

# Molecular Framework Analysis of the Generated Database GDB-13s

Ye Buehler and Jean-Louis Reymond\*

## **Abstract**

The generated databases (GDBs) list billions of possible molecules from systematic enumeration following simple rules of chemical stability and synthetic feasibility. To assess the originality of GDB molecules, we compared their Bemis and Murcko molecular frameworks (MFs) with those in public databases. MFs result from molecules by converting all atoms to carbons, all bonds to single bonds, and removing terminal atoms iteratively until none remain. We compared GDB-13s (99,394,177 molecules up to 13 atoms containing simplified functional groups, 22,130 MFs) with ZINC (885,905,524 screening compounds, 1,016,597 MFs), PubChem50 (100,852,694 molecules up to 50 atoms, 1,530,189 MFs) and COCONUT (401,624 natural products, 42,734 MFs). While MFs in public databases mostly contained linker bonds and 6-membered rings, GDB-13s MFs had diverse ring sizes and ring systems without linker bonds. Most GDB-13s MFs were exclusive to this database, and many were relatively simple, representing attractive targets for synthetic chemistry aiming at innovative molecules.

## Introduction

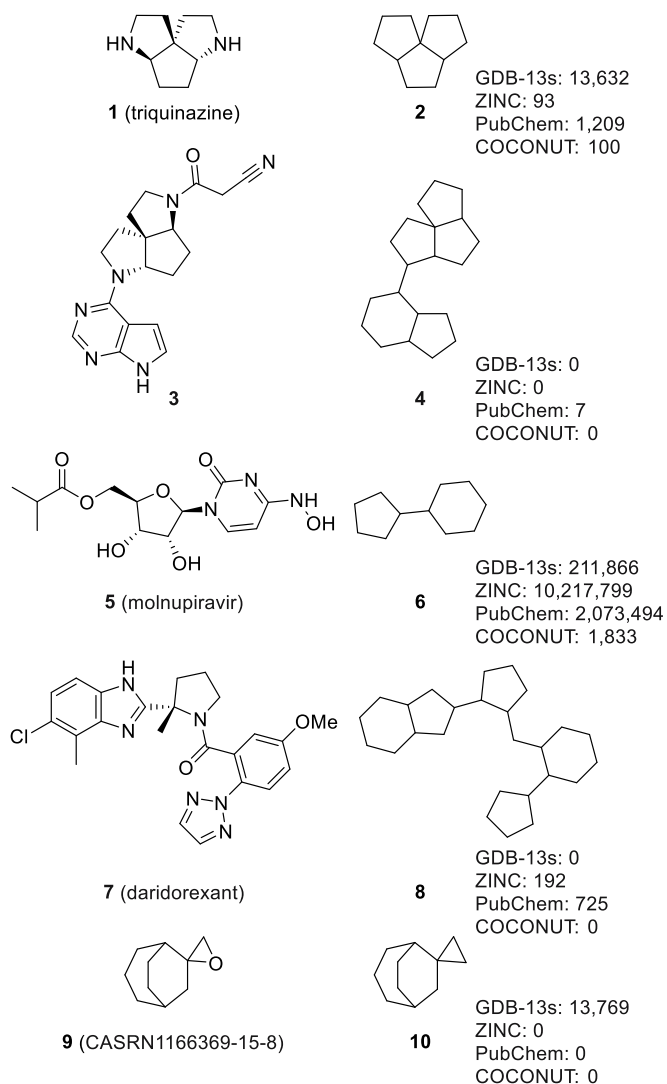
To delineate the chemical space of interest for drug discovery,<sup>1-3</sup> we have reported several generated databases (GDBs) enumerating all possible small molecules up to a given number of non-hydrogen atoms following simple rules of chemical stability and synthetic feasibility.<sup>4,5</sup> These databases contain billions of possible molecules, which are almost all novel since only a few million molecules are known in the size range of the GDBs (up to 17 non-hydrogen atoms).<sup>6</sup> However, defining novelty as non-identity in the context of drug discovery is partly misleading because many similar molecules have comparable properties.<sup>7</sup>

Here we analyze GDB molecules in terms of molecular frameworks (MFs) as proposed by Bemis and Murcko.<sup>8</sup> MFs are the molecular graphs obtained from the structural formula by converting all atoms to carbons, all bonds to single bonds, and removing terminal atoms iteratively until none remain. For our analysis, we considered GDB-13s, a new subset of 99 million possible molecules up to 13 atoms of C, N, O, S and Cl derived from the full GDB-13 (977 million molecules)<sup>9</sup> by restricting allowed functional groups. Although much smaller than GDB-13, GDB-13s contains the complete set of MFs present in GDB-13. We compared MFs of GDB-13s molecules with those derived from 885 million commercially available screening compounds from the ZINC database,<sup>10</sup> from 100 million molecules up to 50 non-hydrogen atoms (heavy atom count HAC  $\leq$  50) in the public database PubChem,<sup>11</sup> and from 400 thousand natural products and natural product-like molecules in the COCONUT database.<sup>12</sup>

MFs define molecular series by their constitutive ring systems and linker bonds compatible with any number of variations in substituents and heteroatoms, and lead to a more demanding definition of novelty.<sup>13,14</sup> For example, our recently reported triquinazine scaffold **1**, inspired from the ring system database GDB4c,<sup>15</sup> represents a new heteroatom variation of the MF of the known angular triquinane **2**, but is not a new MF per se, however the derived Janus kinase inhibitor **3** features an unprecedented MF **4** occurring only in 7 molecules recorded in PubChem which correspond to the record of **3** and related

synthetic intermediate from the original publication (**Figure 1**).<sup>16</sup> Drugs often derive from highly populated MFs, such as molnupiravir (**5**) derived from MF **6** found in most pyrimidine nucleosides and analogs and corresponding to millions of different molecules, including 24 marketed drugs. On the other hand, recently approved drugs, such as the orexin inhibitor daridorexant (**7**), may feature more complex and far less common MFs such as **8**, reflecting the general tendency towards larger and more complex drug structures observed in recent medicinal chemistry trends.<sup>13,14</sup>

As detailed below, we find that, because of the small size of molecules in GDB-13s and the exhaustive enumeration approach taken to create the database, this database features only a relatively few MFs relative to its size compared with ZINC, PubChem and COCONUT. Nevertheless, these MFs are mostly exclusive MF (eMFs) occurring only in GDB-13s and none of the other three analyzed databases, assessing to a vast MF novelty potential in this database. Most remarkably, many eMFs are tricyclic frameworks that should be readily accessible by synthesis. A typical example is the tricyclic MF **10**, for which only a single, non-referenced molecule example is found in Scifinder in form of epoxide **9**.<sup>17</sup>

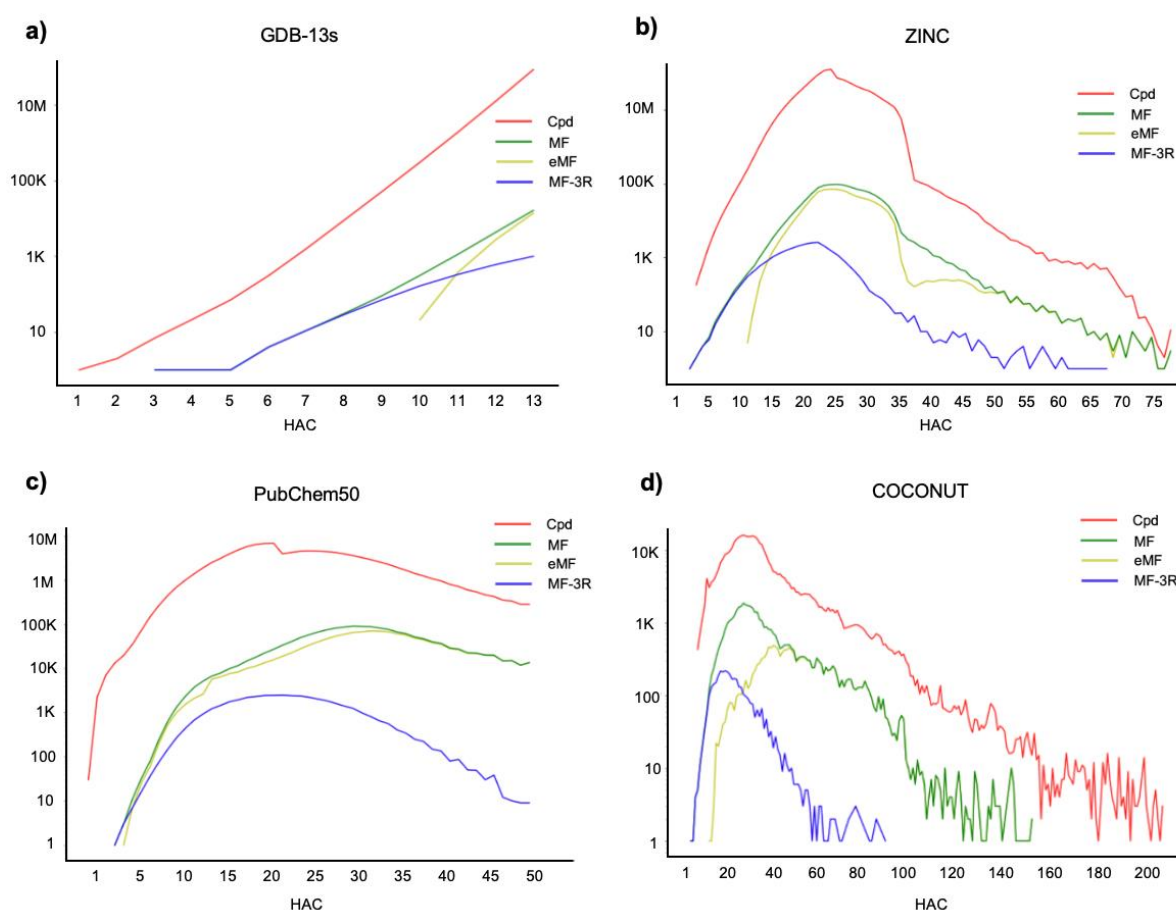


**Figure 1.** Examples of molecules (left) and their constitutive MF (right). The number of occurrences of the MFs in each database is indicated on the right.

## Results and Discussion

**Database selection and MF analysis.** We chose GDB-13 for this analysis because of its manageable size of 977 million molecules. To further restrict the database to molecules resembling those in public databases, we removed functional groups rarely found in medicinal chemistry such as acetals and carbonates, non-aromatic carbon-nitrogen and carbon-carbon double bonds, aziridines, and non-aromatic N-N and N-O bonds. This selection, here named GDB-13s, was reduced by 90% compared to the full GDB-13, further facilitating analysis. Similar to GDB-13, GDB-13s showed an exponential increase in the number of molecules as function of molecule size (**Figure 2a**).

For comparison, we downloaded ZINC, which features 885 million screening compounds available from various providers.<sup>10</sup> Molecules from ZINC are larger than GDB molecules and peak at  $HAC = 26$ , which is a typical drug size, with only very few molecules larger than  $HAC = 36$  (**Figure 2b**). Furthermore, we collected 100 million molecules up to  $HAC = 50$  from PubChem,<sup>11</sup> here named PubChem50. This collection peaked at  $HAC = 21$  but extended more evenly than ZINC up to  $HAC = 50$  (**Figure 2c**). Finally, we considered the recently reported COCONUT, which features 400 thousand natural products or natural product-like molecules.<sup>12</sup> This database is highly populated at  $HAC = 25 - 35$  and contains a few molecules above  $HAC = 100$ , which are mostly glycolipids such as saponins, peptides and polyphenols (**Figure 2d**).



**Figure 2.** Count of molecules (Cpd), molecular frameworks (MF), exclusive molecular frameworks (eMF) and molecular frameworks up to three rings (MF-3R) in GDB-13s (**a**), ZINC (**b**), PubChem50 (**c**) and COCONUT (**d**) as a function of heavy atom count (HAC).

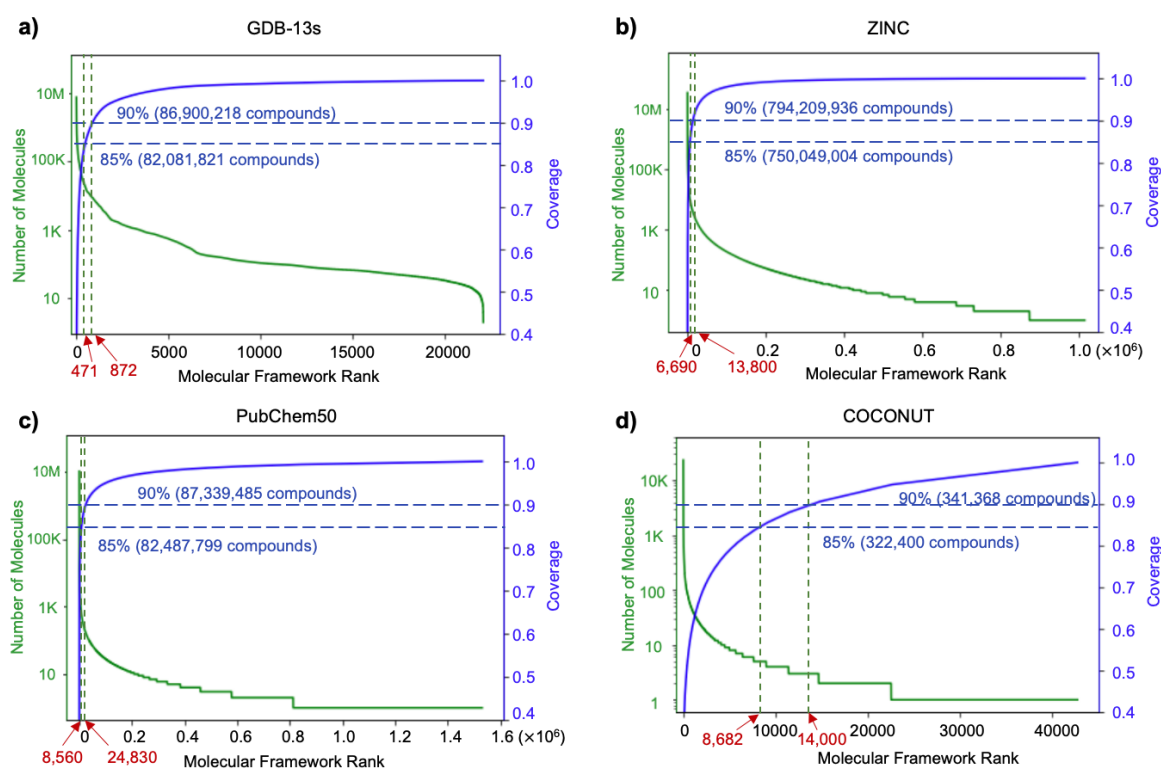
Despite its large size, GDB-13s only contained 22,130 MFs. By contrast, ZINC and PubChem50 both contained over one million MFs, and COCONUT contained 42,734 MFs for only 400 thousand molecules. However, ZINC, PubChem50 and COCONUT contained fewer MFs up to 13 atoms (MF13) than GDB-13s since these databases mostly contain molecules built on MFs larger than 13 atoms (**Table 1a** and **Figure 2a-d**, green lines). The much larger number of molecules per MF in GDB-13s compared to databases of known molecules reflects the exhaustive enumeration approach taken to create the GDB, in contrast to the other three databases collecting known examples. This is also evidenced by the fact that GDB-13s does not contain any MF with only a single molecule example, while 14% of ZINC MFs and 47% of PubChem50 and COCONUT MRs are singletons.

**Table 1.** Molecular framework analysis of GDB-13s, ZINC, PubChem50 and COCONUT.

	<b>GDB-13s</b>	<b>ZINC</b>	<b>PubChem50</b>	<b>COCONUT</b>
<b>a) database size and Cpd/MF</b>				
Cpds	99,394,177	885,905,524	100,852,694	401,624
MF <sup>a)</sup>	22,130	1,016,597	1,530,189	42,734
MF13 <sup>b)</sup>	22,130	1,448	13,422	679
Singletons <sup>c)</sup>	0	141,510	717,917	20,211
% Singletons	0	13.9%	46.9%	47.3%
MF90 <sup>d)</sup>	872	13,800	24,830	14,000
% MF90	3.9%	1.4%	1.6%	32.8%
Cpd/MF90	102,586	57,776	3,656	26
<b>b) MF types</b>				
MF-Ring systems <sup>e)</sup>	17,816	3,841	86,379	6,181
% MF-Ring systems	80.5%	0.4%	0.6%	14.5%
Cpd-Ring systems	96,554,175	73,511,304	21,642,803	136,056
% Cpd-Ring systems	97.1%	8.3%	21.5%	33.9%
MF-5/6 <sup>f)</sup>	3,610	298,901	812,006	24,038
% MF-5/6	16.3%	29.4%	53.1%	56.3%
Cpd-5/6	34,214,845	656,214,620	84,927,019	285,823
% Cpd-5/6	34.4%	74.1%	84.1%	71.2%
<b>c) exclusive MFs</b>				
eMF <sup>g)</sup>	16,936	691,045	1,192,517	16,503
% eMF	76.5%	68.0%	77.9%	38.6%
Cpd-eMF	4,975,340	45,755,635	5,771,217	44,040
% Cpd-eMF	5.0%	5.2%	5.7%	11.0%
<b>d) MFs up to three ring</b>				
MF-3R <sup>h)</sup>	2,215	25,143	40,577	3,670
% MF-3R	10.0%	2.5%	2.7%	8.6%
Cpd-3R	83,472,674	642,704,648	69,406,919	169,647
% Cpd-3R	84.0%	72.5%	68.8%	42.2%
eMF-3R	225	7,794	21,481	317
% eMF-3R	1.0%	0.8%	1.4%	0.7%
Cpd-eMF-3R	209,011	1,648,939	139,368	841
% Cpd-eMF-3R	0.2%	0.2%	0.1%	0.2%

a) MF = molecular framework; b) MF13 = MF up to 13 atoms; c) Singletons = MF with only a single molecule example; d) MF90 = no. of MF covering 90% of the database; e) MF-Ring systems = MF without acyclic bonds; f) MF-5/6 = MF containing only 5- or 6-membered rings; g) eMF = exclusive MF, does not occur in the other three databases; h) MF-3R = MF up to three rings.

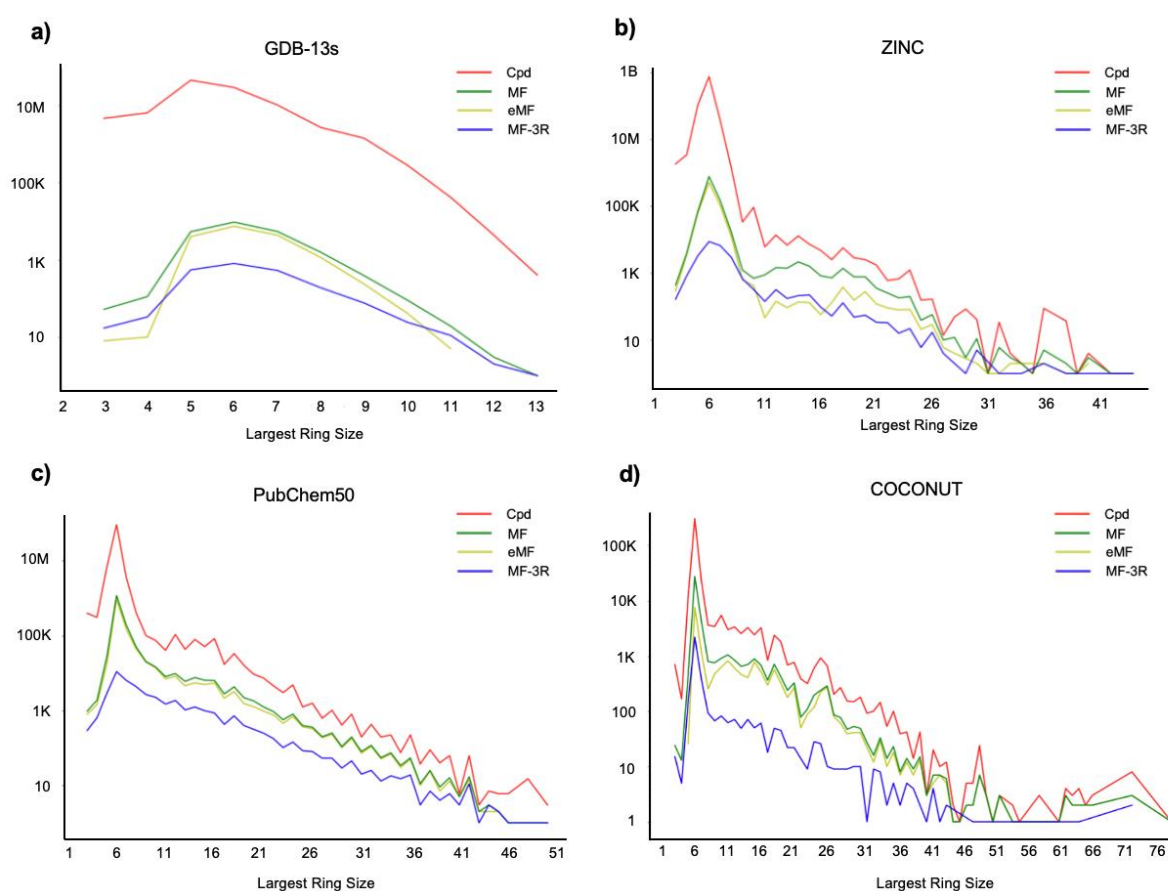
The frequency of molecules per MF followed a typical power law distribution in all four databases (**Figure 3**). This distribution was steepest in ZINC and PubChem 50, where approximately 1.5% of all MFs were sufficient to cover 90% of the database, defined here as MF90 (**Table 1a**). GDB-13s required 3.9% of its MFs to cover 90% of the database, however in this case the number of molecules per MF was higher than in ZINC or PubChem50 due to the lower total number of MF in GDB-13s. A similar coverage of 90% in COCONUT required 32.8% of its MFs, reflecting the large MF diversity of this natural product collection with an average of only 26 compounds per MF90.



**Figure 3.** Frequency distribution of MFs in GDB-13s (a), ZINC (b), PubChem50 (c) and COCONUT (d).

In terms of structural types, the majority of the MFs in GDB-13s (80.5%) were ring systems, which are MFs without any linker bonds, and these ring systems made up almost the entire database (97.1% of all molecules, **Table 1b**). In sharp contrast, the other databases were dominated by MFs containing linker bonds, such that ring systems only composed a small fraction of MFs and molecules in ZINC (0.4% MFs, 8.3% molecules), PubChem50 (0.6% MFs, 21.5% molecules) and COCONUT (14.5% MFs, 33.9% molecules). Furthermore, MFs and molecules containing only 5- or 6-membered

rings were a minority in GDB-13s (16.3% MFs, 34.4% molecules), but made up a much larger fraction of ZINC (29.4% MFs, 74.1% molecules) and dominated in PubChem (53.1% MFs, 84.1% molecules) and COCONUT (56.3% MFs, 71.2% molecules), probably reflecting the fact that 5- and 6-membered rings are easily formed and synthesized. Frequency histograms as function of the largest ring size in fact showed that 5-membered rings were most prevalent in GDB-13s molecules while 6-membered rings dominated in GDB-13s MFs as well as in both molecules and MFs for ZINC, PubChem50 and COCONUT (**Figure 4**).










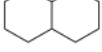
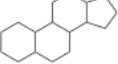


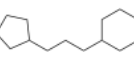
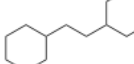

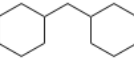
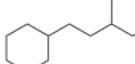

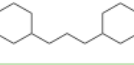
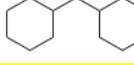
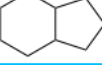
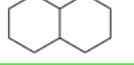
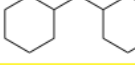

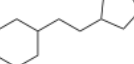
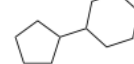
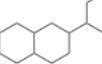

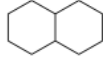

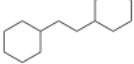
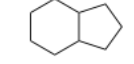

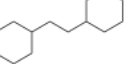
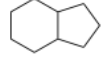

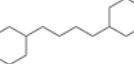
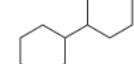
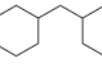
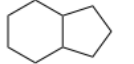
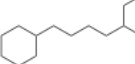

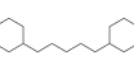
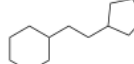
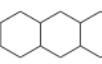
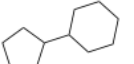
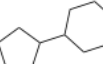

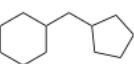
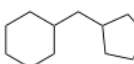
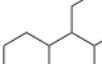
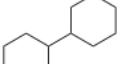
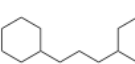

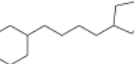
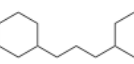
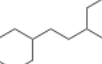
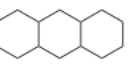
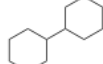


**Figure 4.** Largest ring size histogram of molecules (Cpd), molecular frameworks (MF), exclusive molecular frameworks (eMF) and molecular frameworks up to three rings (MF-3R) in GDB-13s (**a**), ZINC (**b**), PubChem50 (**c**) and COCONUT (**d**).

The relative importance of ring sizes in each database was further illustrated by analyzing the 10 most populated MFs in each database (**11 – 35, Figure 5**). The most populated MF in GDB-13s was cyclopentane (**11**) with 7.3 million molecules, followed by cyclobutane (**12**), cyclopropane (**13**) and



bicyclic fused ring systems (**14** – **17**), with cyclohexane (**18**) appearing in position 8 with 2.3 million molecules. Furthermore, six of the top-10 MFs in GDB-13s (**12**, **13**, **15**, **16**, **19**, **20**) contained a small (3- or 4-membered) ring. By contrast, ZINC, PubChem50 and COCONUT all featured cyclohexane (**18**) as the most populated MF, followed by cyclopentane (**11**) for ZINC and PubChem50 and decalin (**31**) for COCONUT. All top-10 MFs in these databases only contained 5- and 6-membered rings, and were very comparable to the top-10 MFs in CMC as reported by Bemis and Murcko<sup>8</sup> and in the CAS Registry Organic Subset as reported by Lipkus.<sup>18</sup> A similar pattern appeared when considering the top-30 MFs in each set (**Figure S1**).

	GDB-13s	ZINC	PubChem50	COCONUT	CMC	CAS Registry-Organic Subset
	<b>11</b> (7,336,582)					
	<b>12</b> (3,769,528)					
	<b>13</b> (3,386,077)					
	<b>14</b> (3,364,772)					
	<b>15</b> (2,823,509)					
	<b>16</b> (2,412,763)					
	<b>17</b> (2,390,989)					
	<b>18</b> (2,314,855)					
	<b>19</b> (1,737,621)					
	<b>20</b> (1,438,959)					

**Figure 5.** Top-10 most populated MFs in various databases. MFs are numbered by order of appearance in the frequency sorted list across the four databases. The top-30 most populated MFs in various databases are shown in **Figure S1**.

### Exclusive molecular frameworks

To appreciate the uniqueness of each database, we next analyzed which MFs were found only in one of the four databases, here named exclusive MFs (eMFs, **Table 1c**). The much larger number of MF13 (MFs up to 13 atoms) in GDB-13s compared to the other databases ensured that these were mostly eMFs. Indeed 76.5% of MFs in GDB-13s were eMFs. Nevertheless, a comparable percentage of eMFs were present in ZINC (68.0%) and PubChem50 (77.9%), while only 38.6% were eMFs in COCONUT.

Note that eMFs were generally less populated, and the corresponding molecules only made up to approximately 5% of the database for GDB-13s, ZINC and PubChem50, and 10% for COCONUT. A Venn diagram analysis showed that GDB-13s mostly shared MFs with PubChem50, while the overlap with ZINC and COCONUT was much smaller (**Figure 6a**). In all four databases, the most populated eMFs only comprised thousands of molecules, as opposed to up to millions for MFs. eMFs were also generally more complex, featuring polycyclic systems with mostly four or more rings, as illustrated by the two most populated eMFs in each of the four databases analyzed (**36 – 43, Figure 6b**). A similar pattern was visible when surveying the top-10 most populated eMFs (**Figure S2**).

### Molecular frameworks up to three rings

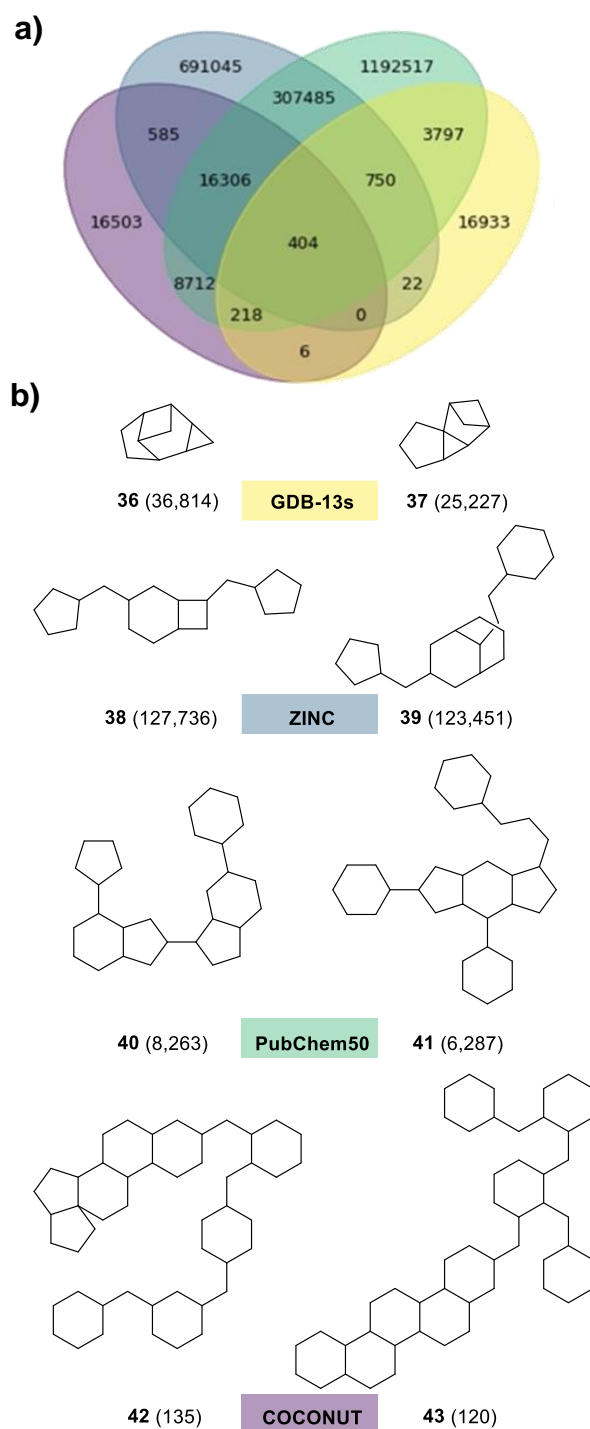
Because the most populated MFs from each of the four databases featured at most three rings, we investigated which percentage of the databases were in fact from MFs with only up to three rings, here named MF-3R, considering all MFs as well as eMFs for each database (**Table 1d**). In line with the most populated MFs, 84% of the molecules in GDB-13s stemmed from MF-3R, although these only made up 10.9% of all MFs. A similar and even more extreme situation in terms of MFs occurred in ZINC (2.5% MF-3R result in 72.5% molecules), PubChem (2.7% MF-3R result in 68.8% molecules) and COCONUT (8.6% MR-3R result in 42.2% molecules). The frequency of molecules with only few rings most likely results from their easier synthesis compared to molecules derived from more complex MFs.

A Venn diagram analysis showed that only very few of MF-3R were exclusive to each database (**Figure 7a**). Furthermore, only approximately 1% all MFs were eMF-3R, and only 0.1% of all

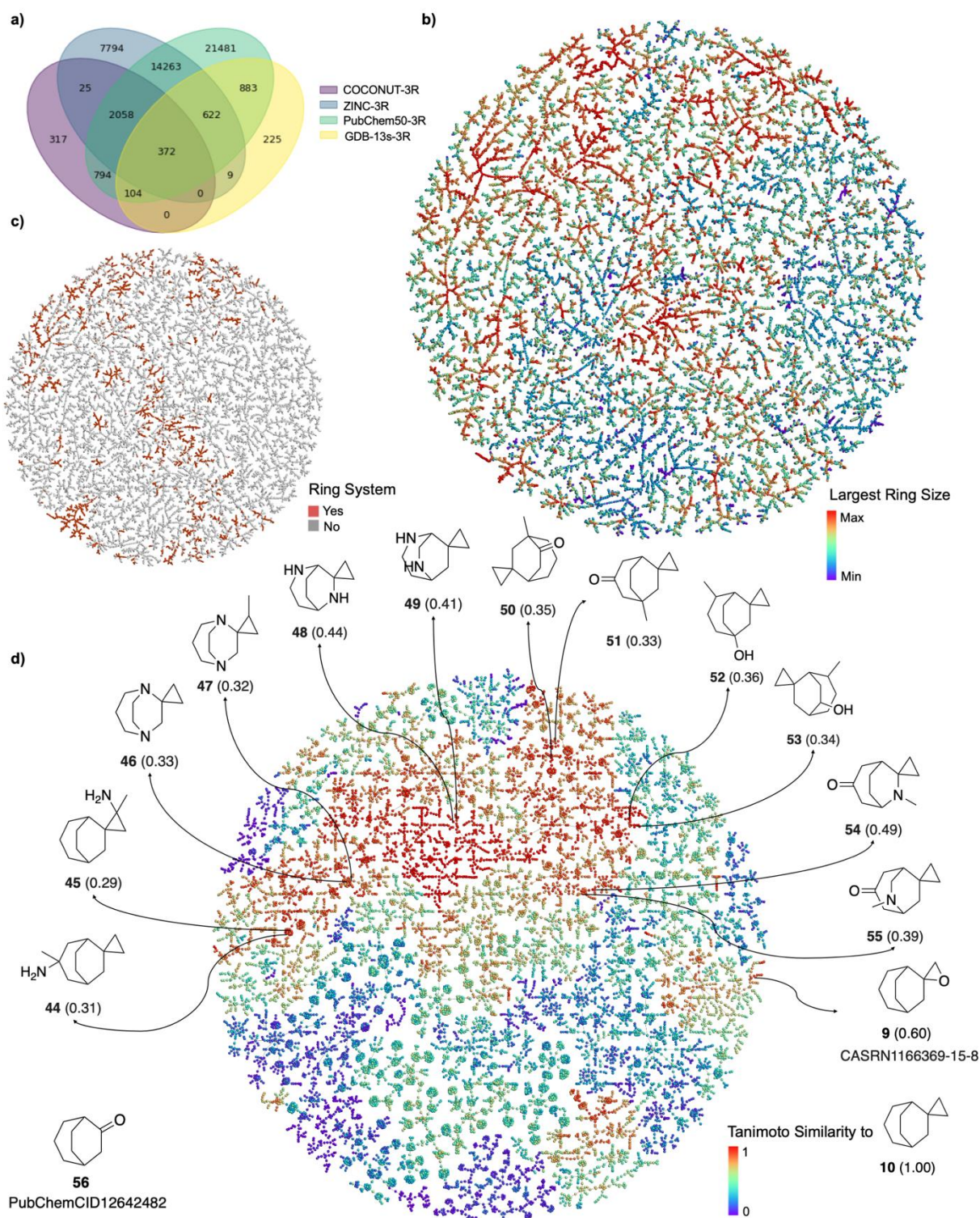
molecules stemmed from eMF-3R in each database (**Table 1d**). Due to database sizes however, this still left a good number of molecules from eMF-3R in each database (>200,000 for GDB-13s and >1,000,000 for ZINC).

In terms of identifying original yet simple molecules, those built from eMF-3R should be the most interesting. To gain an overview of such frameworks, we built a tree-map (TMAP)<sup>19</sup> of MF-3R to visualize their diversity across the different databases. Color-coding the TMAP by the size of the largest ring showed that macrocycles made up a significant fraction of MF-3R ( $\geq 12$ -membered ring, 18.9%), while MF with only small rings (3- or 4-membered) only accounted for 3.0% of MF-3R (**Figure 7b**). Furthermore, color-coding by MF-type (ring system or MF with linker bonds) showed that 14.2% of MF-3R were ring systems (**Figure 7c**).

Although MF group molecules sharing a common structural feature, the diversity accessible from a single MF can be quite substantial, as illustrated by the TMAP displaying the 13,769 possible GDB-13s molecules sharing MF **10**, which is the most populated eMF-3R in GDB-13s (**Figure 7d**). These molecules span the full range of functional groups allowed in GDB-13s and the Tanimoto similarities (Tan) to the parent MF calculated using the standard ECFP4 fingerprint range from almost identical (Tan  $\sim 1$ ) to almost entirely dissimilar (Tan  $\sim 0$ ). Among these molecules, one can readily identify possible analogs of well-known 3D-shaped molecules, such as memantine (**44, 45**), DABCO (**46, 47**), triquinazine **1** (**48, 49**), camphor (**50, 51**), patchoulol (**52, 53**), and tropinone (**54, 55**). Although our definition of eMFs is limited to the comparison of the four databases considered, most eMF in GDB-13s are indeed novel upon checking for novelty in Scifinder.<sup>17</sup> For example, MF **10** contains only a single entry in Scifinder in form of the epoxide **9**, however without any literature reference. This epoxide can probably be synthesized from the parent ketone **56**, which is listed in PubChem.



**Figure 6.** Exclusive molecular frameworks. **(a)** Venn diagram of MF in the different databases. **(b)** The top-2 most populated eMFs in the different databases. The top-10 eMFs in the different databases are shown in **Figure S2**.



**Figure 7.** Molecular framework up to three rings (MF-3R). (a) Venn diagram of MF-3R in the different databases. (b) Tree map (TMAP) visualization of MF-3R in GDB-13s, ZINC, PubChem and COCONUT color-coded by MR-3R size of the largest ring. (c) TMAP color-coded by MF type. (d) TMAP of the 13,769 molecules derived from the most frequent eMF-3R in GDB-13s. An interactive version of the TMAPs with additional color-codes is accessible at <https://tm.gdb.tools/map4> ([MAP4\\_4databases\\_MF3R](#); [MAP4\\_GDB-13s\\_eMF3R\\_Cpd](#)).



## **Conclusion**

The analysis above shows that, despite the large size of GDB-13s, the absolute number of different MFs in GDB-13s is quite low compared to collections such as ZINC, PubChem or COCONUT. In contrast to these collections which contain mostly MFs with 5- and 6-membered rings and including linker bonds, most MFs in GDB-13s feature a broader variety of ring sizes and are ring systems without any acyclic bonds. Most interestingly, many MFs occur only in GDB-13s (eMFs) and feature unprecedented ring combinations. Such eMFs might be the most relevant targets for synthetic chemistry aiming at innovative molecules.

## **Methods**

### **GDB-13s Generation**

The entire GDB-13 (including all C/N/O/Cl/S molecules) dataset was downloaded from our group website (<https://gdb.unibe.ch/downloads>). 977,468,301 entries of the GDB-13 database were filtered by Python programming. Functional groups or substructures were identified by using the Daylight SMARTS language<sup>20</sup>. AlogP (Atomic logP) values using Ghose/Crippen method<sup>21</sup> were calculated by using RDKit<sup>22</sup>. Five rules have been applied to the entire GDB-13 database as follows in order:

- 1) C=O filtration: Only keep the molecules with a double bond as C=O in non-aromatic structures so as to phase out molecules with non-aromatic C=C and C=N. For aromatic rings, all types of double bonds are allowed. There is no restriction for the molecules without any C=O double bonds;
- 2) AlogP filtration: AlogP is the refinement of LogP, it suits smaller molecules. If the AlogP value of a drug is too low, the drug molecule will hardly pass through the cell membrane. In this context, we use this filtration to remove all the molecules with an AlogP value less than 0;
- 3) N-O, N-N in non-aromatic ring filtration: Exclude all N-O and N-N bonds from non-aromatic rings (both atoms are inside the aromatic ring);

- 4) O-C-O filtration: Filter out the molecules containing O-C-O structures;
- 5) N in three-member ring filtration: Eliminate the compounds containing three-member rings with any nitrogen atoms.

GDB-13s can be downloaded from: <https://gdb.unibe.ch/downloads>.

## Data Collection

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 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 PubChem50 database. The COCONUT data adopted in this study is the February 2021 version (<https://github.com/reymond-group/Coconut-TMAP-SVM>). SMILES strings served as inputs and RDKit package was regarded as the main tool for HAC and other properties calculation.

## Molecular Framework Model

The Molecular Framework Model is written in Python 3 and is now distributed on GitHub under the MIT license ([https://github.com/Ye-Buehler/Molecular\\_Framework\\_Model](https://github.com/Ye-Buehler/Molecular_Framework_Model)). It is dependent on several freely available Python packages such as Pandas<sup>23</sup>, Numpy<sup>24</sup> and RDKit. A brief outline of the model is provided here: A molecule as an input will be firstly simplified by converting all its bonds into single bonds and converting all its atoms into carbon atoms. Then all terminal atoms of this molecule will be removed iteratively. The outcome will be a molecular framework as defined by Bemis and Murcko.<sup>8</sup>

## Venn Diagrams and TMAPs

Venn diagrams were computed by using the freely available Python package Venn.<sup>25</sup> TMAPs were generated by specifying standard parameters,<sup>19</sup> and all utilized the MAP4 fingerprint (MinHashed atom-

pair fingerprint up to a diameter of four bonds),<sup>26</sup> which is our lately developed fingerprint suitable for universal classes of molecules, especially preferable for natural product molecules. MAP4 fingerprints were computed with a dimension of 256.

### **Author Information**

#### **Corresponding Author**

**Jean-Louis Reymond** - *Dept. 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** - *Dept. of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

#### **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.

#### **Notes**

The authors declare no competing financial interest.

#### **Acknowledgements**

This work was supported financially by the University of Bern and the Swiss National Science Foundation (Grant no. 200020\_207976). We thank 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.

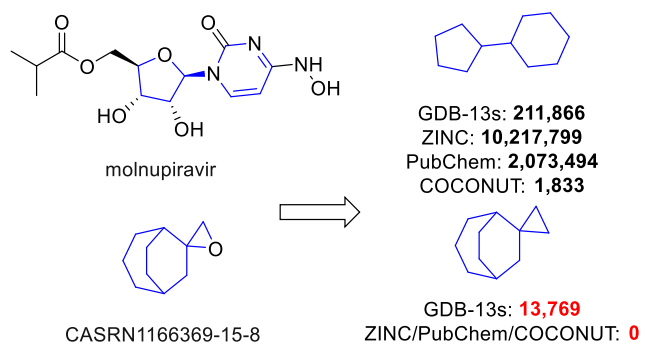


## References

- (1) Mullard, A. The Drug-Maker's Guide to the Galaxy. *Nature* **2017**, *549* (7673), 445–447. <https://doi.org/10.1038/549445a>.
- (2) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discovery Today* **2019**, *24* (5), 1148–1156. <https://doi.org/10.1016/j.drudis.2019.02.013>.
- (3) 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>.
- (4) Reymond, J.-L.; Ruddigkeit, L.; Blum, L.; Deursen, R. van. The Enumeration of Chemical Space. *WIREs Comput. Mol. Sci.* **2012**, *2* (5), 717–733. <https://doi.org/10.1002/wcms.1104>.
- (5) 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), 241–246. <https://doi.org/10.2533/chimia.2020.241>.
- (6) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. <https://doi.org/10.1021/ci300415d>.
- (7) Wermuth, C. G. Similarity in Drugs: Reflections on Analogue Design. *Drug Discovery Today* **2006**, *11* (7–8), 348–354.
- (8) 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>.
- (9) 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.
- (10) 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**. <https://doi.org/10.1021/acs.jcim.0c00675>.
- (11) Gindulyte, A.; Shoemaker, B. A.; Yu, B.; He, J.; Zhang, J.; Chen, J.; Zaslavsky, L.; Thiessen, P. A.; Li, Q.; He, S.; Kim, S.; Cheng, T.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2018**, *47* (D1), D1102–D1109. <https://doi.org/10.1093/nar/gky1033>.
- (12) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminf.* **2021**, *13* (1), 2. <https://doi.org/10.1186/s13321-020-00478-9>.
- (13) Lipkus, A. H.; Watkins, S. P.; Gengras, K.; McBride, M. J.; Wills, T. J. Recent Changes in the Scaffold Diversity of Organic Chemistry As Seen in the CAS Registry. *J. Org. Chem.* **2019**, *84* (21), 13948–13956. <https://doi.org/10.1021/acs.joc.9b02111>.

- (14) Wills, T. J.; Lipkus, A. H. Structural Approach to Assessing the Innovativeness of New Drugs Finds Accelerating Rate of Innovation. *ACS Med. Chem. Lett.* **2020**, *11* (11), 2114–2119. <https://doi.org/10.1021/acsmchemlett.0c00319>.
- (15) Visini, R.; Arus-Pous, J.; Awale, M.; Reymond, J. L. Virtual Exploration of the Ring Systems Chemical Universe. *J. Chem. Inf. Model.* **2017**, *57*, 2707–2718.
- (16) 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. Engl.* **2021**, *60* (4), 2074–2077. <https://doi.org/10.1002/anie.202012049>.
- (17) *SciFinder: Substances Search Service*. <https://scifinder.cas.org> (accessed 2022-08-25).
- (18) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F.; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73* (12), 4443–4451. <https://doi.org/10.1021/jo8001276>.
- (19) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminf.* **2020**, *12* (1), 12. <https://doi.org/10.1186/s13321-020-0416-x>.
- (20) *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns*. <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed 2022-07-25).
- (21) Souza, E. S.; Zaramello, L.; Kuhnen, C. A.; Junkes, B. da S.; Yunes, R. A.; Heinzen, V. E. F. Estimating the Octanol/Water Partition Coefficient for Aliphatic Organic Compounds Using Semi-Empirical Electrotopological Index. *Int J Mol Sci* **2011**, *12* (10), 7250–7264. <https://doi.org/10.3390/ijms12107250>.
- (22) *RDKit: Open-source cheminformatics*. <http://www.rdkit.org> (accessed 2022-07-25).
- (23) McKinney, W. *Data Structures for Statistical Computing in Python*; Austin, Texas, 2010; pp 56–61. <https://doi.org/10.25080/Majora-92bf1922-00a>.
- (24) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
- (25) *Pyven: Venn diagrams for 2, 3, 4, 5, 6 sets*. <https://pypi.org/project/venn> (accessed 2022-07-20).
- (26) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminf.* **2020**, *12* (1), 43. <https://doi.org/10.1186/s13321-020-00445-4>.

## Graphics for the Table of Contents:













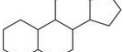


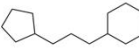
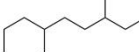
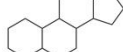
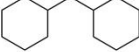
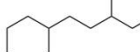
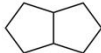
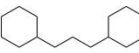
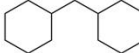
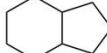

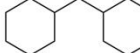

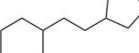

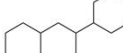



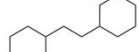
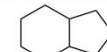

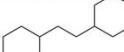

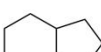
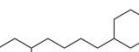
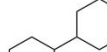
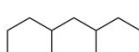
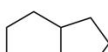
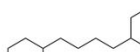

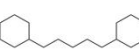
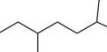
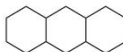
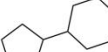
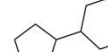

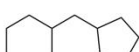
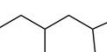
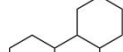
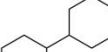


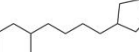
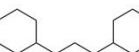
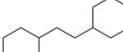
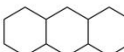




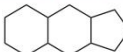
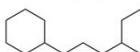
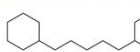

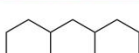
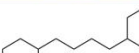
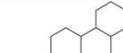
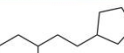
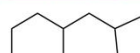
## Supporting Information for

### Molecular Framework Analysis of the Generated Database GDB-13s

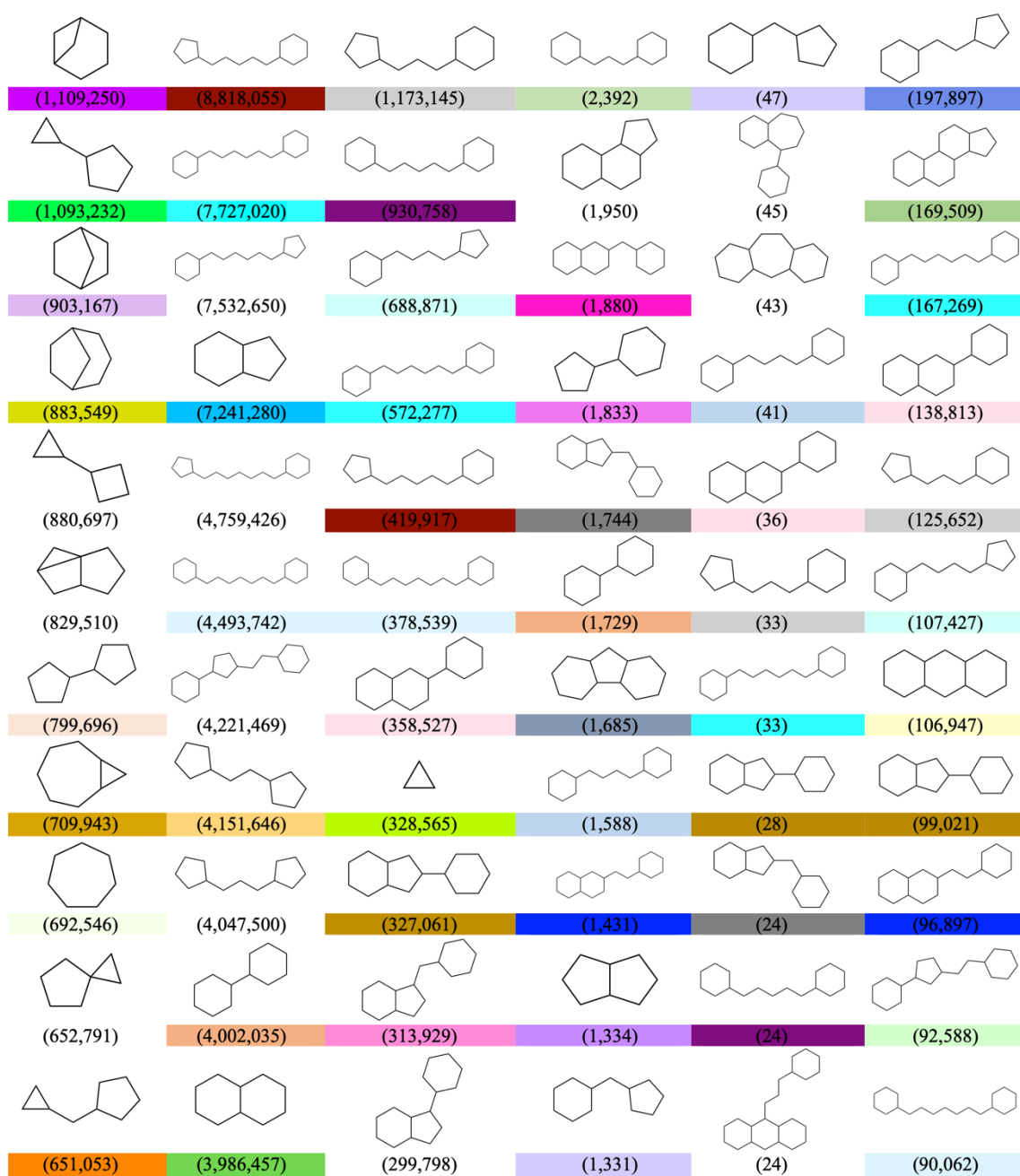
Ye Buehler and Jean-Louis Reymond\*

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

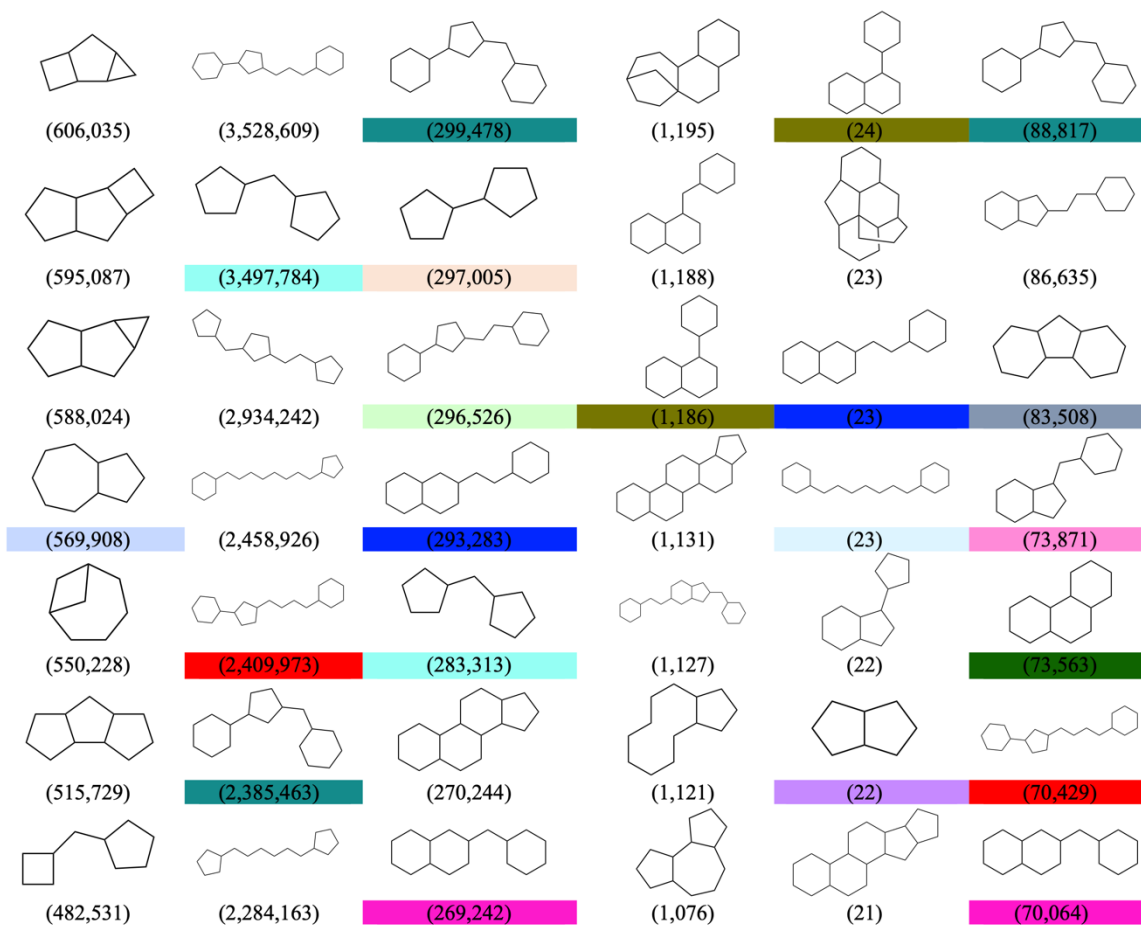
*\*E-Mail: jean-louis.reymond@unibe.ch.*

GDB-13s	ZINC	PubChem50	COCONUT	CMC	CAS Registry-Organic Subset
 (7,336,582)	 (36,134,272)	 (10,651,372)	 (23,365)	 (606)	 (2,335,116)
 (3,769,528)	 (16,288,719)	 (3,144,161)	 (12,582)	 (247)	 (640,741)
 (3,386,077)	 (15,153,460)	 (2,786,150)	 (7,976)	 (195)	 (581,087)
 (3,364,772)	 (14,471,562)	 (2,705,427)	 (7,175)	 (142)	 (471,386)
 (2,823,509)	 (13,650,965)	 (2,073,494)	 (6,188)	 (129)	 (456,212)
 (2,412,763)	 (12,911,391)	 (2,005,437)	 (6,170)	 (119)	 (453,247)
 (2,390,989)	 (12,112,387)	 (1,792,666)	 (4,743)	 (119)	 (370,322)
 (2,314,855)	 (11,076,861)	 (1,779,575)	 (4,004)	 (116)	 (361,675)
 (1,737,621)	 (10,616,792)	 (1,620,038)	 (3,887)	 (108)	 (338,103)
 (1,438,959)	 (10,319,936)	 (1,581,801)	 (3,422)	 (81)	 (324,554)
 (1,408,164)	 (10,217,799)	 (1,488,390)	 (3,232)	 (55)	 (198,673)
 (1,204,793)	 (9,081,289)	 (1,325,261)	 (2,579)	 (57)	 (197,990)

**Figure S1.** The top-30 most populated MFs in GDB-13s, ZINC, PubChem50, COCONUT, CMC and CAS Registry-Organic Subset.

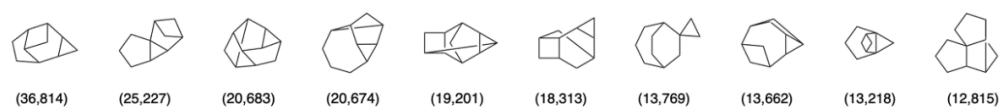


**Figure S1.** The top-30 most populated MFs in GDB-13s, ZINC, PubChem50, COCONUT, CMC and CAS Registry-Organic Subset (Continued).

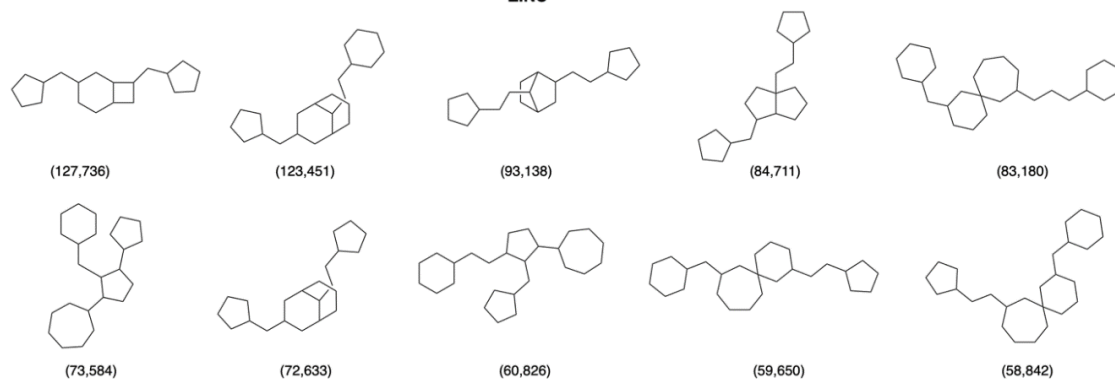


**Figure S1.** The top-30 most populated MFs in GDB-13s, ZINC, PubChem50, COCONUT, CMC and CAS Registry-Organic Subset (Continued).

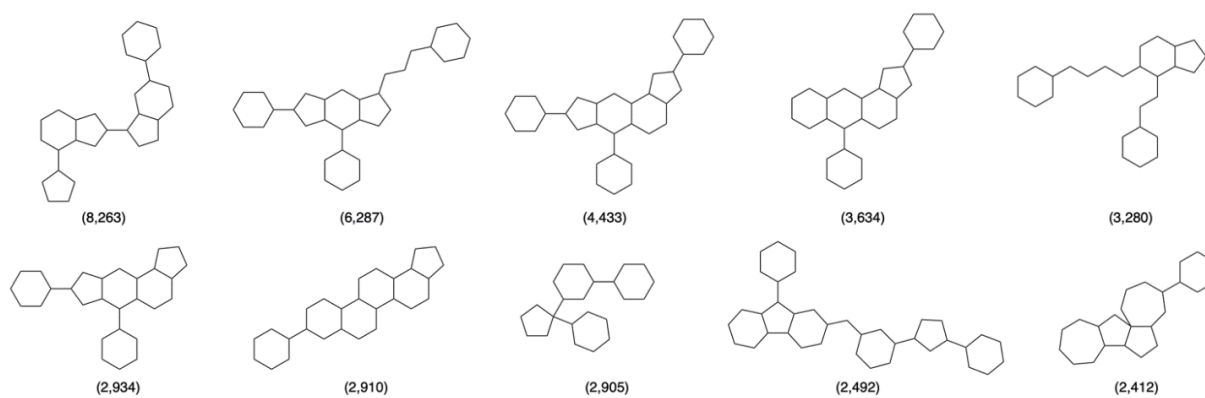
## GDB-13s



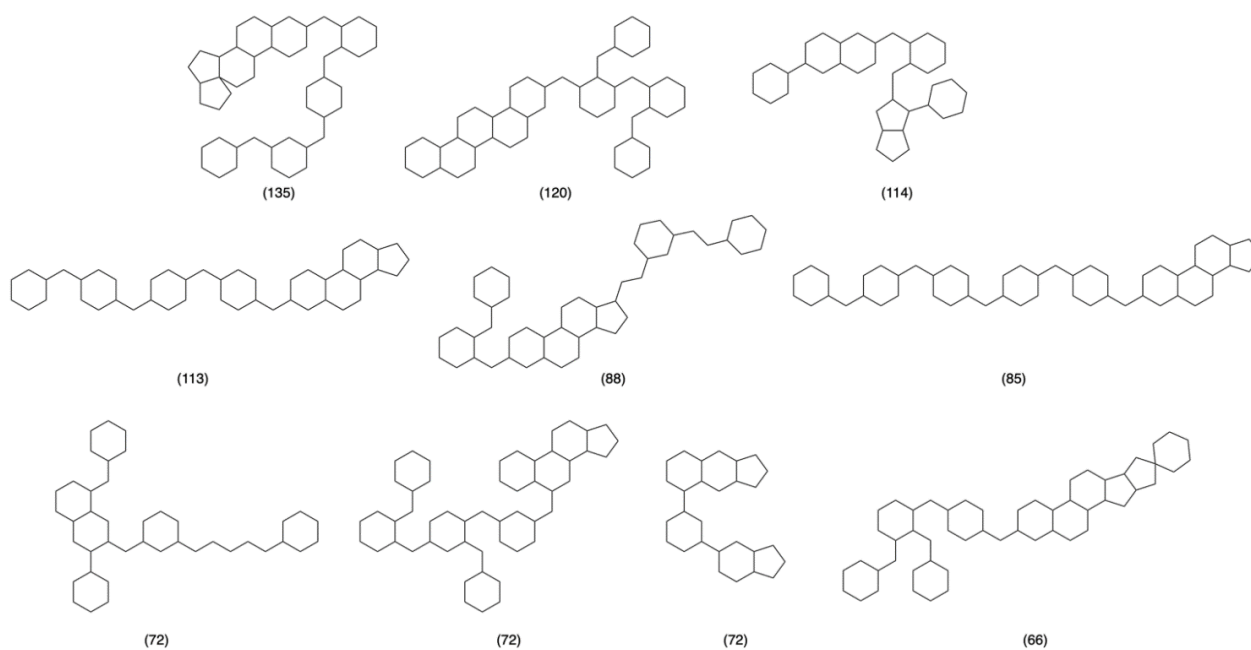
## ZINC



## PubChem50



## COCONUT



**Figure S2.** The top-10 eMFs in GDB-13s, ZINC, PubChem50 and COCONUT.