Data Analysis of Multi-Dimensional Thermophysical Properties of Liquid Substances Based on Clustering Approach of Machine Learning

Gota Kikugawa,[1] Yuta Nishimura,[2] Koji Shimoyama,[1] Taku Ohara,[1]
Tomonaga Okabe, [3] and Fumio S. Ohuchi [4]
[1]Institute of Fluid Science, Tohoku Univ., 2-1-1 Katahira, Aoba-ku, Sendai, 980-8577
[2]School of Engineering, Tohoku Univ., 2-1-1 Katahira, Aoba-ku, Sendai, 980-8577
[3]School of Engineering, Tohoku Univ., Aobayama, Aoba-ku, Sendai, 980-8579
[4]Department of Materials Science and Engineering, University of Washington, Seattle, WA, 98195

In an attempt to develop an efficient framework for visualization of the relationships between various liquids and for global screening in the material exploration, we performed a clustering approach of machine learning for the multi-dimensional thermophysical properties of the liquid substances. Data mining using a self-organizing map (SOM) based on the unsupervised learning was employed to project high-dimensional thermophysical data onto a low-dimensional space. Here we adopted 98 liquid substances with eight thermo-physical properties as input data for the SOM training. We also applied a clustering analysis to output weighted vectors (trained data) on each SOM node in order to group the liquid substances. The present SOM-clustering approach realizes consistent grouping of the liquids in terms of their thermo-fluid properties, i.e., liquid substances are categorized according to the chemical species characterized by the functional groups. In addition, our approach provides useful knowledge for designing and exploring the liquid substances.

Keywords: Self-organizing map, Clustering analysis, Machine learning, Thermophysical properties, Heat medium

1. INTRODUCTION

Technology to design and synthesize liquid substances that have appropriate thermophysical properties for specific applications is an important subject in a wide variety of scientific and technical fields. At the industrial level, efficient development and exploration of heat-transfer fluids and phase-change materials, which are used for heat exchangers in refrigeration and air-conditioning equipment or for thermal storage, are highly required these days. In these areas, prior knowledge-based survey and/or conventional empirical laws for materials properties are utilized for the design guideline; however, when a novel materials with superior properties is developed, a trial-and-error approach is needed since the required specification is so multi-objective, or the empirical laws is not always effective due to the outside of application limit. For example, if required thermophysical properties of the liquids has been realized by mixing multiple liquids, the number of mixing combinations and variation in the composition of each liquid get enormous, which makes the materials exploration almost impossible. Therefore, an alternative approach is highly needed for design and discovery of

novel liquid substances in more efficient manner.

One promising approach is a data-driven machine learning (ML) technique, which is known as materials informatics (MI) [1] recently. Materials informatics has drawn considerable attentions in various material design partly because application of ML rapidly prevails to a wide spectrum of the scientific and technological field with the aid of advancement of computational performance for ML. So far, MI is intensively applied for accelerating search of inorganic materials like shape memory alloys [2-4] and piezoelectrics [5,6], and for predicting properties of perovskite crystal and finding one having optimal properties [7,8]. Other example is ML-assisted materials design and property prediction of solid materials such as elpasolite [9,10], metallic alloys [11-13], transition metal complex [14], layered double hydroxide [15], high dielectric permittivity materials [16], and thermoelectric materials [17]. As such, these studies have already shown effectiveness of materials informatics, i.e., a use of enormous experimental data and the numerical simulation, in particular, first-principle calculations, gives the accurate prediction of physical properties, and development and exploration of improved materials.

Besides the inorganic materials, organic materials and organic-inorganic hybrid systems are getting targeted as an application of MI [18-24]. Discovery and design of polymeric materials, metal organic frameworks (MOFs), and organic/inorganic perovskite are good candidates for MI subject and actually require the ML-assisted screening or development. For example, polymer dielectrics were studied by ML with properties provided by DFT (density functional theory) calculations [18,19]. A Bayesian optimization technique was applied for proposing better experimental settings of polymer fiber synthesis [20]. For construction of a better structure-property relationship, a large dataset of polymeric materials has been provided online [21,22]. ML-assisted materials development and property prediction for soft matters are also reviewed on ref. [24].

Applications of MI protocols on the design and exploration of the liquid substances, on the other hand, have been very limited so far. As mentioned earlier, the efficient development and exploration of heat-transfer fluids like a coolant is, however, getting required significantly. When it comes to a coolant, the multi-objective design is needed since the various requirement should be fulfilled as an industrial product, some of which are basic transport properties like thermal conductivity and viscosity, non-flammability, and global warming potential (GWP). Therefore, it is not an easy task to find the improved substances and optimize the various properties at once.

In our study, we aim at building the overall platform for the efficient screening of materials candidates, which is constructed by a multi-stage screening protocol. This platform involves “global screening” which enables to roughly screen candidates by using data-mining approach based on unsupervised learning, and precise exploration which is realized by commonly adopted MI techniques based on the structure/properties prediction model and/or structure optimization by desired properties. Our vision regarding the multi-stage screening is mostly motivated from the fact that large amount of dataset is not always available for liquid substances. Therefore, in order to build an accurate prediction model with relatively limited dataset, rough screening prior to precise ML-driven materials finding is effective.

In the present paper, we addressed the global screening stage for exploration of the liquid substances

assisted by ML techniques. To this end, we propose to utilize one of the dimensionality reduction techniques called self-organizing map (SOM) to enable visual understanding of the diverse thermophysical properties from various liquid substances. SOM has been developed and progressively applied in a variety of engineering fields, such as optimal design of aircraft wings [25] and development of thermosetting polymer resin [26]. Here, we present a combined approach of SOM and clustering applied to a wide variety of the liquid substances, and a possibility of our framework for materials screening to enable us to design and explore better liquid substances more easily.

2. DATA ANALYSIS

2.1 Self-Organizing Map (SOM)

SOM provides a data-mining technique that helps to understand high dimensional data, such as various thermophysical properties of the liquid species in the present case, by reducing their dimensions of the data to typically a two-dimensional space, so that the relative relationships among input data can be visualized intuitively [27,28]. In this technique, the data are nonlinearly reduced to a low-dimensional space with maintaining the neighborhood relationships among the input data. This data mining technique is now utilized in various fields to clarify relative relationships of complicated data.

In terms of machine learning, SOM is categorized into unsupervised learning, which is based on a neural network (NN). Therefore, it is interpreted as a feed-forward two-layer NN which is composed of an input layer and an output layer. The input layer corresponds to all dimensions of the all input data (thermophysical properties in this study) and all input-layer nodes are connected to all output-layer nodes. Furthermore, each output-layer node has a vector composed of the thermophysical properties. This is called a weighted vector with the same number of dimensions as that of input data, and is corrected through a competitive learning process. Each node of the output layer holds the coordinate information at low dimensions (two dimensions in this study) concurrently, and the neighborhood relationships among nodes in the output layer are expressed by these coordinates, i.e., projection onto the low-dimension map.

Detailed algorithm of the SOM is given in previous literatures, so we only make a brief explanation below. As the first step of a learning process, a weighted vector from all the output-layer nodes, which has the highest similarity to the input data, is chosen and that node is defined as a winner node. The weighted vector on the winner node is updated in such a way that the vector gets closer to the input data; at the same time, the output-layer nodes near the winner node also become closer to the input data, depending on their proximity to the winner node. By repeating this learning process, similarities among the weighted vectors on the output-layer nodes near the winner node increase. This update process is repeated until the weighted vectors converge to obtain the final SOM. Liquid species with similar thermo-physical properties are finally distributed closer on the SOM.

The learning algorithm of SOM can be roughly categorized into two types: online-learning SOM and batch-learning SOM (BLSOM). This study adopted a batch type, in which the learning result is not affected by an input order of the learning data. Algorithm of BLSOM is outlined below. Let total numbers of the input-layer

and output-layer nodes be I and J, respectively, and component vectors of the input-layer nodes and the weighted vectors of the output layer at the t-th learning cycle are expressed as $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathrm{L}\ , \mathbf{x}_I]$ and $\mathbf{m}(t) = [\mathbf{m}_1(t), \mathbf{m}_2(t), \mathrm{L}\ , \mathbf{m}_J(t)]$, respectively. The output-layer node that has the highest similarity with the input data, i.e., the winner node, $\mathbf{m}_c(t)$, is defined as the node that satisfies the following equation.

$$\left\| \mathbf{x}_k - \mathbf{m}_c(t) \right\| = \min_j \left\| \mathbf{x}_k - \mathbf{m}_j(t) \right\|. \tag{1}$$

Here, a Gaussian function is defined as the neighborhood function at the t-th learning cycle.

$$h_{ic}(t) = \exp\left(-\frac{\left\| \mathbf{r}_c - \mathbf{r}_i \right\|^2}{2\sigma^2(t)} \right), \tag{2}$$

where $\mathbf{r}$ denotes a position vector in the two dimensional SOM (not in the high dimensional space formed by $\mathbf{x}$), and thus $\left\| \mathbf{r}_c - \mathbf{r}_i \right\|$ represents the two dimensional distance between the output-layer node c and a surrounding node i, and $\sigma(t)$ is defined as a linear damping function represented as

$$\sigma(t) = \alpha_0 + (\alpha_1 - \alpha_0)\frac{t}{T}, \tag{3}$$

where T is total number of the learning cycles, and α_0 and α_1 are arbitrary positive numbers ($\alpha_1 < \alpha_0$). Thus, $\sigma(t)$ decreases from α_0 to α_1 with progress of the learning cycles. The winner node is picked up over all input nodes based on Eq. (1), and the sum of component vectors of the input-layer nodes on the winner node is expressed as

$$\mathbf{s}_j(t) = \sum_{\alpha=1}^{w_j} \mathbf{x}_{k(\alpha)}, \tag{4}$$

where w_j represents the number of wins in each output-layer node during a single learning cycle and $k(\alpha)$ corresponds to the index of α-th winner input-layer node on the output-layer node j. An update of the weighted vectors for all output-layer nodes is performed repeatedly until the weighted vectors converge by using the following relation at the t-th learning cycle

$$\mathbf{m}_i(t+1) = \frac{\sum_{j=1}^{J} h_{ij}(t)\mathbf{s}_j(t)}{\sum_{j=1}^{J} w_j h_{ij}(t)}. \tag{5}$$

For the input data to be used in the present analysis, we have chosen 98 liquid substances which are categorized into alkane, alcohol, aromatic series, carboxylic acid, amine, ketone, ester, and halide. Under standard temperature and pressure conditions, these substances are in a liquid phase. In the analysis, the following eight thermophysical properties were chosen; they are (mass) density, specific heat at constant pressure, melting point, boiling point, saturated vapor pressure, surface tension, viscosity, and thermal conductivity, each of which was taken from the handbook and database of liquid properties [29-34]. Note that these thermophysical properties do not have the same averaged value and variance over the tested 98 liquids,

i.e., each property has a different data range. Therefore, they were normalized into those with an average of 0 and a variance of 1 before they were used as the input data for learning. All the data analysis codes were implemented using python language (python 3.5), and the SOMPY package [35] was used.

2.2 Clustering Approach

The output-layer nodes after the SOM learning have weighted vectors whose components are multi-dimensional thermophysical properties, and the nodes, which are mutually close in distance in a high dimensional space, are placed in mutually near places in the two dimensional SOM. Furthermore, in order to easily visualize the nodes with similar thermophysical properties, a clustering technique using a k-means method [36] was performed. This unsupervised learning method is one of non-hierarchical clustering techniques, which is used to classify input data having multi-dimensional quantities into preliminarily defined K clusters. In this method, K clusters' gravity center, i.e., centroid, are given first among the data points so as to be as far apart as possible each other (the k-means++ method), and then all data points are engaged in a cluster governed by the centroid to which the Euclidian distance is closest. Next, the centroid is calculated from the data points belonging to each cluster again, and the data points belonging to each cluster are updated. This process is repeated until the positions of centroids and the data points belonging to each cluster converge.

2.3 U-Matrix Method.

A U-Matrix (unified distance matrix) method [37] is used to visualize the similarity between the adjacent nodes based on the distance information between nearest neighbor output-layer nodes. Here, the U-Matrix is given as

$$U_i = \frac{1}{N_i} \sum_{j}^{N_i} \left\| \mathbf{m}_i - \mathbf{m}_j \right\|, \tag{6}$$

where the output-layer node and the nearest neighbor output-layer node are denoted as i and j, respectively, and N_i is the number of nodes adjacent to node i. The U-Matrix gives local distance relationships between the nodes, and a large U-Matrix value would be regarded as a boundary of the clusters where separation between the nodes are large. In this study, the U-Matrix is used for examining "quality" of the clustering result by using the k-means method.

3. RESULTS OF SOM-CLUSTERING APPROACH

The SOM learning was performed on aforementioned 98 liquid substances with eight thermophysical properties. A positioning map where each substance is assigned to the output-layer node that has the weight vector closest to each substance is presented using a 30×30 map in Fig. 1. The clustering result obtained using the k-means method as described in Section 2.2 is represented by the individual color designating each cluster. Originally, an attempt was made to determine the number of clusters through the Elbow method and Silhouette analysis [36]; the number adopted in the present study was 10, selected among those which have relatively good (not best) results of the Silhouette analysis (details not shown here). The liquid substances classified into

clusters are as follows.

1) Water (yellow-green, lower right) and glycerol (red, middle right side) belong to independent clusters.
2) Liquid species having small molecular weights and likely to form hydrogen bonding, such as ethylene glycol and ethanolamine (yellow, lower right)
3) Alcohol and alkane liquids with relatively large molecular weights (blue, right)
4) Alcohol and alkane liquids with relatively small molecular weights (light blue, bottom)
5) Aromatic series and cycloalkane (purple, top center)
6) Alkane smaller than 4), unsaturated alkane, and ketone (pink, lower left)
7) Halogen compounds with relatively small molecular weights (orange, upper left)
8) Halogen compounds with relatively large molecular weights (green, upper left)
9) Heavy halogen compounds substituted for two bromine and/or iodine molecules (light purple, top center)

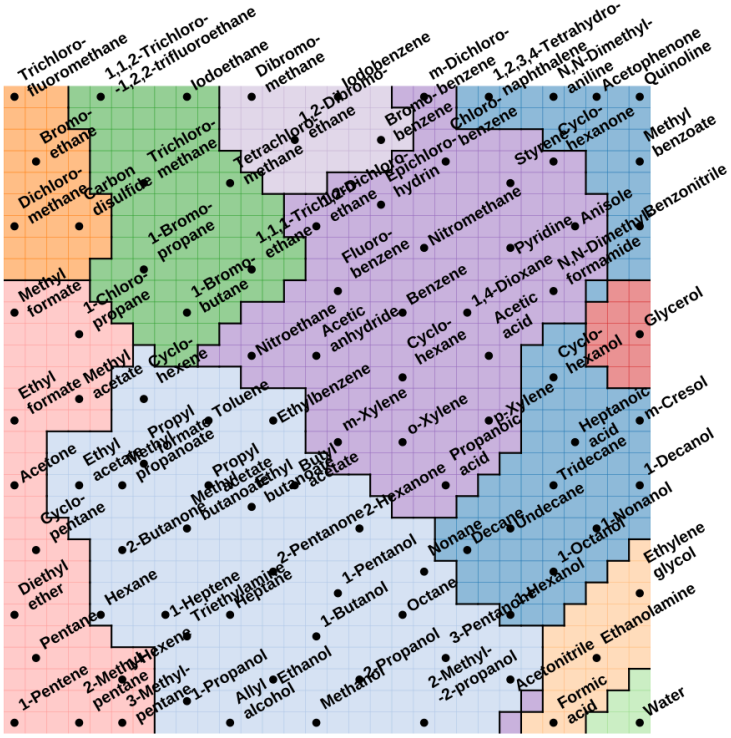

Fig.1 SOM positioning map obtained from thermophysical properties of various liquids. The output nodes are classified via the k-means method and colored separately by each cluster.

As a whole, The SOM indicates that the liquid substances characterized by each functional group are properly classified into groups. It is found that the liquid species with special thermophysical properties such as water or glycerol are automatically distinguished as an individual cluster and that halogen compounds are clearly separated as other clusters. In addition, cyclohexane and benzene are placed in mutually near positions in SOM, although their electronic structures are different with each other.

Figure 2 depicts results from the U-Matrix analysis. Red and blue regions of the figure represent higher and lower values of the U-Matrix, respectively. Therefore, liquid substances in a blue valley separated by the red ridge have the larger deviation in thermophysical properties. The result indicates that some liquid substances, which are classified into the same cluster by the k-means method as shown in Fig. 1, have a large distance in terms of thermophysical properties. This is because a few numbers of the input data are used for learning as compared to the defined number of clusters employed here. It will be necessary to enlarge the number of liquid species for better examination of SOM-clustering approach in future.

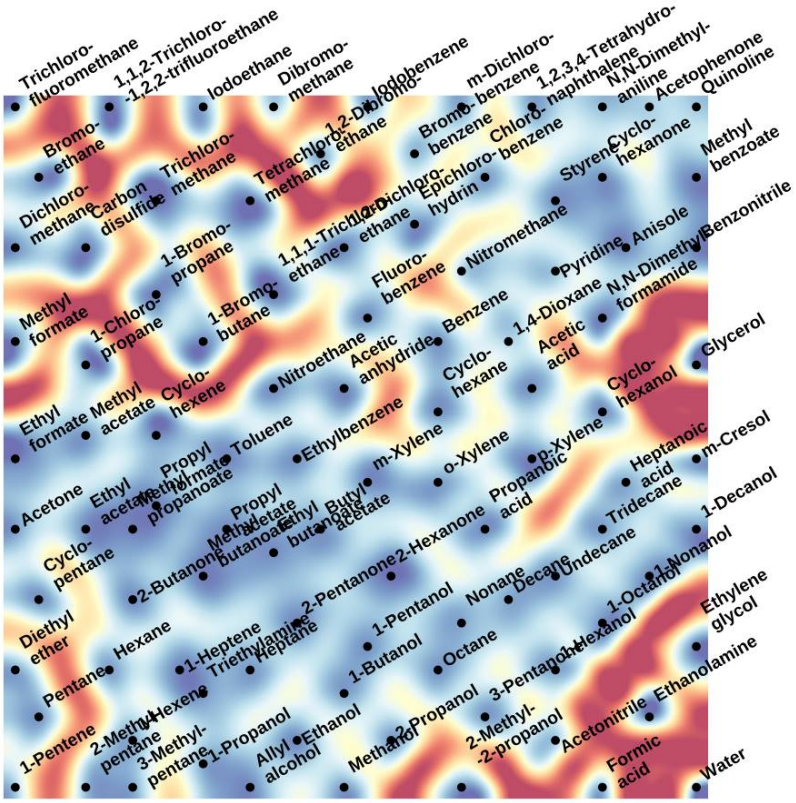

Fig. 2 U-Matrix values on output nodes obtained from the weighted vectors. Red and blue regions of the figure represent higher and lower values of the U-Matrix, respectively. Each liquid species is placed at the same location as in Fig. 1.

4. CONSIDERATION BASED ON EACH THERMOPHYSICAL PROPERTY

While the SOM itself can provide information concerning distant relationships of the input data, more detailed information can be extracted by analyzing distributions of each thermophysical properties, i.e., each component of the weighted vectors, projected on the SOM. To briefly visualize the correlations between various thermophysical properties, a correlation matrix for all the thermophysical properties just obtained from the input data is presented in Fig. 3. In this figure, we can find obvious strong correlations, some of which are the positive correlation between boiling point and melting point and negative correlation between boiling point

and saturated vapor pressure as expected. Meanwhile interesting but not straightforward correlations are also found between the specific heat and thermal conductivity, the surface tension and thermal conductivity, and density and specific heat for the tested liquids. In contrast to the strongly correlated properties, several thermophysical properties are less correlated, for example, between vapor pressure and density. This weak correlation suggests that thermophysical properties are not mutually in a trade-off relationship, which implies a possibility in designing and searching the liquid substances with a higher degree of freedom.

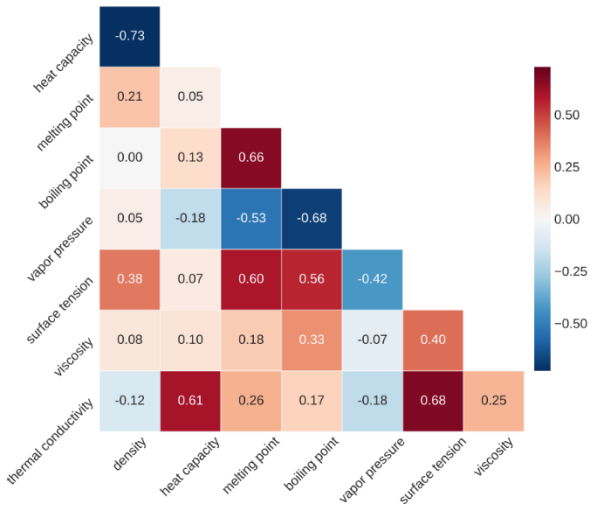

Fig. 3 Correlation matrix for all the thermophysical properties. Each value depicted in the boxes indicates the correlation coefficient between two thermophysical properties.

Figure 4 shows various thermophysical properties represented as “heat maps” after the SOM learning. In each figure, the positions of liquid substances and the cluster boundaries are identical to those shown in the SOM positioning map (Fig. 1). Therefore, one can get a grasp of a whole distribution of various thermophysical properties and correlations of various liquid substances at a glance. For example, these heat maps immediately give a strong negative correlation between density and heat capacity, and specifically this correlation gets stronger near at the cluster of halogen compounds (upper left). The result implies that halogenation of the liquid molecule greatly influences specific thermophysical properties and transport properties. In terms of searching heat-transfer fluids, liquid substances having low viscosity and high thermal conductivity are ideal, i.e, the high heat transfer coefficient can be obtained with low pumping power. Such materials can be readily found in Fig. 4 like methanol and allyl alcohol at the bottom center and formic acid at the lower right as examples. In addition, if heat-transfer fluids are used as a working fluid in a heat pipe (HP), latent heat of vaporization is actively utilized to remove emitted thermal energy effectively. In this case, the boiling temperature should be within an appropriate range, taking into account operating conditions. Assuming a HP is operated under the atmospheric pressure and for cooling of electronic devices, liquid substances having high thermal conductivity and low boiling point (in a temperature range shown in Fig. 4) are desirable like methyl formate or methanol. Although many other restricting conditions, e.g., inflammability, toxicity, and environmental impact must be considered in actual applications, a highly practicable data mining technique can be achieved by analyzing such conditions

and by incorporating them into the framework of SOM.

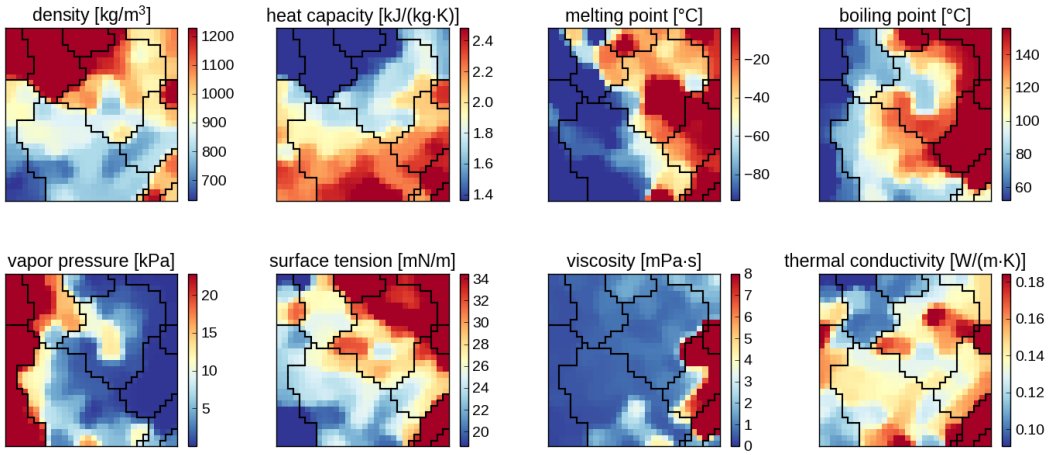

Fig. 4 Heat maps of thermophysical properties projected on the SOM. The cluster boundaries using k-means clustering, which are identical to those in Fig. 1, are also drawn in the figure.

5. CONCLUSION

We applied machine learning protocols to multi-dimensional thermophysical data of liquid substances, and the liquid substances were mapped in two-dimensional space using the SOM and clustering techniques to visualize the proximity relationship. It was demonstrated that this framework enabled us to easily understand relative relationships of the thermophysical properties among liquid substances. Grouping of the liquid substances was successfully achieved via the clustering approach using the k-means method which was applied to the weighted vectors on output-layer nodes in the SOM. Various thermophysical properties of the tested liquid substances were represented as “heat maps” in the SOM. This visualization gives precise correlations among thermophysical properties of a lot of liquid substances at a glance. Thus, the presented framework is useful for quick design and exploration of the candidates of liquid materials for specific applications.

ACKNOWLEDGEMENT

We thank Dr. Yutaka Oya for a fruitful discussion. A part of this study was conducted by the Strategic Innovation Creating Program (SIP) of the Comprehensive Science and Technology Innovation Conference “Development of Manufacture and Quality Control Technologies for High-productivity Innovative PMC for Aircraft” (managing entity: JST). We also acknowledge vitally important encouragement and support made through the University of Washington-Tohoku University: Academic Open Space (UW-TU:AOS).

REFERENCES

[1] Rajan, K., Materials Today, 9, pp. 38-45, (2005).

[2] Xue, D., Balachandran, P.V., Hodgen, J., Theiler, J., Xue, D., and Lookman, T., Nature Comm., 7, 11241, (2016).

[3] Xue, D., Xue, D., Yuan, R., Zhou, Y., Balachandran, P.V., Ding, X., Sun, J., and Lookman, T., Acta Materialia, 125, pp. 532-541, (2017).

[4] Gopakumar, A.M., Balachandran, P.V., Xue, D., Gubernatis, J.E., and Lookman, T., Sci. Rep., 8, 3738, (2018).

[5] Yuan, R., Liu, Z., Balachandran, P.V., Xue, D., Zhou, Y., Ding, X., Sun, J., Xue, D., and Lookman, T., Adv. Mater., 30, 1702884, (2018).

[6] Xue, D., Balachandran, P.V., Yuan, R., Hu, T., Qian, X., Dougherty, E.R., and Lookman, T., PNAS, 113, pp. 13301-13306, (2016).

[7] Schmidt, J., Shi, J., Borlido, P., Chen, L., Botti, S., and Marques, M.A.L., Chem. Mater., 29, pp. 5090-5103, (2017).

[8] Pilania, G., Balachandran, P.V., Kim, C., and Lookman, T., Frontiers in Materials, 3, 19, (2016).

[9] Faber, F.A., Lindmaa, A., von Lilienfeld, O.A., Armiento, R., Phys. Rev. Lett., 117, 135502, (2016).

[10] Pilania, G., Gubernatis, J.E., Lookman, T., Comput. Mater. Sci., 129, pp. 156-163, (2017).

[11] Fischer, C.C., Tibbetts, K.J., Morgan, D., and Ceder, G., Nature Mater., 5, pp. 641-646, (2006).

[12] Sun, Y.T., Bai, H.Y., Li, M.Z., and Wang, W.H., J. Phys. Chem. Lett., 8, pp. 3434-3439, (2017).

[13] Liu, R., Kumar, A., Chen, Z., Agrawal, A., Sundararaghavan, V., and Choudhary, A., Sci. Rep., 5, 11552, (2015).

[14] Janet, J.P. and Kulik, H.J., Chem. Sci., 8, pp. 5137-5152, (2017).

[15] Hu., B., Lu, K., Zhang, Q., Ji, X., and Lu, W., Comput. Mater. Sci., 136, pp.29-35, (2017).

[16] Gao, J., Liu, Y., Wang, Y., Hu, X., Yan, W., Ke, X., Zhong, L., He, Y., and Ren, X., J. Phys. Chem. C, 121, pp. 13106-13113, (2017).

[17] Gaultois, M.W., Oliynyk, A.O., Mar, A., Sparks, T.D., Mulholland, G.J., and Meredig, B., APL Mater., 4, 053213, (2016).

[18] Mannodi-Kanakkithodi, A., Pilania, G., Huan, T.D., Lookman, T., and Ramprasad, R., Sci. Rep., 6, 20952, (2016).

[19] Mannodi-Kanakkithodi, A., Huan, T.D., and Ramprasad, R., Chem. Mater., 29, pp. 9001-9010 , (2017).

[20] Li, C., de Celis Leal, D.R., Rana, S., Gupta, S., Sutti, A., Greenhill, S., Slezak, T., Height, M., and Venkatesh, S., Sci. Rep., 7, 5683, (2017).

[21] PolyInfo; URL: http://polymer.nims.go.jp/

[22] Huan, T.D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G., and Ramprasad, R., Sci. Data, 3, 160012, (2016).; URL: http://khazana.uconn.edu/

[23] Raccuglia, P., Elbert, K.C., Adler, P.D.F., Falk, C., Wenny, M.B., Mollo, A., Zeller, M., Friedler, S.A., Schrier, J., and Norquist, A.J., Nature, 533, pp. 73-77, (2016).

[24] Ferguson, A.L., J. Phys.: Condens. Matter, 30, 043002, (2018).

[25] Shimoyama, K. and Kamisori, K., J. Aircraft, 54, pp. 1317-1327, (2017).

[26] Oya, Y., Kikugawa, G., and Okabe, T., Macromole. Theory Simul., 27, 1600072, (2017).

[27] Kohonen, T., "Self-Organizing Maps 3rd ed.", Springer-Verlag, (2001).

[28] Obermayer, K. and Sejnowski, T.J. (ed.), "Self-Organizing Map Formation: Foundations of Neural Computation", The MIT press, (2001).
[29] Haynes, W.M., "CRC Handbook of Chemistry and Physics 93rd Edition", CRC Press, (2012).
[30] http://tpds.db.aist.go.jp/tpds-web/, accessed on Jan. 15, 2017.
[31] http://webbook.nist.gov/chemistry/, accessed on Jan. 15, 2017.
[32] https://pubchem.ncbi.nlm.nih.gov/, accessed on Jan. 15, 2017.
[33] http://www.chemspider.com/, accessed on Jan. 15, 2017.
[34] https://scifinder.cas.org/scifinder/, accessed on Jan. 15, 2017.
[35] https://github.com/sevamoo/SOMPY, accessed on Dec. 14, 2016.
[36] Raschka, S., "Python Machine Learning", Packt Publishing, (2015).
[37] Ultsch, A., Guimaraes, G., Korus, D., and Li, H., Proc. Transputer Anwender Treffen/World Transputer Congress TAT/WTC 93 Aachen, pp. 194-203, (1993).