High-efficient low-cost characterization of materials properties using domain-knowledge-guided self-supervised learning

Binglin Xie
South China University of Technology
Xianhua Yao
South China University of Technology
Weining Mao
South China University of Technology
Mohammad Rafiei
Johns Hopkins University
Nan Hu (nanhu026@scut.edu.cn)
South China University of Technology https://orcid.org/0000-0003-3536-8414

Article

Keywords:

Posted Date: January 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1241474/v1

nature
machine intelligence

ARTICLES
https://doi.org/xx.xxxx/xxxxxx-021-xxxxx-x

High-efficient low-cost characterization of materials properties using domain-knowledge-guided self-supervised learning

Binglin Xie[1], Xianhua Yao[1], Weining Mao[1], Mohammad H. Rafiei[2,3]✉, and Nan Hu[1]✉

Modern AI-assisted approaches have helped material scientists revolutionize their abilities to better understand the properties of materials. However, current machine learning (ML) models would perform awful for materials with a lengthy production window and a complex testing procedure because only a limited amount of data can be produced to feed the model. Here, we introduce self-supervised learning (SSL) to address the issue of lacking labeled data in material characterization. We propose a generalized SSL-based framework with domain knowledge and demonstrate its robustness to predict the properties of a candidate material with the fewest data. Our numerical results show that the performance of the proposed SSL model can match the commonly-used supervised learning (SL) model with only 5 % of data, and the SSL model is also proven with ease of implementation. Our study paves the way to expand further the usability of ML tools for a broader material science community.

For material scientists, the most time-consuming part of research is nothing but preparing a specimen that could take weeks or longer. After a sample is fabricated, it might take another week or two to assess its properties. Then, you return to your lab and prepare another sample for your next test. The worst part is that you must repeat this cycle again and again to get enough data to be more confident of the accurate material properties. To shorten this never-ending cycle, scientists and engineers have integrated ML models (such as SL) to accelerate the characterization of a given material and/or the reverse design of material microstructure and composition[1-19]. However, the need for a large amount of accurate annotated training data in SL has created a headache for material scientists. They need to conduct sufficient experiments to serve an SL model or "label" them from the ML perspective. This roadblock hampered the usability of ML in material characterization, especially targeting materials with scarce labeled data and difficult to characterize.

In contrast, as a new paradigm between unsupervised and supervised learning, SSL can generate labels based on the property of unlabeled data itself to train the neural network in a supervised manner similar to natural learning experiences. With excellent performance on representation learning and dealing with the issue of unlabelled data, SSL[20-22] has been successfully implemented in a wide range of fields, including image recognition[23], audio representation[24], computer vision[25], document reconstruction[26], atmosphere[27], astronomy[28], medical[29], person re-identification[30], remote sensing[31], robotics[32], omnidirectional imaging[33], manufacturing[34], nano-photonics[35], and civil engineering[36], etc. However, this method has not been formally attempted in material science.

Here, we investigate the reliability and robustness of SSL as a data-efficient technique for characterizing material properties. We chose concrete as a composite trial material because it is the most consumed man-made material worldwide by mass[37], takes at least a month to test its full strength, and has various influencing factors on its properties. Yet, our generalized framework is targeted to apply for a wide range of organic or inorganic materials as long as it is difficult to fabricate and characterize, although only concrete is discussed here. Fig. 1a illustrates the generalized procedure of the material characterization for a given material. With a basic understanding of its constituents, material scientists would use experiments, formulas, or models to estimate material properties. Among three routes, experiments are empirical and labor-intensive; formulas are theoretical and difficult to obtain; numerical models are efficient but highly depend on prior experience. Regardless of the accuracy of each route, scientists and engineers could gain domain knowledge and further fine-tune the characterization process on a given material. In this work, we ultilized a online database and selected a total of 425 experimental datasets[38] on concrete as the ground truth (see Methods for details). Each dataset consists of seven constitutes in concrete mix and its corresponding concrete compressive strength (CCS) after 28 days of curing. Using mix design as inputs and estimating CCS as the output has been implemented[39-42] in ML-based tools (commonly by SL models), but our goal is to showcase the power of SSL using scarce data against its SL counterpart.

Motivated by this goal, we selected an easy-to-train backpropagation neural network model with a simple architecture (see Methods for details). We did compare key parameters of ML models and eventually chose an optimal combination to avoid the issue of overfitting and implement for

[1] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China. [2] Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA [3] Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA. ✉email: mrafiei1@jhu.edu. ✉email: nanhu026@scut.edu.cn

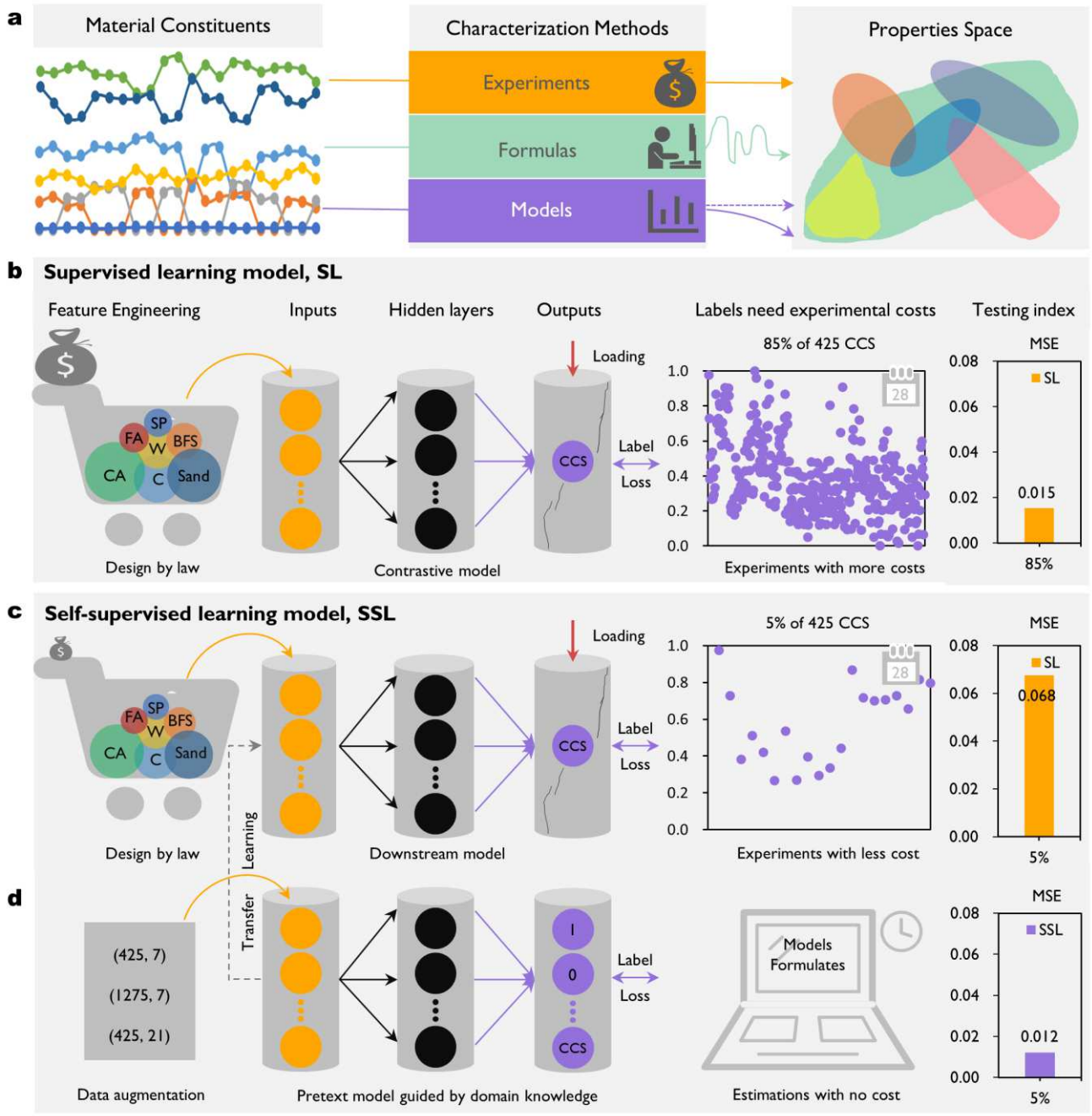

Fig. 1 | General concept and procedure of self-supervised learning for material characterization. a, common characterization approaches for a given material. **b**, model set-up, data distribution and property prediction of a composite material (concrete) using supervised learning, SL. **c**-**d**, model set-up, data distribution and property prediction of a composite material (concrete) using self-supervised learning, SSL. Downstream model in **c** has the same set-up like SL model while pretext model in **d** has a separate pre-trained model using pseudo-data created by domain knowledge.

further studies. In Fig. 1b-d, we compare procedures and predictions by an SL and an SSL model, respectively. Fig. 1b indicates that the SL model with 85% (361 sets of data) of original dataset for training and validation can provide a good CCS prediction where the mean squared error, MSE, is equal to 0.015. In Fig. 1c, the so-called downstream model as part of the SSL model has an identical architecture to the benchmark SL model yet uses only 5% (21 sets of data) of original dataset. Without further treatments, the prediction by this downstream model is unacceptable, which the MSE (0.068) is nearly five times higher than the one by the benchmark SL model, showing that model performance has a high dependence on the size of training data. The major steps of using an SSL model (Fig. 1c-d) include producing pseudo data by data processing and/or domain knowledge, training a pretext model, and transferring learning to the downstream model. Thanks to this procedure, the calculated MSE (0.012) in the SSL model is as low as the MSE of the SL model yet uses 94% less training data (21 vs. 361).

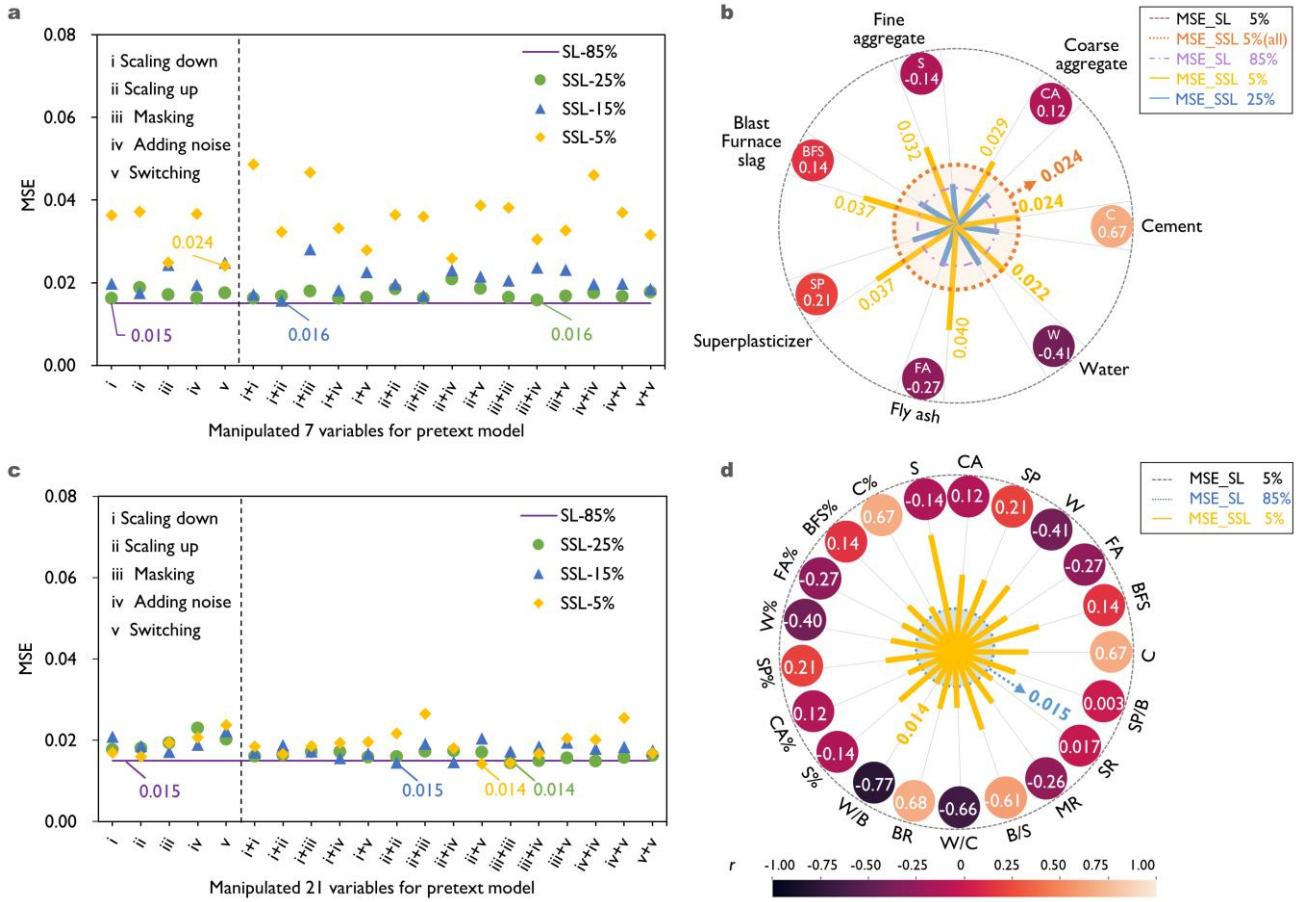

Fig. 2 | SSL model performance with various data processing approaches and domain knowledge of inputs. a, the effect of various augmentation methods on predicting CCS. **b**, the effect of manipulating each variable of concrete on the model performance under three training ratios (with three colour bars, yellow, blue and green). **c**, the model performance by various augmentation methods coupled with additional domain knowledge. **d**, the effect of 21 variables of concrete material on the model performance. Yellow bars refer to MSE values by manipulating seven initial weight value respectively; blue bars refer to seven weight proportion value; green bars refer to seven additional physical values related to concrete mix design.

Having proven the feasibility of using the SSL model to predict CCS significantly reducing training data, we started to establish a generalized SSL-based framework for material scientists without much prior experience on machine learning. For a specific material, the key step of using SSL is to generate pseudo data from the initial unlabeled data in a semi-automated manner during the pretext task. In our case, we proposed several ways to manipulate the original data (Supplementary Fig. S1). Recall that the dimension of the original dataset is [425, 7], referring to 425 entries of mix designs on concrete including seven variables in each entry. First, we generate pseudo-input data through various manipulation approaches such that the dataset dimension can be enlarged to [425×n, 7] in pretext task. Then, in the downstream task, we chose a simple neral network structure (7-7-1) based on exisiting literature to test the model performance. Fig. 2a shows prediction results in terms of MSE values by 20 argumentation methods under three training ratios. On the x-axis, the first five indicate the original dataset is doubled (n=2) under a single manipulation approach, while the next 15 pairs are under dual argumentations whose original dataset is tripled (n=3). The accuracy of the SSL model with 25% training data (cyan dots) can match the one by the benchmark SL model using 85% training data (straight purple line), while we observed a decrease in accuracy by the SSL model with fewer data. When only 5% of initial data was trained, the SSL model (yellow dots) showed deviated scatter points among 20 methods. This result suggests that the combined argumentation methods won't increase the model accuracy, yet using a triplet dataset (scaling down and up the initial data) could be a good start.

After increasing input data points and evaluating the model performance by a purely mathematic approach, we continued to focus on these inputs from a physical perspective. Given that all seven variables in the simulations above (Fig. 2a) were altered together, we wondered how manipulating each concrete variable might change the model performance. In Fig. 2b, we compare MSE values between single variable manipulation and combined manipulation along with the benchmark SL model. It can be seen that seven MSE values under 5% training data (yellow bars) are significantly varied and mostly larger than the one by manipulating all variables together (the orange dash ring) except for cement and water. We then analyzed the correlation coefficient *r* on each variable. We found that cement and water

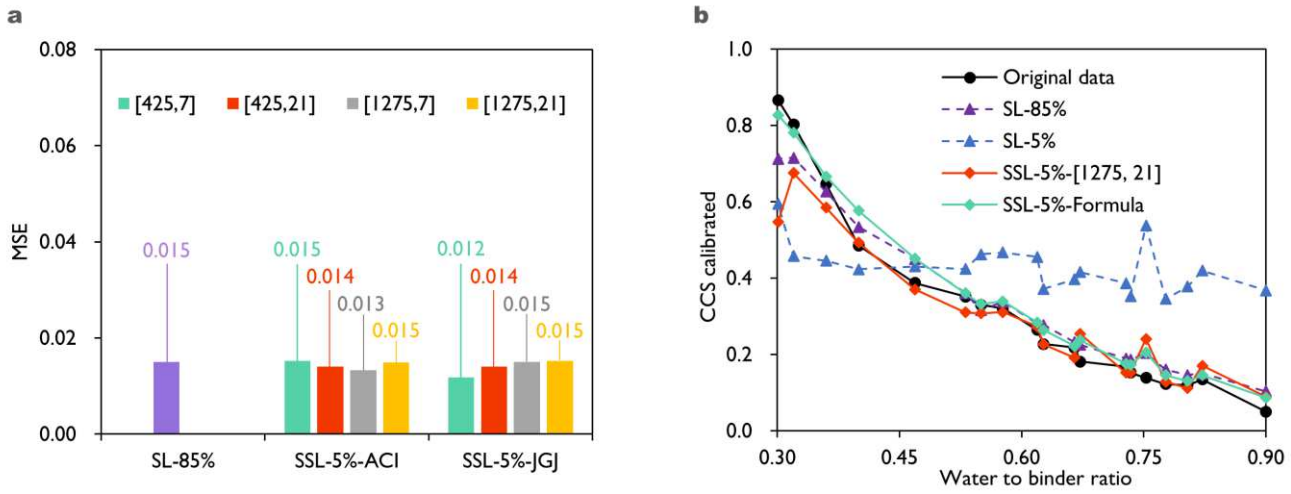

Fig. 3 | SSL model performance with pseudo-output as domain knowledge. a, the effect of output manipulation along with input manipulation. **b**, CCS predictions under various water to binder ratios produced by various SSL models and the benchmark SL model. ACI refers to the US code for structural concrete and JGJ refers to Chinese code.

have the maximum positive *r* (0.67) and negative *r* (-0.41), respectively, indicating a strong connection between these two variables and CCS. When the traing ratio become larger (25%), such observation vanished. Our SSL model somehow identified physical correlations without any prior knowledge of concrete.

Yet, the proposed framework still has tremendous upside in terms of accuracy. We then proposed a total of 14 co-occurring variables related to seven initial inputs (Supplementary Fig. S2) to further increase the data dimension to [425×n, 7+m], see Methods for details. Since the size of input variable increased, we adjusted the neural network architecture in the downstream model and identified an optimal structure for further study (Supplementary Table S1-S2 and Fig. S3-S4). Following the same evaluation in Fig. 2a, we compared the model performance under an expanded data set of [850, 21] and [1275, 21]. Fig. 2c reveals that the overall performance of SSL models exhibited an improved learning capability even with the fewest data possible (5%). In Fig 2d, we manipulated every 21 inputs individually and calculated their *r*. As we expected, the model performed best when manipulating the water-to-binder ratio with the maximum absolute *r* (-0.77). Again, this result (Supplementary Table S3-S4) confirmed that the SSL model might extract the hidden correlation between variables and properties, but further studies are needed to validate this observation.

Based on the confidence gained in manipulating model inputs, we searched for additional low-cost domain knowledge related to model output for the pretext task. Usually, a specific property of a given material can be estimated through a classic theoretical equation, an empirical equation by other researchers, and/or a numerical simulation. In this study, we estimated CCS by an empirical equation and generated a total of 425 pseudo-output for each dataset, see Methods for details. Coupled with the input manipulations, in Fig. 3a, we compare four scenarios under two additional pseudo-outputs. As expected, the SSL model performed better with additional domain knowledge. Variation of MSE values was observed for models using either 7 or 21 variables, but overall, these SSL model has a lower MSE than the one predicted by the benchmark SL model. In Fig. 3b, we plot CCS predictions of 19 data with respect to various water-to-binder ratios and demonstrate the merit of the SSL model by either input or output manipulation (Supplementary Table S5).

In a final demonstration, we proved the robustness and reliability of our SSL-based framework on predicting properties for a different concrete dataset (Supplementary Fig. S5). This dataset has two extra features: fewer data points (103 vs. 425) and two new properties (slump height and flow diameter). With prior experience, we easily designed four pretexting models to manipulate input and/or output to predict a new property (slump height). In Fig 4a-b, we compare MSE values and slump height predictions by the SSL model and found that even using 30% data for training both SL and SSL model has lower accuracy. We believe this issue is associated with fewer data points and complex correlation for slump height (Supplementary Fig. S6 and Table S6). We then went back to CCS prediction and found that with the augmented data dimension of [103, 21] plus the pseudo-output predicted by the empirical equation in ACI, the SSL model can match the benchmark SL model using 65% lesser data. Interesting, when we used slump height as a new input, the SSL model performed even better in some scenarios (Supplementary Fig. S7 and Table S7).

In summary, we develop a robust SSL-based framework to boost the characterization of certain materials that require a lengthy fabrication process. The limited experimental database has impeded the usability of SL models that heavily rely on many training datasets. Instead, the proposed framework is proved to reduce the cost of sample annotation while retaining powerful learning capabilities using minimal data possible. Our SSL model provides a manageable procedure for generating pseudo-data from the initial data and proves its robustness and reliability against the commonly used SL model. We envision this generalized framework that can be applied to a wide range of materials with various levels of domain knowledge. However, we only verified the concept for concrete yet saved at least a 94% decrease in experimental costs. It should note that SSL has been

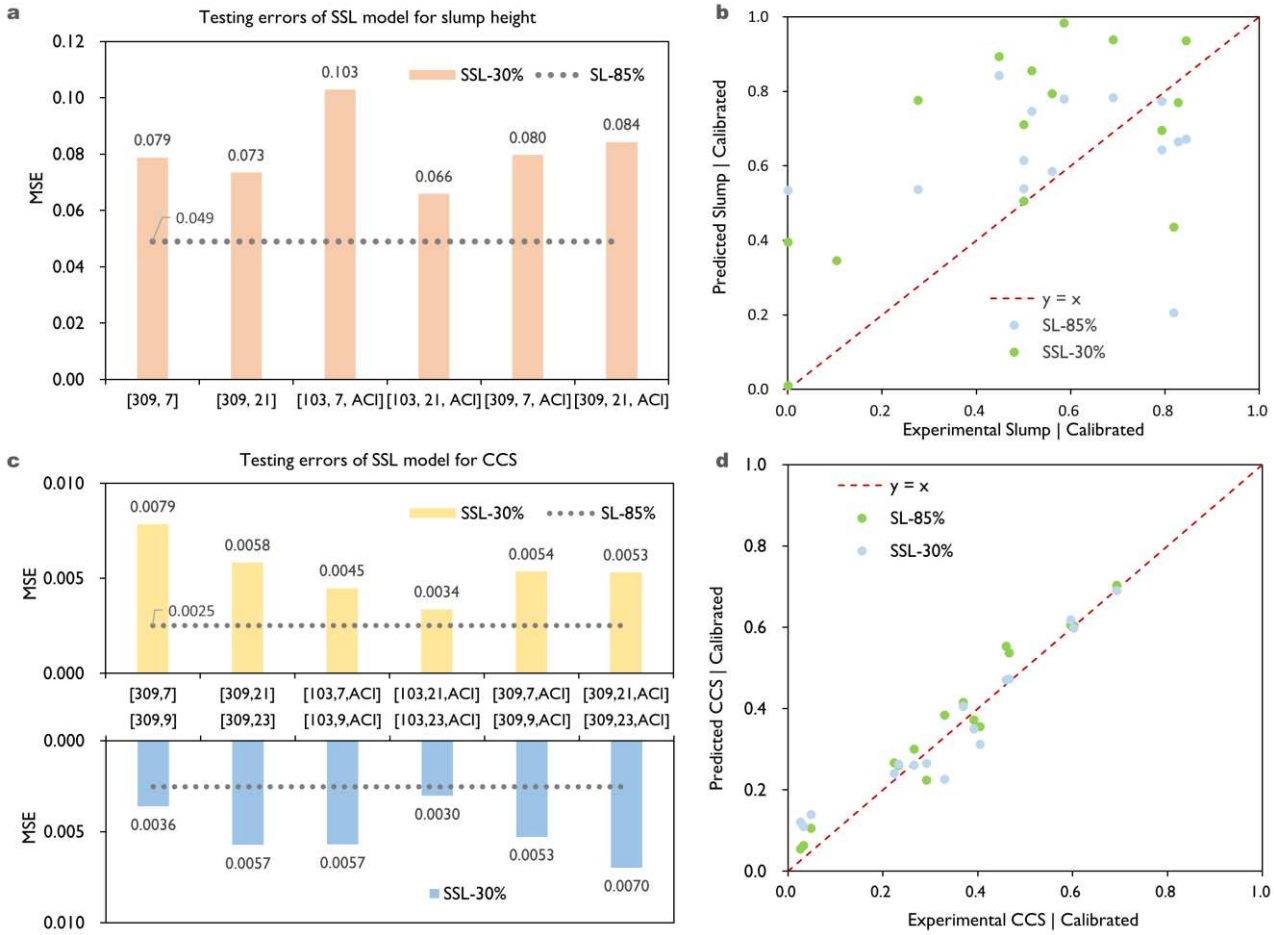

Fig. 4 | The implementation of the proposed SSL framework using a new set of material data. a, model performance using various argumentation methods for predicting slump height of concrete (103 data points). **b**, Slump height predictions between the best SSL model (30%) and the benchmark SL model (85%). **c**, model performance with proposed SSL model with and without using slump and flow as new inputs. **d**, CCS predictions between the best SSL model (30%) and the benchmark SL model (85%).

successfully implemented in fields such as recognition and natural language processing, but predicting material properties appears to be a more complex task. Extensive works can be done, such as developing a rapid search of suitable argumentation approaches for a given material, coupling with deep learning[43] to use complex model architecture for better performance, and complementing pseudo-label generation with active learning[44]. Overall, we recommend this framework to material scientists as a novel and powerful characterization tool and hopefully partially free you from extensive lab routines.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgments, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at XXX

References

1 Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. Big-deep-smart data in imaging for guiding materials design. *Nat Mater* **14**, 973-980, doi:10.1038/nmat4395 (2015).

2 Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. A bridge for accelerating materials by design. *npj Computational Materials* **1**, doi:10.1038/npjcompumats.2015.8 (2015).

3 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547-555, doi:10.1038/s41586-018-0337-2 (2018).

4 Fish, J., Wagner, G. J. & Keten, S. Mesoscopic and multiscale modelling in materials. *Nat Mater* **20**, 774-786, doi:10.1038/s41563-020-00913-0 (2021).

5 Dijkstra, M. & Luijten, E. From predictive modelling to machine learning and reverse engineering of colloidal self-assembly. *Nat Mater* **20**, 762-773, doi:10.1038/s41563-021-01014-2 (2021).

6 Friederich, P., Hase, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat Mater* **20**, 750-761, doi:10.1038/s41563-020-0777-6 (2021).

7 Chen, C.-T. & Gu, G. X. Learning hidden elasticity with deep neural networks. *Proceedings of the National Academy of Sciences* **118**, e2102721118, doi:10.1073/pnas.2102721118 (2021).

8 Daehn, K. *et al.* Innovations to decarbonize materials industries. *Nature Reviews Materials*, doi:10.1038/s41578-021-00376-y (2021).

9 Guo, K. & Buehler, M. J. A semi-supervised approach to architected materials design using graph neural networks. *Extreme Mechanics Letters* **41**, 101029, doi:https://doi.org/10.1016/j.eml.2020.101029 (2020).

10 Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Materials Horizons* **8**, 1153-1172, doi:10.1039/D0MH01451F (2021).

11 Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational Materials* **3**, 53, doi:10.1038/s41524-017-0055-6 (2017).

12 Kim, Y. *et al.* Deep learning framework for material design space exploration using active transfer learning and data augmentation. *npj Computational Materials* **7**, 140, doi:10.1038/s41524-021-00609-2 (2021).

13 Liu, H. *et al.* Predicting the early-stage creep dynamics of gels from their static structure by machine learning. *Acta Materialia* **210**, 116817, doi:https://doi.org/10.1016/j.actamat.2021.116817 (2021).

14 Naser, M. Z. An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. *Automation in Construction* **129**, 103821, doi:https://doi.org/10.1016/j.autcon.2021.103821 (2021).

15 Naser, M. Z. & Alavi, A. H. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Architecture, Structures and Construction*, doi:10.1007/s44150-021-00015-8 (2021).

16 Vahabi, H., Naser, M. Z. & Saeb, M. R. Fire Protection and Materials Flammability Control by Artificial Intelligence. *Fire Technology*, doi:10.1007/s10694-021-01200-3 (2022).

17 Yang, Z., Yu, C.-H. & Buehler, M. J. Deep learning model to predict complex stress and strain fields in hierarchical composites. *Science Advances* **7**, eabd7416, doi:doi:10.1126/sciadv.abd7416 (2021).

18 Zheng, B., Zheng, Z. & Gu, G. X. Uncertainty quantification and prediction for mechanical properties of graphene aerogels via Gaussian process metamodels. *Nano Futures* **5**, 045004, doi:10.1088/2399-1984/ac3c8f (2021).

19 Ma, C. *et al.* Accelerated design and characterization of non-uniform cellular materials via a machine-learning based framework. *npj Computational Materials* **6**, 40, doi:10.1038/s41524-020-0309-6 (2020).

20 Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc Natl Acad Sci U S A* **117**, 30046-30054, doi:10.1073/pnas.1907367117 (2020).

21 Wei, C. *et al.* Self-Supervised Representation Learning for Evolutionary Neural Architecture Search. *IEEE Computational Intelligence Magazine* **16**, 33-49, doi:10.1109/MCI.2021.3084415 (2021).

22 Liu, X. *et al.* Self-supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 1-1, doi:10.1109/tkde.2021.3090866 (2021).

23 Ohri, K. & Kumar, M. Review on self-supervised image recognition using deep neural networks. *Knowledge-Based Systems* **224**, doi:10.1016/j.knosys.2021.107090 (2021).

24 Carr, A. N., Berthet, Q., Blondel, M., Teboul, O. & Zeghidour, N. Self-Supervised Learning of Audio Representations From Permutations With Differentiable Ranking. *IEEE Signal Processing Letters* **28**, 708-712, doi:10.1109/lsp.2021.3067635 (2021).

25 Venator, M., Himer, Y. E., Aklanoglu, S., Bruns, E. & Maier, A. Self-Supervised Learning of Domain-Invariant Local Features for Robust Visual Localization Under Challenging Conditions. *IEEE Robotics and Automation Letters* **6**, 2753-2760, doi:10.1109/lra.2021.3059571 (2021).

26 Paixão, T. M. *et al.* Self-supervised deep reconstruction of mixed strip-shredded text documents. *Pattern Recognition* **107**, doi:10.1016/j.patcog.2020.107535 (2020).

27 Campbell, C. L. *et al.* A self-supervised learning based approach to analyze Martian water–ice cloud properties for planetary atmospheric applications. *Acta Astronautica* **181**, 1-13, doi:10.1016/j.actaastro.2020.12.041 (2021).

28 Hayat, M. A., Stein, G., Harrington, P., Lukić, Z. & Mustafa, M. Self-supervised Representation Learning for Astronomical Images. *The Astrophysical Journal Letters* **911**, doi:10.3847/2041-8213/abf2c7 (2021).

29 Zhao, Q., Liu, Z., Adeli, E. & Pohl, K. M. Longitudinal self-supervised learning. *Med Image Anal* **71**, 102051, doi:10.1016/j.media.2021.102051 (2021).

30 Jiang, K., Zhang, T., Zhang, Y., Wu, F. & Rui, Y. Self-supervised Agent Learning for Unsupervised Cross-Domain Person Re-identification. *IEEE Trans Image Process* **PP**, doi:10.1109/TIP.2020.3016869 (2020).

31 Li, W., Chen, H. & Shi, Z. Semantic Segmentation of Remote Sensing Images With Self-Supervised Multitask Representation Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 6438-6450, doi:10.1109/jstars.2021.3090418 (2021).

32 Sun, P., Yang, Z., Zhang, T., Guo, S. & Chen, F. Primitive-contrastive network: data-efficient self-supervised learning from robot demonstration videos. *Applied Intelligence*, doi:10.1007/s10489-021-02527-8 (2021).

33 Liu, M., Wang, S., Guo, Y., He, Y. & Xue, H. Pano-SfMLearner: Self-Supervised Multi-Task Learning of Depth and Semantics in Panoramic Videos. *IEEE Signal Processing Letters* **28**, 832-836, doi:10.1109/lsp.2021.3073627 (2021).

34 Wang, T., Qiao, M., Zhang, M., Yang, Y. & Snoussi, H. Data-driven prognostic method based on self-supervised learning approaches for fault detection. *Journal of Intelligent Manufacturing* **31**, 1611-1619, doi:10.1007/s10845-018-1431-x (2018).

35 Ma, W. & Liu, Y. A data-efficient self-supervised deep learning model for design and characterization of nanophotonic structures. *Science China Physics, Mechanics & Astronomy* **63**, doi:10.1007/s11433-020-1575-2 (2020).

36 Zhang, K., Zhang, Y. & Cheng, H. D. Self-Supervised Structure Learning for Crack Detection Based on Cycle-Consistent Generative Adversarial Networks. *Journal of Computing in Civil Engineering* **34**, doi:10.1061/(asce)cp.1943-5487.0000883 (2020).

37 Habert, G. *et al.* Environmental impacts and decarbonization strategies in the cement and concrete industries. *Nature Reviews Earth & Environment* **1**, 559-573, doi:10.1038/s43017-020-0093-3 (2020).

38 Yeh, I. C. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research* **28**, 1797-1808, doi:https://doi.org/10.1016/S0008-8846(98)00165-3 (1998).

39 Ben Chaabene, W., Flah, M. & Nehdi, M. L. Machine learning prediction of mechanical properties of concrete: Critical review. *Construction and Building Materials* **260**, doi:10.1016/j.conbuildmat.2020.119889 (2020).

40 Rafiei, M. H., Khushefati, W. H., Demirboga, R. & Adeli, H. Supervised Deep Restricted Boltzmann Machine for Estimation of Concrete. *ACI Materials Journal* **114** (2017).

41 Rafiei, M. H., Khushefati, W. H., Demirboga, R. & Adeli, H. Neural Network, Machine Learning, and Evolutionary Approaches for Concrete Material Characterization. *ACI Materials Journal* **113** (2016).

42 Ouyang, B. *et al.* Using machine learning to predict concrete's strength: learning from small datasets. *Engineering Research Express* **3**, 015022, doi:10.1088/2631-8695/abe344 (2021).

43 Wang, S.-H., Pillai, H. S., Wang, S., Achenie, L. E. K. & Xin, H. Infusing theory into deep learning for interpretable reactivity prediction. *Nature Communications* **12**, 5288, doi:10.1038/s41467-021-25639-8 (2021).

44 Ren, P. *et al.* A Survey of Deep Active Learning. *ACM Computing Surveys* **54**, 1-40, doi:10.1145/3472291 (2022).

Methods

Data preparation. The original datasets for training and testing our model were obtained from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). For SSL model validation, 1030 raw datasets of concrete were collected. Each dataset includes seven material constituents (Content of cement in kg/m^3, denoted as C; Content of blast furnace slag in kg/m^3, denoted as BFS; Content of fly ash in kg/m^3, denoted as FA; Content of water in kg/m^3, denoted as W; Content of superplasticizer in kg/m^3, denoted as SP; Content of coarse aggregates in kg/m^3, denoted as CA; and content of fine aggregates in kg/m^3, denoted as S) and a material property of the compressive strength of experimental concrete specimens at curing 28-day (denoted as CCS, MPa). We selected 425 out of 1030 samples for using CCS after 28-day. Ideally, a new material other than concrete is desirable for model implementation, but we cannot find an appropriate dataset in UCI Machine Learning Repository. Thus, a second concrete data set was used with 103 raw data. In addition to the seven inputs as the first dataset also provide concrete slump height (cm) and flow diameter (cm), and the amount of material constituents are significantly different. Although two datasets contain 425 and 103 test results, we randomly divided them into training, validation, and testing sets under each training ratio. For example, when 85% of the initial data were used to train the benchmark SL model and the downstream model of SSL, while the remaining 15% of initial data were used to test the performance of both SL and SSL models. For example, in the model validation part of this work, 361 out of 425 data were used for the SL model, which has two subsets for training (90%, 325 data) and validation (10%, 36 data). For the SSL model using 5% initial data, these two subsets were reduced to 19 and 2. In the model implementation, 30% of 103 data were used for training, i.e., 31 data including 28 for training and 3 for validation. It should be noted that such data amount has been pushed to the extreme to simulate materials with challenging fabrication and characterization procedures. To identify a threshold that can match the performance of the benchmark SL model, we evaluated other training ratios (10%, 15%, 20%, 25%, and 30%) for the SSL model. All data has been calibrated with Min-Max Normalization into the range [0, 1].

Model architecture. We built a classic Back Propagation Neural Network (NN) model with the Adaptive Moment Estimation learning algorithm based on TensorFlow in Python. The NN architecture comprises inputs, hidden layers, and output. The size of inputs is 7 or 21 variables of the concrete material, while the output is the CCS after casting 28 days. Architecture comparison of the number of hidden layers and neurons are given in Supplementary information. Optimal model architectures are 7-7-1 and 21-56-56-56-1. The ReLU is the activation function for inputs and hidden layers to avoid gradient vanishing and fast train, and the Linear function acts on the output layer. We initialize the network with all weights set to small random values and biases to zero. The mean-square-error (MSE) cost function evaluates training, validation, and testing loss for the benchmark SL and downstream models in SSL. The categorical cross-entropy function evaluates accuracy for the pretext model in SSL. We also adopt the Early Stopping callback to avoid overfitting. Multiple simulations were run to find a generalized model with optimal parameters for the study.

Input manipulation in pretext task. We thought domain knowledge could help the pretext model self-supervised learning more information since the powerful NN model learns what you feed. In this work, the original input data has a dimension of [425, 7], referring to 425 entries and seven variables in each entry. To self-supervised learning more data, we attempted to enlarge them to [425×n, 7+m]. For n, five simple data processing methods were used, including random (i) scaling down, (ii) scaling up, (iii) masking, (iv) adding noise, and (v) switching to generate negative samples for pretext model of SSL. For methods (i) and (ii), inputs were randomly scaled down or up and labeled by a factor (such as 0.25, 0.5, 1.25, or 1.5). For method (iii), we randomly masked inputs by one and zero. For method (iv), we added a noise conforming to the Gaussian distribution to inputs. For method (v), we shuffle the sequence of inputs. When one method is used, the original data can be doubled while multiple methods are combined, the original data can be multiplied n times. Consequently, we could randomly produce many negative and positive samples of original data with the dimension of labels becoming a larger input domain [425×n, 7] to boost the training of the pretext model. Results in Fig. 2 have only considered n=2 and n=3 regardless of manipulating either a single variable or all variables together. Although we also tried to set n up to 9, it did not predict us better. For m, we expand seven initial variables to 14 (m=7) and 21 (m=14) for the pretext model. Note that seven variables in the initial data are the actual weight of each ingredient of a concrete sample. We added another seven variables by considering the weight proportion of each ingredient to total content based on the mass design method of mix proportion of concrete materials. Then, we also found seven extra variables that are known to affect CCS in literature. They are: water to binder ratio, denoted as W/B; The water to cement ratio, denoted as W/C; The percentage of binders denoted as BR; The sand ratio denoted as SR; The binder to sand ratio denoted as B/S; The percentage of mineral admixture denoted as MR; And the percentage of chemical admixture, denoted as SP/B, respectively. It should be noted that these 14 additional variables in the pretext task were calculated entirely based on literature. With further domain

knowledge, more appropriate variables can be proposed, resulting in an even better performance of the pretext model. Thus, the input domain of [425, 7] was augmented to [425, 21]. Coupled with manipulation methods of m and n, the input domain was enlarged up to [425×3, 7+14]. For other materials, m may be limited; as such, n can be enlarged further to complement.

Output manipulation in pretext task. We considered classic formula for the generation of pseudo-output without additional cost to train the pretext model for SSL. It should be noted that all authors have minimal knowledge on concrete without pre-existing or biased preference on formula selection. We selected emprical formula from concrete code in US and China, which both considered water-cement ratio as the key factor on CCS and presents a linear relation. While, two formula are based on different testing standard (ISO by the Chinese code and ASTM by the US code) and Chinese code included more parameters. Therefore, the estimation by two equations are varied.

For ACI 211.1-1991 R2009 standards for mix proportion design procedures in the US, we used the Abrams' formula, a well-known equation that describes the relationship between 28-day CCS and water-cement ratio (W/C). We could fit its table 6.3.4(a) to obtain equation (1).

$$f = 22.212\frac{B}{W} - 12.167 \quad (1)$$

Where f is 28-day CCS (MPa), B is the content of binders, and W is the content of cement.

While Chinese JGJ55-2011 code adopted Bolomys' formula, which was based on Abrams' formula, and developed more influence coefficient for binders, so we could evaluate through a more accurate equation (2).

$$f = \alpha_a f_b \frac{B}{W} - \alpha_a \alpha_b f_b \quad (2)$$

Where, α_a and α_b are the coefficients for kind of coarse aggregate, f_b is the 28-day compressive strength of binders accounts for factors of blast furnace slag, fly ash, and cement.

Data availability

All codes necessary to reproduce results in the main paper are available upon request from the corresponding authors Dr. Nan Hu (nanhu026@scut.edu.cn) and Dr. Mohammad H. Rafiei (mrafiei1@jhu.edu).

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (52008174) and Guangdong Provincial Key Laboratory of Modern Civil Engineering Technology (2021B1212040003). All the sources of support are gratefully acknowledged.

Author contributions

N.H. secured research funding. B.X., M.H.R., and N.H. conceived the project and designed the general concept. M.H.R. initiated the building of SSL model and B.X. further completed the entire framework with the support of X.Y. and W.M. B.X. analysed the data and illustrated all figures in main manuscript and supplementary information. N.H. and M.H.R. supervised the research. All authors wrote and revised the manuscript.

Competing interests

The authors declare no further competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary20220111.pdf