JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 2656−2663

pubs.acs.org/jcim

# Development of R-Group Fingerprints Based on the Local Landscape from an Attachment Point of a Molecular Structure

Shunsuke Tamura,[†] Tomoyuki Miyao,[†,‡] and Kimito Funatsu*[,†,‡,§]

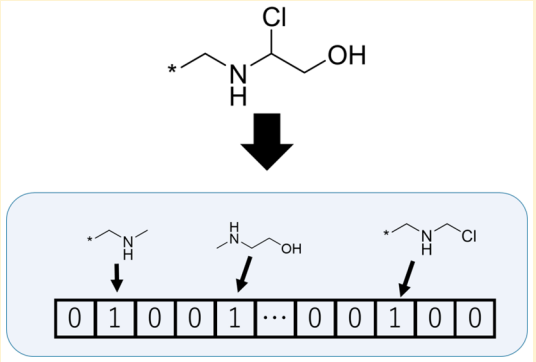
[†]Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

[‡]Data Science Center, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

[§]Department of Chemical System Engineering, School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Supporting Information

**ABSTRACT:** Molecular fingerprints are indispensable in medicinal chemistry for quantifying chemical structures. Fingerprints can be calculated for substructures with attachment points, which are positions where a substructure and a corresponding core structure connect. Because structures with attachment points can be crucial for understanding structure−activity relationships, fingerprints specialized for representing this structural feature are required. R-group fingerprints and R-group descriptors were proposed previously for this purpose; however, these molecular representations have limitations. Current R-group fingerprints do not emphasize information about attachment points, and R-group descriptors are too sensitive to changes in the topological path length from an attachment point. In the present work, we developed novel R-group fingerprints, termed R-path fingerprints, which contain substituent information from an attachment point without being sensitive to small differences in topological distances. The concept of the R-path fingerprints is to describe a chemical substructure from the viewpoint of an attachment point, to distinguish atomistic information around the attachment point and other parts of the substructure. This was achieved by considering all the paths on the shortest path between the attachment point and each atom in a substituent. Benchmark testing was conducted, including comparisons of similarity distributions and potency prediction for R-group substituents. The results showed that R-path fingerprints should be useful for classifying and comparing substructures with attachment points.




## INTRODUCTION

Understanding structure activity-relationships (SARs) is a major goal of compound synthesis campaigns. Analogous molecules are usually synthesized or collected and their biological activity is measured to determine differences in activity among the molecules.[1−4] Visualization and extraction of SAR information from large screening compound data sets by systematically dissecting and organizing compounds have also been reported.[5,6] Although SARs are usually qualitatively evaluated to identify important substituents (substructures) associated with biological activity, methods for quantitative evaluation have also been developed. In Free-Wilson analysis, a linear regression model is constructed using binary variables representing substituents.[7,8] Each variable (substituent) takes a value of one or zero, corresponding to the existence or absence of the substituent in a molecule. This activity prediction (explanation) modeling method has been applied successfully to many pharmaceutical projects to help understand SARs.[9,10] Free-Wilson models, however, cannot predict the biological activity of chemical structures with unseen substituents.

R-group descriptors were developed for quantitative assessment of substituent similarity[11,12] and represent substituents with attachment points. An attachment point of a substituent represents the point where the substituent is attached to a core structure, usually represented by the symbol R. Because substituents can be transformed into numerical vectors by R-group descriptor calculations, comparison of any substituents regarding predicted biological activity becomes possible.

In a previous study, R-group descriptors were calculated by assigning each heavy atom in a molecule to physiochemical feature values and summing up the values based on the topological distances from an attachment point.[11] These R-group descriptors focused on the location of attachment points, to understand the structure of the substituent from the viewpoint of a local landscape of the attachment point. Bioisosteric substituents were successfully distinguished from nonbioisosteric substituents using the R-group descriptors.[12]

**Received:** February 10, 2019
**Published:** May 6, 2019







DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663

Molecular fingerprints are indispensable in medicinal chemistry for quantifying molecular features, including molecular similarity,[13] and this type of molecular representation has been used for R-groups, termed R-group fingerprints. R-group fingerprints using molecular descriptor signatures[14] were investigated and found to be efficient in potency prediction of compounds in patent data sets.[15] Extended connectivity fingerprints (ECFP) were also calculated for substructures with attachment points and succeeded in predicting whether two similar compounds exhibited similar or distinct potencies.[16]

R-group fingerprints should contain information about a substituent from the viewpoint of an attachment point. The attachment point represents another substructure (core structure). Thus, how similar or different substituents are should not be discussed without considering the contribution of the attachment point. Two R-groups can be compared, with respect to SARs, only when the shared core substructure attached to the R-groups takes the same conformation between the whole structures produced by combing the core and the R-groups. Otherwise, it is very hard to extract useful SARs information. The R-group descriptors developed by Martin et al. contain this type of information, which sums feature values based on topological distances from an attachment point.[11] This method, however, is too sensitive to small topological distance changes, which will be explained in the Materials and Methods section.

Herein, we developed novel R-group fingerprints, termed R-path fingerprints. R-path fingerprints contain substituent information from an attachment point without being too sensitive to small differences in topological distances. This was achieved by considering all of the paths on the shortest path between the attachment point and each atom in a substituent. Benchmark testing was conducted, including comparisons of similarity distributions and potency prediction for R-group substituents. The results showed that potency predictions with the proposed R-path fingerprints are similar or slightly better than R-group signature fingerprints.

## MATERIALS AND METHODS

R-group fingerprints should be based on local information around an attachment point, which is the point that connects the substituent and the corresponding core, as well as information describing the whole substituent (global information). The proposed R-group fingerprints approach, termed R-path fingerprints, describes local and global information by accounting for how atoms and bonds are connected from the viewpoint of the attachment point. In the following subsections, R-path fingerprints will be explained in detail, and two other R-group fingerprints and ECFP will also be presented,[17] which were used as controls.

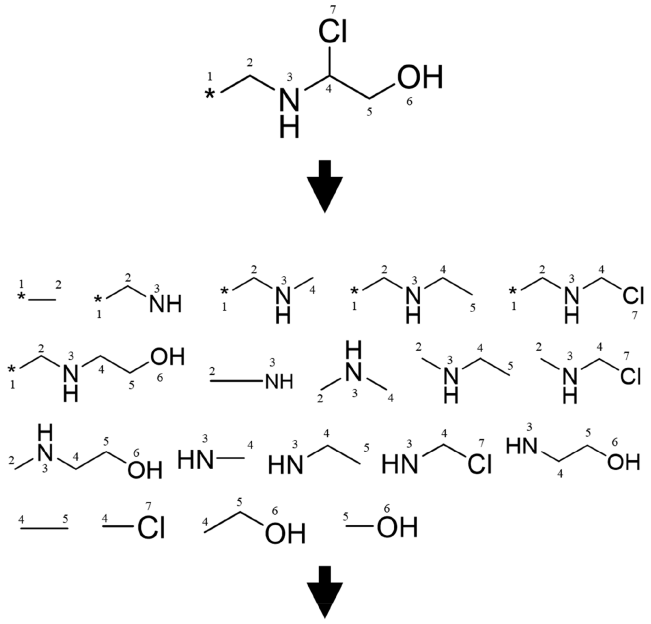
**R-Group Signatures.** R-group signatures are molecular signatures involved in attachment points. Molecular signatures[14] have been used for many projects on quantitative structure−activity/structure−property relationships (QSARs/QSPRs) and virtual screening campaigns, and are useful for capturing local atom environments in a molecule.[18−20] Molecular signatures consist of atom signatures. An atom signature is the canonical representation of a tree with a predefined height whose root is the atom itself. A unique set of atom signatures forms the molecular signatures. Two types of R-group signatures were employed in this work: (1) R-group signatures with heights from 0 to 3 prepared by using the same approach as previous research,[15] termed R-group signatures (binary) and (2) R-group signatures with heights from 0 to 3 that only consist of atom signatures whose root atom is the attachment point, termed R-group signatures (R-root). We used binarized signatures as fingerprints not containing frequencies of atom signatures in a molecule. R-group signatures were calculated by an open-source visual platform for chemo- and bioinformatics software Bioclipse[21] (version 2.6.2) followed by manual curation and extraction of signatures using in-house python scripts.

**Extended Connectivity Fingerprints.** ECFP are feature set fingerprints.[17] These features contain layered atom environment information by merging neighbor atoms' information represented by atom invariants. Features are transformed into integer numbers by applying a hashing function and the set of integers become fingerprints. In this work, the bond diameter 4 was selected for ECFP calculations (ECFP4). Atom invariants for attachment points were calculated by regarding the points as an additional atom type, which is different from any other atoms in the periodic table. ECFP4 fingerprints were folded into a 1024-bit vector via a modulo operation.

**R-Path Fingerprints.** The concept of the proposed fingerprints is to describe a chemical substructure from the viewpoint of an attachment point, thereby distinguishing atomistic information around the attachment point from other parts of the substructure. This is rationalized by the fact that an attachment point in a chemical substructure is recognized as the place for another substructure. Hence, information emphasizing the attachment point might be important when comparing two substructures that are attached to the same core substructure. Figure 1 illustrates the process of calculating R-path fingerprints for the substructure: 2-chloro 2-(methylamino)ethanol with an attachment point connected to the designated carbon atom 2. The shortest path from the attachment point is assigned to each atom in the substructure. The shortest path between the Cl atom (number 7) and the OH group (number 6) is not considered because this path is not associated with the attachment point. Subsequently, all of the paths on each shortest path are extracted, termed subpaths here, and each of them is assigned to a hash value. When assigning a hash value, atom and bond order is preserved on the basis of the nearest atom to the attachment point. In Figure 1, when making feature values for subpaths of the hydroxyl group, all of the subpaths on the path starting from the attachment point and the hydroxyl group are created. There are 14 subpaths on the path excluding individual atoms themselves, (1,6), (1,5), (1,4), (1,3), (1,2), (2,6), (2,5), (2,4), (2,3), (3,6), (3,5), (3,4), (4,6), and (4,5), by pointing the start and end atom numbers on the subpaths. Each subpath is converted into a feature without changing the atom order. Therefore, C−NH and NH−C in Figure 1 are two different features. After all of the subpaths are converted into hashed feature values, they are condensed into a unique set by eliminating duplicated hash feature values. Because many machine learning methods require a fixed-length of independent variables, in this study, 4096 bit-vectors were generated from feature value sets via a modulo operation.

*Atom and Bond Invariants.* Atom and bond invariants (identifiers)[17] can be determined, depending on the resolution required for analyses. For example, when focusing on a substituent as a graph, atom invariants should be the degree of connection to neighboring heavy atoms, and bond invariants


DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663



-2048111647, -1429303803, -1343523130, -1162183666, -979983960, -550037201, -530667557, -161922985, -99585804, 856460422, 886068872, 1393225061, 1504238589, 1586193757, 1614163249, 1679160890, 1710188490, 1971327064, 2055644544, 2115698195

**Figure 1.** Overview of R-path fingerprint generation. The shortest path from an attachment point (* atom) is assigned each atom in a molecule. For each atom, all subpaths on the path are enumerated and assigned to hash values without changing the orders of atoms on the subpaths.

take the same value for all bonds. At the other end of resolution, the Daylight atomic invariants rule can also be employed, e.g., standard ECFPs.[17] In this study, without any specification, atom invariants were based on the standard ECFP rule, which is a combination of atomic number, the number of immediate neighbor heavy atoms, formal charge, the number of hydrogen atoms, number of valence electrons and whether the atom is located in a ring. For bond invariants, the multiplicity of covalent bonds was used, including the distinction of aromatic bonds from single and double bonds.

*Parameters of R-Path Fingerprints.* There are two parameters to be determined before R-path fingerprint calculations: (i) the maximum length of paths and (ii) the minimum length of subpaths. The former determines the maximum topological distance from an attachment point to atoms in a substituent. The latter determines the minimum subpath lengths on paths. Higher values for the minimum subpath lengths are expected to decrease matching elements in fingerprints when comparing two substituents containing relatively large numbers of atoms.

**Benchmark Testing.** *Data Sets.* Four compound-activity data sets were compiled from ChEMBL[22] (version 23) based on the number of active compounds. Equilibrium constants ($pK_i$) were used as potency measurements. For each compound set, matching molecular series (MMS)[23] were systematically generated. An MMS is defined as a group of compounds that share a common core structure with different substituents. Using MMS, SAR information can be effectively extracted.[24] MMS were based on matched molecular pairs (MMPs).[25] An MMP is defined as a pair of chemical structures that only differ by a chemical change at a single site. In this study, MMPs were generated by a computationally efficient algorithm[26] using a program with the help of the OEChem Toolkit.[27] A set of MMPs sharing the same core was compiled as an MMS. For simplicity, the number of attachment points was set to one and MMPs were generated by the following conditions: only cutting single bonds were considered, the maximum size of the substituent was restricted to 13 heavy atoms and the maximum difference between substituents was restricted to eight heavy atoms. In this study, MMSs, which contain at least 30 compounds, was used for enhancing statistical accuracy. Consequently, 86 MMS were eligible and their profiles are reported in Table. 1. It should be noted that the core structure of each MMS had only one attachment point to make benchmark calculation simple. Core structure SMILES for each MMS are reported in Table S1 and profile of compounds in each MMS is reported in Table S2 in the Supporting Information.

**Similarity Distribution.** Target-wise substituents and all of the substituents combined together were used for making pairwise Tanimoto similarity distributions. The total number of unique substituents was 1154. For target-based comparison, R-path fingerprints, R-group signature (binary), R-group signature (R-root), and ECFP4 were employed. For the R-path fingerprints, the maximum path length was set to ten and the minimum subpath length was set to one. Different subpath lengths were studied to examine sensitiveness to parameter values for R-path fingerprints. In this case, the minimum subpath length varied from one to three and the maximum path length was fixed to ten.

**Quantitative Structure−Activity Relationship Modeling.** Another benchmark calculation involved testing predictability using the R-path fingerprints. Local QSAR models of potency prediction for substituents from the five targets were constructed using different fingerprints. Employed R-group fingerprints were R-path fingerprints, R-group signature (binary), R-group signature (R-root), and ECFP4. Each of the 86 MMS in the data sets shown in Table 1 was used for predictability evaluation. QSAR models were constructed by

**Table 1. Benchmarking Dataset Profiles**

| | | | | potency ($pK_i$) | | MW (sub) | | no. of HA (sub) | |
|---|---|---|---|---|---|---|---|---|---|
| ChEMBL ID | target | no. of CPDs | no. of MMS | max | min | max | min | max | min |
| 259 | melanocortin receptor 4 | 736 | 19 | 9.30 | 5.11 | 213.05 | 26.02 | 13 | 2 |
| 217 | dopamine D2 receptor | 1487 | 25 | 8.02 | 4.55 | 246.03 | 26.02 | 13 | 2 |
| 234 | dopamine D3 receptor | 1347 | 22 | 9.74 | 5.39 | 259.09 | 26.02 | 13 | 2 |
| 224 | serotonin 2a (5-HT2a) receptor | 1255 | 20 | 8.69 | 6.65 | 189.23 | 26.02 | 13 | 2 |

[a]For each data set, the CHEMBL ID, target macromolecule name, number of compounds (CPDs), number of MMS, maximum and minimum potencies ($pK_i$ values), maximum and minimum molecular weights (MW), and heavy atom (HA) counts for substituents are reported.

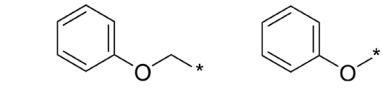


DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663

support vector regressions (SVR).[28] SVR is a variant of a support vector machine (SVM),[29] which is a supervised learning algorithm that aims to identify a hyperplane for separating two classes while maximizing the margin from the hyperplane. SVR was originally proposed as a linear regression method but can be easily extended to nonlinear regression by incorporating a kernel function. In this study, the Tanimoto kernel was used as a kernel function.[30] Predictability was monitored in terms of mean absolute error (MAE) by 10-fold cross validation. Optimization of hyper-parameters, $C$ and $\varepsilon$, was not conducted because of the possibility of overfitting SVR models to fold-out sets. $C$ was set to 1.0, and $\varepsilon$ was set to 0.1 as default parameters in scikit-learn.[31]

## RESULTS AND DISCUSSION

**Exemplary Cases.** Similarity comparison among R-path fingerprints, R-group signatures (binary, R-root), and ECFP4 highlighted characteristics of the R-path fingerprints. In Figure 2, Tanimoto similarity values between methoxybenzene with

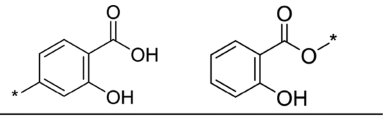


| | |
|---|---|
| R-group signatures (binary) | 0.28 |
| R-group signatures (R-root) | 0.00 |
| ECFP4 | 0.48 |
| R-path fingerprints | 0.36 |

**Figure 2.** Similarity comparison of exemplary structures using different R-group fingerprints. The Tanimoto similarity between the two exemplary structures is reported using distinct R-group fingerprints. R-group fingerprints were binarized R-group signatures (binary), binarized R-group signatures starting with an initial letter [*] (R-root) and R-path fingerprints. The path length for the R-path fingerprints was set to ten.

an attachment point at the terminal methyl and phenol with an attachment point at the oxygen atom are reported. For R-path fingerprints, the similarity value was the second highest following ECFP4. Structural difference between the two substituents was one carbon atom. This small difference, however, made the similarity value zero when using R-group signatures (R-root) because the immediate neighboring atoms to the two attachment points were different from each other. R-group signatures (binary) took a similarity value of 0.28, which is between those for R-path fingerprints and R-group signatures (R-root), because all of the atoms inside each structure were treated equally. Using ECFP4 produced the highest similarity value of 0.48, partly because the majority of the corresponding atoms between the two substituents are in the same atom environment. Figure 3 reports the Tanimoto similarity values between two salicylic acid structures with different attachment points. The similarity value using ECFP4 was the highest followed by R-group signatures (binary). R-path fingerprints were close to zero and R-group signatures (R-root) were zero. ECFP4 and R-group signatures (binary) consider an attachment point no more than a single atom. Because an attachment point implies another substructure, a lower similarity value for the two structures might be preferable in this case.

**Similarity Distributions.** Distributions of the pairwise Tanimoto similarity for target-wise data sets and for all substituents on the basis of different molecular representations



| | |
|---|---|
| R-group signatures (binary) | 0.29 |
| R-group signatures (R-root) | 0.00 |
| ECFP4 | 0.32 |
| R-path fingerprints | 0.09 |

**Figure 3.** Similarity comparison using ECFP4 and R-path fingerprints. The Tanimoto similarity values between the two exemplary structures are reported using ECFP4 and R-path fingerprints. The path length for the R-path fingerprints was set to ten.

are reported in Figure 4. Overall, the frequency of occurrence of substituent pairs monotonically decreased as the pairwise

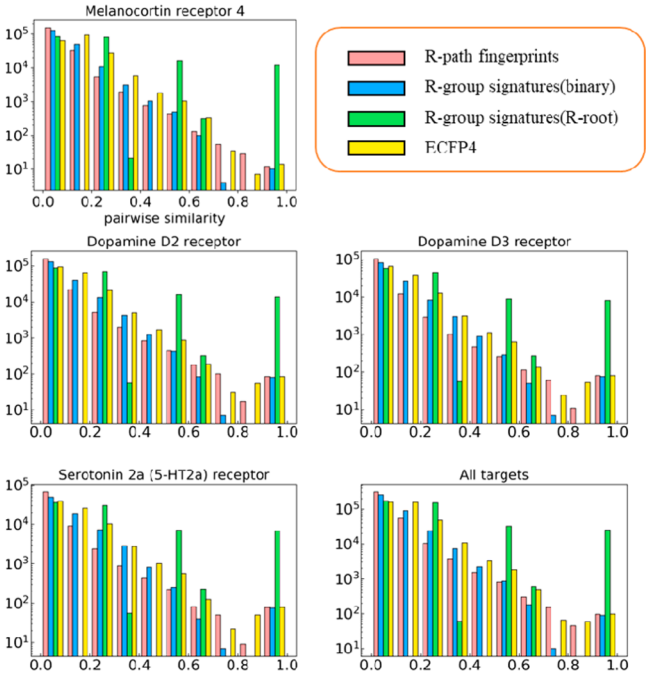



**Figure 4.** Distributions of the pairwise Tanimoto similarity. For each target, distributions of the pairwise Tanimoto similarity for substituents using different fingerprints are reported as histograms. The employed fingerprints were R-path fingerprint, R-group signatures (binary), R-group signatures (R-root), and ECFP4. In addition to target-wise distributions, all the substituents were collected for generating the similarity distribution.

similarity increased. For R-group signatures (R-root), pairwise similarity took discrete values. The signatures considered only the atoms located within three topological distances from an attachment point. Furthermore, atom invariants for R-group signatures are the kind of atoms (i.e., low resolution) that lead to discrete similarity values. The shapes of the distributions among R-path fingerprints, R-group signatures (binary) and ECFP4 were similar to one another, indicating that R-path fingerprints captured chemical structural features similar to the other three fingerprints even though R-path fingerprints considered substituents from the viewpoint of an attachment point.

The next calculation tested the sensitiveness of R-path fingerprints to parameters. Subpath lengths were changed from


DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663

one to three while the maximum path length was fixed at ten. Here, again, pairwise Tanimoto similarity distributions were created for targets. All the substituents were also collected to prepare the pairwise similarity distribution. These distributions are reported in Figure 5. The number of substituents taking

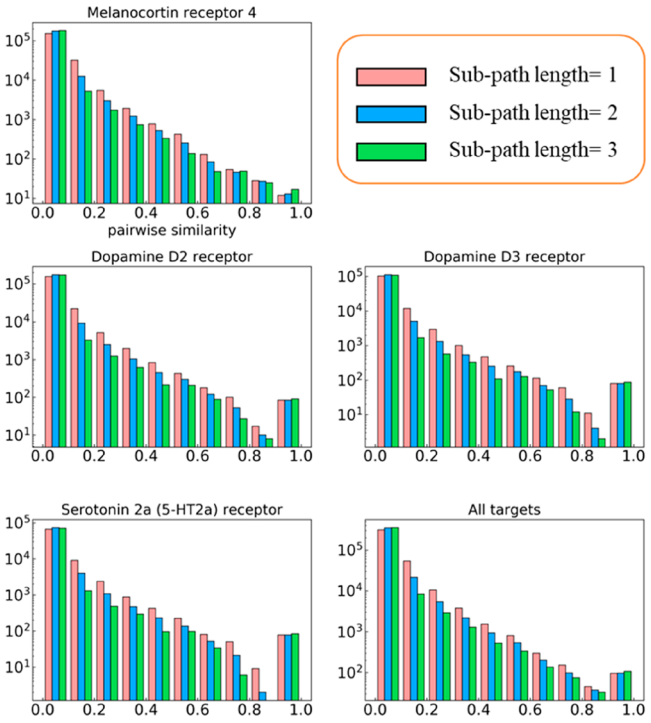



**Figure 5.** Distributions of the pairwise Tanimoto similarity values with a different minimum subpath length. Reported are the histograms of the pairwise Tanimoto similarity values for compounds in a target data set using R-path fingerprints with a different minimum subpath length. The minimum subpath lengths varied from one to three. The path length for the R-path fingerprints was set to ten.

valid R-path fingerprint values decreased when increasing the subpath lengths due to small substituents having fewer atoms than the number of minimum subpath length. The shape of the distributions, however, appeared to be similar to one another, showing that R-path fingerprints is not sensitive to the parameter for the sets of substituents employed in the present study. The pairs of substituents with a similarity value of 1.0 were mostly stereoisomers. Currently, the employed atom and bond invariants cannot distinguish stereoisomers like many other fingerprints. Distributions by changing the maximum path length parameter were also created. Maximum path lengths were also changed from four to 13 at an interval three while the subpath length was fixed at one. The shape of the distributions appeared to be similar to one another, except for the one with the length four due to low resolution of the fingerprints. These distributions are presented in Figure S1 in the Supporting Information.

Scatter plots of a pair of pairwise Tanimoto similarity values between R-path fingerprints and R-group signatures (binary) and between R-path fingerprints and R-group signatures (R-root) are reported in Figures 6 and 7, respectively. As mentioned above, similarities of R-group signatures (R-root) took discrete values (i.e., seven levels, here). The scatter plot in Figure 6 shows a positive correlation between R-path fingerprints and R-group signatures (binary) in terms of

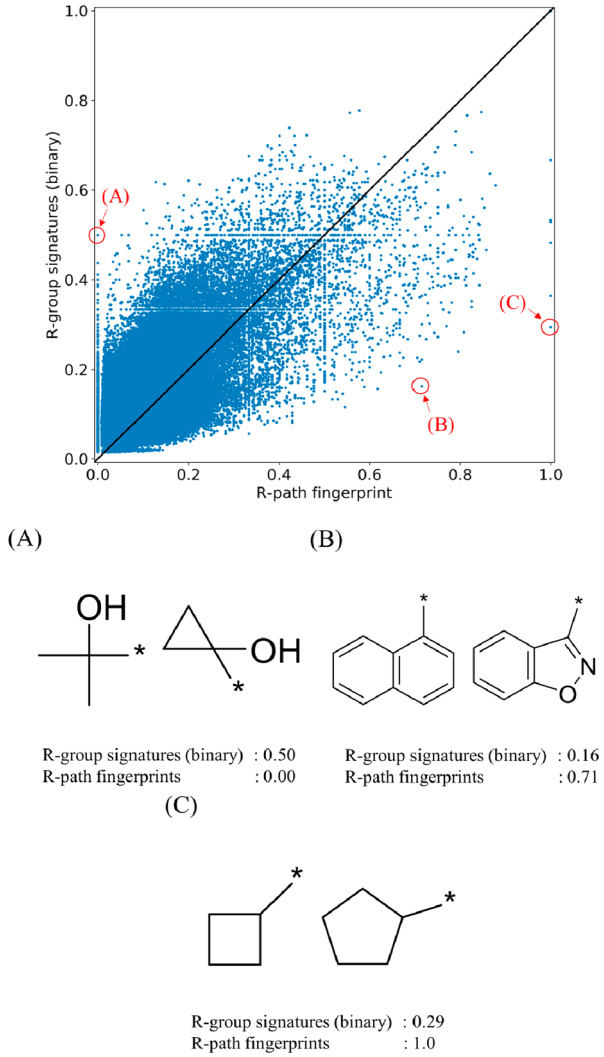



**Figure 6.** Comparison of the pairwise Tanimoto similarity between R-path fingerprints and R-group signatures (binary). Reported is a scatter plot of the pairwise Tanimoto similarities between R-path fingerprints and R-group signatures (binary). Representative pairs, A, B and C, with similarity values are shown below.

pairwise similarity. The correlation coefficient was 0.73. There were, however, several pairs that did not correlate well. These pairs might characterize R-path fingerprints, and exemplary pairs from these are reported in Figure 6. Pair A showed zero similarity between each other using R-path fingerprints, but 0.5 using R-group signatures (binary). This is caused by the difference in the atom invariants. For R-path fingerprints, atom invariants distinguished atoms within and without rings, following the ECFP standard atom invariant rule. In contrast, atom kind is the only atom invariant in R-group signatures, and therefore pair A carbon atoms are indistinguishable. For pair B, R-path fingerprints showed higher similarity between the pair substituents than R-group signatures (binary). The shortest path from the attachment point is assigned to each atom when calculating R-path fingerprints. In the case of pair B, the shortest paths to atoms in the benzene ring that were furthest from the attachment point never passed through the nitrogen and oxygen atoms, leading to relatively high similarity between the pair substituents in terms of R-path fingerprints. In the case of pair C, the similarity between the four-membered and the five-membered rings was 1.0 in terms of R-path fingerprints.


DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663

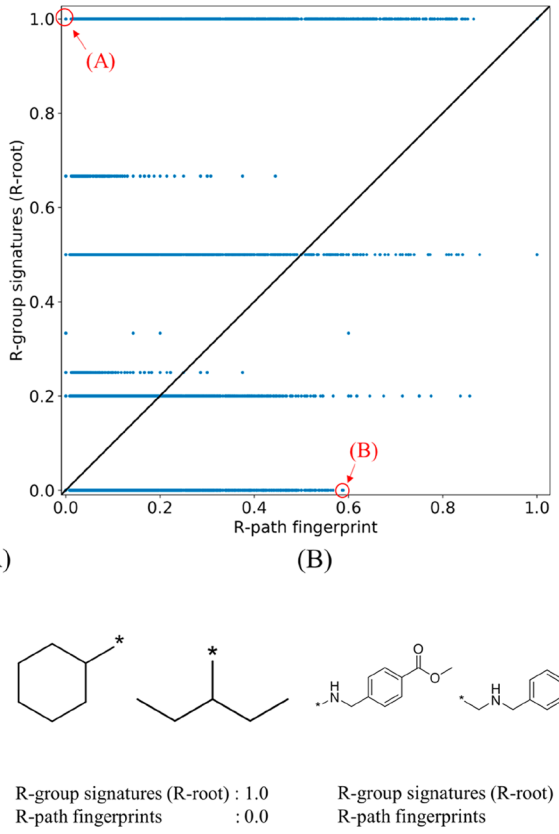



**Figure 7.** Comparison of the pairwise Tanimoto similarity between R-path fingerprints and R-group signatures (R-root). Reported is a scatter plot of pairwise Tanimoto similarities between R-path fingerprints and R-group signatures (R-root). Representative pairs, A and B, with similarity values are shown below.

This also resulted from the shortest path assignment, and thus, there is no distinction between four-membered and five-membered rings. All of the carbon atoms in the five-membered ring (pair C in Figure 6) can be reached within the three topological distances from the attachment point. So can all of the carbon atoms in the four-membered ring of the other substituents of pair C. There was no difference in R-path fingerprints between the two substituents of pair C unless employing the frequency of occurrence of paths rather than binary-type fingerprints. Outlier pairs were also examined by comparing R-path fingerprints and R-group signature (R-root) fingerprints (Figure 7). Pair A showed zero similarity between each other in terms of R-path fingerprints because of the same reason for pair A in Figure 6. Pair B in Figure 7 clarifies the point that it was insufficient to consider only the shortest path to each atom from the attachment point. Using such descriptors (e.g., R-group signatures (R-root)), inserting a single atom between an attachment point and an immediate neighbor atom resulted in zero similarity between the substituents.

**Potency Prediction.** Average MAE values and standard deviation for potency prediction of compounds in MMS are reported in Table 2. The differences in MAE between R-path fingerprints and R-group signatures (binary), between R-path fingerprints and R-group signatures (R-root), and between R-path fingerprints and ECFP4 are reported in Figure 8A–C

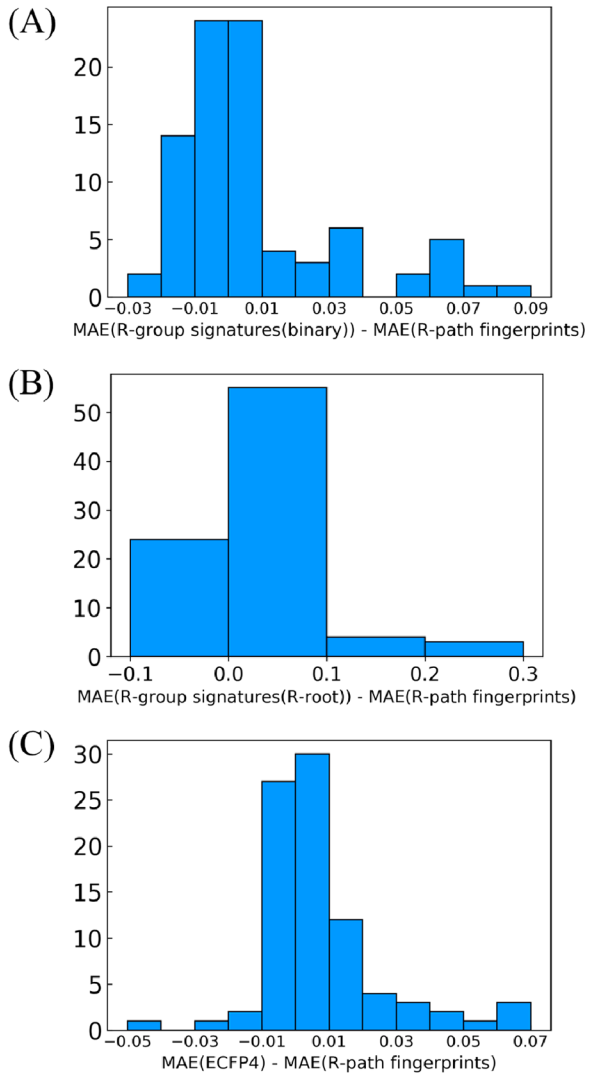



**Figure 8.** Difference in regression performance on the basis of mean absolute error (MAE) values. Reported are histograms of the difference in MAE between R-path fingerprints and R-group signatures (binary; A), R-path fingerprints and R-group signatures (R-root; B), and R-path fingerprints and ECFP4 (C) using support vector regression (SVR) models. MAEs for compounds were calculated via 10-fold cross-validation.

**Table 2. Average Mean Absolute Error in SVR Potency Prediction**[a]

| target | R-group signatures (binary) | R-group signatures (R-root) | R-path fingerprint | ECFP4 |
|---|---|---|---|---|
| melanocortin receptor 4 | 0.35 (0.07) | 0.40 (0.11) | 0.32 (0.07) | 0.34 (0.08) |
| dopamine D2 receptor | 0.22 (0.09) | 0.24 (0.09) | 0.22 (0.09) | 0.22 (0.09) |
| dopamine D3 receptor | 0.23 (0.10) | 0.25 (0.12) | 0.23 (0.08) | 0.24 (0.09) |
| serotonin 2a (5-HT2a) receptor | 0.16 (0.03) | 0.17 (0.03) | 0.17 (0.03) | 0.17 (0.03) |

[a]Standard deviations are provided in parentheses.



DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663

respectively. The number of MMS with a positive MAE difference, meaning that R-path is better, for R-group signatures (binary) and R-path fingerprints was 46 out of 86 MMS, 62 for R-group signatures (R-root) and R-path fingerprints, and 55 for ECFP4 and R-path fingerprints. The performance difference between R-path fingerprints and R-group signatures (binary) and between R-path fingerprints and ECFP4 were marginal; however, R-path fingerprints were superior to R-group signatures (R-root) partly because of the low resolution of R-group signatures (R-root).

Six exemplary substituents are presented in Figure 9, and the measured potency and predicted ones by SVR models using two different fingerprints are provided. Two substituents on the upper row are extracted from the two MMS that showed the most distinct prediction performances by two SVR models using between R-path fingerprints and R-group signatures (binary). For the MMS that substituent A belonged to, potency was predicted better using R-path fingerprints than using R-group signatures (binary). MAE for the MMS using R-path fingerprints was 0.52 and MAE using R-group signatures (binary) was 0.61. For the MMS that substituent B belonged to, potency was predicted with higher accuracy using R-group signatures (binary) than using R-path fingerprints (MAE for R-path fingerprints was 0.20 and MAE for R-group signatures (binary) was 0.18). In the same way, MAE values using R-group signatures (R-root) and R-path fingerprints were compared and exemplary substituents are presented in Figure 9C,D. For the MMS that substituent C belonged to, MAE using R-path fingerprints was 0.30 and using R-group signatures (R-root) was 0.55. Additionally, for the MMS that substituent D belonged to, MAE using R-path fingerprints was 0.20, whereas using R-group signatures (R-root) was 0.17. MAE values using ECFP4 and R-path fingerprints were also compared and exemplary substituents are presented in Figure 9E,F. For the MMS that substituent E belonged to, MAE using R-path fingerprints was 0.36 and using ECFP4 was 0.43. Additionally, for the MMS that substituent F belonged to, MAE using R-path fingerprints was 0.48, whereas using ECFP4 was 0.44. R-path fingerprints were sometimes observed to show low predictability for MMS when the substituents immediately neighboring an attachment point were part of a ring, such as substituent B in Figure 9. In contrast, R-path fingerprints seemed superior to the other R-group signatures in terms of predictability when the immediate neighbor atom was not part of a ring (C and E in Figure 9).

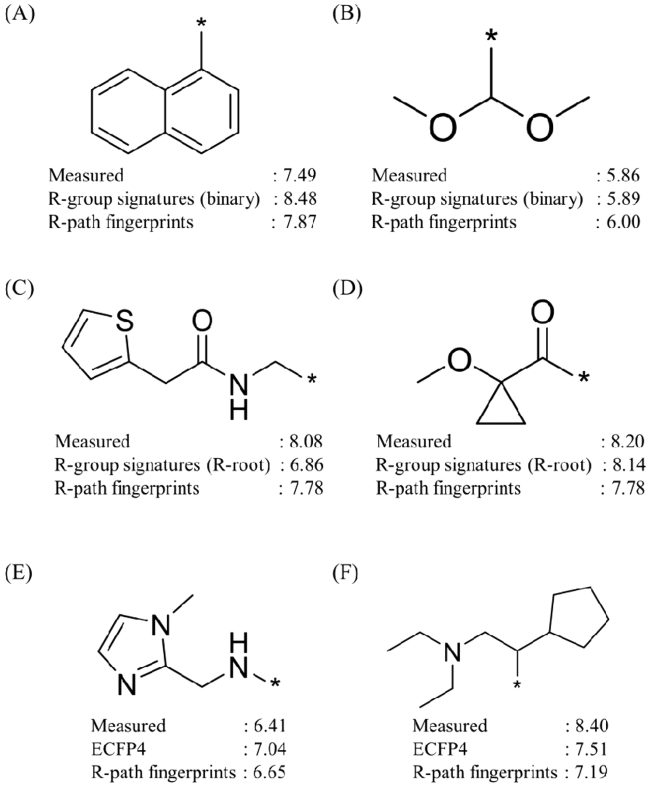



**Figure 9.** Exemplary substituents with predicted and measured potencies. These substituents were extracted from MMS where R-path fingerprints were superior to R-group signatures (binary) by the most margin (A) and vice versa (B), where R-path fingerprints were superior to R-group signatures (R-root; C) and vice versa (D), where R-path fingerprints were superior to ECFP4 (E) and vice versa (F).

## CONCLUSIONS

Herein, we developed a novel R-group fingerprints approach, termed R-path fingerprints, which describe local and global information on a substituent by taking into account how atoms and bonds are connected from the viewpoint of an attachment point. For evaluating R-path fingerprints, similarity distributions of substituents were created and characteristics of the fingerprints were discussed using exemplary substructures. The benchmark calculations for R-path fingerprints clarified that similarity among substituents increased when the attachment points of the substituents were similarly located, and vice versa. This was in accord with our concept of R-group fingerprints. Potency prediction for MMS using the fingerprints was conducted. The results showed that the proposed R-path fingerprints were as efficient as R-group signatures (binary) and ECFP4, which do not take into account the attachment point, and superior to R-group fingerprints (R-root), which take into account the attachment point. Taken together, R-path fingerprints should be a useful tool for classifying and comparing substructures (substituents) with attachment points.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00122.

Figure S1: Distributions of the pairwise Tanimoto similarity values with a different maximum path length; Figure S2: Scatter plot of the pairwise Tanimoto similarities between R-path fingerprints and ECFP4 as well as representative pairs of structures; Figures S3 and S4: Measured versus predicted plots of $pK_i$ for each target using ECFP4 and R-path fingerprints; Tables S1 and S2: Core structure SMILES of each MMS and compound profiles (PDF)

## AUTHOR INFORMATION

### Corresponding Author

*Tel: +81-3-5841-7751. Fax: +81-3-5841-7771. E-mail: funatsu@chemsys.t.u-tokyo.ac.jp.

### ORCID

Tomoyuki Miyao: 0000-0002-8769-2702

Kimito Funatsu: 0000-0002-9368-0302



DOI: 10.1021/acs.jcim.9b00122
J. Chem. Inf. Model. 2019, 59, 2656−2663

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We are grateful to Jürgen Bajorath and Martin Vogt, University of Bonn, Germany, for providing us with codes for MMP fragmentation. We also thank OpenEye Scientific Software, Inc., for providing a free academic license of OpenEye Toolkits.

## ■ REFERENCES

(1) Heim, K. E.; Tagliaferro, A. R.; Bobilya, D. J. Flavonoid Antioxidants: Chemistry, Metabolism and Structure-Activity Relationships. *J. Nutr. Biochem.* **2002**, *13*, 572−584.
(2) Cos, P.; Ying, L.; Calomme, M.; Hu, J. P.; Cimanga, K.; Van Poel, B.; Pieters, L.; Vlietinck, A. J.; Berghe, D. Vanden. Structure−Activity Relationship and Classification of Flavonoids as Inhibitors of Xanthine Oxidase and Superoxide Scavengers. *J. Nat. Prod.* **1998**, *61*, 71−76.
(3) Willson, T. M.; Cobb, J. E.; Cowan, D. J.; Wiethe, R. W.; Correa, I. D.; Prakash, S. R.; Beck, K. D.; Moore, L. B.; Kliewer, S. A.; Lehmann, J. M. The Structure−Activity Relationship between Peroxisome Proliferator-Activated Receptor $\gamma$ Agonism and the Antihyperglycemic Activity of Thiazolidinediones. *J. Med. Chem.* **1996**, *39*, 665−668.
(4) Zarghi, A.; Arfaei, S. Selective COX-2 Inhibitors: A Review of Their Structure-Activity Relationships. *Iran. J. Pharm. Res. IJPR* **2011**, *10*, 655−683.
(5) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *J. Chem. Inf. Model.* **2012**, *52*, 1769−1776.
(6) Gupta-Ostermann, D.; Bajorath, J. The "SAR Matrix" Method and Its Extensions for Applications in Medicinal Chemistry and Chemogenomics. *F1000Research* **2014**, *3*, 113.
(7) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(8) Kubinyi, H. Free Wilson Analysis. Theory, Applications and Its Relationship to Hansch Analysis. *Quant. Struct.-Act. Relat.* **1988**, *7*, 121−133.
(9) Pike, K. G.; Morris, J.; Ruston, L.; Pass, S. L.; Greenwood, R.; Williams, E. J.; Demeritt, J.; Culshaw, J. D.; Gill, K.; Pass, M.; Finlay, M. R. V.; Good, C. J.; Roberts, C. A.; Currie, G. S.; Blades, K.; Eden, J. M.; Pearson, S. E. Discovery of AZD3147: A Potent, Selective Dual Inhibitor of mTORC1 and mTORC2. *J. Med. Chem.* **2015**, *58*, 2326−2349.
(10) Goldberg, F. W.; Leach, A. G.; Scott, J. S.; Snelson, W. L.; Groombridge, S. D.; Donald, C. S.; Bennett, S. N. L.; Bodin, C.; Gutierrez, P. M.; Gyte, A. C. Free-Wilson and Structural Approaches to Co-Optimizing Human and Rodent Isoform Potency for 11$\beta$-Hydroxysteroid Dehydrogenase Type 1 (11$\beta$-HSD1) Inhibitors. *J. Med. Chem.* **2012**, *55*, 10652−10661.
(11) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(12) Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of Intersubstituent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406−411.
(13) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58−63.
(14) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.
(15) Chen, H.; Carlsson, L.; Eriksson, M.; Varkonyi, P.; Norinder, U.; Nilsson, I. Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms. *J. Chem. Inf. Model.* **2013**, *53*, 1324−1336.
(16) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354−2365.
(17) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(18) Kensert, A.; Alvarsson, J.; Norinder, U.; Spjuth, O. Evaluating Parameters for Ligand-Based Modeling with Random Forest on Sparse Data Sets. *J. Cheminf.* **2018**, *10*, 49.
(19) Alvarsson, J.; Lampa, S.; Schaal, W.; Andersson, C.; Wikberg, J. E. S.; Spjuth, O. Large-Scale Ligand-Based Predictive Modelling Using Support Vector Machines. *J. Cheminf.* **2016**, *8*, 39.
(20) Chen, J.; Schmucker, L.; Visco, D.; Chen, J. J.; Schmucker, L. N.; Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules* **2018**, *8*, 24.
(21) Spjuth, O.; Helmus, T.; Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. Bioclipse: An Open Source Workbench for Chemo- and Bioinformatics. *BMC Bioinf.* **2007**, *8*, 59.
(22) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.
(23) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.
(24) Ghosh, A.; Dimova, D.; Bajorath, J. Classification of Matching Molecular Series on the Basis of SAR Phenotypes and Structural Relationships. *MedChemComm* **2016**, *7*, 237−246.
(25) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.
(26) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(27) OEChem TK, version 2.1.4; OpenEye Scientific Software, I: Santa Fe, NM.
(28) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199−222.
(29) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: Berlin, 2000.
(30) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093−1110.
(31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.



DOI: 10.1021/acs.jcim.9b00122
*J. Chem. Inf. Model.* 2019, 59, 2656−2663