Science Bulletin 64 (2019) 1195–1203

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Science Bulletin
www.scibull.com

Article

Predicting the onset temperature (T_g) of Ge_xSe_{1-x} glass transition: a feature selection based two-stage support vector regression method

Yue Liu [a,b], Junming Wu [a], Guang Yang [c,*], Tianlu Zhao [a], Siqi Shi [c,d,*]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[b] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[c] School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China
[d] Materials Genome Institute, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Article history:
Received 29 May 2019
Received in revised form 18 June 2019
Accepted 26 June 2019
Available online 2 July 2019

Keywords:
Onset temperature of glass transition
Machine learning
Support vector machine

ABSTRACT

Despite the usage of both experimental and topological methods, realizing a rapid and accurate measurement of the onset temperature (T_g) of Ge_xSe_{1-x} glass transition remains an open challenge. In this paper, a predictive model for the T_g in Ge_xSe_{1-x} glass system is presented by a machine learning method named feature selection based two-stage support vector regression (FSTS-SVR). Firstly, Pearson correlation coefficient (PCC) is used to select features highly correlated with T_g from the candidate features of Ge_xSe_{1-x} glass system. Secondly, in order to simulate the two-stage characteristic of T_g which is caused by structural variation with a turning point at $x = 0.33$ via the structural analysis, SVR is utilized to build predictive models for two stages separately and then the two achieved models are synthesized using a minimum error based model for T_g prediction. Compared with the topological and other methods based on SVR, the FSTS-SVR gives the highest predictive accuracy with the root mean square error (*RMSE*) and mean absolute percentage error (*MAPE*) of 10.64 K and 2.38%, respectively. This method is also expected to be more efficient for the prediction of T_g of other glass systems with the multi-stage characteristic.

1. Introduction

Ge_xSe_{1-x} glass system has already been extensively studied due to its glasses possessing excellent transparent property in the far-infrared wavelengths range and considered as model binary covalent glasses [1–6]. The influence of temperature on its properties is a key issue, particularly for onset temperature of glass transition (T_g) at which the amorphous material begins converting between viscous or rubbery and glassy states. The T_g is of primary importance not only at various processing stages, but also to estimate the stability of glass parts in service. Furthermore, the T_g of Ge_xSe_{1-x} glass system covers a wide range from 316 to 683 K [7,8]. However, when measuring the T_g of Ge_xSe_{1-x} glass system, the usages of diverse measurement methods and preparation methods always result in different degrees of deviation. To this end, it is of great significance to study the T_g of Ge_xSe_{1-x} glass system.

Compared with that of As_xSe_{1-x} glass system, the T_g of Ge_xSe_{1-x} glass system presents a more apparent two-stage characteristic (in Fig. 1) because of its internal structural variation with the change of component ratio x [2,8–12]. It is widely accepted that the structure of Ge_xSe_{1-x} glass system varies from one type to another obviously and continuously at a turning point ($x = 1/3$) when x increases from 0 to 0.42 [5,6]. The variation tendencies of the two stages are quite different between each other. Therefore, a single stage method cannot accurately capture the two-stage characteristic. In view of this situation, a topological method [7] has been presented for the T_g prediction of Ge_xSe_{1-x} glass system by a two-stage division. As a typical example of theoretical derivation method, it combines the concept of temperature-dependent constraints with the Adam–Gibbs model of viscosity. T_g of a composition of x can be computed from the composition dependence of the topological constraints $n(T_g(x), x)$, viz. $T_g(x)/T_g(x_R) = (d - n(T_g(x_R), x_R))/(d - n(T_g(x), x))$, where d is network dimension, and $n(T_g(x_R), x_R)$ is topological constraints of an initial composition x_R with $T_g(x_R)$ in the range of 0–0.4. As shown in Fig. 1, the topological method firstly takes the composition of $x_R = 0$ as the initial point for calculating values at other points, then divides the T_g prediction process of Ge_xSe_{1-x} glass system into two stages at $x = 0.33$ (more details can be seen in Section 2.1) depending on material structural variation, and finally calculates T_g at each stage. In addition, the topological method has been extended

* Corresponding authors.
E-mail addresses: guangyang@shu.edu.cn (G. Yang), sqshi@shu.edu.cn (S. Shi).

https://doi.org/10.1016/j.scib.2019.06.026
2095-9273/

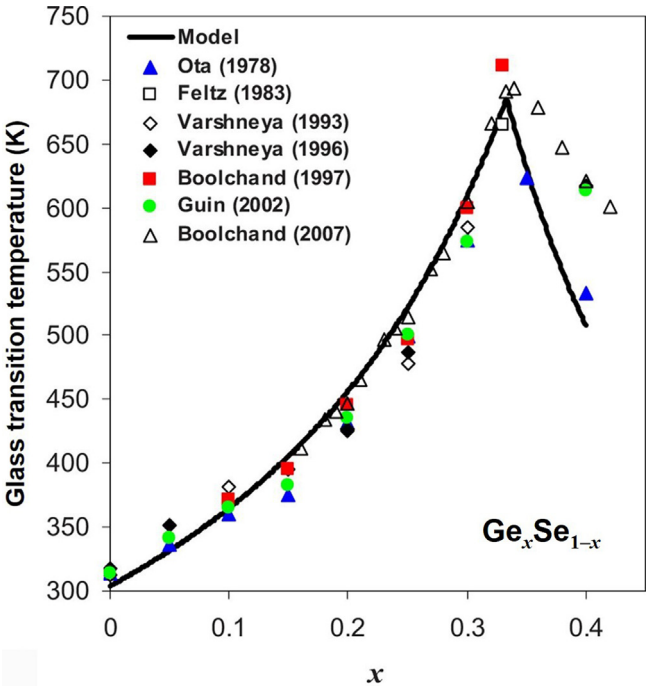

Fig. 1. (Color online) Predicted result (line) and experimental data (scattered symbols) of T_g in the Ge_xSe_{1-x} glass system using the topological model and from previous reports, respectively [7].

to predict the T_g of borate, silicate, phosphate and other glass systems [13–21], such as xNa$_2$O(1–x)B$_2$O$_3$, xLi$_2$O(1–x)B$_2$O$_3$, Na$_2$O-B$_2$O$_3$-SiO$_2$, Na$_2$O-CaO-B$_2$O$_3$ and Na$_2$O-B$_2$O$_3$-P$_2$O$_5$. The topological method constructs the theoretical formula just based on little experimental data, but it tends to have a large absolute prediction error in some composition ranges, particularly in Ge-rich compositions of Ge_xSe_{1-x} glass system [7]. Furthermore, the main shortcomings of the topological method involve a series of complex domain knowledge and poor universal applicability for all types of glasses because of the limited knowledge of glassy structure.

Machine learning is one field of computer science that enables computers to learn without being explicitly programmed [22]. Recently, machine learning has proved its superiority in time efficiency, universal applicability and predictive accuracy of material property prediction, new materials discovery and various other purposes [23–29]. At present, various machine learning methods (e.g., support vector regression (SVR), artificial neural network (ANN), multiple linear regression (MLR), and back-propagation neural network (BPNN)) have been successfully applied for the prediction of macroscopic and microscopic properties [30–35], including the T_g of various kinds of organic glasses. Chen et al. [36] adopted MLR and BPNN to simulate the relationship between the T_g of three classes of vinyl polymers and four condition attributes (rigidness, chain mobility, the molecular average polarizability and the net charge) of the most negative atom. Results indicate that the predicted T_g values are in good agreement with the experimental values. Liu and Cao [37] also used MLR and BPNN to predict the T_g of 113 polymers with *RMSE* of 17.53 K for testing data. Pei et al. [38] employed the particle swarm optimization (PSO) to optimize the SVR for the T_g prediction of random copolymers. They proved that both the predictive accuracy and generalization ability of SVR are superior to those of the quantitative structure-property relationship (QSPR) model [39]. In 2013, Pei et al. [40] used the particle swarm optimization-support vector regression (PSO-SVR) method to predict the T_g of three classes of vinyl polymers, and found that PSO-SVR achieves better performance than the spectral structure-activity relationship analysis and ANN. Furthermore, Alzghoul et al. [41] used several machine learning methods including MLR, partial least-squares, principal component regression, ANN and SVR, for the T_g prediction of drugs, among which SVR gives the best result with *RMSE* of 18.7 K. Above all, machine learning based methods have achieved good prediction performance for T_g values of many kinds of organic glasses, but few references on using machine learning methods to predict the T_g of inorganic glasses are located. Therefore, it is of much practical significance to predict the T_g of Ge_xSe_{1-x} glass system using machine learning methods. Firstly, it is a key step to select an appropriate algorithm or model in the construction of a machine learning system for a specific task, as it will greatly affect the predictive accuracy and generalization ability [42]. As one of the most widely used algorithms, SVR can efficiently solve problems (e.g., small set of samples, nonlinearity, and high dimensionality) with support vectors and kernel functions. Therefore, in our study, SVR is finally selected as the machine learning method to predict the T_g of Ge_xSe_{1-x} glass system. Secondly, the characteristic on the T_g of Ge_xSe_{1-x} glass system needs to be taken into account for higher predictive accuracy. In our previous work, the feasibility and practicality of one-stage methods (ridge regression, SVR and BPNN) for predicting the T_g of As_xSe_{1-x} glass system have been verified [43]. However, the T_g of Ge_xSe_{1-x} glass system presents the more apparent two-stage characteristic than that of As_xSe_{1-x} glass system, implying that one-stage methods do not work well for predicting T_g of Ge_xSe_{1-x} glass system. Thus, an ideal model is adopted for the two-stage characteristic.

Besides the two-stage characteristic, the construction and selection of condition attributes are also of great importance. The T_g of Ge_xSe_{1-x} glass system is affected by many condition attributes and therefore these attributes should be firstly provided by related knowledge or experiment. Then, among these attributes, the ones highly correlated with T_g need to be selected by a feature selection method, enabling a more perfect explanation of T_g property. As one of the critical aspects of machine learning, feature selection (FS) can discover the most relevant attributes from initial condition attributes, so that the dimension of the dataset can be reduced while the description power maintains the same level [44]. Alzghoul et al. [41] reduced the number of descriptors from 284 to 43 by correlation coefficient and then used several machine learning methods to predict the T_g of drugs. Yu [45] used the mean of MLR to pick out four descriptors from 1,664 descriptors and then employed SVR to predict the T_g of polymers. Lu et al. [46] employed the gradient boosting regression (GBR) algorithm to screen out the 14 most important features from the initial 30 features, and then accurately predicted the bandgaps of thousands of hybrid organic-inorganic perovskites (HOIPs). Li et al. [47] used stability selection [48] recursive feature elimination (RFE) [49], and univariate feature selection based on mutual information [50] to select the top 70 features from the initial 791 features and then constructed more accurate models without significant overfitting. In summary, FS can not only rank the importance of material features and discover the most relevant features affecting material properties, but also greatly improve the predictive performance of the model. Thus, a feature selection method is also required to further assess the importance of condition attributes summarized by domain expert knowledge and then more accurately predict the T_g of Ge_xSe_{1-x} glass system.

Concerning with the above two problems, the FSTS-SVR method of Ge_xSe_{1-x} glass system is proposed. Compared with other FS methods, such as RFE, GBR, mutual information and so on, PCC can measure the degree of correlation between two continuous variables in a fast manner [51], which is more suitable for Ge_xSe_{1-x} glass system with small data volume and continuous features. Therefore, PCC is firstly used for the construction of sample dataset

to eliminate the condition attributes having low correlation with T_g. Learning from the two-stage idea of the topological method, the turning point of structural variation in the Ge_xSe_{1-x} glass system is recognized at $x = 0.33$ firstly via the structural analysis. Bounded by this turning point, the variation of T_g can be divided into two stages at which the variation tendencies are different. Then SVR is employed to build models at two stages separately without complex domain knowledge, and the models of two stages can be synthesized through the minimum error based model combination. Finally, the FSTS-SVR method is used to predict the T_g of Ge_xSe_{1-x} glass system with high predictive accuracy when the data volume is small.

2. Materials, methods and calculation procedure setup

2.1. The structural analysis of Ge_xSe_{1-x} glass system

As shown in Fig. 2, typical structural units of Ge_xSe_{1-x} glass system are dependent on composition ratio x: they are Se_n chain unit and $GeSe_4$ tetrahedral unit when $x < 0.33$; almost only $GeSe_4$ tetrahedral unit when $x = 0.33$, and Ge_nGeSe_{4-n} tetrahedral unit and GeSe crystal-like or Ge-rich domain unit when $x > 0.33$ [2,8–12]. By analyzing the available data correlating with structural characteristics, an obvious turning point emerges at $x = 0.33$.

2.2. The collection of experimental data

Referring to the analysis of internal structure and component in relevant reports [7,8,52], the condition attributes, which influence the T_g of Ge_xSe_{1-x} glass system, are summarized into nine groups of data (s_1–s_9) as shown in Table 1, including component ratio x, average coordination number $\langle r\rangle$, Poisson's ratio v, bulk modulus K, mean atomic volume V_0, mean experimental atomic bonding energy U_{0ex} and mean theoretical bonding energy B [8]. $\langle r\rangle$ is calculated according to the component ratio and coordination number of elements. Both v and K are derived from classical elasticity relationships. V_0 is derived from density and mean atomic mass. U_{0ex} can be calculated from the first Grüneisen rule with V_0 and K, and B can be obtained from the known interatomic bonding energy ($U_{Ge\text{-}Ge}$, $U_{Ge\text{-}Se}$, $U_{Se\text{-}Se}$) on the basis of simple structural assumption [8]. Given that the previous T_g values predicted by topological model are well consistent with both dynamic [4,53] and thermodynamic ones [7,13,15,18,20,54,55], and the usual T_g values measured by calorimetry are based on thermodynamics, we adopt the experimental values measured by differential scanning calorimetry (DSC).

2.3. The choice of machine learning method

Support vector machine (SVM), which was proposed by Cortes and Vapnik [56] in 1995, is a statistical learning approach based on the structural risk minimization principle and has been widely applied in materials machine learning [57]. When SVM is applied for regression problem by introducing nonlinear kernel functions, it is called SVR [58]. Given that compared with other methods (e.g., MLR, BPNN), SVR is capable of solving nonlinear problems with small data volume and high dimension [59,60]. Consequently, SVR is selected to predict the T_g of Ge_xSe_{1-x} glass system with only nine groups of sample data in Section 2.2. The schematic plot of SVR is shown in Fig. 3 and its main building process of predictive model is described as follows.

Given the dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where x_i is condition attribute, y_i is decision attribute and m is the number of training samples, we set the regression function $f(x) = \omega \cdot \varphi(x) + b$, where ω is autoregressive coefficient or weight vector, b is error value and $\varphi(x)$ is a nonlinear mapping function to map x into a higher dimensional feature space to conduct a linear regression in the feature space. Then ω and b can be estimated by minimizing the regularized risk function $R(C)$ as Eq. (1).

$$\min R(C) = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{m} l_\varepsilon(f(x_i) - y_i), \tag{1}$$

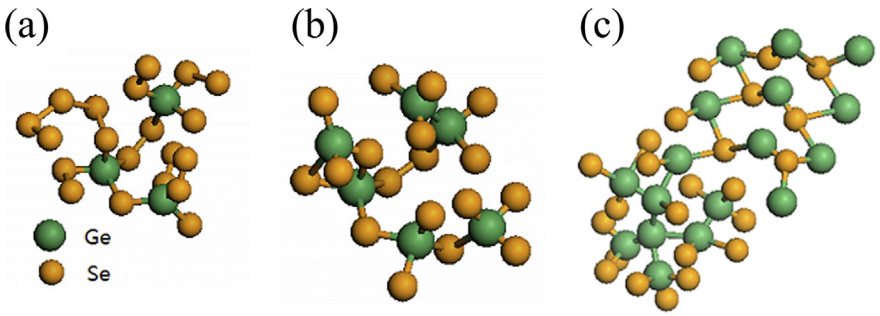

Fig. 2. (Color online) Typical structural units of Ge_xSe_{1-x} glass system. (a) $x = 0.2$, (b) $x = 0.33$ and (c) $x = 0.4$.

Table 1
Nine groups of condition attributes and T_g of Ge_xSe_{1-x} glass system.

	x	$\langle r\rangle$	v (±0.005)	K (±0.01 GPa)	V_0 (±0.05 cm^3 mol^{-1})	U_{0ex} (±0.1 kJ mol^{-1})	B (±0.1 kJ mol^{-1})	T_g (±1 K)
s_1	0.00	2.00	0.347	10.59	18.45	195.5	184.0	316
s_2	0.10	2.20	0.317	11.02	18.08	199.3	214.8	363
s_3	0.20	2.40	0.298	11.91	17.82	212.3	245.6	444
s_4	0.25	2.50	0.291	12.43	17.78	221.0	261.0	509
s_5	0.33	2.67	0.278	12.52	18.06	226.0	286.6	683
s_6	0.36	2.72	0.265	13.92	17.96	250.0	287.8	673
s_7	0.38	2.76	0.259	14.90	17.72	264.0	288.6	646
s_8	0.40	2.80	0.253	15.90	17.41	276.7	289.4	615
s_9	0.42	2.84	0.249	16.91	17.11	289.3	290.2	594

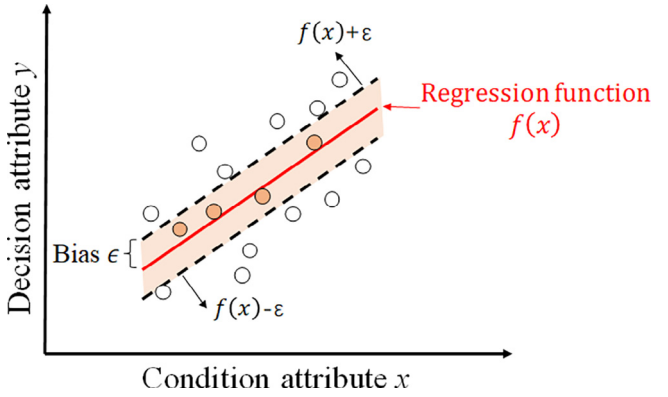

Fig. 3. (Color online) The schematic plot of SVR. The orange circle indicates the sample point as close as possible to the hyperplane $f(x)$, and the hollow circle represents the sample point as close as possible to two segmentation planes $f(x) \pm \varepsilon$ with bias ε from the hyperplane. SVR is to minimize the total deviation of all sample points from the hyperplane.

$$s.t.\ l_\varepsilon(f(x_i)-y_i)=\begin{cases}0, & |f(x_i)-y_i|<\varepsilon,\\ |f(x_i)-y_i|-\varepsilon, & |f(x_i)-y_i|\geqslant\varepsilon,\end{cases} \tag{2}$$

where $\frac{1}{2}\| \omega^2 \|$ is used as a measurement of function flatness, C is regularization constant, l_ε is ε-insensitive loss function and ε is the bias between $f(x)$ and y.

By introducing slack variables ξ_i and ξ_i^* to deal with the data points which dissatisfy Eq. (2), Eqs. (1) and (2) are converted into a primal problem $R(\omega,\ \xi_i,\ \xi_i^*)$:

$$\min R(\omega,\xi_i,\xi_i^*)=\frac{1}{2}||\omega||^2+C\sum_{i=1}^{m}(\xi_i+\xi_i^*), \tag{3}$$

$$s.t.\ f(x_i)-y_i\leqslant\varepsilon+\xi_i,\ y_i-f(x_i)\leqslant\varepsilon+\xi_i^*,\ \xi_i,\xi_i^*\geqslant 0, i=1,2,...,m. \tag{4}$$

In order to obtain ω and b in $f(x)$, Lagrangian multiplier a_i and a_i^* are introduced and the dual problem of SVR can be described as:

$$Q=\frac{1}{2}\sum_{i,j=1}^{m}(a_i^*-a_i)(a_j^*-a_j)K(x_i,x)+\varepsilon\sum_{i=1}^{m}(a_i^*+a_i)-\sum_{i=1}^{m}(a_j^*+a_i), \tag{5}$$

$$s.t.\ \sum_{i=1}^{m}(a_i^*-a_i)=0,\ 0\leqslant a_i,a_i^*\leqslant C, \tag{6}$$

where $K(x_i, x) = \varphi(x_i)\cdot\varphi(x)$ is a kernel function. The resulting regression function $f(x)$ can be described as:

$$f(x)=\sum_{i=1}^{m}(a_i^*-a_i)K(x_i,x)+b. \tag{7}$$

2.4. FSTS-SVR method

2.4.1. Main idea and steps

The main idea of the FSTS-SVR method involves: (1) extracting condition attributes closely related to T_g by PCC; (2) recognizing the turning point via the structural analysis because the T_g of Ge_x-Se_{1-x} glass system exhibits the two-stage characteristic caused by the structural variation; (3) using SVR to build model at each stage separately because of the different variation tendencies of two stages.

Given the original dataset $S_0 = \{s_i | s_i = (X_i, y_i),\ i \in [1, n]\}$, where s_i represents the i^{th} sample, $X_i = (x_1, x_2, \ldots, x_m)$ is the vector of m condition attributes, y_i is the decision attribute and n is the number of samples, the detailed process of the FSTS-SVR method in Fig. 4 is described as follows:

Stage 1: using PCC to extract condition attributes closely related to decision attribute and then constructing sample data $S = \{s_1, s_2, \ldots, s_n\}$.
Stage 2: recognizing the turning point of structural variation by structural analysis and then getting the t^{th} point as turning point T.
Stage 3: depending on T, dividing the dataset S into two subsets, $S_1 = \{s_1, \ldots, s_t\}$ and $S_2 = \{s_t, \ldots, s_n\}$, which represent two stages respectively.
Stage 4: using SVR to build models on S_1 and S_2 separately, $M_{1\text{-SVR}} \rightarrow S_1$ and $M_{2\text{-SVR}} \rightarrow S_2$.
Stage 5: choosing the model which has the minimum error compared with the experimental value at T and combining the models of two stages.
Stage 6: predicting the unknown T_g through the composite model. For each testing sample X_i, judging which stage it belongs to and using the corresponding model M to predict, $y_{\text{predict}} = M(X_i)$.

2.4.2. Sample construction

The better a sample dataset constructed by feature selection is, the higher the predictive accuracy of a machine learning method can obtain. In Table 1, the correlations between condition attributes (x, $\langle r\rangle$, v, K, V_0, U_{0ex}, and B) and decision attribute (T_g) of Ge_x-Se_{1-x} glass system have been investigated in previous works [7,8,52]. Note that one or some condition attributes may have low correlations with T_g. In order to improve the quality of sample data and then to improve the predictive performance, the correlation between each condition attribute and T_g is analyzed by PCC. PCC is defined as:

$$\rho=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})}}, \tag{8}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values of x and y, respectively, and the coefficient ρ ranges from −1 to 1. $\rho > 0$ and $\rho < 0$ mean that the variables are positive and negative correlations, respectively. The closer $|\rho|$ is to 1, the higher correlation degree between variables is.

2.4.3. Turning point recognition and stage division

The internal structural variation rules decide the turning point, leading to the variation of T_g. If the prior knowledge of internal structural variation of glass system has been comprehended, its turning point T can be determined and the sample data can be divided into two subsets $S_1 = \{s_1, \ldots, s_t\}$ and $S_2 = \{s_t, \ldots, s_n\}$, representing two stages. As mentioned in Section 2.1, since the 5th point ($x = 0.33$) of Ge_xSe_{1-x} glass system has been determined as the turning point, the sample data in Table 1 can be divided into $S_1 = \{s_1, \ldots, s_5\}$ and $S_2 = \{s_5, \ldots, s_9\}$.

2.4.4. T_g prediction model built for two stages

The prediction process of T_g using SVR at each stage is shown in Fig. 5, in which X_i represents the vector of condition attributes of subsets S_1 or S_2, and $K(X, X_m)$ represents kernel function, where m is number of kernel products. Models of S_1 and S_2 are built separately as Eqs. (9) and (10), which are derived from Eq. (7).

$$M_{1\text{-SVR}}(X_i)=\sum_{i=1}^{m}(a_i^*-a_i)K(X_i,X)+b,\ X_i\in S_1, \tag{9}$$

$$M_{2\text{-SVR}}(X_i)=\sum_{i=1}^{m}(a_i^*-a_i)K(X_i,X)+b,\ X_i\in S_2. \tag{10}$$

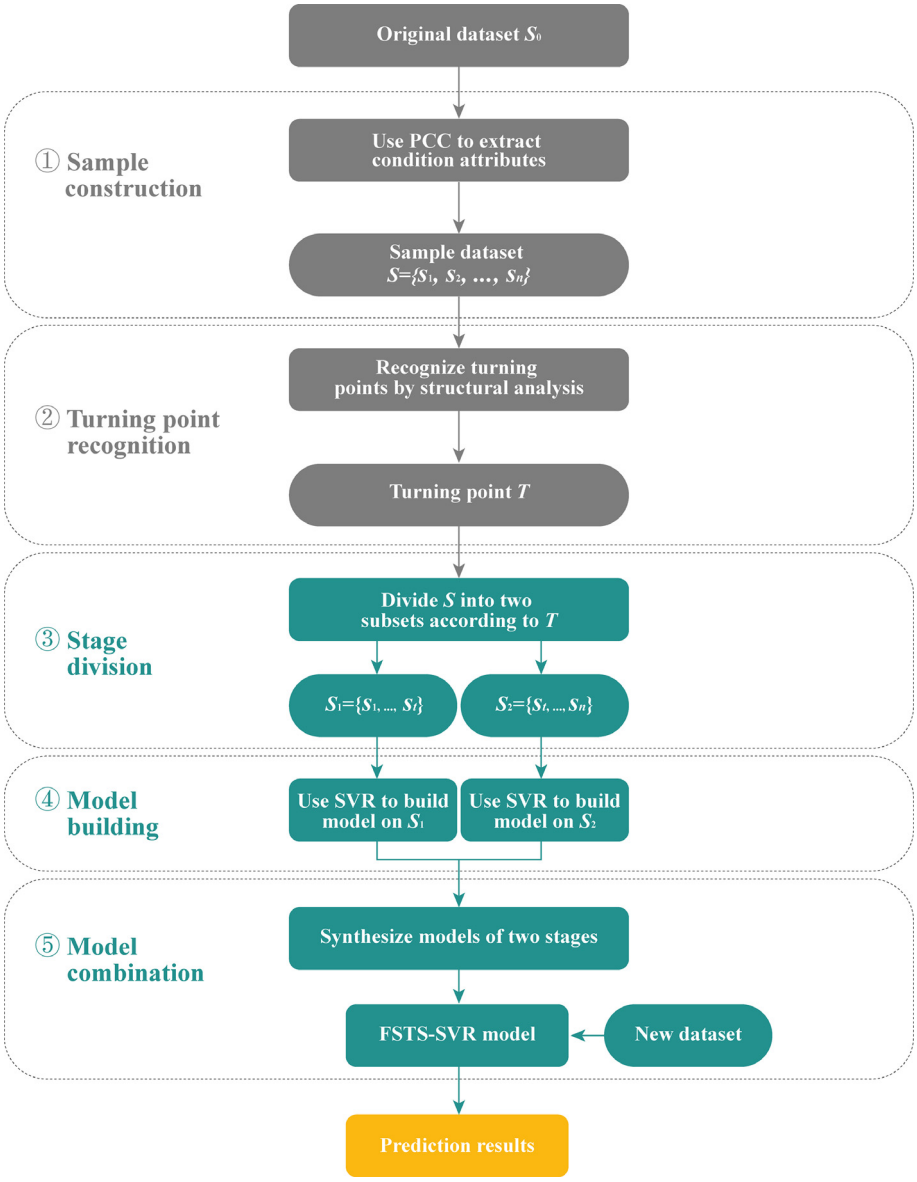

Fig. 4. (Color online) The main process of FSTS-SVR. The FSTS-SVR mainly consists of five key steps: Sample construction; Turning point recognition; Stage division; Model building; and Model combination.

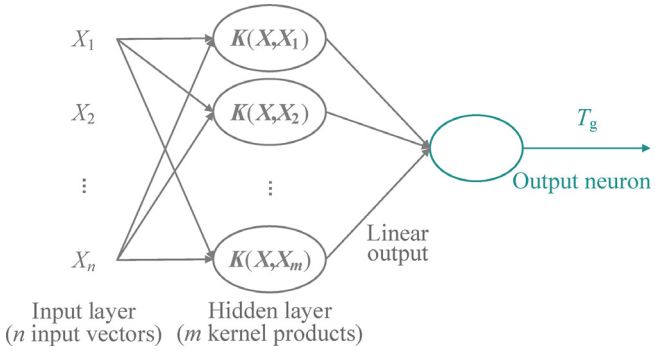

Fig. 5. (Color online) The T_g prediction process using SVR at each stage.

2.4.5. Minimum error based model combination

Given that T_g at the turning point exhibits the two-stage characteristic, the errors between the experimental value and two predicted values at the turning point obtained by $M_{1\text{-SVR}}$ and $M_{2\text{-SVR}}$ are calculated. The model with the minimum error will be chosen, and the composite model can be formulated as Eq. (11).

$$M=\begin{cases} M_{1\text{-SVR}}(X_i), & X_i \in S_1, i \neq t, \\ \min E(M_{1\text{-SVR}}(X_i), M_{2\text{-SVR}}(X_i)), & i=t, \\ M_{2\text{-SVR}}(X_i), & X_i \in S_2, i \neq t, \end{cases} \tag{11}$$

where $M_{1\text{-SVR}}$ and $M_{2\text{-SVR}}$ are the models built for two stages, and $\min E(M_{1\text{-SVR}}(X_i), M_{2\text{-SVR}}(X_i))$ represents the model whose predicted value has the minimum error compared with the experimental value.

2.5. Calculation procedure setup

In order to predict the T_g of Ge_xSe_{1-x} glass system and evaluate the prediction effectiveness of stage division and feature selection, besides FSTS-SVR, SVR, two-stage SVR (TS-SVR) and feature selection based SVR (FS-SVR) are also conducted. Moreover, detailed

comparisons on T_g values are made among these methods, topological method and experiment.

A library for support vector machines (LIBSVM) is adopted as the SVR calculation program and implemented by MATLAB [61], in which the RBF (Gaussian) function is used as the kernel function and grid search is used to optimize hyper-parameters of the model, *C* (Penalty parameter *C* of the error term) and *gamma* (Kernel coefficient for the kernel function RBF). As shown in Fig. 6, since the data volume is only 9, leaving-one-out cross-validation (LOOCV) is adopted as the validation method in the model constructions of SVR to partition the training dataset (a set of examples used to fit the parameters of the model) and testing dataset (a set of examples used only to assess the performance of a fully specified model). LOOCV can accurately evaluate the generalization performance of the machine learning model and reduce the risk of overfitting, especially in the case of small sample sizes [62]. In each iteration, eight samples are partitioned as the training data and one sample independent of the training data is partitioned as the testing data and a total of nine iterations are conducted respectively. Especially, in the 5th iteration calculation of TS-SVR and FSTS-SVR, s_5 is incapable of being taken as the testing data because the 5th point ($x = 0.33$) as s_5 is the turning point. Thus the 5th iteration calculation is unexecuted and other eight iteration calculations are performed. For the convenience of comparison, although s_5 can be used as testing data in SVR and FS-SVR, eight iteration calculations are also considered. Moreover, owing to glassy structural variation at the 5th point, the experimental T_g of this point is taken as the initial point ($x_R = 0.33$) in the topological method for calculating values at other points [7]. The mean predictive accuracy of testing results in each calculation can be taken as the final result.

The *RMSE* and *MAPE* are used to evaluate predictive accuracy and defined as:

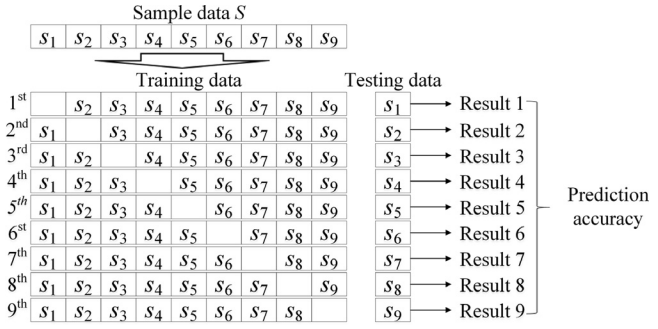

Fig. 6. The schematic plot of LOOCV.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n}}, \tag{12}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i'}{y_i}\right|, \tag{13}$$

where y_i and y_i' represent the i^{th} experimental and predicted values, respectively. The lower the *RMSE* and *MAPE* are, the higher predictive accuracy can obtain.

3. Results and discussion

3.1. T_g of Ge_xSe_{1-x} glass system predicted using SVR and TS-SVR

Table 2 shows the T_g values of Ge_xSe_{1-x} glass system predicted using the SVR and TS-SVR methods based on the dataset of seven condition attributes without feature selection. For the convenience of comparison, results obtained by the topological method [7,13] are also listed. And Fig. 7 shows the composition dependence of T_g of Ge_xSe_{1-x} glass system obtained using experimental and five prediction methods. It is seen from Table 2 and Fig. 7 that T_g values predicted using topological method fit the experimental data effectively when $x \leq 0.33$ with an average absolute prediction error of 9.1 K, but poorly when $x > 0.33$ with an average absolute prediction error of 101.2 K. By contrast, results obtained using SVR are in better agreement with the experimental data at the whole stage. The maximum and minimum absolute prediction errors are 47.7 and 3.1 K, respectively. Notably, when $x > 0.33$, absolute prediction errors using SVR reduce by almost a magnitude with an average absolute prediction error of 14.3 K, which could be caused by the inherent limitation of structural variation assumption in the topological method [7]. Note that the absolute prediction errors at some points ($x = 0$, 0.2 and 0.4) using SVR are still somewhat large. On the other hand, although the absolute prediction errors using TS-SVR further reduce when $x > 0.33$, they are comparatively large when $x < 0.33$ (especially $x = 0.2$) compared with that using SVR. Therefore, it is difficult to compare the predictive accuracies of five methods exactly and clearly by analyzing the errors at various composition values.

Instead, we can evaluate the overall predictive accuracies of five methods by analyzing *RMSE* and *MAPE* as shown in Fig. 8. The *RMSE* and *MAPE* of T_g obtained using SVR are reduced by 66.81% ($|RMSE_{SVR} - RMSE_{topological}|/RMSE_{topological} \times 100\%$) and 54.82% ($|MAPE_{SVR} - MAPE_{topological}|/MAPE_{topological} \times 100\%$) respectively relative to these of T_g obtained using the topological method. More importantly, TS-SVR exhibits a higher predictive accuracy, in which *RMSE* and *MAPE* can reach 13.98 K and 3.12% respectively.

Table 2
Comparison between experimental and predicted results of T_g obtained using topological, SVR, TS-SVR, FS-SVR and FSTS-SVR methods for the testing data.

		1st	2nd	3rd	4th	5th	6th	7th	8th	9th
	X	0	0.1	0.2	0.25	0.33	0.36	0.38	0.4	0.42
	Exp. T_g (K)	316	363	444	509	683	673	646	615	594
Predicted T_g (K)	Topological	303.6	364.3	455.3	520.4	683	599.1	548.6	506.0	469.4
	SVR	268.3	359.9	472.4	519.5	683	657.6	654.4	640.5	601.7
	TS-SVR	335.6	373.4	475.4	488.2	683	661.6	645.7	623.3	603.7
	FS-SVR	305.0	339.8	458.4	531.0	683	650.5	649.9	637.1	609.1
	FSTS-SVR	324.0	385.7	468.0	507.7	683	661.4	645.7	623.6	602.5
Absolute prediction error (K)	Topological	12.4	1.3	11.3	11.4	0	73.9	97.4	109.0	124.6
	SVR	47.7	3.1	28.4	10.5	0	15.4	8.4	25.5	7.7
	TS-SVR	19.6	10.4	31.4	20.8	0	11.4	0.3	8.3	9.7
	FS-SVR	11.0	23.2	14.4	22.0	0	22.5	3.9	22.1	15.1
	FSTS-SVR	8.0	22.7	24.0	1.3	0	11.6	0.3	8.6	8.5

*Exp. = Experimental.

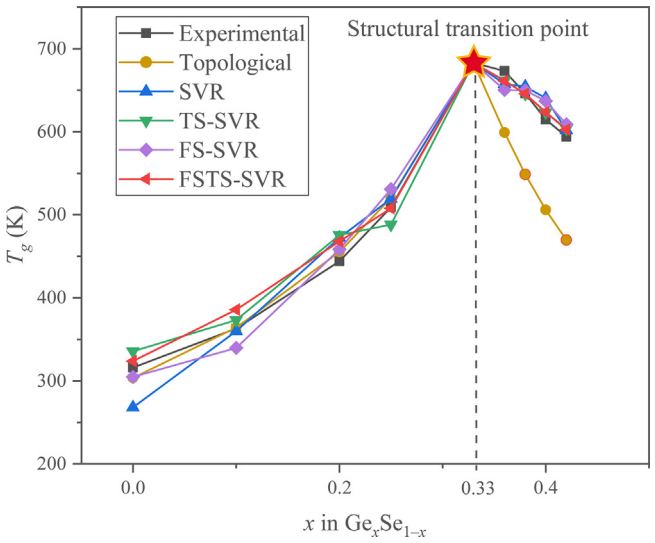

Fig. 7. (Color online) Composition dependence of T_g in the Ge_xSe_{1-x} glass system obtained using experimental, topological, SVR, TS-SVR, FS-SVR, and FSTS-SVR methods.

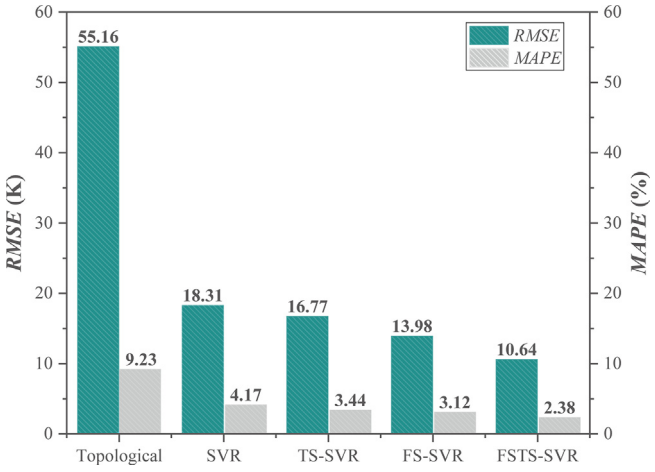

Fig. 8. (Color online) *RMSE* and *MAPE* results of T_g obtained using five methods for the testing data.

Above all, compared with the topological method, machine learning methods are found to achieve higher predictive accuracy for the T_g values of Ge_xSe_{1-x} glass system. Moreover, the consideration of stage division further increases the predictive accuracy. In the next subsection, FSTS-SVR is introduced to model and analyze the T_g values of Ge_xSe_{1-x} glass system to validate the effect of feature selection.

3.2. T_g of Ge_xSe_{1-x} glass system predicted using FS-SVR and FSTS-SVR

As mentioned in Sections 2.1 and 2.2, x can be directly used for the composition control, and $\langle r\rangle$ linearly depends on x and coordination of each element via the equation $\langle r\rangle = 2 + 2x$ in the binary Ge_xSe_{1-x} glass system, so that $\langle r\rangle$ can be considered as a redundant attribute. The high correlation also exists among K, V_0 and U_{0ex} since U_{0ex} can be calculated from the first Grüneisen rule with V_0 and K. Meanwhile, Table 3 exhibits the correlation coefficients ρ between each condition attribute and T_g of Ge_xSe_{1-x} glass system calculated by Eq. (8). It is seen that x, $\langle r\rangle$, K, U_{0ex} and B have positive correlations with T_g while v and V_0 have negative correlations with T_g in all eight iteration calculations. Hereinto, V_0 has the minimum $|\rho|$, which means that V_0 has the weakest correlation with T_g. This is consistent with the fact that the T_g directly relates to the net link energy, while V_0 depends on compaction and volume of atoms in each composition [7,63]. Consequently, the poor correlation between the net link energy and compaction or volume of atom results in the unclear relationship between T_g and V_0. Therefore, considering the redundancy among condition attributes and the correlation between condition attributes and T_g, $\langle r\rangle$ and V_0 are both removed from the dataset in Section 2.2, and a new sample dataset containing five condition attributes (x, v, K, U_{0ex} and B) is constructed for calculation.

In Table 2, the T_g values of Ge_xSe_{1-x} glass system on the dataset of five condition attributes predicted using the FS-SVR and FSTS-SVR methods are also listed. And Fig. 7 shows the composition dependence of T_g of Ge_xSe_{1-x} glass system obtained using experimental, SVR, FS-SVR, TS-SVR and FSTS-SVR methods. As shown in Table 2, absolute prediction errors using FS-SVR are smaller and larger at four points (x = 0, 0.2, 0.38 and 0.4) and four other points (x = 0.1, 0.25, 0.36 and 0.42) respectively compared with that using SVR. On the other side, when $x < 0.33$, the average absolute prediction error using FSTS-SVR is lower by 26.2 K than that using TS-SVR; when $x > 0.33$, absolute prediction errors using FSTS-SVR and TS-SVR are close to each other, and the average value of the corresponding differences is within 0.7 K, which is acceptable or can be ignored. Note that the larger dataset is expected to result in the higher predictive accuracy of FSTS-SVR.

Fig. 8 shows *RMSE* and *MAPE* values obtained using SVR, FS-SVR, TS-SVR and FSTS-SVR methods for the testing data. As for the averaged predictive accuracy shown in Fig. 8, the *RMSE* and *MAPE*obtained using FS-SVR are reduced by 8.41% and 17.51% relative to that obtained using SVR, while the *RMSE* and *MAPE* obtained using FSTS-SVR are reduced by 23.89% and 23.7% relative to that obtained using TS-SVR, respectively. It is clearly seen that the consideration of the feature selection (FS-SVR, FSTS-SVR) causes higher predictive accuracy. Furthermore, given the stage division, FSTS-SVR gives the highest predictive accuracy among all methods, suggesting that compared with seven condition attributes (x, $\langle r\rangle$, v, K, V_0, U_{0ex} and B), five condition attributes (x, v, K, U_{0ex} and B) enable their relationship with T_g to be more accurately established.

Table 3
The correlation coefficients between condition attributes and T_g.

	x	$\langle r\rangle$	v	K	V_0	U_{0ex}	B
1st	0.884	0.889	−0.821	0.603	−0.205	0.655	0.952
2nd	0.909	0.913	−0.873	0.644	−0.444	0.678	0.948
3rd	0.927	0.930	−0.890	0.689	−0.514	0.720	0.963
4th	0.927	0.930	−0.890	0.707	−0.502	0.740	0.961
6th	0.932	0.935	−0.893	0.727	−0.593	0.748	0.961
7th	0.921	0.924	−0.880	0.684	−0.492	0.715	0.955
8th	0.930	0.933	−0.893	0.701	−0.460	0.736	0.958
9th	0.951	0.954	−0.916	0.785	−0.542	0.805	0.965

4. Conclusion

Relative to the topological method, single-stage SVR can predict the T_g of Ge_xSe_{1-x} glass system well with *RMSE* and *MAPE* reduced by 66.81% and 54.82%, respectively. Furthermore, when the structural variation at $x = 0.33$ is considered, the corresponding TS-SVR method exhibits a higher predictive accuracy than single-stage SVR with *RMSE* and *MAPE* reducing from 18.31 to 13.98 K and 4.17% to 3.12%, respectively. More importantly, once PCC is used for sample dataset construction, the resultant FSTS-SVR method can further explore the inherent relationship between T_g and condition attributes of Ge_xSe_{1-x} glass system. By comparing the *RMSE* and *MAPE* obtained using among the topological, SVR, FS-SVR, TS-SVR and FSTS-SVR methods, FSTS-SVR with feature selection and stage division is found to achieve the highest predictive accuracy. Notably when $x > 0.33$, the predictive accuracy almost has been improved up to a magnitude. Consequently, our proposed FSTS-SVR method is expected to be more suitable for predicting the T_g of other glass systems with the multi-stage characteristic.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (2017YFB0701500 and 2017YFB0701600), the National Natural Science Foundation of China (51602187, U1630134, 11874254 and 51622207) and the Shanghai Municipal Education Commission (14ZZ099 and QD2015028). All the computations were performed on the high performance computing platform provided by the High Performance Computing Center of Shanghai University.

Author contributions

Yue Liu, Guang Yang and Siqi Shi conceived and designed the theoritical calculations and experiments. Yue Liu, Junming Wu, Tianlu zhao and Siqi Shi carried out all the theoritical calculations and analyzed the data. Guang Yang performed and alayzed all the experiments. All authors discussed the reults and prepared and revised the manuscript.

References

[1] Boolchand P, Feng X, Bresser W. Rigidity transitions in binary Ge–Se glasses and the intermediate phase. J Non-Cryst Solids 2001;293:348–56.
[2] Bureau B, Troles J, Le Floch M, et al. Germanium selenide glass structures studied by ^{77}Se solid state NMR and mass spectroscopy. J Non-Cryst Solids 2003;319:145–53.
[3] Eggleton BJ, Luther-Davies B, Richardson K. Chalcogenide photonics. Nat Photonics 2011;5:141.
[4] Gueguen Y, Rouxel T, Gadaud P, et al. High-temperature elasticity and viscosity of Ge_xSe_{1-x} glasses in the transition range. Phys Rev B 2011;84:064201.
[5] Mawale R, Halenkovič T, Bouška M, et al. Laser desorption ionization time-of-flight mass spectrometry of Ge_xSe_{1-x} chalcogenide glasses, their thin films, and Ge: Se mixtures. J Non-Cryst Solids 2019;509:65–73.
[6] Sharma D. Fragility and cooperative dynamics correlations in Ge_xSe_{1-x} chalcogenide glasses. J Non-Cryst Solids 2019;516:67–70.
[7] Gupta PK, Mauro JC. Composition dependence of glass transition temperature and fragility. I. A topological model incorporating temperature-dependent constraints. J Chem Phys 2009;130:094503.
[8] Yang G, Gueguen Y, Sangleboeuf J-C, et al. Physical properties of the Ge_xSe_{1-x} glasses in the $0<x<0.42$ range in correlation with their structure. J Non-Cryst Solids 2013;377:54–9.
[9] Boolchand P, Bresser WJ. The structural origin of broken chemical order in $GeSe_2$ glass. Philos Mag Part B 2000;80:1757–72.
[10] Edwards TG, Sen S, Gjersing EL. A combined ^{77}Se NMR and Raman spectroscopic study of the structure of Ge_xSe_{1-x} glasses: towards a self consistent structural model. J Non-Cryst Solids 2012;358:609–14.
[11] Lucas P, King EA, Gulbiten O, et al. Bimodal phase percolation model for the structure of Ge-Se glasses and the existence of the intermediate phase. Phys Rev B 2009;80:214114.
[12] Shatnawi MTM, Farrow CL, Chen P, et al. Search for a structural response to the intermediate phase in Ge_xSe_{1-x} glasses. Phys Rev B 2008;77:094134.
[13] Mauro JC, Gupta PK, Loucks R. Composition dependence of glass transition temperature and fragility. II. A topological model of alkali borate liquids. J Chem Phys 2009;130:234503.
[14] Smedskjaer MM, Mauro JC, Yue Y. Prediction of glass hardness using temperature-dependent constraint theory. Phys Rev Lett 2010;105:115503.
[15] Smedskjaer MM, Mauro JC, Sen S, et al. Quantitative design of glassy materials using temperature-dependent constraint theory. Chem Mater 2010;22:5358–65.
[16] Smedskjaer MM, Mauro JC, Youngman RE, et al. Topological principles of borosilicate glass chemistry. J Phys Chem B 2011;115:12930–46.
[17] Mauro JC. Topological constraint theory of glass. Am Ceram Soc Bull 2011;90:31.
[18] Jiang Q, Zeng H, Liu Z, et al. Glass transition temperature and topological constraints of sodium borophosphate glass-forming liquids. J Chem Phys 2013;139:124502.
[19] Hermansen C, Mauro JC, Yue Y. A model for phosphate glass topology considering the modifying ion sub-network. J Chem Phys 2014;140:154501.
[20] Jiang Q, Zeng H, Li X, et al. Tailoring sodium silicophosphate glasses containing SiO_6-octahedra through structural rules and topological principles. J Chem Phys 2014;141:124506.
[21] Smedskjaer MM. Topological model for boroaluminosilicate glass hardness. Front Mater 2014;1:1–6.
[22] Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 1959;3:210–29.
[23] Shi S, Gao J, Liu Y, et al. Multi-scale computation methods: their applications in lithium-ion battery research and development. Chin Phys B 2016;25:018212.
[24] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater 2016;4:053208.
[25] Hill J, Mulholland G, Persson K, et al. Materials science with large-scale data and informatics: unlocking new opportunities. MRS Bull 2016;41:399–409.
[26] Jain A, Hautier G, Ong SP, et al. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. J Mater Res 2016;31:977–94.
[27] Kalidindi SR, Medford AJ, McDowell DL. Vision for data and informatics in the future materials innovation ecosystem. JOM 2016;68:2126–37.
[28] Ward L, Wolverton C. Atomistic calculations and materials informatics: a review. Curr Opin Solid St M 2017;21:167–76.
[29] Liu Y, Zhao T, Ju W, et al. Materials discovery and design using machine learning. J Materiomics 2017;3:159–77.
[30] Häse F, Valleau S, Pyzer-Knapp E, et al. Machine learning exciton dynamics. Chem Sci 2016;7:5139–47.
[31] Majid A, Khan A, Choi T-S. Predicting lattice constant of complex cubic perovskites using computational intelligence. Comput Mater Sci 2011;50:1879–88.
[32] Hansen K, Montavon G, Biegler F, et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. J Chem Theory Comput 2013;9:3404–19.
[33] Butler KT, Davies DW, Cartwright H, et al. Machine learning for molecular and materials science. Nature 2018;559:547–55.
[34] Wu Y-J, Fang L, Xu Y. Predicting interfacial thermal resistance by machine learning. npj Comput Mater 2019;5:56.
[35] Chandrasekaran A, Kamal D, Batra R, et al. Solving the electronic structure problem with machine learning. npj Comput Mater 2019;5:22.
[36] Chen X, Sztandera L, Cartwright HM. A neural network approach to prediction of glass transition temperature of polymers. Int J Intell Syst 2008;23:22–32.
[37] Liu W, Cao C. Artificial neural network prediction of glass transition temperature of polymers. Colloid Polym Sci 2009;287:811–8.
[38] Pei JF, Cai CZ, Tang JL, et al. Prediction of the glass transition temperatures of styrenic copolymers by using support vector regression combined with particle swarm optimization. J Macromol Sci B 2012;51:1437–48.
[39] Yu X, Wang X, Wang H, et al. Prediction of the glass transition temperatures of styrenic copolymers using a QSPR based on the DFT method. J Mol Struct (Thoechem) 2006;766:113–7.
[40] Pei JF, Cai CZ, Zhu YM. Modeling and predicting the glass transition temperature of vinyl polymers by using hybrid PSO-SVR method. J Theory Comput Chem 2013;12:1350002.
[41] Alzghoul A, Alhalaweh A, Mahlin D, et al. Experimental and computational prediction of glass transition temperature of drugs. J Chem Inf Model 2014;54:3396–403.
[42] Bishop CM. Pattern recognition and machine learning. Springer; 2006.
[43] Liu Y, Zhao T, Yang G, et al. The onset temperature (T_g) of As_xSe_{1-x} glasses transition prediction: a comparison of topological and regression analysis methods. Comput Mater Sci 2017;140:315–21.
[44] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[45] Yu X. Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers. Fiber Polym 2010;11:757–66.
[46] Lu S, Zhou Q, Ouyang Y, et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat Commun 2018;9:3405.

[47] Li W, Jacobs R, Morgan D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. Comput Mater Sci 2018;150:454–63.
[48] Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc B Stat Methodol 2010;72:417–73.
[49] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
[50] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E 2004;69:066138.
[51] Lee Rodgers J, Nicewander WA. Thirteen ways to look at the correlation coefficient. Am Stat 1988;42:59–66.
[52] Sreeram AN, Varshneya AK, Swiler DR. Molar volume and elastic properties of multicomponent chalcogenide glasses. J Non-Cryst Solids 1991;128:294–309.
[53] Angell CA. Formation of glasses from liquids and biopolymers. Science 1995;267:1924–35.
[54] Ji X, Zeng H, Li X, et al. High glass transition temperature barium silicophosphate glasses designed with topological constraint theory. J Am Ceram Soc 2016;99:1255–8.
[55] Micoulaut M, Yue Y. Material functionalities from molecular rigidity: Maxwell's modern legacy. MRS Bull 2017;42:18–22.
[56] Cortes C, Vapnik V. Support-vector networks. Mach learn 1995;20:273–97.
[57] Raccuglia P, Elbert KC, Adler PDF, et al. Machine-learning-assisted materials discovery using failed experiments. Nature 2016;533:73–6.
[58] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.
[59] Eskidere Ö, Ertaş F, Hanilçi C. A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Syst Appl 2012;39:5523–8.
[60] Zhou H, Zhao JP, Zheng LG, et al. Modeling NOx emissions from coal-fired utility boilers using support vector regression with ant colony optimization. Eng Appl Artif Intel 2012;25:147–58.
[61] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. Acm T Intel Syst Tec 2011;2:1–27.
[62] Kohavi R. A study of cross-validation and bBootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on artificial intelligence 1995;2:1137–1143.
[63] Yang G, Bureau B, Rouxel T, et al. Correlation between structure and physical properties of chalcogenide glasses in the As_xSe_{1-x} system. Phys Rev B 2010;82:195206.

Yue Liu finished her Ph.D. degree in control theory and control engineering from Shanghai University (SHU) in 2005. She joined the School of Computer Engineering and Science of SHU as an assistant professor in early 2000 and as an associate professor in 2008 until now. From Sep. 2012 to Sep. 2013, she worked as a visiting scholar at the University of Melbourne. Her current research interests focus on machine learning and its applications in materials science.

Guang Yang finished his Ph.D. degree in chemistry of metrials from University of Rennes 1 (France) in 2012. And then, he spent 2 years as postdoc at ICMCB CNRS and ISM of University of Bordeaux in France. After that, he joined Shanghai University as an assistant professor in early 2014 and as an associate professor in 2017 until now. His current research interests focus on the optical glasses and glass-ceramics.

Siqi Shi obtained his B.S. and M.S. degrees from Jiangxi Normal University in 1998 and in 2001, respectively. He finished his Ph.D. degree from IOP, CAS in 2004. After that, he joined National Institute of Advanced Industrial Science and Technology of Japan and Brown University of USA as a postdoctor until joining Shanghai University as a professor in 2013. His current research interests focus on multiscale calculation of electrochemical energy storage materials and application of machine learning in materials science.