JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS

Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds

Aaron M. Virshup,[†,§] Julia Contreras-García,[†,§,#] Peter Wipf,[‡,§] Weitao Yang,*[,†,§] and David N. Beratan*[,†,§]

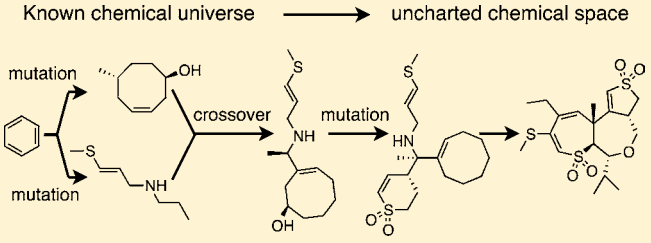
[§]Center for Chemical Methodologies and Library Development, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States

[†]Department of Chemistry, Duke University, Durham, North Carolina 27708, United States

[‡]Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States

Supporting Information

ABSTRACT: The "small molecule universe" (SMU), the set of all synthetically feasible organic molecules of 500 Da molecular weight or less, is estimated to contain over 10^{60} structures, making exhaustive searches for structures of interest impractical. Here, we describe the construction of a "representative universal library" spanning the SMU that samples the full extent of feasible small molecule chemistries. This library was generated using the newly developed Algorithm for Chemical Space Exploration with Stochastic Search (ACSESS). ACSESS makes two important contributions to chemical space exploration: it allows the systematic search of the unexplored regions of the small molecule universe, and it facilitates the mining of chemical libraries that do not yet exist, providing a near-infinite source of diverse novel compounds.



INTRODUCTION

Many grand challenges in science and biomedicine require molecular and materials discovery.[1−4] Yet, the fraction of "chemical space" that has been explored over human history is infinitesimal—less than one part in 10^{50}.[5] The vast unexplored molecular frontier suggests that there is reason for optimism in the face of grand scientific tasks.

Current experimental and theoretical tools are poorly matched to the scale and scope of the molecular discovery undertaking. Enumerating all compounds or materials in the vastness of molecular space is impossible, and assessing their properties is even more unimaginable; even synthetically accessible small organic compounds number over 10^{60}.[5] Further, current compound libraries are notably lacking in diversity, meaning that much of available small molecule chemistry has yet to be explored.[5−7] Chemical libraries that capture the much broader diversity of the entire chemical universe promise to be a more empowering starting point for molecular discovery.[8−11]

Synthetic methods for expanding the diversity of compound collections, collectively known as diversity-oriented synthesis, arose as a reaction to the relatively nondiverse libraries generated by combinatorial synthesis and used in high-throughput screening.[3,8,12] Computational approaches to aid in chemical space exploration can be very broadly classified into molecular optimization techniques, those that aim to identify or design compounds with optimal activity, and enumerative techniques, which aim to explore the full extent of a given chemical space. Optimization techniques include both search techniques based on stochastic and evolutionary algorithm frameworks,[13−17] directed combinatorial library design,[18] and optimization within continuous alchemical spaces.[19−21] Other methods aim to construct molecules that have high similarity to a set of known compounds as a basis for lead discovery.[22,23]

In contrast to these focused design techniques, others have sought to enumerate all structures and to explore the full range of chemistries available within a given chemical space. Reymond has reported enumeration of all possible organic compounds (within a set of rules for synthetic feasibility) of 13 or fewer heavy atoms.[24−26] This "GDB13" database contains nearly 1 billion compounds, most of which are not found in any other compound library. Compound mining in GDB13 has already led to successes in the drug discovery process.[27,28] Oprea has also made considerable progress in quantifying chemical topology, having enumerated all possible ring topologies up to eight rings.[6,29] Both of these studies yield valuable information about the diversity of the small molecule chemical space.

Although the number of compounds in the small molecule universe (SMU) is far too great to be enumerated, we show here that this astronomically large collection can be characterized in a way similar to enumerative techniques but that only requires consideration of a far smaller set of chemical

Received: February 1, 2013
Published: April 2, 2013

   dx.doi.org/10.1021/ja401184g | J. Am. Chem. Soc. 2013, 135, 7296−7303

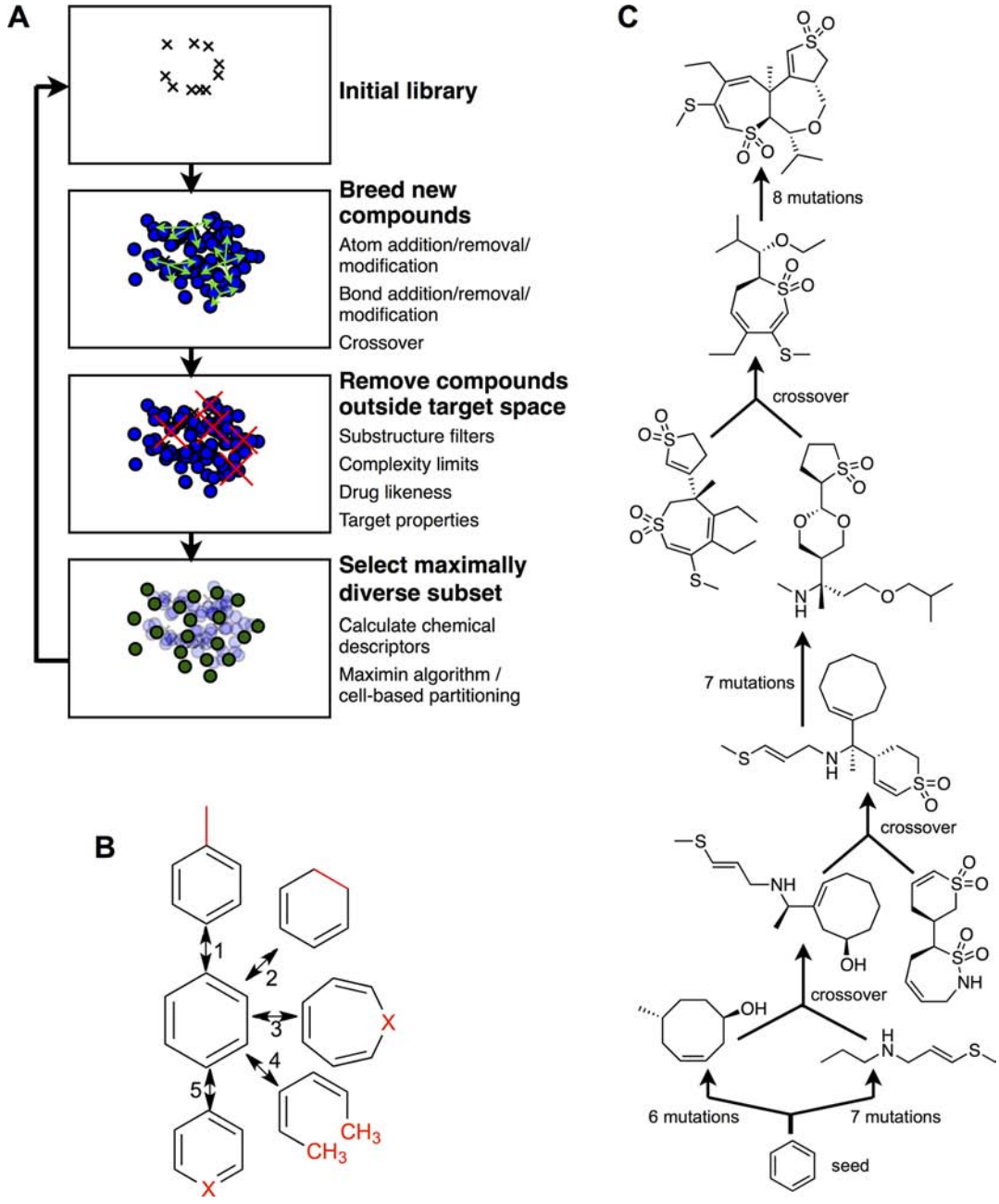


Figure 1. The ACSESS procedure allows the construction of a representative universal library in an arbitrary chemical space. (A) A library of initial molecules is expanded using chemical mutations and crossover; compounds outside the target chemical space are discarded; and a maximally diverse subset of the remaining molecules is selected. This process is repeated until the diversity of the set converges. (B) Chemical structure modifications, which include: (1) addition or deletion of terminal atoms; (2) bond order modifications; (3) addition or deletion of in-chain atoms; (4) removal or addition of cyclic bonds; and (5) modifications of atom type. (C) An example of a chemical space trajectory. The final compound occupies unexplored chemical space in the SMU.

structures. Well-established chemoinformatics techniques allow the construction of "representative sublibraries", maximally diverse collections of compounds that contain as much diversity as the parent library expressed in a much smaller number of compounds.[30−33] Combining these techniques with concepts from chemical evolutionary algorithms, as described below, allows the mapping of humongous chemical spaces such as the SMU.

A schematic overview of the Algorithm for Chemical Space Exploration with Stochastic Search (ACSESS) is shown in Figure 1. By combining stochastic chemical structure mutations with methods for maximizing molecular diversity, ACSESS efficiently produces representative sublibraries of vast chemical spaces. This procedure fundamentally differs from existing chemical genetic algorithms; it is designed to explore rigorously the full diversity available in targeted chemical spaces, including astronomically large ones, such as the space of drug-like molecules in the SMU (*vide infra*) or functional chemical regions that contain molecules with specific desirable properties (see Supporting Information). We term the compound libraries thus generated "representative universal libraries" (RUL), collections of compounds that represent the full extent of chemical diversity within a much larger set of molecular structures.

dx.doi.org/10.1021/ja401184g | *J. Am. Chem. Soc.* 2013, 135, 7296−7303

"Chemical space" is defined here as an M-dimensional Cartesian space in which compounds are located by a set of M physiochemical and/or chemoinformatic descriptors. We focus on chemical spaces defined by selected properties; the SMU, for instance, contains all stable compounds of 500 Da or less. A representative universal library contains chemical compounds that span the full extent of accessible descriptor values in the M-dimensional space. While the choice of the M descriptors and diversity measures may depend on specific applications, this approach remains generally applicable. Unlike previous techniques for selecting a maximally diverse sublibrary, ACSESS is unique in that the parent collection does not need to be enumerated, allowing systematic exploration of uncharted and astronomically large chemical spaces.

METHODS

The ACSESS Algorithm. To map an arbitrary chemical space, ACSESS is seeded with a set of compounds; often, a very small library suffices to initialize the algorithm (a library consisting of benzene and cyclohexane, for example, was used to seed all work shown here). This library is enlarged and diversified over multiple computational generations. In each generation, the library is modified as follows (Figure 1A): (1) The library is expanded by creating new structures using "chemical mutations"; (2) compounds not in the chemical space of interest are removed (including those assessed as not being synthetically feasible or lacking the property of interest); and (3) the size of the library is reduced by selecting a maximally diverse subset of compounds. The qualitative features of the algorithm are discussed below; a complete description is provided in the Supporting Information.

Note that the ACSESS algorithm requires concrete choices of chemical descriptor, diversity function, and target chemical space. As described below, we have chosen descriptors, chemical space filters, and diversity definitions that are relatively general, transferable, and computationally efficient, allowing the construction of a large compound library and exploration of a large compound space. For more focused problems, other descriptors, diversity definitions, filters, or even chemical mutation types can be used as "drop-in" replacements for those described here.

Reproduction and Mutation. ACSESS begins a generation by generating novel chemical structures from the previous generation. First, a set of new compounds is produced by "crossover" mutation (Scheme S1), where two "parent" compounds are copied from the library, and each is split into two fragments by cutting a random acyclic bond. Two of the resulting fragments, one from each parent, are then bonded together, and the resulting structure is added back into the library.

After generating crossover mutants, further novel compounds are generated by copying random individual structures from the existing library, stochastically modifying them, and adding the new, modified structures to the library. These "chemical mutations", as shown in Figure 1B, consist of addition/removal of an atom, either a terminal atom (1) or "within" an existing bond (3); creation/removal of a ring bond (4); modification of atom type (5, for example, changing a carbon atom to a nitrogen atom); and modification of bond order (2). The mutation process and the probabilities of each mutation type are shown in Scheme S2. Because the descriptor set used in this study depends only on molecular connectivity (*vide infra*), stereochemical information was not tracked in these calculations. However, for descriptor sets or property calculations that depend upon absolute or relative configuration, stereochemical mutations can be included as well. For such systems, configurations may be inverted, and *cis*–*trans* diastereomers may be isomerized.

Filters. After new molecular structures are generated using the above chemical mutations, those that fall outside the chemical space of interest must be discarded. In this study, we focus specifically on stable, synthetically accessible drug-like molecules in the SMU. Compounds were therefore screened using a combination of chemical subgroup filters (eliminating compounds that contain reactive or hydrolytically labile moieties, such as strained allenes, cumulenes, hemiacetals, aminals, orthesters, etc.), steric strain filters based on generated 3D conformations (for example, removing compounds with nontetrahedral sp^3 carbons), and simple physiochemical filters (XlogP, Lipinski and Veber rules, among others).[24] A complete list of filters is given in the Supporting Information. Because these calculations did not track stereochemical information, the software used to generate 3D geometries was used to generate any energetically reasonable configuration for each structure. More robust *ab initio* stability filters would be an appealing future alternative to the heuristic ones employed here.[34]

Maximally Diverse Subset Selection. At the final stage of each ACSESS generation, only a maximally diverse subset of the remaining molecular structures is retained; all other compounds are removed from the library. These structures are used to seed the next generation of the ACSESS procedure. Because the new library is chosen from both the new "child" compounds from the current generation and the "parent" compounds from the previous generation, the diversity of the library must necessarily improve or at least remain constant after each generation.

Many quantitative definitions of diversity exist, as do methods for selecting small maximally diverse libraries from larger libraries.[30] One common definition of "diversity" is as the average nearest-neighbor chemical space distance within a set of compounds. Given this definition of diversity, a maximally diverse collection can be selected using the "maximin" algorithm, which creates a representative subset by choosing compounds from a larger library one-by-one, such that each new structure has the largest minimum distance to existing compounds in the subset.[32]

A cell-based definition of diversity can be used if the principal components of the chemical space are known.[35] For cell diversity, chemical space is divided into a discrete, multidimensional grid, and diversity is then simply defined as the number of cells that contain at least one chemical structure. A maximally diverse set of compounds can be selected simply by choosing a single structure from each cell.

Chemical Space Descriptors. Any diversity analysis is highly dependent on the descriptor set chosen, which defines the chemical space coordinates of the structures. A large number of chemical descriptor sets exist,[36] ranging from simple counts of topological properties[37] to measures of 3D shape.[38]

For the mapping presented here, chemical space coordinates were calculated using Moreau–Broto autocorrelation descriptors.[39] This well-established set of descriptors encodes structural information from an arbitrary chemical structure into a fixed-length vector[40] and has been successful in diverse tasks, such as defining biologically relevant similarities in large compound sets[41] and correlating structural diversity with biological activity.[42] Autocorrelation descriptors describe topological correlations between atomic properties on a molecule:

$$\mathrm{AC}(d, p) = \sum_{i \le j} p_i p_j \delta(d_{ij} - d) \tag{1}$$

where

$$\delta(d_{ij} - d) = \begin{cases} 1, & \text{if } d_{ij} = d \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

d_{ij} is the number of bonds separating atoms i and j, and p_i is the value of atomic property p on atom i. Here, the properties p include the atomic number, Gasteiger–Marsili partial charge,[43] atomic polarizability,[44] topological steric index,[45] and unity (i.e., $p_i = 1$ for all i); values of d from 0 to 7 were used, providing a 40D chemical space. Note that these descriptors are based solely on molecular topology and do not reflect stereochemistry or 3D structure.

Principal Component Analysis of the SMU. ACSESS was used to construct a small, 2000-compound representative universal library (RUL), employing the maximin method to select maximally diverse subsets (see SI). The 40 autocorrelation vectors of these 2000 molecules were mean centered and normalized to have unit variance,

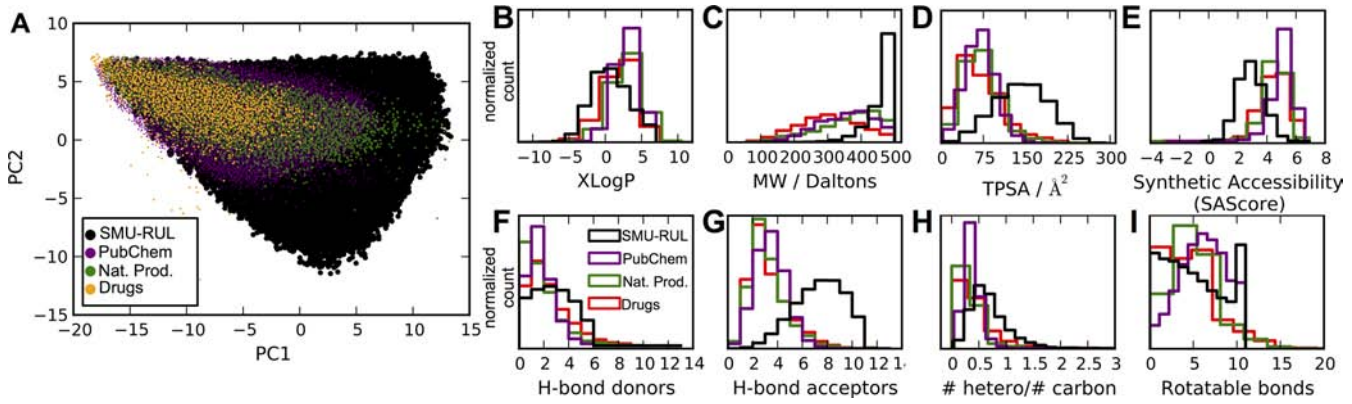


Figure 2. Comparison to existing libraries. The SMU-RUL (black), ZINC natural product library (green), ZINC drug library (orange), and drug-like compounds in PubChem (purple) are shown. (A) Compound locations along the first two principal components of the SMU-RUL library. (B–I) Histograms of physiochemical properties for the four libraries; *y*-axes correspond to normalized compound counts within each library. The properties include: (B) estimated log *P* (XLogP);[54] (C) molecular weight (MW); (D) topologically estimated polar surface area (TPSA);[55] (E) SAScore;[46] (F,G) number of hydrogen-bond donors and acceptors; (H) ratio of noncarbon heavy atoms to carbon atoms; and (I) number of rotatable bonds. Compared to the PubChem database, molecules in the SMU-RUL are, on average, more polar and have a larger molecular weight. Lower synthetic accessibility scores (E) for SMU-RUL compounds are expected because of their novelty and dissimilarity to known compounds.

and principal component analysis (PCA) was performed. Loadings for the first 10 principal components of the SMU are shown in Table S1.

Construction of a Representative Universal Library of the SMU. Next, cell-based diversity was used to construct a large, synthetically optimized representative universal library of the small molecule universe (SMU-RUL). A partitioning scheme was developed based on the largest 10 principal components (PCs) of the SMU. These PCs collectively account for 98.5% of the SMU's chemical space variance. Each PC was then partitioned into bins, with the number of bins proportional to each PC's standard deviation, yielding a $20 \times 15 \times 12 \times 11 \times 9 \times 8 \times 8 \times 6 \times 6 \times 4$ grid that partitions the SMU chemical space into 3.3×10^9 cells. ACSESS was then used to construct the 8.9×10^6 structure SMU-RUL, with maximally diverse subsets selected by choosing one compound per grid cell. In cases where more than one compound was present in a cell, the compound with the highest estimated synthetic accessibility score was selected.

Synthetic Accessibility Scores. The SAScore algorithm estimates a structure's synthetic accessibility based on both its topological complexity and how frequently its substructures appear in large chemical databases.[46] SAScores reported here are based on a substructure analysis of the full ZINC database, and higher SAScores indicate a more facile synthesis. The distribution of SAScores for compounds in the PubChem library is shown in Figure 2E.

Software. ACSESS was implemented in Python 2.7 using OpenEye chemoinformatics toolkits.[47] For screening of 3D geometries, conformers were generated using the OpenEye OMEGA program.[48]

RESULTS

Proof of Principle: the GDB13 Chemical Space. The GDB13 database enumerates all possible compounds of 13 heavy atoms or fewer (within a set of synthetic criteria used as inspiration for the present work) and is currently the largest available database of chemical structures.[24] We first used GDB13 to test the ability of the ACSESS method to capture the diversity of a large molecular ensemble. An RUL of 10 000 compounds in GDB13 (GDB13-RUL) was constructed. Compounds were filtered for synthetic feasibility using the same criteria as in GDB13; the diversity of the set converged after 1000 generations. The 10 000-member GDB13-RUL was as diverse as the fully enumerated GDB13 library but required a factor of 10^4 fewer compounds to be processed computationally than the GDB13 enumeration. Additionally, PCA (and other mappings) of the GDB13-RUL produced diversity metrics similar to those of the fully enumerated library.

These results indicate two important properties of the ACSESS method. First, at convergence of the diversity measure, the RUL captured the full diversity of its parent space. Second, ACSESS generated the GDB13 RUL by enumerating far fewer compounds.

A Representative Library of the SMU. ACSESS was employed to build a representative library of the entire SMU-RUL consisting of 8.9×10^6 structures (database S1), with local optimization for synthetic accessibility. Chemical structures were restricted to 150–500 Da with estimated log $P < 7.0$ and were filtered for reactivity, stability, and drug-likeness. Chemical space coordinates were computed using Moreau–Broto autocorrelation descriptors.[36] A total of 3.6×10^9 structures were screened.

Structures in the SMU-RUL represent a widely spaced mesh over the complete SMU chemical space as defined above. It is important to note, however, that the specific set of compounds in the SMU-RUL is not unique. Instead, each SMU-RUL compound indicates the existence of minimally one, and, on average, $\sim 10^{53}$ related structures (given the estimate of 10^{60} possible SMU structures)[4] in a particular region of chemical space. Convergence of the diversity measure, while not indicating that a global optimum has been reached, shows that we have obtained a set of structures that is representative of the accessible chemical space.

Comparison to Existing Databases. The chemical space coverage of the SMU-RUL was compared to that of three existing databases: PubChem, a database of $>3 \times 10^7$ pure chemical compounds;[49] ZINC natural products, a database of 2×10^5 natural products and metabolites relevant to drug discovery; and ZINC drugs, a database of >7000 approved drugs.[50] For comparison to the SMU-RUL, databases were filtered according to drug-likeness and atom content using a subset of the SMU-RUL filters (see SI).

Although 99.9% of the generated SMU-RUL structures obey Lipinski's rules for drug-likeness,[51] only 11 000 are present in the PubChem database. The scaffolds in the SMU-RUL (defined as the set of atoms that are in or link the molecule's ring systems)[29,49] are also highly novel. Of the 5.1×10^6 unique

dx.doi.org/10.1021/ja401184g | J. Am. Chem. Soc. 2013, 135, 7296–7303

scaffold topologies in the SMU-RUL, only 23 000 are found among the 3.2 × 10^6 scaffold topologies in the PubChem database. Interestingly, the SMU-RUL also contains two known drugs, acetanilide and phenytoin.

Figure 2A shows the chemical space occupied by SMU-RUL compounds along its first two principal components as well as the positions of compounds from the ZINC and PubChem databases. Even in this 2D projection of the 40-dimensional chemical space, the SMU-RUL covers a much larger region of chemical space than existing chemical libraries. Figure 2B–I shows the distribution of eight physiochemical properties in the four sample libraries. Most strikingly, the SMU-RUL, on the whole, contains heavier, more polar and synthetically more challenging members (given current methodologies) than existing compound libraries (Figure 2B–G).

The property distributions in Figure 2 show that much of the available SMU diversity is concentrated at higher molecular weights and more polar structures than currently known compounds. This is not surprising, given the autocorrelation descriptors used to describe diversity here; larger compounds can support a larger range of chemical functions, allowing a wider range of descriptor values to be explored. The relatively low synthetic accessibility scores (SAScores) of the SMU-RUL structures are also expected. These scores indicate that in many regions of chemical space there were few compounds with similar substructure to known compounds.

Self-Organizing Map of the SMU. A self-organizing map (SOM) was constructed from the SMU-RUL to visualize the high-dimensional SMU chemical space (Figure 3). SOMs have a rich history in chemical diversity analyses.[52] An SOM consists of a lattice of "neurons", each associated with a chemical space coordinate.[22,23] The SOM is randomly presented with "cue" coordinates from the training set (here, the SMU-RUL). For each cue, the neuron with coordinates closest to the cue is said to "fire," and it and its neighbors' coordinates are adjusted in the direction of the cue. Iteration of this procedure creates a low-dimensional representation of the high-dimensional chemical space.

A toroidal 300 × 300 SOM was trained using the autocorrelation chemical space coordinates of the SMU-RUL. Each SMU-RUL compound was then assigned to its closest neuron. The compounds were spread relatively evenly throughout the map, with an average of 98.5 ± 25.3 (and at least 16) chemical structures assigned to each neuron. A small region of the map (region EI in Figure 3B) corresponds to relatively low molecular weight structures with 15–20 heavy atoms, while others correspond to higher molecular weights nearer to the 500 Da limit. Figure 3D shows well-defined regions containing either large, fused ring systems or simpler monocyclic and fused bicyclic structures. Variations of other topological and physiochemical properties over the map are shown in Figure 3C–F and Figure S3A.

The autocorrelation vectors of all PubChem library compounds were computed and assigned to neurons on the SOM (Figure 3A). The PubChem compounds were concentrated in a very restricted area compared to the SMU-RUL, with 98% of PubChem compounds assigned to 2% of the neurons. The most significant PubChem compound cluster is centered on region EI and is characterized by compounds with low molecular weights and few rings. The cluster can be further divided into two regions: one corresponding to rigid structures without rotatable bonds and the other to more flexible molecules. A smaller cluster in region DIV corresponds to compounds with higher molecular weights and more complex scaffolds.

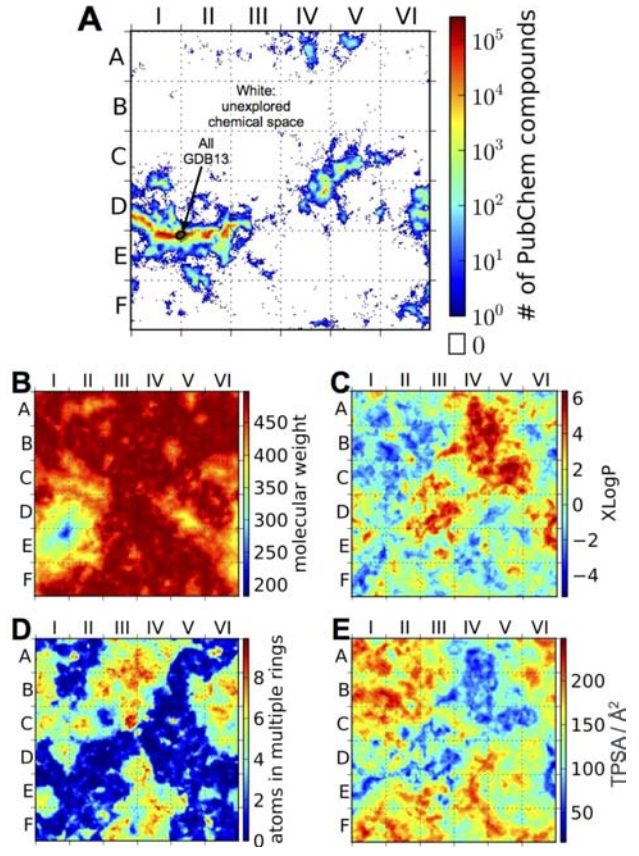


Figure 3. Map of the small molecule universe. A 300 × 300 toroidal SOM was created using normalized autocorrelation descriptors of SMU-RUL compounds. For clarity, the map is divided into 36 labeled sections (AI, BII, etc.), each containing a 50 × 50 grid of neurons. (A) Number of PubChem compounds assigned to a neuron; white indicates neurons which are unoccupied by any PubChem compounds (84% of total). The PubChem compounds are highly clustered to a relatively small region of chemical space; 98% are assigned to only 2% of the neurons. The black circle in region EI encompasses the positions of all GDB13 compounds. (B–D) Molecular properties; each neuron is colored by the median value of its SMU-RUL compounds.

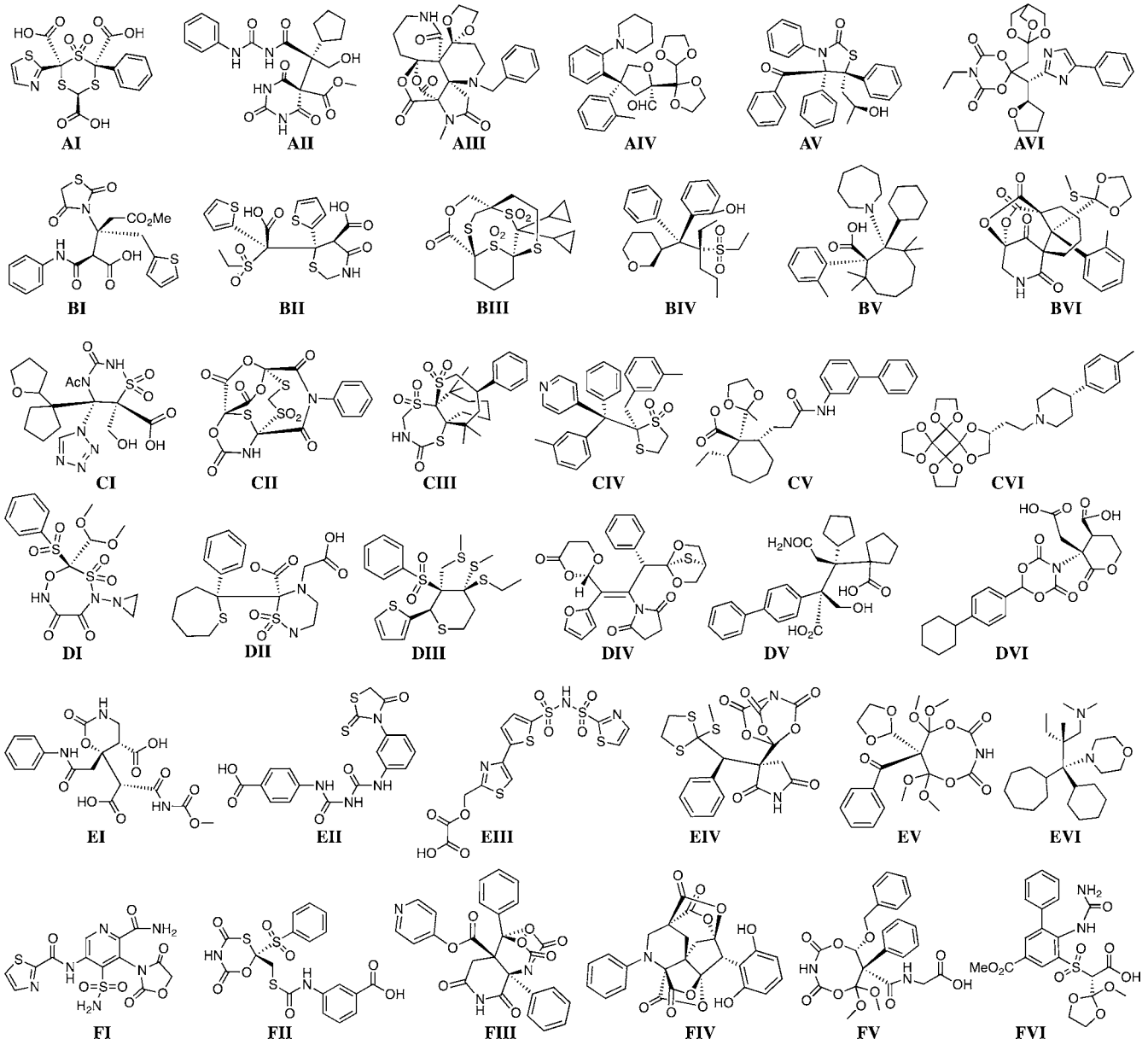
Large regions of chemical space populated by SMU-RUL structures are unrepresented in PubChem (white spaces in Figure 3A). Note that the inverse is not true; a SOM constructed using compounds from both PubChem and the SMU-RUL shows that SMU-RUL structures occupy all of the space occupied by PubChem compounds (Figure S3B). The unexplored regions of chemical space were, like the SMU-RUL in general, almost entirely drug-like based on Lipinski's rules. Examples of SMU-RUL structures from unexplored portions of chemical space (Chart 1 and Figure S4) include complex ring structures (AIII, BVI), many simple ring systems (AV), bridged macrocycles (CIII), and high heteroatom content compounds (CI–CII).

■ CONCLUSIONS

The stochastic exploration described here is a computationally efficient tool for accessing the astronomical number of feasible organic structures. As a comparison of stochastic and enumerative approaches, all 970 × 10^6 compounds from the

dx.doi.org/10.1021/ja401184g | J. Am. Chem. Soc. 2013, 135, 7296–7303

Chart 1. SMU-RUL Compounds from Unexplored Chemical Space[a]



[a]Each compound shown here was selected from a SOM map neuron unoccupied by any PubChem compounds and was among the most synthetically accessible compounds assigned to the neuron. Letters/numerals refer to the regions shown in Figure 3. The stereochemical assignments shown reflect the generated 3D conformations, which are shown as ball-and-stick models in the SI.

enumerated GDB13 library were assigned to the SMU-RUL SOM. In the low-molecular weight portion of the map, 98% of GDB13 compounds were assigned to just 10 neurons, and the GDB13 compounds overall occupy a total of only 61 adjacent neurons, 0.07% of the total (Figure 3A). The combinatorial explosion of new molecules available at higher molecular weights is simply not accessible in the smaller chemical spaces amenable to enumeration.

Importantly, in our stochastic exploration, large gaps were observed in the currently known compound collections. There has never been an attempt made to explore the full range of chemical diversity, either by nature or by man. Nature uses readily available building blocks and biosynthetic tools to develop structural motifs and arguably has employed repetitive patterns and quantum leaps in molecular weight (biopolymers) to address the diversity intense aspects of data storage, immune defense, scaffolding, etc. Laboratory synthesis relies on a nucleation-based building block approach, using iterative bond formations and a limited pool of available reagents. In the absence of obvious incentives otherwise, laboratory synthesis thus emphasizes simplicity and uses small functional group-specific tools to carve out niches around known biologically active scaffolds.

The ACSESS algorithm makes two important contributions to chemical space exploration, both of which are immediately available for further experimentation. First, the gaps identified in the known chemical universe may now be explored systematically. Second, ACSESS allows the mining of chemical libraries that do not yet exist, providing a near-infinite source of novel compounds. For instance, we have used ACSESS to

dx.doi.org/10.1021/ja401184g | J. Am. Chem. Soc. 2013, 135, 7296−7303

search a chemical space of unprecedented size to create a library of compounds with high similarity to bretazenil, a benzodiazepine anxiolytic drug discovered in 1988. Similarity here was defined using the Tanimoto coefficient of the PubChem-format fingerprints.[53] The resulting library represents both a universal library of structural isomers of the target drug and a collection of novel, unpatented candidates for future development. Because of ACSESS's efficiency, more computationally intensive metrics than structural similarity can be employed, affording opportunities for molecular discovery in fields well beyond biology and medicine.

■ ASSOCIATED CONTENT

Supporting Information

SMU-RUL compounds in SMILES format; ACSESS source code; detailed methods and chemical space definitions; computational details of library construction; SOM and PCA analysis of SMU-RUL; construction of bretazenil isomer library; GDB13 proofs-of-principle. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

weitao.yang@duke.edu; david.beratan@duke.edu

Present Address

[#]Laboratoire de Chimie Théorique, Université Pierre et Marie Curie, Paris, France

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Professor Sean Xie for helpful discussion and the NIH for support of the UPCMLD (P50-GM067082).

■ REFERENCES

(1) *Beyond the Molecular Frontier: Challenges for Chemistry and Chemical Engineering*; The National Academies Press: Washington, D.C., 2003.
(2) Sauer, W. H. B.; Schwarz, M. K. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 987.
(3) Schreiber, S. L. *Nature* **2009**, *457*, 153.
(4) Dandapani, S.; Marcaurelle, L. A. *Nat. Chem. Biol.* **2010**, *6*, 861.
(5) Bohacek, R. S.; McMartin, C.; Guida, W. C. *Med. Res. Rev.* **1996**, *16*, 3.
(6) Wester, M. J.; Pollock, S. N.; Coutsias, E. A.; Allu, T. K.; Muresan, S.; Oprea, T. I. *J. Chem. Inf. Model.* **2008**, *48*, 1311.
(7) Triggle, D. J. *Biochem. Pharmacol.* **2009**, *78*, 217.
(8) Tan, D. S. *Nat. Chem. Biol.* **2005**, *1*, 74.
(9) Thomas, G. L.; Wyatt, E. E.; Spring, D. R. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 700.
(10) Hajduk, P. J.; Galloway, W. R. J. D.; Spring, D. R. *Nature* **2011**, *470*, 42.
(11) Brown, L. E.; Cheng, K. C.-C.; Wei, W.-G.; Yuan, P.; Dai, P.; Trilles, R.; Ni, F.; Yuan, J.; MacArthur, R.; Guha, R.; Johnson, R. L.; Su, X.-Z.; Dominguez, M. M.; Snyder, J. K.; Beeler, A. B.; Schaus, S. E.; Inglese, J.; Porco, J. A., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 6775.
(12) Dow, M.; Fisher, M.; James, T.; Marchetti, F.; Nelson, A. *Org. Biomol. Chem.* **2012**, *10*, 17.
(13) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 316.
(14) Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P. *Trends Biotechnol.* **2009**, *27*, 18.
(15) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X.-P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. *Nature* **2012**, *492*, 215.
(16) Zablocki, J. *J. Am. Chem. Soc.* **2007**, *129*, 12586.
(17) Gillet, V. J. *Struct. Bonding (Berlin)* **2004**, *110*, 133.
(18) Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, D. V. S. *J. Mol. Graphics Modell.* **2002**, *20*, 491.
(19) Hu, X. Q.; Beratan, D. N.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 064102.
(20) Balamurugan, D.; Yang, W.; Beratan, D. N. *J. Chem. Phys.* **2008**, *129*, 174105.
(21) Wang, M.; Hu, X. Q.; Beratan, D. N.; Yang, W. *J. Am. Chem. Soc.* **2006**, *128*, 3228.
(22) Brown, N.; McKay, B.; Gasteiger, J. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 761.
(23) van Deursen, R.; Reymond, J.-L. *ChemMedChem* **2007**, *2*, 636.
(24) Blum, L. C.; Reymond, J.-L. *J. Am. Chem. Soc.* **2009**, *131*, 8732.
(25) Fink, T.; Bruggesser, H.; Reymond, J.-L. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504.
(26) Fink, T.; Reymond, J.-L. *J. Chem. Inf. Model.* **2007**, *47*, 342.
(27) Luethi, E.; Nguyen, K. T.; Bürzle, M.; Blum, L. C.; Suzuki, Y.; Hediger, M.; Reymond, J.-L. *J. Med. Chem.* **2010**, *53*, 7236.
(28) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. *ChemMedChem* **2008**, *3*, 1520.
(29) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. *J. Chem. Inf. Model.* **2008**, *48*, 1304.
(30) Farnum, M. A.; Desjarlais, R. L.; Agrafiotis, D. K. In *Handbook of chemoinformatics: from data to knowledge*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. *4*, p 1640.
(31) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169.
(32) Agrafiotis, D. K. *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 841.
(33) Gillet, V. In *Molecular Diversity in Drug Design*; Dean, P., Lewis, R., Eds.; Springer: The Netherlands: 2002, p 43.
(34) Hoffmann, R.; Schleyer, P. v. R.; Schaefer, H. F., III *Angew. Chem., Int. Ed.* **2008**, *47*, 7164.
(35) Xue, L.; Stahura, F. L.; Bajorath, J. In *Methods Molecular Biology*; Bajorath, J., Ed.; Humana Press: New York, 2004; Vol. *275*, p 279.
(36) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2nd ed.; Wiley-VCH: Weinheim, 2009.
(37) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. *ChemMedChem* **2009**, *4*, 1803.
(38) Arteca, G. A. In *Reviews in Computational Chemistry*; John Wiley & Sons, Inc.: Hoboken, NJ, 2007; Vol. *9*, p 191.
(39) Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.
(40) Gasteiger, J. In *Handbook of Chemoinformatics*; Wiley-VCH Verlag GmbH: 2003, p 1034.
(41) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 1205.
(42) Matter, H. *J. Med. Chem.* **1997**, *40*, 1219.
(43) Gasteiger, J.; Marsili, M. *Tetrahedron Lett.* **1978**, *19*, 3181.
(44) Miller, K. J.; Savchik, J. *J. Am. Chem. Soc.* **1979**, *101*, 7206.
(45) Cao, C.; Liu, L. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 678.
(46) Ertl, P.; Schuffenhauer, A. *J. Chemoinf.* **2008**, *1*, 8.
(47) *OEChem 1.7.5; OMEGA 2.4.4; MolProp 2.1.2.*, OpenEye Scientific Software, Inc.: Santa Fe, NM USA, www.eyesopen.com, 2012.
(48) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. *J. Chem. Inf. Model.* **2010**, *50*, 572.
(49) Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887.
(50) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. *J. Chem. Inf. Model.* **2012**, *52*, 1757.
(51) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.
(52) Sadowski, J.; Wagener, M.; Gasteiger, J. *Angew. Chem., Int. Ed.* **1996**, *34*, 2674.
(53) *PubChem Fingerprint - NCBIFTP site*; National Institutes of Health: Bethesda, MD; ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt.

(54) Wang, R.; Ying, F.; Lai, L. J. *Chem. Info. Comput. Sci.* **1997**, *37*, 615.
(55) Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.