UCB 獲得関数と選定記述子を用いたガラス組成のベイズ最適化

中村健作，大谷直也，小池哲也

Bayesian Optimization of Glass Compositions with Upper Confidence Bound and Selected Descriptors

Kensaku NAKAMURA, Naoya OTANI and Tetsuya KOIKE

ニコンは100年以上にわたり光学ガラスの研究開発を実施しており，光学ガラスはカメラや顕微鏡など多数のニコン製品に搭載されている．光学ガラス開発において，“ガラスの組成設計”は，所望の物理特性を有するようにガラス構成元素の種類や量を調整する重要なプロセスであり，専門的な知識や経験に基づく試行錯誤が必要となる．一方，近年，機械学習を用いることで材料開発を加速する試みが注目されている．本研究では，機械学習による光学ガラス開発の高速化を目指す取組みの一環として，機械学習の一手法であるベイズ最適化を組成設計に適用する．具体的には，国際ガラスデータベース INTERGLAD のデータにおいて，ベイズ最適化を用いて高アッベ数組成を探索する．さらに，獲得関数 Upper Confidence Bound のパラメーター調整および，機械学習モデルの入力パラメーター，すなわち，記述子の選定が，ベイズ最適化の探索性能に与える効果を検証する．

Nikon has developed optical glass for over 100 years, and the optical glass has been installed in many Nikon products such as cameras and microscopes. In the development of the optical glass, the composition design of glass is important, in which the types and amounts of constituent elements are adjusted to obtain glass with desirable physical properties (e.g., refractive index and Abbe number), and numerous trials and errors based on the knowledge and experience of experts are required. By contrast, in material sciences, attempts to accelerate material developments using machine learning has been reported recently. In this study, we apply Bayesian optimization, a machine learning method, to the composition design of glass to accelerate the development of optical glass. It is demonstrated that compositions with high Abbe numbers can be identified using Bayesian optimization based on data from the International Glass Database, INTERGLAD. In addition, we discuss the effects of setting the parameter of an acquisition function, upper confidence bounds, and descriptors to the search performance of Bayesian optimization.

Key words 光学ガラス，組成設計，アッベ数，機械学習，ベイズ最適化
optical glass, composition design, Abbe number, machine learning, Bayesian optimization

1 Introduction

Glass is used in various applications, such as camera lenses, windows, and electronic displays. The specifications of glass properties differ depending on the application. For example, in the development of optical products such as camera lenses, the optical properties of glass, such as its refractive index and Abbe number ν_d, must be adjusted to satisfy the specifications [1],[2]. Composition design is a typical method for controlling the physical properties of glass because its properties depend significantly on its composition. Generally, composition design requires the knowledge and experience of experts; furthermore, it is time consuming because the number of element combinations is significant.

Recently, machine learning has garnered attention as an effective tool for accelerating the development of materials, including glass [3]–[10]. We have previously focused on one of the machine learning methods, i.e., Bayesian optimization (BO), which proposes the next experimental condition (e.g., chemical compositions) based on previous experimental data. BO has been applied to the development of various materials such as thermoelectric materials, shape-memory alloys, and oxide glass [7]–[10]. Unlike other optimization methods, BO can search for the next experimental condition in an extrapolated area because it employs acquisition functions that indicate the effectiveness of the experiment based on the predicted values and their uncertainties [11], [12].

Several types of acquisition functions are known, and we have focused on one of the acquisition functions, i.e., the upper confidence bound (UCB), which can achieve a balance between exploitation and exploration by setting a parameter for experiments [13]–[15]. For example, in glass development, if a glass composition that differs significantly from observed ones is required, then the exploration of BO should be enhanced by adjusting the parameter. Hence, BO with the UCB acquisition function is expected to benefit the composition design of glass.

In this study, we applied BO with the UCB acquisition function to optimize glass compositions to identify high ν_d compositions using the International Glass Database (INTERGLAD) [16]. We present the dependence of the search performance of BO on the balance between exploitation and exploration. Subsequently, we discuss the effect on the search performance with respect to the selection of input variables, i.e., descriptors, using random forest (RF) analysis. Generally, the selection of input variables is important to achieve good performances in machine learning [6]–[8], [10].

2 Methods

The composition and ν_d data were obtained from the INTERGLAD [16]. Some compositions that exhibited incorrect values were removed. We used the data of only the silicate system for which the amount of SiO_2 was more than 0 mol%. A total of 7181 compositions were used. The composition included the following 57 components: Al_2O_3, As_2O_3, B_2O_3, BaO, BeO, Bi_2O_3, CaO, CdO, Ce_2O_3, CeO_2, Co_2O_3, CoO, Cs_2O, CuO, Dy_2O_3, Er_2O_3, Fe_2O_3, Ga_2O_3, Gd_2O_3, GeO_2, HfO_2, In_2O_3, K_2O, La_2O_3, Li_2O, Lu_2O_3, MgO, MnO, MnO_2, MoO_2, MoO_3, Na_2O, Nb_2O_3, Nb_2O_5, Nd_2O_3, NiO, P_2O_5, PbO, Pr_2O_3, Rb_2O, SO_3, Sb_2O_3, Sb_2O_5, Sc_2O_3, SiO_2, Sm_2O_3, SnO, SnO_2, SrO, Ta_2O_5, TeO_2, TiO_2, Tl_2O, WO_3, Y_2O_3, ZnO, and ZrO_2. Fig. 1 shows the histogram of ν_d for the compositions. A histogram of the appearance of the components is shown in Fig. 2. We set the target value of ν_d to 70 and analyzed the speed in which BO identifies a composition with a ν_d exceeding 70. Approximately 1% of the total compositions indicated ν_d values exceeding 70.

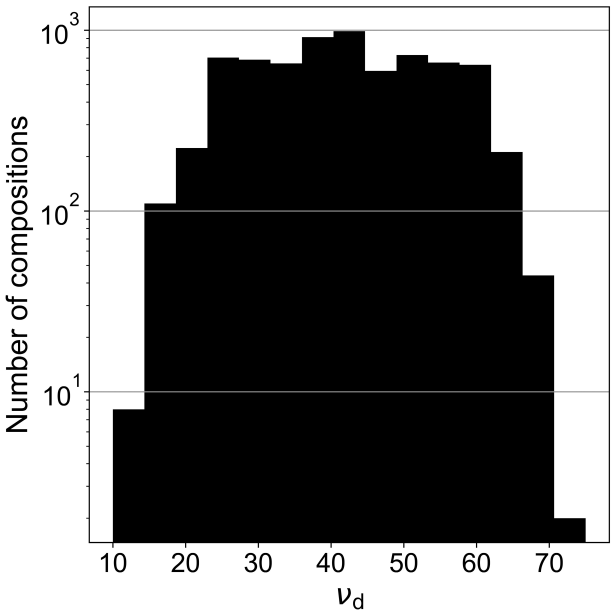

Fig. 1 Histogram of ν_d for collected compositions. Logarithmic scale is used for y-axis.

We used typical descriptors based on elemental physical properties [4], [6]–[10]. The descriptors were calculated from the numbers of elements in the compositions and the following 11 elemental properties: atomic number, Mendeleev number, column and row numbers in the periodic table, covalent radius, Ahrens ionic radius, electronegativity, first ionization energy, melting point, atomic weight, and density [17]–[21]. Specifically, two descriptors, mean x_{mean} and standard deviation x_{std}, were calculated for each property in each composition as follows:

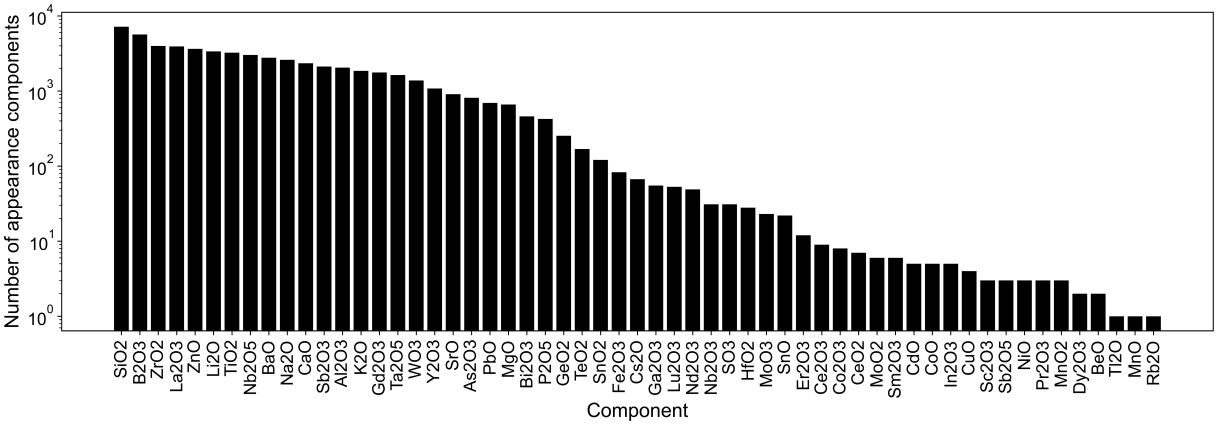

Fig. 2 Histogram of appearance components for collected compositions. Logarithmic scale is used for y-axis.

$$x_{\mathrm{mean}} = \frac{\sum_i y_i x_i}{\sum_i x_i}, \tag{1}$$

$$x_{\mathrm{std}} = \sqrt{\frac{\sum_i (y_i - x_{\mathrm{mean}})^2 x_i}{\sum_i x_i}}, \tag{2}$$

where i represents the element species, x_i the atomic fraction in the composition, and y_i the value of the physical property. In total, 22 descriptors were used for each composition. The set of descriptors was the same as in those in previous studies [7], [8]. The descriptors in the training data were normalized using the mean and standard deviation for each descriptor, i.e., the mean and variance of the values of the descriptors were set to zero and one, respectively. Descriptors are often normalized to equalize their scales [8]. Furthermore, the importance score for each descriptor in the prediction of v_d was calculated by fitting all the data (compositions and their v_d values) via RF regression. RF regression is a decision tree ensemble method that can output the importance of each descriptor [5], [17]. The descriptors with high importance contribute significantly to the prediction of v_d. We executed RF regression using the scikit-learn package [22]. Fig. 3 shows the importance of each descriptor. We analyzed the effect of descriptor selection on the BO search performance by comparing two cases. In the first case, all descriptors were used. In the second case, the following 11 descriptors with higher importance were used: density x_{mean} and x_{std}, Ahrens ionic radius x_{mean} and x_{std}, atomic weight x_{mean} and x_{std}, row numbers in the periodic table x_{mean}, column numbers in the periodic table x_{mean}, atomic number x_{mean} and x_{std}, and melting point x_{std}.

The procedure for BO in this study is as follows: five compositions were randomly selected as initial training data. The remaining compositions were composed as initial test data. The training data were fitted using Gaussian process regression. Gaussian process regression is a Bayesian inference method that outputs the uncertainty of prediction and the predicted value, and it is typically used in BO. We used a GPy library to implement Gaussian process regression [23]. We used a typical kernel function, i.e., the Gaussian kernel, for Gaussian process regression. Using Gaussian process regression, the predicted values and uncertainties (i.e., standard deviations) were obtained for each composition of the test data. Subsequently, the UCB acquisition functions a_{UCB} for the compositions were calculated as the criterion, as follows:

$$a_{\mathrm{UCB}} = \mu + \kappa\sigma, \tag{3}$$

where μ and σ are the predicted values and standard deviation, respectively; κ is a hyperparameter that controls the balance between exploitation and exploration. Although several expressions for the UCB have been proposed [13]–[15], we used a simple one, as shown in Eq. (3), which comprises only three parameters: μ, σ, and κ. By setting κ to a higher value, a composition different from that in the training data is proposed for BO. In this study, we performed BO with different values of κ to evaluate the dependence of BO performance on the balance between exploitation and exploration. We performed an experiment and observed the result under the condition with the highest values of acquisition functions in BO. Subsequently, we observed the v_d of a composition with the highest value of a_{UCB}, i.e., we added the composition and its v_d into the training data and removed them from the test data. This process was repeated until the v_d value of the highest a_{UCB} composition exceeded 70. In this study, when a composition with a high v_d was identified via a small number of observations, the search performance of BO was regarded as superior. We executed the BO search for 50 patterns of the initial training data at each κ value.

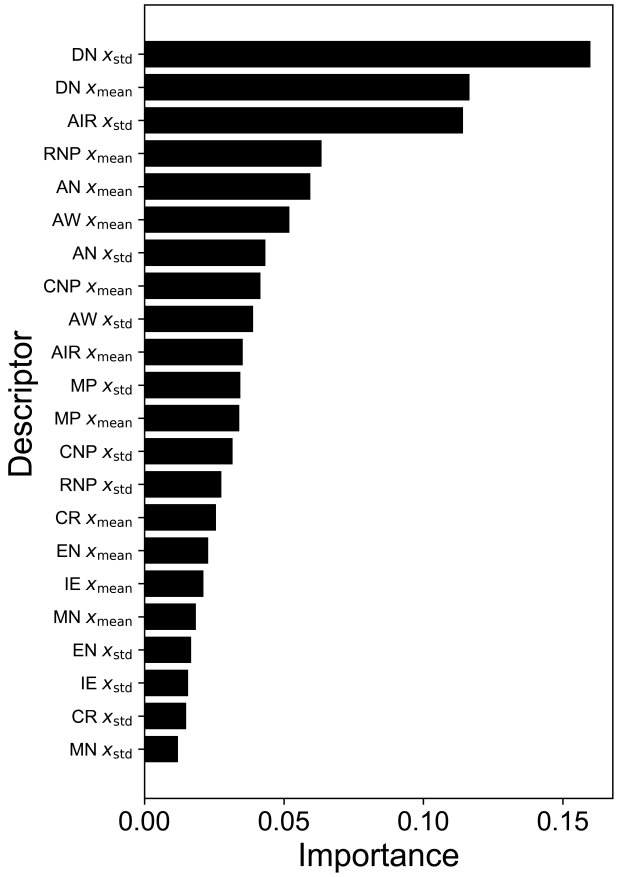

Fig. 3 Importance scores for descriptors calculated using RF. Abbreviations: DN, density; AIR, Ahrens ionic radius; RNP, row numbers in the periodic table; AN, atomic number: CNP, column numbers in the periodic table; AW, atomic weight; MP, melting point; CR, covalent radius; EN, electronegativity; IE, first ionization energy; MN, Mendeleev number.

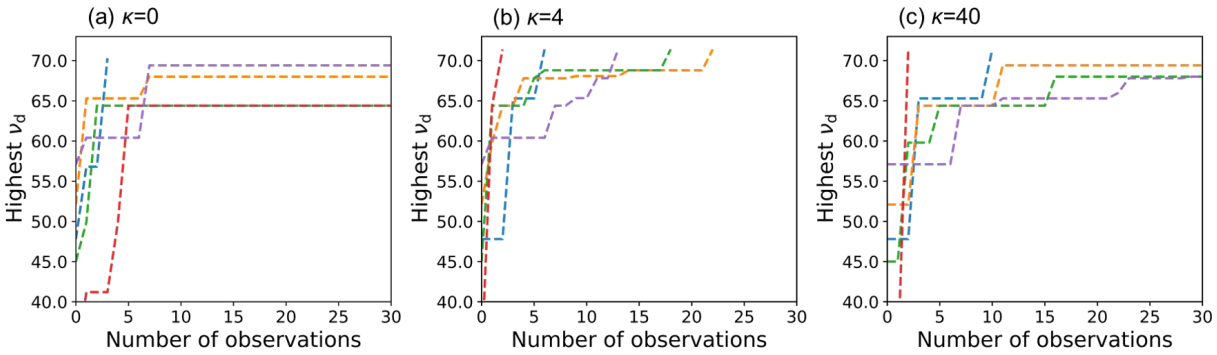

Fig. 4 The highest v_d values until 30th observation with all (non-selected) descriptors with different initial training data for five patterns at the different values of UCB parameter κ. Colors and dotted lines indicate the results for individual initial training data.

3 Results and Discussion

Fig. 4 shows the typical results of the BO in 30 observations for five patterns of the initial training data for different values of κ with 22 descriptors. As shown in Fig. 4, when κ = 4, compositions with v_d > 70 were identified until the 30th observation. However, for the other values of κ, compositions with a high v_d were not identified. BO with appropriate κ values enable compositions with high v_d to be identified rapidly. Fig. 5 shows the relationship between κ and the average number of BO observations required to identify compositions with a high v_d in 50 patterns of the initial training data using all the selected descriptors. In both descriptors, when κ is zero, the number of observations is high. In the case involving all descriptors, when the value of κ was less than 20, the average number of observations became the minimum. Subsequently, when the value of κ exceeded 20, the average number of observations increased. Because a large κ indicates that the uncertainty in the UCB (Eq. (3)) is significant, a vast composition region is searched during BO and compositions with a high v_d cannot be identified. By contrast, in the case involving selected descriptors, when κ is 20 or more, the average number of observations becomes the minimum and is similar for each κ. The average number of observations was smaller when the selected descriptors were used compared with when all descriptors were used. Therefore, these results suggest that tuning the UCB parameter and selecting descriptors can improve the search performance of BO. It is noteworthy that when using the selected descriptors, as κ increases, the average number of observations does not decrease, unlike the case for all descriptors. We speculate that the effect of uncertainty is less prominent when using the selected descriptors than when using all descriptors for a large value of κ in this study because the dimensions of the selected descriptors are smaller than those of all the descriptors.

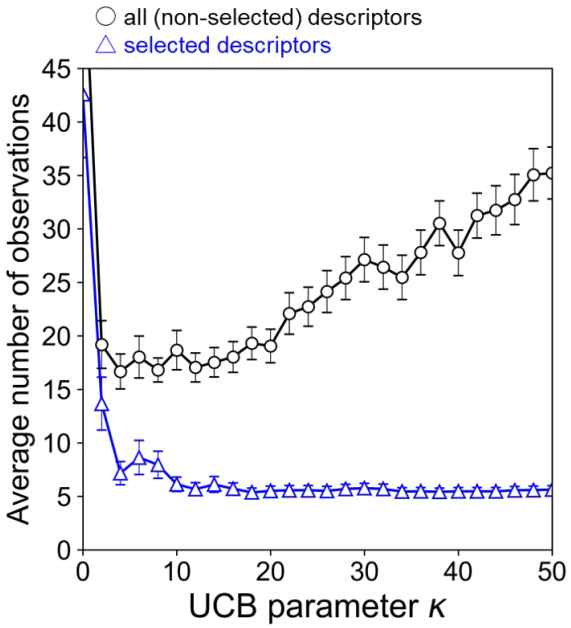

Fig. 5 Average number of observations required to identify v_d exceeding 70 for various values of κ using all (non-selected) and selected descriptors.

4 Conclusion

We demonstrated that BO with a UCB acquisition function enabled compositions with high v_d to be identified using data from the INTERGLAD. We demonstrated that the search performance of BO depended significantly on the UCB. Furthermore, BO with selected descriptors based on their importance scores obtained from RF was more effective in identifying compositions with high v_d than BO with all descriptors. Therefore, parameter tuning and the selection of appropriate descriptors are crucial for rapidly identifying compositions with desirable properties.

Acknowledgements. We would like to thank Dr. M. Mizuguchi, M. Ueda, K. Yoshimoto, and T. Kawashima of the Materials & Advanced Research Laboratory, Nikon Corpora-

tion, for providing advice regarding glass science.

References

[1] K. Yoshimoto, A. Masuno, M. Ueda, H. Inoue, H. Yamamoto, and T. Kawashima, "Thermal and optical properties of La_2O_3–Ga_2O_3–(Nb_2O_5 or Ta_2O_5) ternary glasses," *Journal of the American Ceramic Society*, vol. 101, pp. 3328–3336, 2018.

[2] K. Yoshimoto, A. Masuno, M. Ueda, H. Inoue, H. Yamamoto, and T. Kawashima, "Low phonon energies and wideband optical windows of La_2O_3–Ga_2O_3 glasses prepared using an aerodynamic levitation technique," *Scientific Reports*, vol. 7, p. 45600, 2017.

[3] D. R. Cassar, A. C. P. L. F. de Carvalho, and E. D. Zanotto, "Predicting glass transition temperatures using neural networks," *Acta Materialia*, vol. 159, pp. 249–256, 2018.

[4] D. R. Cassar, "ViscNet: Neural network for predicting the fragility index and the temperature-dependency of viscosity," *Acta Materialia*, vol. 206, p. 116602, 2021.

[5] E. Alcobaça, S. Mastelini, T. Botari, B. Pimentel, D. Cassar, A. Carvalho, and E. Zanotto, "Explainable machine learning algorithms to predict glass transition," *Acta Materialia*, vol. 188, pp. 92–100, 2020.

[6] K. Nakamura, N. Otani, and T. Koike, "PHYSICAL PROPERTY PREDICTING DEVICE, DATA GENERATING DEVICE, PHYSICAL PROPERTY PREDICTING METHOD, AND PROGRAM," (in Japanese), Japan Patent P2020–200213A, 2020.

[7] K. Nakamura, N. Otani, and T. Koike, "Search for oxide glass compositions using Bayesian optimization," *Journal of the Ceramic Society of Japan*, vol. 123, pp. 569–572, 2020.

[8] K. Nakamura, N. Otani, and T. Koike, "Multi-objective Bayesian optimization of optical glass compositions," *Ceramics International*, vol. 47, pp. 15819–15824, 2021.

[9] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, "Accelerated search for materials with targeted properties by adaptive design," *Nature Communications*, vol. 7, p. 11241, 2016.

[10] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization," *Physical Review Letters*, vol. 115, p. 205901, 2015.

[11] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," *Proceedings of the NIPS 2012*, 2012.

[12] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, pp. 148–175, 2016.

[13] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed Bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[14] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[15] D. D. Cox, and S. John, "A statical method for global optimization," *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 2, 1992, pp. 1241–1246.

[16] "INTERGLAD Ver.7," International Glass Database System INTERGLAD Ver. 7, NEW GLASS FORUM. http://www.newglass.jp/interglad_n/. 2017.

[17] L. Breiman, "Random forest," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[18] *Chronological Scientific Tables*, Tokyo: Maruzen Publishing Co., Ltd., 2016.

[19] B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán, and S. Alvarez, "Covalent radii revisited," *Dalton Transactions*, vol. 21, pp. 2832–2838, 2008.

[20] R. D. Shannon, and C. T. Prewitt, "Effective ionic radii in oxides and fluorides," *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry*, vol. 25, pp. 925–946, 1969.

[21] P. Villars, K. Cenzual, J. Daams, Y. Chen, and S. Iwata, "Data-driven atomic environment prediction for binaries using the Mendeleev number," *Journal of Alloys and Compounds*, vol. 367, pp. 167–175, 2004.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] "Sheffield ML/GPy - Gaussian process framework in python," GitHub. http://github.com/ShefieldML/GPy. 2020.

中村健作　Kensaku NAKAMURA
研究開発本部　数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

大谷直也　Naoya OTANI
研究開発本部　数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

小池哲也　Tetsuya KOIKE
研究開発本部　数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

中村健作
Kensaku NAKAMURA

大谷直也
Naoya OTANI

小池哲也
Tetsuya KOIKE