**Supporting Information**

# 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

## Data

The boric acid derivatives explored in this study were identified by conducting substructure extraction with Open Babel[1] from the SDF catalog of Matrix Scientific,[2] yielding 3003 varieties. The specific types of sugars involved are detailed in Table S1. Furthermore, the reaction products were automatically retrieved using the Reaction SMARTS feature of RDKit.[3]

**Table S1.** The target sugars

| Number | Name | SMILES |
|---|---|---|
| 0 | beta-D-glucopyranose | O[C@H]1[C@@H]([C@H]([C@@H]([C@@H](CO)O1)O)O)O |
| 1 | alpha-D-glucopyranose | O[C@@H]1[C@@H]([C@H]([C@@H]([C@@H](CO)O1)O)O)O)O |
| 2 | beta-D-glucofuranose | O[C@H]1[C@@H]([C@H]([C@@H]([C@@H](CO)O)O1)O)O |
| 3 | alpha-D-glucofuranose | O[C@@H]1[C@@H]([C@H]([C@@H]([C@@H](CO)O)O1)O)O)O |
| 4 | beta-D-fructopyranose | C1[C@H]([C@H]([C@@H]([C@](O1)(CO)O)O)O)O |
| 5 | alpha-D-fructopyranose | OC[C@]1([C@H]([C@@H]([C@@H](CO1)O)O)O)O |

| | | |
|---|---|---|
| 6 | beta-D-fructofuranose | OC[C@@]1([C@H]([C@@H]([C@@H](CO)O1)O)O)O |
| 7 | alpha-D-fructofuranose | OC[C@]1([C@H]([C@@H]([C@@H](CO)O1)O)O)O |
| 8 | beta-D-ribopyranose | C1[C@H]([C@H]([C@H]([C@@H](O1)O)O)O)O |
| 9 | alpha-D-ribopyranose | O[C@@H]1[C@@H]([C@@H]([C@@H](CO1)O)O)O |
| 10 | beta-D-ribofuranose | O[C@H]1[C@@H]([C@@H]([C@@H](CO)O1)O)O |
| 11 | alpha-D-ribofuranose | O[C@@H]1[C@@H]([C@@H]([C@@H](CO)O1)O)O |
| 12 | beta-D-arabinopyranose | O[C@H]1[C@H]([C@@H]([C@@H](CO1)O)O)O |
| 13 | alpha-D-arabinopyranose | O[C@@H]1[C@H]([C@@H]([C@@H](CO1)O)O)O |
| 14 | beta-D-arabinofuranose | O[C@H]1[C@H]([C@@H]([C@@H](CO)O1)O)O |
| 15 | alpha-D-arabinofuranose | O[C@@H]1[C@H]([C@@H]([C@@H](CO)O1)O)O |
| 16 | C4a_1 | O=C[C@@H]([C@H](CO)O)O |
| 17 | C4a_2 | O=C[C@@H]([C@@H](CO)O)O |
| 18 | C4a_3 | O=C[C@H]([C@@H](CO)O)O |
| 19 | C4a_4 | O=C[C@H]([C@H](CO)O)O |
| 20 | C4k_1 | C([C@H](C(=O)CO)O)O |
| 21 | C4k_2 | O=C(CO)[C@@H](O)CO |
| 22 | C3k+C3k_1 | OCC(O)(CO)[C@H](C(CO)=O)O |
| 23 | C3k+C3k_2 | OCC(O)(CO)[C@@H](C(CO)=O)O |
| 24 | C4k+C1+C1_1 | OCC(O)(CO)C([C@H](O)CO)=O |
| 25 | C4k+C1+C1_2 | OCC(O)(CO)C([C@@H](O)CO)=O |
| 26 | C4a+C2_1 | O[C@H]([C@@](O)(C=O)[C@H](O)CO)CO |
| 27 | C4a+C2_2 | O[C@H]([C@](O)(C=O)[C@H](O)CO)CO |
| 28 | C4a+C2_3 | O[C@H](C(O)(C=O)[C@@H](O)CO)CO |
| 29 | C4a+C2_4 | O[C@@H](C(O)(C=O)[C@H](O)CO)CO |

| | | |
|---|---|---|
| 30 | C4k+C2_1_1 | [H]C([C@H]([C@@](O)(CO)[C@H](CO)O)O)=O |
| 31 | C4k+C2_1_2 | [H]C([C@@H]([C@@](O)(CO)[C@H](CO)O)O)=O |
| 32 | C4k+C2_1_3 | [H]C([C@@H]([C@](O)(CO)[C@H](CO)O)O)=O |
| 33 | C4k+C2_1_4 | [H]C([C@H]([C@@](O)(CO)[C@@H](CO)O)O)=O |
| 34 | C4k+C2_1_5 | [H]C([C@H]([C@](O)(CO)[C@@H](CO)O)O)=O |
| 35 | C4k+C2_1_6 | [H]C([C@H]([C@](O)(CO)[C@H](CO)O)O)=O |
| 36 | C4k+C2_1_7 | [H]C([C@@H]([C@@](O)(CO)[C@@H](CO)O)O)=O |
| 37 | C4k+C2_1_8 | [H]C([C@@H]([C@](O)(CO)[C@@H](CO)O)O)=O |
| 38 | C4k+C2_2_1 | OC[C@@H]([C@@](O)(C(CO)=O)CO)O |
| 39 | C4k+C2_2_2 | OC[C@@H]([C@](O)(C(CO)=O)CO)O |
| 40 | C4k+C2_2_3 | OC[C@H]([C@@](O)(C(CO)=O)CO)O |
| 41 | C4k+C2_2_4 | OC[C@H]([C@](O)(C(CO)=O)CO)O |

**DFT calculations**

DFT calculations were conducted using Gaussian 16 software.[4] The geometry of each compound was optimized at the B3LYP/6-31G+(d,p) level of theory, allowing for the identification of the lowest energy conformers. For these optimized structures, Gibbs free energies, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, and Mulliken charges were determined.

**Machine learning**

To construct the machine learning model, 2,926 results from DFT calculations under the conditions previously described were employed. The feature variables pertaining to sugars incorporated categorical variables encoded with type labels, in addition to the maximum, minimum, range, average, and standard deviation of the HOMO, LUMO, and Mulliken charges.

For the boric acid derivatives and their reaction products, feature variables were computed using RDKit's structural descriptors.[3] The Boruta algorithm was utilized for feature selection to build the model[5], and the XGBoostRegressor algorithm[6] was chosen for the modeling process. To assess the model's performance, cross-validation with a fold number of 10 was carried out. The performance metrics employed to evaluate the regression tasks include mean absolute error (MAE) and coefficient of determination (RCV2).

**Descriptors**

**・Chi2n**

The Chi2n value calculation process[7] can be illustrated using benzene as an example. Specific examples of Chi2n values, including that for benzene, are detailed in Table S2.

Step 1: Calculation of nVal

For each carbon atom in benzene, calculate the nVal (the number of valence electrons excluding bonds with hydrogen). For benzene, the nVal for each carbon atom is 3 (since a carbon atom typically has 4 valence electrons and is bonded to one hydrogen atom).

Step 2: Calculation of delta values

Calculate the delta value for each atom. The delta is calculated as $1/\sqrt{(nVal)}$. Therefore, the delta for each carbon atom in benzene is $1/\sqrt{3}$.

Step 3: Path Exploration

Explore all paths of length 3 in benzene (paths of atoms connected through 2 bonds). In the case of benzene, there are 6 paths. These paths are between adjacent carbon atoms (e.g., C1-C2-C3, C2-C3-C4, etc.).

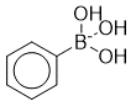
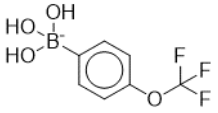
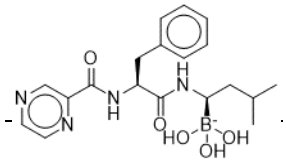
Step 4: Calculation of the Product of Delta Values

For each path, calculate the product of the delta values of the atoms included in the path. In the case of benzene, each path consists of 3 atoms. Therefore, the product of the delta values is $(1/\sqrt{3})^3$.

Step 5: Accumulation of the Product of Delta Values

Accumulate the product of delta values for all paths to obtain the final Chi2n value. The Chi2n for benzene is 1.155.

**Table S2.** Examples of Chi2n values

| Structure | Chi2n value |
|---|---|
|  | 2.144 |
|  | 2.851 |
|  | 7.345 |

**· PEOE_VSA9, PEOE_VSA10**

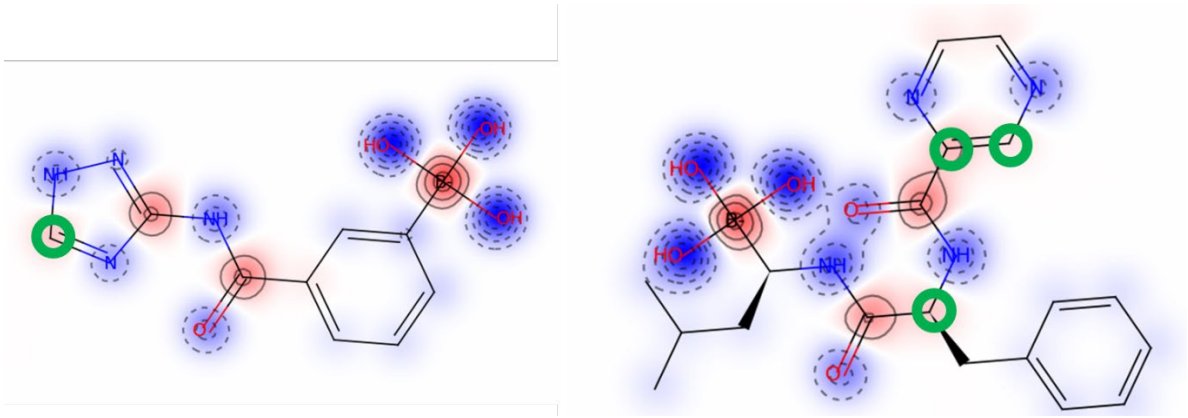


**Figure S1.** The distribution of Gasteiger charges within the molecule (red for positive, blue for negative) and the atoms corresponding to PEOE_VSA9 and PEOE_VSA10 (green).

• **VSA_ESTAE6**

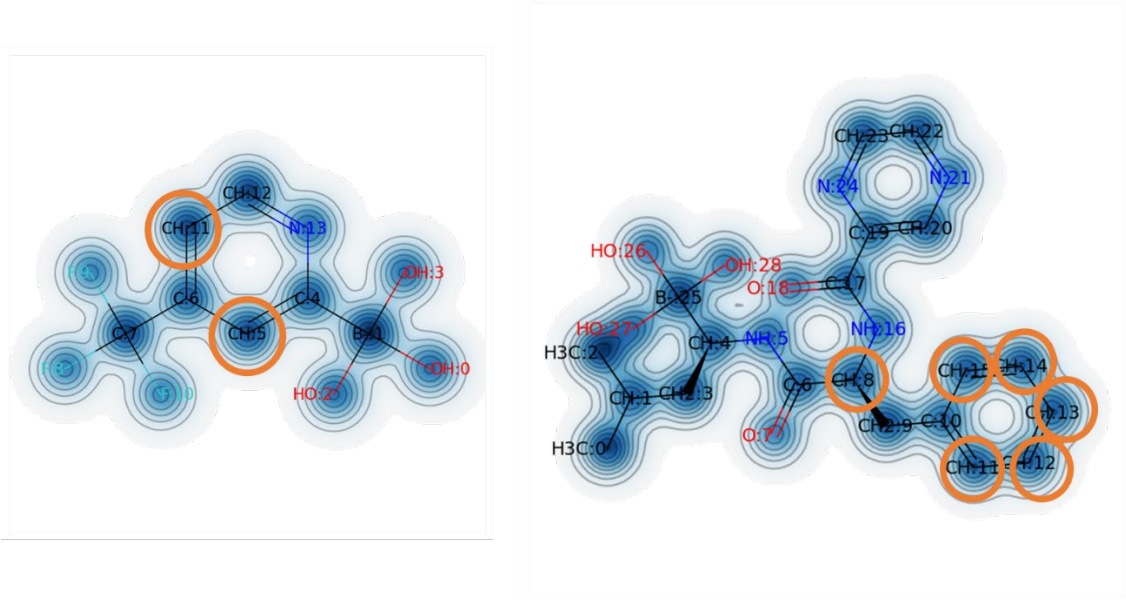
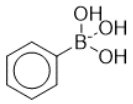
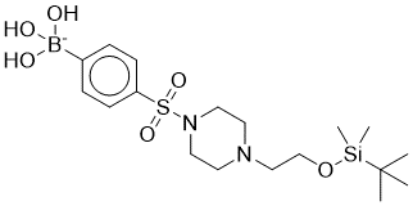
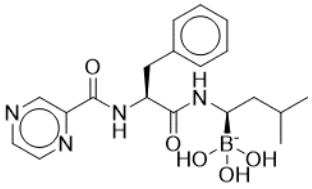



**Figure S2.** The van der Waals surface area of each atom within the molecule (blue) and the atoms corresponding to VSA_ESTATE6 (orange).

**Suggested boronic acid derivatives**

**Table S3.** $\Delta_r G_{\mathbf{diff}}$ values of each boric acid derivative

| Name | Structure | $\Delta_r G_{\mathrm{diff}}$ value |
|---|---|---|
| Phenylboronic acid |  | 3.344 |
| (4-((4-(2-((Tert-butyldimethylsilyl)oxy)ethyl)piperazin-1-yl)sulfonyl)phenyl)boronic acid (Randomly sampled) |  | 1.513 |
| **Bortezomib (Suggested)** |  | **5.913** |

**References**


1 Open Babel development team, *Open Babel*, **2016**.

2 Research Chemicals | Building Blocks | Matrix Scientific, https://www.matrixscientific.com/.

3 RDKit : Open-source cheminformatics., https://www.rdkit.org/.

4 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, J. E. Jr.; Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 16*, D. J. Gaussian, Inc., Wallingford CT, **2016**.

5 M. B. Kursa, A. Jankowski, W. R. Rudnicki, *Fundam. Inform.* **2010**, *101*, 271.

6 T. Chen, C. Guestrin, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, *13-17-August-2016*, 785.

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