Journal of
Chemical Information
and Modeling

Article

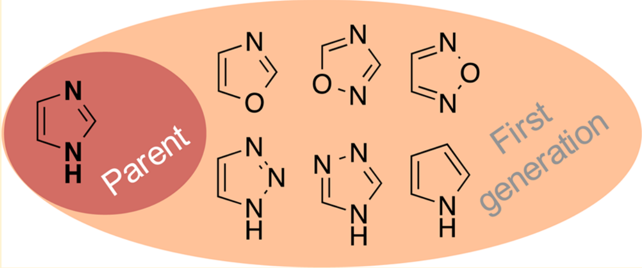
pubs.acs.org/jcim

Applications of Systematic Molecular Scaffold Enumeration to Enrich Structure−Activity Relationship Information

N. Yi Mok* and Nathan Brown

Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London, SM2 5NG, U.K.

Supporting Information



ABSTRACT: Establishing structure−activity relationships (SARs) in hit identification during early stage drug discovery is important in accelerating hit confirmation and expansion. We describe the development of *EnCore*, a systematic molecular scaffold enumeration protocol using single atom mutations, to enhance the application of objective scaffold definitions and to enrich SAR information from analysis of high-throughput screening output. A list of 43 literature medicinal chemistry compound series, each containing a minimum of 100 compounds, published in the *Journal of Medicinal Chemistry* was collated to validate the protocol. Analysis using the top representative Level 1 scaffolds this list of literature compound series demonstrated that *EnCore* could mimic the scaffold exploration conducted when establishing SAR. When *EnCore* was applied to analyze an HTS library containing over 200 000 compounds, we observed that over 70% of the molecular scaffolds matched extant scaffolds within the library after enumeration. In particular, over 60% of the singleton scaffolds with only one representative compound were found to have structurally related compounds after enumeration. These results illustrate the potential of *EnCore* to enrich SAR information. A case study using literature cyclooxygenase-2 inhibitors further demonstrates the advantage of *EnCore* application in establishing SAR from structurally related scaffolds. *EnCore* complements literature enumeration methods in enabling changes to the physicochemical properties of molecular scaffolds and structural modifications to aliphatic rings and linkers. The enumerated scaffold clusters generated would constitute a comprehensive collection of scaffolds for scaffold morphing and hopping.

INTRODUCTION

Establishing the relationships between molecular structures and their modulatory activity against biological targets of interest represents an essential activity of medicinal chemistry in drug discovery. Structure−activity relationship (SAR) information correlates chemical structures with important parameters during hit- and lead-optimization, including both physicochemical properties of and biological responses to ligand molecules. Its widespread application to understand quantitatively the biological effects of chemical structural modifications and the development of predictive models from SAR information are collectively referred to as quantitative SAR (QSAR).

High-throughput screening (HTS) represents a common hit discovery activity in early stage drug discovery.[1] SAR information is often cumulatively enriched when screening-hit matter is progressed toward a clinical candidate through various strategies such as analog-by-catalog, structure-based molecular design, and array chemistry. However, it would be desirable to derive SAR information during the hit discovery stage when classifying hit matter into clusters of compound series. Such SAR would further inform prioritization for hit selection and hit-to-lead activities.

To establish SAR, it is often necessary to define suitable representations of the chemical series that is relevant to the majority of the compounds synthesized and tested, commonly referred to as a molecular scaffold.[2−4] Substituents attached to such defined scaffolds represent structural variants that can be correlated to observed property changes. Objective scaffold representations such as the Murcko framework[5] and the Scaffold Tree classification[2] have been reported. Such scaffold definitions are data set independent, and allow for easy interpretation by expert chemists, hence are commonly applied in establishing SAR information.[4,6,7] In addition, objective scaffold definitions have also been applied in assembling screening libraries[8,9] and analyzing chemical space of fragment screening.[10]

The application of objective scaffold definitions in clustering screening hits represents a structure-oriented approach to understand the SAR of HTS data. Screening compounds that share the same molecular scaffold can be clustered together and SAR information can be readily derived. This approach can quickly identify compound clusters with promising screening results that can be prioritized for further hit-to-lead studies. On the contrary, clustering methods such as molecular fingerprint similarity clustering[11] and HTS fingerprints[12] may cluster

Received: July 2, 2016
Published: December 19, 2016

together molecules of low structural resemblance to each other, which may hinder the assembly of SAR information when prioritizing compound series and developing medicinal chemistry design strategies. Nonetheless, objective scaffold definitions may sometimes be too stringent and result in molecular scaffolds that are only represented by single exemplars without revealing other structural analogs in a screening library. These "singleton" scaffolds may be more challenging for hit confirmation and expansion,[13] and may often be deprioritized when shortlisting screening-hit matter for hit-to-lead activity.

To enhance the application of objective scaffold definitions in HTS compound clustering and triaging, it would be beneficial to introduce some controlled fuzziness in the molecular scaffold representation that is structurally relevant and interpretable for synthetic medicinal chemistry. Literature publications have reported the utilization of molecular enumeration methods in designing medicinal chemistry reagents,[14] enhancing de novo and structure-based molecular design[15,16] and exploring heterocyclic regioisomers (HREMS).[17] Herein, we report the development of *EnCore*, an enumeration approach based upon molecular core scaffolds, that incorporates single atom mutations by systematically introducing elemental atom changes within a defined molecular scaffold. The mutated scaffolds represent structurally related scaffolds to the parent scaffold and are grouped in an *enumerated scaffold cluster*. The method is validated using medicinal chemistry compound series derived from the literature and the DrugBank approved drugs data set to demonstrate its relevance to medicinal chemistry molecular design. Its application in enriching SAR information was investigated using an in-house compound library designed for HTS and a case study of literature cyclooxygenase-2 (COX-2) inhibitors.

■ MATERIALS AND METHODS

***EnCore* Molecular Scaffold Enumeration.** The canonical representation of the input molecular scaffold in simplified molecular-input line-entry system (SMILES) format is used for enumeration implemented using Pipeline Pilot client v9.5.[18] Any explicit hydrogen atoms in the canonical SMILES are removed prior to enumeration. Carbon, nitrogen, and oxygen atoms within an input SMILES, including all aromatic rings, aliphatic rings, and structural linkers, are interchangeable to either of the other two elements, and a single atom change is allowed per mutation. In each generation, all single atom mutations are generated. The mutated SMILES are then converted to molecular representation and the mutated scaffolds are checked for valid valence. Sulfur atom mutation was not included in the default parameters since many of the sulfur mutations produced scaffolds that violated the valence check. Mutated scaffolds violating the valence check are removed before the number of aromatic atoms in the mutated scaffold is compared to the parent scaffold. Any mutated scaffolds with a different number of aromatic atoms (i.e., a single atom mutation would violate the Hückel's rule of aromaticity)[19] are removed. This aromaticity check is analogous to that described in the MORPH algorithm that only alters aromatic ring species in an input molecule.[15] Finally, any duplicate mutated scaffolds are removed. All unique mutated scaffolds within the same enumerated scaffold cluster are used for the next generation of mutations (Figure 1). Figure 2 shows two exemplary *EnCore* enumerated scaffold clusters of imidazole and 3-(4*H*-1,2,4-triazolo-4-yl)pyridine.

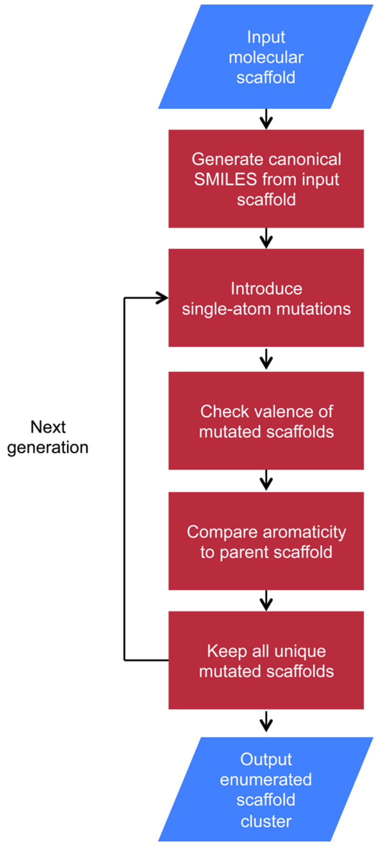

Input molecular scaffold
Generate canonical SMILES from input scaffold
Introduce single-atom mutations
Check valence of mutated scaffolds
Compare aromaticity to parent scaffold
Keep all unique mutated scaffolds
Next generation
Output enumerated scaffold cluster

Figure 1. *EnCore* systematic molecular scaffold enumeration workflow.

Literature Medicinal Chemistry Series. ChEMBL21[20] was mined for compound series reported in the *Journal of Medicinal Chemistry*. To ensure reasonable chemical space has been explored for the selected series, each series was required to contain a minimum of 100 compounds in a single publication, with reported IC_{50} values against a defined molecular target. 62 publications satisfying these criteria were retrieved. Inspection of this publication shortlist removed 19 publications that did not focus on synthetic medicinal chemistry designs (e.g., publications on virtual screening, CoMFA/QSAR analyses of literature compounds, model developments and reviews). The top Level 1 scaffold, defined as the Scaffold Tree Level 1 scaffold[2,13] representing the largest number of compounds within the series, in each of the remaining 43 publications were enumerated (see Supporting Information Table S1 for the list of publications and their top Level 1 scaffolds). All other Level 1 scaffolds explored within each series were then compared to the mutated scaffolds output in the corresponding enumerated scaffold cluster.

DrugBank Data Set. Here, 1826 compounds from the DrugBank approved drugs data set were analyzed.[21] After stripping salts, 962 compounds compliant with Lipinski's rule-of-five[22] and containing at least two rings were retained to generate 475 unique Level 1 scaffolds using the Pipeline Pilot component *Generate Scaffold Tree*.[18] The data set of unique Level 1 scaffolds was used for scaffold enumeration. EPFP7 fingerprints (extended path fingerprint with a distance of seven bonds), analogous to the Daylight molecular fingerprints,[23,24] were generated using Pipeline Pilot[18] and the mutated scaffolds were compared to their parent scaffold. The mutated scaffold with the highest Tanimoto fingerprint similarity to the parent

DOI: 10.1021/acs.jcim.6b00386
J. Chem. Inf. Model. 2017, 57, 27−35

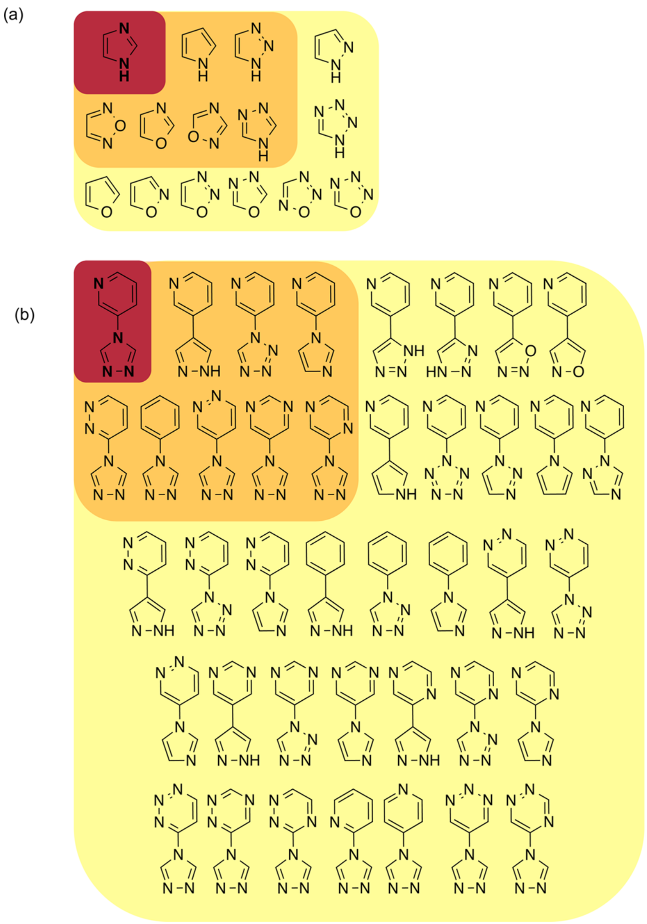


Figure 2. Exemplary *EnCore* enumerated scaffold clusters of (a) imidazole and (b) 3-(4*H*-1,2,4-triazolo-4-yl)pyridine. Scaffold generations are illustrated in different colored backgrounds: parent scaffold (red), first generation mutated scaffolds (orange), and second generation mutated scaffolds (yellow).

scaffold and the average Tanimoto fingerprint similarity of all mutated scaffolds per generation were used to evaluate the scaffold diversity of the enumerated scaffold clusters over generations.

ICR/CRT Screening Library Data Set. The Institute of Cancer Research/ Cancer Research Technology (ICR/CRT) in-house hit discovery screening collection, containing 214 540 compounds, was used as an HTS library data set. Level 1 scaffolds were generated for 214 540 compounds using Pipeline Pilot component *Generate Scaffold Tree*.[18] 23 319 unique Level 1 scaffolds were generated. 11 662 of these scaffolds had only one representative screening compound (singleton scaffold). All unique Level 1 scaffolds were used to generate enumerated scaffold clusters.

Literature Cyclooxygenase-2 (COX-2) Inhibitors Data Set. All small molecules with reported IC_{50} values against cyclooxygenase-2 (COX-2) published in the *Journal of Medicinal Chemistry* were retrieved using ChEMBL21.[20] Level 1 scaffolds were generated for the 1289 unique compounds retrieved from 93 publications (see Supporting Information Table S2 for the list of publications). Where a molecule has multiple IC_{50} values reported in multiple publications, the highest value corresponding to weakest COX-2 inhibition was retained.

RESULTS AND DISCUSSION

To validate the *EnCore* scaffold enumeration protocol, we first investigated its relevance to medicinal chemistry molecular design using the literature medicinal chemistry series data set. The protocol was then applied to the DrugBank data set to define a rule-of-thumb for the number of generations of enumeration. After these criteria are defined, the application of *EnCore* scaffold enumeration in enriching SAR information was investigated using the ICR/CRT library as an HTS compound data set. The capability of *EnCore* scaffold enumeration in increasing the number of structurally related compounds associated with each scaffold was analyzed. Finally, a case study of literature COX-2 inhibitors was used to illustrate the advantage of *EnCore* applications in enriching SAR information from structurally related scaffolds.

All experiments presented here used the Level 1 scaffolds generated following the Scaffold Tree algorithm that systematically deconstructs molecules based on ring-focused dis-

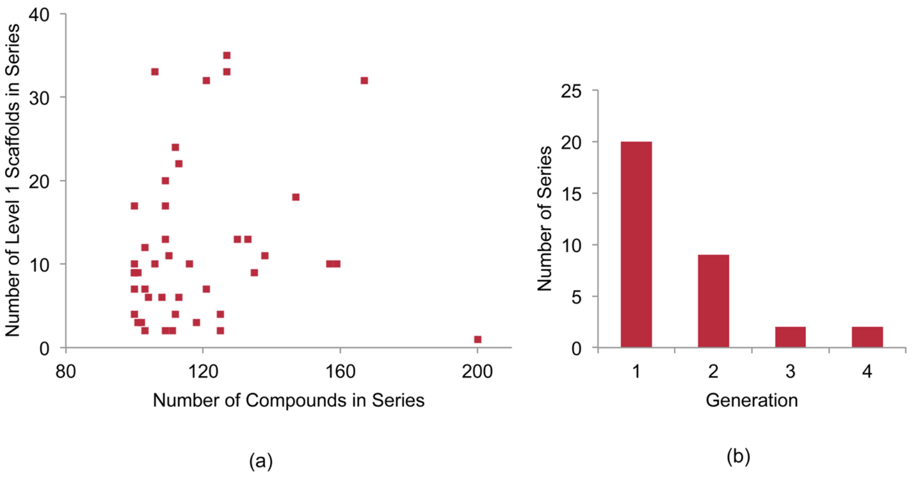


Figure 3. (a) Number of Level 1 scaffolds within the 43 literature medicinal chemistry compound series against the number of compounds within the series. (b) Breakdown of the generation of enumerations in which additional Level 1 scaffolds explored within the same compound series were identified using the enumerated scaffold clusters. For some compound series, multiple Level 1 scaffolds reported in the publication matched the enumerated scaffolds in multiple generations.

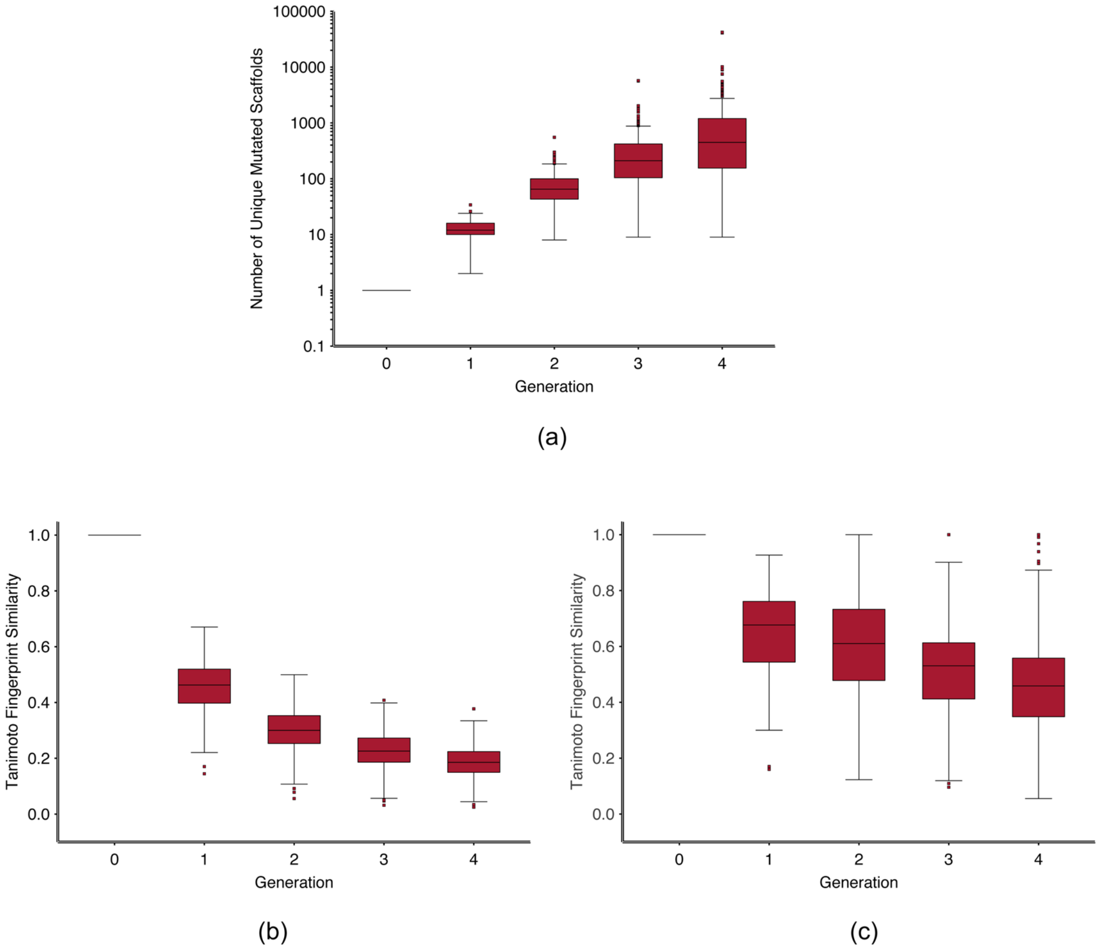


Figure 4. Box-and-whisker plots of distributions of mutated scaffolds from the DrugBank approved drugs data set. (a) Number of new molecular scaffolds obtained from scaffold enumeration over four generations of mutations. (b) Average Tanimoto fingerprint similarity (EPFP7) of the mutated scaffolds to the parent scaffold per generation of mutations. (c) Tanimoto fingerprint similarity (EPFP7) of the mutated scaffolds most similar to their respective parent scaffolds per generation of mutations.

connection rules.[2] In a recent publication on the scaffold diversity of exemplified medicinal chemistry,[13] it was observed that Level 1 of the Scaffold Tree typically represents an appropriate objective and invariant scaffold definition. We therefore utilizes the Level 1 scaffolds as the scaffold representation to illustrate the features of *EnCore* applications.

Theoretically, to exhaustively generate all possible mutated scaffolds of an input scaffold, the enumeration protocol by definition needs to compile as many generations as the number of heavy atoms in the molecular scaffold. As the size of the enumerated scaffold cluster increases over generations, the number of mutated scaffolds increases, thereby increasing the time required to generate these mutations. Therefore, it is necessary to evaluate the number of unique mutated scaffolds generated and their structural diversity per generation in comparison to the parent scaffold, and define a balance between the time required to generate mutated scaffolds and the increase in chemical space sampled by the mutated scaffolds.

The following two experiments analyze the relevance of *EnCore* to medicinal chemistry and the structural diversity of the mutated scaffolds over generations, respectively. We apply our results to validate the *EnCore* protocol and to define the optimal number of enumeration generations as a rule-of-thumb for *EnCore* applications.

Relevance to Medicinal Chemistry Molecular Design. One of the potential applications of *EnCore* aims to mimic the exploration of molecular scaffolds when developing SAR in a drug discovery program. To demonstrate that *EnCore* is relevant for medicinal chemistry molecular design, published compound series from the medicinal chemistry literature were analyzed. This was facilitated by mining the ChEMBL database[25] to identify compound series published in the *Journal of Medicinal Chemistry* that are likely to originate from a medicinal chemistry program.

Figure 3a shows the homogeneity of the 43 collated literature compound series investigated, by plotting the number of unique Level 1 scaffolds against the total number of compounds within each series. This data set represents a diverse profile of medicinal chemistry compound series, from homogeneous series containing only one scaffold representing the entire series of 200 compounds published in one paper, to heterogeneous series represented by over 30 Level 1 scaffolds. Out of the 43 input scaffolds, 23 enumerated scaffold clusters were successful in identifying Level 1 scaffolds explored within the same compound series, i.e. the mutated scaffolds generated in the enumerated scaffold clusters matched other Level 1 scaffolds reported in their respective publication. For some compound series, multiple Level 1 scaffolds reported in the publication matched the enumerated scaffolds in multiple generations. The

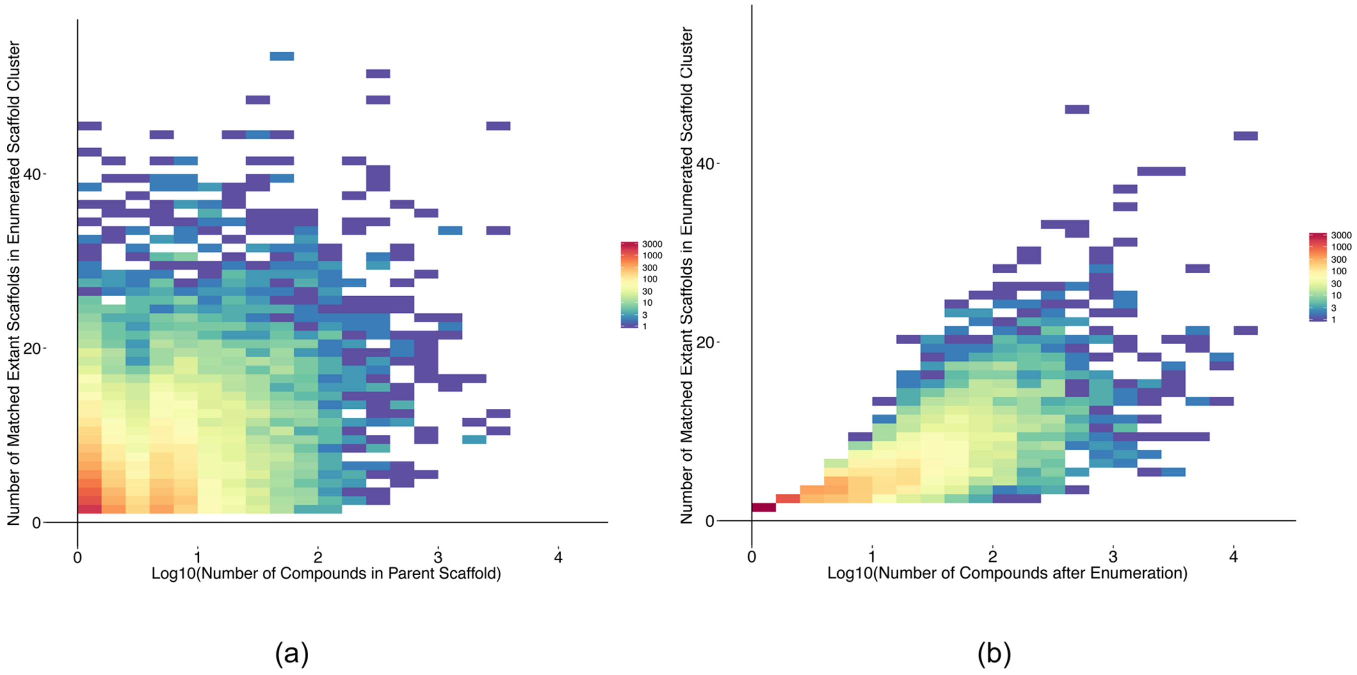


Figure 5. Heat maps showing (a) the number of matched extant scaffolds after scaffold enumeration plotted against the original number of screening compounds represented by the parent Level 1 scaffolds within the ICR/CRT screening library (n = 23 319). The color key shows the density of unique parent scaffolds that have recorded the same number of matched extant scaffolds after enumeration. (b) The number of matched extant scaffolds plotted against the number of structurally related screening compounds associated after scaffold enumeration for all singleton scaffolds within the ICR/CRT screening library (n = 11 662). The color key shows the density of unique parent scaffolds in each bin.

breakdown of the generation of enumerations where other Level 1 scaffolds were matched is shown in Figure 3b. For 20 compound series, mutated scaffolds in the first generation could match other explored molecular scaffolds within the series; for nine series, their second generation mutated scaffolds identified additional molecular scaffolds within the series, whereas only two compound series have mutated scaffolds in the third and fourth generations that matched any explored molecular scaffolds.

The observations in Figure 3 support the hypothesis that the enumerated scaffold clusters can identify additional molecular scaffolds that have been explored within the same compound series reported in literature medicinal chemistry programs. Even though only a subset of the enumerated scaffold clusters could match explored molecular scaffolds reported (23 out of 43 compound series), the results here suggest that *EnCore* can partially mimic the exploration of molecular scaffolds when developing SAR in medicinal chemistry projects.

Structural Diversity of the Mutated Scaffolds over Generations. To further assist defining the optimal number of enumeration generations as a rule-of-thumb for *EnCore* applications, the number of unique mutated scaffolds generated and their structural diversity were assessed using the Level 1 scaffolds derived from the DrugBank approved drugs data set.[21] Figure 4a shows the box-and-whisker plots of the distribution of the number of unique mutated scaffolds generated from the 475 DrugBank Level 1 scaffolds per generation. The median number of unique mutated scaffolds generated is 12 in the first generation, increasing to 65 in the second, 210 in the third, and 448 in the fourth generations. Since the average number of heavy atoms in the parent Level 1 scaffolds for the entire DrugBank data set is 12 ± 3, the increase in the number of unique mutated scaffolds per generation observed is not surprising. It is noteworthy that some of the outliers in the fourth generation, reaching over 10 000 unique mutated scaffolds, originated from parent scaffolds with greater than 25 heavy atoms exemplified by two rings connected by a very long alkyl chain.

EPFP7 fingerprints, analogous to the Daylight molecular fingerprints,[10,23,24] were used to assess the structural diversity of the unique mutated scaffolds in comparison to their parent scaffold per generation. Figure 4b shows the average Tanimoto fingerprint similarity for all unique mutated scaffolds to their respective parent scaffold per generation, whereas Figure 4c shows the Tanimoto fingerprint similarity of the mutated scaffold most similar to its parent scaffold per generation. From these plots, it can be observed that the structural similarity of the unique mutated scaffolds to their parent scaffolds decreases sharply over generations, with the median of the distribution at only 0.46 in the first generation when considering the average of all unique mutated scaffolds generated, and below 0.3 beyond the second generation. Even when considering the most similar mutated scaffold in each generation, the median fingerprint similarity value drops below 0.6 after two generations of mutations. All the median values fall below the threshold Tanimoto similarity cutoff of 0.85, above which any molecular pairs may be considered chemically similar.[24] Together with the analysis on literature medicinal chemistry compound series, it can be concluded from these two experiments that two generations of enumeration is a reasonable balance between chemical space sampling and the likelihood of the mutated scaffolds within the enumerated scaffold cluster having been explored in medicinal chemistry. All further *EnCore* applications described within this paper will refer to two generations of enumerations as the optimal number defined here.

Relevance to Enriching SAR in HTS Library. To demonstrate the applicability of *EnCore* in enriching SAR, we

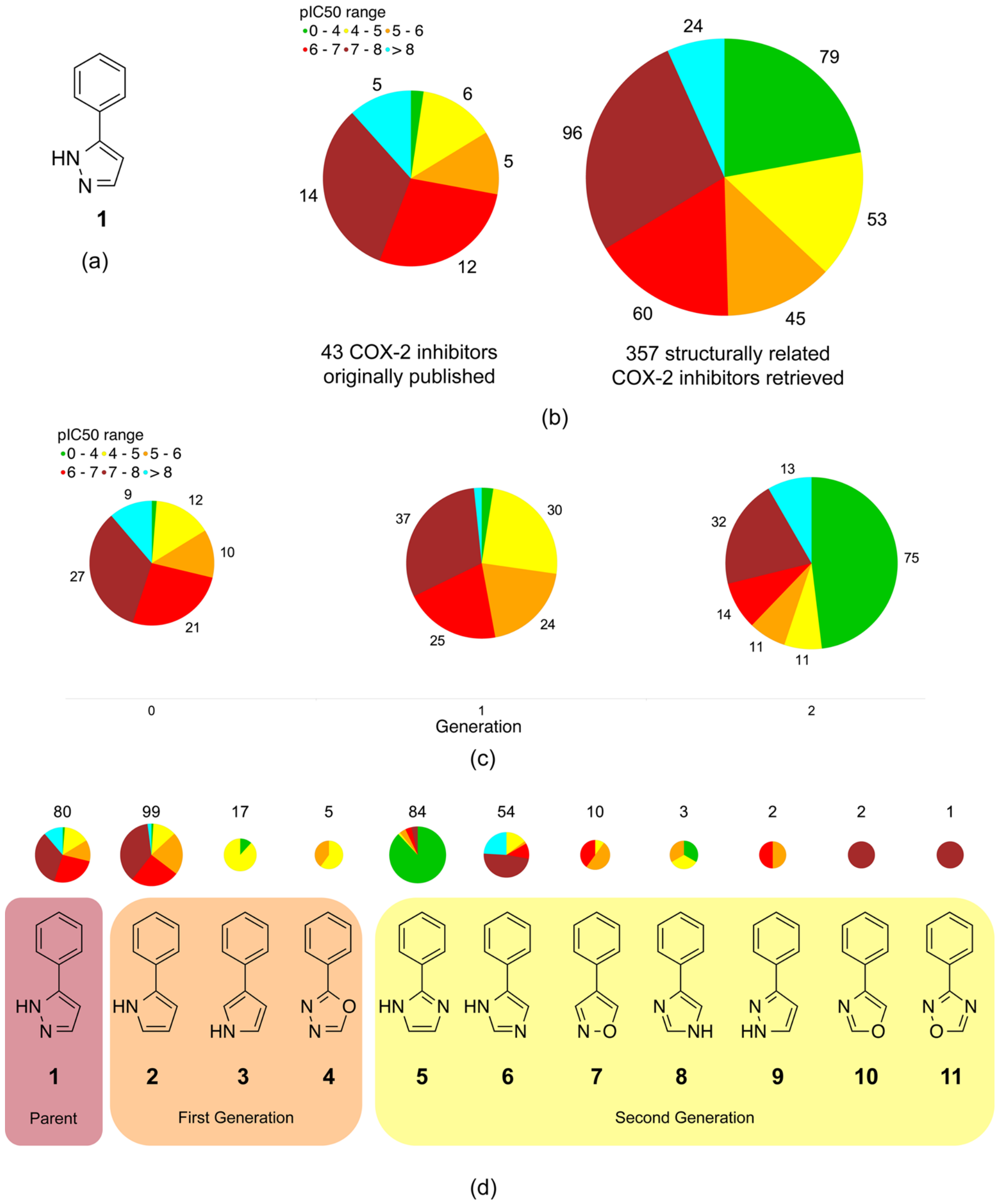


Figure 6. Expansion of SAR in COX-2 inhibitors is enabled through applying *EnCore* scaffold enumeration. (a) Top Level 1 scaffold, 5-phenyl-1*H*-pyrazole (**1**), identified in the COX-2 inhibitors published by Penning et al.[26] (b–d) Distribution of the COX-2 pIC_{50} values (b) from the inhibitors published originally (n = 43) and structurally related inhibitors retrieved from ChEMBL21 (n = 357); (c) grouped by generations of mutated scaffolds (n = 357); and (d) grouped by individual Level 1 scaffolds (n = 357). Numbers in pie charts correspond to compound count in each pIC_{50} range category unless stated otherwise. Numbers in pie charts in part d correspond to compound count represented by each scaffold.

evaluated a screening library that has been designed for HTS applications. When conducting an HTS, it is common practice to quickly identify hit series that both deliver the desirable biological responses and indicate initial SAR to guide medicinal chemistry molecular design. However, commonly, SAR information from primary HTS output can be limited; for example, when a screening hit is classified as a "singleton" where no structurally related analogs are represented within the library.

Using the ICR/CRT screening library containing over 200 000 compounds, we explored the application of *EnCore* to enrich SAR information by classifying the library compounds into multiple enumerated scaffold clusters. Enumerated scaffolds for each individual parent scaffold were compared to other extant scaffolds present in the screening library. A match between extant scaffolds and those within the enumerated scaffold cluster of the parent scaffold would indicate an increase in the number of structurally related screening compounds associated with that particular parent scaffold, hence the possibility of enriching SAR information.

All 23 319 Level 1 scaffolds from the entire ICR/CRT screening library were enumerated over two generations and the enumerated scaffold clusters were assessed. Figure 5a shows a heat map indicating the number of molecular scaffolds after enumeration against the original number of screening compounds associated with each parent scaffold in the screening library. Each bin in the heat map contains a population of unique parent scaffolds that recorded the same number of matched extant scaffolds after enumeration, and the color key represents the density of parent scaffolds in each bin. Out of the 23 319 parent scaffolds, extant scaffolds within the screening library have been identified in the enumerated scaffold clusters of 17 199 scaffolds, whereas the remaining 6120 scaffolds (the bottom row in the heat map in Figure 5a) did not match any extant scaffolds after enumeration. Two parent scaffolds had the largest number of matched extant scaffolds in their enumerated scaffold clusters, matching 54 extant scaffolds representing 3976 screening compounds in total. Since each of these two parent scaffolds represent only around 50 screening compounds individually, the increase in the number of screening compounds associated with these parent scaffolds after *EnCore* application represents a substantial expansion of structurally related compounds within the screening library for which SAR information may be readily available.

Another striking observation is that out of the 11 662 singleton parent scaffolds (the leftmost column in the heat map in Figure 5a), 7369 parent scaffolds matched with extant scaffolds after *EnCore* enumeration. This indicates that structurally related screening compounds within the library can be associated with over 60% of the singleton scaffolds that each represent only one screening compound. Since by definition the compounds in these singleton scaffolds have no structural analogs sharing the same molecular scaffold, the application of *EnCore* identifies structurally related analogs and enhances the possibility of obtaining readily available SAR information from within the screening library. A more detailed heat map for singleton scaffolds (Figure 5b) shows the number of matched extant scaffolds against the number of screening compounds after *EnCore* enumeration. This further illustrates that many singleton scaffolds have a substantial increase in the number of structurally related screening compounds after *EnCore* enumeration. Two singleton scaffolds have over 10 000 screening compounds associated after *EnCore* enumeration, represented by 21 and 43 extant scaffolds (the rightmost column in the heat map in Figure 5b).

Case Study Using Literature COX-2 Inhibitors. To demonstrate the expansion of SAR enabled through applying *EnCore* scaffold enumeration in HTS data analysis, literature COX-2 inhibitors were retrieved and analyzed as a simulated screening data set.

In the literature medicinal chemistry series collated, the 5-phenyl-1*H*-pyrazole scaffold **1** (Figure 6a) was identified as the top Level 1 scaffold representing 43 compounds in the COX-2 inhibitors published by Penning et al.[26] Its enumerated scaffold cluster after two generations of *EnCore* enumeration contains 71 unique scaffolds (Table S1). 1289 unique compounds with reported IC_{50} values against COX-2 published in 93 *Journal of Medicinal Chemistry* publications were retrieved from ChEMBL21 (Table S2). Eleven Level 1 scaffolds, representing 357 unique compounds out of the 1289 COX-2 inhibitors retrieved, matched scaffolds in the *EnCore* enumerated scaffold cluster from **1**. Figure 6b compares the distribution of the pIC_{50} values among the 43 compounds originally published in Penning et al.[26] and the expanded list of 357 compounds retrieved using the *EnCore* enumerated scaffold cluster. The number of compounds with COX-2 $pIC_{50} \geq 6$ increased from 31 to 180 compounds, whereas the number of weak inhibitors with COX-2 $pIC_{50} < 4$ also increased from one to 79 compounds. These observations illustrate that SAR information can be readily enriched from compounds with structurally related scaffolds as a result of applying *EnCore* scaffold enumeration.

Grouped by generations of mutated scaffolds (Figure 6c), 80 compounds (57 with COX-2 $pIC_{50} \geq 6$) are represented by the parent scaffold **1**, whereas 121 compounds (64 with COX-2 $pIC_{50} \geq 6$) are represented by first generation mutated scaffolds and another 156 compounds (59 with COX-2 $pIC_{50} \geq 6$) are represented by second generation mutated scaffolds in the enumerated scaffold cluster. Further analysis of these 357 compounds by individual Level 1 scaffolds (Figure 6d) suggested that many of the 153 analogs represented by the 2-phenyl-1*H*-pyrrole **2** and 5-phenyl-1*H*-imidazole **6** scaffolds could inhibit COX-2, whereas the majority of the 84 analogs represented by 2-phenyl-1*H*-imidazole **5** were inactive against COX-2 (74 with COX-2 $pIC_{50} < 4$). This demonstrates the advantage of *EnCore* application in instantly establishing SAR information from structurally related scaffolds. Although this is a retrospective analysis of published COX-2 inhibitors, this method is also applicable to prospective analysis, for instance, of compound screening data from an HTS campaign.

Our analyses of literature medicinal chemistry compound series and an HTS library illustrate two potential applications of *EnCore* to enable medicinal chemistry design by enriching SAR information. Using the top Level 1 scaffolds in the 43 literature medicinal chemistry compound series retrieved from the *Journal of Medicinal Chemistry*, *EnCore* is able to identify other explored scaffolds within the same publication for 23 series. This satisfies one of the aims of developing *EnCore* to enrich SAR information in medicinal chemistry molecular design.

Although our molecular scaffold diversity analysis suggests an optimum of two generations of mutations as a default setup, *EnCore* can generate multiple generations of mutated scaffolds if the enumerated scaffold clusters require broader diversity from the parent scaffold. The protocol can theoretically enumerate as many generations of mutated scaffolds as the number of heavy atoms in the scaffold. The resultant enumerated scaffold cluster constitutes a comprehensive collection of structurally related molecular scaffolds for scaffold morphing and hopping.[15,27] In comparison to manual molecular scaffold designs, the enumerated scaffold cluster would be advantageous in removing potential bias against less common but structurally feasible molecular scaffolds, whereas synthetically intractable molecular scaffolds may stimulate the development of innovative synthetic methods.[28]

Our experimental observations from enumerating molecular scaffolds of the HTS library offer interesting insights into the classification of compounds into structurally related enumerated scaffold clusters. While various methods of compound clustering have been extensively applied in SAR analyses, *EnCore* offers a comprehensive method to identify structurally related analogs that are relevant to SAR exploration. Structurally related screening compounds within the library can be associated with over 60% of the singleton scaffolds after molecular scaffold enumeration. Analogously, over 70% of all the scaffolds in the compound library matched extant scaffolds within the library after enumeration. Even though the increase in the number of screening compounds differs for each parent scaffold, the results demonstrated that *EnCore* can readily identify screening compounds containing structurally related molecular scaffolds to enrich SAR information from HTS output. The ability of *EnCore* application in enriching SAR information was further demonstrated in the case study using literature COX-2 inhibitors, where molecules represented by structurally related scaffolds could be readily compared to identify favorable and to deprioritize unfavorable scaffolds for COX-2 inhibition. While objective scaffold definition is certainly structurally intuitive and easily interpretable, a major drawback of its application in analyzing HTS output is that it may introduce a sizable population of structural singletons that are considered unfavorable for hit confirmation and expansion.[13] However, when applied in conjunction with the molecular scaffold enumeration approach, we demonstrate that SAR information can be readily enriched. This approach will also be useful when selecting a subset of structurally related molecules as prioritized representatives of a large screening collection, and when designing compound libraries where structurally related compounds can be incorporated to enhance intrinsic SAR within a screening collection.

EnCore is complementary to literature enumeration tools. In comparison to the literature regioisomer enumeration tool HREMS,[17] *EnCore* allows changes to the physicochemical properties of the molecular scaffold such as topological polar surface area when introducing heteroatom changes, which can be beneficial in molecular scaffold design and optimization. The method here is not limited to aromatic heterocycles, as opposed to the literature MORPH algorithm,[15] hence capturing additional design ideas that introduce changes to aliphatic ring systems and structural linkers in molecular scaffolds.

■ CONCLUSIONS

Establishing SAR in hit identification during early stage drug discovery is important in accelerating hit confirmation and expansion. While the application of objective scaffold definitions in analyzing HTS output can assist in achieving this goal, its stringent definitions may result in singleton scaffolds being deprioritized. To enhance the application of objective scaffold definitions in establishing SAR, a systematic molecular scaffold enumeration approach *EnCore* has been developed to enrich SAR information.

Using the top Level 1 scaffolds in the literature medicinal chemistry compound series retrieved from the *Journal of Medicinal Chemistry*, *EnCore* was able to match additional molecular scaffolds explored within the same compound series. This demonstrates that *EnCore* could mimic the scaffold exploration conducted when establishing SAR in a drug discovery program.

In addition, the enumerated scaffold clusters constitute a comprehensive collection of scaffolds for scaffold morphing and hopping when designing new molecules. Since *EnCore* enumerates all possible mutations over generations, potential bias against less common but structurally feasible molecular scaffolds is minimized, whereas synthetically intractable molecular scaffolds may stimulate the discovery of innovative synthetic methods.

When *EnCore* was applied in conjunction with objective scaffold definitions to the analysis of an HTS library, an enrichment in readily available SAR information was demonstrated. Over 70% of the molecular scaffolds observed an increase in the number of structurally related screening compounds associated with each scaffold after enumeration. In particular, over 60% of the singleton scaffolds with only one representative compound were found to have structurally related compounds after enumeration. This will be particularly useful in expanding design opportunities when encountering structural singletons during hit identification.

EnCore offers additional capabilities to literature enumeration tools including changes to the physicochemical properties of molecular scaffolds and structural modifications to aliphatic rings and linkers. This offers a complementary tool for molecular scaffold exploration in medicinal chemistry to establish and enrich SAR information.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00386.

List of the 43 literature medicinal chemistry compound series and their top Level 1 scaffolds tabulated in Table S1 and the list of 93 publications used in the COX-2 inhibitors case study tabulated in Table S2. (PDF)

EnCore enumeration protocol as a Pipeline Pilot protocol (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Yi.Mok@icr.ac.uk.

ORCID

N. Yi Mok: 0000-0002-2827-3735

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is funded by the Cancer Research U.K. Grant C309/A11566. We thank Prof. Julian Blagg for his helpful comments on the manuscript.

■ REFERENCES

(1) Paricharak, S.; Ijzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data. *ACS Chem. Biol.* **2016**, *11*, 1255−1264.

(2) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(3) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inf.* **2010**, *29*, 366−385.

(4) Ertl, P. Intuitive Ordering of Scaffolds and Scaffold Similarity Searching Using Scaffold Keys. *J. Chem. Inf. Model.* **2014**, *54*, 1617−1622.

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

(6) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581−583.

(7) Agrafiotis, D. K.; Wiener, J. J. M. Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure-Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* **2010**, *53*, 5002−5011.

(8) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3*, 435−444.

(9) Mok, N. Y.; Brenk, R. Mining the ChEMBL Database: An Efficient Chemoinformatics Workflow for Assembling an Ion Channel-Focused Screening Library. *J. Chem. Inf. Model.* **2011**, *51*, 2449−2454.

(10) Mok, N. Y.; Brenk, R.; Brown, N. Increasing the Coverage of Medicinal Chemistry-Relevant Space in Commercial Fragments Screening. *J. Chem. Inf. Model.* **2014**, *54*, 79−85.

(11) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.

(12) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.;

Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399−1409.

(13) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51*, 2174−2185.

(14) Ward, R. A.; Kettle, J. G. Systematic Enumeration of Heteroaromatic Ring Systems as Reagents for Use in Medicinal Chemistry. *J. Med. Chem.* **2011**, *54*, 4670−4677.

(15) Beno, B. R.; Langley, D. R. Morph: A New Tool for Ligand Design. *J. Chem. Inf. Model.* **2010**, *50*, 1159−1164.

(16) Miyao, T.; Kaneko, H.; Funatsu, K. Ring System-Based Chemical Graph Generation for de novo Molecular Design. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 425−446.

(17) Tyagarajan, S.; Lowden, C. T.; Peng, Z. W.; Dykstra, K. D.; Sherer, E. C.; Krska, S. W. Heterocyclic Regioisomer Enumeration (HREMS): A Cheminformatics Design Tool. *J. Chem. Inf. Model.* **2015**, *55*, 1130−1135.

(18) *Pipeline Pilot*, v 9.5.0.831; Biovia, 2015.

(19) Schleyer, P. V. Introduction: Aromaticity. *Chem. Rev.* **2001**, *101*, 1115−1117.

(20) ChEMBL v21. https://www.ebi.ac.uk/chembl/ (accessed March 8, 2016).

(21) DrugBank. http://www.drugbank.ca (accessed February 12, 2016).

(22) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(23) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(24) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(25) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(26) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]-benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.

(27) Brown, N. Bioisosteres and Scaffold Hopping in Medicinal Chemistry. *Mol. Inf.* **2014**, *33*, 458−462.

(28) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic Rings of the Future. *J. Med. Chem.* **2009**, *52*, 2952−2963.