Developing a Design Guideline of Boronic Acid Derivatives to Scavenge Targeted Sugars in the Formose Reaction Products using DFT-based Machine Learning

Nanako Ishihara[1], Genta Chikatani[1], Hiroaki Nishijima[1], Hiro Tabata[1], Yoko Hase[1], Yoshiharu Mukouyama[1,2], Shuji Nakanishi[1,3,*], Shiho Mukaida[1,*]

[1] Research Center for Solar Energy Chemistry, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
[2] Division of Science, College of Science and Engineering, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan
[3] Innovative Catalysis Science Division, Institute for Open and Transdisciplinary Research Initiatives (ICS-OTRI), Osaka University, Suita, Osaka 565-0871, Japan

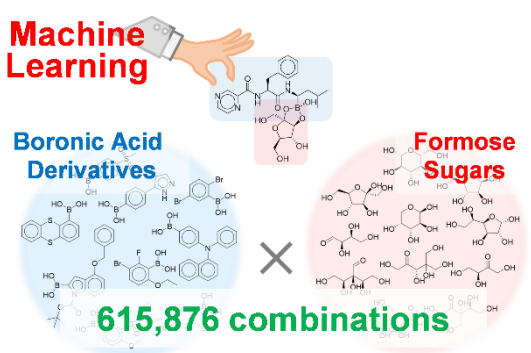
*Corresponding author: Research Center for Solar Energy Chemistry, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan Email: nakanishi.shuji.es@osaka-u.ac.jp (S.Nakanishi), mukaida@rcsec.chem.es.osaka-u.ac.jp (S.Mukaida)

Abstract

Formose reaction facilitates the synthesis of sugars from HCHO, yet the valuable sugars constitute only a small portion of the total products. This necessitates the need for a chemical scavenger capable of selectively capturing only valuable sugars. With over 600,000 potential combinations of boronic acid-based scavengers available, pursuing a deductive search approach is unfeasible. This study aims to derive guidelines for designing scavengers that readily bind with target sugars while avoiding non-target ones, via machine learning informed by DFT calculations.

Keywords: Formose reaction, Machine learning, DFT calculation

Graphical abstract



In this study, we clarified the characteristics that a selective scavenger for bio-assimilable sugars contained in formose reaction products must possess, utilizing machine learning informed by DFT calculations.

Since its discovery by Butlerov *et al.* in 1861,[1] the formose reaction—a non-enzymatic pathway for synthesizing sugars from formaldehyde (HCHO) under basic conditions—has attracted considerable interest for its implications in origin of life and food production.[2-6] Recently, our research successfully demonstrated that *Corynebacterium glutamicum* can be cultured using formose sugar, highlighting its potential as a substrate for the bioproduction of valuable compounds.[7] The formose reaction involves a myriad of reaction combinations featuring carbonyl groups, including aldol reactions, retro-aldol reactions, and aldose-ketose isomerization reactions, thereby forming a highly complex chemical reaction network.[8] Therefore, achieving high selectivity for desired products within this network is challenging.

Developing selective catalysts has been proposed as a strategy to address the challenge inherent in the formose reaction.[9-14] As an approach different from catalyst development, a strategy has also been proposed that involves adding scavengers that bind to the target sugar. Through the protection of hydroxy groups in these target sugars by the chemical scavenger, further reactions of the target sugars can be prevented, thereby increasing their yields. Boronic acids (BA) emerge as promising materials for these scavengers, as

they can form boronic esters with the diols of sugars in aqueous solutions, thereby stabilizing the products.[15] In fact, Ricardo *et al.* have discovered that boronic acid, protect ribose in the formose reaction system.[16] However, due to their general property of forming cyclic esters with diols, BA can bind not only to the target sugars in the formose reaction products but also to non-target sugars, which are unnecessary. Thus, simply adding BA to the formose reaction system captures non-target sugars as well, potentially clogging the entire reaction.[17]

This issue can be addressed by employing boronic acid derivatives that have strong binding affinity for the target sugars while having weak binding affinity for the non-target sugars. To find suitable boronic acid derivatives for the formose reaction, exploration of all combination of boronic acid derivatives and sugars is needed. However, there are more than 3,000 types of commercially available boronic acid derivatives. Additionally, the formose reaction can yield up to 23 types of sugars, not even accounting for stereoisomers, presenting a significant challenge in identifying the most effective derivatives for selective sugar capture. Moreover, given that a single sugar molecule can have multiple pairs of hydroxyl groups capable of interacting with boronic acid derivatives[18], there is more than one potential interaction pattern between a sugar and a boronic acid derivatives; in some cases, there can be as many as eight varieties. Consequently, even when restricting the variety of sugars under consideration, the total number of potential combinations can surpass 600,000. Exploring these combinations through deductive methods is unfeasible. In addressing these challenges, employing machine learning proves to be a valuable approach.[19-23] Hence, this study aims to derive guidelines for designing scavengers that enhance selectivity for target sugars by investigating boronic acid derivative-based scavengers through machine learning, which is informed by density functional theory (DFT) calculations. There have been several examples where machine learning has been applied to formose reaction systems[24,25]; however, this study is the first to explore scavenger materials using this approach.

Figure 1(a) shows an overview of the formose reaction network, highlighting both the target and non-target sugars investigated in this study. All sugar structure used in this work is detailed in Table S1. In this study, bio-assimilable sugars such as glucose, fructose, ribose, and arabinose, which possess five (C5) and six (C6) carbon atoms, were defined as target sugars. On the other hand, aldoses (C4a) and ketoses (C4k), containing four carbon atoms, were classified as non-target sugars. This classification stems from the fact that C4a and C4k serve as precursors to C5 and C6 sugars, and their interception by scavengers might inhibit the generation of the larger C5 and C6 sugars. Furthermore, the reaction products of C4k and C4a with glycolaldehyde (C2) (C4k+C2, C4a+C2), the compound formed by the reaction of C4k with two molecules of HCHO (C1) (C4k+C1+C1), and the product resulting from the combination of two molecules of C3k (C3k+C3k) were designated as non-target sugars in this study. This is due to their potential to transform into glucose or fructose through subsequent reactions within the formose reaction network.

Figure 1(b) depicts the conceptual framework of this study. We computed the Gibbs free energy change ($\Delta_r G°$) for 2,927 reactions from the possible 615,876 combinations of ester formation reactions between boronic acid derivatives and sugars using DFT calculations. Utilizing these computations as training data, we developed a machine learning model to predict the $\Delta_r G°$ for the entire set of 615,87616 ester formation reactions. Given the aim of this study to identify boronic acid derivatives that selectively bind to target sugars, our focus is on the difference between the average $\Delta_r G°$ for ester formation reactions involving a specific boronic acid with the 4 types of target sugars

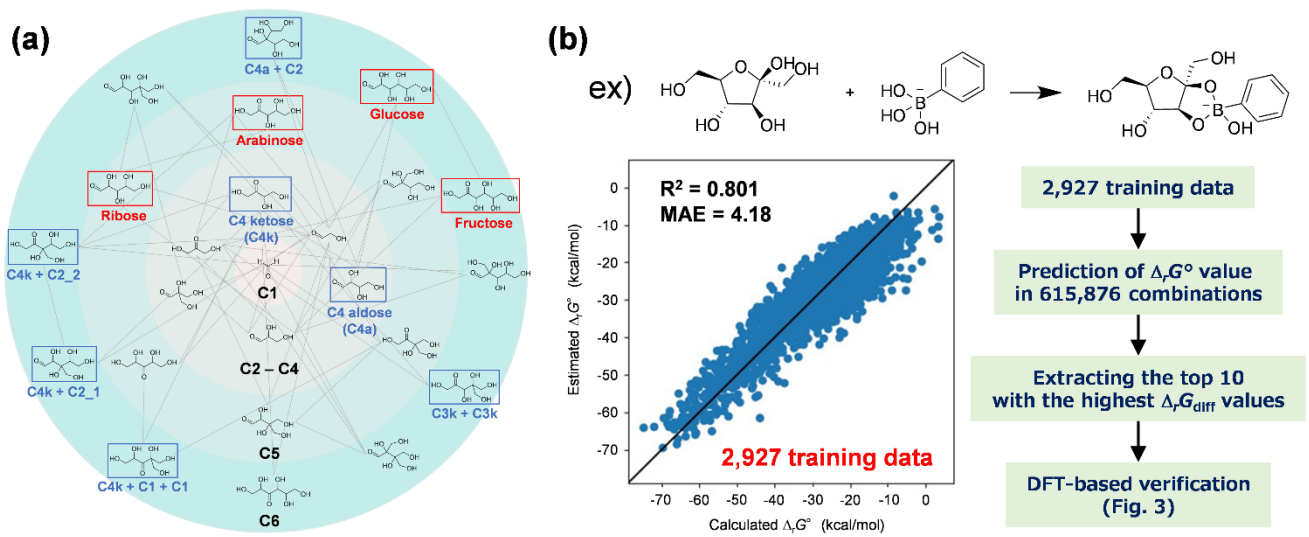


Figure 1. (a) The formose reaction network. Species surrounded with red and blue squares represents target sugars and non-target sugars, respectively. (b) The conceptual framework of this study and the actual versus predicted plot from the constructed machine learning model.

($\overline{\Delta_r G°}_{\text{target}}$) and the average $\Delta_r G°$ for ester formation reactions between that same boronic acid and the 7 types of non-target sugars ($\overline{\Delta_r G°}_{\text{non-target}}$). This difference is represented as $\Delta_r G_{\text{diff}}$.

$$\Delta_r G_{\text{diff}} = \overline{\Delta_r G°}_{\text{non-target}} - \overline{\Delta_r G°}_{\text{target}} \quad (1)$$

The greater the $\Delta_r G_{\text{diff}}$ value, the more it exhibits characteristics that align with the objective of this study. $\Delta_r G_{\text{diff}}$ values are derived from the $\Delta_r G°$ values predicted by the machine learning model, and boronic acid derivatives that appear promising are identified based on these values.

Finally, the $\Delta_r G°$ values for the ester formation reactions between the boronic acid derivative identified as optimal by the machine learning model and both target and non-target sugars are calculated using DFT. These values are then compared with those of a standard boronic acid (i.e., phenylboronic acid (PBA)) to validate the suitability of the boronic acid derivative proposed by the machine learning model. We would like to re-emphasize here that the purpose of this study is to perform a first screening to extract the general trends that promising boronic acid derivatives should possess, through comprehensive analysis of over 600,000 combinations. Therefore, the accuracy of the machine learning and DFT calculations presented hereafter, as well as the interpretation of their results, are conducted in light of this objective.

The actual versus predicted plot for the constructed machine learning model is displayed in Figure 1b. The coefficient of determination (R^2) exceeded 0.8, as indicated in Figure 1(b), verifying that the model has adequate performance. The results of the analysis of feature importance by SHAP values[26] are shown in Figure 2(a). Notably, among the different factors considered, the sugar species emerged as the most significant feature, suggesting that the structure of the sugar plays a predominant role in determining the $\Delta_r G°$, while the contribution from boronic acid functional groups is relatively minor. This finding is consistent with the discussions in the field of glucose sensing, where it has been argued that the sugar species, more specifically its structure, significantly influences its affinity with boronic acid derivatives.[16] The fact that the sugar species, rather than the boronic acid derivatives or the boronic ester formation products, was identified as the most significant factor in the $\Delta_r G°$ calculation further supports the validity of the machine learning approach employed in this study.[27,28]

The findings from the SHAP analysis further revealed that the descriptor Chi2n, indicative of the topological characteristics of molecules, also played a significant role. The Chi2n value, understood to be calculated by considering interatomic paths, increases in response to greater structural complexity, such as branching, and with the enlargement of molecular size (Table S2).[29] The positive correlation between the Chi2n values and the predicted $\Delta_r G°$ values, as confirmed by the SHAP analysis depicted in Figure 2(b), suggests that steric hindrance is a primary contributing factor. Additionally, the analysis revealed that descriptors associated with the C=N bond in boronic acid derivatives, specifically PEOE_VSA9 (Figure S1), and VSA_ESTATE6, which represents the carbon of the phenyl group (Figure S2), did not significantly contribute. Taken all together, it is suggested that the steric

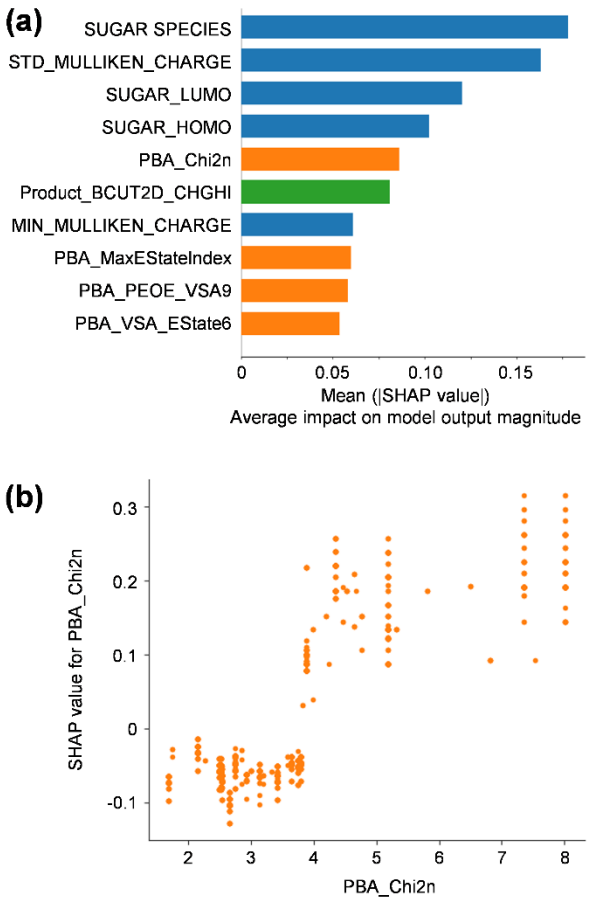


Figure 2. (a) The SHAP importance values of the machine learning model constructed. Blue, orange, and green represents sugars, boronic acid derivatives, and ester bond formation products, respectively. (b) A dependence plot between Chi2n and SHAP values.

effects associated with boronic acid derivatives play a more substantial role in influencing the reaction $\Delta_r G°$ than the presence of specific functional groups or bonds.

As previously discussed, boronic acid derivatives were assessed based on their $\Delta_r G_{\text{diff}}$ value (equation (1)). In comparison to the $\Delta_r G_{\text{diff}}$ value of 3.344, associated with standard PBA, the $\Delta_r G_{\text{diff}}$ value for the boronic acid derivative Bortezomib—identified as an optimal molecule by machine learning—was significantly higher, registering at 5.913, as detailed in Table S3. Thus, through DFT calculations, it has been confirmed that Bortezomib, as suggested by the machine learning model, indeed exhibits better selectivity towards target sugars compared to PBA.

The increase in the $\Delta_r G_{\text{diff}}$ value for Bortezomib may result from two scenarios: (1) higher affinity for the target sugar, or (2) lower affinity for the non-target sugars. Consequently, we investigated which scenario, (1) or (2), applies. The $\Delta_r G°$ values, as estimated by DFT calculations, for the ester formation reactions involving Bortezomib and PBA with their target and non-target sugars are shown in Figure 3a and 3b, respectively. It should be noted here that we plot the average $\Delta_r G°$ values of the ester formation

reactions between various isomers of each sugar and the

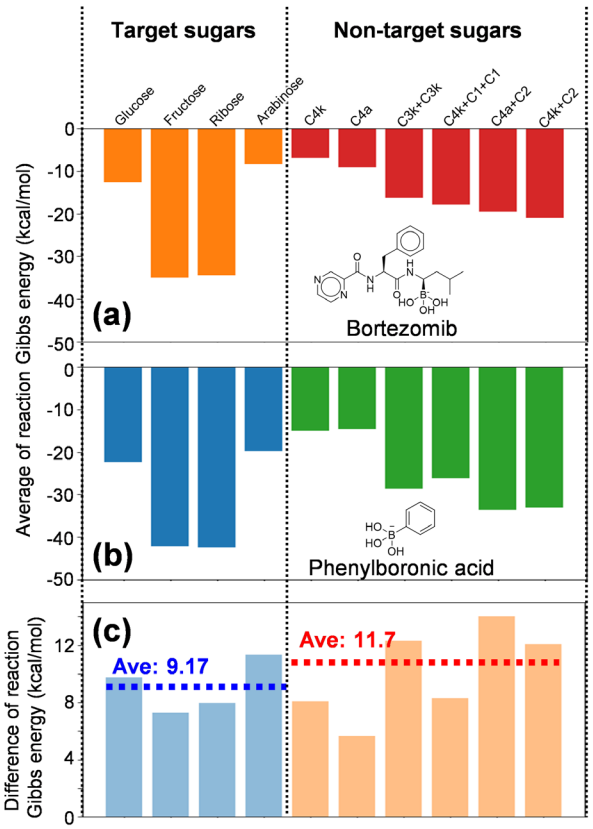


Figure 3. The average reaction Gibbs energy of (a) Bortezomib and (b) PBA for each sugar, and (c) the difference in reaction Gibbs energy between Bortezomib and PBA.

boronic acid derivatives. For the target sugars, the values were around -30 kcal/mol with Bortezomib and -40 kcal/mol with PBA. In contrast, for the non-target sugars, the values exceeded -20 kcal/mol with Bortezomib and were around -30 kcal/mol with PBA. The bar plot of Figure 3c illustrates the difference in $\Delta_r G°$ values between the two boronic acid derivatives. On the other hand, the dotted lines in Figure 3c represent the average differences in the $\Delta_r G°$ values between Bortezomib and PBA for each target and non-target sugars. (That is, they do not correspond to the simple average of each bar shown in Fig. 3c.) This average yields positive values for both target and non-target sugars, indicating that the larger $\Delta_r G_{\mathrm{diff}}$ value in Bortezomib, compared to PBA, was attributable to reason (2) rather than (1).

Next, we will explore the chemical reasons behind the increased difference in the $\Delta_r G_{\mathrm{diff}}$ value. As shown in Figure 3(c), focusing on the magnitude of the $\Delta_r G°$ for each sugar with PBA and Bortezomib, it is apparent that Bortezomib tends to exhibit larger $\Delta_r G°$ compared to PBA. Given the discussion regarding Chi2n in Figure 2, the primary reason for this is believed to be the steric hindrance resulting from the molecular complexity of Bortezomib. On the other hand, the aforementioned increase in $\Delta_r G°$ in the Bortezomib varies depending on the sugar species, being smaller for target sugars compared to non-target sugars. This variation is thought to stem from the inherent structure of the sugars themselves. Specifically, cyclic fructose and cyclic ribose in target sugars have many *cis*-configured OH groups advantageous for the formation of boronic acid esters, and the sugar itself has small steric hindrance.[30] Conversely, the branched sugars contained in the non-target sugars. In other words, in the target sugars, the $\Delta_r G°$ mentioned earlier is mitigated, and such an effect cannot be expected in the non-target sugar. Therefore, it is speculated that in Bortezomib, the difference in $\Delta_r G°$ between the target sugars and non-target sugars becomes significant, suggesting its potential function as a selective scavenger for sugars.

In this study, we aimed to obtain a general guideline for designing boronic acid-based scavengers that are efficacious in selectively capturing target sugars. From a comprehensive pool of 615,876 possible combinations, involving boronic ester products derived from 42 types of monosaccharides and 3003 commercially available monoboronic acid molecules, Bortezomib emerged as an optimal candidate. The insights from SHAP analysis and DFT calculations indicate that controlling the spatial structure is more crucial than tuning electronic properties, such as the selection of functional groups, in the context of boronic acid derivatives. In this study, we searched for the ideal scavenger under the assumption that one molecule of a boronic acid derivative forms an ester bond with one sugar molecule, as a first screening. Through machine learning, Bortezomib was proposed as a candidate molecule. This study has established a fundamental guideline demonstrating that effective control of stereo-configuration is beneficial. Following this guideline, future work will likely find that enhancing reaction selectivity can be effectively achieved by combining multiple mono-boronic acid molecules or poly-boronic acids to a target sugar.

Supplementary data

Supplementary material is available at *Chemistry Letters*

Acknowledgements

This work was achieved through the use of SQUID at the Cybermedia Center, Osaka University. I would like to thank Kaito Nagita for advice about conducting DFT calculations on SQUID and useful discussions.

References

1 A. Butlerov, *Justus Liebigs Ann. Chem.* **1861**, *120*, 295.
2 N. W. Gabel, C. Ponnamperuma, *Nature* **1967**, *216*, 453-455.
3 C. Reid, L. E. Orgel, *Nature* **1967,** *216*, 455.
4 J. B. García Martínez, K. A. Alvarado, X. Christodoulou, D. C. Denkenberger, *J. CO2 Util.* **2021**, *53*, 101726.
5 S. Cestellos-Blanco, S. Louisia, M. B. Ross, Y. Li, N. E. Soland, T. C. Detomasi, J. N. Cestellos Spradlin, D. K. Nomura, P. Yang, *Joule* **2022,** *6*, 2304.
6 I. V. Delidovich, A. N. Simonov, O. P. Taran, V. N. Parmon, *ChemSusChem* **2014,** *7,* 1833.
7 H. Tabata, H. Nishijima, Y. Yamada, R. Miyake, K. Yamamoto, S. Kato, S. Nakanishi, *ChemBioChem* **2024,** *25,* e202300760.
8 W. E. Robinson, E. Daines, P. van Duppen, T. de Jong, W. T. S. Huck, *Nature Chem.* **2022,** *14,* 623.

9 J. Castells, F. López-Calahorra, F. Geijo, *Carbohydr. Res.* **1983,** *116,* 197.

10 H. Tabata, G. Chikatani, H. Nishijima, T. Harada, R. Miyake, S. Kato, K. Igarashi, Y. Mukouyama, S. Shirai, M. Waki, Y. Hase, S. Nakanishi, *Chem. Sci.* **2023,** *14*, 13475.

11 M. Waki, S. Shirai, Y. Hase, Dalton Transactions 2024, 53, 2678.

12 Z. Iqbal, S. Novalin, *Curr. Org. Chem*. **2012,** *16,* 769.

13 T. I. Khomenko, M. M. Sakharov, O. A. Golovina, *Russ. Chem. Rev.* **1980**, *49*, 570.

14 T. Matsumoto, H. Yamamoto, S. Inoue, *J. Am. Chem. Soc*. **1984**, *106,* 4829.

15 J. P. Lor, J. O. Edwards, *J. Org. Chem.* **1959,** *24,* 769.

16 A. Ricardo, M. A. Carrigan, A. N. Olcott, S. A. Benner, *Science* **2004,** *303,* 196.

17 T. Imai, T. Michitaka, A. Hashidzume, *Beilstein J. Org. Chem.* **2016,** *12*, 2668.

18 T. D. James, K. R. A. Samankumara Sandanayake, S. Shinkai, *Angew Chem. Int. Ed. Engl.* **1996,** *35,* 1910.

19 Y. Guo, X. He, Y. Su, Y. Dai, M. Xie, S. Yang, J. Chen, K. Wang, D. Zhou, C. Wang*, J. Am. Chem. Soc.* **2021**, *143*, 5755.

20 D. P. Metcalf, Z. L. Glick, A. Bortolato, A. Jiang, D. L. Cheney, C. D. Sherrill, *J. Chem. Inf. Model.* **2023,** *64*.

21 T. Ando, N. Shimizu, N. Yamamoto, N. N. Matsuzawa, H. Maeshima, H. Kaneko, *J. Phys. Chem. A* **2022,** *126,* 6336.

22 S. Samizo, H. Kaneko, *ACS Omega* **2023,** *8,* 27247.

23 A. Yada, K. Nagata, Y. Ando, T. Matsumura, S. Ichinoseki, K. Sato, *Chem. Lett.* **2018,** *47,* 284.

24 H. J. Cleaves, G. Hystad, A. Prabhu, M. L. Wong, G. D. Cody, S. Economon, R. M. Hazen, *Proc. Natl. Acad. Sci. USA* **2023,** *120*, e2307149120.

25 S. Asche, G. J. T. Cooper, G. Keenan, C. Mathis, L. Cronin, *Nat. Commun.* **2021,** *12,* 1.

26 S. M. Lundberg, P. G. Allen, S.-I. Lee, *Adv. Neural Inf. Process. Syst.* **2017,** *10,* 4768–4777.

27 H. Fang, G. Kaur, B. Wang, *J. Fluoresc.* **2004,** *14*, 481.

28 N. Fujita, S. Shinkai, T. D. James, *Chem. Asian J.* **2008,** *3,* 1076.

29 L. H. Hall, L. B. Kier, *Rev. Comput. Chem.* **1991,** 367.

30 Y. Suzuki, M. Shimizu, T. Okamoto, T. Sugaya, S. Iwatsuki, M. Inamo, H. D. Takagi, A. Odani, K. Ishihara, *ChemistrySelect* **2016,** *1,* 5141.

1