

Expanding Bioactive Fragment Space with the Generated Database GDB-13s

Ye Buehler and Jean-Louis Reymond*

*Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern,
Freiestrasse 3, 3012 Bern, Switzerland*

Abstract

Identifying innovative drug-like small molecules is critically important in medicinal chemistry to address new targets and overcome limitations of classical molecular series. By deconstructing molecules into ring fragments (RFs, ring and ring-adjacent atoms) and acyclic fragments (AFs, only acyclic atoms), we find that molecules in public databases of drug-like compounds (ZINC, PubChem) and natural products (COCONUT) mostly consist of RFs and AFs up to 13 atoms, and that many RFs and AFs are enriched in bioactive compared to inactive molecules in ChEMBL. We then search the 28,246,012 RFs and 2,640,023 AFs in the generated database GDB-13s (99,394,177 molecules up to 13 atoms following simple functional group and ring strain criteria) for subsets resembling ChEMBL bioactive RFs and AFs. Many of these RFs and AFs are structurally simple, have favorable synthetic accessibility scores, and represent opportunities for synthetic chemistry to contribute to drug innovation in the context of fragment-based drug discovery.

Introduction

Medicinal chemistry becomes an increasingly retrospective activity as known drug-like molecules and their biological activity accumulate in public databases such as PubChem¹ and ChEMBL.² Nevertheless, discovering novel molecules remains critically important to address new target types and overcome limitations of classical molecular series in terms of physico-chemical properties, selectivity, toxicity and metabolism, as well as to secure intellectual property and the possibility of commercial development.³⁻⁶

We have shown previously at the example of the generated databases (GDBs) that systematic enumeration of molecules from mathematical graphs using simple rules of chemical stability and synthetic feasibility opens up an extremely large chemical space,⁷⁻¹⁰ which is potentially more diverse than that accessible by combining available building blocks through established reactions or by sampling generative models trained with known molecules.¹¹⁻¹⁵ For instance, the GDBs feature molecules with many unprecedented molecular frameworks (graphs including rings and linker bonds).^{16,17} However, identifying GDB-molecules that are both significantly novel and relevant for medicinal chemistry is challenging.

Here we propose an approach to this problem taking accumulated knowledge of bioactive compounds into account through an analysis of fragments. First, we assess the known chemical space by deconstructing molecules in the public databases ZINC (screening compounds),¹⁸ PubChem (published molecules),¹ and COCONUT (natural products and NP-like molecules)¹⁹ into ring fragments (RFs), obtained by removing all atoms not directly connected to a ring, and acyclic fragments (AFs), obtained by removing all ring atoms (**Figure 1**). This fragmentation is inspired by computational retrosynthetic analysis such as RECAP,²⁰ rdScaffoldNetwork,²¹ DAIM,²² BRICS,²³ CCQ,²⁴ eMolFrag,²⁵ molBLOCKS,²⁶ or

Fragmenter.²⁷ In the present context, our deconstruction into RFs and AFs is designed to simplify molecules and focus on structural types, for instance by converting all substituents of a ring to a single atom for RFs and isolating acyclic groups from their rings in AFs. Interestingly, most molecules in ZINC, PubChem and COCONUT break down into RFs and AFs of 13 atoms or less.

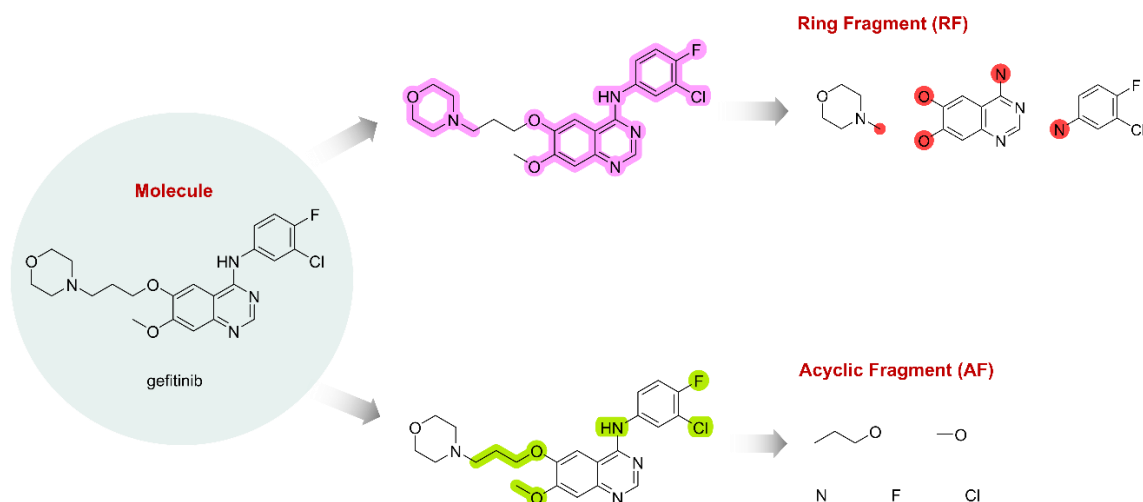


Figure 1. Fragmentation of molecules into ring fragments (RFs) and acyclic fragments (AFs). General principle at the example of drug gefitinib. For RFs acyclic atoms are labeled in red.

In the second part of our approach, we identify RFs and AFs which are strongly enriched in bioactive compared to inactive molecules in ChEMBL (target annotated compounds),² and search for analogs of these fragments in RFs and AFs derived from the generated database GDB-13s. This database lists 99,394,177 small molecules up to 13 atoms exhaustively enumerated from mathematical graphs following simple rules of chemical stability and synthetic feasibility, and contains many unprecedented molecular frameworks (graphs including rings and linker bonds).^{16,17} Many of the bioactive-like RFs and AFs identified in GDB-13s are structurally relatively simple, have favourable synthetic accessibility score (SAscore),²⁸ and therefore represent opportunities for synthetic chemistry to contribute to drug innovation in the context of fragment-based drug discovery.^{29,30}

Results and Discussion

Fragments analysis of known molecules and GDB-13s

To assess the known chemical space, we extracted RFs and AFs from 885,905,524 molecules in the ZINC database,¹⁸ 100,852,694 molecules up to 50 non-hydrogen atoms in PubChem,¹ and 401,624 natural products (NPs) and NP-like molecules in COCONUT.¹⁹ We also extracted RFs and AFs from the 99,394,177 molecules in GDB-13s,¹⁷ to be used as source of novelty later in the study.

In all these databases, the number of molecules per RF and AF followed a typical power law distribution, with few RFs and AFs occurring in many molecules and a relatively large number of RFs and AFs occurring only once, referred to as singletons (**Figure 2a-b, Table 1**). The most frequent RFs and AFs in each database were rather small, featuring mono- and disubstituted benzene rings and azacycles for RFs in known molecules, cyclopropanes for RFs in GDB-13s, and single atom groups for AFs in all databases (**Figure S1 and S2**). In fact, although the size distribution of molecules, RFs and AFs in known molecules extended far beyond 13 atoms (**Figure 2c-f**), RFs and AFs up to 13 atoms were sufficient to cover most molecules, except for natural products in COCONUT which featured many molecules with RFs larger than 13 atoms (**Table 1**, entries 2-4). While fragments shared by the four databases were often structurally simple, those occurring in only one of the four databases analyzed (exclusive fragments, eRF and eAF) were generally more complex, as exemplified by the most frequent cases (**Figure S3 and S4**).

Within the space covered by RFs and AFs up to 13 atoms, GDB-13s largely outnumbered the known molecules in terms of RFs, resulting in a high percentage of exclusive RFs (99.2% eRF_{≤13}). Most AF_{≤13} in GDB-13s were also exclusive (92.7% eAF_{≤13}), although the absolute number of AFs in GDB-13s was comparable to AFs in

ZINC and smaller than AFs in PubChem. In fact, PubChem, ZINC and COCONUT also contained many exclusive $eRF \leq 13$ and $eAF \leq 13$, reflecting that the enumeration of GDB-13s excluded strained rings (*e.g.* cubane, prismane) and certain functional groups (*e.g.* non-aromatic olefins, isocyanide, anhydride, acetal, hemi-acetal, aminal, hemi-aminal, enol ether, peroxide, nitro, azide, thiol, thioether), and only considered C, N, O, S, and Cl as elements. Nevertheless, the above analysis showed that GDB-13s contained a very large number of both eRFs and eAFs and could therefore serve as a source of novel RFs and AFs to expand the space of known molecules.

Table 1. Molecule and fragment count in different databases.

No.		ZINC		PubChem		COCONUT		GDB-13s	
1	Cpds	885,905,524		100,852,694		401,624		99,394,177	
2	Cpds from $RF \leq 13^a$	743,430,899	83.9%	68,876,892	68.3%	132,432	33.0%	99,394,177	100%
3	Cpds from $AF \leq 13^b$	818,548,834	92.4%	94,526,506	93.7%	357,976	89.1%	99,394,177	100%
4	Cpds from $ARF \leq 13^c$	678,518,591	76.6%	62,998,179	62.5%	98,990	24.6%	99,394,177	100%
5	RF	2,838,201		9,037,484		115,381		28,246,012	
6	eRF ^d	2,165,176	76.3%	8,139,719	90.1%	45,448	39.4%	28,011,035	99.2%
7	RF-Singleton ^e	1,115,630	39.3%	6,111,177	67.6%	78,920	68.4%	23,842,697	84.4%
8	$RF \leq 13^f$	158,576	5.6%	1,746,923	19.3%	17,211	14.9%	28,246,012	100%
9	eRF $\leq 13^g$	17,578	0.6%	1,333,179	14.8%	1,863	1.6%	28,011,035	99.2%
10	$RF \leq 13$ -Singleton ^h	58,749	2.1%	1,048,461	11.6%	10,244	8.9%	23,842,697	84.4%
11	AF	2,756,691		5,466,187		45,816		2,640,023	
12	eAF ^d	2,319,553	84.1%	4,722,488	86.4%	18,608	40.6%	2,447,627	92.7%
13	AF-Singleton ^e	688,408	25.0%	4,256,810	77.9%	34,243	74.7%	2,576,927	97.6%
14	$AF \leq 13^f$	338,990	12.3%	2,225,960	40.7%	17,216	37.6%	2,640,023	100%
15	eAF $\leq 13^g$	145,340	5.3%	1,805,294	33.0%	2,131	4.7%	2,447,627	92.7%
16	$AF \leq 13$ -Singleton ^h	52,606	1.9%	1,535,039	28.1%	9,950	21.7%	2,576,927	97.6%

a) Cpds from $RF \leq 13$ = molecules covered by RF up to HAC = 13. b) Cpds from $AF \leq 13$ = molecules covered by AF up to HAC = 13. c) Cpds from $ARF \leq 13$ = molecules covered by both RF and AF up to HAC = 13. d) eRF/eAF = exclusive RF/AF, absent from the other three databases. e) RF/AF-Singleton = RF/AF with only a single molecule example. f) $RF \leq 13/AF \leq 13$ = RF/AF up to HAC = 13. g) eRF $\leq 13/eAF \leq 13$ = exclusive RF13/AF13, absent from the other three databases. h) $RF \leq 13$ -Singleton / $ARF \leq 13$ -Singleton = $RF \leq 13/AF \leq 13$ with only a single molecule example. RF and AF subcategories are calculated relative to total RF and AF, respectively.

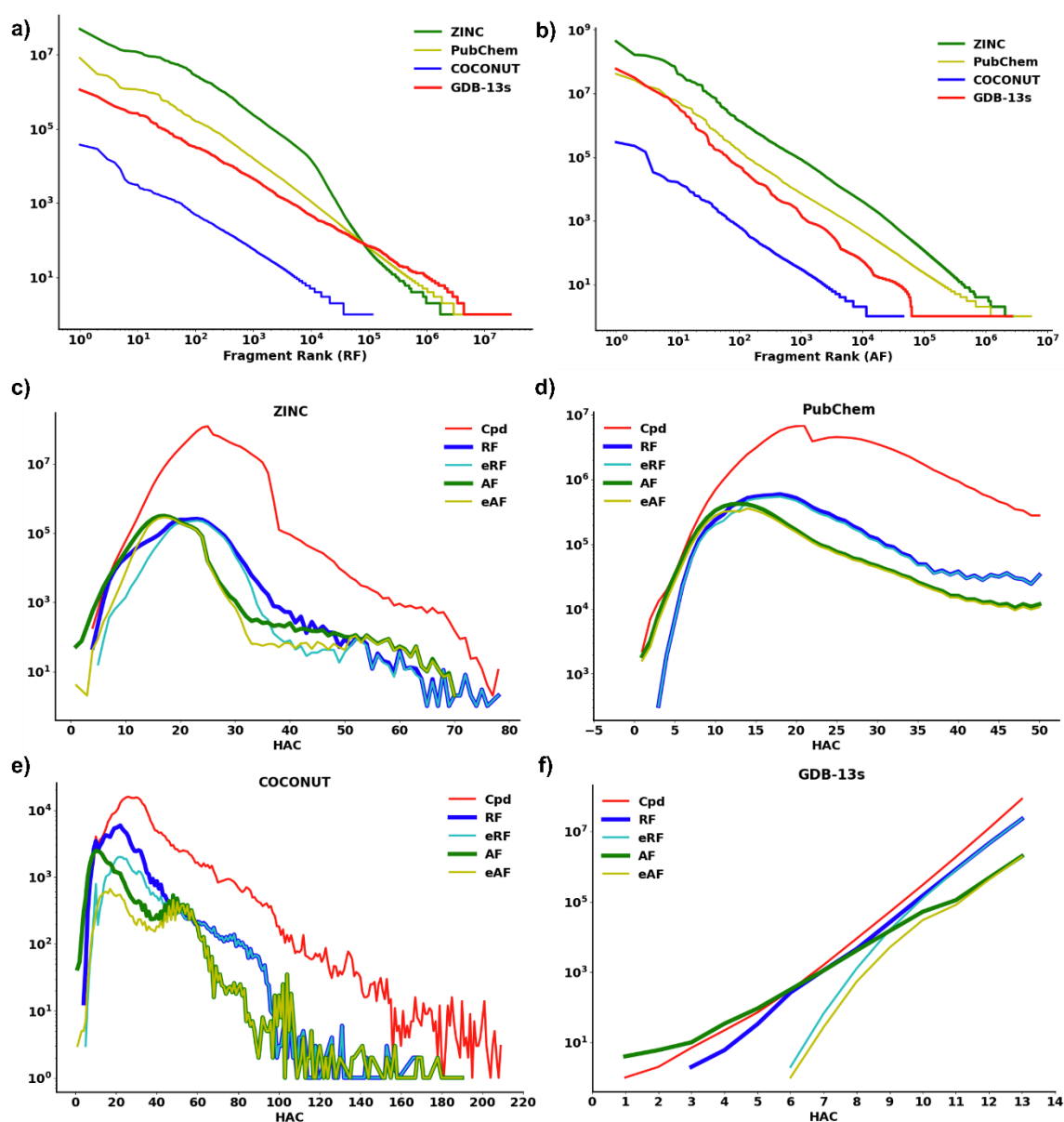


Figure 2. Frequency distribution of RFs (a) and AFs (b) in ZINC, PubChem, COCONUT, and GDB-13s. Count of compounds (Cpds), RFs, eRFs, AFs and eAFs in ZINC (c), PubChem (d), COCONUT (e), and GDB-13s (f) as a function of HAC. The curves of RF and AF are depicted thicker to help visualize the distribution in the regions with high overlap.

Comparative analysis of RFs and AFs in ChEMBL active and inactive molecules

Aiming to select novel fragments in GDB-13s by exploiting knowledge on bioactive compounds, we analyzed molecules from the ChEMBL database to test if different RFs and AFs were associated with active or inactive compounds.² We selected the 2,136,218 ChEMBL molecules with $HAC \leq 50$, separated them into 560,230 actives (IC_{50} or $EC_{50} \leq 10 \mu M$, ChEMBLa) and 1,575,988 inactives (all others, ChEMBLi), and extracted the corresponding RFs and AFs. For each RF and AF, we computed its total occurrence as the number of ChEMBL molecules containing this RF or AF, its relative occurrence in active molecules (%active) and inactive molecules (%inactive), and an activity ratio $R_{bioact} = \%active/\%inactive$.

A volcano scatter plot of the total occurrence of each RF or AF as function of R_{bioact} showed that RFs and AFs spanned a broad range of R_{bioact} values and total occurrences (**Figure 4a-b**). The situation was similar when analyzing only fragments up to 13 atoms (**Figure 4c-d**). From this analysis, we partitioned ChEMBL fragments according to their R_{bioact} values into active ($R_{bioact} \geq 4$), inactive ($R_{bioact} \leq 0.25$) or non-preferential fragments (intermediate values, $R_{bioact} \sim 1$). While the most frequent fragments were small and non-preferential, many fragments, including all singletons, occurred exclusively in either ChEMBLa or ChEMBLi subsets, and were accordingly assigned to either active ($R_{bioact} \geq 4$) or inactive ($R_{bioact} \leq 0.25$) subsets, respectively (**Table 2**). The top-10 most frequent active ($R_{bioact} \geq 4$) and inactive ($R_{bioact} \leq 0.25$) RFs and AFs in ChEMBL were all with the size range of GDB-13s. Four of these top-10 active RFs featured halogenated benzene rings, while four of the top-10 inactive RFs were saturated heterocycles (**Figure S5**). For AFs, fluorine prevailed in four of the top-10 active AFs, while sulfur occurred in four the top-10 inactive AFs (**Figure S6**).

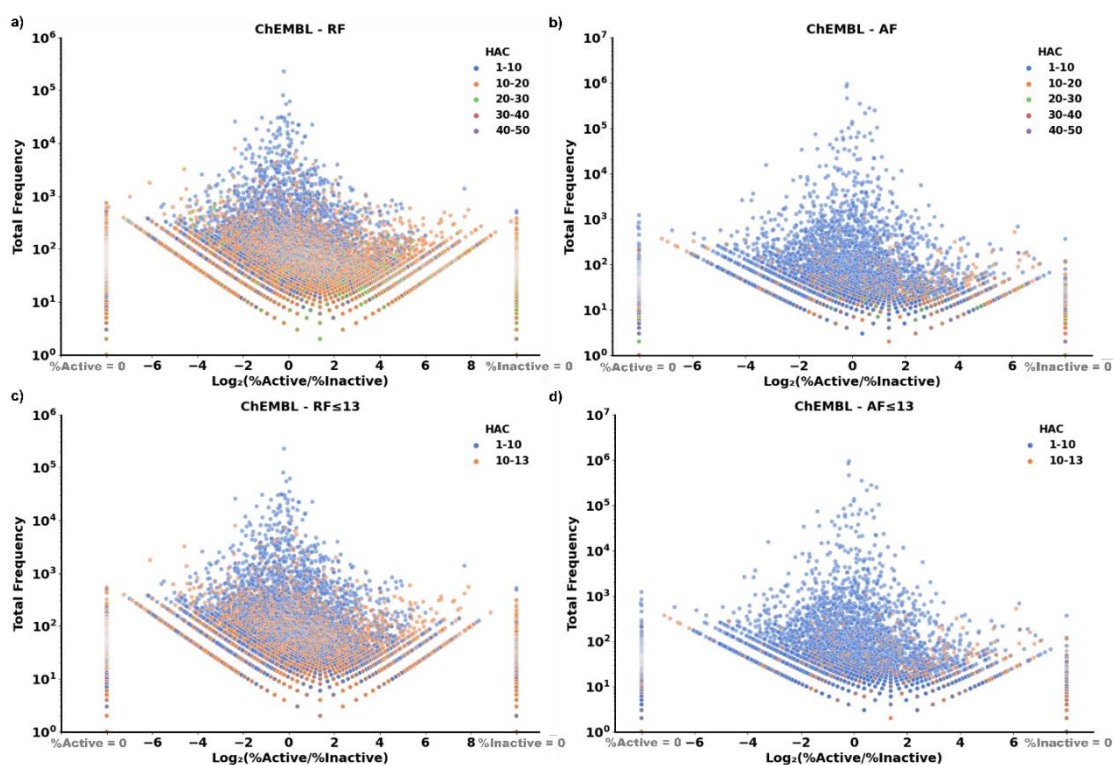


Figure 4. Volcano plots visualizing all active and inactive fragments extracted from ChEMBL. The logarithm value (base 2) of the ratio of the proportion of fragments in all active molecules and the proportions of fragments in all inactive molecules, namely $\log_2(\% \text{Active}/\% \text{Inactive})$, was plotted on the x-axis, and the total frequency (the sum of the occurrences of fragment in active molecules and in inactive molecules) was plotted on the y-axis. Colors of the data points indicating the HAC range of the fragments. Occurrences of fragments that only appeared in inactive compounds ($\% \text{Active} = 0$) were displayed vertically in a straight line at the left end of the plot, while occurrences of fragments that only appeared in active compounds ($\% \text{Inactive} = 0$) were displayed vertically in a straight line at the right end of the plot.

Table 2. RFs/AFs analysis of ChEMBLa, and ChEMBLi.

No.		ChEMBLa	ChEMBLi	$R_{\text{bioact}} \geq 4$	$R_{\text{bioact}} \sim 1$	$R_{\text{bioact}} \leq 0.25$
1	Cpds	543,971	1,575,988			
2	Cpds from RF $\leq 13^a$)	215,243 39.6%	870,442 55.2%			
3	Cpds from AF $\leq 13^b$)	523,674 96.3%	1,509,677 95.8%			
4	Cpds from ARF $\leq 13^c$)	198,367 36.5%	813,618 51.6%			
5	RF	145,174	300,613	116,023	25,197	266,255
6	eRF ^d)	106,862 73.6%	262,301 87.3%	106,862 92.1%	0 0%	262,301 98.5%
7	RF-Singleton ^e)	93,023 64.1%	193,248 64.3%	78,758 67.9%	0 0%	182,620 68.6%
8	RF $\leq 13^f$)	28,309 19.5%	55,143 18.3%	15,211 13.1%	10,883 43.2%	40,930 15.4%
9	eRF $\leq 13^g$)	11,881 8.2%	38,715 12.9%	11,881 10.2%	0 0%	38,715 14.5%
10	RF ≤ 13 -Singleton ^h)	12,260 8.5%	23,463 7.8%	7,642 6.6%	0 0%	20,699 7.8%
11	AF	26,482 4.7%	81,690 5.2%	16,567	8,605	71,125
12	eAF ^d)	14,613 55.2%	69,817 85.5%	14,613 88.2%	0 0%	69,817 98.2%
13	AF-Singleton ^e)	15,773 59.6%	49,745 60.9%	11,252 67.9%	0 0%	46,974 66.0%
14	AF $\leq 13^f$)	16,137 60.9%	45,091 55.2%	7,875 47.5%	7,063 82.1%	36,498 51.3%
15	eAF $\leq 13^g$)	6,347 24.0%	35,301 43.2%	6,347 38.3%	0 0%	35,301 49.6%
16	AF ≤ 13 -Singleton ^h)	8,008 30.2%	22,540 27.6%	4,638 28.0%	0 0%	20,689 29.1%

a) Cpds from RF ≤ 13 = molecules covered by RF up to HAC = 13. b) Cpds from AF ≤ 13 = molecules covered by AF up to HAC = 13. c) Cpds from ARF ≤ 13 = molecules covered by both RF and AF up to HAC = 13. d) eRF/eAF = exclusive RF/AF, absent from the other three databases. e) RF/AF-Singleton = RF/AF with only a single molecule example. f) RF ≤ 13 /AF ≤ 13 = RF/AF up to HAC = 13. g) eRF ≤ 13 /eAF ≤ 13 = exclusive RF ≤ 13 /AF ≤ 13 , absent from the other three databases. h) RF ≤ 13 -Singleton /ARF ≤ 13 -Singleton = RF ≤ 13 /AF ≤ 13 with only a single molecule example. RF and AF subcategories are calculated relative to total RF and AF, respectively.

While many RFs and AFs occurred preferentially in either ChEMBL active or ChEMBL inactive molecules, these fragments did not differ strongly from each other or from RFs and AFs in known molecules (PubChem, ZINC and COCONUT) in terms of overall structural features. Indeed, the different datasets of known molecules had quite similar property profiles for RFs up to 13 atoms in terms of the number of rings, largest ring size, number of acyclic atoms and heteroatoms (**Figure 5a-d**). Similarly, AFs up to 13 atoms in these datasets had comparable property profiles concerning the number of quaternary centers, triple bonds, heteroatoms, and terminal atoms (**Figure S7a-d**).

On the other hand, the property profiles of GDB-13s RFs and AFs were clearly different from those of known molecules. For instance, RFs from GDB-13s had a broader distribution in terms of number of rings and largest ring size, and fewer heteroatoms than the different RF datasets of known molecules. Furthermore, GDB-13s AFs stood out with a larger number of triple bonds and terminal atoms compared to AF datasets of known molecules. These differences probably explained the less favorable synthetic accessibility score (SAscore) of GDB-13s RFs and AFs (**Figure 5e** and **S7e**).²⁸ Indeed, the SAscore is based on the presence of substructures frequently found in known molecules. Note that GDB-13s RFs and AFs had relatively high natural product likeness scores (NPscore),³¹ comparable to those of COCONUT molecules (**Figure 5f** and **S7f**). The high NPscore of GDB-13s RFs and AFs probably reflects the high percentage of non-aromatic, stereochemically complex structures in GDB-13s since the NPscore assigns higher values for the presence of such structural features.

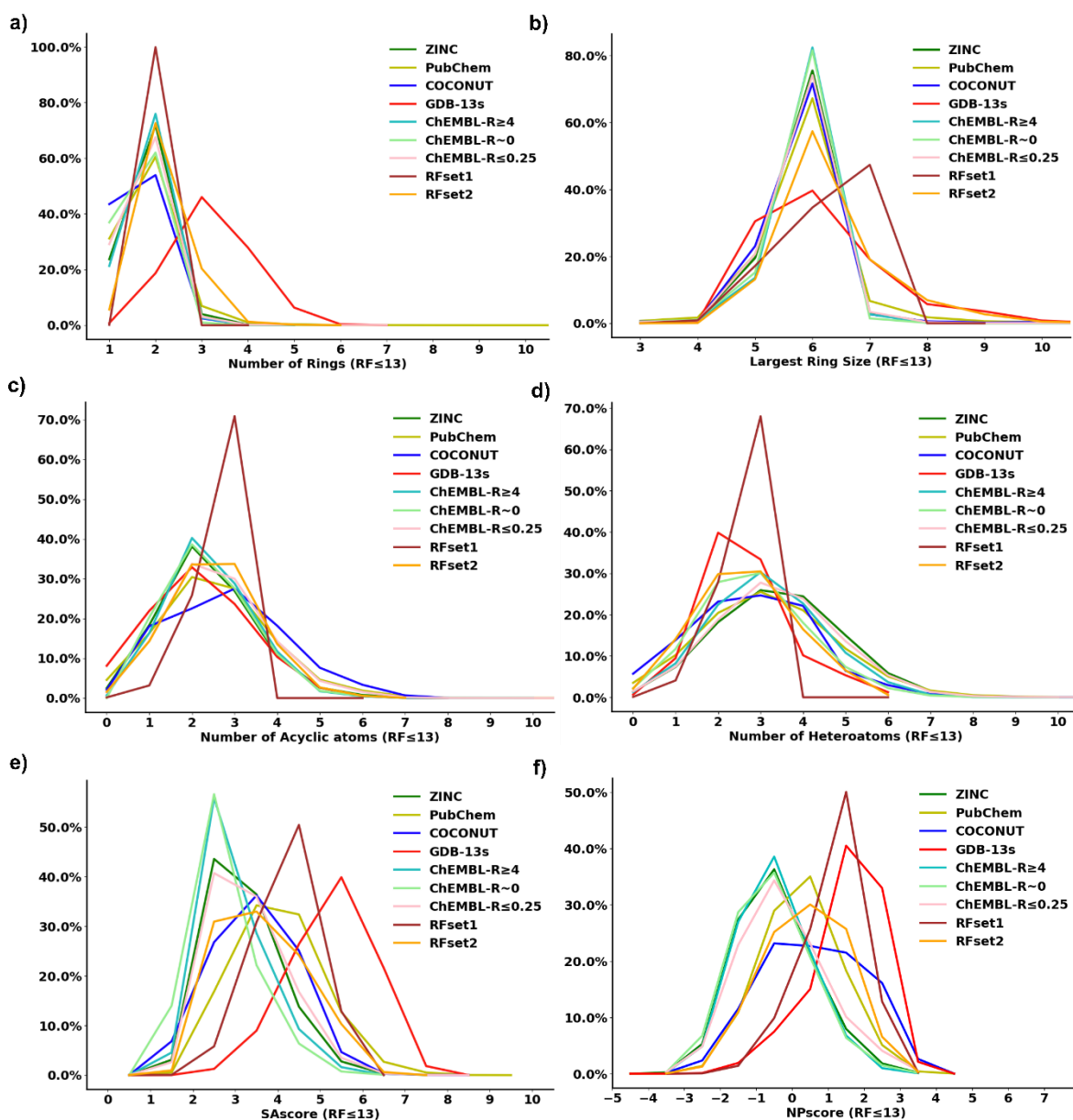


Figure 5. Frequency histograms of RFs from the various databases and subsets for (a) number of rings, (b) largest ring size, (c) number of acyclic atoms, (d) number of heteroatoms, (e) SAScore, and (f) NPscore.

Bioactivity guided selection of RFs and AFs in GDB-13s

The analysis above suggested two possible approaches to select RFs and AFs from GDB-13s for drug design. First, the narrower structural parameter ranges covered by RFs and AFs from known molecules, active or inactive, which correlated with their more favorable SAScores compared to GDB-13s RFs and AFs, suggested to select GDB-13s fragments with limited structural complexity, which would certainly help in view of a possible synthesis. Following up on this idea, we selected a subset of GDB-13s RFs and AFs by constraining structural parameters closer to known molecules but considering only those exclusive to GDB-13s to ensure novelty. To our delight, this selection resulted in a sizable number of GDB-13s fragments. Indeed, we obtained 960,587 GDB-13s eRFs with up to two rings, ring size up to seven, up to three heteroatoms and three acyclic atoms, named **RFset1**. For the selection of AFs from GDB-13s, we obtained 462,439 GDB-13s eAFs without any quaternary center and up to one triple bond, up to four heteroatoms and up to four terminal atoms, named **AFset1**.

In a second, narrower selection, we assumed that ChEMBL derived RFs and AFs in the $R_{\text{bioact}} \geq 4$ value range (defined as active fragments) reflected privileged structural types, while those in the $R_{\text{bioact}} \leq 0.25$ value range (defined as inactive fragments) marked undesirable structural types, in terms of possible bioactivities. To expand the scope of ChEMBL active fragments, we retrieved all GDB-13s RFs and AFs within a Jaccard distance $d_J \leq 0.6$ of any of the ChEMBL active fragments using the MAP4 fingerprint as similarity measure.³² In this manner, we obtained 97,664 RFs and 43,704 AFs, from which we removed 25,162 RF and 15,484 AF found within $d_J \leq 0.6$ of any inactive fragments, leaving 72,502 RFs, named **RFset2**, and 28,220 AFs, named **AFset2**, as bioactive-like fragments from GDB-13s. In these sets, many fragments were also exclusive to GDB-13s, ensuring novelty (51,303 eRFs, 70.8% and 17,620 eAFs, 62.4%).

The property profiles of **RFset1** and **AFset1**, which both resulted from constraining structural parameters, remained substantially different from those of known molecules because frequency peaked at the highest parameter value selected. This distribution reflects the combinatorial enumeration used to generate GDB-13s, which provides many more possible molecules at the largest values of structural parameters. Therefore, the SAScore remained less favorable and the NPscore relatively high in both sets. On the other hand, the property profiles of **RFset2** and **AFset2**, selected by substructure similarity to ChEMBL bioactive fragments, were like those of known molecules, reflecting the structural similarity selection used to compose these sets (**Figure 5a-d** and **S7a-d**). **RFset2** and **AFset2** also displayed lower SAScore and NPscore values than the full sets of GDB-13s RFs and AFs, indicating that they were generally less complex and closer to RFs and AFs from known molecules (**Figure 5e-f** and **S7e-f**).

To gain a detailed insight into the bioactivity selected subset of GDB-13s RFs and AFs, we computed interactive TMAPs (tree-maps)³³ using the MinHashed fingerprint MAP4 as similarity measure (**Figure 6**).³² These interactive TMAPs allow one to browse through the two databases and search for interesting RFs and AFs using various color-coded properties as guide. To illustrate the available options, we searched for novel analogs of the three most frequent active ($R_{\text{bioact}} \geq 4$) RFs in ChEMBL, one of which occurs in the kinase inhibitor drug gefitinib, revealing potentially interesting analogs (**Figure 7**). Further interesting GDB-13s eRFs are exemplified as analogs of triquinazine, an eRF from GDB-13s previously used as scaffold for a Janus kinase inhibitor analog of the known drug tofacitinib.³⁴ In principle, the same selection can also be made with GDB-13s analogs of AFs, as exemplified for the most frequent AFs in active ($R_{\text{bioact}} \geq 4$) AFs from ChEMBL (**Figure S8**). In this case however, the selection of interesting AFs is less obvious since the chemistry of AFs highly depends on their connection to RFs.

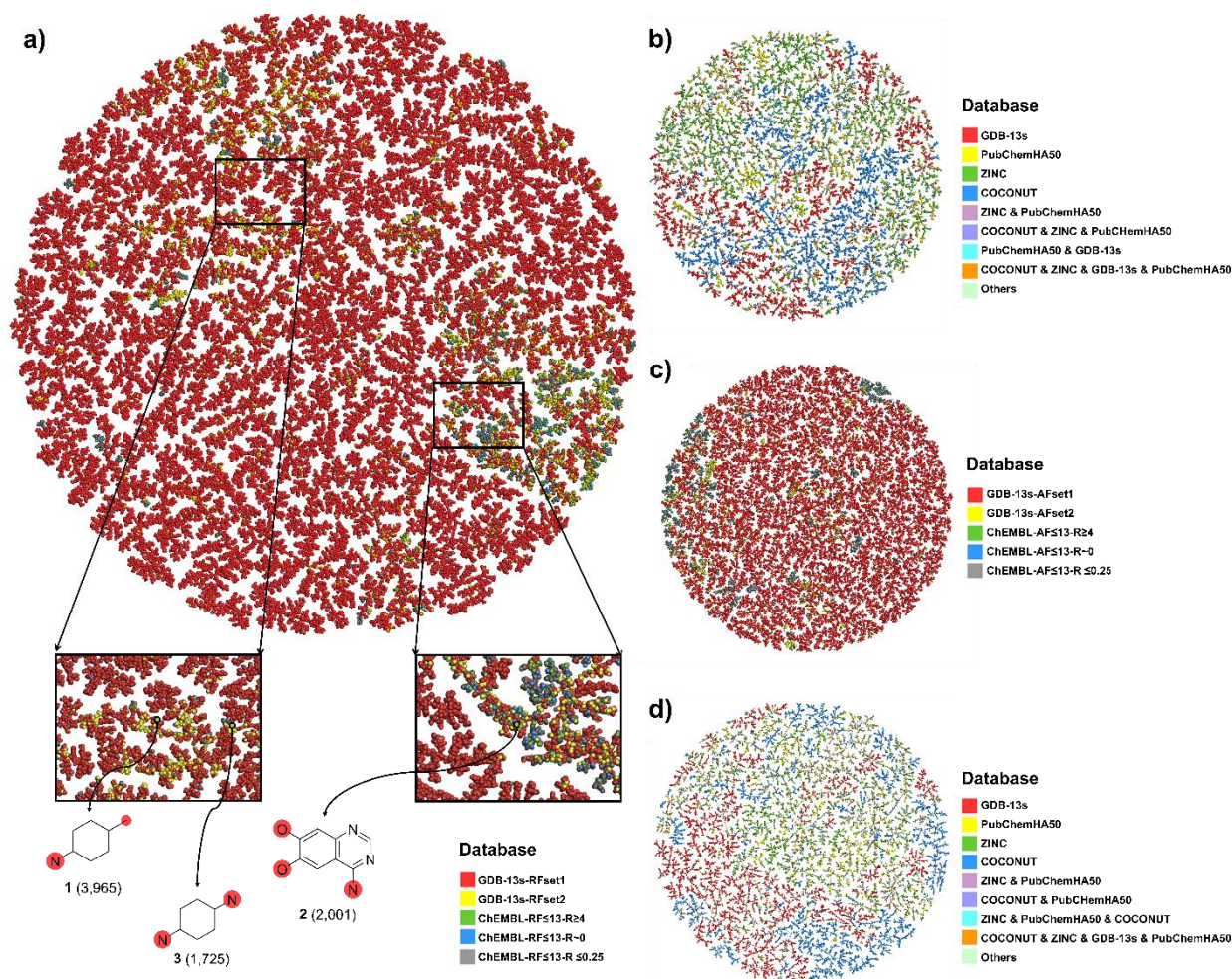


Figure 6. TMAP visualization of (a) 1,042,610 RFs from **RFset1**, **RFset2** and ChEMBL (b) top-10,000 RFs in ZINC, PubChem, COCONUT and GDB-13s (c) 533,153 AFs from **AFset1**, **AFset2** and ChEMBL (d) top-10,000 AFs in ZINC, PubChem, COCONUT and GDB-13s, color-coded by the source datasets, SAscore and different properties. An interactive version of the TMAPs is accessible at <https://tm.gdb.tools/map4> (MAP4_fused_GDB-13s_RFset1_RFset2_and_ChEMBL; MAP4_4databases_top10k_RF; MAP4_fused_GDB-13s_AFset1_AFset2_and_ChEMBL; MAP4_4databases_top10k_AF).

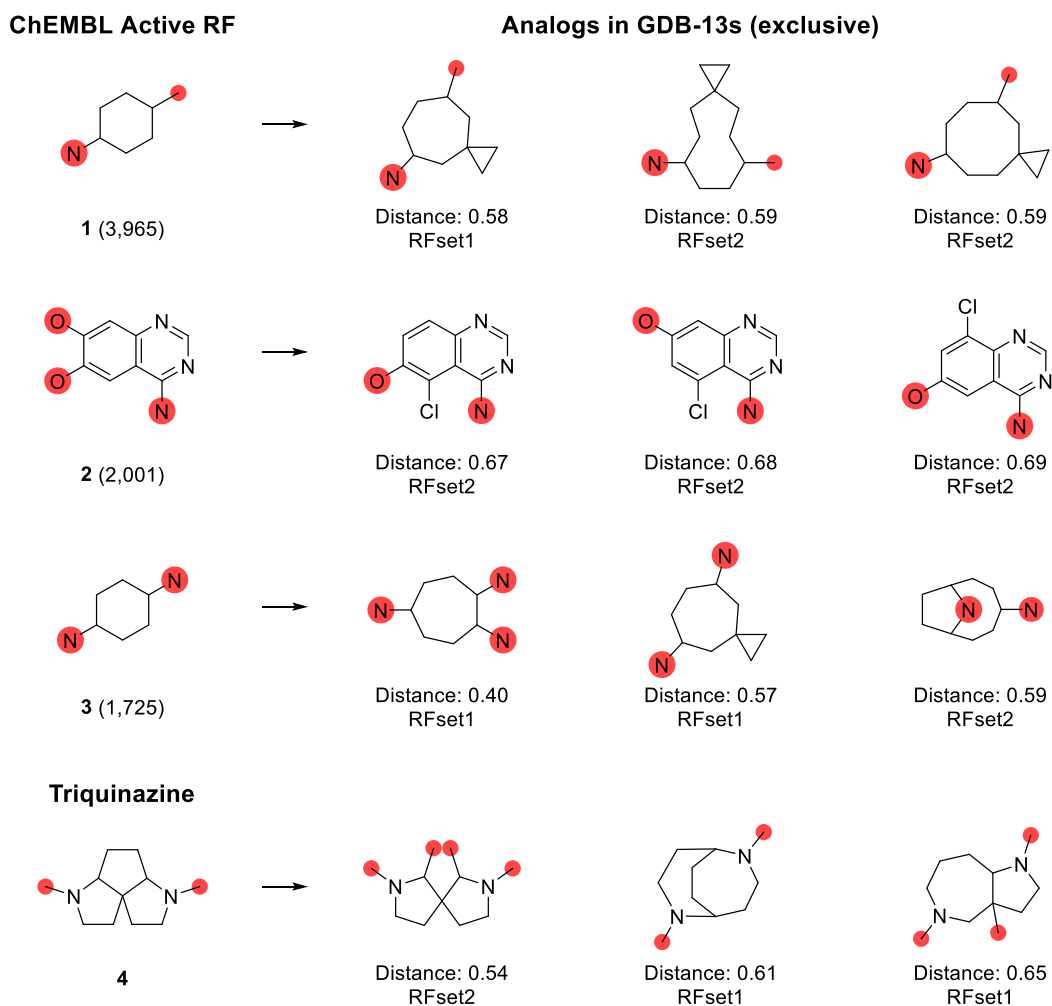


Figure 7. Analogs of highly active ChEMBL RFs and triquinazine found in the subsets of GDB-13s (**RFset1** / **RFset2**). Total occurrences of the ChEMBL RFs, or the distances between the analogs and the targets are indicated in parentheses.

Conclusion

The goal of this study was to focus attention on the most relevant part of the vast small molecule chemical space revealed by the enumeration of the GDB databases by considering the need for novelty combined with structural simplicity and a certain level of similarity to known bioactive molecules. The narrowing of the many millions GDB molecules to approximately one million RFs and half a million AFs represents an enormous reduction in the number of molecules to be considered but opens more than enough opportunities for novelty to be realized by chemical

synthesis, considering that the practical synthesis of novel building blocks is resource and time intensive. Focusing on novel yet simple structures is essential to follow the evidence of more than a century of medicinal chemistry showing that the most useful drug molecules do not need to be very complex.

Methods

Extracting RFs and AFs from molecules

RFs and AFs were obtained from molecules by processing their SMILES³⁵ using RDkit³⁶ as follows (**Figure 1**). RFs: break all bonds between any two acyclic atoms and remove all acyclic atoms not directly attached to the rings. Acyclic atoms directly connected to more than one ring system are disconnected and reattached to each ring system separately. AFs: break all bonds between cyclic atoms and acyclic atoms and remove all cyclic atoms.

TMAPs

Tree-maps (TMAPs) were generated by specifying standard parameters,³³ using the MAP4 fingerprint (MinHashed atom-pair fingerprint up to a diameter of four bonds).³² MAP4 fingerprints were computed with a dimension of 256.

Data and Software Availability

GDB-13 and GDB-13s are hosted on the open-access repository Zenodo and can be downloaded free of charge at <https://doi.org/10.5281/zenodo.7041051>. All the molecules are stored in dearomatized, canonized SMILES format and compressed as a GNU zip archive. The ZINC data used in this study is the February 2022 version (<https://zinc.docking.org>). The PubChem data with a version of October 2021, was first downloaded from the NCBI (The National Center for Biotechnology Information), NIH (National Institutes of Health) via an FTP server (<https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full>). Then

the compounds with HACs not greater than 50 were extracted to build the PubChem database. The COCONUT data adopted in this study is the February 2021 version (<https://github.com/reymond-group/Coconut-TMAP-SVM>). ChEMBL active and inactive data sets were extracted from ChEMBL31 (<https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest>). The Molecule Breakdown Model is made freely available and under the MIT license. It is distributed in a GitHub repository upon publication of this manuscript: https://github.com/Ye-Buehler/Molecule_Breakdown_Model.

Associated Content

Supporting Information

The Supporting Information is available free of charge at <https://xxx>.

Top-10 most populated RFs/AFs in GDB-13s, ZINC, PubChem and COCONUT; Top-20 most frequent RFs shared by the different databases; Top-10 eRFs in the different databases; Top-10 most frequent RFs and AFs in the active and inactive ChEMBL subsets; Frequency histograms of AFs from the various databases and subsets for number of quaternary centers, number of triple bonds, number of heteroatoms, number of terminal atoms, SAScore, and NPscore; Analogs of highly active ChEMBL AFs found in GDB-13s **AFset1/AFset2** (PDF).

Author Information

Corresponding Author

Jean-Louis Reymond - *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*; **ORCID**: 0000-0003-2724-2942; ***E-Mail**: jean-louis.reymond@unibe.ch.

Other Author

Ye Buehler - *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*; **ORCID**: 0000-0002-8139-830X; **E-Mail**: ye.buehler-feng@unibe.ch.

Author Contributions

Y.B. designed, realized the study and wrote the paper. J.L.R. co-designed and supervised the study and wrote the paper.

Funding

This work was funded by the Swiss National Science Foundation, grant number 200020_207976.

Notes

The authors declare no competing financial interest.

Acknowledgments

We thank Dr. Sacha Javor for critical reading of the manuscript and helpful suggestions. We also thank UBELIX (<http://www.id.unibe.ch/hpc>), the HPC cluster at the University of Bern, for providing free computing service.

References

- (1) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. <https://doi.org/10.1093/nar/gky1033>.
- (2) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.;

- Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. <https://doi.org/10.1093/nar/gky1075>.
- (3) Taylor, R. D.; MacCoss, M.; Lawson, A. D. G. Rings in Drugs. *J. Med. Chem.* **2014**, *57* (14), 5845–5859. <https://doi.org/10.1021/jm4017625>.
- (4) Ivanenkov, Y. A.; Zagribelnyy, B. A.; Aladinskiy, V. A. Are We Opening the Door to a New Era of Medicinal Chemistry or Being Collapsed to a Chemical Singularity? *J. Med. Chem.* **2019**, *62* (22), 10026–10043. <https://doi.org/10.1021/acs.jmedchem.9b00004>.
- (5) Krieger, J.; Li, D.; Papanikolaou, D. Missing Novelty in Drug Development*. *Rev. Financ. Stud.* **2022**, *35* (2), 636–679. <https://doi.org/10.1093/rfs/hhab024>.
- (6) Bhutani, P.; Joshi, G.; Raja, N.; Bachhav, N.; Rajanna, P. K.; Bhutani, H.; Paul, A. T.; Kumar, R. U.S. FDA Approved Drugs from 2015–June 2020: A Perspective. *J. Med. Chem.* **2021**, *64* (5), 2339–2381. <https://doi.org/10.1021/acs.jmedchem.0c01786>.
- (7) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131* (25), 8732–8733. <https://doi.org/10.1021/ja902302h>.
- (8) Reymond, J.-L.; Ruddigkeit, L.; Blum, L.; van Deursen, R. The Enumeration of Chemical Space. *WIREs Comput. Mol. Sci.* **2012**, *2* (5), 717–733. <https://doi.org/10.1002/wcms.1104>.
- (9) Meier, K.; Bühlmann, S.; Arús-Pous, J.; Reymond, J.-L. The Generated Databases (GDBs) as a Source of 3D-Shaped Building Blocks for Use in Medicinal Chemistry and Drug Discovery. *CHIMIA* **2020**, *74* (4), 241–241. <https://doi.org/10.2533/chimia.2020.241>.
- (10) Mullard, A. The Drug-Maker's Guide to the Galaxy. *Nature* **2017**, *549* (7673), 445–447. <https://doi.org/10.1038/549445a>.
- (11) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inform.* **2018**, *37* (1–2), 1700123. <https://doi.org/10.1002/minf.201700123>.

- (12) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *J. Cheminformatics* **2019**, *11* (1), 71. <https://doi.org/10.1186/s13321-019-0393-0>.
- (13) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today* **2019**, *24* (5), 1148–1156. <https://doi.org/10.1016/j.drudis.2019.02.013>.
- (14) Zhang, J.; Mercado, R.; Engkvist, O.; Chen, H. Comparative Study of Deep Generative Models on Chemical Space Coverage. *J. Chem. Inf. Model.* **2021**, *61* (6), 2572–2581. <https://doi.org/10.1021/acs.jcim.0c01328>.
- (15) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (9), 2021–2034. <https://doi.org/10.1021/acs.jcim.2c00224>.
- (16) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. <https://doi.org/10.1021/jm9602928>.
- (17) Buehler, Y.; Reymond, J.-L. Molecular Framework Analysis of the Generated Database GDB-13s. *J. Chem. Inf. Model.* **2023**, *63* (2), 484–492. <https://doi.org/10.1021/acs.jcim.2c01107>.
- (18) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065–6073. <https://doi.org/10.1021/acs.jcim.0c00675>.
- (19) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminformatics* **2021**, *13* (1), 2. <https://doi.org/10.1186/s13321-020-00478-9>.
- (20) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAPsRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying

Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry.
12.

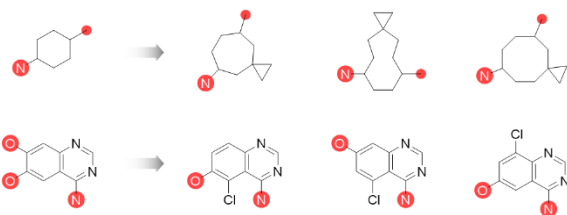
- (21) Kruger, F.; Stiefl, N.; Landrum, G. A. RdScaffoldNetwork: The Scaffold Network Implementation in RDKit. *J. Chem. Inf. Model.* **2020**, *60* (7), 3331–3335. <https://doi.org/10.1021/acs.jcim.0c00296>.
- (22) Kolb, P. Decomposition And Identification of Molecules. **2010**.
- (23) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using “Drug-Like” Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503–1507. <https://doi.org/10.1002/cmdc.200800178>.
- (24) Heikamp, K.; Zuccotto, F.; Kiczun, M.; Ray, P.; Gilbert, I. H. Exhaustive Sampling of the Fragment Space Associated to a Molecule Leading to the Generation of Conserved Fragments. *Chem. Biol. Drug Des.* **2018**, *91* (3), 655–667. <https://doi.org/10.1111/cbdd.13129>.
- (25) Liu, T.; Naderi, M.; Alvin, C.; Mukhopadhyay, S.; Brylinski, M. Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with EMolFrag. *J. Chem. Inf. Model.* **2017**, *57* (4), 627–631. <https://doi.org/10.1021/acs.jcim.6b00596>.
- (26) Ghersi, D.; Singh, M. MolBLOCKS: Decomposing Small Molecule Sets and Uncovering Enriched Fragments. *Bioinformatics* **2014**, *30* (14), 2081–2083. <https://doi.org/10.1093/bioinformatics/btu173>.
- (27) Chemaxon. <https://chemaxon.com> (accessed 2022-12-05).
- (28) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. <https://doi.org/10.1186/1758-2946-1-8>.
- (29) Erlanson, D. A.; McDowell, R. S.; O’Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **2004**, *47* (14), 3463–3482. <https://doi.org/10.1021/jm040031v>.

- (30) Hajduk, P. J.; Greer, J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discov.* **2007**, *6* (3), 211–219. <https://doi.org/10.1038/nrd2220>.
- (31) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J Chem Inf Model* **2008**, *48* (1), 68–74. <https://doi.org/10.1021/ci700286x>.
- (32) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminformatics* **2020**, *12* (1), 43. <https://doi.org/10.1186/s13321-020-00445-4>.
- (33) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminformatics* **2020**, *12* (1), 12. <https://doi.org/10.1186/s13321-020-0416-x>.
- (34) Meier, K.; Arús-Pous, J.; Reymond, J.-L. A Potent and Selective Janus Kinase Inhibitor with a Chiral 3D-Shaped Triquinazine Ring System from Chemical Space. *Angew. Chem. Int. Ed.* **2021**, *60* (4), 2074–2077. <https://doi.org/10.1002/anie.202012049>.
- (35) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. <https://doi.org/10.1021/ci00057a005>.
- (36) *RDKit: Open-source cheminformatics*. <http://www.rdkit.org> (accessed 2022-07-25).

Graphics for the table of contents:

ChEMBL Active Ring
Fragments (116,023)

Analogs in GDB-13s
(1,018,741)



Supporting Information for

Expanding Bioactive Fragment Space with the Generated Database GDB-13s

Ye Buehler and Jean-Louis Reymond*

*Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern,
Freiestrasse 3, 3012 Bern, Switzerland*

*E-Mail: jean-louis.reymond@unibe.ch.

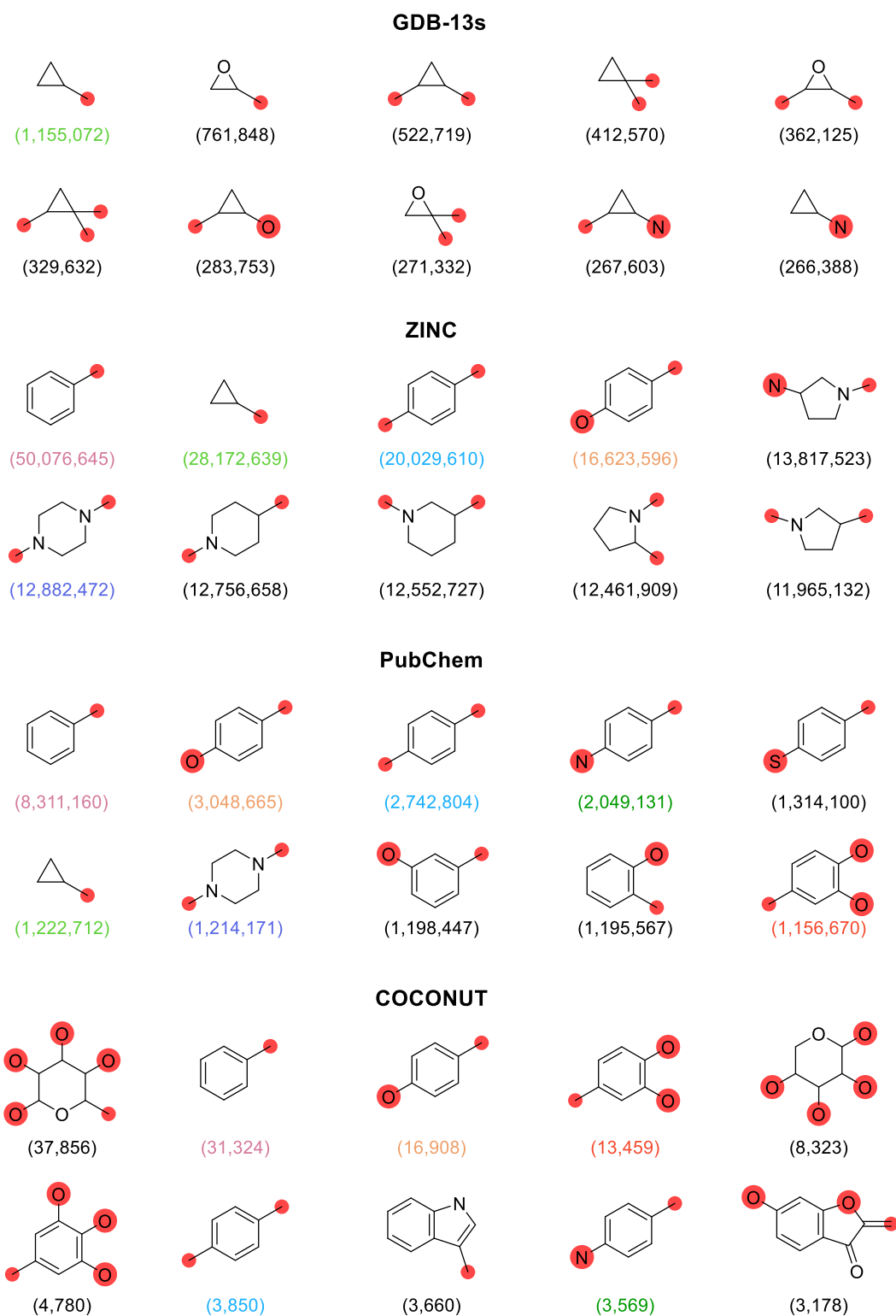


Figure S1. Top-10 most populated RFs in various databases. RFs are displayed by order of appearance in the frequency-sorted list across the four databases. A color-code has been added to the numbering of RFs appearing several times to facilitate comparison across different databases.









GDB-13s				
C (59,728,857)	O (31,359,435)	N (17,833,068)	— (12,838,313)	—O (10,272,140)
—N (8,217,373)	≡ (7,472,170)	≡N (5,646,458)	=O (4,527,252)	 (3,990,243)
ZINC				
C (439,571,674)	O (163,746,752)	=O (157,963,731)	 (137,113,332)	F (122,930,376)
—O (108,087,576)	Cl (90,508,032)	— (76,131,053)	 (70,240,782)	 (41,327,773)
PubChem				
C (41,483,903)	O (26,310,947)	—O (16,911,266)	Cl (15,291,554)	F (13,322,309)
N (9,796,147)	 (7,419,007)	Br (6,628,254)	— (6,326,128)	=O (5,571,694)
COCONUT				
O (295,447)	C (225,321)	—O (146,339)	 (33,600)	 (28,438)
— (23,490)	 (18,133)	N (17,310)	=O (16,899)	Cl (16,406)

Figure S2. Top-10 most populated AFs in various databases. AFs are displayed by order of appearance in the frequency-sorted list across the four databases.

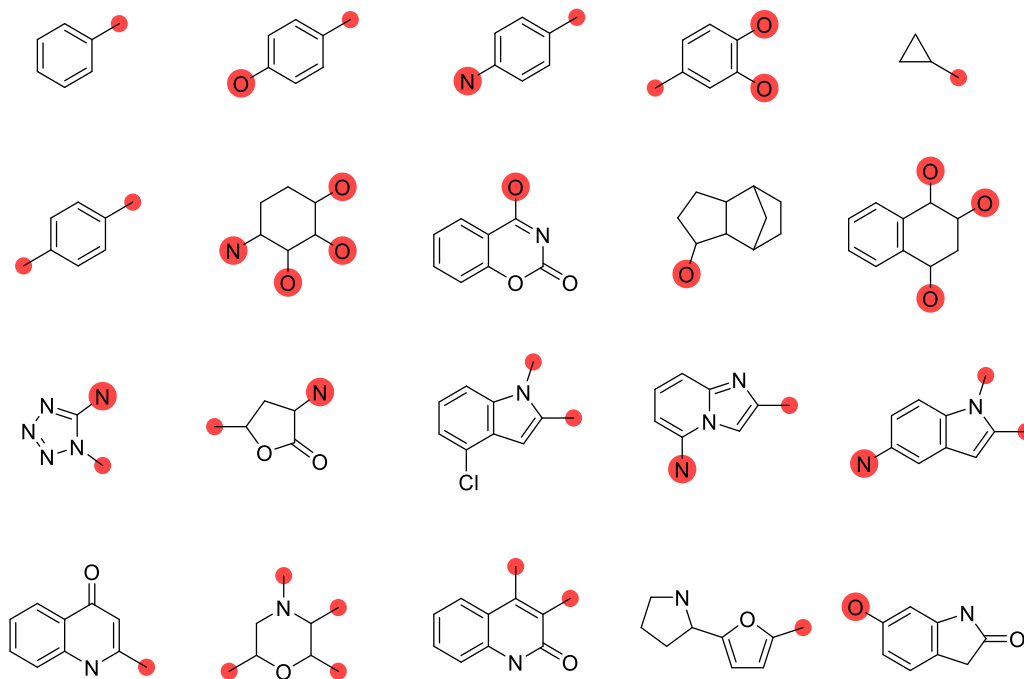


Figure S3. Top-20 most populated RFs shared by the different databases (GDB-13s, ZINC, PubChem, and COCONUT).

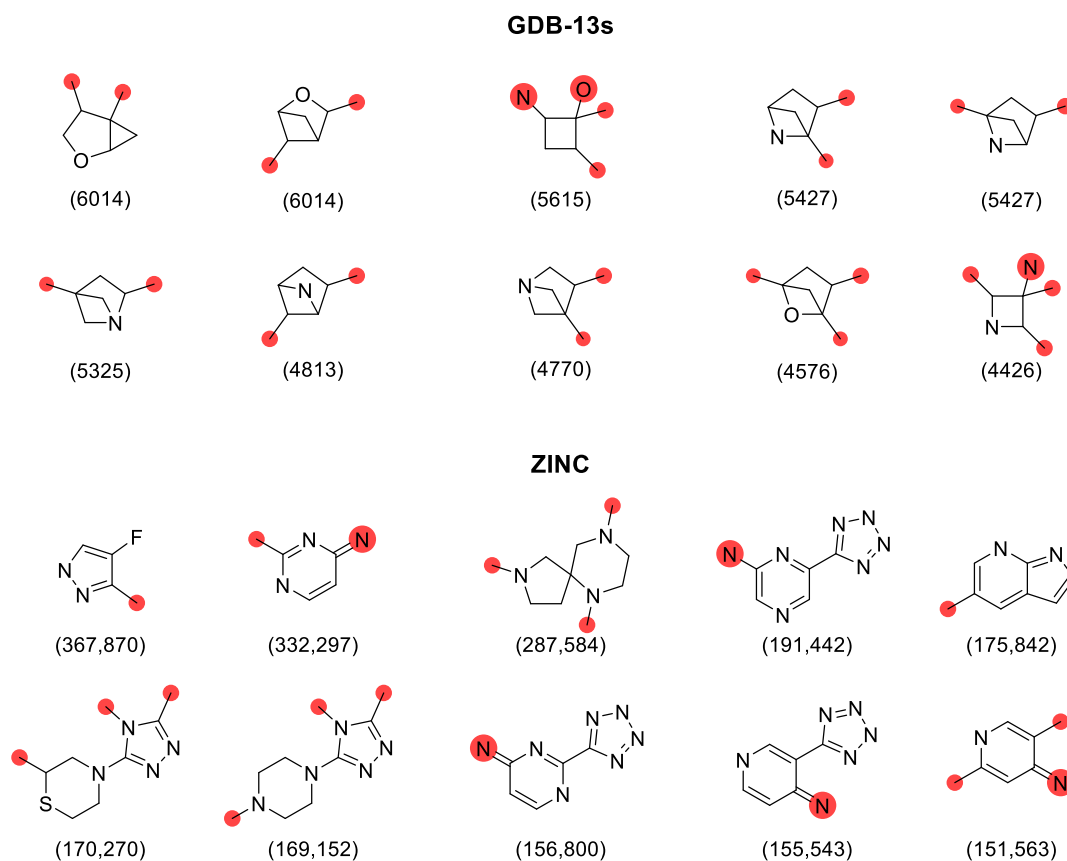


Figure S4. (a) Top-10 eRFs in GDB-13s and ZINC, occurrences of the obtained eRFs are indicated in parentheses.

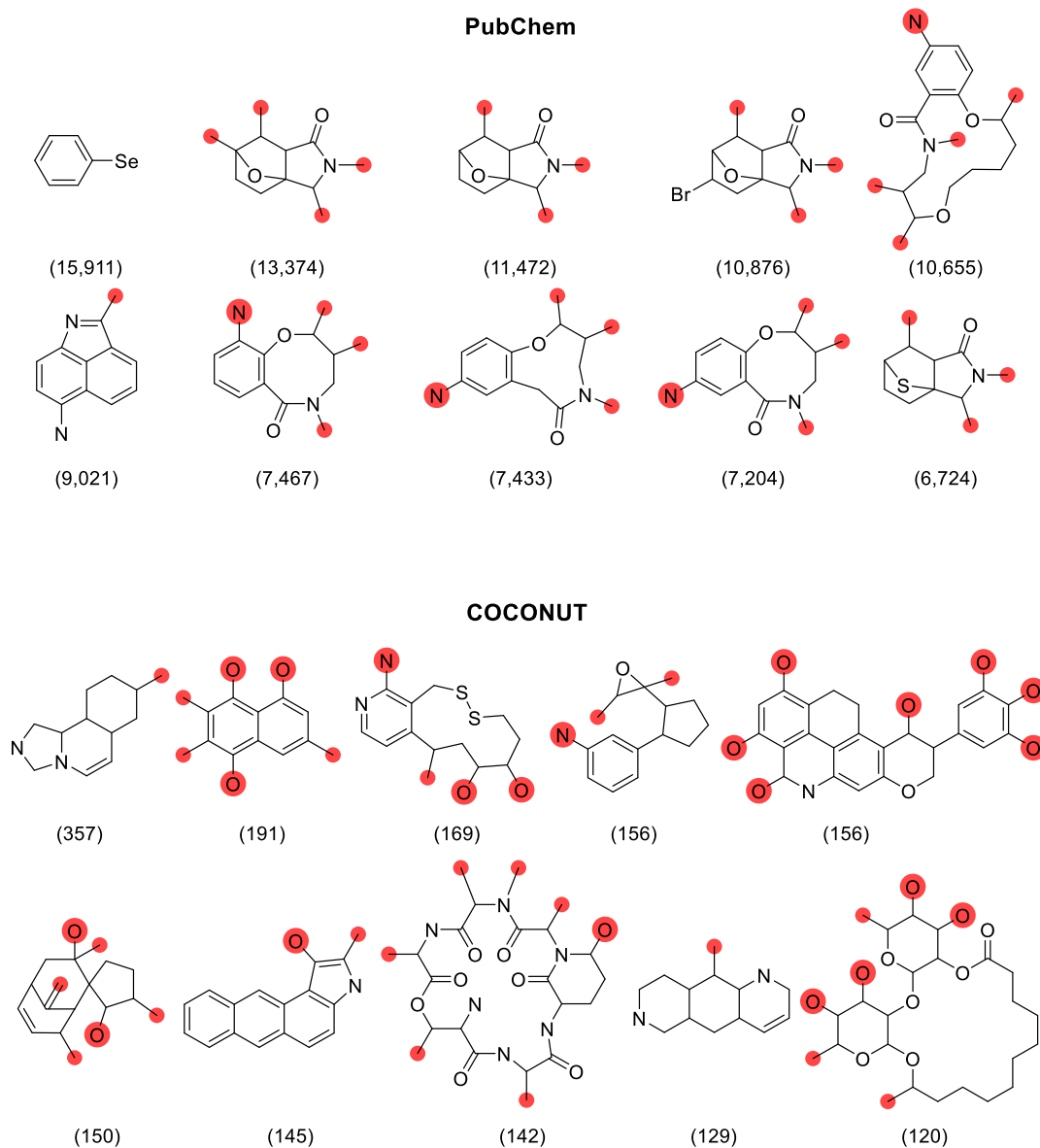


Figure S4. (b) Top-10 eRFs in PubChem and COCONUT, occurrences of the obtained eRFs are indicated in parentheses.

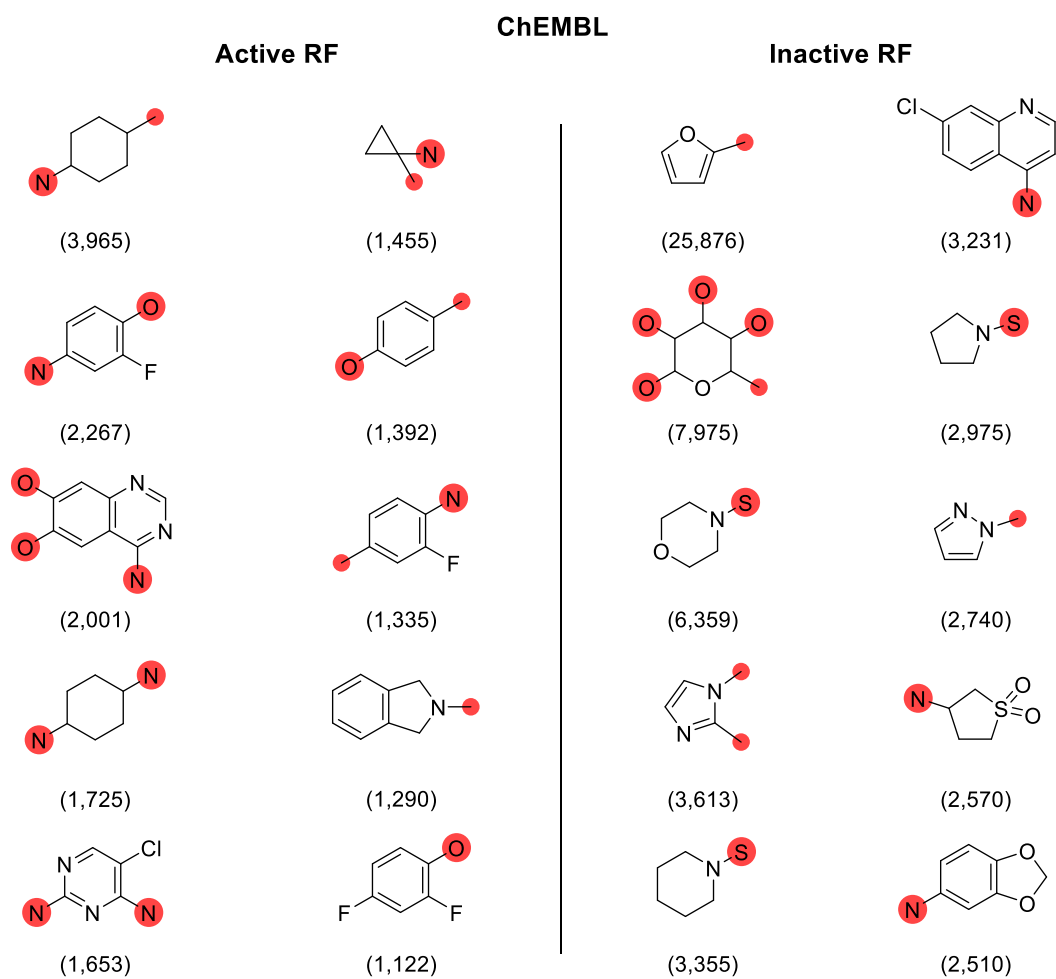


Figure S5. Top-10 most frequent RFs in the active ($R_{\text{bioact}} \geq 4$) and inactive ($R_{\text{bioact}} \leq 0.25$) ChEMBL subsets annotated with total occurrences of the RF in ChEMBL.

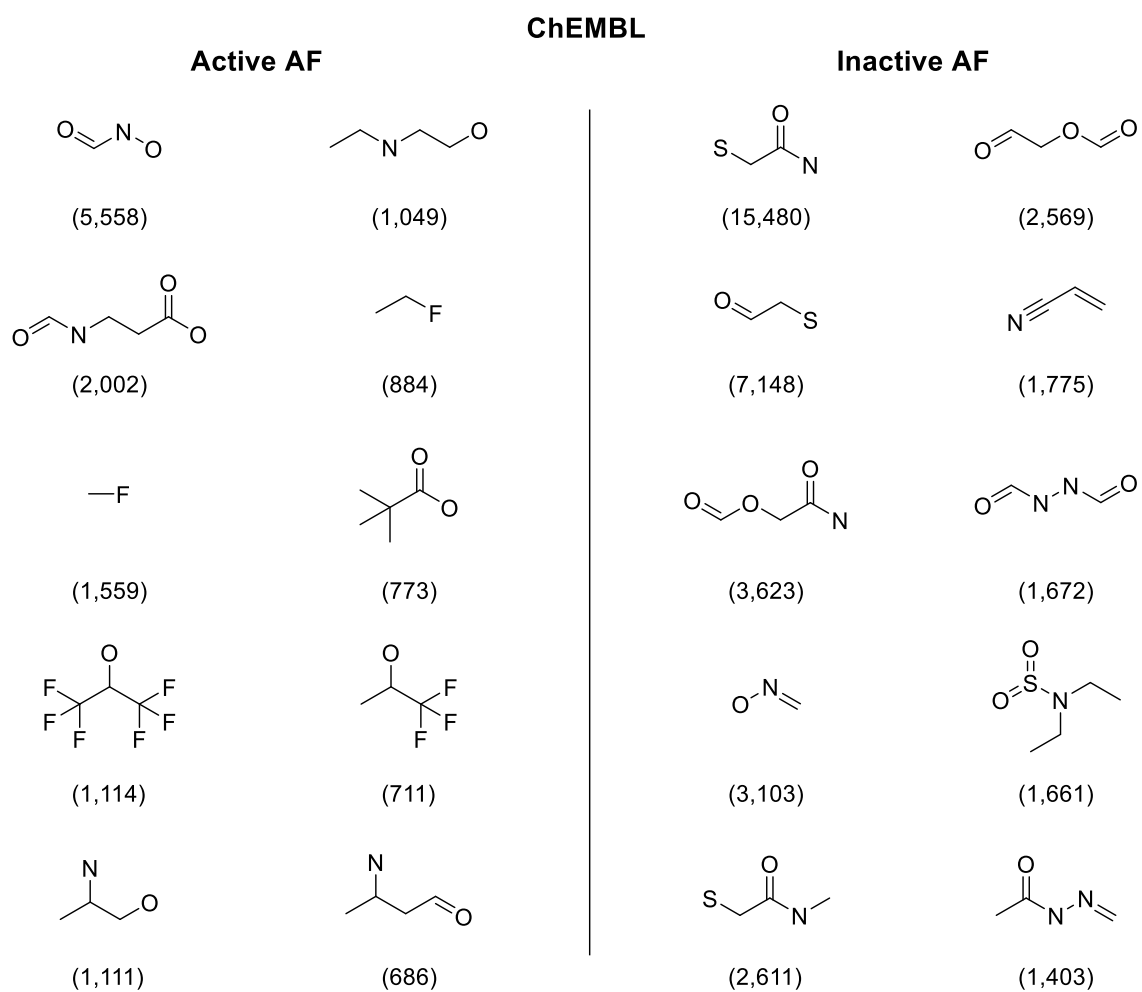


Figure S6. Top-10 most frequent AFs in the active ($R_{\text{bioact}} \geq 4$) and inactive ($R_{\text{bioact}} \leq 0.25$) ChEMBL subsets annotated with total occurrences of each AF in ChEMBL.

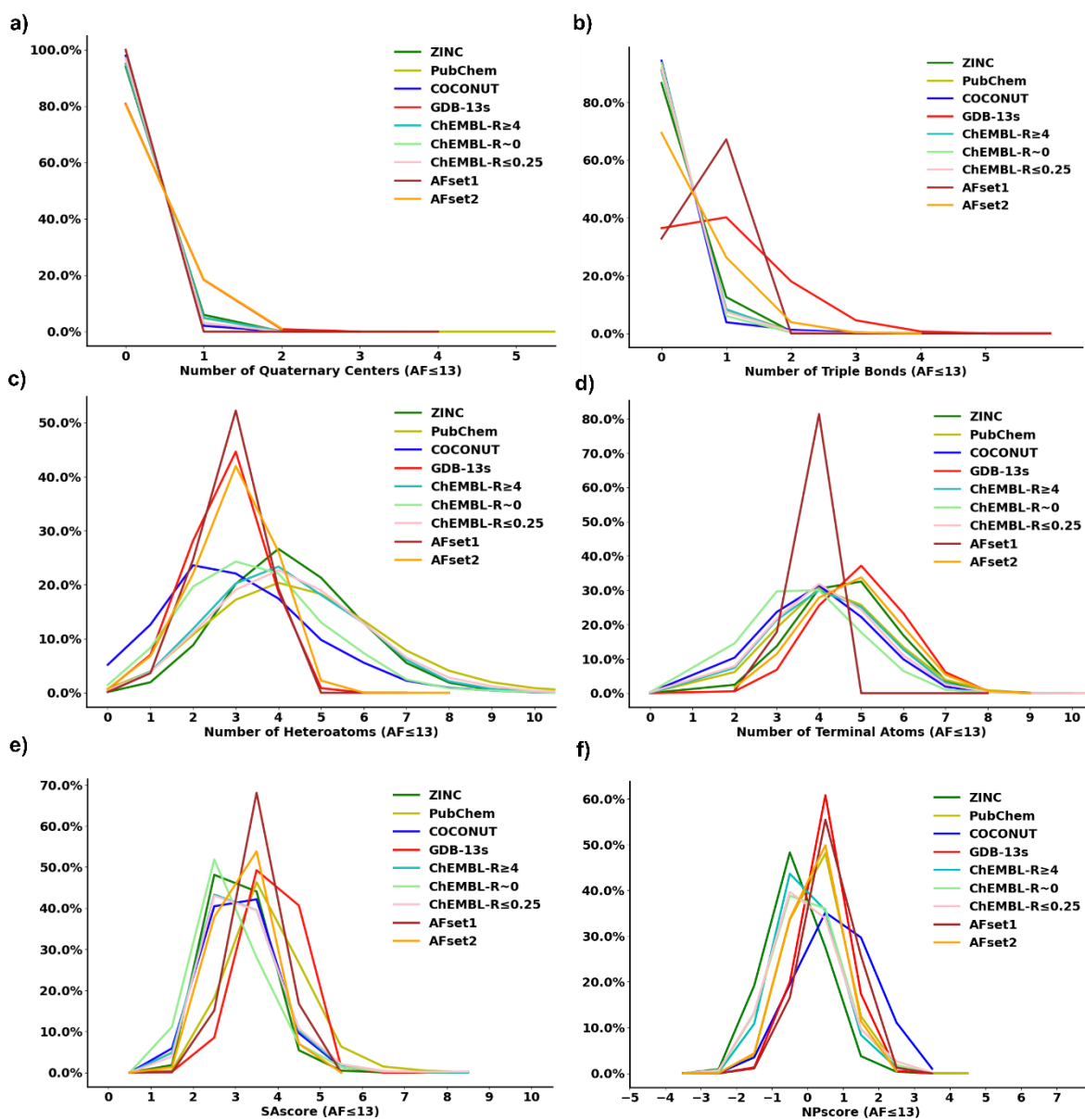


Figure S7. Frequency histograms of AFs from the various databases and subsets for (a) number of quaternary centers, (b) number of triple bonds, (c) number of heteroatoms, (d) number of terminal atoms, (e) SAScore, and (f) NPscore.

ChEMBL Active AF

Analogues in GDB-13s (exclusive)

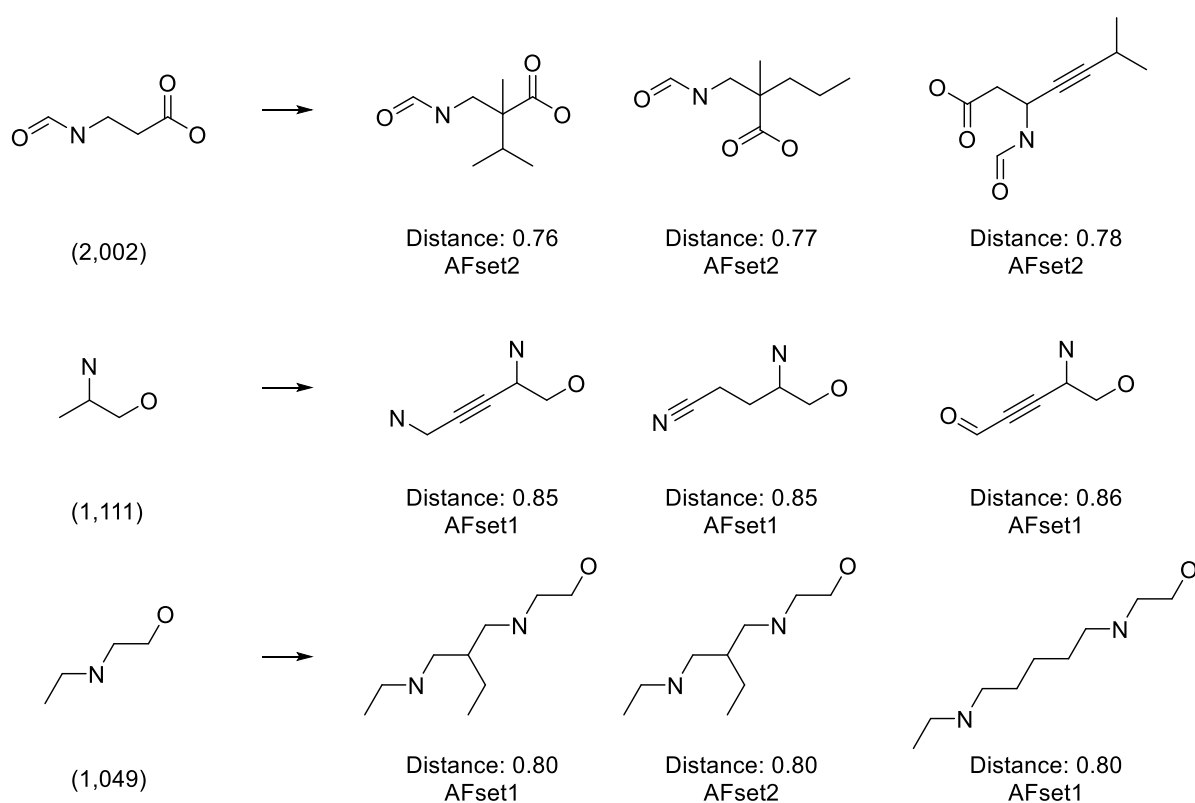


Figure S8. Analogues of highly active ChEMBL AFs found in the subsets of GDB-13s (**AFset1/AFset2**). Total occurrences of the ChEMBL AFs, or the distances between the analogues and the targets are indicated in parentheses.