Large-Scale Comparison of Alternative Similarity Search Strategies with Varying Chemical Information Contents

Oliver Laufkötter, † Tomoyuki Miyao, ‡ and Jürgen Bajorath*,†

†Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany
‡Data Science Center and Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

ABSTRACT: Similarity searching (SS) is a core approach in computational compound screening and has a long tradition in pharmaceutical research. Over the years, different approaches have been introduced to increase the information content of search calculations and optimize the ability to detect compounds having similar activity. We present a large-scale comparison of distinct search strategies on more than 600 qualifying compound activity classes. Challenging test cases for SS were identified and used to evaluate different ways to further improve search performance, which provided a differentiated view of alternative search strategies and their relative performance. It was found that search results could not only be improved by increasing compound input information but also by focusing similarity calculations on database compounds. In the presence of multiple active reference compounds, asymmetric SS with high weights on chemical features of target compounds emerged as an overall preferred approach across many different activity classes. These findings have implications for practical virtual screening applications.

1. INTRODUCTION

Chemical similarity searching (SS) using molecular fingerprints is one of the mainstay methods in the chemoinformatics field1,2 and continues to be one of the most popular approaches for ligand-based virtual screening.3–5 Fingerprints are bit string representations of molecular structure and properties and produce molecule-specific linear bit patterns.1,2 In SS, fingerprint representations of query (reference) and target (database) compounds are compared using similarity metrics, first and foremost, the Tanimoto coefficient (Tc).1 Fingerprint overlap is quantified as a measure of molecular similarity.1,2 In virtual screening, query compounds typically are known actives that are used to search databases and rank database compounds according to decreasing similarity to reference molecules.3 Calculated fingerprint similarity is then used as an indicator of activity similarity.2 Despite its conceptual simplicity, SS has been successful in many practical applications to identify novel active compounds1–5 and often rivals computationally more complex screening methods.6

A long-investigated issue in SS has been the question of how to best increase the information content of search calculations and maximize the recall of active compounds in benchmark settings as well as the identification of new chemical entities in prospective applications.3 Over the years, this question has been addressed in methodologically different ways. One of the first approaches operates at the level of reference compounds. Compared to search calculations using single reference compounds, the use of multiple references usually increases the recall of active compounds.7 These observations can be rationalized to result from neighborhood behavior of similarity calculations.8 This means that the use of multiple related yet distinct reference molecules expands the chemical neighborhood of given active compounds and increases the likelihood of identifying structurally variable target compounds having similar activity. This neighborhood principle in virtual screening even applies if additional reference molecules are used whose activity status is unknown,9 as long as they are sufficiently similar to known actives and complement their chemical neighborhoods. The use of multiple reference molecules including similar compounds with unknown activity states (presumed inactive compounds) is referred to as “turbo” SS.9,10 This term was coined in analogy to turbochargers that increase engine power through the use of exhaust gases of an engine. Accordingly, turbo similarity searching (TSS) is expected to increase search performance by using inactive compounds that are structural neighbors of known actives. If multiple reference compounds are used, regardless of their activity states, data fusion techniques such as k-nearest neighbor (k-NN) calculations must be applied to merge the results for individual reference molecules and generate a final
similarity score. For example, in 5-NN calculations, the five largest similarity values obtained for k reference molecules are averaged to yield the similarity score of a given database compound, and in 1-NN calculations, the largest of k values is chosen as the final score.

In addition to increasing the number of reference compounds, SS can also be tuned by considering alternative similarity measures. For example, while Tc calculations are symmetrical in nature (i.e., the comparison of the fingerprint of molecule A to the one of B produces the same similarity value as the comparison of B to A). By contrast, calculation of the Tversky index (Tv) makes it possible to induce asymmetry in similarity assessment. By appropriately adjusting weighting factors, increasing weight can be put on molecular representations of query or target compounds, as further discussed below. For example, fingerprint settings of the reference compounds might be preferentially weighted relative to those of database compounds or vice versa.

Another class of methods addresses the issue of similarity search information at the level of molecular representations. In so-called fingerprint or bit profile scaling, bit patterns of multiple reference compounds are compared, and consensus bits are identified that are preferentially set on in reference molecules, given a certain threshold (e.g., 80% of available references). Then, a scaling factor (sf) is applied to consensus bits to increase their relative impact on Tc calculations.

Although different categories of methods to increase the effectiveness of SS have been individually explored in confined benchmark calculations, these approaches have so far not been systematically compared. Therefore, we have revisited the question of how to increase the information content and effectiveness of SS through a large-scale analysis and comparison of different strategies using hundreds of compound activity classes, hence going far beyond typical benchmark settings. In particular, we have been interested in addressing the question to what extent increasing reference compound information or emphasizing characteristic features of query and/or target compounds might contribute to similarity search performance. In the following, the results of our systematic investigation are presented.

## 2. RESULTS AND DISCUSSION

### 2.1. Analysis Concept

Standard SS, TSS, Tversky similarity searching (TvSS), and SS with scaled consensus bits (CBSS) have been compared using varying numbers of reference compounds. The CBSS approach is illustrated in Figure 1. We note that for structural fingerprints, consensus bits can be mapped to reference and hit compounds, and the resulting structural patterns can be interpreted. Such feature mapping can identify important substructures in active compounds that determine structure–activity relationships.

For SS, TSS, and CBSS, the symmetrical Tc similarity metric was applied. SS with single reference molecules served as a baseline for search performance. Calculation details are provided in the Materials and Methods section. Initially, different search strategies were applied to 609 qualifying compound activity classes. On the basis of the results, a subset of 100 activity classes was selected for which SS using single reference compounds approached its limits (corresponding to randomly generated compound rankings). Different search strategies were then compared in greater detail using these “difficult” activity classes, which yielded a further refined view of their relative performance.

We note that TvSS and CBSS depend on predefined parameter settings. In the case of TvSS, these parameters include the \( \alpha \) and \( \beta \) factors that determine relative weights on the bit settings of reference and database compounds, respectively, in Tv calculations. In CBSS, parameters include the cut-off value for determining consensus bits for a set of reference compounds (e.g., bits set on in at least eight of 10 references, corresponding to a cut-off value of 80%) and the sf that is applied to consensus bits in Tc calculations. An initial systematic parameter search to optimize search performance for a sample of activity classes revealed that there were no generally preferred parameter settings, but that best parameters often varied for individual classes, as one might expect. Therefore, alternative parameter settings were applied, as further specified in the Materials and Methods section.

### 2.2. Global Similarity Search Comparison

All similarity-based compound rankings were evaluated using the area under the receiver operating characteristic curve (AUC ROC). Figure 2 shows boxplots of mean AUC ROC for all 609 activity classes and different search strategies. Boxplots were grouped by the number of reference compounds that were used and color-coded according to different strategies and parameter settings. All search strategies displayed an enrichment of correctly identified active compounds at high ranks, as reflected by median AUC ROC values of 0.7 or greater. With increasing number of reference compounds, there was a general increase in search performance across all strategies. For 20 reference compounds, mean AUC ROC values greater 0.9 were consistently observed. On a global scale, differences in search performance between different strategies were only small, which was attributed to averaging effects over many different activity classes. Therefore, we next analyzed the distributions of activity classes over different AUC ROC values for standard similarity search calculations using increasing numbers of reference compounds. Figure 3 shows these distributions, which again emphasize the strong gain in similarity search performance when reference compound information was increasing. Hence, SS using single reference compounds provided a baseline performance level. The corresponding SS (1ref) AUC ROC distribution in Figure 3 had a mean value of 0.718 and a standard deviation of 0.122. Thus, an AUC ROC value of 0.6 corresponded almost exactly to one standard deviation below the mean of the SS (1ref)
distribution \((1\sigma = 0.596)\). Activity classes falling into the AUC ROC interval \([0.5, 0.6]\) were categorized as "difficult" because standard SS approached its limits in these cases, with only marginally better performance than random compound ranking (corresponding to an AUC ROC value of 0.5). Thus, these classes represented challenging test cases and were
expected to benefit most from the application of increasingly information-rich search strategies. Therefore, activity classes falling into the AUC ROC interval \([0.5, 0.6]\) were selected for further analysis, yielding exactly 100 classes.

### 2.3. Search Results for Difficult Activity Classes

Figure 4 shows boxplots of mean AUC ROC values for the subset of 100 activity classes and different search strategies. Compared to the global view, a much more differentiated picture was obtained concerning the relative performance of the different similarity search strategies. As expected, search performance was generally lower for these activity classes. However, for 20 reference compounds, median AUC ROC values above 0.8 were observed in all cases except one, hence clearly indicating successful calculations. Compared to SS, TSS did not provide a detectable advantage, although larger numbers of references including "turbo" compounds were used in each case. Hence, the search calculations were largely determined by available active reference compounds, rather than additional decoys. In this context, we note that previous investigations of TSS were carried out on comparably small benchmark sets. By contrast, clear differences were observed for TvSS calculations with an opposing asymmetry and also for CBSS with different parameter settings. Notably, \((Tv-\alpha = 0.01)\) calculations were consistently superior to \((Tv-\alpha = 10)\), especially when fewer than 10 reference compounds were used. For searches with single reference molecule, the median of the AUC ROC distribution for \((Tv-\alpha = 10)\) was close 0.5 (random compound ranking). However, for \((Tv-\alpha = 0.01)\), an increase in the median exceeding 0.6 was observed. Furthermore, the \((Tv-\alpha = 0.01)\) strategy also reached the overall largest median AUC ROC value exceeding 0.85 for 20 reference compounds. The results also showed that there was overall no significant advantage of CBSS over SS and TSS when a large sf of 5 or 10 was used. By contrast, the \((CBSS-sf = 1, \text{cut-off} = 0.05)\) combination exceeded the search performance of SS, TSS, and CBSS with other parameter settings.

### 2.4. Preferred Search Strategies

Figure 4 shows that \((Tv-\alpha = 0.01)\) and \((CBSS-sf = 1, \text{cut-off} = 0.05)\) were the overall best performing search strategies. Performance advantages over other approaches were largest for 5 or fewer reference compounds but also observed for 10 and 20 references, albeit by a smaller margin. This was an unexpected finding because both approaches prioritized characteristics of target compounds over reference molecules. In the case of \((Tv-\alpha = 0.01)\), this was achieved by the asymmetric nature of the similarity calculations, which put 100-fold more weight on the bit settings of database compounds compared to the reference compounds. Corresponding effects were produced by \((CBSS-sf = 1, \text{cut-off} = 0.05)\) through a special form of bit profile scaling. In this case, all bits that were set on in fingerprints of reference compounds were scaled down by their frequency (given sf = 1) for Tc calculations. For example, if a bit was present in half of the reference compounds, it was considered with a value of 0.5 during Tc calculations. Only bits consistently set on in all reference compounds obtained a regular weight of 1. Thus, the \((CBSS-sf = 1, \text{cut-off} = 0.05)\)
strategy de-emphasized bit settings of reference compounds during Tc calculations, thus achieving a similar net effect as (Tv-α = 0.01), which highly weighted bit settings of database compounds during Tv calculations.

2.5. Structural Heterogeneity. Figure 5 (top) shows the distribution of scaffold-to-compound ratios for all activity classes compared to the subset of 100 difficult classes. Scaffold-to-compound ratios were calculated as a measure of intraclass structural heterogeneity. A large scaffold-to-compound ratio indicates the presence of many compounds with distinct core structures. Such structural heterogeneity typically complicates SS. Hence, one might assume that activity classes with low standard similarity search performance might simply be structurally more diverse than others. However, the comparison of scaffold-to-compound ratios in Figure 5 shows that differences in search performance could not solely be attributed to different degrees of structural heterogeneity because many difficult classes fell into the area proximal to the mode of the global distributions. Nonetheless, there was a tendency for classes with increasing scaffold-to-compound ratio to yield reduced search performance, as also revealed in Figure 5 (bottom). The scatter plot reports result for SS (20refs) while highlighting difficult classes (selected on the basis of SS (1ref) calculations). However, the observation that differences in similarity search performance could not simply be attributed to varying degrees of structural heterogeneity among active compounds emphasized the need for an in-depth analysis of relative search performance using different strategies and relevant parameters.

2.6. Focusing on Target Compounds. For any set of active reference compounds, the (Tv-α = 0.01) and (CBSS-sf = 1, cut-off = 0.05) strategies were found to be the generally preferred approaches. Albeit methodologically distinct, these strategies had in common that they emphasized characteristic features of target compounds. Figure 6 shows mean Δ AUC ROC values for these strategies relative to standard SS using three reference compounds for the subset of 100 activity classes. The results are representative for varying numbers of reference compounds. For the majority of activity classes (87/100), the performance of (Tv-α = 0.01) was superior to (CBSS-sf = 1, cut-off = 0.05), albeit often only by a slight margin. However, comparing true positives among highly ranked database compounds for different activity classes showed that these strategies often detected overlapping yet distinct sets of active compounds, as illustrated in Figure 7a. The overlap between active compounds in the top 1% of the ranking generated using the different search strategies is reported in Figure 7b. Hence, although asymmetric SS achieved overall higher performance, both strategies were complementary in detecting active compounds.

Figure 7a also shows that correctly identified active compounds using (Tv-α = 0.01) had a tendency to be smaller and less complex than reference compounds or were substructures of reference compounds. Preferential detection of substructures of active compounds is a direct consequence of highly weighted bit settings of database molecules. It follows that reduced search performance of asymmetric SS with high weights on the bit settings of database compounds relative to SS is likely to result from the presence of potential hits that are larger and chemically more complex than reference compounds. In such cases, putting high weights on the bit settings of reference compounds reverses this tendency and favors the detection of larger hits that contain reference compounds as substructures. Differences in molecular size and complexity typically lead to differences in fingerprint bit density. If fingerprints of potential hits have higher bit density than those of reference compounds, asymmetric SS with high weights on bit settings of reference compounds favors the detection of such hits. Hence, molecular size and complexity effects can be directly related to fingerprint bit densities. Moreover, systematic differences in bit density between active and inactive compounds also affect SS. Such bit density differences can also be exploited through asymmetric SS with appropriate weights, leading, for example, to the preferential de-selection of reference compounds having lower fingerprint bit density than optimized active compounds.

For the 100 activity classes, the median molecular weight of correctly identified hits was 406.1 for SS, 392.1 for (CBSS-sf = 1, cut-off = 0.05), and 386.2 for (Tv-α = 0.01). Furthermore, the median potency of correctly identified hits was 670 nM, for SS, 790 nM for (CBSS-sf = 1, cut-off = 0.05), and 786 nM for (Tv-α = 0.01).

2.7. Concluding Remarks. In this work, we have investigated alternative similarity search strategies on an unprecedentedly large scale and analyzed their relative performance. Figure 7 illustrates that different search strategies often detect overlapping yet distinct sets of active compounds, which are essentially impossible to predict, and hence
emphasize the need for a detailed analysis of search calculations and factors that determine their outcome. In accordance with our analysis concept, identifying activity classes that provided challenging test cases for standard SS and attempting to improve search performance in these cases provided a differentiated picture of alternative methods and strategies. The systematic search trials discussed above revealed that the accuracy of SS could be improved by increasing reference compound information as well as by emphasizing characteristic features of target compounds. First, a general gain in search performance was achieved when increasing numbers of active reference compounds were used, which is intuitive and has been observed previously. Interestingly, further expansion of reference neighborhoods through the addition of turbo compounds had only little effects when assessed systematically. Hence, characteristic features of compounds having a specific activity were more important than adding decoys. Second, however, for any given number of active reference compounds, the \((T_{v} - \alpha = 0.01)\) and \((\text{CBSS-sf} = 1, \text{cut-off} = 0.05)\) search strategies emerged as best-performing approaches. This was unexpected because these strategies emphasized bit settings of target compounds, either directly through asymmetric search calculations, \((T_{v} - \alpha = 0.01)\), or indirectly by de-emphasizing bit settings of reference compounds, \((\text{CBSS-sf} = 1, \text{cut-off} = 0.05)\). By contrast, putting high weight on bit settings of reference molecules through CBSS was less effective. Hence, the analysis revealed a dual role of reference and target compound information for optimizing search performance and identified preferred strategies for advanced SS. TvSS with multiple reference compounds and high weights on target compounds was the overall best approach across all activity classes and should thus be an attractive choice for practical virtual screening applications. Given the large number of activity classes that were investigated, this strategy should be an attractive initial choice. Considering the influence of both reference and target compounds on the search results, asymmetric SS should generally be a preferred approach for virtual screening because it makes it possible to carry out complementary search trials, with high weights on reference or database compounds. Importantly, if highly optimized active compounds from medicinal chemistry are used as references, which is often the case, typical screening hits are expected to be smaller and chemically less complex. Hence, asymmetric SS with high weights on bit settings of database compounds, the preferred strategy identified in our large-scale investigation, favors the detection of such hits.

3. MATERIALS AND METHODS

3.1. Activity Classes. From ChEMBL (release 24),\(^{18}\) compound activity classes were extracted that contained at least 100 compounds with available high-confidence activity data for which the following selection criteria were applied:\(^{16}\) Species “\textit{Homo sapiens}”, relationship type “D”, confidence score “9”, target type “single protein”, activity type “Ki” or “IC\(_{50}\)”, activity relation “\(=\)”, activity unit “nM”; if an activity comment “inactive”, “inconclusive” or “not active” was detected, the compound was removed.

On the basis of these selection criteria, at total of 609 activity classes comprising 259 099 unique compounds were obtained. From compounds of each class, conventional Bemis–Murcko scaffolds\(^{20}\) were extracted and scaffold-to-compound ratios per class determined as an indicator of intraclass structural diversity. On the basis of these ratios, maximally diverse compound sets have a value of 1 (i.e., each compound contains a unique scaffold/core structure).

3.2. Molecular Representation. Compounds were consistently represented as extended connectivity fingerprints with bond diameter 4 (ECFP4)\(^{21}\), a preferred fingerprint design for SS and other chemoinformatics applications. ECFP4 is a feature set fingerprint consisting of layered atom environments encoded using a hashing function. The size of the feature set is compound-dependent but can be folded through modulo to yield a fingerprint with constant number of bits per molecule. A folded 1024-bit version of ECFP4 was generated using RDKit.\(^{22}\)
3.3. Similarity Searching Setup. For each search trial, 1, 3, 5, 10, or 20 reference compounds were randomly selected from each activity class. Then, the search was carried out for the remaining active compounds per class using all other 608 activity classes as the background database (representing potential false positives in compound rankings). Each trial with a single or multiple reference compounds was carried out 30 and 10 times, respectively, with independently selected compounds, and search results were averaged. Calculations with multiple reference compounds were carried out using 1-NN or k-NN data fusion selecting the largest similarity value or the average of the similarity values of all reference compounds as the final similarity score for each database compound, respectively. We found that there were no significant differences in search performance between 1-NN and k-NN calculations and hence report the results of 1-NN calculations herein.

For all search strategies except TvSS, fingerprint Tanimoto similarity was calculated.

\[
Tc = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}
\]

Equation 1 shows the formula for calculating the Tc where A and B refer to the fingerprint representations of query compound A and database compound B, respectively. Tc is defined as the intersection of the fingerprints of A and B divided by the union of the fingerprints A and B.

Database compounds were ranked on the basis of decreasing similarity scores to reference molecule(s). Rankings were evaluated using AUC ROC calculations that relate true and false positive rates to each other.23

For individual similarity search strategies, specific calculations were carried out, as described in the following.

3.3.1. Tversky Similarity Searching. In TvSS, Tc calculations are replaced with the asymmetric Tv similarity measure, which makes it possible to differentially weigh the bit settings of query and target compounds.

\[
Tv = \frac{|A \cap B|}{\alpha|A| + |B| - |A \cap B|} \tag{2}
\]

Equation 2 shows the Tv formula. As above, A and B refer to the fingerprints of reference compound A and target compound B. The \(\alpha\) and \(\beta\) weighting factors are used to introduce the asymmetry into similarity calculations. For \(\alpha = \beta = 1\), Tv corresponds to the symmetric Tc.

A systematic parameter search was carried out for a subset of 30 activity classes and a total of 135 activity classes and sets of 3 reference compounds were randomly selected (yielding a maximum of 40 turbo compounds). For multiple reference compounds \((n = 3, 5, 10, \text{or } 20)\), the top \(n\) or \(2n\) neighbors were selected. For individual similarity search strategies, specifi-
cations were carried out, as described in the following.

3.3.2. Turbo Similarity Searching. For TSS, an initial SS (1ref) search was carried out in the database to identify nearest neighbors of chosen reference compounds, which were then added to the reference set. For single reference compounds, the top three or 10 neighbors were added for different search trials. For multiple reference compounds \((n = 3, 5, 10, \text{or } 20)\), the top \(n\) or \(2n\) neighbors were selected (yielding a maximum of 40 turbo compounds).

3.3.3. Similarity Searching with Consensus Bit Scaling. For CBSS, consensus bits were determined on the basis of their frequency in reference sets and subjected to scaling. For example, if a cut-off value of 80% was applied and two bits were identified that qualified as consensus bits, these bits were scaled taking their reference set frequency into account. If one bit was set on in 80% of the references and the other in 90%, applying a \(sf\) of 5 produced an effective weight of 4.0 for the former and of 4.5 for the latter during Tc calculations. Therefore, two parameters must be set including the frequency cut-off for consensus bits and the \(sf\) applied to them during Tc calculations. For CBSS, a parameter search was carried out for a subset of 30 activity classes and sets of 5 reference compounds to optimize search performance. Cut-off values ranging from 0.1 to 1.0 and \(sf\) ranging from 0 to 10 were investigated. For global search calculations, \((sf, \text{ cut-off})\) combinations were selected that yielded high search performance in several cases including \((5, 0.8), (10, 0.8), \text{and } (0.05, 1)\).

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-7369-100.

ORCID

Tomoyuki Miyao: 0000-0002-8769-2702
Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

The study was carried out and the manuscript written with contributions of all the authors. All the authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project leading to this report has received funding (for O.I.) from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, “Big Data in Chemistry” (“BIG-CHEM”, http://bigchem.eu). The article reflects only the authors’ view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

REFERENCES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. 1998, 38, 983–996.
(2) Stumpfe, D.; Bajorath, J. Similarity searching. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 2011, 1, 260–282.
(3) Willett, P. Similarity-based Virtual Screening Using 2D Fingerprints. Drug Discovery Today 2006, 11, 1046–1053.
(4) Ripphanus, P.; Nisius, B.; Pet teson, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. J. Med. Chem. 2010, 53, 8461–8467.
(5) Ripphanus, P.; Nisius, B.; Bajorath, J. State-of-the-art in Ligand-based Virtual Screening. Drug Discovery Today 2011, 16, 372–376.
(6) Vaneck, A.; Baskin, I. Machine Learning Methods for Property Prediction in Cheminformatics: Quo Vadis? J. Chem. Inf. Model. 2012, 52, 1413–1437.
(7) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azaaou, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-based Methods for
Virtual Screening Using Multiple Bioactive Reference Structures. J. Chem. Inf. Comput. Sci. 2004, 44, 1177–1185.

(8) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior – A Useful Concept for Validation of Molecular Diversity Descriptors. J. Med. Chem. 1996, 39, 3049–3059.

(9) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzouzi, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-based Virtual Screening Using Nearest-Neighbour Information. J. Med. Chem. 2005, 48, 7049–7054.

(10) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N.; Patel, Y.; Willett, P. Turbo Similarity Searching: Effect of Fingerprint and Dataset on Virtual-Screening Performance. Stat. Anal. Data Min. 2009, 2, 103–114.

(11) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. J. Chem. Inf. Comput. Sci. 2004, 44, 1840–1848.

(12) Horvath, D.; Marcou, G.; Varnek, A. Do Not Hesitate to Use Tversky - And Other Hints for Successful Active Analogue Searches with Feature Count Descriptors. J. Chem. Inf. Model. 2013, 53, 1543–1562.

(13) Tversky, A. Features of Similarity. Psychol. Rev. 1977, 84, 327–352.

(14) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. J. Chem. Inf. Comput. Sci. 2001, 41, 746–753.

(15) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. J. Chem. Inf. Comput. Sci. 2003, 43, 1218–1225.

(16) Heikamp, K.; Bajorath, J. How Do 2D Fingerprints Detect Structurally Diverse Active Compounds? Revealing Compound Subset-specific Fingerprint Features through Systematic Selection. J. Chem. Inf. Model. 2011, 51, 2254–2265.

(17) Wang, Y.; Eckert, H.; Bajorath, J. Apparent Asymmetry in Fingerprint Similarity Searching is a Direct Consequence of Differences in Bit Densities and Molecular Size. ChemMedChem 2007, 2, 1037–1042.

(18) 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.

(19) Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. J. Chem. Inf. Model. 2014, 54, 3056–3066.

(20) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. 1996, 39, 2887–2893.

(21) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742–754.

(22) RDKit: Cheminformatics and Machine Learning Software. 2013. http://www.rdkit.org (accessed Jan 2, 2019).

(23) Bradley, A. P. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognit. 1997, 30, 1145–1159.