Modelling and Simulation in Materials Science and Engineering

A novel structure-property relationship model based on machine learning

To cite this article before publication: Huiran Zhang *et al* 2020 *Modelling Simul. Mater. Sci. Eng.* in press https://doi.org/10.1088/1361-651X/ab6bb7

A novel structure-property relationship model based on machine learning

HUIRAN ZHANG[1,2,3], ZHITING GUO[1], HONGQING HU[1], GAOFENG ZHOU[1], QING LIU[1], YAN XU[4], QUAN QIAN[1], DONGBO DAI[1,*]

[1]*School Computer Engineering and Science, Shanghai University, Shanghai 200444, China;*

[2]*Materials Genome Institute of Shanghai University, Shanghai University, Shanghai 200444, China;*

[3]*Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China;*

[4]*College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China;*

Keywords: *Binary alloys; Structure–property relationship; Solid solution; Machine learning;*

Abstract

In materials science, the relationship between the material internal structure and its associated macroscale properties can be used to guide the design of materials. In this study, we constructed an interpretative machine learning model to capture the structure-property relationship and predict the solid solubility in binary alloy systems. To do this, we used a dataset containing about 1,843 binary alloys and corresponding experiment values of solid solubility. We designed a common function to represent the relationship between individual descriptor and solid solubility, and a deep neural network to integrate the multiple functions. The resulting model can correctly predict the solid solubility value than other machine learning models. What's more, based on this model, it's feasible to analyze the effect of structures on target property.

1. Introduction

Data-driven methods, including machine learning (ML) methods, could use materials datasets to discover the relationship between structures and properties [1]. Usually, the relationship is obtained via material experiments combined with domain knowledge [2–5]. Based on material experiments and computer simulations, there has accumulated many materials datasets, which contain existing experimental structures and properties databases [6–9], computed structures and properties databases [10–13], and the datasets which come from references [14,15]. Incorporating the ML methods and these materials datasets, it is considered that the procedure of the discovery can be accelerated in most situations [16–21]. Thus, there has formed a mature framework focusing on applying ML methods to materials research including structure-property calculation, crystal structure prediction and statistically driven design [22].

Too many researches show that ML methods have been combined with quantitative structure-property relationship (QSPR) models. In 1991, Bolis et al. proposed an AI algorithm based on ML to construct a structure-activity relationship (SAR, which is equivalent to QSPR when a chemical property is modeled as the response variable) model for drug design [23]. In 1998, Zheng et al. proposed a novel automated variable selection QSPR approach based on K-nearest neighbors (kNN, an ML

method) to predict the activity of compounds, and the experiments based on several datasets showed that kNN-QSPR models were supported by the statistical hypothesis testing [24]. In 2005, Liu et al. used a least-squares support vector machine (LSSVM) to build an accurate QSPR model for the prediction of the C_{60}'s solubility in various solvents [25]. In 2012 & 2014, Golmohammadi et al. and Fernandez et al. also applied SVM to their own QSPR models [26,27]. However, these QSPR models, which are based on black box, cannot tell the researchers the relationship between target property and each descriptor contained in the original dataset.

In this paper, we proposed a QSPR model based on least-squares and deep neural network (DNN). This model could explain the influence of each descriptor on target property. We constructed some mathematical expressions to capture the statistical characteristics of the descriptors and then these expressions were used as inputs of the neural network to predict the target property. A binary alloy dataset is used in our work and solid solubility is the target property. The result shows that the accuracy of our model is higher than other general ML methods and our model also has better interpretability. The code has been made available at: https://github.com/frearb/DNN-QSPR.

2. Data collection

We collected solid solubility values of binary alloys from the *Volume 3 of the ASM Handbook*, which was published in 1992 [28]. The dataset contains most of the binary alloy systems. Most features (or structures) of atoms in the binary alloy systems are obtained from *The Materials Project (https://materialsproject.org)* via a python library pymatgen (Python Materials Genomics) [29]. The features include Fermi level electronic number, bond energy, atomic mass, atomic number, atomic radius, boiling point, melting point, Mendeleev number, atomic volume, and electronegativity. Then we obtained a dataset include 1,843 examples with 10 features.

We adopted the following strategies in our dataset to regenerate descriptors:

1) The difference of a feature between solute and solvent atoms is used as descriptor. The size factor has an effect on the atomic arrangement in the alloy solid. A difference in the Pauling electronegativity between the two constituent elements determined the degree of charge transfer between neighboring unlike atoms or the degree of iconicity but not that of covalency. It's more likely to favor compound formation when a solute has a large difference in electronegativity relative to the host [30]. According to Hume-Rothery's rules [31,32], it states that solute and solvent are likely to form solid solution while they have the same valency and similar crystal structure. So, it is true that the difference between solute and solvent affects the formation of solid solution.

2) The features which are highly correlated or have similar meaning are excluded. The similar features are worthless in building models, such as the atomic mass and atomic number. These features, which are considered redundant features, may slow down the learning process or cause the classifier to over-fit the training data [33].

3) The descriptors that are easy to characterize will be preferred. We visualized the two-dimensional distribution of each descriptor and target property. Some distribution maps are disorderly and irregular, which could jeopardize our analysis.

After applying these strategies, we obtained a dataset with four descriptors including radius descriptor, Mendeleev descriptor, Fermi descriptor and electronegativity descriptor. Radius descriptor is represented as $\Delta r = r_{solute} - r_{solvent}$. Mendeleev descriptor is represented as $\Delta MN = MN_{solute} - MN_{solvent}$,

where MN means Mendeleev number. Fermi descriptor is represented as $\Delta FE = FE_{solute} - FE_{solvent}$, where FE means the number of electrons above the Fermi level. Electronegativity descriptor is represented as $\Delta\chi = \chi_{solute} - \chi_{solvent}$. The solid solubility is defined as $m_{Solute}/(m_{Solute} + m_{solvent})$.

Fig. 1 complements the characterization of our dataset by showing the distribution of the descriptors and solid solubility respectively. In the whole 1,843 binary alloy systems, about half of them cannot form solid solutions, in which the solid solubility is zero. About 20% of these binary alloys can form solid solutions in any ratio, in which the solid solubility is 1. The last 30% can form solid solutions with solubility limits. When the value of a feature in the solvent is greater than in the solute, the corresponding value of descriptor is a negative number, and vice versa. It's certain that the distribution of solid solubility and other descriptors is not completely random, in other words, the closer a descriptor trends to zero the higher the solid solubility value is.

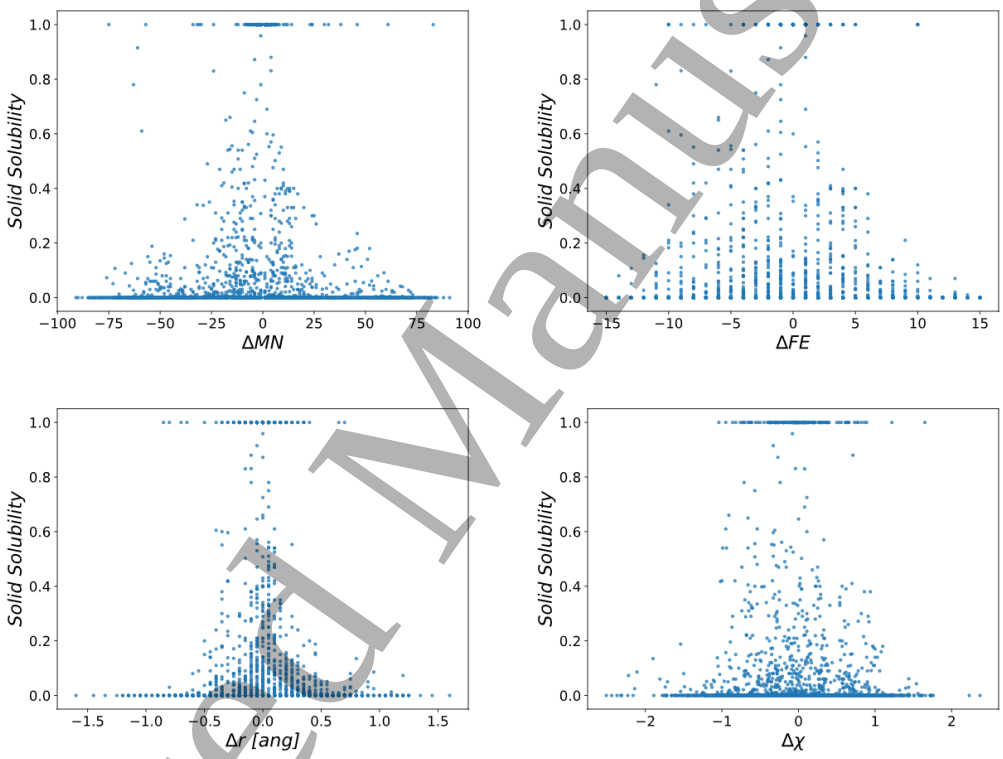

Fig. 1. Distribution of solid solubility and individual descriptors. Each point represents a binary alloy.

3. Designing the QSPR model

In the prediction of solid solubility in binary alloy systems, there exist one target property and several related descriptors, and the goal is to predict the target property values with the descriptors. In this model, we firstly analyzed the relationship between the target property (marked as y) and each related descriptor (marked as x_i) to find out a common function f which can capture the statistical characteristics between y and every x_i as $y \approx f(x_i, \theta_i)$ where θ_i is the model argument for x_i. Then we used another function g to integrate multiple functions into one as $y = g\big(f(x_1, \theta_1), f(x_2, \theta_2), \dots, f(x_n, \theta_n)\big)$ to take all the related descriptors into consideration. So, the key point is to find the appropriate function f and function g.

3.1. The construction of the common function f

According to the fact that the four descriptors in Fig. 1 have a similar distribution, thus we can choose one of the descriptors as an example to show how to construct the function common f. In Fig. 2, for each value of Δr, a series of solid solubility values are mapped. So, we calculated the average values of the solid solubility for each Δr, and the average values were marked as blue stars in Fig. 2, as the expected value of the solid solubility. Then we constructed the common function f with unknown parameters to fit the blue stars. Based on the function f, a nonlinear least-squares method is used to determine the values of the parameters for each descriptor. The nonlinear least-squares method is used to fit a set of observations with a model, that is nonlinear in unknown parameters. The basis of the method is to approximate the model by a linear one and to refine the parameters by successive iterations. The nonlinear least-squares method in this work use the trust region reflective (RFT) algorithm to perform minimization.

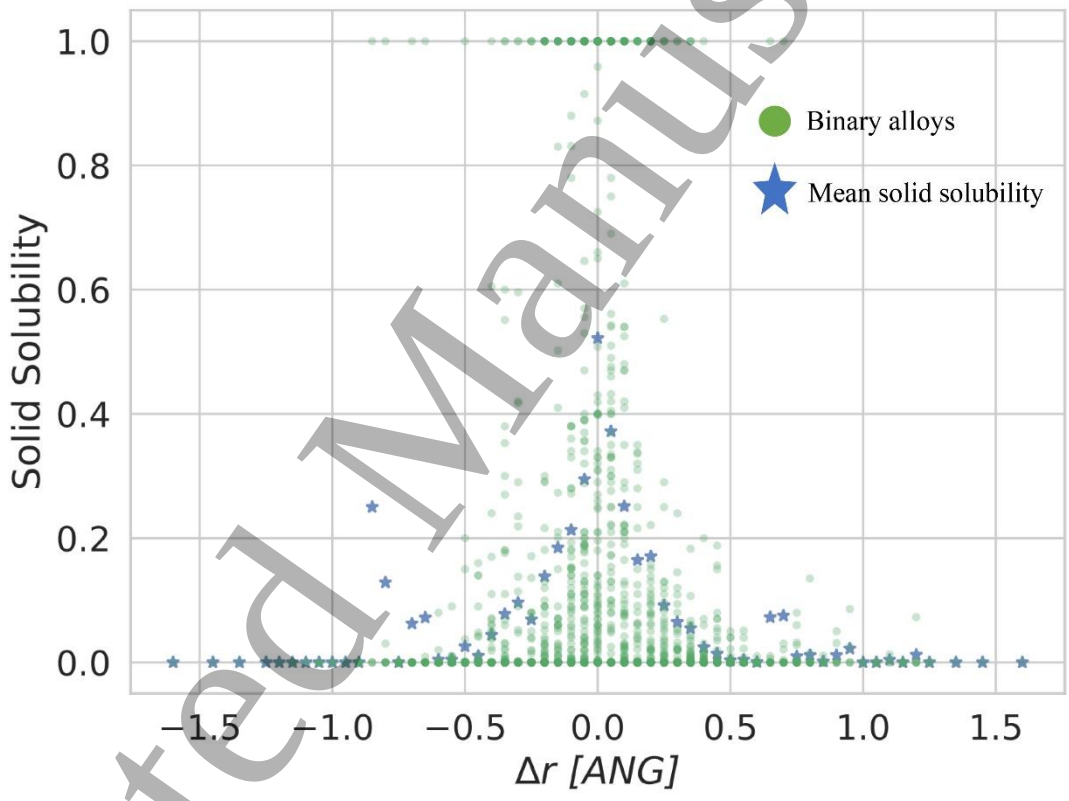

Fig. 2. Distribution of radius descriptor and solid solubility. The radius descriptor represents the difference of atom radius between solute and solvent in whole 1843 binary alloy systems. The blue stars represent the mean solid solubility of the alloys which share the same radius descriptor.

3.2. The construction of the integration function g

As shown in Fig.3, we constructed a DNN as the function g. A DNN is an artificial neural network (ANN) with multiple layers between the input and output layers [34]. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a nonlinear relationship, and its network moves through the layers calculating the probability of each output. The network constructed in this work contains two hidden layers to integrate the multiple functions. In each hidden layer, there is also a dropout layer to avoid overfitting and a rectified linear unit (ReLU) layer to capture nonlinear relationships

[35,36]. Optimization routines is used to search for a good combination of the DNN's hyperparameter values, including the number of neurons and the dropout percentage in the first and second hidden layers. The Adam algorithm is used as the optimizer in our model. DNN diagram in Fig. 3 is a simplified representation. In reality, the number of neurons in first hidden layer is 64 and in second hidden layer is 128.

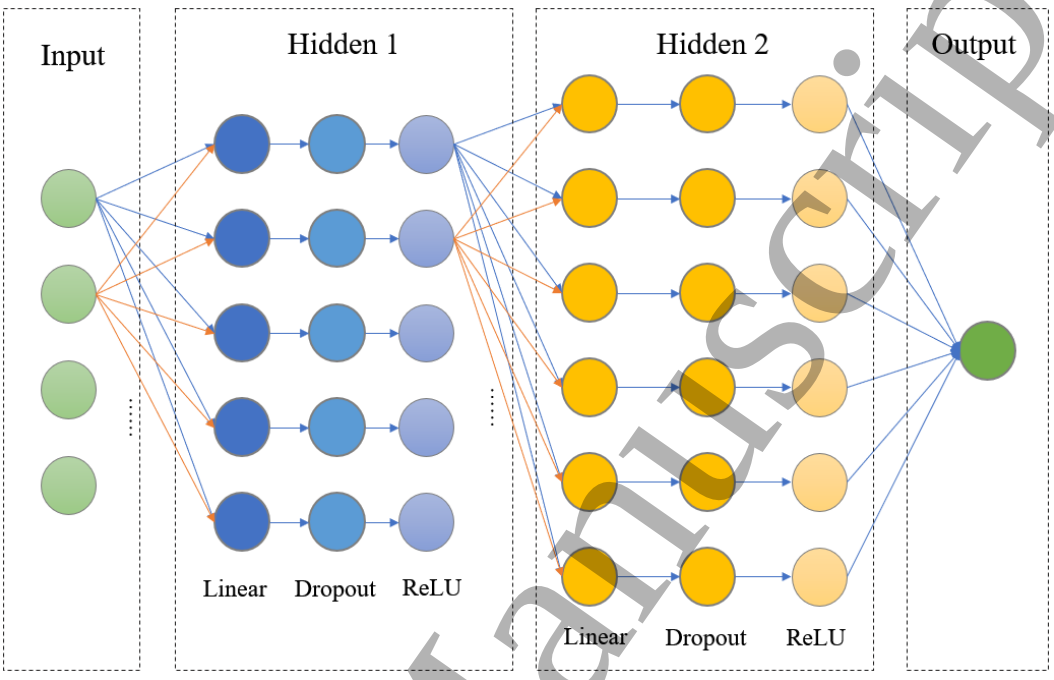

Fig. 3. The neural network structure is a simplified representation. The hidden layer 1 has 64 nodes and the hidden layer 2 has 128 nodes.

4. Result and discussion

Firstly, all the data is used to fit a function f, and the result is expressed in the form of equation and plotted in feature-target map. Then, the fitting process is combined with DNN model to test the accuracy of this method. In step 2, the fitting process only uses the training set instead of all the data. In our experiment, 20% of the data is divided into test set, and the remaining data is used for 4-fold cross validation. In each iteration of cross-validation, 75% of the remaining data is used as training set and 25% is used as validation set. The selection of the test set and the data split in cross validation both use stratified sampling to eliminate errors caused by data imbalances.

4.1. The result of the function f fitting

To represent the distribution between individual descriptors and solid solubility, the function f must conform to at least two characteristics: when x approaches 0, $f(x)$ approaches a constant, and when x approaches infinity, y approaches 0. A basic function can be used to meet the above requirements:

$$f(x) = e^{-|x|} \qquad (1)$$

In order to apply the function to different descriptors, it must have these characteristics:

1) The constant can be adjusted;
2) The curvature of the curve can be adjusted;

3) The function can be asymmetric.
So, the function f can be modified to

$$f(x) = \begin{cases} ae^{bx}, & x < 0 \\ ae^{-cx}, & x \geq 0 \end{cases} \qquad (2)$$

In experiments, we also made a nonlinear transformation on y to make the function f fits the data better. The transformation is

$$y' = e^{y} - 1 \qquad (3)$$

Then we can obtain the final function *f*:

$$f(x) = \begin{cases} log(1 + ae^{bx}), & x < 0 \\ log(1 + ae^{-cx}), & x \geq 0 \end{cases} \qquad (4)$$

Using the nonlinear least-squares method, the unknown parameters can be determined. In this method, the bounds of the parameters are defined as [0, ∞] and the initial values of the parameters are set to 1. At last, we got the functions as follows. The relationship between radius descriptor and solid solubility is represented as:

$$f(x)_{\Delta r} = \begin{cases} log(1 + 0.60180946e^{7.33871449x}), & x < 0 \\ log(1 + 0.60180946e^{-7.01234062x}), & x \geq 0 \end{cases}. \qquad (5)$$

We also can get the other three functions below:

$$f(x)_{\Delta MN} = \begin{cases} log(1 + 1.01287113e^{0.15080559x}), & x < 0 \\ log(1 + 1.01287113e^{-0.18376909x}), & x \geq 0 \end{cases}. \qquad (6)$$

$$f(x)_{\Delta FE} = \begin{cases} log(1 + 0.45995616e^{0.2861688x}), & x < 0 \\ log(1 + 0.45995616e^{-0.38705189x}), & x \geq 0 \end{cases}. \qquad (7)$$

$$f(x)_{\Delta \chi} = \begin{cases} log(1 + 0.42070734e^{2.18825612x}), & x < 0 \\ log(1 + 0.42070734e^{-2.80316508x}), & x \geq 0 \end{cases}. \qquad (8)$$

The fitted function curve of the radius descriptor is drawn in Fig. 4a (red line). Most of the blue stars are near the curve, which means that it's possible to represent the relationship between the descriptor and target values with a mathematical expression. However, several blue stars are far from the curve. Because for these points, the solid solubility values used to calculate the average value are not enough, resulting that the average values cannot represent the expected values of the solid solubility. It also means that the function f is inherently resistant to overfitting.

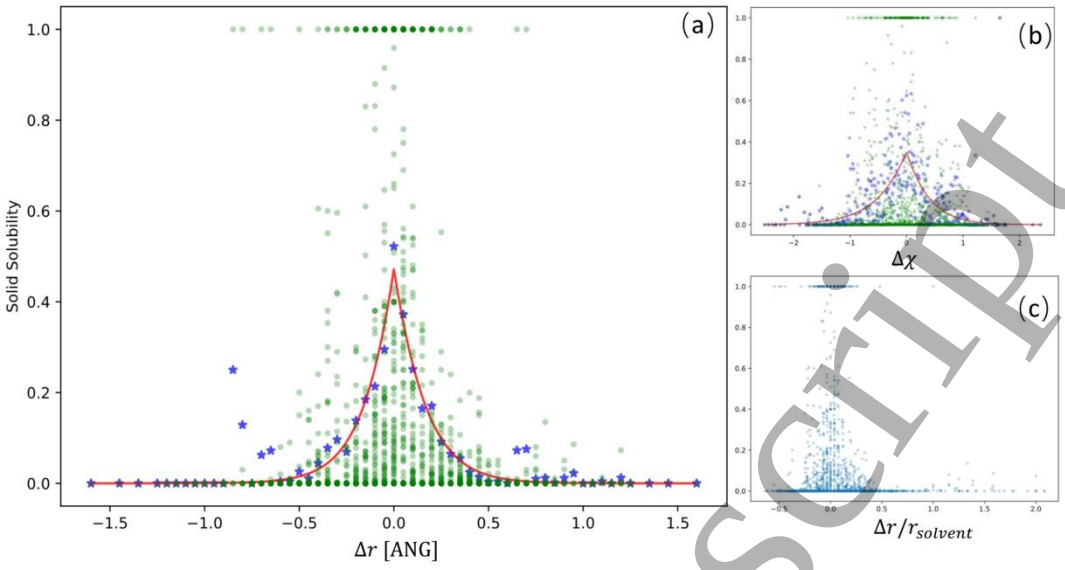

Fig. 4. (a) The function fitting in the distribution of solid solubility and radius descriptor. Blue stars point the mean solid solubility values for each value of radius descriptor and the red plus represent the fitted function curve. (b) The function fitting for electronegativity descriptor. (c) The distribution of solid solubility and radius difference.

In Hume-Rothery's rules, to form substitutional solid solutions, the atomic radius of the solute and solvent atoms must differ by no more than 15%. Obviously, when the rules were proposed, Hume-Rothery only used the alloys of copper and silver. In this kind of alloys, the 15% threshold value of the difference of atomic size has been validated [30]. In addition, the electronegativity descriptor, which was proposed that in the solid solubility system, the electronegativity of the solute and solvent atoms should differ by no more than 0.4[37]. It was only verified in alloy systems of aluminum and magnesium[38]. In this work, a more comprehensive dataset was used. The figures show that there don't exist segmentation points at 15% and 0.4 (Fig. 4b & 4c). Thus, continuity functions are more suitable to represent solid solubility than segmentation points.

It's also found that some descriptors which are not mentioned in Hume-Rothery's rules, such as Mendeleev number and Fermi level electronic number, can affect solid solubility. The Mendeleev number is proposed by D.G. Pettifor as a new chemical scale [39], which has been proven to play an important role in atomic environment prediction[1] and compound former prediction [40]. Such a scale was optimized and simplified in the later work [41,42]. The scale was set up under a rule that the variation of the chemical coordinate within a group does not mix and overlap with neighboring groups. The magnitude of the chemical scale is fixed by requiring it to take the Pauling electronegativity values for Be to F. It means Mendeleev Number includes electronegativity, thus in our model, the former is better than the latter.

4.2. The result of the DNN-QSPR model

In the DNN-QSPR model, a network was constructed to take the functions as inputs and solid solubility as outputs. As comparisons, two more DNN models were built, one takes raw descriptors as inputs and another takes the union of raw descriptors and the functions as inputs. Three models share a similar network structure.

As shown in Fig. 5, the test error of the DNN-QSPR model is lower than that of the DNN model with raw descriptors. And the test error of the DNN model with the union inputs is similar to the DNN-QSPR model. It means that the function f can help the network to capture the intrinsic relationship between individual descriptors

and property, because the accuracy is higher when the network inputs are functions than the inputs are raw descriptors. What's more, the function f can effectively represent the descriptors, because there is almost no increase in accuracy when the network inputs are the union of raw descriptors and functions than only functions. The discrepancy between validation error and test error is small means this model is non-overfitting. Further speaking, this model can avoid overfitting because the function f tries to generalize the distribution of dataset rather than fit with every data point, and the network structure is deliberately designed to combat with overfitting. So, the DNN-QSPR model, which uses mathematical expression to capture statistical characteristics of the descriptor and uses neural network to integrate the result of multiple functions, is useful in the discovery and interpretation of structure-property relationships in materials science.

The same dataset was also used in the other ML models and the results are listed in Table 1. In all these models, the accuracy is the highest when the functions are taken as inputs. The accuracy of the models, which used the functions as inputs, is slightly improved compared to the models using raw descriptors. It should be noted that the accuracy of linear models (Ridge and Lasso) have been greatly improved because it's hard for linear model to capture the complex relationship with raw descriptors. In the KNN model and the SVR model, the MAEs are smaller when using the functions as inputs than using the union of functions and raw descriptors. Because the functions and the raw descriptors are not independent of each other and their correlation affects the performance of these models. The result also shows that the DNN-QSPR model performs best in this comparison.

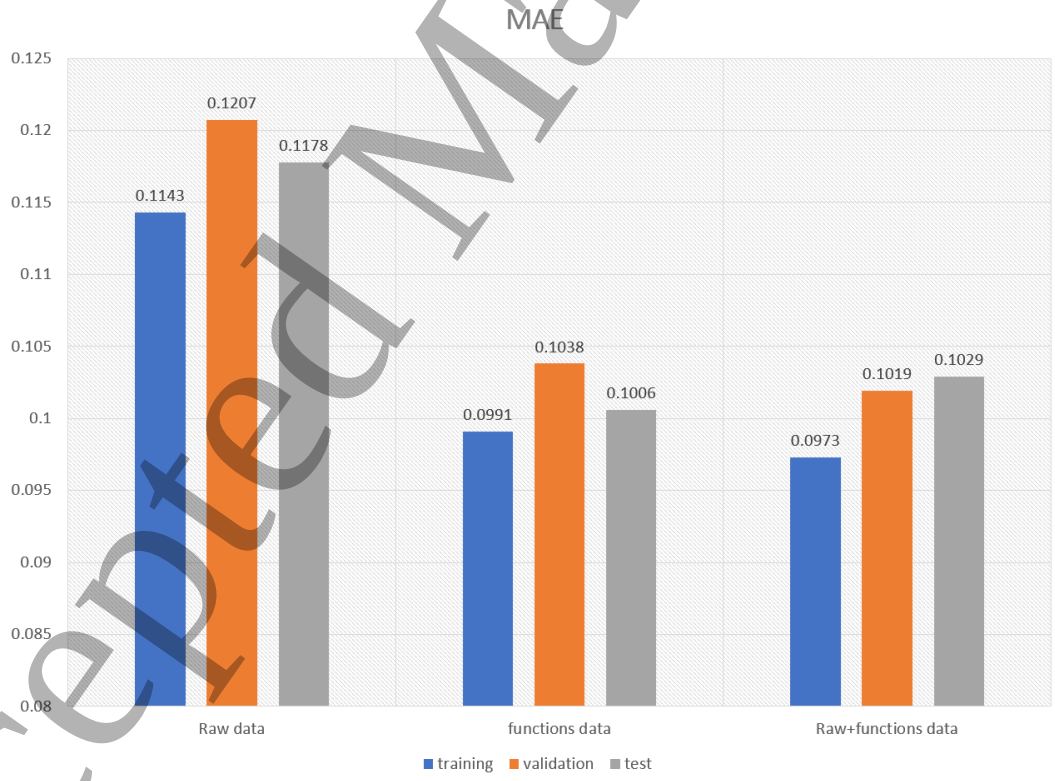

Fig. 5. The MAE in different models. The 'Raw' means the network inputs are raw descriptors, and the 'functions' means the network inputs are the functions obtained in section A. The result in training dataset is marked as blue, in validation dataset is marked as orange and in test dataset is marked as gray.

Table 1

The MAEs of other ML models with different input sets. The 'Raw' means the inputs are raw descriptors, the 'Functions' means the inputs are the functions obtained in section A. The full names of the following models are K Neighbors Regressor, Random Forest Regressor, Decision Tree

Regressor, Support Vector Regression, Ridge Regression, Lasso Regression.

Models	Raw data			Functions data			Raw+Functions data		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
KNN	0.1044	0.1284	0.1242	0.0943	0.1213	0.1141	0.1029	0.1285	0.1267
RF	0.1202	0.1358	0.1397	0.0991	0.1201	0.1191	0.0983	0.1204	0.1216
DT	0.1253	0.1422	0.1423	0.0990	0.1197	0.1238	0.0966	0.1200	0.1272
SVR	0.1345	0.1631	0.1616	0.1273	0.1286	0.1247	0.1321	0.1514	0.1521
Ridge	0.2158	0.2164	0.2151	0.1531	0.1545	0.1527	0.1533	0.1553	0.1536
Lasso	0.2158	0.2164	0.2151	0.1529	0.1543	0.1524	0.1530	0.1550	0.1533

The above results indicate that the constructed functions do effectively represent the original descriptor and can improve the accuracy of the ML models. Moreover, when combined materials science with ML, different models have different performance for the same dataset, and different data processing methods will result in different accuracy in the same ML model. So, it's necessary to find a suitable method for a specific material problem, such as the DNN-QSPR model for the solid solubility of binary alloy systems.

5. Conclusion

In this study, we proposed the DNN-QSPR model. The experiment results show that based on a well-designed neural network, the model performs well in the prediction of solid solubility and it is none-overfitting by itself. Compared to several common ML methods, this model can predict solid solubility with a higher accuracy. Besides, this model analyzes each descriptor separately and then combines them together, considering the interaction between these descriptors. Different from the general machine learning model, it could help the researchers to interpret what leads to the result and find the effect between material structures and properties easily. This work also gives an exploration for representing the relationship between structure and property with a mathematical expression.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No.2018YFB0704400)

References

[1] Villars P, Cenzual K, Daams J, Chen Y and Iwata S 2004 Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB *J. Alloys Compd.* **367** 167–75

[2] Park J H and Jana S C 2003 The relationship between nano- and micro-structures and mechanical properties in PMMA-epoxy-nanoclay composites *Polymer (Guildf).* **44** 2091–100

[3] Bec S, Tonck A, Georges J M, Coy R C, Bell J C and Roper G W 1999 Relationship between mechanical properties and structures of zinc dithiophosphate anti–wear films *Proc. R. Soc. London. Ser. A Math. Phys. Eng. Sci.* **455** 4181–203

[4] Lang A-D, Zhai J, Huang C-H, Gan L-B, Zhao Y-L, Zhou D-J and Chen Z-D 1998 Relationship between Structures and Photocurrent Generation Properties in a Series of Hemicyanine Congeners *J. Phys. Chem. B* **102** 1424–9

[5] Alabort E, Barba D, Sulzer S, Lißner M, Petrinic N and Reed R C 2018 Grain boundary properties of a nickel-based superalloy: Characterisation and modelling *Acta Mater.* **151** 377–94

[6] Belkly A, Helderman M, Karen V L and Ulkch P 2002 New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research and design *Acta Crystallogr. Sect. B Struct. Sci.* **58** 364–9

[7] Gaulton A, Bellis L J, Bento A P, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B and Overington J P 2012 ChEMBL: A large-scale bioactivity

database for drug discovery *Nucleic Acids Res.* **40** 1100–7
[8] Pence H E and Williams A 2010 ChemSpider: An Online Chemical Information Resource *J. Chem. Educ.* **87** 1123–4
[9] Graulis S, Chateigner D, Downs R T, Yokochi A F T, Quirós M, Lutterotti L, Manakova E, Butkus J, Moeck P and Le Bail A 2009 Crystallography Open Database - An open-access collection of crystal structures *J. Appl. Crystallogr.* **42** 726–9
[10] Wolverton C, Meredig B, Saal J E, Aykol M and Kirklin S 2013 Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD) *Jom* **65** 1501–9
[11] Nelson L J, Yang K, Mingo N, Setyawan W, Curtarolo S, Taylor R H, Levy O, Xue J, Hart G L W, Wang S, Buongiorno-Nardelli M and Sanvito S 2012 AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations *Comput. Mater. Sci.* **58** 227–35
[12] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera R S, Gold-Parker A, Vogt L, Brockway A M and Aspuru-Guzik A 2011 The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid *J. Phys. Chem. Lett.* **2** 2241–51
[13] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 Commentary: The materials project: A materials genome approach to accelerating materials innovation *APL Mater.* **1**
[14] Collings E 1975 *Physics of Solid Solution*
[15] Tan Y, Li J, Tang Z, Wang J and Kou H 2018 Design of high-entropy alloys with a single solid-solution phase: Average properties vs. their variances *J. Alloys Compd.* **742** 430–41
[16] Ouyang Y, Zhang Z, Li D, Chen J and Zhang G 2019 Emerging Theory, Materials, and Screening Methods: New Opportunities for Promoting Thermoelectric Performance *Ann. Phys.* **1800437** 1800437
[17] Pilania G, Liu X-Y and Wang Z 2019 Data-enabled structure–property mappings for lanthanide-activated inorganic scintillators *J. Mater. Sci.*
[18] Amabilino S, Bratholm L A, Bennie S J, Vaucher A C, Reiher M and Glowacki D R 2019 Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality 1–28
[19] Rajak P, Kalia R K, Nakano A and Vashishta P 2018 Neural Network Analysis of Dynamic Fracture in a Layered Material
[20] Ward L, O'Keeffe S C, Stevick J, Jelbert G R, Aykol M and Wolverton C 2018 A machine learning approach for engineering bulk metallic glass alloys *Acta Mater.* **159** 102–11
[21] Xue D, Balachandran P V., Hogden J, Theiler J, Xue D and Lookman T 2016 Accelerated search for materials with targeted properties by adaptive design *Nat. Commun.* **7** 11241
[22] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547–55
[23] Bolis G, Di Pace L and Fabrocini F 1991 A machine learning approach to computer-aided molecular design *J. Comput. Aided. Mol. Des.* **5** 617–28
[24] Zheng W and Tropsha A 2000 Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle *J. Chem. Inf. Comput. Sci.* **40** 185–94
[25] Liu H, Yao X, Zhang R, Liu M, Hu Z and Fan B 2005 Accurate quantitative structure-property relationship model to predict the solubility of C60 in various solvents based on a novel approach using a least-squares support vector machine *J. Phys. Chem. B* **109** 20565–71
[26] Golmohammadi H, Dashtbozorgi Z and Acree W E 2012 Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine *Eur. J. Pharm. Sci.* **47** 421–9
[27] Fernandez M, Boyd P G, Daff T D, Aghaji M Z and Tom K 2014 Machine Learning Virtual Screening for Rapid and Accurate Recognition of High Performing MOFs for CO2 Capture *J Phys Chem Lett* **Submitted** 1087–92
[28] Anon 1992 *Volume 3: Alloy Phase Diagrams (Asm Handbook)*
[29] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis *Comput. Mater. Sci.* **68** 314–9
[30] Li S, Zhang H, Dai D, Ding G and Wei X 2019 Study on the factors affecting solid solubility in binary alloys : An exploration by Machine Learning *J. Alloys Compd.* **782** 110–8
[31] Hume-Rothery W, Mabbott G W and Evans K M C 1935 Errata: The Freezing Points, Melting Points, and Solid Solubility Limits of the Alloys of Silver and Copper with the Elements of the B Sub-Groups *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **234** 0–0
[32] Hume-Rothery W 1966 Atomic diameters, atomic volumes and solid solubility relations in alloys *Acta Metall.* **14** 17–20
[33] Yu L and Liu H 2004 Efficient Feature Selection via Analysis of.pdf *J. Mach. Learn. Res.* **5** 1205–24
[34] Lecun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[35] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: A simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[36] Ito Y 1991 Representation of functions by superpositions of a step or sigmoid function and their

applications to neural network theory *Neural Networks* **4** 385–94
[37] Darken L S and GURRY R W Physical chemistry of metals
[38] Alonso J A and Simozar S 1980 Prediction of solid solubility in alloys *Phys. Rev. B* **22** 5583–9
[39] Pettifor D G 1984 A chemical scale for crystal-structure maps *Solid State Commun.* **51** 31–4
[40] Villars P, Brandenburg K, Berndt M, LeClair S, Jackson A, Pao Y H, Igelnik B, Oxley M, Bakshi B, Chen P and Iwata S 2001 Binary, ternary and quaternary compound former/nonformer prediction via Mendeleev number *J. Alloys Compd.* **317**–**318** 26–38
[41] Pettifor D G and Podloucky R 1985 Pettifor and Podloucky respond *Phys. Rev. Lett.* **55** 261–261
[42] Pettifor D G 1988 Structure maps for. Pseudobinary and ternary phases *Mater. Sci. Technol.* **4** 675–91