Science Advances
AAAS

Supplementary Materials for

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

Tanjin He *et al.*

Corresponding author: Gerbrand Ceder, gceder@berkeley.edu

Sci. Adv. **9**, eadg8180 (2023)
DOI: 10.1126/sciadv.adg8180

This PDF file includes:

Supplementary Text

Model training and validation

To train the PrecursorSelector encoding model, 44,736/2,254/2,934 synthesis reactions were used as the training/validation/test set as discussed in Section "Data preparation". Each reaction consists of a target material and precursor materials as extracted from the literature. For the purpose of training and validation, a random subset of precursor materials is selected to be replaced with a placeholder "[MASK]" (*21*) in each reaction, referred to as the masked reaction. Because a combinatorial number of masked reactions can be generated from the same reaction, the sampling space of masked reactions is much larger than that of original reactions. To sample as many different masked reactions during the training phase, we employ a dynamic masking strategy (*74*) that randomly samples a batch of reactions and re-generates the masking pattern in every training step. Different from the training samples, the validation samples are generated using static masking during data preprocessing because keeping the validation set unchanged is necessary for model selection afterwards. In this work, we trained with a batch size of 8 masked reactions for 500,000 steps, or 50 epochs with 10,000 steps per epoch. The optimizer used was Adam (*75*) with learning rate of 5×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Starting from the 2,254 original reactions in the validation set, we applied the masking procedure and randomly sampled 3,320 masked reactions for validation. The optimal model selected was the one with minimal loss on the validation samples to minimize overfitting (*76*).

Two tasks are implemented in the representation learning model: (1) the masked precursor completion (MPC) task that predicts the complete precursor set based on the target material and the synthesis context provided by the unmasked precursors, and (2) the composition recovery task that predicts the chemical composition of the target material from the encoded target vector. The loss function in the MPC task, denoted as L_1, is the circle loss (*71*) to maximize the within-class similarity and minimize the between-class similarity in multi-label classification. Here, the within-class similarity corresponds to the similarity of precursor materials present in the same reaction, while the between-class similarity corresponds to the similarity between used and unused precursor materials. The loss function in the composition recovery task, denoted as L_2, is the mean squared error (MSE) loss to compare the difference between predicted composition and the real composition of the target material. The total loss, denoted as L_{multi}, is an adaptive multi-task loss (*73*) to automatically weigh L_1 and L_2 with

$$L_{multi} = \frac{1}{\sigma_1^2} L_1 + \frac{1}{\sigma_2^2} L_2 + \log \sigma_1^2 + \log \sigma_2^2 \quad \text{(S1)}$$

where σ_1 and σ_2 are the model's observation noise parameters which are learned alongside other model parameters. The training and validation losses for each task and the total are shown in Fig. S1. The training loss is averaged every 500 training steps to estimate performance on a substantial number of training samples, which in this study is 4,000. The validation loss is evaluated before training and at the end of each training epoch. As the training loss continues to decrease, the validation loss initially decreases and then increases. The minimal total validation loss is achieved at the end of 15th epoch, leading to the optimal model. The final performance of the optimal model is tested by predicting precursors for 2,654 unseen target materials in the test set. Our similarity-based recommendation strategy successfully reproduces a known precursor set 82% of the time in five attempts or less.

Computation time for similarity evaluation

In this work, all 24,034 materials in the knowledge base are converted to 32-dimensional vectors in advance, forming a 24,034 × 32 matrix. For the 2,654 test materials, we monitored the time required to vectorize them one by one and calculate their cosine similarity to the target vectors in the pre-stored matrix. The similarity evaluation took merely 26 seconds for all the test materials (i.e. 0.01 seconds/material) because of the fast matrix multiplication.

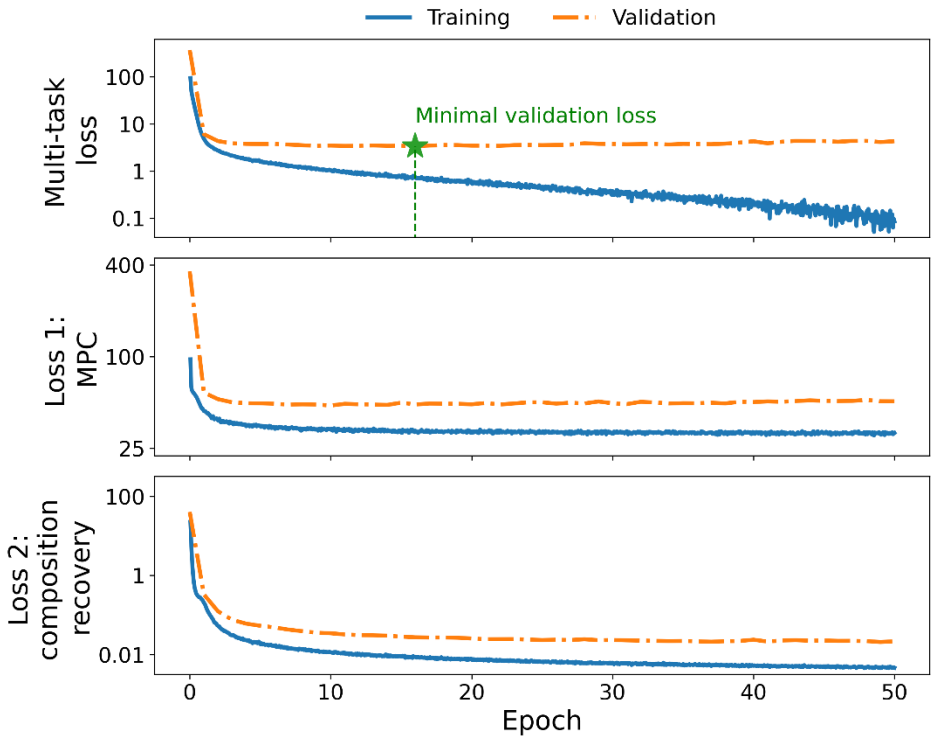

Fig. S1. Evolution of training and validation loss while training the PrecursorSelector encoding model. Top: total multi-task loss, $\boldsymbol{L_{multi}}$. Middle: loss for the MPC task, $\boldsymbol{L_1}$. Bottom: loss for the composition recovery task, $\boldsymbol{L_2}$.

REFERENCES AND NOTES

1. J. C. Hemminger, J. Sarrao, G. Crabtree, G. Flemming, M. Ratner, Challenges at the frontiers of matter and energy: Transformative opportunities for discovery science, *Tech. rep.*, USDOE Office of Science (SC) (United States) (2015).

2. A. Miura, C. J. Bartel, Y. Goto, Y. Mizuguchi, C. Moriyoshi, Y. Kuroiwa, Y. Wang, T. Yaguchi, M. Shirai, M. Nagao, N. C. Rosero-Navarro, K. Tadanaga, G. Ceder, W. Sun, Observing and modeling the sequential pairwise reactions that drive solid-state ceramic synthesis. *Adv. Mater.* **33**, 2100312 (2021).

3. M. Bianchini, J. Wang, R. J. Clément, B. Ouyang, P. Xiao, D. Kitchaev, T. Shi, Y. Zhang, Y. Wang, H. Kim, M. Zhang, J. Bai, F. Wang, W. Sun, G. Ceder, The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nat. Mater.* **19**, 1088–1095 (2020).

4. Z. Jiang, A. Ramanathan, D. P. Shoemaker, In situ identification of kinetic factors that expedite inorganic crystal formation and discovery. *J. Mater. Chem. C* **5**, 5709–5717 (2017).

5. E. J. Corey, Robert Robinson lecture. Retrosynthetic thinking—essentials and examples. *Chem. Soc. Rev.* **17**, 111–133 (1988).

6. A. Stein, S. W. Keller, T. E. Mallouk, Turning down the heat: Design and mechanism in solid-state synthesis. *Science* **259**, 1558–1564 (1993).

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

8. J. R. Chamorro, T. M. McQueen, Progress toward solid state synthesis by design. *Acc. Chem. Res.* **51**, 2918–2925 (2018).

9. H. Kohlmann, Looking into the black box of solid-state synthesis. *Eur. J. Inorganic Chem.* **2019**, 4174–4180 (2019).

10. H. Schäfer, Preparative solid state chemistry: The present position. *Angew. Chem. Int. Ed. Engl.* **10**, 43–50 (1971).

11. O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).

12. E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).

13. M. C. Swain, J. M. Cole, Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).

14. A. M. Hiszpanski, B. Gallagher, K. Chellappan, P. Li, S. Liu, H. Kim, J. Han, B. Kailkhura, D. J. Buttler, T. Y.-J. Han, Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *J. Chem. Inf. Model.* **60**, 2876–2887 (2020).

15. M. Aykol, J. H. Montoya, J. Hummelshøj, Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).

16. M. J. McDermott, S. S. Dwaraknath, K. A. Persson, A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat. Commun.* **12**, 3097 (2021).

17. E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, E. Olivetti, Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).

18. H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain, G. Ceder, Machine-learning rationalization andprediction of solid-state synthesis conditions. *Chemistry of Materials* **34**, 7323–7336 (2022).

19. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013).

20. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **26**, 3111–3119 (2013).

21. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018).

22. T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, G. Ceder, Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem. Mater.* **32**, 7861–7873 (2020).

23. X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, J. Schrier, Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).

24. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

25. C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).

26. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).

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

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

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

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

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

32. Z.-y. Mao, Y.-c. Zhu, Q.-n. Fei, D.-j. Wang, Investigation of 515nm green-light emission for full color emission $LaAlo_3$ phosphor with varied valence Eu. *JOL* **131**, 1048–1051 (2011).

33. E. Mendoza-Mendoza, K. P. Padmasree, S. M. Montemayor, A. F. Fuentes, Molten salts synthesis and electrical properties of Sr- and/or mg-doped perovskite-type $LaAlo_3$ powders. *J. Mater. Sci.* **47**, 6076–6085 (2012).

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

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

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

37. E. Zvereva, O. Savelieva, Y. D. Titov, M. Evstigneeva, V. Nalbandyan, C. Kao, J.-Y. Lin, I. Presniakov, A. Sobolev, S. Ibragimov, M. Abdel-Hafiez, Yu. Krupskaya, C. Jähne, G. Tan, R. Klingeler, B. Büchner, A. N. Vasiliev, A new layered triangular antiferromagnet Li_4FeSbO_6: Spin order, field-induced transitions and anomalous critical behavior. *Dalton Trans.* **42**, 1550–1566 (2013).

38. J. Luan, L. Zhang, K. Ma, Y. Li, Z. Zou, Preparation and property characterization of new Y_2FeSbO_7 and In_2FeSbO_7 photocatalysts. *Solid State Sci.* **13**, 185–194 (2011).

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

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

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

42. F. Lalère, V. Seznec, M. Courty, J. Chotard, C. Masquelier, Coupled X-ray diffraction and electrochemical studies of the mixed Ti/V-containing NASICON: $Na_2TiV(PO_4)_3$. *J. Mater. Chem. A* **6**, 6654–6659 (2018).

43. P. Feng, W. Wang, K. Wang, S. Cheng, K. Jiang, $Na_3V_2(PO_4)_3$/C synthesized by a facile solid-phase method assisted with agarose as a high-performance cathode for sodium-ion batteries. *J. Mater. Chem. A* **5**, 10261–10268 (2017).

44. Y. Fang, L. Xiao, X. Ai, Y. Cao, H. Yang, Hierarchical carbon framework wrapped $Na_3V_2(PO_4)_3$ as a superior high-rate and extended lifespan cathode for sodium-ion batteries. *Adv. Mater.* **27**, 5895–5900 (2015).

45. B. Wang, X. Li, Q. Zeng, G. Yang, J. Luo, X. He, Y. Chen, Efficiently enhanced photoluminescence in Eu^{3+}-doped $Lu_2(MoO_4)_3$ by Gd^{3+} substituting. *Mater. Res. Bull.* **100**, 97–101 (2018).

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

47. T. Yasunaga, M. Kobayashi, K. Hongo, K. Fujii, S. Yamamoto, R. Maezono, M. Yashima, M. Mitsuishi, H. Kato, M. Kakihana, Synthesis of $Ba_{1-x}Sr_xYSi_2O_5N$ and discussion based on structure analysis and dft calculation. *J. Solid State Chem.* **276**, 266–271 (2019).

48. Y. Kitagawa, J. Ueda, M. G. Brik, S. Tanabe, Intense hypersensitive luminescence of Eu^{3+}-doped $YSiO_2N$ oxynitride with near-uv excitation. *Opt. Mater.* **83**, 111–117 (2018).

49. M. Markina, K. Zakharov, E. Ovchenkov, P. Berdonosov, V. Dolgikh, E. Kuznetsova, A. Olenev, S. Klimin, M. Kashchenko, I. Budkin, *et al.*, Interplay of rare-earth and transition-metal subsystems in $Cu_3Yb(SeO_3)_2O_2Cl$. *Phys. Rev. B* **96**, 134422 (2017).

50. D. Zhang, H. Berger, R. K. Kremer, D. Wulferding, P. Lemmens, M. Johnsson, Synthesis, crystal structure, and magnetic properties of the copper selenite chloride $Cu_5(SeO_3)_4Cl_2$. *Inorg. Chem.* **49**, 9683–9688 (2010).

51. H. Zhuang, Y. Bao, Y. Nie, Y. Qian, Y. Deng, G. Chen, Synergistic effect of composite carbon source and simple pre-calcining process on significantly enhanced electrochemical performance of porous $LiFe_{0.5}Mn_{0.5}PO_4$/C agglomerations. *Electrochim. Acta* **314**, 102–114 (2019).

52. L. Wang, Y. Li, J. Wu, F. Liang, K. Zhang, R. Xu, H. Wan, Y. Dai, Y. Yao, Synthesis mechanism and characterization of $LiMn_{0.5}Fe_{0.5}PO_4$/C composite cathode material for lithium-ion batteries. *J. Alloys Compd.* **839**, 155653 (2020).

53. Q.-Q. Zou, G.-N. Zhu, Y.-Y. Xia, Preparation of carbon-coated $LiFe_{0.2}Mn_{0.8}PO_4$ cathode material and its application in a novel battery with $Li_4Ti_5O_{12}$ anode. *J. Power Sources* **206**, 222–229 (2012).

54. H. Yi, C. Hu, H. Fang, B. Yang, Y. Yao, W. Ma, Y. Dai, Optimized electrochemical performance of $LiMn_{0.9}Fe_{0.1-x}Mg_xPO_4$/C for lithium ion batteries. *Electrochim. Acta* **56**, 4052–4057 (2011).

55. G. Heymann, E. Selb, M. Kogler, T. Götsch, E.-M. Köck, S. Penner, M. Tribus, O. Janka, $Li_3Co_{1.06(1)}TeO_6$: Synthesis, single-crystal structure and physical properties of a new tellurate

compound with Co^{II}/Co^{III} mixed valence and orthogonally oriented Li-ion channels. *Dalton Trans.* **46**, 12663–12674 (2017).

56. J. S. Ndzila, S. Liu, G. Jing, J. Wu, L. Saruchera, S. Wang, Z. Ye, Regulation of Fe^{3+}-doped $Sr_4Al_6SO_{16}$ crystalline structure. *J. Solid State Chem.* **288**, 121415 (2020).

57. N. G. Dorbakov, V. V. Titkov, S. Y. Stefanovich, O. V. Baryshnikova, V. A. Morozov, A. A. Belik, B. I. Lazoryak, Barium-induced effects on structure and properties of β-$Ca_3(PO_4)_2$-type $Ca_9Bi(VO_4)_7$. *J. Alloys Compd.* **793**, 56–64 (2019).

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

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

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

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

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

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

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

65. H. Peng, X. Luan, L. Li, Y. Zhang, Y. Zou, Synthesis and ion conductivity of $Li_7La_3Nb_2O_{13}$ Ceramics with cubic garnet-type structure. *J. Electrochem. Soc.* **164**, A1192–A1194 (2017).

66. L. van Wüllen, T. Echelmeyer, H.-W. Meyer, D. Wilmer, The mechanism of li-ion transport in the garnet $Li_5La_3Nb_2O_{12}$. *Phys. Chem. Chem. Phys.* **9**, 3298–3303 (2007).

67. K. Park, H. Lim, S. Park, G. Deressa, J. Kim, Strong blue absorption of green Zn_2SiO_4: Mn^{2+} phosphor by doping heavy Mn^{2+} concentrations. *Chem. Phys. Lett.* **636**, 141–145 (2015).

68. C. Chen, Y. Zhuang, D. Tu, X. Wang, C. Pan, R.-J. Xie, Creating visible-to-near-infrared mechanoluminescence in mixed-anion compounds $SrZn_2S_2O$ and SrZnSO. *Nano Energy* **68**, 104329 (2020).

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

70. J. Thirumalai, R. Krishnan, I. Shameem Banu, R. Chandramohan, Controlled synthesis, formation mechanism and lumincence properties of novel 3-dimensional $Gd_2(MoO_4)_3$:Eu^{3+} nanostructures. *J. Mater. Sci. Mater. Electron.* **24**, 253–259 (2013).

71. R. Alcantara, J. Jumas, P. Lavela, J. Olivier-Fourcade, C. Pérez-Vicente, J. Tirado, X-ray diffraction, ^{57}Fe Mössbauer and step potential electrochemical spectroscopy study of $LiFe_yCo_{1-y}O_2$ compounds. *J. Power Sources* **81–82**, 547–553 (1999).

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

73. I. Radosavljevic, J. A. Howard, A. W. Sleight, J. S. Evans, Synthesis and structure of $Bi_3Ca_9V_{11}O_{41}$. *J. Mater. Chem.* **10**, 2091–2095 (2000).

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

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

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