Electronic Supplementary Information (ESI):

Predicting Synthesis Recipes of Inorganic Crystal Materials using Elementwise Template Formulation

Seongmin Kim,[a] Juhwan Noh,[a] Geun Ho Gu,[b] Shuan Chen,[a] and Yousung Jung*[a,c]

a. Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon, 34141, South Korea.

b. School of Energy Technology, Korea Institute of Energy Technology, 200 Hyuksin-ro, Naju, 58330, South Korea.

c. School of Chemical and Biological Engineering, Institute of Chemical Processes, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea.

*Corresponding author e-mail: yousung.jung@snu.ac.kr

Table of Contents

Supplementary Text

Data distribution

Fig. S1 shows the population of the curated dataset. Except artificial elements (Tc, Po, several Period-7 elements, etc.), every element for comprising the inorganic compositions could be treated by *ElemwiseRetro* as shown in Fig. S1a. Furthermore, we sorted the precursor templates that frequently appear in the dataset, which shows the biased distribution particularly on oxides(-O2) and carbonates(-CO3) (Fig. S1b). Fig. S2 shows the data distribution split by publication-year-test.

Ablation study

To elucidate the crucial components of the model, we constructed ablation models. The schematic architectures for the ablation models were illustrated in Fig. S3. To investigate the pooling effect that combines the updated atomic node features to one global descriptor, pooling layer was added to *ElemwiseRetro* which denoted as the global aggregated prediction model (Global agg.), as shown in Fig. S3a. 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 (Source elem-wise w. GLA), as shown in Fig. S3b. 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 Table S1. Two source element-wise ablation models outperformed the conventional global aggregated model from top-1 to top-5, 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.

Experimental Details

For curating inorganic database, we used *Pymatgen*[1] library, which is an open-source Python library for materials analysis. The model was constructed using *Pytorch*,[2] 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.

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-loss function of the individual atomic node was used to train the *ElemwiseRetro*. For training the synthetic temperature prediction model, robust L1 loss function was used, which has been suggested to be more robust than the conventional L1 loss by considering the aleatoric variance as well as the predictive values (heating temperature in this work).[3] 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 Fig. S4.

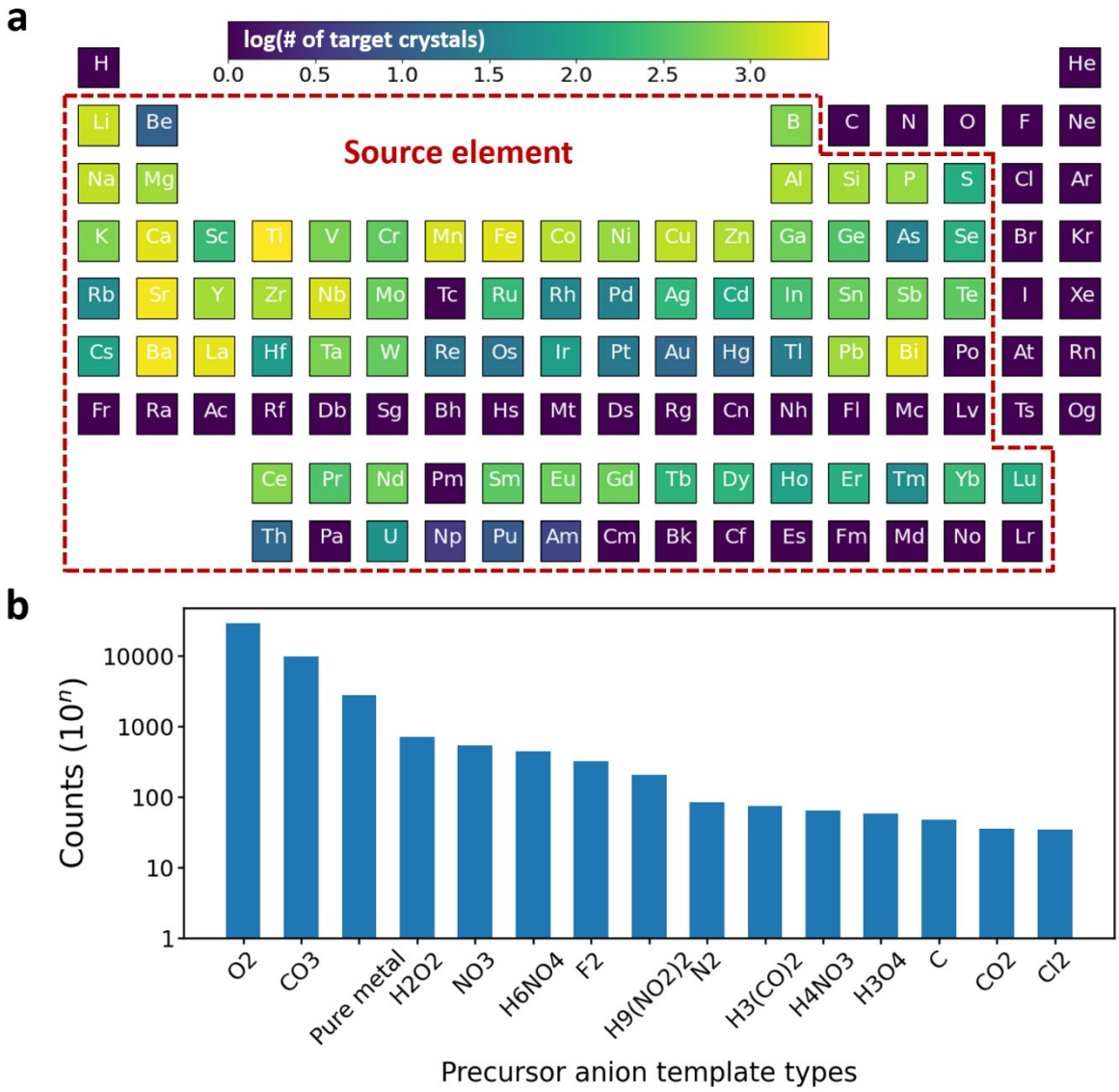

Figure S1. (a) By the definition of the formulated source elements, the element distribution in the periodic table for making up inorganic target material compositions are shown as the colored map counted by log-scaled numbers. (B) The precursor anion template distribution for the frequent types (up to the 15th) among the total 60 extracted list are shown.

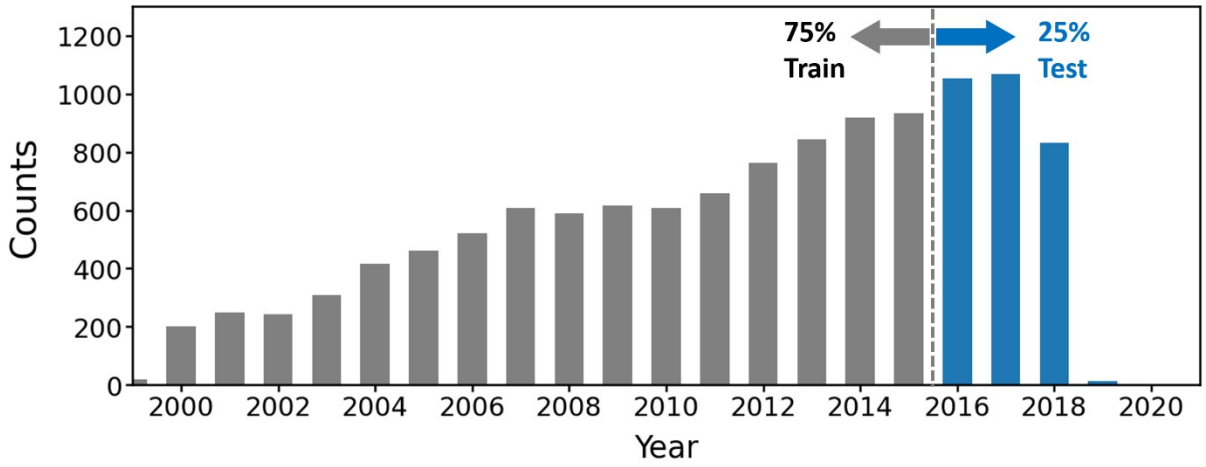

Figure S2. Inorganic reaction dataset counted by published years are shown. After training the model only using the dataset before 2016, the dataset after 2016 were tested to validate the model transferability up to the afterward time space.

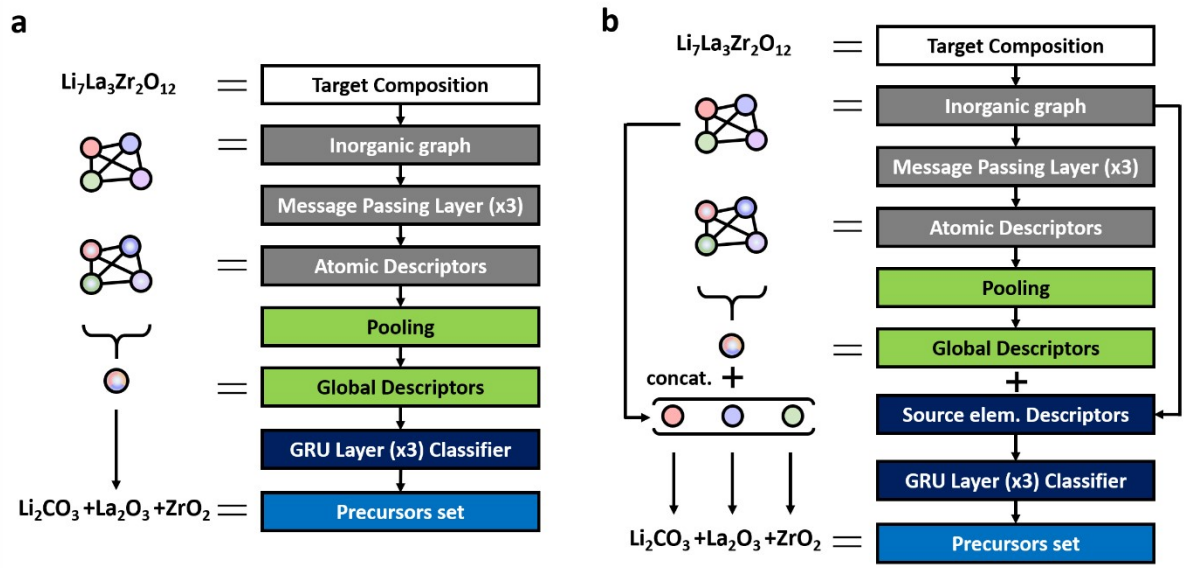

Figure S3. Schematic diagram of the two retrosynthetic ablation model architectures for (a) the conventional global aggregated prediction using pooling layer and for (b) the source element-wise with global aggregated prediction model.

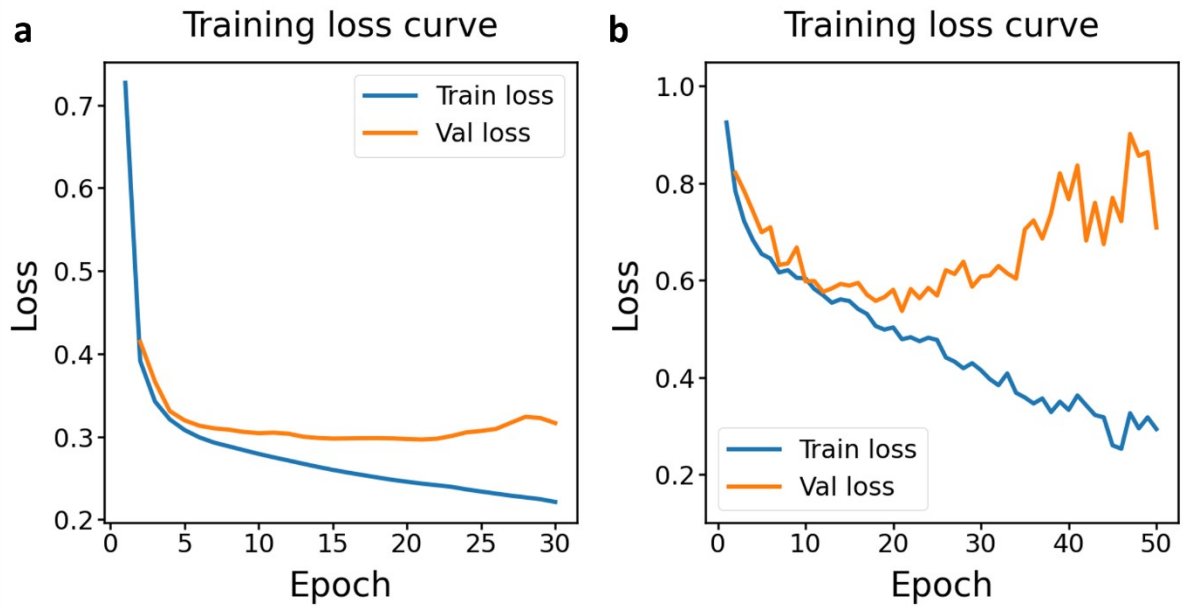

Figure S4. Training and validation loss curve of (a) the retrosynthetic model (*ElemwiseRetro*) and (b) the synthetic temperature prediction model during the training process.

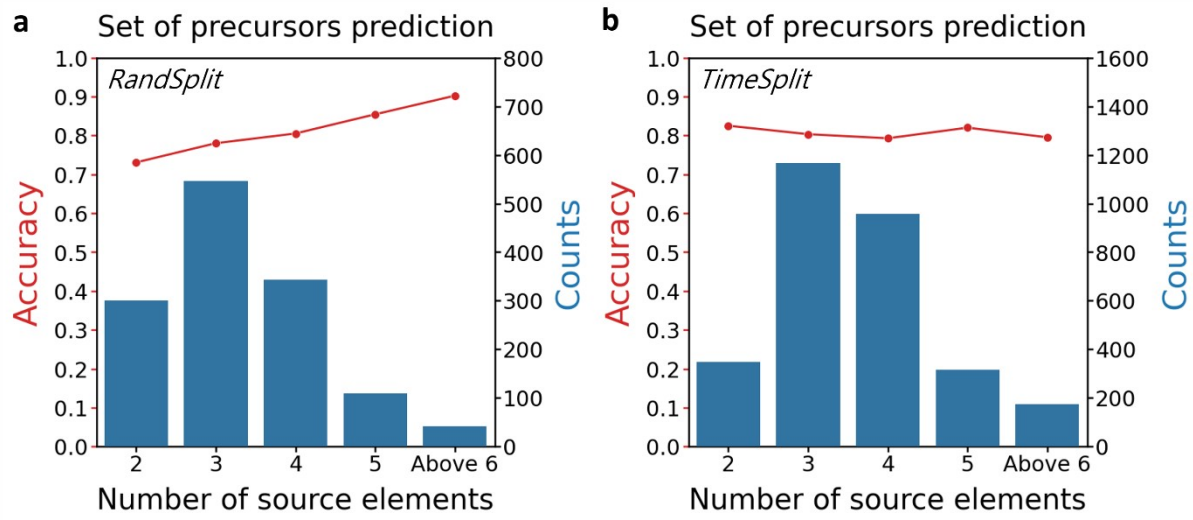

Figure S5. *Top-1* accuracy depending on the number of source elements contained in target compositions for (a) the randomly split and (b) the publication-year-split test dataset, along with each histogram. (blue bar)

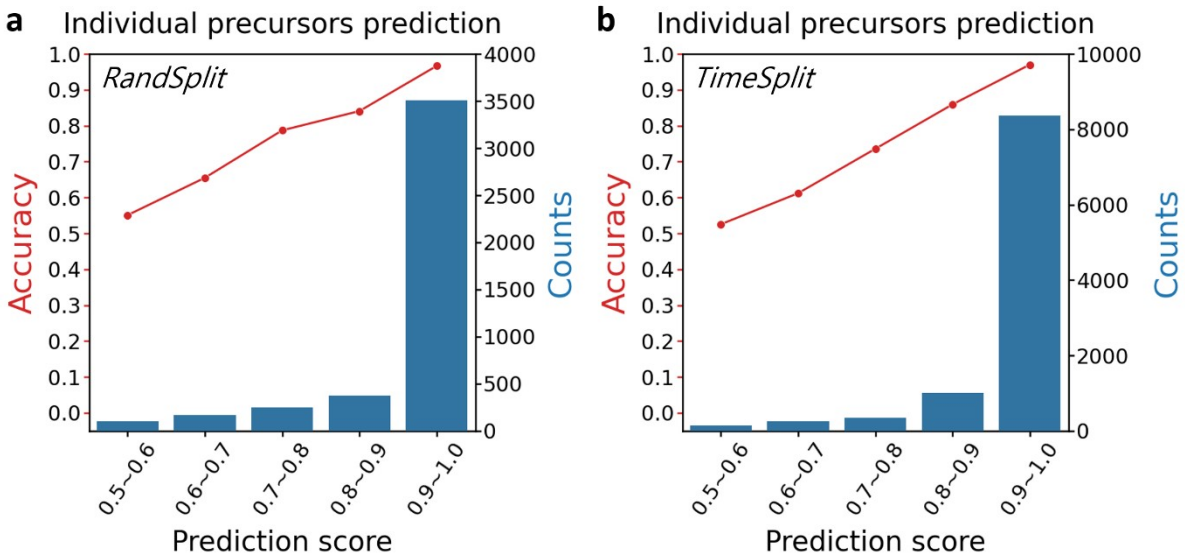

Figure S6. The prediction accuracy (red marked line) of individual precursors as a function of model individual prediction scores for (a) the randomly split and (b) the publication-year-split test dataset, along with each histogram. (blue bar)

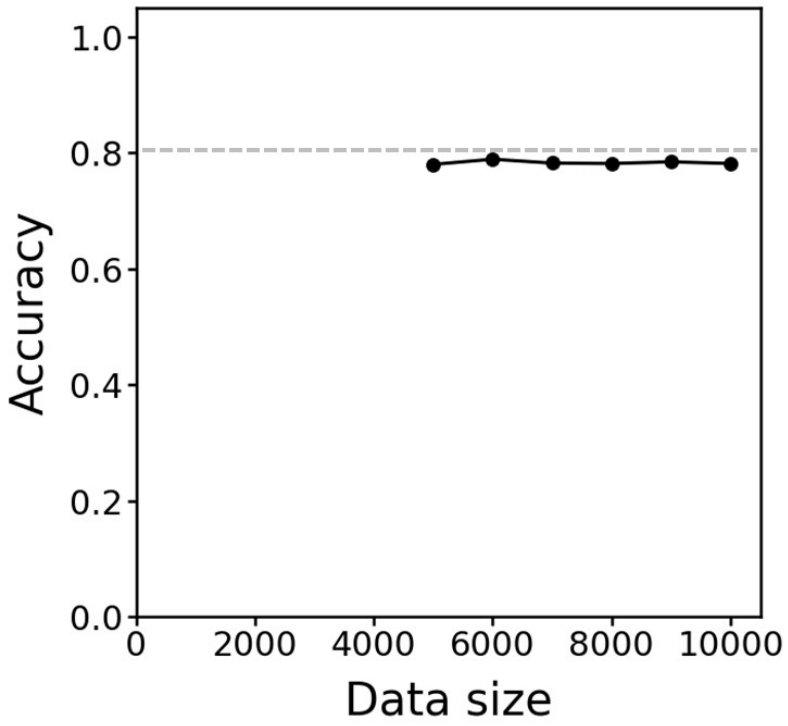

Figure S7. The *top-1* exact match accuracy depending on the training data size.

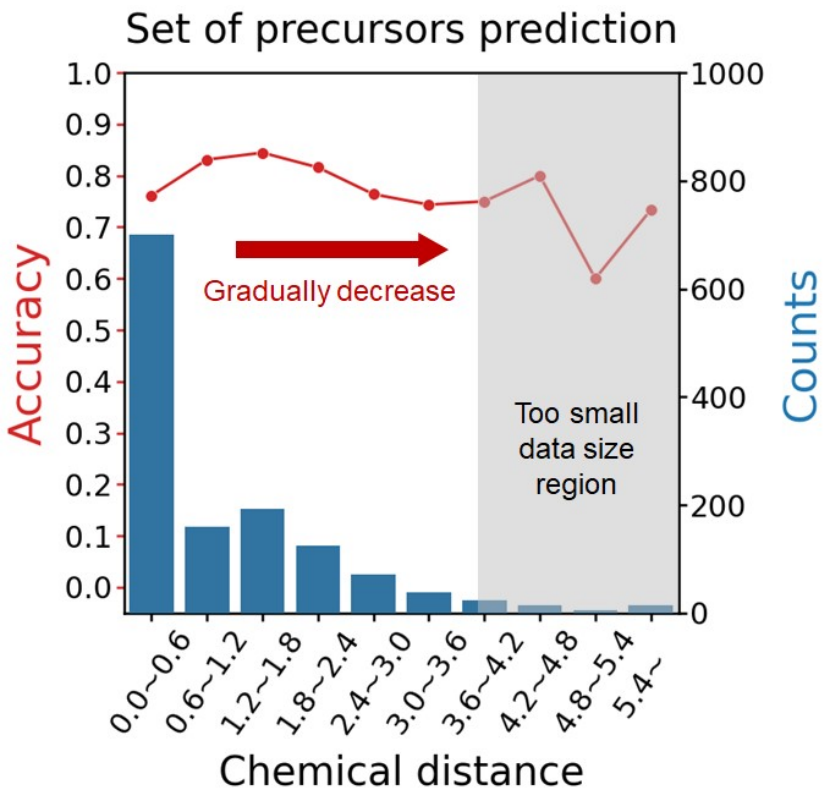

Figure S8. The *top-1* prediction accuracy vs. chemical distance plot with their histograms. The chemical distance was measured by combining the 1st-neighboring Euclidean distance of atomic descriptors from the training data.

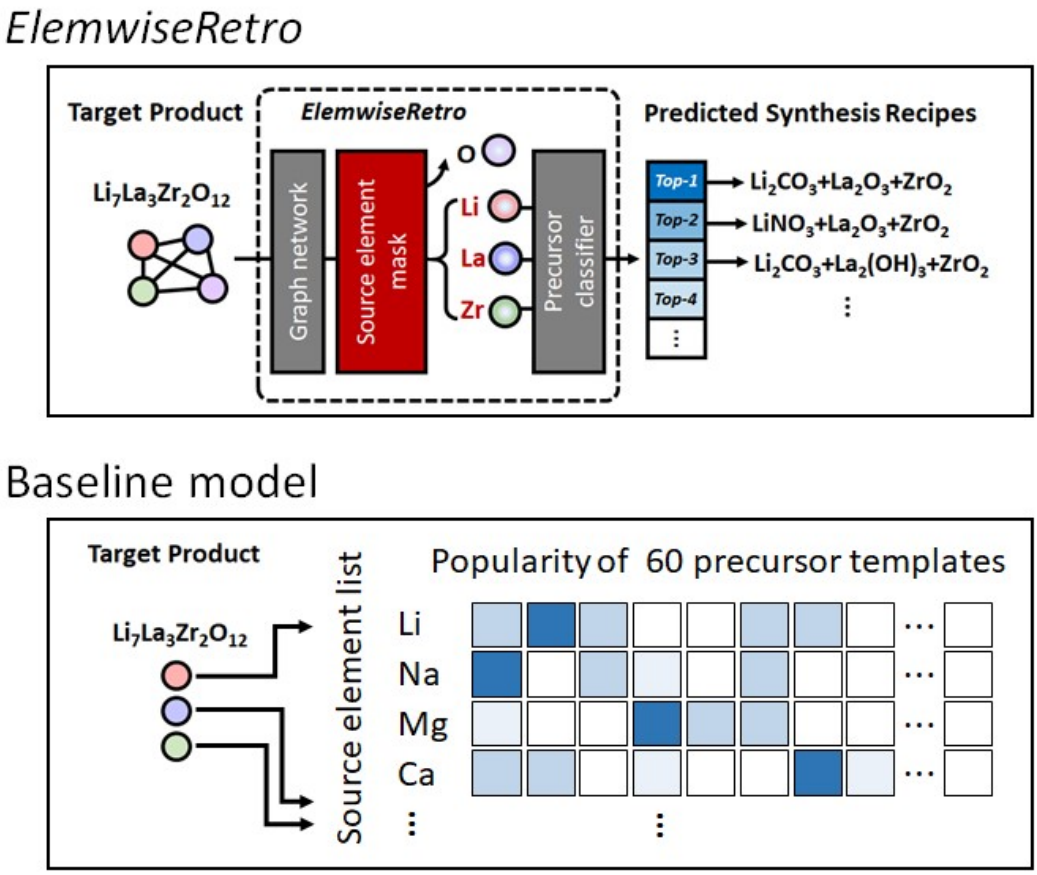

Figure S9. Schematic diagram of the *ElemwiseRetro* and the baseline model architectures. *ElemwiseRetro* combines the compositions of given target materials through the graph neural network and predicts the precursors. However, the baseline only produces outcomes by looking at individual trends for the elements of the target compositions.

Table S1. The *top-k* exact match accuracy for the prediction of inorganic synthesis precursors by three retrosynthetic ablation models.

Top-k accuracy	Ablation model		
	Source Elemwise (%)	*Source Elemwise* w. GA. (%)	GA. (%)
k = 1	78.5	78.4	64.5
k = 2	87.7	87.3	72.5
k = 3	92.1	92.2	80.3
k = 4	94.0	93.6	84.4
k = 5	95.0	95.0	88.5

REFERENCES

1. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Computational Materials Science*, 2013, **68**, 314-319.
2. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, 2017.
3. R. E. Goodall and A. A. Lee, *Nature communications*, 2020, **11**, 6280.