

# PREDICTION OF SUPERCONDUCTING TRANSITION TEMPERATURE USING A MACHINE-LEARNING METHOD

## NAPOVED TEMPERATURE PREHODA V SUPERPREVODNOST Z UPORABO METODE STROJNEGA UČENJA

Yao Liu<sup>1</sup>, Huiran Zhang<sup>1,2,3</sup>, Yan Xu<sup>4</sup>, Shengzhou Li<sup>1</sup>, Dongbo Dai<sup>1</sup>, Chengfan Li<sup>1</sup>, Guangtai Ding<sup>1,2</sup>, Wenfeng Shen<sup>1,3</sup>, Quan Qian<sup>1,2</sup>

<sup>1</sup>Shanghai University, School of Computer Engineering and Science, no. 99 Shangda Road, Baoshan District, Shanghai 200444, China

<sup>2</sup>Materials Genome Institute of Shanghai University, Shanghai 200444, China

<sup>3</sup>Shanghai University, Shanghai Institute of Advanced Communication and Data Science, Shanghai 200444, China

<sup>4</sup>Shanghai University of Electric Power, College of Mathematics and Physics, Shanghai 200090, China  
hrzhangsh@shu.edu.cn

Prejem rokopisa – received: 2018-03-12; sprejem za objavo – accepted for publication: 2018-05-18

doi:10.17222/mit.2018.043

A high-transition-temperature (high- $T_C$ ) superconductor is an important material used in many practical applications like magnetically levitated trains and power transmission. The superconducting transition temperature  $T_C$  is determined by the layered crystals, bond lengths, valency properties of the ions and Coulomb coupling between electronic bands in adjacent, spatially separated layers. The optimal  $T_C$  can be attained upon doping and applying the pressure for the optimal compounds. There is an algebraic relation for the optimal  $T_C$  of the optimal compounds,  $T_{CO} = K_B^{-1}\beta/(t\xi)$ , where  $t$  and  $\xi$  are two structural parameters,  $K_B$  is Boltzmann's constant,  $\beta$  is a universal constant and  $T_{CO}$  is the optimal transition temperature. Nevertheless, the  $T_C$  of the non-optimum compounds is smaller than  $T_{CO}$ . To predict the  $T_C$  for the all compounds, we developed a prediction model based on the machine-learning method called support vector regression (SVR) using structural and electronic parameters to predict  $T_C$ . In addition, the principal component analysis (PCA) was applied to reduce dimensions and interdependencies among the parameters, and particle swarm optimization (PSO) was utilized to search for the optimal parameters of SVR for an improved performance of the prediction model. The results showed that the proposed PCA-PSO-SVR model takes advantage of the machine-learning method to directly predict  $T_C$  and theoretically provide guidance on measuring  $T_C$ .

Keywords: superconducting transition temperature  $T_C$ , machine learning, structural and electronic parameters, PCA-PSO-SVR

Visoka temperatura prehoda v superprevodnost ( $T_C$ ) je pomembna funkcionalna lastnost materiala za mnoge vrste praktične uporabe, kot je naprimer uporaba magnetne levitacije za vlake ali prenos moči. Temperaturo prehoda v superprevodnost  $T_C$  določa plastovitost kristala, dolžina medatomskih vezi, valenčne lastnosti ionov in Coulombovo sklapljanje med sosednjimi valenčnimi pasovi prostorsko ločenih plasti. Optimalna  $T_C$  se lahko doseže z dopiranjem (dodajanjem, legiranjem) in uporabo tlaka za optimalno kemijsko sestavo. Obstaja algebraična zveza za optimalno  $T_C$  optimalne spojine,  $T_{CO} = K_B^{-1}\beta/(t\xi)$ , kjer sta  $t$  in  $\xi$  dva strukturna parametra,  $K_B$  je Boltzmannova konstanta,  $\beta$  je univerzalna konstanta in  $T_{CO}$  je optimalna temperatura prehoda. Vendar je  $T_C$  neoptimalne spojine vedno manjša kot  $T_{CO}$ . Avtorji tega prispevka so za napoved  $T_C$  vseh spojin razvili model na osnovi metode strojnega učenja. Za napoved  $T_C$  so uporabili vektorsko regresijo (SVR) z odgovarjajočimi strukturnimi in elektronskimi parametri. Dodatno so uporabili osnovno komponentno analizo (PCA, angl.: Principal Component Analysis), da so lahko zmanjšali soodvisnosti med parametri. Uporabili so še optimizacijo množic delcev (PSO; angl.: Particle Swarm Optimization) za iskanje optimalnih parametrov SVR in izboljšanje lastnosti modela. Raziskave avtorjev tega prispevka so pokazale, da predlagani model PCA-PSO-SVR s pridom izkorišča prednosti metode strojnega učenja za neposredno napoved  $T_C$ , in tudi zagotavlja teoretično podlago za merjenje  $T_C$ .

Ključne besede: temperatura prehoda v superprevodnost  $T_C$ , strojno učenje, strukturni in elektronski parametri, PCA-PSO-SVR

## 1 INTRODUCTION

As important functional materials, high-transition-temperature (high- $T_C$ ) superconductors<sup>1</sup> have some typical physical parameters, such as transition temperature  $T_C$ , magnetic susceptibility and critical current density ( $J_C$ ), which make them very useful in many practical applications like magnetically levitated trains and power transmission.<sup>2-5</sup> Previous researches showed that the high- $T_C$  superconductors are generally characterized by a two-dimensional layered superconducting condensate with unique features that are not traditional superconducting metals.<sup>6</sup> Their important property,  $T_C$ , is determined by their layered crystals, bond lengths, valency properties of the ions, and Coulomb coupling

between electronic bands in adjacent, spatially separated layers.<sup>7</sup> The optimal  $T_C$  can be attained upon doping with other external materials or applying pressure for the optimal high- $T_C$  superconducting compounds.<sup>8</sup> There is an algebraic relation for the optimal  $T_C$  of the optimal compounds:<sup>7,9</sup>

$$T_{CO} = K_B^{-1}\beta/(t\xi) \quad (1)$$

Here,  $t$  is related to the mean spacing between interacting charges in the layers,  $\xi$  is the distance between interacting electronic layers,  $K_B$  is Boltzmann's constant,  $\beta$  is a universal constant and  $T_{CO}$  is the optimal transition temperature.

Formula (1) is a good way to predict the  $T_{CO}$  of optimal high- $T_C$  superconducting compounds. However,

non-optimum compounds, in which sample degradation is evident, typically show that  $T_C$  is smaller than  $T_{C0}$ .<sup>7</sup> In other words, it is critical to predict  $T_C$  of various high- $T_C$  superconducting compounds. In our present work, we developed a prediction model based on a machine-learning method to predict  $T_C$  of various high- $T_C$  superconducting compounds using structural and electronic parameters. The results of the prediction model show that the model can predict  $T_C$  quickly and accurately.

Recently, in order to accelerate the process of discovery and deployment of new materials, more and more researchers have used machine-learning methods to find new materials, classify them and predict their properties.<sup>10–13</sup> For high- $T_C$  superconductors and their  $T_C$ , researchers developed a computational-intelligence-based model via SVR<sup>14</sup> to estimate the  $T_C$  of YBCO superconductors using lattice parameters as the descriptors<sup>15–17</sup> and manually found the optimal parameters of SVR one by one with the trend charts of the effects of the parameters on the experimental results. It is a very feasible way to estimate the  $T_C$  of YBCO superconductors, but the manual parameter optimization may take a lot of time.

In this paper, in order to predict  $T_C$  of various high- $T_C$  superconductors, we established a PCA-PSO-SVR model based on a machine-learning method using structural and electronic parameters. These parameters, including  $\zeta$  (the distance between interacting electronic layers),  $A$  (the distance between interacting electronic layers),  $d$  (the periodicity),  $\eta$  (the number of type II layers),  $\nu$  (the number of type I layers),  $\sigma$  (the fractional charge per type I layer) and  $\gamma$  (the factor for calculating  $\sigma$ ), related to 31 kinds of high- $T_C$  superconductors that form the dataset from the literature.<sup>7,9</sup> The dataset has only 31 samples and each sample has only 7 features, which is obviously a small sample set, but the SVR shows many unique advantages of processing small sample sets because of the theory of statistical learning and the minimum principle of structural risk. Hence, we chose the SVR as the regression algorithm of the prediction model. To achieve a higher performance of the model, we adopted automatic optimization with a simple and efficient PSO<sup>18</sup> optimization algorithm instead of the manual optimization used in the previous studies when searching for the optimal SVR parameters. Meanwhile, we found that some parameters are interdependent by analysing the crystal structure and parameters of the high- $T_C$  superconductors, so we used PCA<sup>19</sup> to reduce dimensions and interdependencies in the data pre-processing for a better accuracy of the prediction model. In addition, we also trained the PSO-SVR model and the back-propagation neural network (BPNN)<sup>20</sup> with the dataset for comparison. The corresponding experimental results showed that the PCA-PSO-SVR prediction model is more accurate when predicting  $T_C$ . Meanwhile, we used additional data to validate the prediction model, and the results were also reasonable. It means that this

prediction model, based on the machine-learning method, can directly predict  $T_C$ .

## 2 ESTABLISHMENT OF THE PREDICTION MODEL

In order to identify the feasibility and validity of the new model, 31 kinds of high- $T_C$  superconductors, including cuprates, ruthenates, ruthenocuprates, iron pnictides and organics, whose  $T_C$  values are in a range of [10.5, 145], were selected from the literature<sup>7</sup> as the dataset. These materials are independent of the locations of two carrier types, of which type I is defined with the BaO-CuO-BaO (or equivalent) layers and type II is defined with the cuprate-plane CuO<sub>2</sub>-Y-BuO<sub>2</sub> (or equivalent) layers. The details of the dataset were presented in the Data.docx file. In the process of establishing the prediction model, the structural ( $Z$ ,  $A$ ,  $d$ ,  $\eta$ ,  $\nu$ ) and electronic ( $\sigma$ ,  $\gamma$ ) parameters were scaled to [0,1] with the min-max normalization, and taken as the input vectors, while  $T_C$  was the output value for the regression. Given that the parameters are interdependent (e.g.,  $\gamma$  is related to  $\sigma$ , and  $Z$  is a part of  $d$  according to the definition of these two parameters), we used the widely applicable PCA method to reduce the dimensions and interdependencies of the parameters. In the PCA process, first, the covariance matrix of the dataset is calculated, then the eigenvalues of the covariance matrix are calculated, and finally the top  $d$  eigenvalue of all the eigenvalues is selected, while the corresponding feature vectors form the solution of the PCA. The selected reduced dimensions are based not only on the contribution rate that can be calculated with Equation (2) but also on the errors of the predicted results of the PCA-PSO-SVR model.

$$C = \frac{\sum_{i=1}^n \lambda_i}{\sum_{j=1}^m \lambda_j} \quad (2)$$

Here,  $\lambda_i$  denotes the  $i$ th eigenvalue,  $n$  denotes the chosen dimension amount and  $m$  denotes the entire dimension amount.

Detailed results of the dimension reduction are discussed in the next section. After reducing the dimensions, the dataset was divided into two parts via the leave-one-out cross-validation (LOOCV)<sup>21</sup> method. 30 samples were used to train the model and the last one was used to validate the model. Because the dataset was a small sample set and the parameters were nonlinear, we chose SVR as the regression algorithm and a radial basis function (RBF)<sup>22</sup> as the kernel function. In the parameter optimization of SVR, the insensitive loss coefficient  $\#$  was empirically set as 0.05, and the penalty coefficient  $C$  and the width coefficient  $\gamma$  could be optimized with PSO. After searching for the optimal parameters, the corresponding SVR was optimal and the PCA-PSO-SVR prediction model was established as well.

### 3 RESULTS AND DISCUSSION

During the data pre-processing, we adopted the PCA to process the dataset for reducing the dimensions and interdependencies among the parameters. In order to select the optimal reduced dimensions, the calculated eigenvalues of the covariance matrix of the parameters and the corresponding contribution rates  $C$  were sorted and listed in **Tables 1** and **2**. From **Table 2** and **Figure 1**, we can see that with an increase in the reduced dimensions from 1 to 7, the contribution rates also increase. There is a significant improvement from 3 dimensions to 4 dimensions, that is, the contribution rate of 4 dimensions reaches 96.23 % while the contribution rate of 3 dimensions reaches 89.57 %. Meanwhile, when the number of reduced dimensions is more than 4, the contribution rate is close to 100 %, which means that the loss rate is close to 0. Generally speaking, when the contribution rate is over 95 %, the corresponding reduced dimensions of the parameters can represent the original parameters well. In addition, we also made a holistic performance analysis of the impact of the dimensions after adopting the PCA for the proposed model.

Different reduced dimensions from 1 to 7 were used to train and establish the PCA-PSO-SVR prediction model, and every sample of the dataset was used to test each model and obtain the predicted values. Although we do not show those specific predicted values of each sample, we show the mean absolute error (MAE) and root

mean square error (RMSE) of the proposed model, with the dimensionality varying from 1 to 7 in **Figure 2**. When the dimensionality gradually increases to 4, the MAE and RMSE decrease. However, when the dimensionality is bigger than 4, the MAE and RMSE are larger than in the case of the dimensionality being 4. In other words, both MAE and RMSE are minimal when the dimensionality is 4. The reason why 4 dimensions are the best according to the PCA can be explained as follows: when the number of reduced dimensions is smaller than 4, the corresponding contribution rate is lower than 90 %. Therefore, it loses too much information hidden in the original dataset and the result is certainly not accurate. When the number of reduced dimensions is over 4, though the contribution rate is obviously higher than that of 4-dimension, more parameters mean more noise and interference. Thus, 4-dimensional parameters were selected for the regression process based on the comprehensive result analysis of the contribution rate and errors.

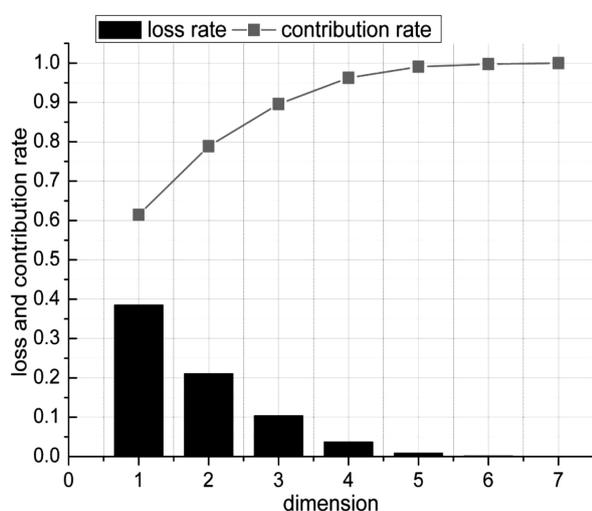
In addition to trianing and validating the PCA-PSO-SVR model with the processed dataset, we trained and validated the PSO-SVR and BPNN model with the original dataset. The predicted  $T_C$  of each sample obtained with three different models is shown in **Figure 3**; the specific values listed in the Data.docx file and the corresponding absolute errors are also presented. It can be seen that many sample points deviate from the standard line in **Figure 3a**; in other words, the absolute

**Table 1:** Eigenvalues of the covariance matrix after sorting

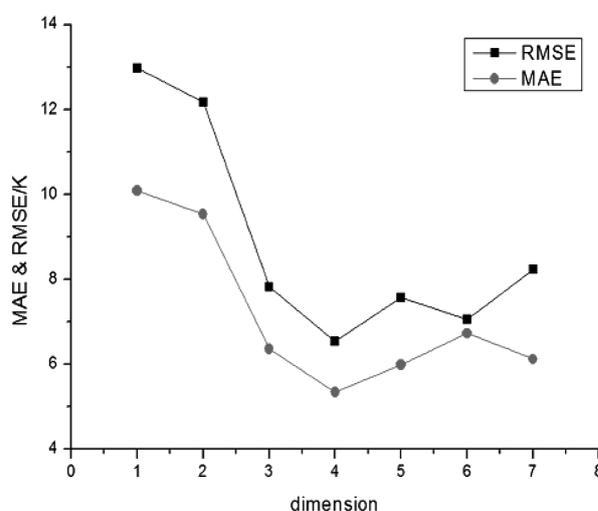
Number	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\lambda_5$	$\lambda_6$	$\lambda_7$
Eigenvalue	0.4257	0.1207	0.0743	0.0462	0.0197	0.0047	0.0017

**Table 2:** Contribution rates of the dimensions from 1 to 7

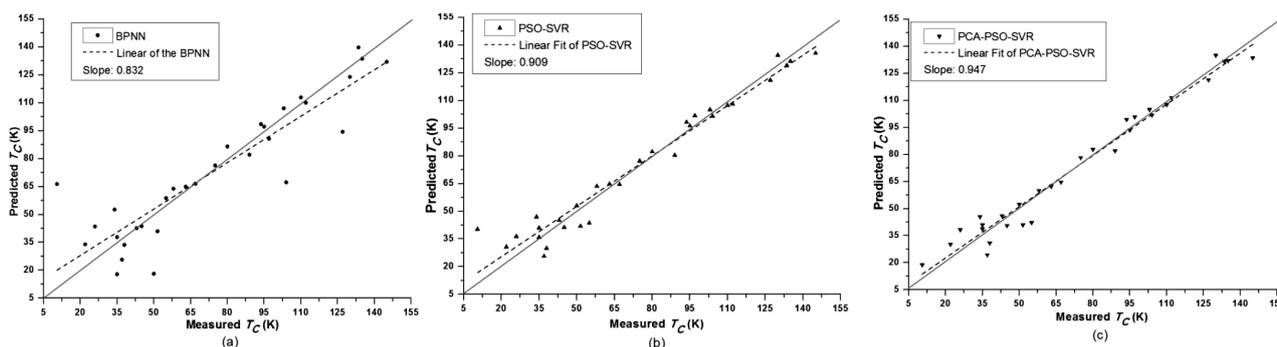
Dimension	1	2	3	4	5	6	7
Contribution rate	0.6143	0.7885	0.8957	0.9623	0.9907	0.9976	1.0000



**Figure 1:** Loss rate and contribution rate of the dimensionality from 1 to 7



**Figure 2:** MAE and RMSE of the proposed model with dimensionality from 1 to 7



**Figure 3:** a), b) and c) show the correlation between the measured  $T_C$  and  $T_C$  predicted by BPNN, PSO-SVR and PCA-PSO-SVR, respectively. 31 kinds of high- $T_C$  superconductors were used with three different methods to predict  $T_C$ , represented by circles, up-triangles and down-triangles in every subfigure; the black dashed line represents the corresponding fit line and the red solid line is the standard line. The slope of the fit line was used to determine the performance of the corresponding method, and the three fit-line slopes are 0.832, 0.909 and 0.947, respectively.

**Table 3:** Comparison of the prediction performance of BPNN, PSO-SVR and PCA-PSO-SVR

methods	MAE/K	MAPE/%	RMSE/K	$R$
BPNN	10.59	23.46 %	16.44	0.8972
PSO-SVR	6.15	12.56 %	8.23	0.9745
PCA-PSO-SVR	5.34	11.85 %	6.54	0.9843

**Table 4:** Measured  $T_C$ , the  $T_C$  predicted with the PCA-PSO-SVR and the corresponding absolute error

No	1	2	3	4	5	6	7
Measured $T_C$ /K	18.0	10.2	6.3	6.9	17.0	16.0	9.5
Predicted $T_C$ /K	22.6697	10.2721	6.7415	6.5941	21.6752	21.1690	9.9879
MAE/K	4.6697	0.0721	0.4415	0.3059	4.6752	5.1690	0.4879

errors of the samples predicted with BPNN are so large that this prediction model is not suitable. Comparing **Figure 3b** with **Figure 3c**, more sample points are presented with the PCA-PSO-SVR model and they are closer to the standard line than the sample points of the PSO-SVR model. Specifically, the accuracy of 19/31 samples obtained with PCA-PSO-SVR is better than that of PSO-SVR; especially for the leftmost sample point, the absolute error dropped from 30 K to 8 K. Based on the singularity of the leftmost sample point, we can say that it is the only organic superconductor that is very different from the others in the dataset and the values of some of its parameters are much bigger than those of the corresponding parameters of the other samples, leading to a big prediction error. Because of the data pre-processing with PCA, the influence of the parameters with large values on the predicted results becomes smaller after projection, so the corresponding absolute error dropped a lot. Meanwhile, we can also see that the fit line of PCA-PSO-SVR is closer to 1 than the other two fit lines, which means that its accuracy is better. The performance of each model can also be analysed statistically as shown with **Table 3**, which includes MAE, the mean absolute percentage error (MAPE), RMSE and the correlation coefficient ( $R$ ). It can be found that the PCA-PSO-SVR index is the best in all three models, being 5.34 K, 11.85 %, 6.54 k and 0.9843, respectively. Based on the above analysis, the proposed PCA-PSO-SVR model is very suitable to predict the  $T_C$  for the dataset.

We added seven  $A_x(S)_yTiNCl$  compound high- $T_C$  superconductors,<sup>22</sup> which had structural characteristics similar to those of the preceding dataset. By reading and analysing the literature, we extracted the required data, included in the Data.docx file. We used the new data to validate the proposed PCA-PSO-SVR prediction model, and the corresponding predicted values and MAE are included in **Table 4**. Very small MAEs were found for four of the seven samples. Compared with the previous predicted results of the developing prediction model, the current predicted results for all the samples are reasonable. In other words, the PCA-PSO-SVR prediction model exhibited a good accuracy for the above additional data.

## 4 CONCLUSIONS

In this paper, we provided a PCA-PSO-SVR model for predicting  $T_C$  from structural and correlative electronic parameters of high- $T_C$  superconductors. SVR was adopted to deal with the dataset, which was a small sample set, and the PSO algorithm was utilized to search for its optimal parameters to achieve a good performance. The PCA was employed to reduce dimensions and interdependencies between the parameters, and the selected optimal dimensions of the parameters were subsequently utilized in PSO-SVR to train and validate the regression model. In addition, we also trained a PSO-SVR model without the PCA and BPNN, with the

dataset used for comparison. According to the assessment results and comparison, the PCA-PSO-SVR model provided a better accuracy of prediction than the other models for the dataset, and the corresponding MAE was 5.34 k. At last, additional data was used to validate the prediction, and the results were also reasonable. In a word, machine-learning methods can be applied to some domains of materials and the PCA-PSO-SVR ensemble method may be used to predict the  $T_C$  of new high- $T_C$  superconductors.

## Acknowledgment

We acknowledge the support from the National Key Research and Development Program of China (no. 2016yfb0700502) and the research grants (no. 14dz2261200, no. 15dz2260300, no. 16142203000 and no. 16511101200) from the Science and Technology Commission of Shanghai Municipality and the Innovation Program of Shanghai Municipal Education Commission (14YZ024).

## 5 REFERENCES

- <sup>1</sup> J. G. Bednorz, K. A. Müller, Possible high  $T_C$  superconductivity in the Ba-La-Cu-O system, *Z. Phys. B: Condensed Matter*, 64 (1986) 2, 189–193
- <sup>2</sup> F. Ben Azzouz et al., Structure, microstructure and transport properties of B-doped YBCO system, *Physica C*, 442 (2006), 13–19, doi:10.1016/j.physc.2006.03.135
- <sup>3</sup> H. Su, D. O. Welch, The effects of space charge, dopants, and strain fields on surfaces and grain boundaries in YBCO compounds, *Supercond. Sci. Technol.*, 18 (2005), 24–34, doi:10.1088/0953-2048/18/1/005
- <sup>4</sup> D. Volochová et al., Time dependent changes in Ag doped YBCO superconductors, *Acta Physica Polonica A*, 118 (2010), 1047–1048, doi:10.12693/APhysPolA.118.1047
- <sup>5</sup> P. Paturia, H. Palonen, H. Huhtinen, Properties of Pr- and BZO-doped YBCO multilayers, *Physics Procedia*, 36 (2012), 661–664, doi:10.1016/j.phpro.2012.06.263
- <sup>6</sup> D. R. Harshman, A. P. Mills Jr., Concerning the nature of high- $T_C$  superconductivity: survey of experimental properties and implications for interlayer coupling, *Phys. Rev. B*, 45 (1992) 18, 10684
- <sup>7</sup> D. R. Harshman, A. T. Fiory, J. D. Dow, Theory of high- $T_C$  superconductivity: transition temperature, *J. Phys.: Condens. Matter*, 23 (2011), 295701, doi:10.1088/0953-8984/23/29/295701
- <sup>8</sup> J. J. Ying, X. F. Wang, X. G. Luo et al., Superconductivity and magnetic properties of single crystals of  $K_{0.75}Fe_{1.66}Se_2$  and  $Cs_{0.81}Fe_{1.61}Se_2$ , *Phys. Rev. B*, 83 (2011) 21, 212502
- <sup>9</sup> D. R. Harshman, A. T. Fiory, Charge compensation and optimal stoichiometry in superconducting  $(Ca_x La_{1-x})(Ba_{1.75-x}La_{0.25+x})Cu_3O_y$ , *Phys. Rev. B*, 86 (2012), 144533
- <sup>10</sup> G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning, *Frontiers in Materials*, 3 (2016), doi:10.3389/fmats.2016.00019
- <sup>11</sup> P. Raccuglia et al., Machine-learning-assisted materials discovery using failed experiments, *Nature*, 533 (2016), 73–76, doi:10.1038/nature17439
- <sup>12</sup> G. Pilania, P. V. Balachandran, J. E. Gubernatis, T. Lookman, Classification of  $ABO_3$  perovskite solids: a machine learning study, *Acta Crystallographica*, 71 (2015), 507–513, doi:10.1107/S2052520615013979
- <sup>13</sup> J. Ling, R. Jones, J. Templeton, Machine learning strategies for systems with invariance properties, *Journal of Computational Physics*, 318 (2016), 22–35, doi:10.1016/j.jcp.2016.05.003
- <sup>14</sup> V. N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995
- <sup>15</sup> T. O. Owolabi, K. O. Akande, S. O. Olatunji, Estimation of superconducting transition temperature  $T_C$  for superconductors of the doped MgB2 system from the crystal lattice parameters using support vector regression, *Journal of Superconductivity and Novel Magnetism*, 28 (2015), 75–81, doi:10.1007/s10948-014-2891-7
- <sup>16</sup> T. O. Owolabi, K. O. Akande, S. O. Olatunji, Application of computational intelligence technique for estimating superconducting transition temperature of YBCO superconductors, *Applied Soft Computing*, 43 (2016), 143–149, doi:10.1016/j.asoc.2016.02.005
- <sup>17</sup> R. Eberhart, J. Kennedy, A new optimizer using particle swarm Theory, *International Symposium on Micro Machine and Human science*, IEEE, 1995, 39–43
- <sup>18</sup> I. T. Jolliffe, *Principal component analysis*, Springer, Berlin, 87.100 (1986), 41–64
- <sup>19</sup> D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, *Neurocomputing: foundations of research*, MIT Press, (1988), 533–536
- <sup>20</sup> S. Peng et al., Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, *FEBS Letters*, 555 (2003), 358–362
- <sup>21</sup> M. D. Buhmann, *Radial basis functions: Theory and implementations*, Cambridge Monographs on Applied and Computational Mathematics, 2003, 259
- <sup>22</sup> D. R. Harshman, A. T. Fiory, Comment on "superconductivity in electron-doped layered  $TiNiCl$  with variable interlayer coupling", *Phys. Rev. B*, 90 (2014) 18, 186501, doi:10.1103/PhysRevB.90.186501