

1 Article

## 2 How Size Matters: Diversity for Fragment Library Design

3 Yun Shi <sup>1,\*</sup> and Mark von Itzstein <sup>1,\*</sup>

4 <sup>1</sup> Institute for Glycomics, Griffith University, Gold Coast Campus, Gold Coast, Queensland 4222, Australia;  
5 y.shi@griffith.edu.au (Y.S.); m.vonitzstein@griffith.edu.au (M.v.I.)

6 \* Correspondence: y.shi@griffith.edu.au (Y.S.); m.vonitzstein@griffith.edu.au (M.v.I.)

7

8 **Abstract:** Fragment-based drug discovery (FBDD) has become a major strategy to  
9 derive novel lead candidates for various therapeutic targets, as it promises efficient  
10 exploration of chemical space by employing fragment-sized (MW < 300)  
11 compounds. One of the first challenges in implementing a FBDD approach is the  
12 design of a fragment library, and more specifically, the choice of its size and  
13 individual members. A diverse set of fragments is required to maximise the  
14 chances of discovering novel hit compounds. However, the exact diversity of a  
15 certain collection of fragments remains underdefined, which hinders direct  
16 comparisons among different selections of fragments. Based on structural  
17 fingerprints, we herein introduced quantitative metrics for the structural diversity  
18 of fragment libraries. Structures of commercially available fragments were  
19 retrieved from the ZINC database, from which libraries with sizes ranging from  
20 100 to 100,000 compounds were selected. The selected libraries were evaluated  
21 and compared quantitatively, resulting in interesting size-diversity relationships.  
22 Our results demonstrated that while library size does matter for its diversity, there  
23 exists an optimal size for structural diversity. It is also suggested that such  
24 quantitative measures can guide the design of diverse fragment libraries under  
25 different circumstances.

26 **Keywords:** diversity; fragment-based drug discovery; library design; library size.

27

### 28 1. Introduction

29 Fragment-based drug discovery (FBDD) has been developed in the past twenty years as an  
30 approach to derive novel lead compounds for various therapeutic targets [1–4]. It features the use  
31 of fragment-sized compounds that mostly comply with the ‘Rule-of-3’ [5] for the identification of hits,  
32 which can be subsequently developed into potent lead compounds. Compared to the more  
33 traditional high-throughput screening that employs drug-like compounds following the ‘Rule-of-5’  
34 [6], the smaller sizes of fragments used in FBDD lead to more efficient sampling of the relevant  
35 chemical space and thus better chances of identifying novel hits [7]. The smaller sizes also result in  
36 higher ligand efficiency [8] and more efficient structural optimization of fragment hits [9]. With  
37 these advantages, FBDD has gained popularity in both academia and industry in recent years [10],  
38 and led to the discovery of three FDA-approved drugs [11–14].

39 The first and foremost step in FBDD is the design of a fragment library, as library compositions  
40 directly influence the outcome of FBDD projects. One of the most frequently discussed topic for its  
41 design is the size of the fragment library, which has a substantial impact on the early stages as it  
42 affects the time and monetary costs in addition to the outcome of FBDD projects. Interestingly, the  
43 majority of respondents in recent polls had up to 2000 compounds in their fragment libraries [15,16].

Besides an optimal library size, consideration is also given to the structural complexity, physicochemical profile, and shape profile of fragments [17]. However, the *diversity* of a fragment library should be the most critical factor, because it affects the sampling efficiency of the relevant chemical space as well as the novelty of potential hit compounds. Of note, better diversity should decrease the screening hit rate, which appears to be high for many FBDD campaigns [4,18]. Hence, the size of a fragment library should be discussed in conjunction with its diversity.

Diversity needs to be characterized by descriptors, which can be classified mostly into three categories. The first are functional (performance) descriptors based on the bioactivities of compounds towards a panel of (functionally dissimilar) biological targets [19]. Although regarded as the most relevant category of diversity descriptors for drug discovery [20,21], acquisition of bioactivity data can be very resource-demanding [22,23]. In addition to a lack of bioactivity data for fragment-sized compounds in the literature, their activities would also be difficult to detect and measure due to their weak affinities [7]. The second are physicochemical (property-based) descriptors, including common physicochemical properties such as molecular weight and hydrophobicity and even electronic properties [24]. The third are structural descriptors, among which molecular fingerprints (structural features) are routinely used to represent chemical structures. The extended-connectivity (radial) fingerprints [25] is effective at retrieving bioactive compounds [26], therefore it was chosen as the descriptor of diversity in our study.

There are currently two major types of quantitative metrics for structural diversity [27]. The first type of metrics assess the similarity (and thus difference) between pairs of chemical structures. The most notable metric of this type is the Jaccard index [28], later referred to as the popular Tanimoto index (similarity) [29]. The second type of metrics calculates the coverage of the relevant chemical space by a library of compounds, and the most straightforward one is a ratio based on richness, defined as the number of unique fingerprints (structural features) [30]. In this work, we propose the adoption of a third type of metrics, i.e. a diversity index that takes into account not only the number of unique structural fingerprints but also their proportional abundances [31–33], for the quantitative measurement of diversity. True diversity, or the effective number of structural features, is a commonly used metric of this type and can be defined by the following equation [33]:

$$D = \frac{1}{\prod_{i=1}^R p_i^{p_i}}, \quad (1)$$

where  $D$  stands for true diversity,  $R$  is richness (the total number of fingerprints), and  $p_i$  represents the proportional abundance of the  $i$ th fingerprint. It can be deduced from equation (1) that, for the same richness, a library with a more even distribution of proportional abundances will have a larger true diversity than a library with a less even distribution. These diversity indexes have been used in ecological studies for decades, yet they have not been applied to the measurement of diversity of fragment libraries to date. Although there are other plot-based methods to illustrate diversity in more visually appealing ways, such as principle component analysis [34] and principal moments of inertia [35,36], these three quantitative metrics, i.e. Tanimoto similarity, number of fingerprints, and true diversity, are more suited for direct comparison of libraries with different sizes.

To provide insights into how the library size affects the structural diversity, we herein compare fragment libraries of different sizes, selected from commercially available fragments, and demonstrate interesting size-diversity relationships. Such relationships indicated the presence of an optimal library size for structural diversity. We also extend this investigation to a more restrictive scenario, in which only fluorinated fragments are considered and consequentially similar size-diversity relationships were observed. Certain cost-effective sizes that capture significant proportions of the overall diversity available with very small portions of available fragments are also proposed. Our results demonstrated that these quantitative metrics could assist in the design of fragment libraries under various circumstances.

## 2. Results

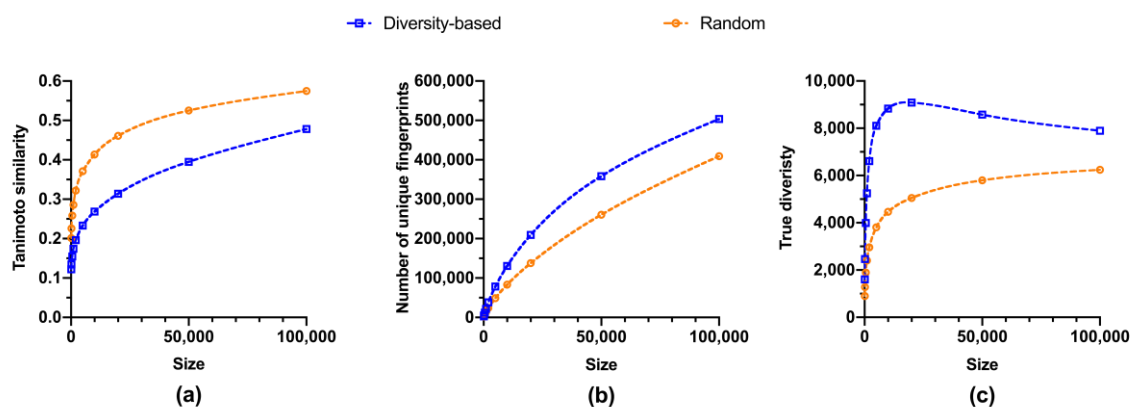
### 2.1. Library selection

92 To generate libraries for comparison, both diversity-based selections and random selections  
93 were performed from 227,787 commercially available fragments that had undergone filtering by the  
94 'Rule-of-3' criteria [5]. Libraries with sizes of 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000, and  
95 100000 were selected. Both diversity-based selections and random selections were performed, with  
96 the latter in triplicate. To demonstrate that our approach can be applied to different circumstances,  
97 selections were also performed on a fluorinated subset of the 227,787 commercially available  
98 fragments, consisting of 47,708 fragments that has 1-3 fluorine atoms. Such restriction on the  
99 number of fluorine atoms captured the majority of fluorinated fragments, which are commonly used  
100 for FBDD projects employing  $^{19}\text{F}$  NMR as the screening method [37,38]. Fluorinated libraries with  
101 sizes of 100, 200, 500, 1000, 2000, 5000, 10000, and 20000 were selected in similar fashions.

## 102 2.2. Size-diversity relationship of regular fragment libraries

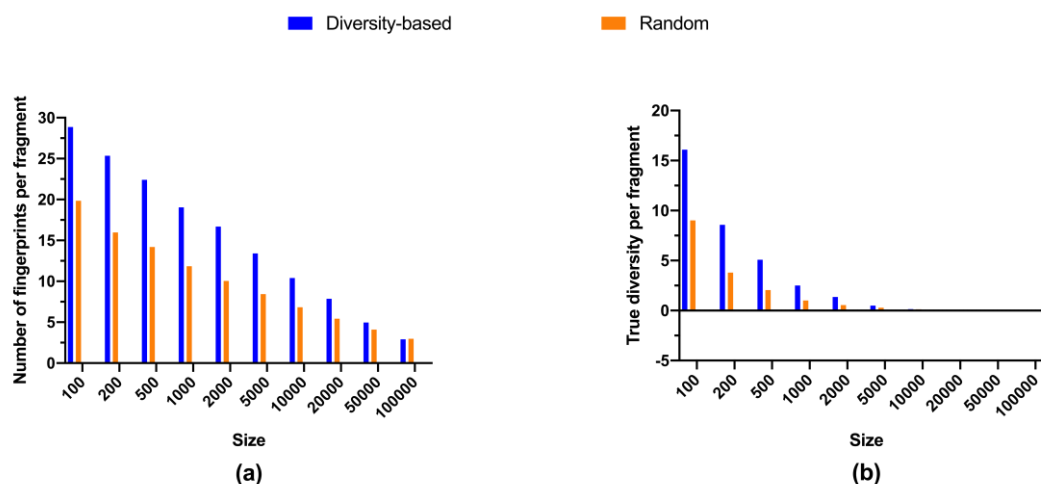
103 To understand the relationship between the size of fragment libraries and their structural  
104 diversity, quantitative metrics were calculated for selected libraries (Figure 1). As expected,  
105 fragments became more similar to each other as the library size increased, and the diversity-based  
106 selection did lead to more dissimilar fragments than random selections (Figure 1a). Richness of  
107 fragment library also rose with its size, with diversity-based selections outperforming random  
108 selections (Figure 1b). However, marginal richness, i.e. the additional number of unique  
109 fingerprints per additional fragment, was declining while library size grew (Figure 2a). For  
110 diversity-based selections, the average efficiency of adding unique fingerprints from 2,000 fragments  
111 to 5,000 fragments, 13.4 fingerprints per compound, was less than half of that from nothing to 100  
112 fragments, 28.9 fingerprints per compound. Similar trends were observed for randomly selected  
113 libraries, although the gap between diversity-based and random selections became smaller when  
114 library sizes grew excessively large, i.e. beyond 5,000 compounds. Thus, it is more efficient to have  
115 relatively small library for richness and we estimated the number of fragments required to  
116 accomplish two arbitrary degrees of coverage, 5% and 10%, respectively (Table 1).

117



118 **Figure 1.** Structural diversity vs size of fragment libraries, with the former measured by: (a) Average  
119 of the similarity of each compound to its closest neighbour; (b) total number of unique fingerprints  
120 (richness); (c) true diversity calculated by equation (1). Dash curves are generated from cubic spline  
121 fitting. Metrics for random selections are average values of triplicates.

122



123 **Figure 2.** Efficiency in adding diversity: (a) average number of unique fingerprints (richness) per  
 124 compound; (b) average value of true diversity per compound. Metrics for random selections are  
 125 average values of triplicates.

126

127 **Table 1.** Library sizes (diversity-based selection) required to achieve certain values of structural  
 128 diversity.

Structural diversity (value)	Minimum size (ratio of total 227,787 fragments) <sup>1</sup>
5% total richness <sup>2</sup> (33,834)	1,715 (0.75%)
10% total richness <sup>2</sup> (67,669)	4,103 (1.80%)
Overall true diversity (6,662.4)	2,052 (0.90%)
Maximum true diversity <sup>1</sup> (9097.6)	17,666 (7.76%)

129 <sup>1</sup> Values are estimated by cubic spline fitting with 99,901 segments.

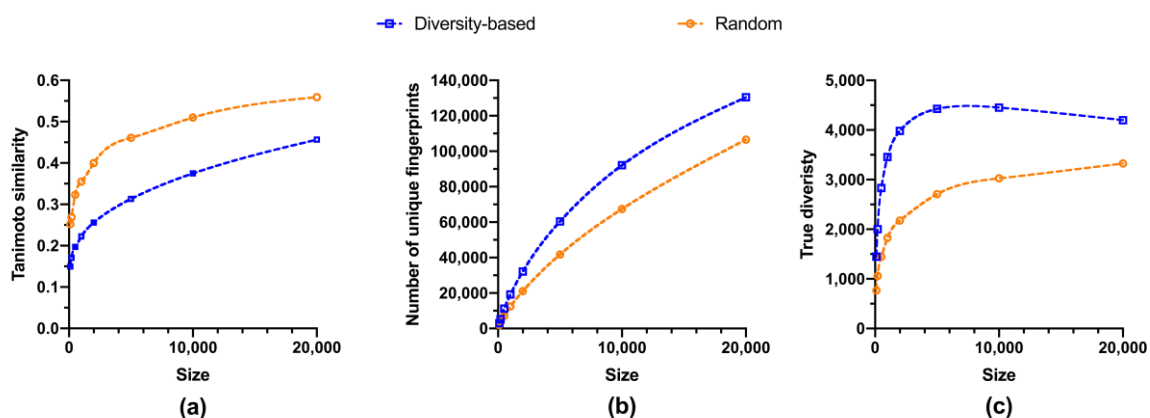
130 <sup>2</sup> Total richness (number of unique fingerprints) is 676,686.

131 Surprisingly, values of true diversity exhibited different trends between diversity-based  
 132 selections and random selections (Figure 1c). While the latter showed a constantly rising movement,  
 133 the former reached a maximum at about 18,000 fragments, representing less than 8% of the overall  
 134 available fragments (Table 1), before starting to decline (Figure 1c). In addition, marginal true  
 135 diversity experienced a more drastic decline in comparison with the marginal richness (Figure 2).  
 136 For diversity-based selections, the average efficiency of adding true diversity from 2,000 fragments  
 137 to 5,000 fragments, 1.4 per compound, was an order of magnitude less than that from nothing to 100  
 138 fragments, 16.1 per compound. Consistent with the decline of true diversity after the library size  
 139 from diversity-based selections reached about 18,000, the marginal true diversity became negative  
 140 after 20,000 compounds (Figure 2b). More strikingly, only approximately 2,000 fragments, i.e. less  
 141 than 1%, are required to attain the same level of true diversity as all the 227,787 fragments available  
 142 for selection (Table 1).

### 143 2.3. Size-diversity relationship of fluorinated fragment libraries

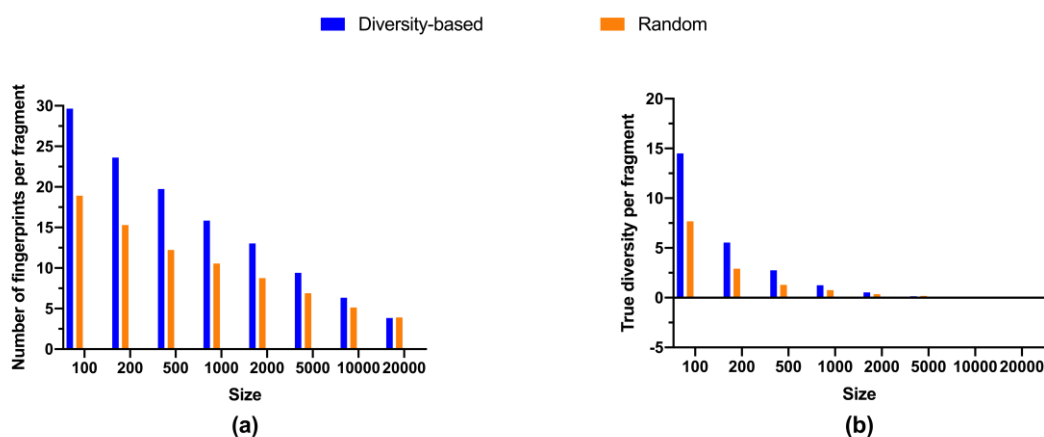
144 Libraries selected from fluorinated fragments presented similar size-diversity relationships as  
 145 those from regular fragments (Figures 3 and 4, Table 2). Both similarity to the closest neighbor and  
 146 richness illustrated growing trends (Figures 3a and 3b), whereas the true diversity for libraries subject  
 147 to diversity-based selection also reached a maximum at about 7,500 fragments (Figure 3c and Table  
 148 2). Analogously, both marginal richness and marginal true diversity diminished with increasing  
 149 library size, while the gap between diversity-based and random selections in efficiency became  
 150 smaller for larger library sizes, i.e. beyond 500 compounds (Figure 4). Nevertheless, it required

151 relatively more fluorinated fragments to achieve the same level of diversity than that for regular  
 152 fragments. About 3.4% of total fluorinated fragments were needed to attain 10% coverage (Table 2),  
 153 much higher than that for regular fragments, about 1.8%. Additionally, it took close to 15.7% of total  
 154 fluorinated fragments to reach maximum true diversity, while for regular fragments only about 7.8%  
 155 were required. Further, 2.5% of total fluorinated fragments were required to achieve the same level  
 156 of true diversity as all the 47,708 fluorinated fragments, whereas less than 1% of regular fragments  
 157 were required. These observations can be explained by the constant presence of fluorine atoms, and  
 158 thus fluorine-containing fingerprints, in all fluorinated compounds. Inevitably, there would be a  
 159 larger overlap of fluorine-associated fingerprints among fluorinated compounds, rendering the  
 160 distribution of proportional abundances for fingerprints less even and thereby a smaller value of true  
 161 diversity calculated by equation (1). Such a phenomenon can also be expected for other restrictive  
 162 circumstances demanding the presence of certain functional groups and/or pharmacophores.  
 163



164 **Figure 3.** Structural diversity vs size of fluorinated fragment libraries, with the former measured by:  
 165 (a) Average of the similarity of each compound to its closest neighbour; (b) total number of unique  
 166 fingerprints (richness); (c) true diversity calculated by equation (1). Dash curves are generated from  
 167 cubic spline fitting. Metrics for random selections are average values of triplicates.

168



169 **Figure 4.** Efficiency in adding diversity: (a) average number of unique fingerprints (richness) per  
 170 fluorinated compound; (b) average value of true diversity per fluorinated compound. Metrics for  
 171 random selections are average values of triplicates.

172

173



174 **Table 2.** Fluorinated library sizes (diversity-based selection) required to achieve certain values of  
 175 structural diversity.

Structural diversity (value)	Minimum size (ratio of total 47,708 fluorinated fragments) <sup>1</sup>
5% total richness <sup>2</sup> (8,992)	675 (1.41%)
10% total richness <sup>2</sup> (17,983)	1,616 (3.39%)
Overall true diversity (3,621.9)	1,203 (2.52%)
Maximum true diversity <sup>1</sup> (4485.5)	7,483 (15.69%)

176 <sup>1</sup> Values are estimated by cubic spline fitting with 19,901 segments.

177 <sup>2</sup> Total richness (number of unique fingerprints) is 179,833.

178

### 179 3. Discussion

180 The exact size-diversity relationships for fragment libraries are affected by several factors,  
 181 including the fragments available for selection, the selection method, and the diversity metric.  
 182 Using fluorinated fragments as an example, we have shown that similar size-diversity relationships  
 183 are observed for this subset of available fragments. Thus, we speculate that different but similar  
 184 size-diversity relationships could be observed for a different set of fragments available for selection.  
 185 This could be either more restrictive, such as a set of fragments from a certain vendor, or more  
 186 inclusive, such as a virtual set of all theoretically possible fragments [39]. Moreover, we expect that  
 187 a different selection method, such as a clustering method [40], would offer somewhat different  
 188 results. Yet it should be noted that clustering methods are much less efficient than the directed  
 189 sphere exclusion method used in this study [41], which features good computational performance on  
 190 large data sets and enabled our calculations to be carried out on a desktop computer. Furthermore,  
 191 our results illustrated that different diversity metrics could indeed show very different size-diversity  
 192 relationships. While both similarity and richness increased with the size of fragment library, the rate  
 193 of increase experienced a more significant decline in the former than in the latter, resulting in larger  
 194 curvatures of the fitted lines for similarity. In contrast, the true diversity of libraries from diversity-  
 195 based selections started to decrease after a certain size, highlighting the uneven distribution of  
 196 structural fingerprints as the library size grew excessively large.

197 Not unexpectedly, our results showed that the marginal diversity diminishes while the library  
 198 size increases, the extent and significance of which depends on the choice of diversity metrics. This  
 199 indicates that it is unnecessary and possibly counterproductive to play numbers game and build  
 200 excessively large libraries, and that cost-effective sizes of fragment library exist for structural  
 201 diversity. For regular fragments selected from commercially readily available compounds, we  
 202 propose a library size of ~2,000, corresponding to 0.9 % of total available fragments in this study.  
 203 This size covers more than 5% of richness, approximates the true diversity of all available fragments,  
 204 and (perhaps coincidentally) matches the most popular fragment library size [15,16]. For the  
 205 fluorinated subset, a library size of ~1,200 achieves similar coverage of richness and true diversity.  
 206 However, better selection methods may even reduce these proposed numbers.

207 In addition to structural diversity, considerations should also be given to practical factors such  
 208 as experimental solubility, (absence of) aggregation, and stability for fragment library design [17].  
 209 These factors are essential for the success of FBDD campaigns, yet they are difficult to predict without  
 210 experimental data. Hence, it would be more pragmatic to slightly increase the library size in the  
 211 initial *in silico* design and perform necessary quality checks after procurement of fragments.

212 In summary, we have introduced quantitative metrics to evaluate the structural diversity of  
 213 fragment libraries, investigated their size-diversity relationships, and demonstrated the existence of  
 214 an optimal library size for structural diversity depending on specific situations. Based on our  
 215 results, we propose the use of relatively small library sizes and the application of these quantitative  
 216 measures to the design of diverse fragment libraries under various circumstances.

#### 217 4. Materials and Methods

218 Structures of commercially-available, fragment-sized compounds were retrieved from the ZINC  
219 15 database [42] (<https://zinc15.docking.org/tranches/home/>) in SMILES format on 2 Jan 2019. A  
220 subset was chosen with the following criteria: Anodyne for Reactivity; In-Stock for Purchasability;  
221 up to 300 Daltons for Molecular Weight; up to 3 for LogP. These criteria resulted in 1,413,973  
222 compounds. The Canvas program (Schrödinger, LLC) was used for subsequent calculations.  
223 Physicochemical properties were calculated by canvasMolDescriptors and compounds violating an  
224 adapted version of the 'Rule-of-3' [5], i.e.  $100 \leq MW \leq 300$ ,  $\log P \leq 3$ , number of rings  $\leq 3$ , number of  
225 hydrogen bond donors (HBD)  $\leq 3$ , number of hydrogen bond acceptors (HBA)  $\leq 3$ , number of  
226 rotatable bonds (RB)  $\leq 3$ , and polar surface area  $\leq 60 \text{ \AA}^2$  were removed. HBD, HBA, and RB are  
227 custom defined according to a previous work [43]. Any compound with reactive groups was filtered  
228 by the ligfilter functionality and duplicate structures were eliminated by the uniqueness  
229 functionality. Finally, 227,787 compounds were left for selection of fragment libraries.

230 Radial fingerprints [25] were generated by canvasFPGen, with 64-bit precision ( $2^{64}$ ) to avoid  
231 fingerprint collisions, Daylight invariant atom types [44], and three radial iterations. Based on the  
232 these fingerprints, diversity-based selections were performed with canvasDBCS, using directed  
233 sphere exclusion method [41] and Tanimoto similarity [29]. An exclusion sphere size of 0.4 was used  
234 to select libraries with a maximum size of 100,000 compounds. In parallel, random selections of  
235 fragment libraries as control were carried out in triplicate by the UNIX command *shuf*. To quantify  
236 the diversity of selected libraries, three different metrics were calculated as follows: maximum  
237 Tanimoto similarity [29] was computed by canvasFPHist; total number of unique fingerprints [30]  
238 was counted by canvasFPBinary2CSV; and true diversity [33] was determined by the UNIX  
239 command *awk* using equation (1).

240 For fluorinated fragments, the ligfilter functionality was used to filter the 227,787 compounds  
241 with a criterion of  $1 \leq \text{number of fluorine atoms} \leq 3$ , and the resulting 47,708 fragments were subject  
242 to analogous calculations and selections with a maximum library size of 20,000 compounds.

243 Prism 8 (GraphPad Software, Inc.) was employed to generate plots of the aforementioned three  
244 metrics against the size of selected libraries, and the cubic spline function was used to fit spine curves.

245 **Supplementary Materials:** The following are available online at [www.mdpi.com/xxx/s1](http://www.mdpi.com/xxx/s1), File S1: List of SMILES  
246 structures of all 227,787 fragments used in this study.

247 **Author Contributions:** Conceptualization, Y.S. and M.v.I.; methodology, Y.S.; software, Y.S.; formal analysis,  
248 Y.S.; investigation, Y.S.; data curation, Y.S.; writing—original draft preparation, Y.S.; writing—review and  
249 editing, Y.S. and M.v.I.; visualization, Y.S.; supervision, M.v.I.; funding acquisition, Y.S. and M.v.I

250 **Funding:** This research was funded by the National Health and Medical Research Council, grant number  
251 1071659 (M.v.I.). Y.S. was a recipient of Griffith University Postdoctoral Fellowship.

252 **Acknowledgements:** We gratefully acknowledge the support of the Griffith University eResearch Services Team  
253 and the use of the High Performance Computing Cluster "Gowonda" to complete this research.

254 **Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the  
255 study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to  
256 publish the results.

257

258 **References**

- 259 1. Shuker, S.B.; Hajduk, P.J.; Meadows, R.P.; Fesik, S.W. Discovering High-Affinity Ligands for Proteins:  
260 SAR by NMR. *Science* **1996**, *274*, 1531–1534.
- 261 2. Erlanson, D.A.; McDowell, R.S.; O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **2004**, *47*, 3463–  
262 3482.
- 263 3. Rees, D.C.; Congreve, M.; Murray, C.W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discov.*  
264 **2004**, *3*, 660–672.
- 265 4. Erlanson, D.A.; Fesik, S.W.; Hubbard, R.E.; Jahnke, W.; Jhoti, H. Twenty years on: the impact of fragments  
266 on drug discovery. *Nat. Rev. Drug Discov.* **2016**, *15*, 605–619.
- 267 5. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug*  
268 *Discov. Today* **2003**, *8*, 876–877.
- 269 6. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to  
270 estimate solubility and permeability in drug discovery and development settings1PII of original article:  
271 S0169-409X(96)00423-1. The article was originally published in *Advanced Drug Delivery Reviews* 23  
272 (1997) 3–25.1. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.
- 273 7. Hann, M.M.; Leach, A.R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding  
274 Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.
- 275 8. Hopkins, A.L.; Groom, C.R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov.*  
276 *Today* **2004**, *9*, 430–431.
- 277 9. Murray, C.W.; Rees, D.C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187–192.
- 278 10. Romasanta, A.K.S.; van der Sijde, P.; Hellsten, I.; Hubbard, R.E.; Keseru, G.M.; van Muijlwijk-Koezen, J.;  
279 de Esch, I.J.P. When fragments link: a bibliometric perspective on the development of fragment-based  
280 drug discovery. *Drug Discov. Today* **2018**, *23*, 1596–1609.
- 281 11. Tsai, J.; Lee, J.T.; Wang, W.; Zhang, J.; Cho, H.; Mamo, S.; Bremer, R.; Gillette, S.; Kong, J.; Haass, N.K.; et  
282 al. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. *Proc.*  
283 *Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3041–3046.
- 284 12. Bollag, G.; Hirth, P.; Tsai, J.; Zhang, J.; Ibrahim, P.N.; Cho, H.; Spevak, W.; Zhang, C.; Zhang, Y.; Habets,  
285 G.; et al. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma.  
286 *Nature* **2010**, *467*, 596–599.
- 287 13. Souers, A.J.; Levenson, J.D.; Boghaert, E.R.; Ackler, S.L.; Catron, N.D.; Chen, J.; Dayton, B.D.; Ding, H.;  
288 Enschede, S.H.; Fairbrother, W.J.; et al. ABT-199, a potent and selective BCL-2 inhibitor, achieves  
289 antitumor activity while sparing platelets. *Nat. Med.* **2013**, *19*, 202–208.
- 290 14. Perera, T.P.S.; Jovcheva, E.; Mevellec, L.; Vialard, J.; Lange, D.D.; Verhulst, T.; Paulussen, C.; Ven, K.V.D.;  
291 King, P.; Freyne, E.; et al. Discovery and Pharmacological Characterization of JNJ-42756493 (Erdafitinib),  
292 a Functionally Selective Small-Molecule FGFR Family Inhibitor. *Mol. Cancer Ther.* **2017**, *16*, 1010–1020.
- 293 15. Erlanson, D. Practical Fragments: Poll results: affiliation, fragment-finding methods, and library size.  
294 *Practical Fragments* 2014.
- 295 16. Erlanson, D. Practical Fragments: Poll results: the modern fragment library. *Practical Fragments* 2018.
- 296 17. Keserű, G.M.; Erlanson, D.A.; Ferenczy, G.G.; Hann, M.M.; Murray, C.W.; Pickett, S.D. Design Principles  
297 for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery  
298 (FBDD) Programs for Use in Academia. *J. Med. Chem.* **2016**.
- 299 18. Chen, I.-J.; Hubbard, R.E. Lessons for fragment library design: analysis of output from multiple screening  
300 campaigns. *J. Comput. Aided Mol. Des.* **2009**, *23*, 603–620.



- 301 19. Dixon, S.L.; Villar, H.O. Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting. *J.*  
302 *Chem. Inf. Comput. Sci.* **1998**, *38*, 1192–1203.
- 303 20. Roth, H.-J. There is no such thing as ‘diversity’! *Current Opinion in Chemical Biology* **2005**, *9*, 293–295.
- 304 21. Koutsoukas, A.; Paricharak, S.; Galloway, W.R.J.D.; Spring, D.R.; IJzerman, A.P.; Glen, R.C.; Marcus, D.;  
305 Bender, A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking  
306 of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2014**, *54*, 230–242.
- 307 22. Kauvar, L.M.; Higgins, D.L.; Villar, H.O.; Sportsman, J.R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K.E.;  
308 Dilley, H.; Rocke, D.M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**,  
309 *2*, 107–118.
- 310 23. Wawer, M.J.; Li, K.; Gustafsdottir, S.M.; Ljosa, V.; Bodycombe, N.E.; Marton, M.A.; Sokolnicki, K.L.; Bray,  
311 M.-A.; Kemp, M.M.; Winchester, E.; et al. Toward performance-diverse small-molecule libraries for cell-  
312 based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U.S.A.*  
313 **2014**, *111*, 10911–10916.
- 314 24. Carbó, R.; Leyda, L.; Arnau, M. How similar is a molecule to another? An electron density measure of  
315 similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
- 316 25. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- 317 26. Duan, J.; Dixon, S.L.; Lowrie, J.F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into  
318 database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **2010**, *29*, 157–170.
- 319 27. Gillet, V.J. Diversity selection algorithms. *Wiley Interdisciplinary Reviews: Computational Molecular Science*  
320 **2011**, *1*, 580–589.
- 321 28. Jaccard, P. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines.  
322 *Bulletin de la Societe Vaudoise des Sciences Naturelles* **1901**, *37*, 241–272.
- 323 29. Tanimoto, T.T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business  
324 Machines Corporation, 1958;
- 325 30. Martin, E.J.; Blaney, J.M.; Siani, M.A.; Spellmeyer, D.C.; Wong, A.K.; Moos, W.H. Measuring Diversity:  
326 Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431–1436.
- 327 31. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
- 328 32. Simpson, E.H. Measurement of Diversity. *Nature* **1949**, *163*, 688.
- 329 33. Hill, M.O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **1973**, *54*, 427–432.
- 330 34. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**,  
331 *24*, 417–441.
- 332 35. Sylvester, J.J. XIX. A demonstration of the theorem that every homogeneous quadratic polynomial is  
333 reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. *The*  
334 *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1852**, *4*, 138–142.
- 335 36. Sauer, W.H.B.; Schwarz, M.K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for  
336 Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987–1003.
- 337 37. Jordan, J.B.; Poppe, L.; Xia, X.; Cheng, A.C.; Sun, Y.; Michelsen, K.; Eastwood, H.; Schnier, P.D.; Nixey, T.;  
338 Zhong, W. Fragment Based Drug Discovery: Practical Implementation Based on <sup>19</sup>F NMR Spectroscopy.  
339 *J. Med. Chem.* **2012**, *55*, 678–687.
- 340 38. Vulpetti, A.; Dalvit, C. Design and Generation of Highly Diverse Fluorinated Fragment Libraries and their  
341 Efficient Screening with Improved <sup>19</sup>F NMR Methodology. *ChemMedChem* **2013**, *8*, 2057–2069.
- 342 39. Visini, R.; Awale, M.; Reymond, J.-L. Fragment Database FDB-17. *J. Chem. Inf. Model.* **2017**, *57*, 700–709.

- 343 40. Downs, G.M.; Barnard, J.M. Clustering Methods and Their Uses in Computational Chemistry. In *Reviews*  
344 *in Computational Chemistry*; John Wiley & Sons, Ltd, 2003; pp. 1–40 ISBN 978-0-471-43351-4.
- 345 41. Gobbi, A.; Lee, M.-L. DISE: Directed Sphere Exclusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 317–323.
- 346 42. Sterling, T.; Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- 347 43. Gillet, V.J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural  
348 Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165–179.
- 349 44. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES  
350 notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- 351