Electronic Supplementary Material (ESI) for Nanoscale Advances.

SISSO-assisted prediction and design of mechanical properties of porous graphene with a uniform nanopore array

Anran Wei [a], Han Ye [b,*], Zhenlin Guo [c], Jie Xiong [d,e*]

[a] School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[b] State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China

[c] Mechanics Division, Beijing Computational Science Research Center, Beijing 100193, China

[d] School of Materials Science and Engineering, Harbin Institute of Technology, Shenzhen 518055, China

[e] Shenzhen Research Institute, The Hong Kong Polytechnic University, Shenzhen 518057, China

* Han_ye@bupt.edu.cn (H.Y.); george-jie.xiong@connect.polyu.hk (J.X.)

1. Introduction of other forecasting methods for comparison

Typical forecasting methods available in the python library scikit-learn [1] are employed in the present work to compare with the SISSO algorithm [2] adopted in this study, which are briefly described as following:

1.1. Linear regression

Linear regression (LR) [3] is one of the simplest and oldest methods to build the relationship between different quantities, in which the targeted property is a linear combination of the features. The least-squares fitting, which fits the linear model to minimize the residual sum of squares between the predicted values and actual values of the targeted property, is used in this work.

1.2. Support vector regression

The support vector regression (SVR) [4] is based on the structural risk minimization principle. It employs kernel functions to convert the features into a higher dimensional space according to the targeted property. Two kernel functions are used in this work, the linear function (SVR_lin) [5] and the radial basis function (SVR_rbf) [6].

1.3. Decision tree regression

The decision trees (DTs) [7], including the ID3, C4.5, CART algorithms, are non-parametric supervised learning algorithms used for classification and regression. One of the DT algorithms, the classification and regression trees (CART) [8], which constructs binary trees using the feature and threshold that yield the minimum mean squared error at each node, is utilized in the present work.

1.4. Random forest regression

Random forest (RF) regression [9] is a parallel ensemble learning approach using DT as the base learner. Individual DTs usually exhibit high variance and tend to overfit. RF can achieve a reduced variance by taking an average of predictions of individual DTs with a slight increase of bias.

2. Performance of other forecasting methods for comparison

The LR is first employed to fit linear models for the relationship between targeted properties and features based on the least squared approximation. The d and λ affect strength and work of fracture are expressed as

$$\sigma_s = 0.1637d - 66.88\lambda + 66.81 \#(S1)$$

$$W_c = -0.0238d - 2.382\lambda + 2.578 \#(S2)$$

Fig. S1a and Fig. S1b plot the predicted results of LR models against the MD-calculated values of strength and work of fracture, respectively. It is found that the linear models cannot predict targeted properties with d and λ well. The averaged R^2 is 0.7294, in which $R^2 = 0.8974$ for strength and $R^2 = 0.5613$ for work of fracture.

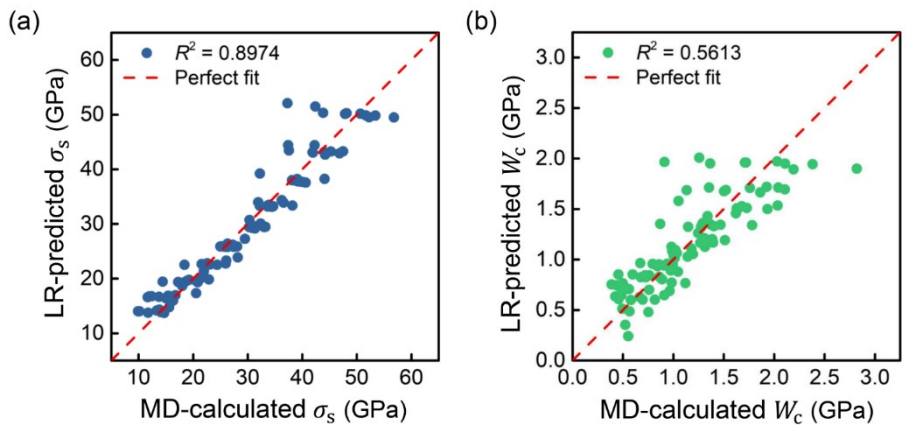

Fig. S1 Predictions of (**a**) strength (σ_s) and (**b**) work of fracture (W_c) by the linear regression (LR) model against real values calculated by molecular dynamics (MD) simulations.

Table S1 Hyperparameters that should be tuned for each machine learning algorithm.

ML model	hyperparameters	Set of values
SVR_lin	*C*	0.1, 1, 10, 100, 1000, 10000, 100000
SVR_rbf	*C*	0.1, 1, 10, 100, 1000, 10000, 100000
	gamma	"auto", "scale"
DT	*max_depth*	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
	max_features	1, 2
RF	*n_estimators*	65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125

Defining the architecture of an ML model is important before training it. Those parameters defining the architecture of an ML model are termed as hyperparameters. The hyperparameters that need to be tuned for good performance of the next four ML algorithms are listed in Table S1. Only one or two hypermeters affect the performance of SVR_lin, SVR_rbf, and DTs, while there are more hypermeters related to the performance of an RF model. Three dominant hyperparameters of the RF model, *n_estimators*, *max_depth*, and *max_feature*, are considered in this work. Since an RF is assembled by individual DTs, we use the DTs with tuned hyperparameters as the base learner of RF for convenience, i. e., the *max_depth* and *max_features* of RF are the same as the searched values of DT. Only the *n_estimators*, the number of trees that form the forest, is tuned for the RF model. The grid search is applied to tune those hyperparameters within the sets of values listed in Table S1. The hyperparameter tuning processes for SVR_lin, SVR_rbf, DT, and RF are displayed in Fig. S2. The maximum averaged R^2 and the corresponding hyperparameters for each algorithm are listed in Table S2.

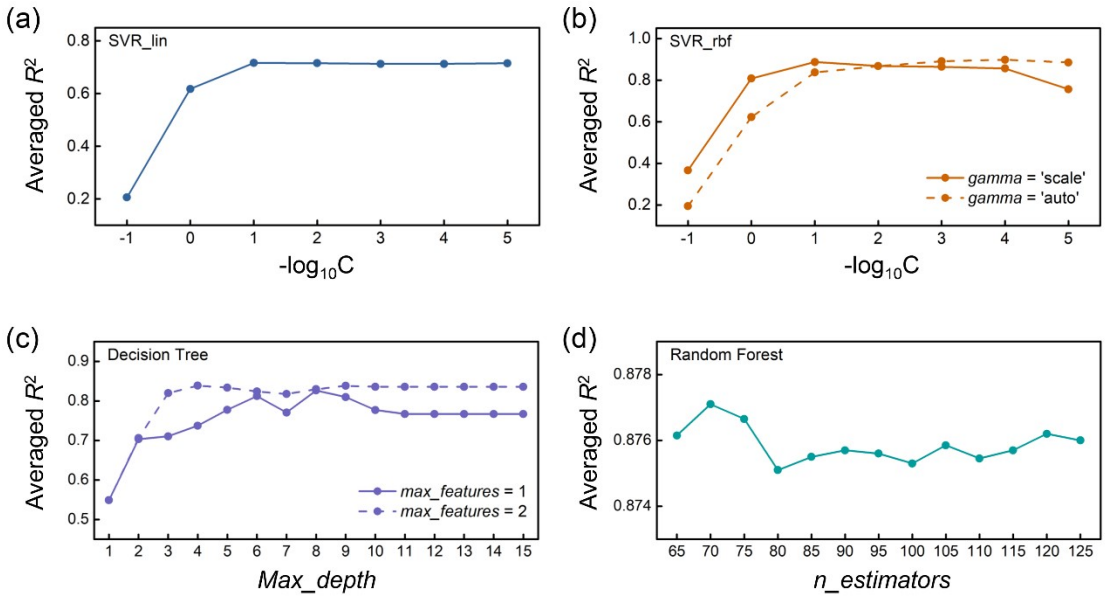

Fig. S2 Performance of (**a**) SVR_lin, (**b**) SVR_rbf, (**c**) DT, and (**d**) RF models with different hyperparameters.

Table S2 Maximum R^2 of each machine learning algorithm and corresponding hyperparameters.

ML model	Maximum R^2	Hyperparameters
SVR_lin	0.7162	C = 10
SVR_rbf	0.8980	C = 10000; *gamma* = "auto"
DT	0.8385	*max_depth* = 4; *max_features* = 2
RF	0.8771	*n_estimators* = 70

It is noticed from the training results that the SVR_rbf outperforms other ML algorithms with an averaged R^2 of 0.898 when $C = 10000$ and $gamma =$ "auto". The DT and RF also show superior performance than the linear models. The weakest model SVR_lin has similar performance with the linear models. Fig. S3 shows the predicted strength and work of fracture in all test sets using the SVR_rbf model against the MD-calculated values. It performs much better than the linear models on the predictions of high strength and high work of fracture.

The tree-based algorithm can provide feature importance to interpret the ML models. For instance, in the current work, the feature importance of two fed features in RF models are given as Fig. S4.

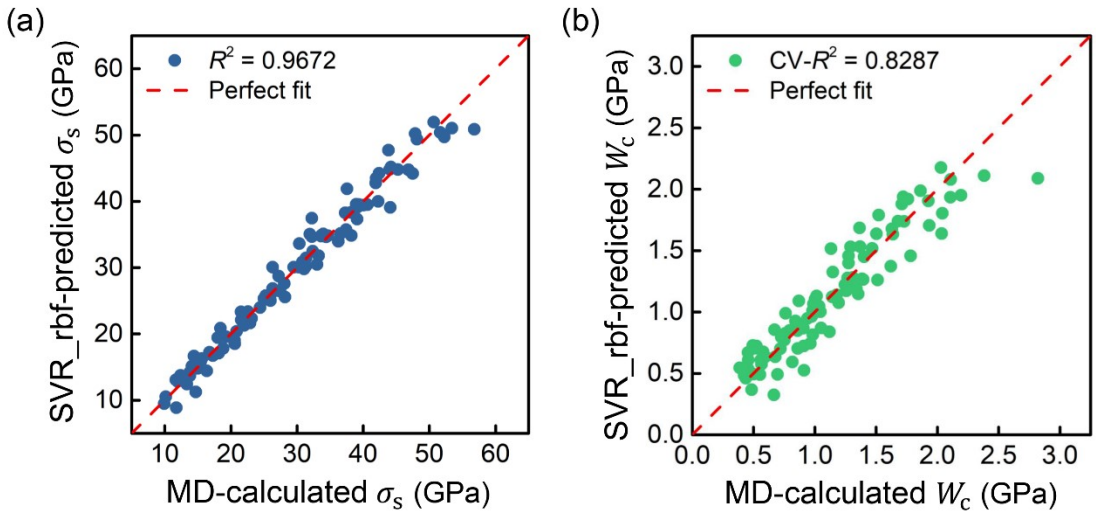

Fig. S3 Predictions of (**a**) strength (σ_s) and (**b**) work of fracture (W_c) by the SVR_rbf model against real values calculated by molecular dynamics (MD) simulations.

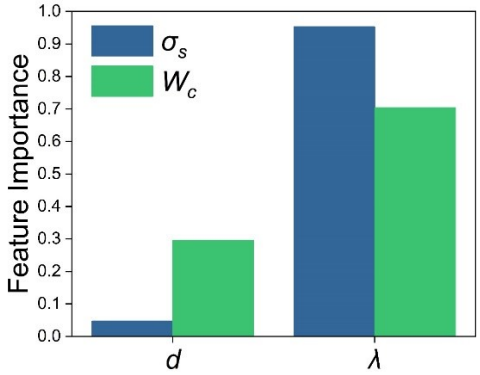

Fig. S4 Feature importance obtained by random forest method.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825-2830.

[2] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Mater. 2 (2018) 083802.

[3] D.C. Montgomery, E.A. Peck, G.G. Vining, Introduction to linear regression analysis, John Wiley & Sons, Hoboken, New Jerrsy, 2012.

[4] M. Awad, R. Khanna, Support Vector Regression, in: M. Awad, R. Khanna (Eds.) Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, Apress, Berkeley, CA, 2015, pp. 67-80.

[5] C.-H. Ho, C.-J. Lin, Large-scale linear support vector regression, J. Mach. Learn. Res. 13 (2012) 3323–3348.

[6] B. Kuo, H. Ho, C. Li, C. Hung, J. Taur, A Kernel-Based Feature Selection Method for SVM With RBF Kernel for Hyperspectral Image Classification, IEEE J-STARS 7 (2014) 317-326.
[7] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81-106.
[8] G. De'ath, K.E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology 81 (2000) 3178-3192.
[9] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947-1958.