Search for oxide glass compositions using Bayesian optimization with elemental-property-based descriptors

Kensaku NAKAMURA$^{1,†}$, Naoya OTANI$^1$ and Tetsuya KOIKE$^1$

$^1$Mathematical Sciences Research Laboratory, Nikon Corporation, 2–15–3 Konan, Minato-ku, Tokyo 108–6290, Japan

Our study shows that machine learning technique, Bayesian optimization (BO) can efficiently find high-refractive-index glasses from a large number of candidate compositions using data from the INTERGLAD database. The effect of the parameters (i.e., descriptors) input to the BO algorithm on search performance is described. The results show that elemental-property-based (EPB) descriptors, recently applied in materials science, are more effective than the component-amount-based ones traditionally used in the study of glass. The results suggest that BO with EPB descriptors can accelerate the search for glass compositions with desirable properties.

©2020 The Ceramic Society of Japan. All rights reserved.

Key-words : Glass, Refractive index, Machine learning, Bayesian optimization, Descriptors

Oxide glass has many applications, such as lenses, optical fibers, and electric displays. The control of glass composition is a key process in the development of glass materials, because compositions strongly determines the physical properties of glass, such as the refractive index, density, and glass transition temperature.$^{1,2}$ The search for new glass compositions with desirable physical properties generally requires a lot of experiments and is thus time-consuming.

Machine learning techniques are expected to accelerate the study of materials. They have been applied to many fields of materials science.$^{3–11}$ Bayesian optimization (BO), a data-driven method, has been extensively studied.$^{3–5}$ BO finds materials with the desirable properties by repeating the recommendation of the next candidate (e.g., compound composition or synthetic condition) through machine learning and the observation of the candidate from an experiment or simulation. The advantage of BO is that BO can find materials in the extrapolation area (the outside of observed data) by using uncertainties. Other machine learning methods focus on the interpolate area (the inside of observed data). Therefore, BO is expected to find new materials that are not derived from only experience and knowledge. In the BO algorithm, the candidate is chosen with predicted values and their uncertainties (variances or standard deviations) calculated from previous observed data with specific regressions such as Gaussian process regression. By considering uncertainties, the algorithm can recommend a candidate that is very different from the observed candidates.

Here, we apply BO to the search for glass compositions with targeted properties and examine its performance in the study of glass. We seek compositions with a high refractive index ($n_d$) of 587.6 nm. Because control of the refractive index by composition modification is necessary for manufacturing optical glass, it is worthwhile to apply BO for finding compositions with specific optical properties. To the best of our knowledge, there are no previous reports on the application of BO to the study of oxide glass. We discuss the types of descriptors that are input as parameters to machine learning algorithms. Generally, the selection of descriptors is an important factor that determines the prediction performance of machine learning.$^{10,12}$ Recent studies on materials with machine learning, have used descriptors based on elemental properties (e.g., atomic number, electronegativity, and melting temperature) to predict material properties.$^{3,6,10–12}$ The component content in compositions has traditionally been used as a descriptor in glass science, and glass scientists have guessed the physical properties of glass assuming a linear relation between the properties and such descriptors.$^{8,13–15}$ Here, we refer to these descriptors as elemental-property-based (EPB) and component-amount-based (CAB) descriptors, respectively. Some reports have recently shown that glass properties can be well predicted using neural network regression, a nonlinear method, with CAB descriptors.$^{7,8}$ EPB descriptors is considered to improve the performance of machine learning, because the physical properties of glass depend on those of elements in the glass. For example, the refractive index relates to the polarization of the elements thorough the Lorentz-Lorentz equation.$^{1,2}$ Although some studies have used EBP descriptors for investigating metallic glass,$^{9,11}$ here are
no reports on the application of EPB descriptors to the study of oxide glass.

We obtained data for oxide compositions and \( n_d \) values from the database INTERGLAD.\(^{16}\) Some compositions with obviously wrong \( n_d \) values were found in the data. Specifically, there were compositions with \( n_d \) values of more than 10 and the values are unreasonably high. Therefore, those compositions were eliminated as wrong data. Data for elemental physical properties for creating EBP descriptors were obtained the literature.\(^{17)}\) We used compositions that contained elements whose physical properties were available. For example, the element Eu was missing electron negativity values in the handbook we used for collecting elemental property data, and thus compositions that included Eu were not used.\(^{17)}\) We collected a total of 13,400 compositions and the corresponding \( n_d \) values. The range of \( n_d \) values is between 1.45 and 2.65. Figure 1 shows a histogram of the \( n_d \) values for these compositions.

In this study, the Python library GPy was used for implementing nonlinear regression, namely, Gaussian process regression with a Gaussian kernel.\(^{20)}\) Gaussian process regression is a Bayesian-based approach that is typically used in BO to calculate predicted values and their uncertainties (standard deviations) from data. To obtain the next candidate, we used the expected improvement \( a_{EI} \) as the criterion. It is calculated from the predicted value \( \mu \) and the standard deviation \( \sigma \) by the Gaussian process as follows:\(^{3,5)}\)

\[
a_{EI} = \left[ \frac{\phi(z)}{z} + \Phi(z) \right],
\]

\[
z = \frac{(\mu - \mu^*)}{\sigma},
\]

where \( \mu^* \) is the maximum value in the observed training data, \( \phi \) is the probability density function, and \( \Phi \) is the cumulative distribution function for the normal distribution. The composition with the highest value of \( a_{EI} \) becomes the next candidate.

22 EPB descriptors were prepared for each composition.\(^{10,12)}\) We used 11 elemental properties: atomic number, Mendeleev number, the column and row numbers in the periodic table, covalent radius, Arrhenius ionic radius, electronegativity, first ionization energy, melting temperature, atomic weight and density. The reason for choosing these properties is that these properties are used in the studies of material informatics in common.\(^{3,6)}\) Also, because most of these physical properties are usually listed in the periodic table, it is easy to obtain the values. Furthermore, the glass compositions contain various elements and therefore it is convenient for our study that these physical properties are known for many elements. Then, we calculated two descriptors, namely the mean \( f_{\text{mean}} \) and the standard deviation \( f_{\text{std}} \), for each elemental property as follows:

\[
f_{\text{mean}} = \sum_i f_i \frac{x_i}{\sum_i x_i},
\]

\[
f_{\text{std}} = \sqrt{\frac{\sum_i (f_i - f_{\text{mean}})^2 x_i}{\sum_i x_i}},
\]

where \( x_i \) is the atomic fraction in each composition expressed as an atomic percentage, \( f_i \) is the value of the elemental property, and \( i \) represents the element species. We also used 66 CAB descriptors. The number of CAB descriptors corresponds to the number of oxide component types in the studied compositions expressed as a molar fraction. The used oxide component types are follows: \( \text{Ag}_2\text{O}, \text{Al}_2\text{O}_3, \text{As}_2\text{O}_3, \text{B}_2\text{O}_3, \text{BaO}, \text{BeO}, \text{Bi}_2\text{O}_3, \text{CaO}, \text{CdO}, \text{Ce}_2\text{O}_3, \text{CeO}_2, \text{Co}_2\text{O}_3, \text{Co}_3\text{O}_4, \text{CoO}, \text{Cr}_2\text{O}_3, \text{Cs}_2\text{O}, \text{CuO}, \text{Dy}_2\text{O}_3, \text{Er}_2\text{O}_3, \text{Fe}_2\text{O}_3, \text{FeO}, \text{Ga}_2\text{O}_3, \text{Gd}_2\text{O}_3, \text{Ge}_2\text{O}_3, \text{HfO}_2, \text{Ho}_2\text{O}_3, \text{In}_2\text{O}_3, \text{K}_2\text{O}, \text{La}_2\text{O}_3, \text{Li}_2\text{O}, \text{Lu}_2\text{O}_3, \text{MgO}, \text{MnO}, \text{Mn}_2\text{O}_3, \text{MoO}_2, \text{MoO}_3, \text{Na}_2\text{O}, \text{Nb}_2\text{O}_5, \text{Nd}_2\text{O}_3, \text{NiO}, \text{P}_2\text{O}_5, \text{PhO}, \text{Pr}_2\text{O}_3, \text{Pr}_{11}\text{O}_{21}, \text{Rb}_2\text{O}_3, \text{SO}_3, \text{Sb}_2\text{O}_5, \text{Sb}_2\text{O}_4, \text{Sc}_2\text{O}_3, \text{SiO}_2, \text{Sm}_2\text{O}_3, \text{SnO}, \text{SnO}_2, \text{SrO}, \text{Ta}_2\text{O}_5, \text{Te}_2\text{O}_6, \text{ThO}_2, \text{TiO}_2, \text{Ti}_2\text{O}_3, \text{Ti}_3\text{O}_4, \text{V}_2\text{O}_5, \text{WO}_3, \text{Y}_2\text{O}_3, \text{ZnO}, \text{ZrO}_2). \]

When using the EPB descriptors, the descriptors in the observed dataset were scaled using the mean and standard deviation for each descriptor, that is, the mean was subtracted, and the result was divided by the standard deviation to equalize the magnitudes between descriptors. The EPB descriptors in the candidate dataset were scaled using the mean and standard deviation estimated from the observed dataset. This scaling is commonly conducted in machine learning studies when the descriptors vary in magnitude.\(^{21)}\) For example, the magnitude of \( f_{\text{mean}} \) calculated from electronegativity (e.g., 1.90 for Si) is greatly different from that calculated from the melting temperature (e.g., 1000 K for Si).

Here, we describe our BO procedure for finding high-\( n_d \) compositions. First, five compositions are randomly chosen from all data as the initial observed data and the remaining data are used as the candidate data. Next, the Gaussian process regression learns the observed dataset and predicts the \( \mu \) and \( \sigma \) values for the compositions in the candidate dataset. Then, we calculate \( a_{EI} \) for the candidate compositions using Eq. (1) and observe the \( n_d \) value of the
composition with the highest $a_{EI}$. In other words, the composition data (i.e., the composition and the corresponding $n_d$ value) are added to the observed dataset and removed from the candidate dataset. We repeat the learning of the observed dataset by the Gaussian process regression and the observation of the $n_d$ value with the highest $a_{EI}$ 15 times for 50 patterns. The five initial compositions are randomly chosen for each pattern. We examine how many observations with high-$n_d$ compositions are found by BO from the candidate dataset.

Figures 2 and 3 show the results of the search for high $n_d$ compositions with EPB and CAB descriptors. Figure 2 shows the highest values of $n_d$ until the $n$-th observation for five typical patterns of the initial observed datasets. It can be seen that using EPB descriptors is more effective than using CAB descriptors for finding high-$n_d$ compositions. The composition $27.66\text{Bi}_2\text{O}_3-69.96\text{PbO}-2.38\text{SiO}_2$ with a high $n_d$ value of 2.64, is found within 15 observations for all five patterns of the initial observed datasets. It was found for all 50 patterns of the initial observed datasets with the EPB descriptors with an average number of observations of only 4.8 (standard deviation: 3.0). In contrast, this composition is not found by BO with the CAB descriptors. Here, it should be noted that there are only two compositions with a refractive index of above 2.6, i.e., $27.66\text{Bi}_2\text{O}_3-69.96\text{PbO}-2.38\text{SiO}_2$ ($n_d = 2.64$) and $37.5\text{Gd}_2\text{O}_3-37.5\text{SiO}_2-25\text{B}_2\text{O}_3$ ($n_d = 2.65$), in our data. The latter composition wasn’t found by BO with both EPB and CAB descriptors. The latter composition is rich in the components of SiO$_2$ and B$_2$O$_3$ that are contained in many compositions. In our data, about 70 percent of all compositions contains SiO$_2$ and about 65 percent contains B$_2$O$_3$. Because BO tends to find rare compositions using uncertainties, we speculate that BO didn’t find the latter compositions with much SiO$_2$ and B$_2$O$_3$ within the 15 observations.

Figure 3 shows that BO with the EPB descriptors tends to find compositions with a higher refractive index compared with those found with the CPB descriptors. Generally, when a training dataset is small, an excessive number of descriptors tends to decrease prediction performance. In our study, the number of CAB descriptors 66 was higher than that EPB descriptors 22. Because the initial observed
datasets had only five compositions and the number of compositions in the observed datasets increased by only one at each BO step, the size of the observed dataset (training dataset) was very small. Therefore, the small number of EPB descriptors contributed to the higher search performance obtained with EPB descriptors. The BO search performance also depends on factors such as the selection of descriptors, target physical properties training dataset size, or their combinations. For example, there is no guarantee that our 11 EPB descriptors are the most effective to the BO for glass composition search. Therefore, we need to understand the detailed relation between performance and these factors and to develop more effective descriptors for BO glass composition search in the future.

In summary, we showed that high-$n_d$ glass compositions can be found from a large number of candidates in a low number of observations using BO. Furthermore, we examined the dependence of BO search performance on descriptors type and confirmed that BO with EPB descriptors can find high-$n_d$ compositions more rapidly than that with CAB descriptors. Our study is the first to use BO and EPB descriptors for investigating oxide glass. The results suggest that BO with EPB descriptors is a powerful tool for accelerating the search for compositions with desirable physical properties.

Acknowledgments We would like to thank Dr. M. Mizuguchi, M. Ueda, K. Yoshimoto, and T. Kawashima of Materials & Advanced Research Laboratory, Nikon Corporation, for providing advice regarding glass science.

References
1) A. Masuno, H. Inoue, K. Yoshimoto and Y. Watanabe, Opt. Mater. Express, 4, 710–718 (2014).
2) K. Yoshimoto, A. Masuno, M. Ueda, H. Inoue, H. Yamamoto and T. Kawashima, Sci. Rep., 7, 45600 (2017).
3) D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, Nat. Commun., 7, 11241 (2016).
4) A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput and I. Tanaka, Phys. Rev. Lett., 115, 205901 (2015).
5) B. R. Leduc, K. Barros, T. Lookman and C. J. Humphreys, Sci. Rep., 6, 24862 (2016).
6) L. Ward, S. C. O’Keeffe, J. Stevick, G. R. Jelbert, M. Aykol and C. Wolverton, Acta. Mater., 159, 102–111 (2018).
7) D. R. Cassar, A. Carvalho and E. D. Zanotto, Acta. Mater., 159, 249–256 (2018).
8) J. C. Mauro, A. Tandia, K. D. Vargheese, Y. Z. Mauro and M. M. Smorskjaer, Chem. Mater., 28, 4267–4277 (2016).
9) Y. Deng, H. Zeng, Y. Jiang, G. Chen, J. Chen and L. Sun, Mater. Res. Express, 5, 035205 (2018).
10) D. Xue, D. Xue, R. Yuan, Y. Zhou, P. V. Balachandran, X. Ding, J. Sun and T. Lookman, Acta. Mater., 125, 532–541 (2017).
11) L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, npj Comput. Mater., 2, 16028 (2016).
12) K. Matsumoto and T. Horide, Appl. Phys. Express, 12, 073003 (2019).
13) A. Makishima and J. D. Mackenzie, J. Non-Cryst. Solids, 17, 147–157 (1975).
14) A. Fluegel, Glass. Technol. Part A, 50, 25–46 (2009).
15) K. Ishii, T. Tsuneoka, S. Sakida, Y. Benino and T. Nanba, J. Ceram. Soc. Jpn., 120, 98–103 (2012).
16) INTERGLAD, ver. 7, New Glass Forum, http://www.newglass.jp/interglad.n/.
17) National Astronomical Observatory of Japan, Chronological Scientific Tables; Maruzen Co., Ltd, Tokyo (2016).
18) B. Cordero, V. Cômez, A. E. P. Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán and S. Alvarez, Dalton T., 21, 2832–2838 (2008).
19) R. D. Shannon and C. T. Prewitt, Acta Crystallogr. B, 25, 925–946 (1969).
20) GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy.
21) G. Pilania, P. V. Balachandran, C. Kim and T. Lookman, Front. Mater., 3, 19 (2016).