ACS Editors' Choice

Journal of
Chemical Information
and Modeling

Article

Cite This: J. Chem. Inf. Model. XXXX, XXX, XXX-XXX

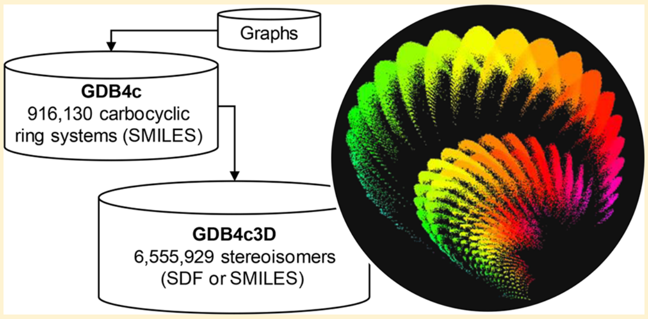
pubs.acs.org/jcim

Virtual Exploration of the Ring Systems Chemical Universe

Ricardo Visini,† Josep Arús-Pous,† Mahendra Awale, and Jean-Louis Reymond*

Department of Chemistry and Biochemistry, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland

Supporting Information


Graphs
GDB4c
916,130 carbocyclic ring systems (SMILES)
GDB4c3D
6,555,929 stereoisomers (SDF or SMILES)

ABSTRACT: Here, we explore the chemical space of all virtually possible organic molecules focusing on ring systems, which represent the cyclic cores of organic molecules obtained by removing all acyclic bonds and converting all remaining atoms to carbon. This approach circumvents the combinatorial explosion encountered when enumerating the molecules themselves. We report the chemical universe database GDB4c containing 916 130 ring systems up to four saturated or aromatic rings and maximum ring size of 14 atoms and GDB4c3D containing the corresponding 6 555 929 stereoisomers. Almost all (98.6%) of these ring systems are unknown and represent chiral 3D-shaped macrocycles containing small rings and quaternary centers reminiscent of polycyclic natural products. We envision that GDB4c can serve to select new ring systems from which to design analogs of such natural products. The database is available for download at www.gdb.unibe.ch together with interactive visualization and search tools as a resource for molecular design.

INTRODUCTION

Innovation at the level of molecular structures is essential to the progress of chemistry.[1] While over 100 million organic molecules have already been synthesized, many more are in principle possible, perhaps as many as 10^{60} below 500 Da.[2−5] Beyond simple counting, the development of cheminformatics methods such as SMILES to write 2D-molecular information in compact format,[6] and 3D-generators to convert 2D-structures to 3D-models considering all possible stereoisomers and conformers,[7,8] has made it possible to explicitly enumerate millions of virtual molecules.[9] While most algorithms produce lists of possible molecules following specific criteria on demand,[10−17] we have gained global insight into the entire chemical universe by enumerating all possible molecules up to a given size following simple rules of chemical stability and synthetic feasibility.[18−22] This strategy succeeded up to the Generated DataBase GDB-17 containing 166.4 billion molecules up to 17 atoms of C, N, O, S, and halogens.[23,24]

By visualizing GDB-17 in various property spaces, we found that the chemical universe up to 17 atoms is mostly populated by novel chiral 3D-shaped molecules, among which bioactive compounds can be identified by virtual screening, synthesis, and testing.[25−34] However, we were unable to enumerate molecules beyond 17 atoms due to the combinatorial explosion of possibilities. Herein, we report an approach to overcome this limitation by focusing on ring systems, which represent the cyclic cores of organic molecules obtained by removing all acyclic bonds and converting all remaining atoms to carbon, as defined by Bemis and Murcko.[35] Enumerating ring systems leaves aside molecules arising from combinations of smaller fragments such as those in virtual libraries listing the coupling products of known building blocks with known reactions.[36−42] Considering only a single representative for each ring system furthermore reduces database size to manageable numbers while retaining a central aspect of the structural identity of molecules and therefore of their potential for novelty. Our enumeration reveals that known ring systems as well as previously reported collections of aromatic and heteroaromatic rings[43,44] only form a very small part of the chemical universe of ring systems, which is strongly dominated by chiral and 3D-shaped macrocycles. Such ring systems represent a defining feature of natural products and an opportunity for expanding chemistry into novel chemical space.[45−47]

RESULTS AND DISCUSSION

Database Assembly. We started our enumeration with the 3 282 214 777 graphs listed by the program GENG[48] set to produce all possible planar graphs up to 16 nodes with degree two, three, or four, which was large enough to cover all topological possibilities of tetracyclic ring systems yet small enough for convenient handling. Next, we filtered this output to retain only the 93 463 graphs up to four rings without acyclic edges and converted them to saturated hydrocarbon SMILES. We then used an iterative ring enlargement algorithm inserting methylene groups in each single bond[16] to exhaust the possibilities under a set of maximum ring size criteria. This procedure resulted in a total of 728 391 cyclic hydrocarbons up to 30 atoms, which was the maximum size allowed by the enumeration rules. The 151 largest ring systems allowed by these rules featured all possible combinations of one 14-membered, one seven-membered, and two six-membered rings

Received: July 31, 2017
Published: October 11, 2017

connected by spiro centers. By contrast to our previous GDB databases where many rules were applied to restrict ring strain, the only rule used here was to remove tetracyclic ring systems where an atom is shared by two small rings, which are frequent but highly strained structures. We did not apply this rule up to tricyclic ring systems because they form only a very small part of the database, and such unusually strained ring systems might be of interest for applications of our database related to quantum chemistry calculations.[49]

We further diversified the ring systems by combinatorially aromatizing all five- and six-membered rings, eliminating small rings fused to aromatic rings in the process, which added 187 740 ring systems containing one to four aromatic rings. For five-membered aromatic rings, we introduced a nitrogen atom at the first possible position in the ring to form a pyrrole as one of several possible chemically allowed forms of five-membered aromatic rings, such as to obtain molecules that can be properly handled by cheminformatics software. The complete database of 916 130 ring systems was named GDB4c and stored as SMILES (Figure 1).

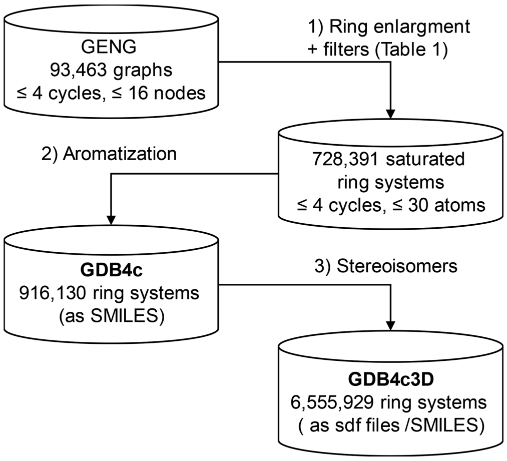


Figure 1. Assembly procedure for GDB4c and GDB4c3D.

We also generated a second database, GDB4c3D, listing the 3D-structures of our ring systems. We succeeded in converting 895 375 2D ring systems (97.7% of GDB4c) to 3D structures representing 6 535 174 individual stereoisomers using the program CORINA,[7] which enumerates all possible stereoisomers of any given 2D molecule. This corresponds to an average of 7.3 stereoisomers per 2D ring system. Additionally, we found that the ChemAxon 3D builder with "fine" cleanup mode was able to generate one 3D structure each for 20 755 of the remaining 20 756 2D ring systems which had not been accepted by CORINA. We also used the ChemAxon 3D builder to optimize the geometry of the 6 535 174 3D structures produced by CORINA and correct for some inaccuracies in bond angles. The total size of the resulting GDB4c3D database was 6 555 929 structures, stored in SDF and chiral SMILES format. The database comprised 2 736 737 pairs of enantiomers and 1 082 455 achiral ring systems.

For comparison purposes, we assembled a reference database of known ring systems starting from all organic molecules in the public databases DrugBank,[50] ChEMBL,[51] SureChEMBL, ZINC,[52] PubChem,[53] and Reaxys.[54] For each molecule, we removed all acyclic atoms and bonds and converted all remaining cyclic atoms to carbon and all nonaromatic bonds to single bonds, which produced one or more ring systems for each molecule. Five-membered aromatic rings were handled the same way as when generating GDB4c, converting them all into pyrroles. The ring systems were then combined, and doubles were eliminated, leaving 79 502 ring systems. The frequency of occurrence of ring systems in molecules followed a power law. One ring system, benzene, accounted for 50% of the occurrences, and the 10 most frequent ring systems accounted for nearly 90% of all occurrences, which illustrates that the ring system diversity of organic molecules is in fact quite low (Figure 2, Table S1).

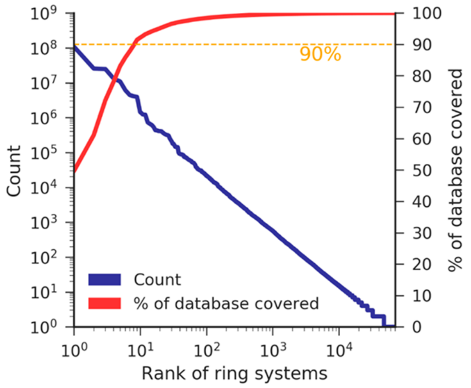


Figure 2. Frequency of occurrence of ring systems in known molecules. Ring systems are sorted by decreasing frequency of occurrence (logarithmic x axis) and the number of occurrences in known molecules (left logarithmic y axis, blue curve), and the cumulative coverage of known molecules by ring systems (right y-axis, red curve) is shown.

To form our reference database of known ring systems, here called RDB, we only considered the 12 536 ring systems which were also present in GDB4c, leaving out 57 637 ring systems containing more than four rings (48 728) and/or rings larger than 14 (16 629). We also generated a database of reference 3D ring systems, here called RDB3D, by collecting the corresponding 95 309 possible stereoisomers from GDB4c3D. The restriction to RDB ring systems also found in GDB4c was necessary because structural analysis (ring counts, etc.) as well as 3D structure generation largely failed for RDB ring systems not contained in GDB4c, mostly due to the presence of very large rings.

GDB4c versus RDB. Here, we illustrate the composition and novelty of GDB4c in comparison to known ring systems in RDB by measuring the distribution of ring systems according to various structural parameters (Figures 3 and 4, Table S2). Note that since RDB only contains 12 536 ring systems following the enumeration rules in Table 1, 98.6% of the 916 130 ring systems in GDB4c are novel.

The heavy atom count (hac) histograms show that GDB4c peaks at hac = 20, while RDB peaks at a smaller value of hac = 15 (Figure 3A). The larger size of GDB4c ring systems compared to RDB reflects the fact that, compared to RDB, GDB4c contains a much higher percentage of tetracyclic ring systems (GDB4c, 98.8%; RDB, 74.6%; Figure 3B) and ring systems containing at least one macrocycle (≥eight-membered ring; GDB4c, 92.6%; RDB, 36.6%; Figure 3C). GDB4c is overwhelmed by macrocycles because these offer more ring connection possibilities than smaller rings and therefore provide the largest number of ring systems in the exhaustive

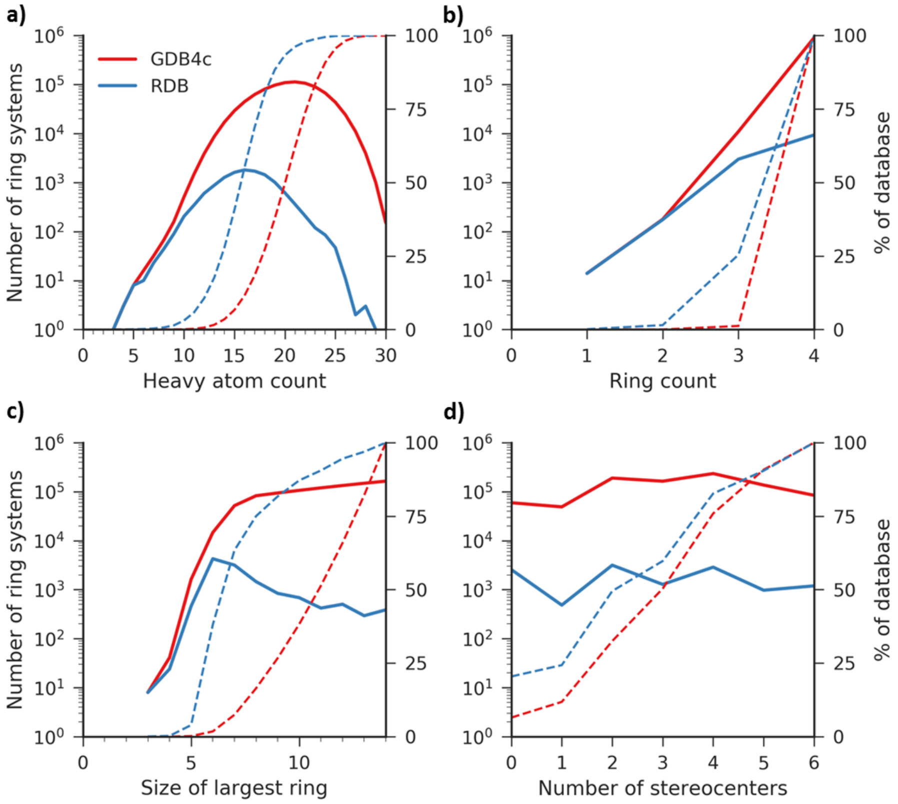


Figure 3. Property histograms for ring systems in GDB4c and RDB. Values (left axis, solid lines) and cumulative % (right axis, dashed lines) are reported for GDB4c (red) and RDB (blue). (a) Heavy atom count; (b) ring count; (c) size of the largest ring; (d) number of stereocenters (calculated from the 2D structure). See Figure S1 for the corresponding plots for GDB4c3D and RDB3D.

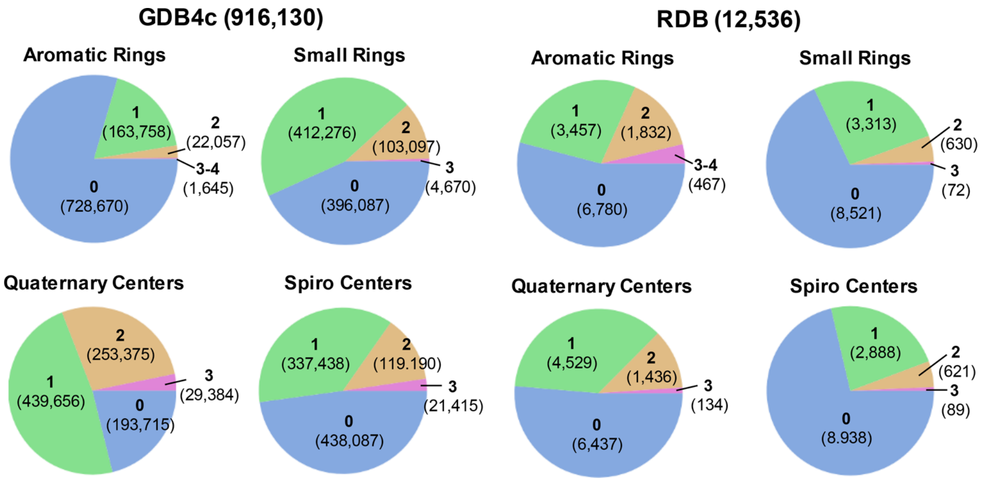


Figure 4. Pie charts of GDB4c and RDB as a function of the number of aromatic rings, number of small (three- or four-membered) rings, number of quaternary centers, spiro centers. The value of each slice is in bold and the total number of ring systems is between parentheses. See Figure S2 for the corresponding pie charts for GDB4c3D and RDB3D.

enumeration. By contrast, ring systems with a six-membered ring as their largest ring are most abundant in RDB (34.1% of the database) because six-membered rings dominate synthetic chemistry, while macrocycles are generally more difficult to synthesize and therefore more rarely explored.[55]

Both databases show a comparable distribution of ring systems in terms of stereocenters, although GDB4c contains a higher percentage of ring systems with stereocenters (856 692, 93.5%) compared to RDB (9969, 79.5%, Figure 3D). In both cases, we observe an intriguing dominance of ring systems with an even number of stereocenters (even/odd/no stereocenter, GDB4c: 507 758/348 934/59 438, RDB: 7232/2738/2566) reminiscent of the dominance of ring systems with even numbers of carbon atoms in databases of known molecules.[56] We believe that the dominance of ring systems with even numbers of stereocenters in GDB4c and RDB reflects the fact that a connection between two adjacent rings involves most often zero (aromatic rings, spiro centers) or two stereocenters (other bicyclic systems). The stronger dominance of ring systems containing stereocenters in GDB4c compared to RDB

Table 1. Rules for Selecting Ring Systems in GDB4c

rule	comment
(i) rules applied during ring enlargement:	
≤4 rings	sets the maximum number of rings to 4
≤2 rings larger than 6 atoms	a maximum of two rings larger than 6 are allowed
≤1 ring larger than 7 atoms	only one ring may be larger than 7 atoms
ring size ≤14 atoms	rings allowed only up to 14-membered
for tetracyclic systems only: no atom in two small rings	restricts appearance of fused small rings
(ii) rules for aromatization:	
only 5- or 6-membered aromatic rings	other ring sizes cannot be aromatic
no aromatic bonds in small rings	excludes 3- and 4-membered rings fused to aromatics

reflects the smaller percentage of ring systems containing aromatic rings in GDB4c (187 460, 20.5%) compared to RDB (5707, 45.9%, Figure 4, Table S2). We further observe that, compared to RDB, GDB4c has more ring systems containing at least one quaternary center (GDB4c: 722 415, 78.9%; RDB: 6099, 48.6%) or at least one small (three- or four-membered) ring (GDB4c: 520 043, 56.8%; RDB: 4015, 32.0%, Figure 4, Table S2). Note that the higher abundance of quaternary centers in GDB4c compared to RDB is partly caused by a higher fraction of ring systems containing at least one spiro center (GDB4c: 478 043, 52.2%; RDB: 3598, 28.7%).

Taken together, the above comparisons show that GDB4c stands out by being more abundant than RDB in terms of macrocycles as well as stereochemically rich ring systems containing quaternary centers, including spiro centers, and small rings. These differences reflect the difficulty in preparing ring systems containing these structural elements but also indicate opportunities for ring system novelty that can arise if one is willing to synthesize these more challenging features.

GDB4c3D versus RDB3D. The trends observed in the composition of GDB4c versus RDB are preserved when comparing the 6 555 929 stereoisomers in GDB4c3D with the 95 309 stereoisomers in RDB3D (Figures S1 and S2, Table S3). Note that for both GDB4c and RDB the enumeration of all possible diastereomers in the corresponding 3D database results in a smaller percentage of ring systems containing aromatic rings or spiro centers because these structural features are almost always nonstereogenic. An additional analysis in terms of molecular shape as measured by the principal moment of inertia[57] (PMI) shows that both GDB4c3D and RDB3D feature ring systems across the entire shape triangle (Figure 5A,B). Frequency histograms along the PMI axes show that compared to RDB3D, GDB4c3D is slightly shifted away from rod-shaped (lower PMI1 values) toward sphere-shaped (high PMI1 values) and disk-shaped (lower PMI2 values) ring systems, reflecting the higher frequency of ring systems with quaternary centers and macrocycles (Figure 5C,D).

Strained Rings in GDB4c3D. To test whether the ring systems enumerated in GDB4c contained strained rings, we analyzed GDB4c3D in comparison to RDB3D and to organic molecules up to 50 atoms in the Cambridge Structure Database (CSD) in terms of the deviation of bond angle values from their optimal values. Measuring all bond angles in GDB4c3D (217 836 431 records), RDB3D (2 702 528 records), and in CSD (4 776 156 records) and grouping them according to atom hybridization (sp2/sp3), ring size (3, 4, 5, 6, > 6), and position (endo/exocyclic) produced 10 different bond angle categories. In each of these categories, the bond angles followed a Gaussian-like distribution and mean value close to the textbook value (Figure 6A). The distributions were very similar between the three databases except in three cases, namely,

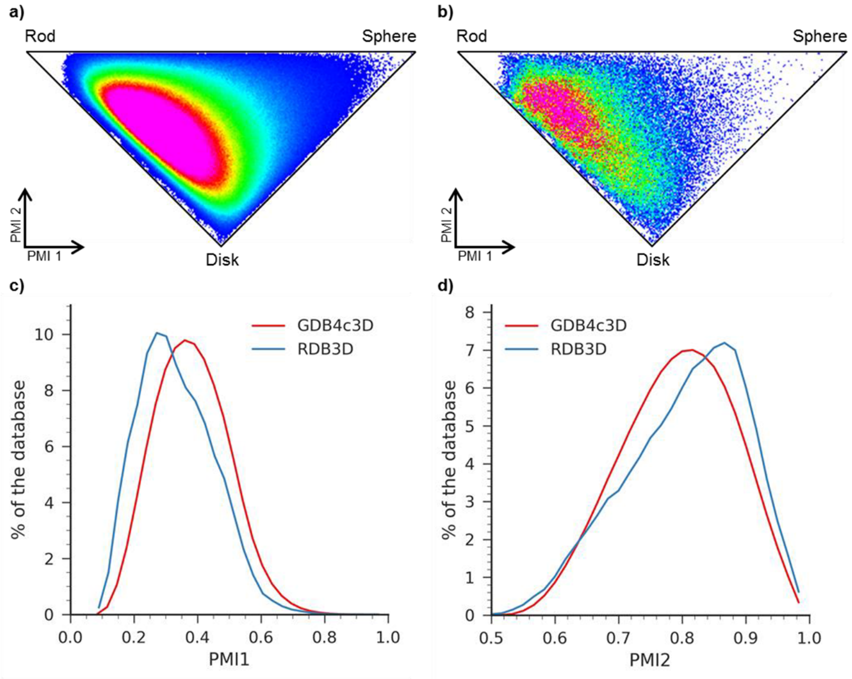


Figure 5. Molecular shape analysis of GDB4c3D against RDB3D. (a,b) PMI-plot of GDB4c3D (a) and RDB3D (b). Colors show the number of ring systems per pixel and range from blue (smallest) to magenta (largest). In GDB4c3D, the range is between 1 and 800, and in RDB3D, it is between 1 and 20. (c) Normalized PMI1 frequency histogram. (d) Normalized PMI2 frequency histogram.

DOI: 10.1021/acs.jcim.7b00457

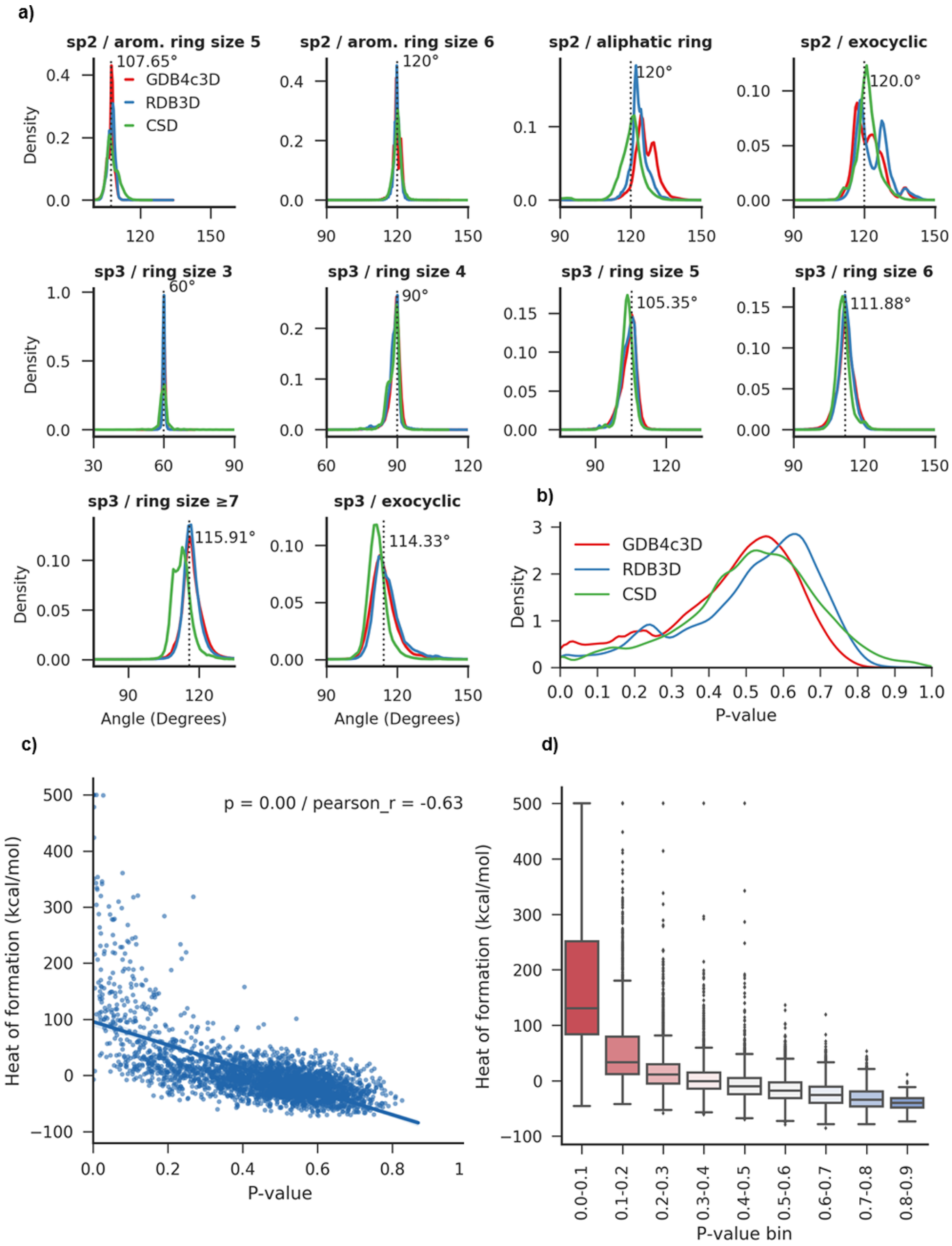


Figure 6. Estimating ring system strain from bond angles. (a) Kernel density estimation (KDE) of the distribution of the angles in each of the 10 categories for the three databases: GDB4c3D, RDB3D, and CSD. The mean used for the normal distribution fitting is specified on a vertical dashed black line (see Table S4 for more information). (b) KDE of the p value distribution in each database. (c) Scatter plot with regression line showing the correlation between p value and heat of formation of a subsample of 3000 ring systems of GDB4c3D obtained from the bigger 65 300 ring system sample. Notice that the linear regression test is from the whole 65 300 ring system sample and is significant with p value = 0.0, and the Pearson's correlation is −0.63. (d) Box plots of the heat of formation values categorized by binning the p value distribution in 0.1 intervals for a sample of 65 300 ring systems in GDB4c3D.

angles at sp^2 centers when they are in aliphatic rings or exocyclic and angles at sp^3 centers in at least seven-membered rings, which occur in the more unusual and diverse structural types.

To obtain a measure of ring strain from the bond angle value distribution, we first fitted a normal distribution in each category using as a guide the simplest molecules of each group, from which a p value was computed for each angle (Table S4). We then assigned a p value to each ring system by taking the lowest p value across all its angles. This analysis showed a comparable distribution of p values in all three databases, with cases covering the entire p value range, including cases with very low p values reflecting the presence of structures with unusual bond angles. Both GDB4c3D and CSD molecules peaked at p = 0.55, while RDB3D peaked at a higher value of p = 0.65 (Figure 6B, note that for CSD the statistics is done per molecule and not per ring system). Comparing p values with the calculated heat of formation obtained from MOPAC for a selection of 65 300 ring systems showed a good correlation between both values, suggesting that bond angle deviation

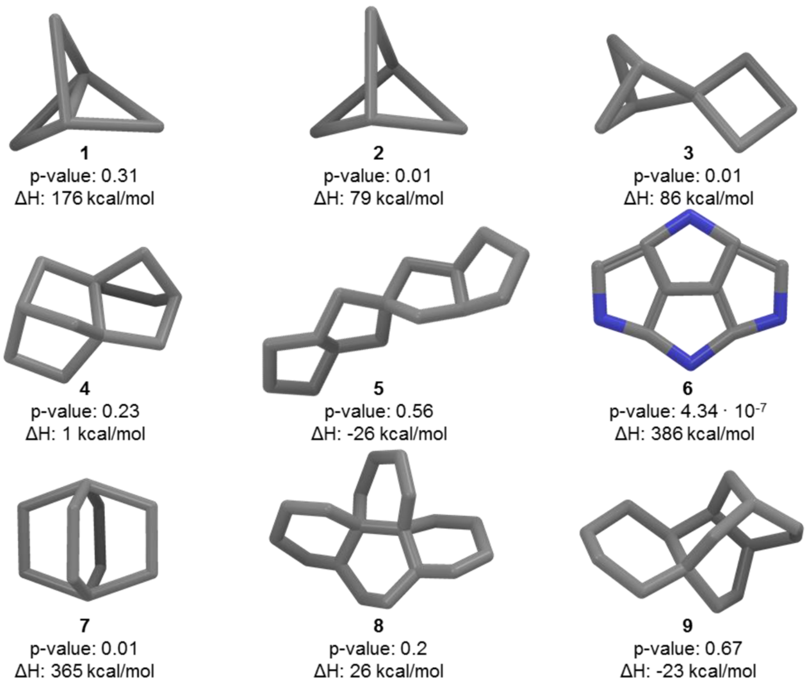


Figure 7. Examples of ring systems shown in 3D annotated with calculated *p* value and heat of formation. SMILES: (**1**) C1[C@]23C[C@]12C3. (**2**) C1[C@H]2C[C@@H]1C2. (**3**) C1CC2(C1)[C@H]1C[C@@H]2C1. (**4**) C1C[C@]23C[C@H]1C[C@]21CC[C@@H](C1)C3. (**5**) C1C[C@H]2CC3(C[C@H]2C1)C[C@H]1CCC[C@@H]1C3; (**6**) N1C=C2NC3=CNC4=C3C2=C1N4. (**7**) C1CC23CCC1(CC2)CC3. (**8**) C1CC[C@@]23CCCC[C@]22CCCC[C@H]2CC[C@@H]3C1. (**9**) C1CC[C@@]23CC[C@@H]4CC[C@@H](C[C@@H]4[C@@H]-2C1)C3.

provides a good measure for overall ring system strain (Figure 6C,D). Since unusually strained ring systems are relatively rare in the overall GDB4c, we have selected to leave them in the databases and label all ring systems with the *p* value as a structural warning. Ring systems with unusually strained geometries might be of interest in the field of theoretical chemistry.

Selected examples of ring systems with same ring sizes across the *p* value range are shown in Figure 7. The difficulty to decide on whether a ring system is synthetically possible or not on the basis of the calculated numbers is illustrated by 1,1,1-propellane (**1**) and bicyclo[1.1.1]pentane (**2**), which are both strained, reactive yet synthetically accessible ring systems for which practical synthetic routes have recently been discovered.[58] On the other hand, the spiro ring system **3** containing three four-membered rings is unknown and probably not synthesizable as it combines the difficult bicyclo[1.1.1]pentane with a strained quaternary center at the spirocyclobutane. Among five-membered ring systems, many innovative examples such as **4** and **5** do not present unusual ring strain; however, the polycyclic aromatic five-membered ring system **6** has some of the lowest *p* values and highest calculated energies in GDB4c due to the presence of four very unusual exocyclic sp^2 angles. Among ring systems with six-membered rings, the highly strained tricyclo[2.2.2.2]decane (**7**) is representative of ring systems with pyramidalized quaternary centers that are clearly not synthetically feasible. On the other hand, complex ring systems such as **8** and **9** are not strained and constitute attractive 3D-shaped scaffolds.

Interactive Chemical Space Maps. To facilitate the understanding and exploration of GDB4c and GDB4c3D, we have generated color-coded interactive chemical space maps in the form of Java applets, called "mapplets," following our previously reported approach.[31] These chemical space maps are obtained by placing the databases in multidimensional chemical spaces and performing dimensionality reduction.[59,60] Here, we consider chemical spaces whose dimensions correspond to the individual bit values of atom-pair fingerprints counting the number of atom pairs in different categories (here, aliphatic and aromatic atoms) separated by increasing topological distances for 2D structures[61] or through-space distances for 3D structures.[62] These atom-pair fingerprints represent molecular shape and pharmacophores, which are two key parameters determining the biological activity of drug-like molecules.[8] The 2D maps are produced by similarity mapping as a dimensionality reduction method,[63,64] calculating similarity values to 173 reference ring systems, each selected randomly from one of the 173 occupied value pairs (heavy atom count, largest ring). The GDB4c-mapplet features interactive maps for GDB4c, GDB4c3D, and the reference databases RDB and RDB3D and can be downloaded at www.gdb.unibe.ch for curiosity-driven exploration of the ring system universe.

The different color-coded similarity maps in the GDB4c mapplet illustrate the structural diversity of GDB4c and GDB4c3D (Figure 8). The map of GDB4c distributes ring systems in concentric series of crescent-shaped groups of leaves (Figure 8, left, and Figures S3 and S4). Each concentric group of leaves contains ring systems with an increasing number of nonaromatic rings, and each leaf contains a group of ring systems of the same number of heavy atoms. GDB4c is most densely populated in the upper right portion of the map, representing nonaromatic ring systems with the largest rings. By contrast, RDB mostly populates the lower left portion of the map containing the smaller ring systems, in particular those containing aromatic rings.

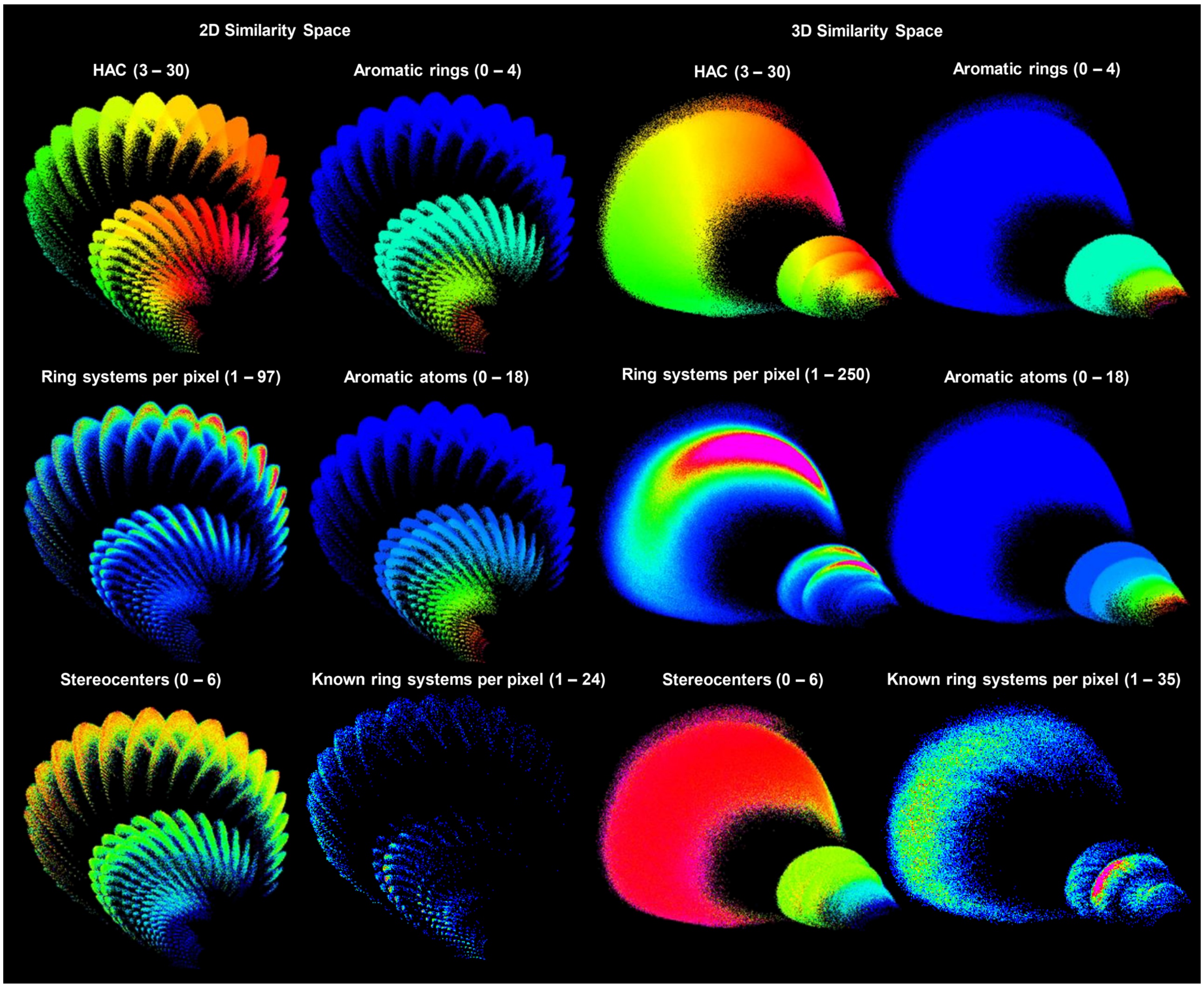
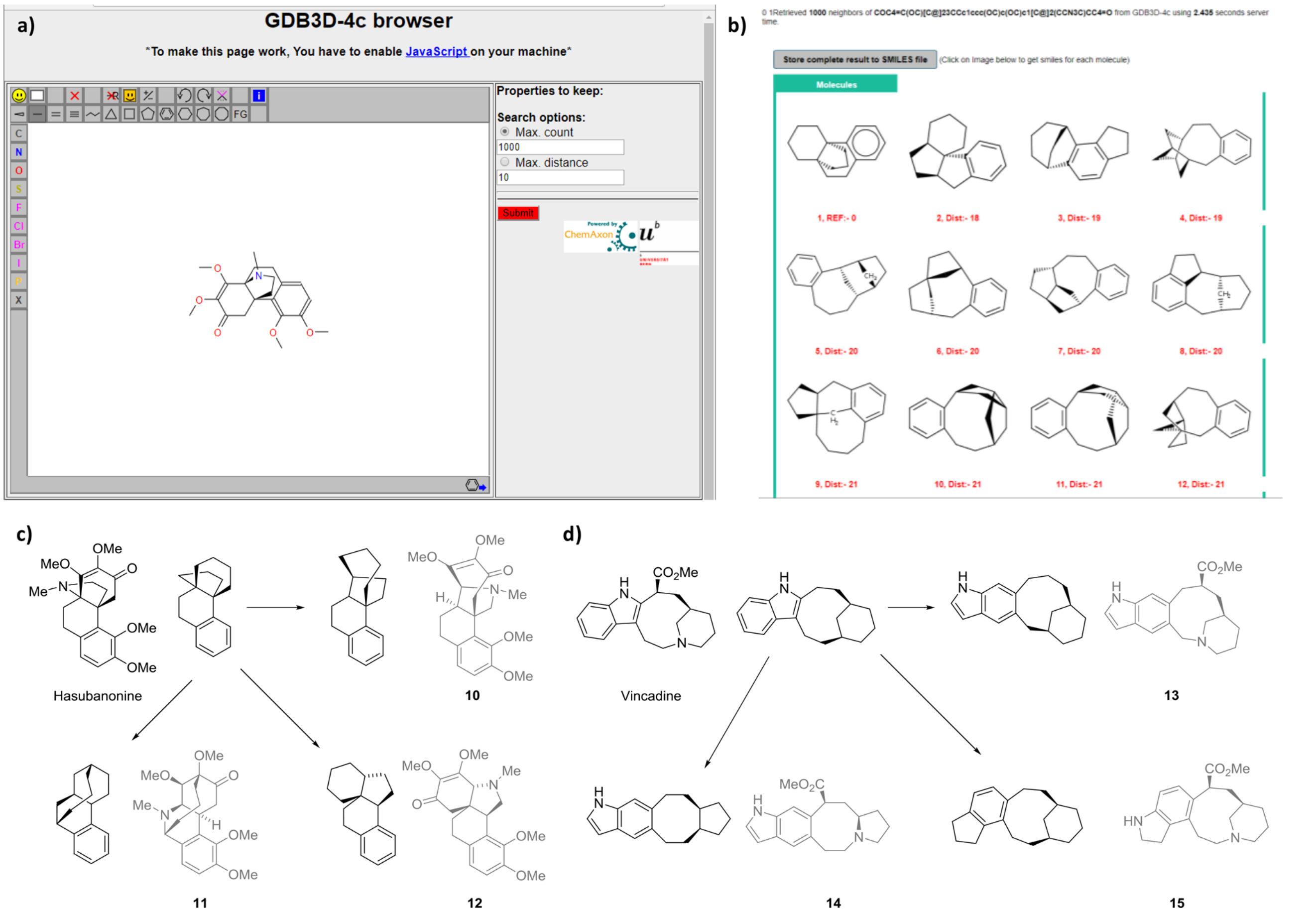


Figure 8. Color-coded chemical space maps of GDB4c (left) and GDB4c3D (right) obtained by similarity mapping. The maps are color coded according to the indicated value from lowest (blue) to highest (magenta) value in the indicated value range. Interactive versions of these maps can be downloaded as a "mapplet" Java application at www.gdb.unibe.ch. See Figures S3−S6 for all images available in the mapplet for GDB4c, RDB, GDB4c3D, and RDB3D color-coded by HAC, frequency, aromatic ring count, aromatic atom count, stereocenter count, *p* value (ring strain), and minimum ring count.

The map of GDB4c3D shows concentric crescent shaped regions similar to the 2D ring system maps but with much fewer geometrical features, reflecting the fact that the 3D atom pair fingerprint is encoded from through-space distances between atoms and does not perceive bond connectivity details (Figure 8, right and Figure S5 and S6). As for the 2D ring systems map, the largest crescent displays aliphatic ring systems, while aromatic ring systems are grouped in the smallest crescent at center. GDB4c3D is mostly occupying the upper part of the outermost crescent containing nonaromatic ring systems with large rings, while RDB3D has its highest density in the left part of the third crescent featuring small ring systems with a single aromatic six-membered ring.

Virtual Screening. Virtual screening guided by shape and pharmacophore similarity is particularly successful in identifying so-called "scaffold-hopping" analogs, which are molecules with the same biological activity as the parent drug but with a very different scaffold.[65,66] Such similarity searching should be possible for ring systems in GDB4c and GDB4c3D using the atom-pair fingerprints Xfp or 3DXfp described above in the context of our chemical space maps and might produce interesting suggestions for new ring systems conserving essential features of a known ring system as a help for designing new analogs of known bioactive molecules.

To enable Xfp/3DXfp similarity searching in GDB4c/GDB4c3D, we have created the corresponding Web-based nearest neighbor search tools,[33] which are accessible at www.gdb.unibe.ch. The GDB4c browser takes any 2D structure as input, extracts the largest parent ring system, and searches for the Xfp-nearest neighbors of this ring system in GDB4c. The GDB4c3D browser first generates a 3D structure from the input molecule with CORINA using either the specified stereochemistry or a single stereoisomer, extracts the largest parent 3D ring system, and searches for 3DXfp-nearest neighbors in GDB4c3D. Both browsers use the city-block distance as a similarity measure because it enables preorganization of the database for very fast searching.[33] Note that the geometry of the ring system is kept as in the parent functional molecule; therefore if analogs of the ring system itself are desired, the user should input the ring system directly and not



GDB3D-4c browser

To make this page work, You have to enable JavaScript on your machine

Properties to keep:

Search options:

● Max. count: 1000

○ Max. distance: 10

0 1Retrieved **1000** neighbors of **COC4=C(OC)[C@]23CCc1ccc(OC)c(OC)c1[C@]2(CCN3C)CC4=O** from GDB3D-4c using **2.435** seconds server time.

(Click on image below to get smiles for each molecule)

Figure 9. Virtual screening of GDB4c3D. (a) GDB4c3D Web-browser search window interface for nearest 3DXfp-neighbor searching shown with the natural product hasubanonine. (b) Browser results window showing the extracted parent scaffold of hasubanonine and the 3DXfp-nearest neighbors. (c) Structure of hasubanonine, its ring system, ring system analogs identified by 3DXfp-similarity searching, and designed natural product analogs **10**−**12**. (d) Same as c for the natural product vincadine with design analogs **13**−**15**. See methods and Supporting Information Figures S7 and S8 for details.

the functionalized molecule. In both browsers, the search results are displayed as a list of images, and the SMILES list can be downloaded as a text file (Figure 9A,B).

These similarity browsers might help in designing new analogs of known drugs, in particular polycyclic natural products, as illustrated here with hasubanonine and vincadine (Figure 9C,D). Starting with the chemical structures of these natural products, we used the above-described browsers to extract approximately 10 000 nearest neighbors of their parent ring system from GDB4c and GDB4c3D. We then scored these ring systems by similarity to the parent ring systems using the ROCS shape Tanimoto,[66] which is a more precise measure of 3D similarity than 3DXfp alone, and by Tanimoto similarity for a binary 1024-bit Daylight type substructure fingerprint to perceive bond connectivity details (Figures S7 and S8).[67] We identified in each case three new ring systems scoring relatively high in both ROCS score and Tanimoto similarity but which were not found either in our RDB or in Scifinder. We then distributed the necessary atoms and functional groups to form analogs of the reference natural products, placing these functional groups manually at equivalent positions in the ring system compared to the parent natural product. The resulting virtual molecules **10**−**15** represent possible synthetic targets with attractive natural product-like structures. While elegant, these molecules would clearly be challenging to synthesize. These examples are meant to illustrate how GDB4c or GDB4c3D might serve as a source of inspiration for molecular design.

■ CONCLUSION

In summary, we have enumerated all possible ring systems up to four rings following a set of maximum ring size criteria and obtained the databases GDB4c containing 916 130 ring systems as 2D structures and GDB4c3D listing the corresponding 6 555 929 stereoisomers as 3D structures. By comparison, known molecules exemplify only 79 502 ring systems, only 12 536 of which fall within the boundaries of GDB4c in terms of ring count and sizes, implying that the vast majority of the ring system chemical universe as enumerated in GDB4c is yet unknown. The new ring systems in GDB4c are stereochemically rich tetracyclic macrocycles containing small rings and quaternary centers including spiro centers. These features are more challenging to synthesize than aromatic six-membered rings which dominate in known ring systems. Tools for interactive visualization and virtual screening of GDB4c and GDB4c3D by similarity searching together with the databases themselves are publicly available at www.gdb.unibe.ch. We hope that these tools will serve as an aid and inspiration for synthetic chemists to explore the vast and mostly unknown universe of ring systems. Considering the vast number of possibilities to

DOI: 10.1021/acs.jcim.7b00457

design molecules from ring systems by building scaffolds with various topologies,[68,69] and functionalizing these scaffolds with heteroatoms and substituents,[35] exploring new ring systems opens essentially unlimited opportunities to enrich the structural diversity of drug-like compounds.

■ METHODS

Database Generation. All of the required software to generate the 2D database was written in Java and uses the JChem libraries by ChemAxon (www.chemaxon.com). For GDB4c, exhaustive filtering using the tools specified before was done after each step, to reduce to a minimum the number of graphs/molecules to process at the next step. For GDB4c3D, stereoisomers were obtained from CORINA (www.mn-am.com/products/corina) in SDF form and all structures optimized using the ChemAxon Fine 3D cleaner.

RDB was computed by grouping all molecules from DrugBank, ChEMBL, SureChEMBL, ZINC, PubChem, and Reaxys; stripping all molecules from acyclic bonds; converting cyclic atoms to carbon; and reintroducing one N atom per five-membered ring at the first available position (the aromatization algorithm iteratively tries to add the nitrogen atom in each of the positions: the first position that generates a valid pyrrole is chosen, and the algorithm stops). All ring systems with a maximum of 50 heavy atoms were retained in the database. The RDB3D reference database containing 3D structures of ring systems was created as follows: Xfp fingerprints were calculated for ring systems in RDB and GDB4c databases (see below for details of Xfp fingerprint). Each of the ring systems in the RDB database was compared to ring systems in GDB4c by matching their Xfp fingerprints. Whenever the Xfp fingerprint of the RDB ring system was identical to the Xfp fingerprint of the ring system in GDB4c, the respective ring system from GDB4c was selected, and then all stereoisomers of that ring system in GDB4c3D were extracted. In the case of comparisons, where the Xfp fingerprint from RDB matched to more than one ring system in GDB4c (due to fingerprint collision multiple rings systems may have identical Xfp fingerprints), the GDB4c ring systems that were considered were those that had the same canonical smiles or, if none was found, the first one.

For the estimation of ring strain, the reference organic CSD subset was created by considering all CSD molecules up to 50 heavy atoms composed of elements no other than C, H, N, O, F, Br, or Cl. The counterions were removed, and ionization states of molecules were adjusted to pH 7.4 using an in-house built Java program as the starting point. If the compound was available in complex form, only one of the largest fragments was retained.

***p* Values.** Calculations were done using the cluster computing platform Apache Spark, with both Scala and Python as programming languages. The chemistry libraries used were RDKit in Python and ChemAxon JChem in Scala. For every structure, each angle was measured and annotated with the following properties: whether the angle is in a ring, angle ring size, angle ring aromaticity, and carbon atom hybridization state. The angles were then grouped in the categories specified in Table S4. The fitted normal distributions of each category had the mean obtained in one of three ways: (1) If there is a minimal structure in which all angles are of that category, then the mean of all these angles is considered; (2) if not, the textbook value is used. (3) If there is no simple way of determining the mean, the GDB4c3D mean is used (Table S4). For the standard deviation, the value chosen was always 3 times the standard deviation of all the molecules in GDB4c3D (Table S4). Next, a z score was computed for each angle, and a two-tailed p value was obtained. The angles were then grouped by structure, and the lowest p value from all angles was selected. The p value was annotated in the GDB4c3D SDF and SMILES file.

Heat of Formation. The heat of formation of a sample of 65 300 ring systems extracted from GDB4c3D was calculated using MOPAC2016 (openmopac.net). Hydrogen atoms were first added to each ring system, and geometry was optimized using ChemAxon fast. If the optimization failed, the hydrogens were left at the default positions. Next, each structure was given as input to MOPAC to compute the heat of formation using the PM7 method with SYMMETRY and no structure optimization (NOOPT). The output from MOPAC, alongside information from each structure, was analyzed. Last, in the case of structures with a heat of formation higher than 500 kcal/mol, the energy was set to this value.

Fingerprint Calculation. To make the 2D database searchable, we used a modified version of our previously reported Xfp fingerprint[61] considering only three atom categories: all atoms, aromatic and aliphatic atoms, and a maximum topological distance of 15 bonds, producing a 48-bit fingerprint (including distance 0). These atom categories allow the perception of aromaticity versus aliphatic rings and overall molecular shape but exclude considering properties of the nitrogen atom, which are only added in aromatic five-membered rings to obtain valid SMILES for the ring systems. For each atom category, atom pairs are summed for each topological distance, and the sum values are divided by the number of atoms in the category. The final bit values are multiplied by 100 and rounded to the nearest integer value.

To make the 3D database searchable, we similarly adapted our previously reported 3DXfp fingerprint[62] considering only two atom categories: all atoms and aromatic atoms and a maximum through-space distance of 20 Å, resulting in a 32 bit-fingerprint. For each of the atom pairs AB in the molecule, a Gaussian function was generated centered at the atom pair distance d_{AB} with width of 0.18 × d_{AB}, and the function was sampled at 1.45, 1.71, 2.02, 2.38, 2.81, 3.32, 3.91, 4.62, 5.45, 6.43, 7.59, 8.96, 10.57, 12.47, 14.71, and 17.36 Å (16 bit values at $d_{n+1} = d_n \times 1.18$). For each of the 16 bits, values were summed across all atom pairs; the sum was divided by $\mathrm{hac(category)}^{1.5}$, multiplied by 100, and rounded to the integer value.

PMI-Maps Calculation. The shape analysis was carried out after the protocol of Sauer and Schwartz with in-house software written in Java.[57]

Principal Component Analysis (PCA) and Similarity Maps Creation. For the GDB4c and GDB4c3D data sets, principal component analysis (PCA) was performed using in-house software written in Java which uses the JSci science library (jsci.sourceforge.net/) to compute the eigenvalues and eigenvectors. Similarity maps and the corresponding color coded mapplets were generated as described previously.[64]

To generate a similarity fingerprint, we selected 173 reference ring systems by picking one compound randomly from each of the 173 available (HAC, largest ring) value pair bins. For each database compound, we then calculated the 173 city-block distances (CBD) using the modified Xfp and 3DXfp described above and converted the CBD to a similarity value S as $S_i = \mathrm{CBD}_i/(\mathrm{CBD}_i + X)$, where X is the median CBD across

DOI: 10.1021/acs.jcim.7b00457

compound pairs in the database (obtained from sampling 1 million distance pairs randomly).

To obtain similarity maps, we the performed PCA of the similarity fingerprint (S_1, ...S_i, ...S_{173}) data set on the whole GDB4c respectively on GDB4c3D. We then binned the (PC1, PC2) values into a 500 × 500 pixel image and color-coded each pixel by the average descriptor value for all compounds in that pixel. The similarity maps of RDB and RDB3D were created using the same eigenvectors obtained from the PCA of GDB4c and GDB4c3D and binning the (PC1, PC2) plane into a 300 × 300 pixel image.

Virtual Screening. The Web-based Xfp and 3DXfp browsers for GDB4c and GDB4c3D were assembled as previously reported for other databases.[22,70] Three similarity search examples (set 1−3), each for hasubanonine and vincadine, were performed. *Set 1*: The similarity search was performed using the full natural product as input to the GDB4c browser. We retrieved 2500 analogs in the case of hasubanonine and 4000 analogs in the case of vincadine. Extracting the corresponding 3D stereoisomers from GDB4c3D yielded approximately 10 000 ring systems in each case. *Sets 2 and 3*: Searches were performed using the extracted ring system of the natural product and full natural product (preserves the geometry of the ring system as in the natural product) as inputs to the GDB4c3D browser, respectively. For each set 2 and 3, we collected 5000 analogs in the case of hasubanonine and 6000 analogs in the case of vincadine, yielding approximately 10 000 ring systems in each case when considering both enantiomers of chiral ring systems. *Set 4*: A control set of ∼10 000 randomly selected ring systems from GDB4c3D within hac ± 2 of the ring system of the natural product. For each of the four sets, ROCS similarity calculations were performed using the ROCS shape Tanimoto and the 1024-bit Daylight type substructure fingerprint from JChem as a similarity measure. The scatter plots of similarity values are shown in Figures S3 and S4.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00457.

> Ten most common ring systems in the known ring system database (Table S1). Details of the composition of GDB4c (Table S2) and GDB4c3D (Figures S1 and S2 and Table S3). Categories of bond angles used for ring strain calculations (Table S4). Color-coded images of similarity maps for GDB4c, RDB, GDB4c3D, and RDB3D (Figures S3−S6). Scatter plots of similarity values of ring system analogs of hasubanonine and vincadine identified in GDB4c and GDB4c3D (Figures S7 and S8) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Fax: +41 31 631 80 57. E-mail: jean-louis.reymond@dcb.unibe.ch.

ORCID

Jean-Louis Reymond: 0000-0003-2724-2942

Author Contributions

[†]Equal contributions

Author Contributions

R.V. and J.A.-P. realized the project and wrote the paper. M.A. computed the MQN- and SMIfp search mapplets, performed virtual screening, and wrote the paper. J.-L.R. designed and supervised the project and wrote the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported financially by the University of Berne, the Swiss National Science Foundation. J.A.-P. is supported financially by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 676434, "Big Data in Chemistry" ("BIGCHEM," http://bigchem.eu). We thank OpenEye Scientific Software Pvt. Ltd. for providing free academic licenses for Flipper/Omega/Rocs and ChemAxon Pvt. Ltd. for providing free academic and web licenses for their products, Dr. David Evans of RELX Intellectual Properties SA for providing access to the Reaxys database, and Dr. Sacha Javor for sharing his expertise in calculating the heat of formation.

■ REFERENCES

(1) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

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

(3) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(4) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-Like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.

(5) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432*, 823−823.

(6) Weininger, D. Smiles, a Chemical Language and Information-System 0.1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(7) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to 3-Dimensional Atomic Coordinates - Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(8) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.

(9) Reymond, J. L.; Ruddigkeit, L.; Blum, L. C.; Van Deursen, R. The Enumeration of Chemical Space. *WIREs comput. Mol. Sci.* **2012**, *2*, 717.

(10) Buchanan, B. G.; Smith, D. H.; White, W. C.; Gritter, R. J.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. 22. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-Dendral Program. *J. Am. Chem. Soc.* **1976**, *98*, 6168−6178.

(11) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. Sprout: Recent Developments in the De Novo Design of Molecules. *J. Chem. Inf. Model.* **1994**, *34*, 207−217.

(12) Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413−419.

(13) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective

Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079−1087.

(14) Brown, N.; McKay, B.; Gasteiger, J. The De Novo Design of Median Molecules within a Property Range of Interest. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 761−771.

(15) Lameijer, E. W.; Kok, J. N.; Back, T.; Ijzerman, A. P. The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545−552.

(16) van Deursen, R.; Reymond, J. L. Chemical Space Travel. *ChemMedChem* **2007**, *2*, 636−640.

(17) Virshup, A. M.; Contreras-Garcia, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.

(18) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual Exploration of the Small-Molecule Chemical Universe Below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−1508.

(19) Fink, T.; Reymond, J. L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.

(20) 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*, 8732−8733.

(21) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical Space as a Source for New Drugs. *MedChemComm* **2010**, *1*, 30−38.

(22) Blum, L. C.; van Deursen, R.; Reymond, J. L. Visualisation and Subsets of the Chemical Universe Database Gdb-13 for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637−47.

(23) 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*, 2864−2875.

(24) Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722−730.

(25) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Discovery of Nmda Glycine Site Inhibitors from the Chemical Universe Database Gdb. *ChemMedChem* **2008**, *3*, 1520−4.

(26) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4*, 1803−1805.

(27) Luethi, E.; Nguyen, K. T.; Burzle, M.; Blum, L. C.; Suzuki, Y.; Hediger, M.; Reymond, J. L. Identification of Selective Norbornane-Type Aspartate Analogue Inhibitors of the Glutamate Transporter 1 (Glt-1) from the Chemical Universe Generated Database (Gdb). *J. Med. Chem.* **2010**, *53*, 7236−7250.

(28) Garcia-Delgado, N.; Bertrand, S.; Nguyen, K. T.; van Deursen, R.; Bertrand, D.; Reymond, J.-L. Exploring Alpha 7-Nicotinic Receptor Ligand Diversity by Scaffold Enumeration from the Chemical Universe Database Gdb. *ACS Med. Chem. Lett.* **2010**, *1*, 422−426.

(29) Brethous, L.; Garcia-Delgado, N.; Schwartz, J.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Synthesis and Nicotinic Receptor Activity of Chemical Space Analogues of N-(3r)-1-Azabicyclo[2.2.2]Oct-3-Yl-4-Chlorobenzamide (Pnu-282,987) and 1,4-Diazabicyclo[3.2.2]-Nonane-4-Carboxylic Acid 4-Bromophenyl Ester (Ssr180711). *J. Med. Chem.* **2012**, *55*, 4605−4618.

(30) Schwartz, J.; Awale, M.; Reymond, J.-L. Smifp (Smiles Fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules. *J. Chem. Inf. Model.* **2013**, *53*, 1979−1989.

(31) Awale, M.; van Deursen, R.; Reymond, J. L. Mqn-Mapplet: Visualization of Chemical Space with Interactive Maps of Drugbank, Chembl, Pubchem, Gdb-11, and Gdb-13. *J. Chem. Inf. Model.* **2013**, *53*, 509−518.

(32) Ruddigkeit, L.; Awale, M.; Reymond, J. L. Expanding the Fragrance Chemical Space for Virtual Screening. *J. Cheminf.* **2014**, *6*, 27−39.

(33) Ruddigkeit, L.; Blum, L. C.; Reymond, J.-L. Visualization and Virtual Screening of the Chemical Universe Database Gdb-17. *J. Chem. Inf. Model.* **2013**, *53*, 56−65.

(34) Visini, R.; Awale, M.; Reymond, J. L. Fragment Database Fdb-17. *J. Chem. Inf. Model.* **2017**, *57*, 700−709.

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

(36) Danziger, D. J.; Dean, P. M. Automated Site-Directed Drug Design: A General Algorithm for Knowledge Acquisition About Hydrogen-Bonding Regions at Protein Surfaces. *Proc. R. Soc. London, Ser. B* **1989**, *236*, 101−113.

(37) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. Recap-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(38) Leach, A. R.; Hann, M. M. The in Silico World of Virtual Libraries. *Drug Discovery Today* **2000**, *5*, 326−336.

(39) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to De Novo Design Using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163−1184.

(40) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. Leap into the Pfizer Global Virtual Library (Pgvl) Space: Creation of Readily Synthesizable Design Ideas Automatically. *Methods Mol. Biol.* **2011**, *685*, 253−276.

(41) Chevillard, F.; Kolb, P. Scubidoo: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, *55*, 1824−1835.

(42) Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, *56*, 1253−1266.

(43) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings. In Silico Exploration of Ring Universe to Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006**, *49*, 4568−4573.

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

(45) Beckmann, H. S. G.; Nie, F.; Hagerman, C. E.; Johansson, H.; Tan, Y. S.; Wilcke, D.; Spring, D. R. A Strategy for the Diversity-Oriented Synthesis of Macrocyclic Scaffolds Using Multidimensional Coupling. *Nat. Chem.* **2013**, *5*, 861−867.

(46) Xie, J.; Bogliotti, N. Synthesis and Applications of Carbohydrate-Derived Macrocyclic Compounds. *Chem. Rev.* **2014**, *114*, 7678−7739.

(47) Valeur, E.; Gueret, S. M.; Adihou, H.; Gopalakrishnan, R.; Lemurell, M.; Waldmann, H.; Grossmann, T. N.; Plowright, A. T. New Modalities for Challenging Targets in Drug Discovery. *Angew. Chem., Int. Ed.* **2017**, *56*, 10294−10323.

(48) McKay, B. D. Practical Graph Isomorphism. *Congressus Numerantium* **1981**, *30*, 45−87.

(49) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.

(50) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. Drugbank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.

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

(52) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. Zinc: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(53) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. Pubchem: A Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.

(54) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D., The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; American Chemical Society: Washington, DC, 2014; Vol. *1164*, pp 127−148.
(55) Donald, J. R.; Unsworth, W. P. Ring-Expansion Reactions in the Synthesis of Macrocycles and Medium-Sized Rings. *Chem. - Eur. J.* **2017**, *23*, 8780−8799.
(56) Sarma, J. A. R. P.; Nangia, A.; Desiraju, G. R.; Zass, E.; Dunitz, J. D. Even-Odd Carbon Atom Disparity. *Nature* **1996**, *384*, 320−320.
(57) Sauer, W. H.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987−1003.
(58) Gianatassio, R.; Lopchuk, J. M.; Wang, J.; Pan, C. M.; Malins, L. R.; Prieto, L.; Brandt, T. A.; Collins, M. R.; Gallego, G. M.; Sach, N. W.; Spangler, J. E.; Zhu, H.; Zhu, J.; Baran, P. S. Organic Chemistry. Strain-Release Amination. *Science* **2016**, *351*, 241−246.
(59) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in Visual Representations of Chemical Space. *Expert Opin. Drug Discovery* **2015**, *10*, 959−973.
(60) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2015**, *55*, 84−94.
(61) Awale, M.; Reymond, J. L. Atom Pair 2d-Fingerprints Perceive 3d-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of Zinc and Gdb-17. *J. Chem. Inf. Model.* **2014**, *54*, 1892−1897.
(62) Awale, M.; Jin, X.; Reymond, J. L. Stereoselective Virtual Screening of the Zinc Database Using Atom Pair 3d-Fingerprints. *J. Cheminf.* **2015**, *7*, 3.
(63) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-Based Data-Fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chem. Biol. Drug Des.* **2007**, *70*, 393−412.
(64) Awale, M.; Reymond, J. L. Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and Chembl in High Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2015**, *55*, 1509−1516.
(65) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. ″Scaffold-Hopping″ by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(66) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(67) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Model.* **1992**, *32*, 515−521.
(68) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. Scaffold Topologies. 1. Exhaustive Enumeration up to Eight Rings. *J. Chem. Inf. Model.* **2008**, *48*, 1304−1310.
(69) Wester, M. J.; Pollock, S. N.; Coutsias, E. A.; Allu, T. K.; Muresan, S.; Oprea, T. I. Scaffold Topologies. 2. Analysis of Chemical Databases. *J. Chem. Inf. Model.* **2008**, *48*, 1311−1324.
(70) Awale, M.; Reymond, J. L. A Multi-Fingerprint Browser for the Zinc Database. *Nucleic Acids Res.* **2014**, *42*, W234−W239.