are displayed.

or incorrectly parsed elements and stoichiometry, which reduces the data size from 31,782 to 25,873. Next, the entries with the inconsistency between the target crystal elements and the source elements in precursors were removed, resulting in 22,837 data. The data are further trimmed by selecting data where only one source element is present for each precursor, which is the case for the most of the data (20,843) due to its affordability in real experimental synthesis for any type of inorganic synthesis. The duplicate cases were removed (13,445), and few cases of the unstable or metastable precursors like Li_2O_2 were also filtered out due to unmanageability in real synthetic environments.

For synthesis step with several experiment, the synthetic temperature was calculated by averaging. The data with the synthesis temperature less than $300°C$ and more than $1600°C$ were removed, as they are outliers. For multi-step reaction cases which have more than one heating step, we took the average temperature to represent the overall reaction. Those with the high standard deviation data were removed. In this study, several other reaction conditions (e.g. sequence of action verbs, type of mixing device, heating atmosphere, etc.) which might be up to each laboratorial standardized procedure would not be considered, incorporating the conditions is a topic of future work.

Through the aforementioned preprocessing, our final dataset size is 11,122 for the precursors set prediction from the targets and 7,541 for the synthetic temperature prediction. The whole dataset was divided into training: validation: test (8: 1: 1) to separate test data from training process. Figure 3 shows the coverage of the inorganic reaction data based on the formulated source elements and precursors templates. The element coverage (Figure 3a) in the periodic table which makes up the target crystals and the reaction coverage (Figure 3b) depending on the target types in the inorganic reaction dataset were shown, which represents our inorganic reaction domain. Although the total reaction coverage from our template-based approach is 74.6 $\%$, which should be further developed, our formulated concepts still have a possibility to handle reactions involving most elements and the broad types of popular inorganic materials (e.g. oxide, composite, alloy, halide, and etc.).

A.2 Precursor templates extraction

In predicting the retrosynthetic precursors for given inorganic materials, we used source element-wise precursors templates to determine each types of precursor compounds. After thoroughly investigating the whole 11,122 inorganic synthetic dataset which was curated by the abovementioned preprocessing, we obtained the list of the precursors templates (e.g. -CO_3 in $Li_2(CO_3)$, $Na_2(CO_3)$, and -OH in $Li(OH)$, $Al(OH)_3$). The full list of total 39 retrosynthetic precursor templates are shown in Table 2, which appeared at least once in the whole curated dataset. Based on this precursor templates, our retrosynthetic model can predict each precursor per one source element within the pre-defined 39 template space.

Table 2: The list of 39 templates of precursor frameworks which appeared in the inorganic synthetic dataset.

Formula	Nomenclature	Formula	Nomenclature
-O_n	- oxide	-$(OOH)_n$	- oxy-hydroxide
-$(CO_3)_n$	- carbonate	‘ ‘	(Pure source element)
-$(OH)_n$	- hydroxide	-$(CO)_n$	- carbonyl
-$(NO_3)_n$	- nitrate	-$(CH_3O)_n$	- methoxide
-F_n	- fluoride	-$(C_2H_5O)_n$	- ethoxide
-$(OF)_n$	- oxyfluoride	-$(C_3H_7O)_n$	- isopropoxide
-Cl_n	- chloride	-$(C_4H_9O)_n$	- tert-butoxide
-$(ClO)_n$	- hypochlorite	-$(C_6H_5O_7)_n$	- citrate
-Br_n	- bromide	-$(CH_3COO)_n$	- acetate
-$(BrO)_n$	- oxybromide	-$[(CH_3COO)O]_n$	- yl acetate
-In	- iodide	-$(C_5H_7O_2)_n$	- acetylacetonate
-H_n	- hydride	-$[(C_5H_7O_2)_2O]_n$	- oxy-acetylacetonate
-C_n	- carbide	-H_nO_m	- ic acid
-$(CN)_n$	- cyanide	-$(NH_4)_nO_m$	- ammonium ate
-$(C_{0.7}N_{0.3})_n$	- carbo-nitride	-$(NH_4)_nH_mO_l$	- ammonium ate acid
-N_n	- nitride	-$[(NH_4)_2(NO_3)_6]_n$	- ammonium nitrate
-$(ON)_n$	- oxynitride	-$[(CO_3)_4(OH)_2]_n$	- carbonate hydroxide
-$(NH_2)_n$	- amide	-$(NO_3)_nO_m$	- nitrate oxide
-$[C_2H_3(NH_2)]_n$	- acetamide	-$(NO_3)_nO_m(OH)_l$	- oxy-nitrate-hydroxide
-$(C_2O_4)_n$	- oxalate	-	-

B Model details

B.1 Retrosynthetic model

We denote our retrosynthetic precursor prediction model as *ElemwiseRetro*, and the other for the synthetic temperature prediction. The overall schematics of the two model architectures are illustrated in Figure 4. In order to find the plausible set of precursors that could synthesize the target product, the composition of the inorganic target material was converted to a graph representation, referred to as $Roost$.[34] The atomic feature vectors learned from the $ElemNet$[36] were embedded as the initial node states of the inorganic graph. We then apply the message passing neural network (MPNN)[37] to the graph representation, which updates the initial atomic features by the surrounding environmental information. Typically, after passing the MPNN, the pooling operation is used to gather the updated node vectors in order to obtain single inorganic descriptor as one-to-one mapping with target input. Instead of using the pooling layer, however, we extracted the node vectors, which correspond to the source elements, from the updated atomic features to solve the retrosynthetic one-to-many (target-to-precursors) problem. Then the source element descriptors were separately entered into the prediction network (element-wise prediction).

After the training with the precursor templates, the retrosynthetic model classified each source elements to infer their precursor template classes. At the end of the classifier, probability score distributions of the precursor templates for each source elements were obtained by the SoftMax layer. Using this individual probability, we can automatically compute the joint probability, resulted in the set of precursors outcome. These probability concept enables the model to derive the most synthetically probable precursors for inorganic retrosynthesis by ranked as descending probability scores.

After predicting the set of precursors by *ElemwiseRetro*, both the target and precursors were inputted in the second model for predicting their synthetic temperature (Figure 4b). The compositions of the target and precursors were converted to inorganic graph by aforementioned $Roost$. To distinguish information between the target and precursors, the atomic nodes in the inorganic graph were only intra-connected within the target and precursor set, separately. Therefore, the target (or precursor) atomic features were updated only from the surrounding target (or precursor) information. After the MPNN, the attention pooling layer was applied to extract the target and precursor descriptors from

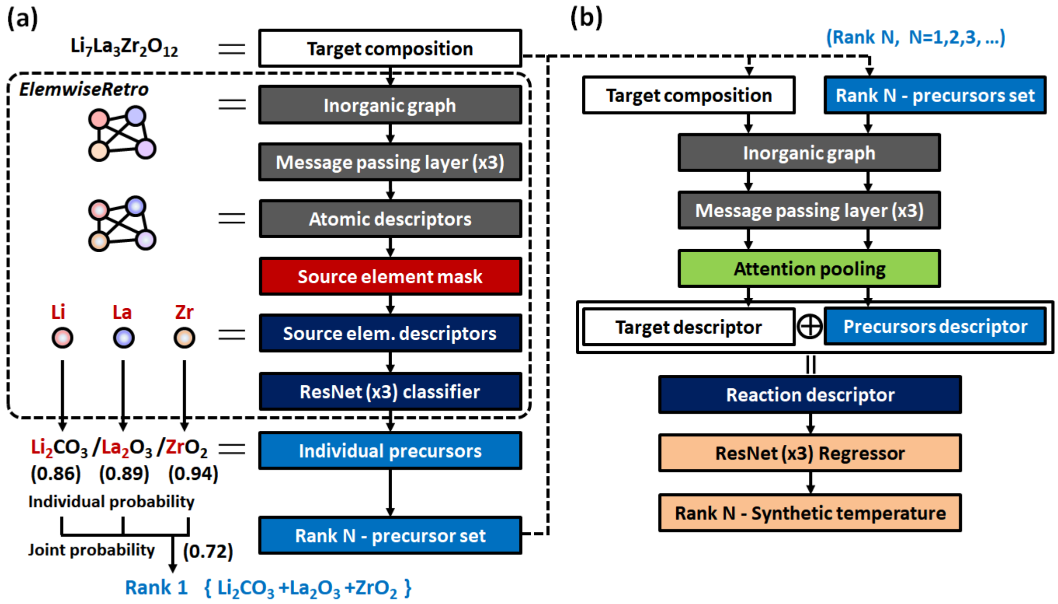

Figure 4: Schematic diagram of (a) the retrosynthetic model (*ElemwiseRetro*) architecture for the precursor set prediction and (b) the synthetic temperature prediction. When predicting the set of precursors for $Li_7La_3Zr_2O_{12}$, the updated source element feature vectors (Li, La, and Zr) are used for the precursor template classifier, resulting in the individual precursors outcome with each probability scores. Finally, the set of precursors were derived from the joint probability (The rank-1 result came from the highest joint probability).

Table 3: The *top-k* exact match accuracy for the prediction of inorganic synthesis precursors by three retrosynthetic the ablation models.

	Model		
Top-k accuracy (%)	*Elem-wise*	*Elem-wise* w.GLA	Global agg.
$k = 1$	**83.7**	83.0	66.8
$k = 2$	90.9	**91.0**	74.6
$k = 3$	**94.5**	**94.5**	82.4
$k = 5$	**97.0**	96.9	89.0
$k = 10$	**98.3**	**98.3**	95.8

the updated target and precursor graphs, respectively. Then the two descriptors were concatenated and fed into the regressor network to predict their synthetic temperature.

B.2 Model construction

The atomic feature vectors learned from the $ElemNet$[36] were embedded as the initial node vectors of the inorganic graph. The atomic embedding dimension is 136, which is mapped to 63 dimensions by one linear layer. The stoichiometric weight is concatenated to each mapped atomic vector, resulting that the initial node dimension is 64. We used three MPNN layers to update the node features. The three hidden layers of prediction network have 512/ 512/ 512 nodes. At the end of the prediction network, SoftMax layer is used in *ElemwiseRetro*.

B.3 Model ablation study

To elucidate the crucial components of the model, we constructed ablation models. The schematic architectures for the ablation models were illustrated in Figure 6. To investigate the pooling effect that combines the updated atomic node features to one global descriptor, pooling layer was added to

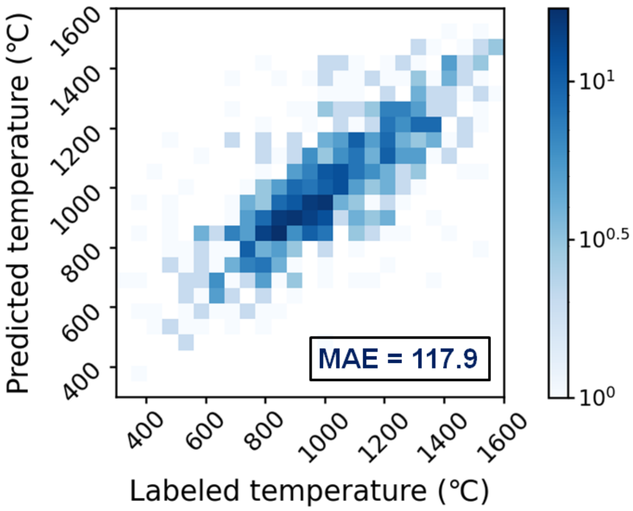

Figure 5: The result of 2D parity heat map from the synthetic temperature prediction model.

ElemwiseRetro which denoted as the global aggregated prediction model (Global agg.), as shown in Figure 6a. To further develop the source element-wise prediction model, the global descriptor is concatenated with the initial atomic features, the model of which is denoted as the source element-wise with global aggregated prediction model (Elem-wise w. GLA), as shown in Figure 6b. All prediction networks for the ablation test were composed of identical GRU layers, comparing the effect of two different types of descriptor network. The top-k exact match accuracy of precursors set prediction for each model was tabulated in 3. Two source element-wise ablation models outperformed the conventional global aggregated model from top-1 to top-10, suggesting that the superior model performance was derived from the usage of element-wise descriptors to predict the set of precursors. However, the model performance did not depend on the usage of the global factor. This means that using the concept of source element-wise prediction is the important feature to predict the inorganic retrosynthetic reactions.

C Experimental details

C.1 Implementation

For curating inorganic database, we used $Pymatgen$[38] library, which is an open-source Python library for materials analysis. The model was constructed using $Pytorch$[39], the deep learning libraries. All experiments were conducted under the machine, which has an Intel Core i9-12900K @ 3.20 GHz, 128 GB of RAM, and NVIDIA GeForce RTX 3090 GPU.

C.2 Training conditions

Train and validation dataset were used to train the two sequentially-connected models (*ElemwiseRetro* and the synthetic temperature prediction model). The learning rate for both models was 3e-4, weight decay coefficient was 1e-6, and the batch size was 128. The cross entropy and robust $L1$ loss functions were used to train these two models, respectively. The weight parameters of the best validation loss during the training process (within 50 epoch) were used as the optimized model parameters. The training curve of the trainset and the validation set for two models were shown in Figure 7.

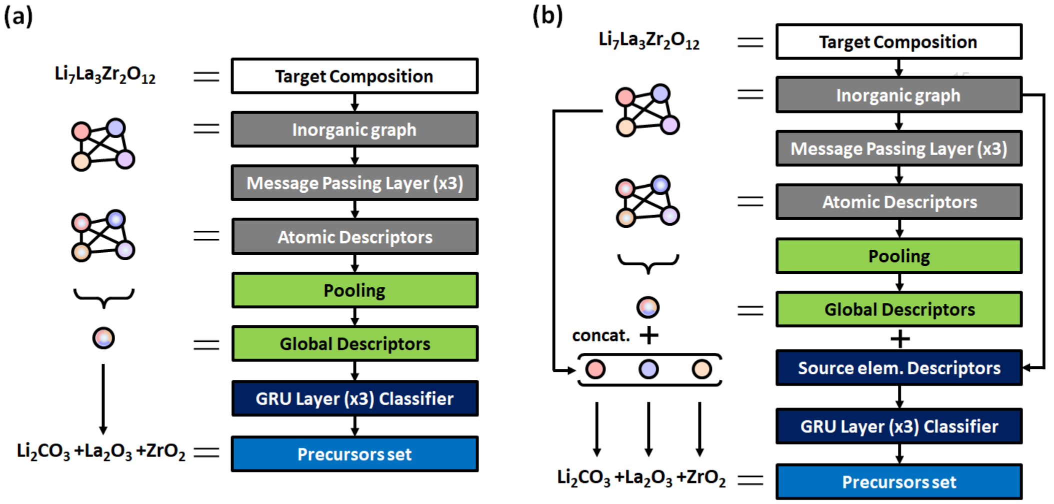

Figure 6: Schematic diagram of the two retrosynthetic ablation model architectures for (a) the conventional global aggregated prediction using pooling operation and for (b) the source element-wise with global aggregated prediction model.

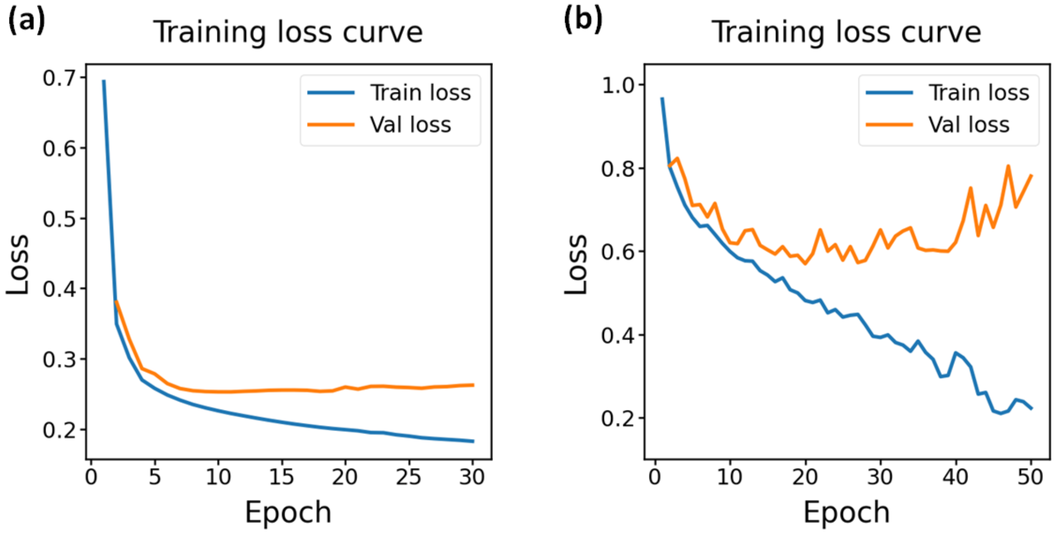

Figure 7: Training and validation loss curve of (a) the retrosynthetic model (*ElemwiseRetro*) and (b) the synthetic temperature prediction model during the training process.