

ARTICLE OPEN



Deep learning framework for material design space exploration using active transfer learning and data augmentation

Yongtae Kim^{1,4}, Youngsoo Kim^{2,4}, Charles Yang³, Kundo Park¹, Grace X. Gu³ and Seunghwa Ryu¹✉

Neural network-based generative models have been actively investigated as an inverse design method for finding novel materials in a vast design space. However, the applicability of conventional generative models is limited because they cannot access data outside the range of training sets. Advanced generative models that were devised to overcome the limitation also suffer from the weak predictive power on the unseen domain. In this study, we propose a deep neural network-based forward design approach that enables an efficient search for superior materials far beyond the domain of the initial training set. This approach compensates for the weak predictive power of neural networks on an unseen domain through gradual updates of the neural network with active transfer learning and data augmentation methods. We demonstrate the potential of our framework with a grid composite optimization problem that has an astronomical number of possible design configurations. Results show that our proposed framework can provide excellent designs close to the global optima, even with the addition of a very small dataset corresponding to less than 0.5% of the initial training dataset size.

npj Computational Materials (2021)7:140; <https://doi.org/10.1038/s41524-021-00609-2>

INTRODUCTION

In order to discover or design novel materials having outstanding properties, significant effort has been paid to devise various material design approaches such as biomimicry, design of experiment methods, and other conventional optimization methods^{1–13}. However, these approaches often require in-depth physics-based analysis of the relationship between materials descriptors and properties. Hence, a fundamental understanding on the underlying physical mechanisms determining the material properties is a primer for the material design. Machine learning models are alternative promising tools for materials design, because they enable design space exploration only with a database representing the relationship between the descriptors of material (inputs) and the properties (outputs). Trained machine learning models can infer the relationship with several orders of magnitude speedup compared to actual data generation from experiments or physics-based simulations tools^{14–22}. In many applications, the machine learning models, such as Gaussian process regression, radial basis function network, support vector machine, and deep neural network (DNN), are adopted as surrogate forward models, which predict the outputs from the corresponding inputs^{23–25}. These models are combined with high-throughput screening and various optimization methods to obtain new materials with targeted properties^{26–29}. However, it requires a lot of effort to find desired materials in vast design space with a forward design approach, because a large number of candidates must be tested to search for the optimal material due to the absence of the gradient of predicted output with respect to input features^{19,30–32}.

In this regard, inverse design methods, which adapt machine learning models as a designer directly suggesting promising candidate materials based on target properties, are being

intensively studied to avoid the aforementioned arduous design space exploration process^{18,33–38}. Autoencoder (AE), variational autoencoder (VAE), and generative adversarial network (GAN) are three of the most commonly used DNN based generative models^{18,36,38}. These methods are very efficient if the target design is located within or close to the seen domain, i.e., the domain of the training dataset^{34,39–42}. However, they cannot generate data in the unseen domain, i.e., the domain outside the ranges of the training dataset defined by the input feature space and output values. Hence, their applicability is limited, because it is computationally infeasible to generate training data large enough to cover the entire high-dimensional design space of most material design problems.³⁵ Moreover, DNN is likely to have lower predictive performance on unseen domains unless the DNN learns a governing equation representing the relationship between the inputs and the outputs which require domain expertise and some heuristics to create an appropriate mathematical formulation^{43–49}.

Thus, in most cases, it is inevitable to rely on the lower predictive power of DNN and repeat the validation of many suggested promising candidates through laborious simulations or experiments until the material with the desired property is obtained³⁵. In addition, even if active learning is adopted to increase the prediction reliability of a DNN in the domain near the target design, a large number of data should be validated and utilized to update the DNN^{35,50}. This is because the proposed data is likely to be positioned far from the target design due to the poor reliability of the DNN originally trained with the initial training dataset. Hence, the design process requires a very large computational cost.

In this study, we propose a systematic neural network-based forward design approach to efficiently search for the desired

¹Department of Mechanical Engineering & KI for the NanoCentury, Korea Advanced Institute of Science and Technology (KAIST), Yuseong-gu, Daejeon, Republic of Korea.

²Department of Nature-Inspired System and Application, Korea Institute of Machinery and Materials, Daejeon, Republic of Korea. ³Department of Mechanical Engineering, University of California, Berkeley, CA, USA. ⁴These authors contributed equally: Yongtae Kim, Youngsoo Kim. ✉email: ryush@kaist.ac.kr

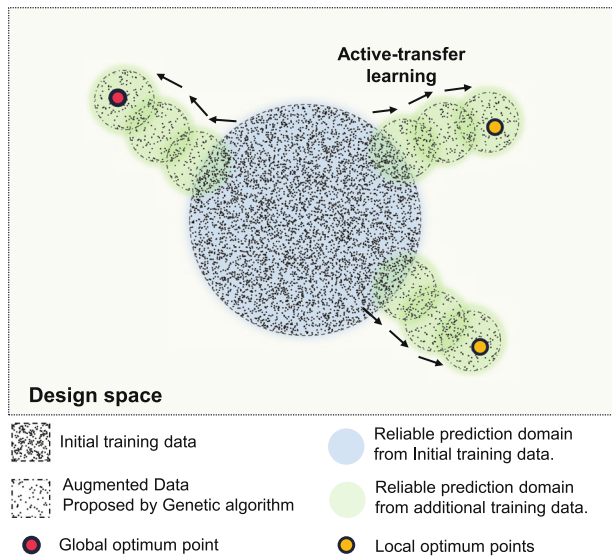


Fig. 1 Schematic of deep learning framework for material design space exploration. Schematic of gradual expansion of reliable prediction domain of DNN based on the addition of data generated from the hyper-heuristic genetic algorithm and active transfer learning.

materials outside the training data domain which overcomes the aforementioned limitations of the existing methods. It is known that DNN trained by gradient descent algorithms with unbounded activation function is capable of making reliable predictions on the unseen domain close to the training dataset with linear approximation^{44,51,52}. Hence, our framework gradually expands the reliable prediction domain of DNN toward the region of desired properties by updating DNN via active transfer learning. Relatively sparse and small additional datasets including materials with incrementally superior properties are iteratively added to the training set based on a data augmentation technique to increase generalization of DNN, i.e., ability to make accurate and stable predictions on unseen data⁵³. The limitation of a forward design approach is mitigated by using a hyper-heuristic genetic algorithm on top of the updated DNN. Our study demonstrates that materials with desired properties can be designed out of inferior original training datasets with small dataset augmentation and validation.

RESULTS

Schematic of the forward design framework

The schematic of our framework is depicted in Fig. 1. DNN trained with the initial training dataset is capable of making a reliable prediction on the design space slightly larger than the training data domain, as represented in the bluish region. To find the materials with desired properties, which are positioned outside the domain of initial training data, DNN should be able to make a reliable prediction on the domain containing the desired design. In this regard, using the trained DNN to predict the properties of new material designs proposed by the genetic algorithm, a relatively small set of materials superior to those in the existing dataset is suggested. Since the newly proposed materials are outside of the current training set and the DNN predictions on them are not accurate, their properties are evaluated again with accurate physics-based simulations (if high-throughput experimental facility is available, one can use experiments). Those data are integrated to the training dataset with a data augmentation technique, and the DNN is updated based on the new training data with active transfer learning as represented with a black

arrow. This process is repeated until the DNN is able to make a reliable prediction in the domain close to the optimal point represented as the large redpoint. The DNN after the last update is used to find the optimal design. A detailed explanation of our framework is provided in the following sections.

Architectures of the deep neural network (DNN)

To leverage the prediction of DNN on the unseen domain, the DNN architecture consists of an unbounded activation function, i.e., leaky rectified linear unit activation function (leaky ReLU) with coefficient 0.1^{46,52} (See the comparison with other activation functions in Supplementary Figure 1). The architectures are constructed based on residual network (Resnet) with full pre-activation, which is known for good generalization performance with a sufficient number of learnable parameters, with batch normalization layer as regularization methods^{54,55} (See the details of DNN architecture in Supplementary Figure 2). We check the predictive performance of the DNN in the seen and unseen domains by setting the randomly chosen 10% of data as validation sets, and the dataset with the highest 10% output values as validation sets, respectively. In the optimization procedure, one has to explore the dataset having output values higher than those of the initial training set. Hence, validation on the dataset with the highest 10% output values for the DNN trained on the dataset with the lowest 90% output values would represent the predictive performance of the DNN in the unseen domain during the optimization procedure. The flowchart representing the process for constructing the DNN architecture is depicted in Supplementary Figure 3.

Prediction results upon seen domain and unseen domain

In this study, we demonstrate the applicability of our framework by solving a representative problem with a large design space - the design of composite microstructures with superior mechanical properties, i.e., stiffness, strength, and toughness, which are close to the global optimum located far beyond the domain of initial training data. The details of data generations are presented in the Methods section and Fig. 2. The training results of DNN regarding stiffness and strength upon seen and unseen domains are represented in Fig. 3. The prediction accuracy gradually decreases as data is located further away from the training data, as reported in the literature^{44,52}. The training results on seen domain generally show better results compared to the results on the unseen domain, as expected. Still, despite the mismatch in the absolute values, the DNN network is able to distinguish relative magnitudes to some extent, as shown in Fig. 3c–d. However, the training results for toughness show that it is infeasible to make predictions on unseen output value range (Supplementary Figure 4). We suspect that the poor predictive performance on toughness originates from the complexity in determining toughness from the entire stress-strain curve encompassing the full failure process involving complex crack propagation and branching process. We expect that this challenge might be overcome with the sequential learning methods that can learn the complex and nonlinear material behavior beyond the onset of yield or failure^{56,57}. However, because it is beyond the scope of this study, we leave the optimization of toughness as future work. In the stiffness and strength training results in an unseen domain, the DNN predictions gradually deviate more as the data is positioned further away from the training dataset.

Material design process

A flowchart of the proposed material design framework is depicted in Fig. 4. DNN tested with randomly selected validation sets are adopted to allow reliable prediction in a broader domain of output values. For the genetic algorithm, 30 microstructures

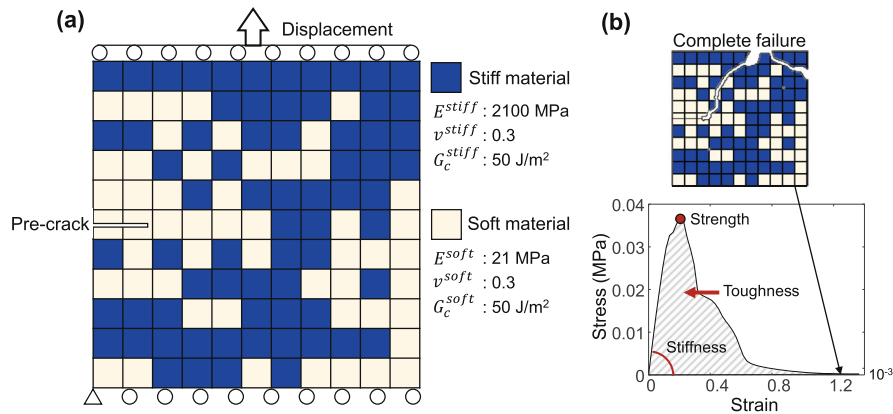


Fig. 2 Schematic of 11 × 11 grid composite data generation. **a** Schematic of 11 × 11 grid composite composed of stiff and soft materials, and **(b)** the process for obtaining mechanical properties of composite based on the stress-strain curves. The tensile deformations are applied until the complete failure of composite, i.e., no load-bearing capacity.

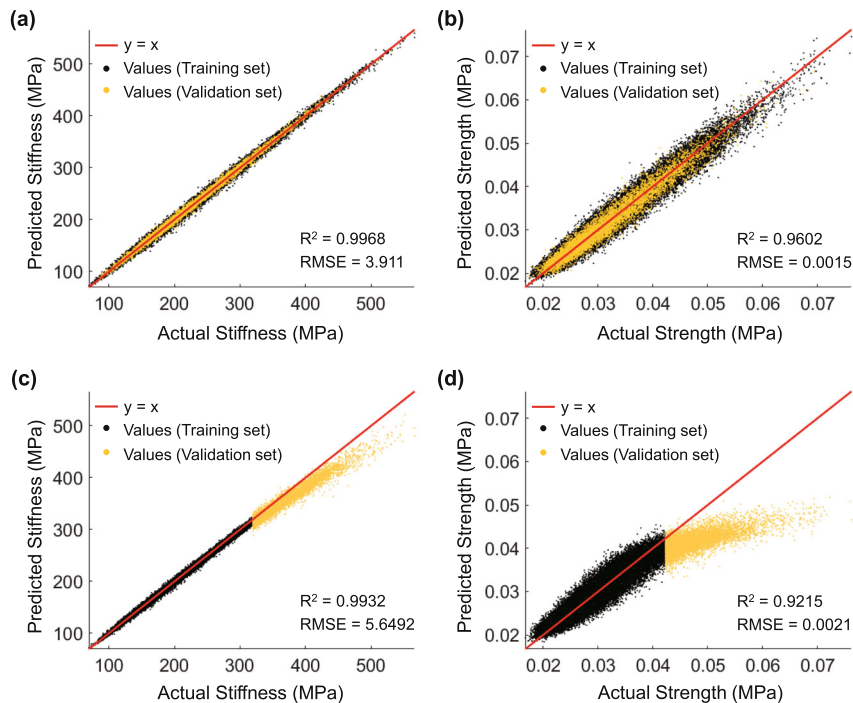


Fig. 3 DNN training results. **a–b** DNN training results in terms of stiffness and strength with the validation sets randomly selected 10% of data in the data distribution. **c–d** DNN training results to check generalization on an unseen domain by setting the training sets and the validation sets as 90% of data with lowest output value and 10% of data having highest output value, respectively. The coefficient of determination (R^2) are calculated by the following equation: $R^2 = 1 - \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - f_i)^2}$. y_i , f_i and \bar{y} represent the actual value, fitted value, and mean of actual values. The root mean squared error (RMSE) is calculated by the following equation: $RMSE = \sqrt{\sum (y_i - f_i) / N}$.

having the highest properties are selected for the mating pool as a greedy sampling method to enable sufficient genomic variations at each generation. We note that selecting the proper amount of data for the mating pool is important when considering the trade-off between the risk of being stuck in local minima and computational time. Additionally, we utilize the intuition from solid mechanics that the symmetrical microstructure is beneficial for the load-bearing capacity and that soft material at the crack tip is able to relieve the stress concentration at the crack tip^{1,4}. A hyper-heuristic genetic algorithm combining this domain knowledge is implemented by constraining the explored microstructure design to satisfy the prescribed conditions. The constraints accelerated the optimization process compared to conventional particle swarm optimization and genetic algorithm (See the details

of the comparison for each optimization method in Supplementary Figure 6). In a genetic algorithm, the crossovers are implemented by selecting two microstructures from the mating pool as parents and randomly assigning stiff material to the area occupied by stiff materials in parent configurations. The mutations are applied by randomly switching the position of stiff material block and soft material block by keeping the ratio between the stiff and soft blocks. Approximately 4×10^4 unique candidate microstructures are generated from the mating pool at each generation.

The output values of the candidate microstructures are predicted with the DNN. Because the microstructures proposed by the genetic algorithm are located close to the microstructures from the previous generation in terms of feature space and output

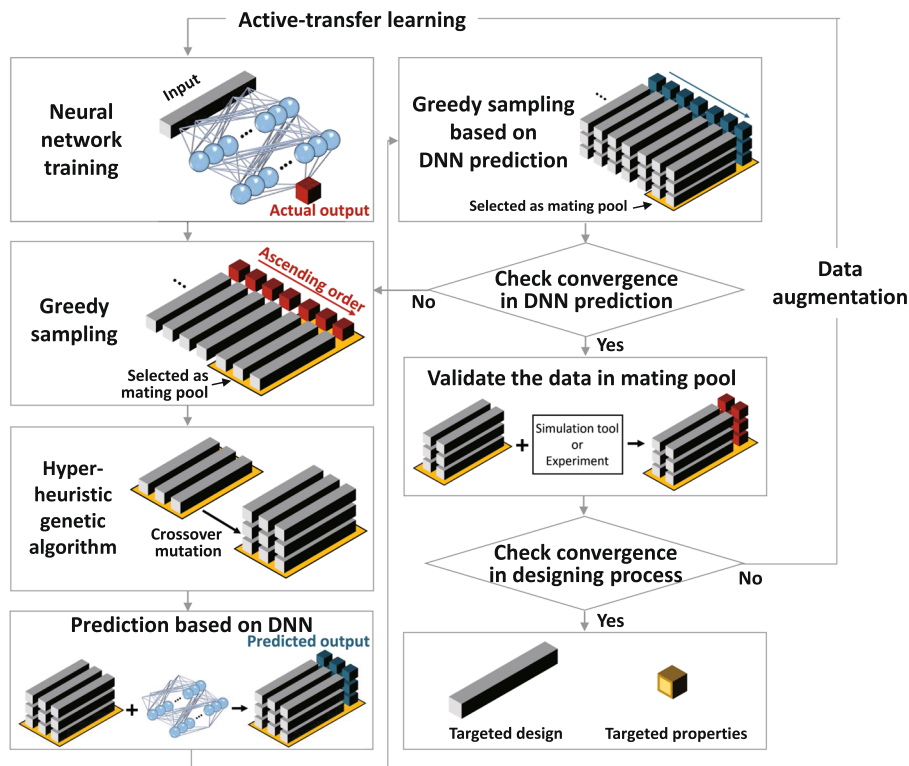


Fig. 4 Flowchart of design framework. Flowchart of the hyper-heuristic genetic optimization algorithm based on active transfer learning.

values, the DNN could make reliable predictions on the microstructures suggested by the genetic algorithm for certain generations. Based on those predictions, we select candidate microstructures, which are expected to have output values closer to the target design than the existing microstructures, for the mating pool, and continue this process until DNN-predicted output values do not show improvement, i.e., convergence to the predictable limit of DNN. After the convergence, we calculate the true output values of the candidate microstructures in the mating pool with FEA simulations and integrate them to the training data with data augmentation techniques. The data augmentation is implemented by oversampling the new data, i.e., replicating the new data by 50 times which is split into training and validation sets with a 9:1 ratio. By employing the data augmentation methods, emphasis is placed on the generalization on the unseen domain, as suggested by the previous study^{53,58,59}. Without the data augmentation, the added data are diluted in the training data because it is relatively sparser and smaller compared to the existing training dataset. We note that other data augmentation techniques, such as synthetic minority oversampling technique, are difficult to be adapted due to the characteristic of the input features, i.e., binary matrix, for the composite optimization problem considered in the study, and that different data augmentation schemes can be adapted depending on the characteristics of the design problem. The update of DNN is conducted by re-training the DNN based on the integrated training data with reduced learning rate (10^{-6}) and reduced training epoch (10) in the transfer learning scheme. Because DNN requires relatively moderate updates on learnable parameters to expand the reliable prediction domain, it takes a dramatically reduced time for DNN updates compared to the initial training. The details of the DNN training history are included in Supplementary Figures 8–9.

Composite microstructures with maximum mechanical properties

The design process for maximizing stiffness and strength is represented in Fig. 5. The data points at the 0th update represent the initial training dataset. In the design process for maximizing stiffness shown in Fig. 5 (a), approximately 1.6×10^5 unique microstructures are investigated within 4 h through DNN. This corresponds to the several orders of magnitude speedup compared to FEA simulation tools which would take more than 8 weeks for the generation of 10^5 data. In the design process for maximum stiffness, a significant improvement is observed with the initial training of DNN, because the stiffness prediction upon unseen output value range indicates that the DNN is able to distinguish the relative magnitudes of stiffness despite some mismatch in absolute values. We hypothesize that the moderate level of extrapolation beyond the seen domain is made possible due to the piecewise linear unbound function from leaky ReLU used in the DNN. Indeed, our test on various activation functions shows that unbounded ReLU-based DNN has reasonable predictive power on the unseen domain, while bounded activation functions are inefficient for the task (Supplementary Figure 1). The iterative update of DNN is also beneficial in making robust improvements on local minima explorability with the DNN and in escaping from local minima, as demonstrated by the gradual improvement of converged values shown in Fig. 5a. Given that almost identical microstructure designs come out from the 3rd and 4th updates of DNN, we believe that the design process is converged at the optimization process after the 4th update. An identical process is also applied for finding the microstructure with the maximum strength, which is shown in Fig. 5b. The maximum strength value increases until we update the DNN iteratively five times.

The material designs having the six highest stiffness and strength at the last DNN update are depicted in Fig. 6. The best microstructures for stiffness and strength are positioned at the

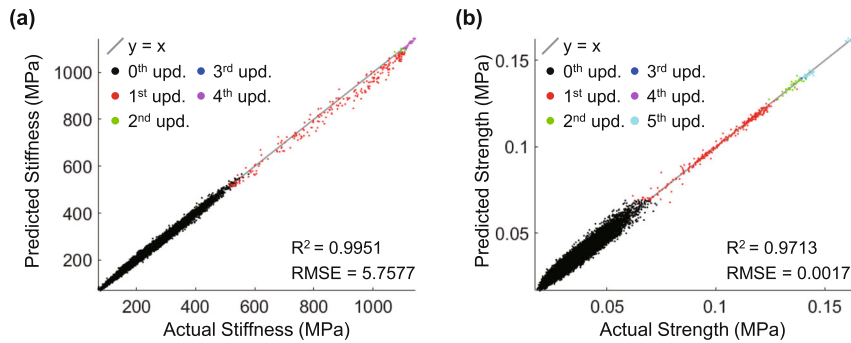


Fig. 5 Designing progress plot. Designing progress plot for maximum (a) stiffness and (b) strength. The values are predicted from DNN trained at the last update. The data obtained from each iteration are represented with a different color.

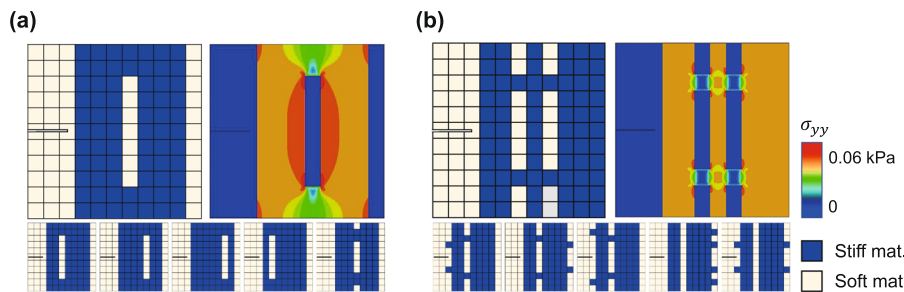


Fig. 6 Optimized composite structures. Top 6 composite structures optimized for maximum (a) stiffness and (b) strength. The enlarged structures on the top are structures having the highest stiffness and strength value, and the corresponding axial stress (σ_{yy}) distribution is colored at $\varepsilon_{yy} = 2.2727 \times 10^{-5}$. The maximum stiffness and strength of composite having optimized microstructures are 1144.7 and 0.16276 MPa, respectively. The five structures having the following highest value of stiffness and strength are represented below.

upper left corners of Fig. 6a and Fig. 6b, respectively. Microstructures ranked at 2nd to 6th are visualized at the bottom. The theoretical upper bound of elastic modulus (E^{max}) of composite having a volume fraction of stiff materials (f) can be achieved with the following equation: $E^{max} = E^{stiff} f + E^{soft} (1 - f)$. For our design problem with $f = 71/121$, E^{max} is 1223.7 MPa if it were not for the pre-existing crack. The stiffness of optimized microstructures (1144.7 MPa) is close to the theoretical upper bound even though a pre-crack exists that inherently reduces the stiffness of the structure. Therefore, we can infer that the maximum stiffness design found from our framework is close to the global optimum. For the composite with maximum strength, the strength is significantly improved compared to the best design in the training set. In those designs, a majority of stiff blocks are located in the region far from the pre-crack (right side), because such unbalanced material distribution leads to the increased stiffness while reducing the negative effect from the pre-crack. At the same time, since the crack initiation may occur at any point with a high-stress concentration other than the pre-existing crack tip, it is important to reduce the stress concentration at the vertices made by different material blocks. Figure 6b shows that the stress concentration at the crack tip is relieved owing to the optimized structure compared to some randomly chosen composite structures in the initial training data (Supplementary Figure 10). As a result, we could obtain a design that is more than twice as strong as the best design in the original dataset, (i.e., the maximum strength increases from 0.076165–0.16276 MPa throughout the optimization process). The proposed designs also outperform the composite structures which are manually designed based on the physical intuitions (See the details in Supplementary Figure 11). The results for maximum stiffness and strength are obtained only by augmenting 366 and 424 microstructures, respectively. The augmented dataset size is about 0.4% of the initial training dataset size. The histograms for the additional data are depicted in Fig. 7.

DISCUSSION

In this study, we propose a systematic forward material design framework to obtain superior design far beyond the domain of the initial training dataset. The framework is applied to a composite microstructure design problem for obtaining the maximum mechanical properties of 11×11 grid composite, which has an astronomically high number of possible configurations. Because this type of design process inherently requires efficient search in the unseen domains, a forward design approach is adopted by gradually expanding the reliable prediction domain of DNN with active transfer learning and data augmentation. The better design candidates are firstly proposed by the genetic algorithm based on DNN predictions. The properties of the candidate are obtained with FEA calculations before they are augmented to the training dataset, in order to secure the predictive performance of the DNN for a newly added dataset. The limitations in a forward design approach, such as being stuck in local minima, are mitigated by updating DNN and controlling the mutation methods in a hyper-heuristic genetic algorithm.

We emphasize that the iterative and gradual expansion of the reliable prediction domain assisted by the genetic algorithm is key to the superior efficiency of our framework. In contrast to candidates obtained directly from advanced generative models in previous studies, candidates suggested from the genetic algorithm are located relatively closer to the dataset from the previous generation in terms of feature space and output values. Hence, the prediction accuracy of DNN is maintained in some degree during the gradual expansion of the reliable prediction domain. Also, expansion proceeds along relatively narrow but correct routes towards the target design as depicted in Fig. 1. The superior efficiency of our design framework can be more clearly demonstrated by comparing the size of the additional datasets required to get the final design. In our study, only $\sim 4 \times 10^2$ additional data are validated to update the DNN originally trained

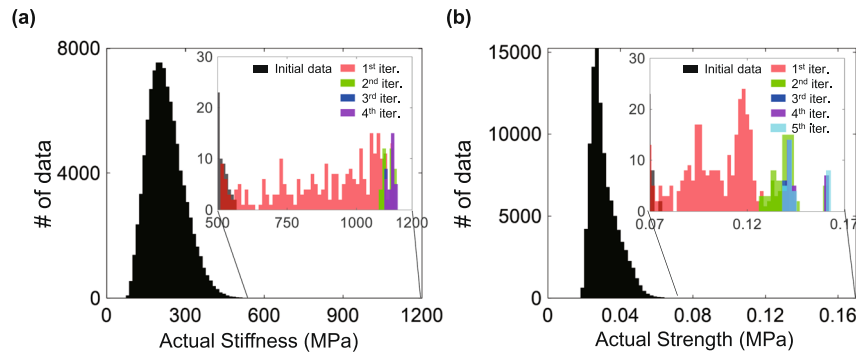


Fig. 7 Histogram of optimization results. Histogram of (a) stiffness and (b) strength after the optimization. The initial dataset are colored black, the data from 1st, 2nd, 3rd, 4th, and 5th updates of DNN are colored red, green, blue, purple, and cyan, respectively.

with the 10^5 initial training data in the design space having $\sim 1.8 \times 10^{34}$ possible configurations.

We expect that our framework is applicable to a wide range of optimization problems in other science and engineering disciplines with astronomically large design space, because it provides an efficient way of gradually expanding the reliable prediction domain toward the target design while avoiding the risk of being stuck in local minima. Especially, being a less-data-hungry method, design problems in which data generation is time-consuming and expensive will benefit most from our framework.

METHODS

Data generation based on finite element analysis (FEA)

As a training dataset, the mechanical properties, i.e., stiffness, strength, and toughness, of two-dimensional 11×11 grid composites are obtained from finite element analysis (FEA) under plane stress conditions. The composites consist of perfectly bonded 70 stiff material blocks and 51 soft material blocks containing a pre-existing crack. The input features are formatted as one-hot binary encoding representing the position of stiff and soft materials, and the outputs are defined as corresponding mechanical properties. The 100,000 unique microstructures are constructed, and the corresponding stress-strain curves are obtained by applying uniaxial tension with quasi-static infinitesimal strain increment of 0.0000227 until the complete fracture occurs. 100,000 data correspond to the fraction of $\sim 10^{-29}$ out of the total number of available combinations, $C(120, 70) \sim 10^{34}$. The stiffness, strength, and toughness are measured by the initial gradient of a stress-strain curve, the maximum stress point, and the total area under the stress-strain curve, respectively. The materials are assumed as linear elastic materials with the following elastic modulus (E), Poisson's ratio (ν) and critical strain energy release rate (G_c), respectively, for stiff and soft materials: $E^{\text{stiff}} = 2100 \text{ MPa}$, $\nu^{\text{stiff}} = 0.3$, $G_c^{\text{stiff}} = 50 \text{ J/m}^2$, $E^{\text{soft}} = 21 \text{ MPa}$, $\nu^{\text{soft}} = 0.3$, $G_c^{\text{soft}} = 50 \text{ J/m}^2$. For the simulation of the crack growth, the hybrid crack phase-field (CPF) model are adopted in commercial FEA software ABAQUS with user-element subroutine⁶⁰. The hybrid CPF model enables the modeling of complex crack nucleation and propagation by describing the failure of material with a continuous scalar field called crack phase field $d(x)$. The material completely loses its load-bearing capacity when $d(x)$ reach to 1, whereas $d(x) = 0$ indicates the absence of any damage. A detailed explanation and implementation of the model can be found in our previous study⁶⁰.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The codes developed in this study are available from the authors upon reasonable request.

Received: 19 April 2021; Accepted: 10 August 2021;

Published online: 02 September 2021

REFERENCES

- Libonati, F., Gu, G. X., Qin, Z., Vergani, L. & Buehler, M. J. Bone-inspired materials by design: toughness amplification observed using 3D printing and testing. *Adv. Eng. Mater.* **18**, 1354–1363 (2016).
- Sigmund, O. & Torquato, S. Design of materials with extreme thermal expansion using a three-phase topology optimization method. *J. Mech. Phys. Solids* **45**, 1037–1067 (1997).
- Sigmund, O. & Torquato, S. Design of smart composite materials using topology optimization. *Smart Mater. Struct.* **8**, 365 (1999).
- Kim, Y., Kim, Y., Libonati, F. & Ryu, S. Designing tough isotropic structural composite using computation, 3D printing and testing. *Compos. Part B: Eng.* **167**, 736–745 (2019).
- Guan, A.-Y., Liu, C.-L., Sun, X.-F., Xie, Y. & Wang, M.-A. Discovery of pyridine-based agrochemicals by using intermediate derivatization methods. *Bioorg. Med. Chem.* **24**, 342–353 (2016).
- Zhu, H. et al. Discovery of TaFeSb-based half-Heuslers with high thermoelectric performance. *Nat. Commun.* **10**, 1–8 (2019).
- Carraturo, M. et al. Graded-material design based on phase-field and topology optimization. *Comput. Mech.* **64**, 1589–1600 (2019).
- Chen, C.-T., Chrzan, D. C. & Gu, G. X. Nano-topology optimization for materials design with atom-by-atom control. *Nat. Commun.* **11**, 1–9 (2020).
- Holder, A. M. et al. Novel phase diagram behavior and materials design in heterostructural semiconductor alloys. *Sci. Adv.* **3**, e1700270 (2017).
- Kim, Y., Jeong, H., Gu, G. X. & Ryu, S. A three-dimensional fracture pattern diagram of staggered platelet structures. *Compos. Struct.* **220**, 769–775 (2019).
- Radman, A., Huang, X. & Xie, Y. Topological design of microstructures of multi-phase materials for maximum stiffness or thermal conductivity. *Comput. Mater. Sci.* **91**, 266–273 (2014).
- de Kruijff, N., Zhou, S., Li, Q. & Mai, Y.-W. Topological design of structures and composite materials with multiobjectives. *Int. J. Solids Struct.* **44**, 7092–7109 (2007).
- Osanov, M. & Guest, J. K. Topology optimization for architected materials design. *Annu. Rev. Mater. Res.* **46**, 211–233 (2016).
- Pruksawan, S., Lambard, G., Samitsu, S., Sodeyama, K. & Naito, M. Prediction and optimization of epoxy adhesive strength from a small dataset through active learning. *Sci. Technol. Adv. Mater.* **20**, 1010–1021 (2019).
- Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* **20**, 318–331 (2015).
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- Liu, X., Athanasiou, C. E., Pature, N. P., Sheldon, B. W. & Gao, H. A machine learning approach to fracture mechanics problems. *Acta Mater.* **190**, 105–112 (2020).
- Hanakata, P. Z., Cubuk, E. D., Campbell, D. K. & Park, H. S. Forward and inverse design of kirigami via supervised autoencoder. *Phys. Rev. Res.* **2**, 042006 (2020).
- Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 1–6 (2013).
- Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 1–13 (2017).
- Chen, C. T. & Gu, G. X. Effect of constituent materials on composite performance: exploring design strategies via machine learning. *Adv. Theory Simul.* **2**, 1900056 (2019).
- Gu, G. X., Chen, C.-T., Richmond, D. J. & Buehler, M. J. Bioinspired hierarchical composite design using machine learning: simulation, additive manufacturing, and experiment. *Mater. Horiz.* **5**, 939–945 (2018).

23. Gu, G. X., Chen, C.-T. & Buehler, M. J. De novo composite design based on machine learning algorithm. *Extrem. Mech. Lett.* **18**, 19–28 (2018).
24. Kim, Y., Yang, C., Kim, Y., Gu, G. X. & Ryu, S. Designing an adhesive pillar shape with deep learning-based optimization. *ACS Appl. Mater. Inter.* **12**, 24458–24465 (2020).
25. Mansouri Tehrani, A. et al. Machine learning directed search for ultra-incompressible, superhard materials. *J. Am. Chem. Soc.* **140**, 9844–9853 (2018).
26. Noh, J., Gu, G. H., Kim, S. & Jung, Y. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* **11**, 4871–4881 (2020).
27. Janet, J. P., Chan, L. & Kulik, H. J. Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J. Phys. Chem. Lett.* **9**, 1064–1071 (2018).
28. Chen, W.-C., Schmidt, J. N., Yan, D., Vohra, Y. K. & Chen, C.-C. Machine learning and evolutionary prediction of superhard BCN compounds. *npj Comput. Mater.* **7**, 1–8 (2021).
29. Avery, P. et al. Predicting superhard materials via a machine learning informed evolutionary structure search. *npj Comput. Mater.* **5**, 1–11 (2019).
30. Simpson, T. W., Mauery, T. M., Korte, J. J. & Mistree, F. Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA J.* **39**, 2233–2241 (2001).
31. Jeong, S., Murayama, M. & Yamamoto, K. Efficient optimization design method using kriging model. *J. Aircr.* **42**, 413–420 (2005).
32. Li, Z., Gerdroodbary, M. B., Valipour, P., Moradi, R. & Babazadeh, H. The optimization via response surface method for micro hydrogen gas actuator. *Int. J. Hydrog. Energy* **44**, 31633–31643 (2019).
33. Peurifoy, J. et al. Nanophotonic particle simulation and inverse design using artificial neural networks. *Sci. Adv.* **4**, eaar4206 (2018).
34. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
35. Chen, C. T. & Gu, G. X. Generative deep neural networks for inverse materials design using backpropagation and active learning. *Adv. Sci.* **7**, 1902607 (2020).
36. Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324 (2020).
37. Hsu, Y.-C., Yu, C.-H. & Buehler, M. J. Tuning mechanical properties in polycrystalline solids using a deep generative framework. *Adv. Eng. Mater.* **23**, 20201339 (2021).
38. Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminformatics* **10**, 1–9 (2018).
39. Kollmann, H. T., Abueidda, D. W., Koric, S., Guleryuz, E. & Sobh, N. A. Deep learning for topology optimization of 2D metamaterials. *Mater. Des.* **196**, 109098 (2020).
40. Kim, K. et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput. Mater.* **4**, 1–7 (2018).
41. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
42. Yao, Z. et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **3**, 76–86 (2021).
43. Marcus, G. Deep learning: A critical appraisal. Preprint at <https://arxiv.org/abs/1801.00631> (2018).
44. Martius, G. & Lampert, C. H. Extrapolation and learning equations. Preprint at <https://arxiv.org/abs/1610.02995> (2016).
45. Barnard, E. & Wessels, L. Extrapolation and interpolation in neural network classifiers. *IEEE Control Syst. Mag.* **12**, 50–53 (1992).
46. Mitchell, J., Minervini, P., Stenetorp, P. & Riedel, S. Extrapolation in NLP. Preprint at <https://arxiv.org/abs/1805.06648> (2018).
47. Greydanus, S., Dzamba, M. & Yosinski, J. Hamiltonian neural networks. *Adv. Neural Inf. Process. Syst.* **32**, 15379–15389 (2019).
48. Long, Z., Lu, Y. & Dong, B. PDE-Net 2.0: learning PDEs from data with a numeric-symbolic hybrid deep network. *J. Comput. Phys.* **399**, 108925 (2019).
49. Kloss, A., Schaal, S. & Bohg, J. Combining learned and analytical models for predicting action effects. Preprint at <https://arxiv.org/abs/1710.04102> (2017).
50. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, 1–17 (2019).
51. Lopedoto, E. & Weyde, T. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. 159–165 (Springer).
52. Xu, K. et al. How neural networks extrapolate: from feedforward to graph neural networks. Preprint at <https://arxiv.org/abs/2009.11848> (2020).
53. Hooker, G. & Rosset, S. Prediction-based regularization using data augmented regression. *Stat. Comput.* **22**, 237–249 (2012).
54. He, K., Zhang, X., Ren, S. & Sun, J. In *European conference on computer vision*. 630–645 (Springer, 2016).
55. He, K., Zhang, X., Ren, S. & Sun, J. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778 (Springer, 2016).
56. Mozaffar, M. et al. Deep learning predicts path-dependent plasticity. *Proc. Natl Acad. Sci.* **116**, 26414–26420 (2019).
57. Abueidda, D. W., Koric, S., Sobh, N. A. & Sehitoglu, H. Deep learning for plasticity and thermo-viscoplasticity. *Int. J. Plasticity* **136**, 102852 (2021).
58. Li, L. & Ma, C. In *IOP Conference Series: Earth and Environmental Science*. 022095 (2019).
59. Campagna, R. & Perracchione, E. Data-driven extrapolation via feature augmentation based on variably scaled thin plates splines. *J. Sci. Comput.* **88**, 1–15 (2021).
60. Jeong, H., Signetti, S., Han, T.-S. & Ryu, S. Phase field modeling of crack propagation under combined shear and tensile loading with hybrid formulation. *Comput. Mater. Sci.* **155**, 483–492 (2018).

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program (2019R1A2C4070690) and Creative Materials Discovery Program (2016M3D1A1900038) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea, as well as the KAIST-funded Global Singularity Research Program for 2019 (N11190118). Additionally, the authors acknowledge funding support from 3M and Inc. to HP Labs.

AUTHOR CONTRIBUTIONS

Yongtae Kim and Youngsoo Kim conducted the research under the guidance of Seunghwa Ryu. Yongtae Kim and Seunghwa Ryu wrote this manuscript. Charles Yang and Kundo Park contributed to the results analysis. Grace X. Gu revised the paper. All authors revised and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00609-2>.

Correspondence and requests for materials should be addressed to S.R.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021