Computational Materials Science 140 (2017) 315–321

Contents lists available at ScienceDirect

Computational Materials Science

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

The onset temperature (T_g) of As_xSe_{1-x} glasses transition prediction: A comparison of topological and regression analysis methods

Yue Liu [a], Tianlu Zhao [a], Guang Yang [b,*], Wangwei Ju [a], Siqi Shi [b,c,*]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[b] School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China
[c] Materials Genome Institute, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Article history:
Received 9 June 2017
Received in revised form 2 September 2017
Accepted 4 September 2017

Keywords:
Onset temperature of glass transition
Topological method
Regression analysis methods

ABSTRACT

As_xSe_{1-x} glasses are promising candidates as matrix for mid-infrared applications, but it is usually invasive, costly and time-consuming or even impossible to measure the onset temperature (T_g) of glass transition of each composition in the system for glass preparation and fiber processing by experimental methods. In this paper, topological and regression analysis (ridge regression, support vector regression and back-propagation neural network) methods are used to predict the T_g of As_xSe_{1-x} glass system and compared with each other. The topological method predicts the T_g of As_xSe_{1-x} glass system by composition dependence of quantitative structure, although its calculation range is limited in the composition range of $0 \le x \le 0.5$ due to no enough knowledge of quantitative structures and their variation. In contrast, regression analysis methods can model the relationships between physical attributes and T_g without complex domain knowledge, thus extending the calculation range to $x = 0.6$ and achieving much higher prediction accuracy. Among them, back-propagation neural network achieves the highest prediction accuracy with an *RMSE* of 1.21 K (7.87 K) and *MAPE* of 0.33% (1.96%) for training data (testing data). Significantly, a three-attribute correlation equation based on ridge regression is obtained, possessing much higher prediction accuracy than that of the topological method.

1. Introduction

Glass transition temperature is a temperature range at which the amorphous material is converted between viscous or rubbery and glassy states. As an important physical property, its accurate acquisition is not only one of the frontier problems of kinetic and thermodynamics, but also an important and difficult issue of condensed matter physics theory [1,2]. As_xSe_{1-x} glasses are promising candidates as matrix for mid-infrared applications, such as a step-index fiber for mid-infrared supercontinuum generation light sources [3], and it is of great significance to study its T_g for glass preparation and fiber processing. Although the T_g of glass sample can be measured by different experimental methods, such as differential scanning calorimetry (DSC), differential thermal analysis (DTA), thermal expansion and the high temperature elastic moduli of glass [4], it is invasive, costly, and time-consuming or even impossible to measure the T_g of each composition ($0 \le x \le 0.6$) in As_xSe_{1-x} glass system.

To overcome the disadvantages of experimental methods, theoretical derivation methods in which the variation of glass structure is combined with its change rule of microscopic attributes in form of the theoretical formulas, such as the modified Gibbs–DiMarzio equation [5], are carried out for T_g prediction. Recently, as an important theoretical derivation method, the topological method has been extended from the initial Phillips–Thorpe theory to quantitatively account for the effect of temperature, integrating the concept of temperature-dependent constraints with the Adam–Gibbs model of viscosity [6–8]. It is successful in predicting the T_g and fragility of the basic binary Ge_xSe_{1-x} glass system [6]. Furthermore, due to the advantages of simplicity and the ability to obtain concise analytical expressions for the scaling of macroscopic properties with composition, it has been extended to predict the T_g, fragility and hardness of borate, silicate, phosphate and other glass systems [8–16], such as $xNa_2O(1-x)B_2O_3$, $xLi_2O(1-x)B_2O_3$, Na_2O-B_2O_3-SiO_2, Na_2O-CaO-B_2O_3 and Na_2O-B_2O_3-P_2O_5. In this way, not only can the topological method be used for predicting the glass properties, but also optimizing the glass production process [7]. However, no application on the topological model in the prediction of macroscopic properties of the basic binary As_xSe_{1-x} glass system. One possible reason is that the coordination numbers of

* Corresponding authors at: School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China (S. Shi).
E-mail addresses: guangyang@shu.edu.cn (G. Yang), sqshi@shu.edu.cn (S. Shi).

http://dx.doi.org/10.1016/j.commatsci.2017.09.008
0927-0256/

arsenic and germanium are 3 and 4, resulting in the two-dimensional and three-dimensional network structures of As_xSe_{1-x} and Ge_xSe_{1-x} glass systems, respectively. Therefore, it is necessary to verify the prediction effectiveness and applicability of the topological method in the T_g prediction of As_xSe_{1-x} glass system.

Different from theoretical derivation methods, regression analysis methods in machine learning predict a continuous-valued attribute associated with a target object via a regression model relating output variable y to a function of input variables X, $y = f(X)$. What's more, regression analysis methods not only involve no great complexity of domain knowledge but also exhibit the high prediction accuracy, applicability and expandability. The applications of these machine learning methods in materials property prediction have been comprehensively reviewed [17]. By using regression analysis methods for the T_g prediction of Ge_xSe_{1-x} glass system, high prediction accuracies have been achieved in our previously unpublished work, proving that regression analysis methods has higher prediction accuracy than that of the topological method [18]. In addition, many regression analysis methods, such as multi linear regression (MLR), back-propagation neural network (BPNN), support vector regression (SVR) and so on, have been applied to predict the T_g of some kinds of glasses. Chen et al. [19] and Liu et al. [20] used both MLR and BPNN to predict the T_g of polymers and obtained predicted values in good agreement with the experimental data. Pei et al. [21] employed the particle swarm optimization (PSO) to optimize SVR for the T_g prediction of random copolymers and proved that both the prediction accuracy and generalization ability of SVR are superior to those of the QSPR model. In 2013, Pei et al. [22] utilized the PSO-SVR method to predict the T_g of three classes of vinyl polymers, and found that SVR has the better prediction performance than the spectral structure activity relationship analysis (S-SAR) and artificial neural network (ANN). Furthermore, Alzghoul et al. [23] applied several machine learning methods (MLR, partial least-squares (PLS), principal component regression (PCR), ANN and SVR) for the T_g prediction of drugs, among which SVR gives the best results with *RMSE* of 18.7 K. In this paper, the regression analysis methods are introduced for the T_g prediction of As_xSe_{1-x} glass system to verify their applicability and to obtain higher prediction accuracy, and are compared with the topological method.

The remainder of this paper is as follows. In Section 2, structural analysis and learning sample set of As_xSe_{1-x} glass system are presented. Meanwhile, the modeling process of topological and regression analysis methods are described. Section 3 analyzes and discusses the obtained results. Major developments are summarized in the final section.

2. Materials and methods

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

Until now, glass structure of As_xSe_{1-x} glass system has been well understood by Raman, XPS and NMR analyses [24–29]. As shown in Fig. 1, As_xSe_{1-x} glass system mainly consists of two type structural units (Se_n chain unit and $AsSe_{3/2}$ pyramidal unit) when $x < 0.4$, only $AsSe_{3/2}$ pyramidal unit when $x = 0.4$ and four types structural units ($AsSe_{3/2}$ pyramidal unit, As_4Se_4 unit, As_4Se_3 unit and As_4 unit) when $x > 0.4$ [25,26]. By analyzing the available data correlating with structure characteristics, an obvious turning point emerges at $x = 0.4$ and the whole system of T_g variation can be divided into two stages, $0 \le x \le 0.4$ and $0.4 \le x \le 0.6$.

2.2. The acquisition of learning sample set

Referring to the analysis of internal structure and components of As_xSe_{1-x} glass system [6,25], six physical attributes involving three calculation attributes and three experimental attributes, which correlate with the T_g of As_xSe_{1-x} glass system, are taken as the input variables in regression models. Three calculation attributes can be obtained or calculated directly, including the component ratio x, the average coordination number $\langle r \rangle$ and the mean theoretical bonding energy b, whereas three experimental attributes need to be experimentally measured, including the Poisson's ratio v, the bulk modulus K and the mean experimental atomic bonding energy U_{0ex} [25]. Here, x is taken from 0 to 0.55, and values of the corresponding attributes and T_g (measured by DSC test) are listed in Table 1. (one line of data group represents one sample s_i). Eleven samples (s_1–s_{11}) of six attributes and T_g are taken as $Dataset_{six\text{-}attribute}$, which will be used as the learning sample set for modeling and prediction for T_g. Note that v, K and U_{0ex} cannot always be obtained because that measurements of longitudinal and transverse ultrasonic wave velocities failed at some compositions, such as $x = 0.6$. On the other hand, to calculate the T_g in the whole composition rage, sample s_{12} without v, K and U_{0ex} is added in Table 1, and twelve samples (s_1–s_{12}) of three calculation attributes and T_g are also established and taken as $Dataset_{three\text{-}attribute}$.

2.3. The topological method

First, since the extrapolated infinite temperature viscosity is generally independent of composition x [30], $A(x) = \log_{10}\eta_\infty$ and the well-known Adam–Gibbs relationship [31] becomes

$$\log_{10}\eta(T_g, x) = A(x) + \frac{B(x)}{T_g \cdot S_c(T_g, x)} \tag{1}$$

where $A(x)$ and $B(x)$ do not dependent on temperature but on the composition, and $S_c(T_g, x)$ is the configurational entropy.

Then, according to that the viscosity of a supercooled liquid at the glass transition temperature is independent of the composition [1], the T_g of a certain composition x can be calculated with respect to some other reference compositions x_R.

$$\eta(T_g(x), x) = \eta(T_g(x_R), x_R) \tag{2}$$

By combining Eq. (1), we get

$$\frac{T_g(x)}{T_g(x_R)} = \frac{B(x)}{B(x_R)} \frac{S_c(T_g(x_R), x_R)}{S_c(T_g(x), x)} \tag{3}$$

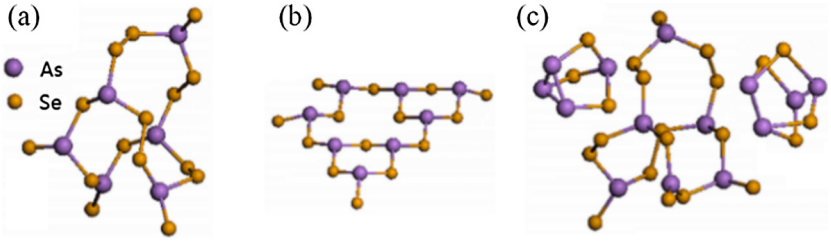

Fig. 1. Typical structure units of As_xSe_{1-x} glass system. (a) $x = 0.3$, (b) $x = 0.4$ and (c) $x = 0.5$.

Table 1
The learning sample set of As_xSe_{1-x} glass system.

	x	$\langle r\rangle$	b (±0.1 kJ/mol)	v (±0.005)	K (±0.01 Pa)	U_{0ex} (±0.1 kJ/mol)	T_g (±1 K)
s_1	0.00	2.00	184.2	0.331	8.28	152.8	313
s_2	0.10	2.10	191.4	0.320	9.82	177.3	346
s_3	0.15	2.15	195.0	0.316	10.69	190.9	360
s_4	0.20	2.20	198.6	0.312	11.44	201.5	370
s_5	0.25	2.25	202.1	0.310	12.28	213.7	380
s_6	0.30	2.30	205.7	0.307	13.13	225.1	398
s_7	0.35	2.35	209.3	0.305	13.93	235.7	422
s_8	0.40	2.40	212.9	0.304	14.80	248.8	458
s_9	0.45	2.45	211.9	0.306	13.23	223.6	459
s_{10}	0.50	2.50	211.0	0.316	11.96	203.8	446
s_{11}	0.55	2.55	210.0	0.330	10.78	184.8	419
s_{12}	0.60	2.60	209.1				389

Next, supposing that the barrier height is a slowly varying function of x [32,33], $B(x) \approx B(x_R)$, Eq. (3) can be derived as

$$\frac{T_g(x)}{T_g(x_R)} = \frac{S_c(T_g(x_R), x_R)}{S_c(T_g(x), x)} \tag{4}$$

which is a general result. To calculate $T_g(x)$ using Eq. (4), $S_c(T_g, x)$ needs to be modeled.

The configurational entropy $S_c(T_g, x)$ is given by,

$$S_c(T_g, x) = fNk \ln \Omega \tag{5}$$

where f is the degree of freedom, $f = d - n(T_g, x)$, related to the average number of constraints per atom $n(T_g, x)$ and the network dimensionality d, N is the number of atoms, k is Boltzmann constants and Ω is the number of degenerate configurations per floppy mode (proportional to the volume of the phase space explored by a floppy mode).

Finally, due to the fact that N, k, and Ω are constants, Eq. (4) can be written as,

$$\frac{T_g(x)}{T_g(x_R)} = \frac{f(T_g(x_R), x_R)}{f(T_g(x), x)} = \frac{d - n(T_g(x_R), x_R)}{d - n(T_g(x), x)} \tag{6}$$

Therefore, the T_g of a composition x can be calculated through composition dependence of topological constraints, which is well known as the topological method [6,8].

2.4. The regression analysis methods

Regression analysis methods can be divided into two categories: linear (e.g., MLR, PLS and ridge regression (RR)) and nonlinear (e.g., kernel ridge regression, SVR and BPNN) methods. Here, linear method based on RR, and nonlinear methods based on SVR and BPNN are implemented on both $Dataset_{six\text{-}attribute}$ and $Dataset_{three\text{-}attribute}$, respectively, to identify the attributes and algorithm that predict the T_g of As_xSe_{1-x} glass system best. The attributes and T_g in $Dataset_{six\text{-}attribute}$ and $Dataset_{three\text{-}attribute}$ are taken as the input and output variables in each regression model, respectively. All regression analysis methods are implemented by machine learning classes in scikit-learn toolkit [34], which is a free software machine learning library for the *Python* programming language, and the specific model constructions of these methods are described as follows:

2.4.1. The construction of RR model

RR addresses some of the problems of ordinary least squares regression by imposing a penalty on the size of coefficients, and is suitable for prediction problem in which the input variables have a high correlation with each other [35]. In this paper, the RR is implemented by the Ridge class. In the process of model construction, the GridSearchCV class is used as grid searching method to optimize the parameter, *alpha* (controls the amount of shrinkage). When the RR model is constructed, the corresponding calculation equation can be extracted from the model.

2.4.2. The construction of SVR model

Extended from Support Vector Machine (SVM) [36], SVR uses nonlinear kernel functions to map the input data into a high dimensional feature space and is capable of solving nonlinear problems with small dataset and high dimension [37,38]. Here, the attributes are mapped into a higher dimensional feature space using the RBF function. SVR is implemented by the SVR class, and the GridSearchCV class is also used to optimize parameters, C (the penalty parameter of the error term) and *gamma* (the kernel coefficient for RBF function).

2.4.3. The construction of BPNN model

With the advantages of self-learning, adaptation, resistance to noise and so on, ANN has been applied widely in materials science [39]. As one of the most commonly used ANNs, BPNN is a feed forward neural network practiced by back propagation algorithm. In theory, a simple three-layer (input, hidden and output layers) BPNN using simple nonlinear function can approximate any nonlinear functions with any precision. In this paper, a three-layer BPNN is implemented by the MLPRegressor class. The number of neurons in the input and output layers are equal to the number of input and output variables, respectively, and the number of neurons in the hidden layer is adjusted to minimize the prediction risk. During the training phase, the logistic sigmoid function and LBFGS method are used as the activation function and the solver to optimize the weight between neurons, respectively. Given that the initialized weights and threshold are random in BPNN and the predicted results of BPNN in every iteration are different, the average results of 200 iterations are taken as the final results.

The Leaving-One-Out Cross-Validation (LOOCV) method is adopted in the model constructions of three methods to partition the training and testing data. In each iteration, testing data has one sample and training data has $n - 1$ other samples. A total of n iterations are conducted and then *RMSE* and *MAPE* are used to evaluate the prediction accuracy.

$$RMSE = \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n} \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - y_i'}{y_i} \tag{8}$$

where n denotes the amount of data, y_i and y_i' represent the ith experimental and predicted values, respectively. The lower the *RMSE* and *MAPE* are, the higher prediction accuracy of regression analysis methods can obtain.

3. Results and discussion

3.1. T_g of As_xSe_{1-x} glass system predicted using the topological method

As described in Section 2.1, the T_g of As_xSe_{1-x} glass system can be divided into two stages at $x = 0.4$ due to the structural transition. In the composition range of $0 \leq x \leq 0.4$, the number of constraints per atom n is

$$n(T, 0 \leqslant x \leqslant 0.4) = x[1.5q_\alpha(T) + 3q_\beta(T)] + (1-x)[q_\alpha(T) + q_\gamma(T)] + q_\delta(T) \quad (9)$$

where $q(T)$ gives the hardness of a particular constraint, $q(T) = \theta(T_q - T)$, and T_q represents the temperature below which a particular constraint becomes rigid, $q(T) = 1(T_q > T)$. The subscripts denote different types of constraints. α represents a As-Se or Se-Se linear bond constraint, β represents a Se-As-Se, or Se-As-As, or As-As-As angular constraint, γ denotes the angular constraints centered at Se, and δ is the Van der Waals bond. Hence, the first term accounts for one point five linear and three angular constraints at each arsenic atom, the second term gives one linear and one angular constraint for each selenium atom, and the last term provides for Van der Waals bonding. According to Refs. [6,8], we assume that the relative strengths of the bonds are such that $T_\delta \leq T_\gamma \leq T_\beta \leq T_\alpha$: The As or Se linear bond constraints (α) providing the backbone of the glassy network are strongest and frozen at the highest temperature. The As angular constraint (β) is the second strongest constraint and are frozen in just below the lowest T_g, owing to the sp^2 hybridization of the As orbitals, which produces planar pyramidal bond angels. The Se angular constraint (γ) is known to be quite soft [40]. Obviously, Van der Waals (δ) forces are the weakest constraints. Above all, the T_g must fall between T_β and T_α across the full range of x values, i.z.,

$$T_\delta \leqslant T_\gamma \leqslant T_\beta \leqslant T_g(0 \leqslant x \leqslant 0.4) \leqslant T_\alpha \quad (10)$$

With Eq. (10), Eq. (9) can be simplified as

$$n(T, 0 \leqslant x \leqslant 0.4) = 1 + 0.5x \quad (11)$$

For a network in three-dimensional space ($d = 3$), the average number of degrees of freedom per atom is

$$f(T_g(x), 0 \leqslant x \leqslant 0.4) = d - n(T_g(x), 0 \leqslant x \leqslant 0.4) = 2 - 0.5x \quad (12)$$

By combining Eqs. (6) and (12), the T_g can be calculated as

$$\frac{T_g(0 \leqslant x \leqslant 0.4)}{T_g(0.4)} = \frac{f(T_g(0.4), 0.4)}{f(T_g(x), 0 \leqslant x \leqslant 0.4)} = \frac{1.8}{2 - 0.5x} \quad (13)$$

In the composition range of $0.4 \leq x \leq 0.5$, there are $AsSe_{3/2}$ and As-rich units in forms of As_4Se_4 and As_4Se_3 cage-like molecules. According to previous report that the fractions of As_4Se_4 and As_4Se_3 are almost same, a dummy variable v can be introduced [25], and the fraction of As_2Se_3 unit is set as another dummy variable u. Then the quantitative composition can be written as,

$$As_xSe_{1-x} = (AsSe_{3/2})_u(AsSe)_v(AsSe_{3/4})_v \quad (14)$$

To balance the numbers of As and Se atoms,

$$u + v + v = x, \quad (15)$$

$$\frac{3}{2}u + v + \frac{3v}{4} = 1 - x, \quad (16)$$

they give the relations

$$u = 1.6 - 3x \quad (17)$$

$$v = 2x - 0.8 \quad (18)$$

Substituting Eqs. (17) and (18) into Eq. (14), we can obtain

$$As_xSe_{1-x} = (AsSe_{3/2})_{1.6-3x}(AsSe)_{2x-0.8}(AsSe_{3/4})_{2x-0.8} \quad (19)$$

Because the T_g is directly related to the net link energy, the influence of two cage-like molecules is omitted. Then, similar to the constraints in Se-rich range, the number of constraints per atom n in As-rich range can be

$$n(T_g(x), 0.4 \leqslant x \leqslant 0.5) = (1.6 - 3x)\{[1.5q_\alpha(T) + 3q_\beta(T)] + 1.5[q_\alpha(T) + q_\gamma(T)]\} + q_\delta(T) \quad (20)$$

With Eq. (10), Eq. (20) can be simplified as

$$n(T_g(x), 0.4 \leqslant x \leqslant 0.5) = 4.8 - 9x \quad (21)$$

For a three-dimensional network ($d = 3$), the average number of degrees of freedom per atom is

$$f(T_g(x), 0.4 \leqslant x \leqslant 0.5) = d - n(T_g(x), 0.4 \leqslant x \leqslant 0.5) = 9x - 1.8 \quad (22)$$

Similar to Se-rich case, by combining Eqs. (6) and (22), the T_g can be calculated as

$$\frac{T_g(0.4 \leqslant x \leqslant 0.5)}{T_g(0.4)} = \frac{f(T_g(0.4), 0)}{f(T_g(x), 0.4 \leqslant x \leqslant 0.5)} = \frac{1.8}{9x - 1.8} \quad (23)$$

Combining Eqs. (13) and (23), we therefore can obtain the final result

$$T_g(x) = \begin{cases} \frac{1.8}{2-0.5x}T_g(0.4), & 0 \leqslant x \leqslant 0.4 \\ \frac{1.8}{9x-1.8}T_g(0.4). & 0.4 \leqslant x \leqslant 0.5 \end{cases} \quad (24)$$

The T_g of As_xSe_{1-x} glasses system ($0 \leq x \leq 0.4$) can be well fitted by the modified Gibbs-Marzio equation as $y = 315.5/(1–0.74x)$. However, according to the topological method, the T_g can be simulated as

$$y = \begin{cases} \frac{1.8*458}{2-0.5x}, & 0 \leqslant x \leqslant 0.4 \\ \frac{1.8*458}{9x-1.8}, & 0.4 \leqslant x \leqslant 0.5 \end{cases} \quad (25)$$

Eq. (25) in the composition range of $0 \leq x \leq 0.4$ deviates from the modified Gibbs-Marizo equation, which differs from that of the topological method for T_g prediction in Ge_xSe_{1-x} glass system [26]. This means that the topological method for the T_g prediction cannot be extended well to all glass systems, even a basic binary As_xSe_{1-x} glass system. This may be caused by the limited knowledge of quantitative structures and variation of coordination number from 4-coordinated germanium to 3-coordinated arsenic, resulting in an uncertain average number of constraints and also an uncertain average number of degrees in As_xSe_{1-x} glass system. Furthermore, note that the quantitative structures of As_xSe_{1-x} glass system containing many inclusions, such as $AsSe_3$, As_4, As_4Se_4 and As_4Se_3 in the composition range of $0.5 < x \leq 0.6$ are much more complex and uncertain than these in the composition range of $0 < x \leq 0.5$ [25,26]. Therefore, the topological method is unable to calculate the T_g when $x > 0.5$.

The T_g values of As_xSe_{1-x} glass system obtained using the topological method are listed in Table 2 (the experimental T_g at $x = 0.4$ is taken as the initial point in the topology method, so the predicted T_g is equal to the experiment one (shown in italic) and the predicted error at this point is none). It can be seen that the predicted errors increase from 29.73 K to 140.67 K, which illustrates that the T_g values obtained using the topological method are in poor agreement with the experimental T_g. Fig. 2 shows the composition dependence of T_g of As_xSe_{1-x} glass system obtained using experimental and topological methods. It is also shown that especially when x is far away from 0.4, the predicted errors get much worse with 92.60 K and 140.67 K at $x = 0.45$ and 0.5, respectively.

Table 2
Predicted results of T_g obtained using the topological method.

x	Experimental T_g (K)	Predicted T_g (K)	Predicted error (|K|)
0.00	313	412.20	99.20
0.10	346	422.77	76.77
0.15	360	428.26	68.26
0.20	370	433.89	63.89
0.25	380	439.68	59.68
0.30	398	445.62	47.62
0.35	422	451.73	29.73
0.40	*458*	*458.00*	
0.45	459	366.40	92.60
0.50	446	305.33	140.67

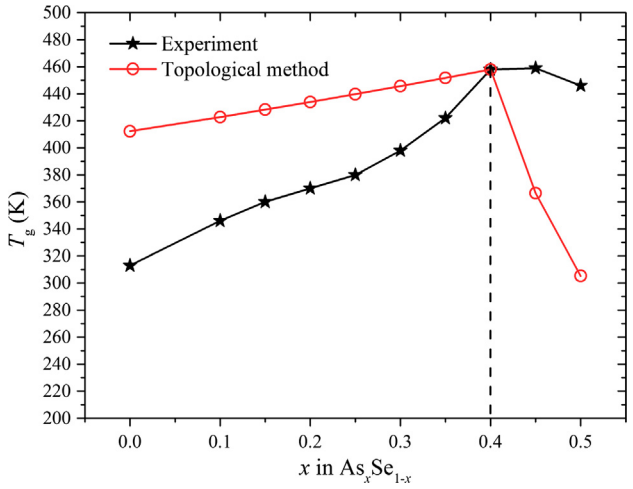

Fig. 2. Composition dependence of T_g of As_xSe_{1-x} glass system obtained using experimental and topological methods.

Consequently, the *RMSE* and *MAPE* results of topological method reach 77.17 K and 17.65%, respectively, which are unaccepted.

Obviously, although the topological method obtains good performance in some glass systems [3–12], it shows poor prediction effectiveness and applicability in the As_xSe_{1-x} glass system, which may be caused by the limited knowledge of quantitative structures and its variation in the system.

3.2. T_g of As_xSe_{1-x} glass system predicted using the regression analysis methods

RR, SVR and BPNN are implemented to establish the relationships between attributes and T_g of As_xSe_{1-x} glass system for the higher prediction accuracy. The relevant parameters in each method are set as follows: (1) *alpha* in RR is 0.1; (2) *C* and *gamma* in SVR are 100 and 0.01, respectively; (3) The number of neurons in the hidden layer in BPNN is 10.

3.2.1. Results obtained on $Dataset_{six\text{-}attribute}$

Table 3 lists that the T_g values of As_xSe_{1-x} glass system in eleven iterations of LOOCV using the ridge regression, SVR and BPNN methods for testing data on $Dataset_{six\text{-}attribute}$. The minimum and maximum predicted errors of these three methods reach 0.62 K and 31.26 K, respectively, which are much smaller than those of topological method (29.73 K and 140.67 K). Fig. 3 shows the composition dependence of T_g of As_xSe_{1-x} glass system obtained using RR, SVR and BPNN methods for testing data on $Dataset_{six\text{-}attribute}$ as well as experimental and topological methods. We can clearly see that as compared with those of topological method, the T_g values obtained using the regression analysis methods are all in a much better agreement with the experimental T_g in the whole stage, especially when $x > 0.4$. More importantly, un-restricted to the limited knowledge of quantitative structures and its variation in the topological method, the regression analysis methods have the ability to predict the T_g even when $x = 0.55$.

Now, we turn to analyze and compare the predicted results of three regression analysis methods. It is seen from Table 3 that predicted errors of RR at x=0, 0.25, 0.3, 0.4, 0.45, 0.5 and 0.55 (average is 19.94 K) are relative larger than the errors at other points (average is 5.46 K). In Fig. 3, RR shows a poor prediction effectiveness in the composition range of $0.4 \leq x \leq 0.55$ with an average of 22.79 K. Particularly when $x = 0.55$, the T_g value obtained by RR is up to 31.26 K and deviates from the normal variation tendency largely. This is because the amount of training data (s_1–s_{10}) in the 11st iteration of LOOCV is so small that the RR cannot learn the approximation of variation tendency in the composition range of $0.4 \leq x \leq 0.55$ well, leading to its under-fitting. The average *RMSE* and *MAPE* values in eleven iterations of LOOCV of three regression analysis methods on $Dataset_{six\text{-}attribute}$ are shown in Fig. 4. Compared with SVR and BPNN, RR exhibits relative lower prediction accuracy with an *RMSE* of 10.25 K and *MAPE* of 2.38% for training data, and an *RMSE* of 14.67 K and *MAPE* of 3.61% for testing data.

On the other hand, nonlinear regression analysis methods both achieve better prediction results than RR. As shown in Table 3, compared with RR, SVR also has relative large predicted errors at some points (average is 14.36 K), such as x = 0, 0.25, 0.3, 0.4, 0.45 and 0.55, in which the maximum error reaches 21.84 K. However, Fig. 3 illustrates that SVR learns a much better approximation of the variation tendency in the range of $0.4 \leq x \leq 0.55$ with an average predicted error of 13.23 K. Meanwhile, SVR reduces the *RMSE* by 29.03% ($|RMSE_{RR}\text{-}RMSE_{SVR}|/RMSE_{RR}$) and *MAPE* by 28.25% ($|MAPE_{RR}\text{-}MAPE_{SVR}|/MAPE_{RR}$) for testing data while shows a bit worse *RMSE* and *MAPE* for training data (see Fig. 4). On the other hand, the T_g values obtained using BPNN fit the experimental data extremely well in the whole stage. However, the average predicted error at x = 0, 0.4 and 0.55 reaches up to 21.17 K, which is much higher than that at other points (2.87 K). Particularly when x = 0.4, the predicted error reaches 30.87 K, reflecting the shortcoming of BPNN which is easily falling into local optimum. Meanwhile, as compared with SVR, BPNN reduces the *RMSE* (*MAPE*) by 24.40% (24.32%) for testing data and further reduces one magnitude of *RMSE* and *MAPE* for training data.

In addition, given the advantage of the interpretability of linear regression models, a six-attribute correlation equation based on RR can be obtained as Eq. (27).

$$T_g = 0.26711x + 0.26711\langle r\rangle - 0.09773v + 0.12346K + 0.07695U_{oex} + 0.26309b + 0.04908 \quad (0 \leqslant x \leqslant 0.55) \tag{26}$$

where the six coefficients and one constant are the average values of eleven iterations of LOOCV on $Dataset_{six\text{-}attribute}$. The composition dependence of T_g of As_xSe_{1-x} glass system obtained using the six-attribute equation is shown in Fig. 5. The obtained *RMSE* and *MAPE* can respectively reach 12.81 K and 2.32%, which are reduced by 86.56% and 85.53% compared with the topological method. However, it is clearly seen from Fig. 5 that the T_g values obtained using six-attribute equation deviate from the normal variation tendency largely when $x \geq 0.4$ because of the problem of small amount of dataset.

3.2.2. Results obtained on $Dataset_{three\text{-}attribute}$

Although regression analysis methods can obtain high prediction accuracy on $Dataset_{six\text{-}attribute}$, measurements of the three experimental attributes are costly, time-consuming and influenced by experimental conditions, inducing Eq. (27) hard to be applied in

Table 3
Comparison between predicted results of T_g in eleven iterations obtained using RR, SVR and BPNN methods for testing data on $Dataset_{six\text{-}attribute}$.

Iteration	x	Experimental T_g (K)	Predicted T_g (K)			Predicted error (|K|)		
			RR	SVR	BPNN	RR	SVR	BPNN
1st	0.00	313	297.25	302.51	330.82	15.75	10.49	17.82
2nd	0.10	346	341.67	340.72	346.63	4.33	5.28	0.63
3rd	0.15	360	360.76	359.38	356.66	0.76	0.62	3.34
4th	0.20	370	379.28	377.40	369.25	9.28	7.40	0.75
5th	0.25	380	396.56	396.08	381.82	16.56	16.08	1.82
6th	0.30	398	414.07	412.38	399.86	16.07	14.38	1.86
7th	0.35	422	429.45	429.31	425.51	7.45	7.31	3.51
8th	0.40	458	435.59	446.05	427.13	22.41	11.95	30.87
9th	0.45	459	435.70	437.16	454.85	23.30	21.84	4.15
10th	0.50	446	431.80	438.32	439.07	14.20	7.68	6.93
11th	0.55	419	450.26	430.44	433.87	31.26	11.44	14.87

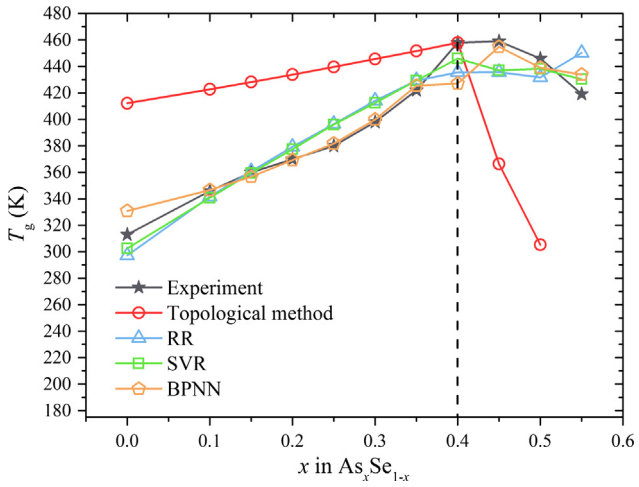

Fig. 3. Composition dependence of T_g of As_xSe_{1-x} glass system obtained using RR, SVR and BPNN methods for testing data on $Dataset_{six\text{-}attribute}$ as well as experimental and topological methods.

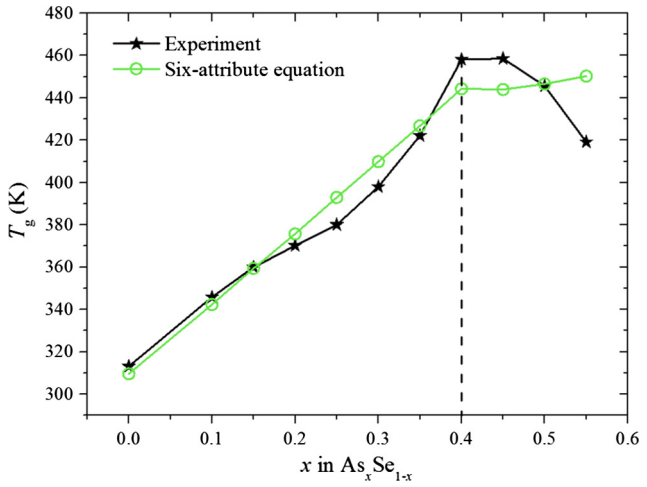

Fig. 5. Composition dependence of T_g of As_xSe_{1-x} glass system obtained using six-attribute equation based on RR.

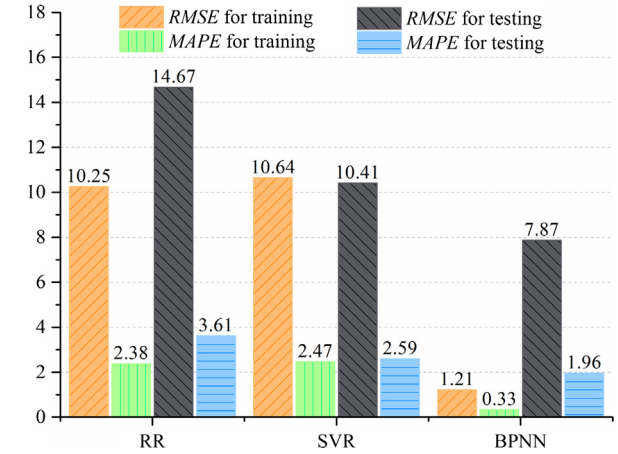

Fig. 4. Average *RMSE* and *MAPE* results of T_g in eleven iterations obtained using RR, SVR and BPNN methods on $Dataset_{six\text{-}attribute}$.

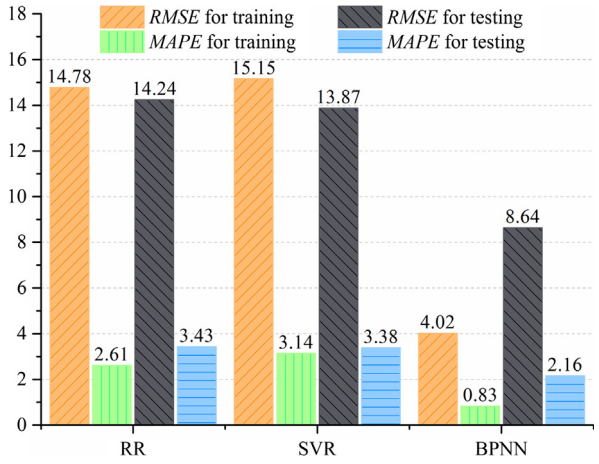

Fig. 6. Average *RMSE* and *MAPE* results of T_g in twelve iterations obtained using RR, SVR and BPNN methods on $Dataset_{three\text{-}attribute}$.

practice. Thus, RR, SVR and BPNN are also implemented on $Dataset_{three\text{-}attribute}$ for better practicability.

The average *RMSE* and *MAPE* results in twelve iterations of LOOCV of these three methods on $Dataset_{three\text{-}attribute}$ are shown in Fig. 6. Compared with the values on $Dataset_{six\text{-}attribute}$ in Fig. 4, RR reduces the *RMSE* by 0.43 K and *MAPE* by 0.18% for testing data on $Dataset_{three\text{-}attribute}$ slightly, but increases the *RMSE* by 4.53 K and *MAPE* by 0.23% for training data. SVR obtains an average *RMSE* of 14.51 K and average *MAPE* of 3.26% for training and testing data on $Dataset_{three\text{-}attribute}$, both higher than those (10.53 K and 2.53%) on $Dataset_{six\text{-}attribute}$. For BPNN, the *RMSE* and *MAPE* values on $Dataset_{three\text{-}attribute}$ are a bit larger than those on $Dataset_{six\text{-}attribute}$, in which the *RMSE* (*MAPE*) for training data on $Dataset_{three\text{-}attribute}$ is larger by 2.81 K (0.50%) than that on $Dataset_{six\text{-}attribut}$.

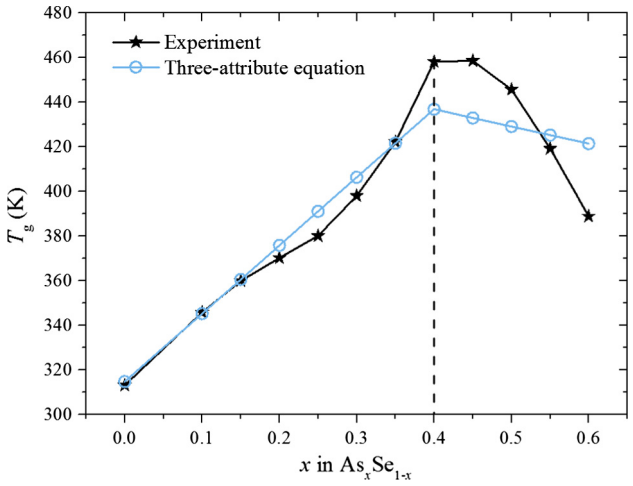

Fig. 7. Composition dependence of T_g of As_xSe_{1-x} glass system obtained using three-attribute equation based on RR.

A three-attribute correlation equation based on RR on $Dataset_{three\text{-}attribute}$ is also obtained as Eq. (27)

$$T_g = 0.00569x + 0.00569\langle r\rangle + 0.83153b + 0.01152 \quad (0 \leqslant x \leqslant 0.6) \tag{27}$$

where three coefficients and one constant are the average values of twelve iterations of LOOCV on $Dataset_{three\text{-}attribute}$. The composition dependence of T_g of As_xSe_{1-x} glass system obtained using the three-attribute equation is shown in Fig. 7. The obtained *RMSE* and *MAPE* are 15.07 K and 2.63%, respectively, which are a bit higher than those of Eq. (26). However, from Figs. 5 and 7, it is obvious that the three-attribute correlation equation exhibits a better simulation of the T_g variation tendency than the six-attribute correlation equation when $x \geq 0.4$ and has the ability to calculate the T_g at $x = 0.6$. This is because the amount of $Data_{three\text{-}attribute}$ is bigger than that of $Data_{six\text{-}attribute}$, causing the better learning of RR for the approximation of variation tendency. Meanwhile, given that the three calculation attributes can be calculated directly, the three-attribute correlation equation has better practicability.

4. Conclusion

In this study, a systematic comparison of topological and regression analysis methods for T_g prediction of As_xSe_{1-x} glass system is presented. Given that the limited knowledge of quantitative structures and its variation, the topological method cannot be extended to As_xSe_{1-x} glass system with a high *RMSE* of 77.17 K and *MAPE* of 17.65%, respectively, in the composition range of $0 \leq x \leq 0.5$. In contrast, BPNN not only overcomes that limitation, but also achieves the highest prediction accuracy on the dataset containing three experimental and three calculation attributes. More importantly, a correlation equation with three calculation attributes based on RR is obtained. It possesses the advantages of simplicity and practicability similar to the calculation equation of topological method but has much higher prediction accuracy. Our investigation indicates that in the case of the limited knowledge, although the precisions of regression methods are limited to small data volume and few attributes in the As_xSe_{1-x} glass system, regression methods still exhibit better prediction effectiveness and practicability from the perspective of data characteristic. When the data volume is sufficient, the data quality is high and attributes are easily obtained, the regression methods can obtain higher prediction accuracy and better applicability. What's more, the regression methods are able to expand to mixed glass systems as previous work in organic glass systems where the topological method is not.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. U1630134, 51622207 and 51372228), the National Key Research and Development Program of China (Nos. 2017YFB0701600 and 2017YFB0701500), Shanghai Institute of Materials Genome from the Shanghai Municipal Science and Technology Commission (No. 14DZ2261200), the Shanghai Municipal Education Commission (No. 14ZZ099). All the computations were performed on the high performance computing platform of Shanghai University. Guang Yang thanks the National Natural Science Foundation of China (No. 51602187), the Training Scheme of Shanghai Young Teachers, and the Young Eastern Scholars Program 2015 from Shanghai Municipal Education Commission.

References

[1] C.A. Angell, Science 267 (1995) 1924–1935.
[2] P.G. Debenedetti, F.H. Stillinger, Nature 410 (2001) 259–267.
[3] C.R. Petersen, U. Møller, I. Kubat, B. Zhou, S. Dupont, J. Ramsay, T. Benson, S. Sujecki, N. Abdel-Moneim, Z. Tang, D. Furniss, A. Seddon, O. Bang, Nat. Photon. 8 (2014) 830–834.
[4] T. Rouxel, J. Chem. Phys. 135 (2011) 184501.
[5] A.N. Sreeram, A.K. Varshneya, D.R. Swiler, J. Non-Cryst. Solids 128 (1991) 294–309.
[6] P.K. Gupta, J.C. Mauro, J. Chem. Phys. 130 (2009) 094503.
[7] M. Micoulaut, Y. Yue, MRS Bull. 42 (2017) 18–22.
[8] J.C. Mauro, P.K. Gupta, R. Loucks, J. Chem. Phys. 130 (2009) 234503.
[9] M.M. Smedskjaer, J.C. Mauro, S. Sen, Y. Yue, Chem. Mater. 22 (2010) 5358–5365.
[10] M.M. Smedskjaer, J.C. Mauro, Y. Yue, Phys. Rev. Lett. 105 (2010) 115503.
[11] J.C. Mauro, Am. Ceram. Soc. Bull. 90 (2011) 31.
[12] M.M. Smedskjaer, J.C. Mauro, R.E. Youngman, C.L. Hogue, M. Potuzak, Y. Yue, J. Phys. Chem. B 115 (2011) 12930–12946.
[13] Q. Jiang, H. Zeng, Z. Liu, J. Ren, G. Chen, Z. Wang, L. Sun, D. Zhao, J. Chem. Phys. 139 (2013) 124502.
[14] C. Hermansen, J.C. Mauro, Y. Yue, J. Chem. Phys. 140 (2014) 154501.
[15] Q. Jiang, H. Zeng, X. Li, J. Ren, G. Chen, F. Liu, J. Chem. Phys. 141 (2014) 124506.
[16] M.M. Smedskjaer, Front. Mater. 1 (2014) 23.
[17] Y. Liu, T. Zhao, W. Ju, S. Shi, J. Materiomics 3 (2017) 159–177.
[18] Y. Liu, T. Zhao, G. Yang, W. Ju, S. Shi, Comput. Mater. Sci. (2017) (in preparation).
[19] X. Chen, L. Sztandera, H.M. Cartwright, Int. J. Intell. Syst. 23 (2008) 22–32.
[20] W. Liu, C. Cao, Colloid Polym. Sci. 287 (2009) 811–818.
[21] J.F. Pei, C.Z. Cai, J.L. Tang, S. Zhao, F.Q. Yuan, J. Macromol. Sci., Part B 51 (2012) 1437–1448.
[22] J.F. Pei, C.Z. Cai, Y.M. Zhu, J. Theor. Comput. Chem. 12 (2013) 1350002.
[23] A. Alzghoul, A. Alhalaweh, D. Mahlin, C.A.S. Bergström, J. Chem. Inf. Model. 54 (2014) 3396–3403.
[24] M. Deschamps, C. Roiland, B. Bureau, G. Yang, L. Le Polles, D. Massiot, Solid State Nucl. Magn. Reson. 40 (2011) 72–77.
[25] G. Yang, B. Bureau, T. Rouxel, Y. Gueguen, O. Gulbiten, C. Roiland, E. Soignard, J. L. Yarger, J. Troles, J.-C. Sangleboeuf, P. Lucas, Phys. Rev. B 82 (2010) 195206.
[26] G. Yang, O. Gulbiten, Y. Gueguen, B. Bureau, J.-C. Sangleboeuf, C. Roiland, E.A. King, P. Lucas, Phys. Rev. B 85 (2012) 144107.
[27] B. Bureau, J. Troles, M. LeFloch, F. Smektala, G. Silly, J. Lucas, Solid State Sci. 5 (2003) 219–224.
[28] R. Golovchak, A. Kovalskiy, A.C. Miller, H. Jain, O. Shpotyuk, Phys. Rev. B 76 (2007) 125208.
[29] M. Deschamps, C. Genevois, S. Cui, C. Rolland, L. LePolles, E. Furet, D. Massiot, B. Bureau, J. Phys. Chem. C 119 (2015) 11852–11857.
[30] C.A. Angell, J. Non-Cryst. Solids 73 (1985) 1–17.
[31] G. Adam, J.H. Gibbs, J. Chem. Phys. 43 (1965) 139–146.
[32] M.J. Toplis, Journal 83 (1998) 480.
[33] M.J. Toplis, Chem. Geol. 174 (2001) 321–331.
[34] <http://scikit-learn.org>.
[35] A.E. Hoerl, R.W. Kennard, Technometrics 12 (1970) 55–67.
[36] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[37] Ö. Eskidere, F. Ertaş, C. Hanilçi, Expert Syst. Appl. 39 (2012) 5523–5528.
[38] H. Zhou, J.P. Zhao, L.G. Zheng, C.L. Wang, K.F. Cen, Eng. Appl. Artif. Intell. 25 (2012) 147–158.
[39] Z. Zhang, K. Friedrich, Compos. Sci. Technol. 63 (2003) 2029–2044.
[40] R. Kerner, Glass Phys. Chem 26 (2000) 313–324.