Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds.

Aaron M. Virshup[1], Julia Contreras-García[1], Peter Wipf[2], Weitao Yang[1*] and David N. Beratan[1*]

[1,2]Center for Chemical Methodologies and Library Development, [1]Department of Chemistry, Duke University, Durham, NC 27708, [2]Department of Chemistry, University of Pittsburgh, Pittsburgh, PA 15260

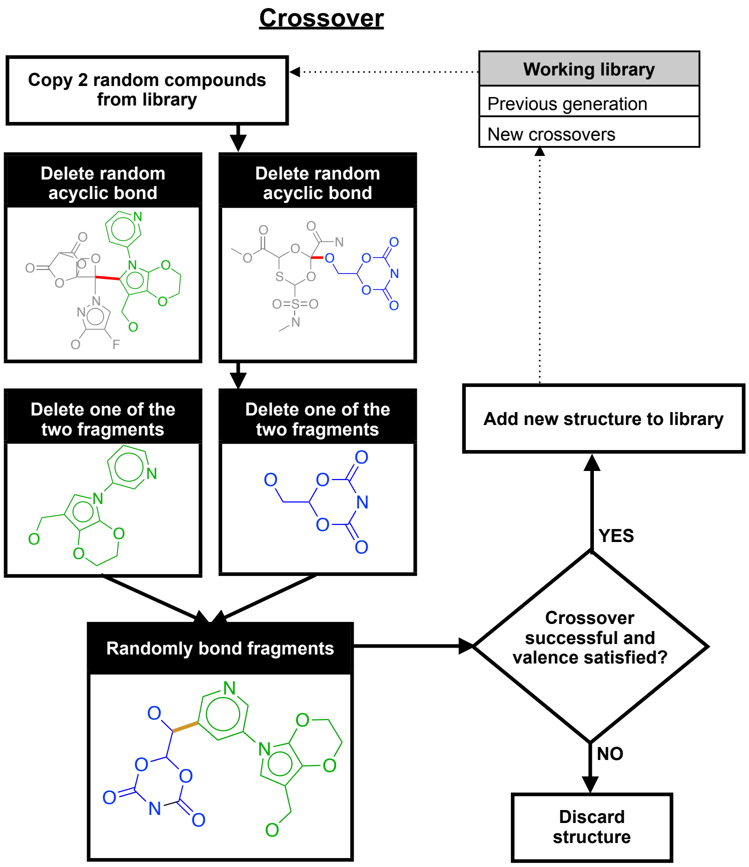
S1. SI Methods

The ACSESS algorithm is based on a genetic algorithm-like framework. However, given a "fitness function", genetic algorithms (GAs) search for individual solutions with maximum fitness through selective breeding and mutation. ACSESS, in contrast, generates an optimal *set* of individuals – a set of chemical compounds that span the chemical space of interest. The algorithm is based on repetition of three steps: 1) chemical reproduction and mutation, 2) filtering, and 3) selection of a maximally diverse subset.

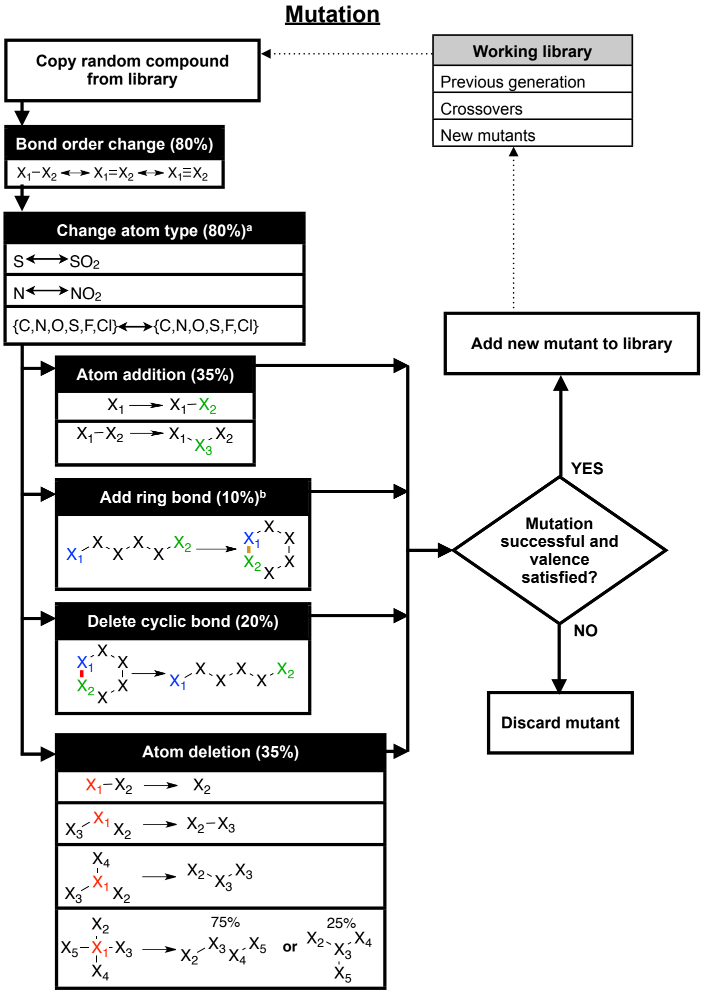
1) Reproduction and mutation. ACSESS explores chemical space through the repeated application of a set of "genetic operators" that take the form of chemical structure modifications.

Each generation of the ACSESS algorithm begins with a series of crossover mutations. These operations are common features of GAs, employed to speed traversal of the space of interest. For crossovers in ACSESS, two molecules are randomly selected from the library, and then copied; each of the copies is split into two fragments by breaking a randomly chosen acyclic bond; and a new molecule is created by bonding one of the fragments of the first molecule to one of the second. This newly created hybrid molecule is then added to the library (Scheme S1).

Following these crossovers, random chemical mutations are employed to expand the diversity of the library and explore each compound's local chemical space. For each mutation, a molecule is randomly chosen and copied from the current library and a random chemical mutation applied to it. The mutated copy is then added to the library (Scheme S2).


Crossover
Copy 2 random compounds from library
Working library
Previous generation
New crossovers
Delete random acyclic bond
Delete random acyclic bond
Delete one of the two fragments
Delete one of the two fragments
Add new structure to library
YES
Randomly bond fragments
Crossover successful and valence satisfied?
NO
Discard structure

Scheme S1. Method for generating “crossover” structures.



Scheme S2. Method for generating mutants. Each mutation type is shown with its probability of being applied. After each compound is copied from the library, a single bond order may be modified (80%) and a single atom type may be changed (80%). Finally, ONE of the final mutation types may be applied. In each case, genetic operators may only be applied if allowed by valence rules; otherwise, the mutation will fail and the mutant is discarded. **a)** In **atom type mutation**, sulfur and nitrogen may also be respectively changed into, or created from, sulfone and nitro groups, which are represented internally as specialized atom types. Note that the oxygen atoms in these two groups may not be bonded to or modified by any other genetic operator. **b)** In creating new **ring bonds**, the randomly selected atom is bonded to another that is within 5-7 bonds, creating a new ring of 6-8 atoms.

The mutations employed take the form of chemical structure modifications (Figure 1B), including creation or deletion of an atom, creation or deletion of a cyclic bond, elemental mutation (e.g., changing a carbon to a nitrogen), and changes of bond order. A "crossover" operation is also employed to speed up chemical space exploration, in which a new compound is created by bonding fragments of two "parent" compounds. Several of the mutation types were inspired by Reymond's chemical space travel algorithm (1).

Allowed atoms (and their corresponding valences) are C (4), N (3), O (2), S (2), F (1) and Cl (1). Exceptions to these valence restrictions are made in the cases of sulfone and nitro groups. Hydrogens are added as necessary to fill any empty valence. With the exception of the nitrogen in nitro groups, all atoms are assumed to have a formal charge of 0.

2) Compound filtering. After generating new compounds by mutation and crossover, the new compounds are filtered; any compound that falls outside the chemical space of interest is removed (see below for chemical space and filter definitions). In addition to testing each molecule for inclusion in the target chemical space, molecules are checked for stability, as assessed by a set of filters, and targeted property values. For instance, in mapping the SMU, compounds with molecular weight > 500 Daltons and/or containing unstable functional groups were removed from the library. For mapping the set of benzodiazepine-like compounds, those with similarity coefficients of 70% or less were removed.

3) Diverse subset selection. A maximally diverse subset of the candidate library is selected, and a numerical measure of its chemical space coverage computed. In general, almost any method for selecting a representative subset (RS) and a corresponding measure of diversity may be used in this step; many methods for both measurement and optimization of library diversity exist (2). RSs are small collections of compounds that retain the full chemical space diversity of a larger compound library.

Chemical space coordinates were calculated using Moreau-Broto autocorrelation descriptors, which are based on correlations between atomic properties in a molecule:

$$AC(d,p) = \sum_{i \leq j} p_i p_j \delta\left(d_{ij} - d\right) \quad , \qquad (1)$$

where 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* included atomic number, Gasteiger charge, polarizability, topological steric index, and unity (i.e., $p_i = 1$ for all *i*); values of *d* from 0 to 7 were used, giving a 40-dimensional chemical space.

Two types of diverse subset selections were employed and are described below: the maximin algorithm and cell-based partitioning. Cell-based partitioning is appealing because it is computationally efficient for large compound libraries. However, it requires prior PCA of the chemical space. Therefore, in practice, the maximin algorithm is first used to construct a small RUL of the chemical space, which permits PCA. Cell-based diversity can then be used, allowing the construction of much larger libraries.

3.1) Chemical space distances and maximin selection. Because the 40 components of this vector have disparate physical units and numerical magnitudes, each was periodically rescaled by a multiplicative factor to set its standard deviation within the RS to unity. The Cartesian chemical space distance r_{nm} between compounds *n* and *m* is therefore given by:

$$r_{nm} = \sqrt{\sum_{k=1}^{40} \left(\frac{r_{i,n} - r_{i,m}}{\sigma_i} \right)^2} , \tag{2}$$

where $r_{i,n}$ is the value of the i^{th} descriptor for compound n, and σ_i is the standard deviation of the i^{th} descriptor. The definition of distance within chemical space therefore changes over the course of the calculation as the variance of each descriptor converges to its final value (e.g., Figure S1B and Figure S2B).

The RS was selected to maximize the diversity of the library as measured as the root-mean-squared nearest-neighbor distance between compounds (2):

$$D = \frac{1}{N} \sqrt{\sum_{n=1}^{N} \min_{n \neq m} \left(r_{nm}{}^2 \right)} , \tag{3}$$

where the sum over indices n runs over each compound in the library and N is the number of compounds in the library.

For ease of implementation and computational efficiency, the “maximin” algorithm (2) was employed to generate small RSs from the “parent” library. Maximin first seeds the RS with a randomly chosen compound from the larger library. The algorithm then adds new compounds to the RS one by one. Each additional compound selected is the one with the maximal distance to all compounds already in the RS. This protocol therefore constructs a set with near-maximal nearest-neighbor distances or dissimilarities that approximately maximizes the diversity measure in eq. (3).

3.1) Cell-based partitioning. The maximin algorithm has a computational complexity of N^2, which, in practice, creates severe computational bottlenecks in the ACSESS algorithm for libraries larger than ~10^5 compounds. Cell-based partitioning, which scales linearly with the number of compounds, is therefore used to create the larger RULs. Cell-based partitioning is acommonly used method for selecting a maximally diverse subset of compounds (3).

The partitioning here is based on dimensionality reduction of the chemical space via PCA. Each of the chosen PCs is split into a number of bins, proportional to the square root of the variance associated with each PC. This effectively partitions chemical space into a grid of hyper-rectangular cells. To create a maximally diverse set of compounds, a maximum of one compound from each cell is selected. Diversity here is defined as the proportion of occupied cells.

This scheme also allows local property optimization within each cell. If more than one compound occupies a given cell, the one with the most optimal chosen property (e.g., synthetically accessibility) is selected.

4) Implementation. A Python implementation of this algorithm, which utilizes OpenEye software (4), is available as supporting database S5.

S2. Substructure filters and property limits

In the filtering step of the ACSESS algorithm, compounds are discarded if they fall outside the target chemical space. In this work, the SMU chemical space is defined both by molecular weight (compounds with MW 150-500 Da) and a series of filters for synthetic accessibility and drug-likeness.

GDB13 filters. All filters used in the construction of GDB13 were also employed here; an extensive description of these filters may be found in the original GDB11 and GDB13 publications (5, 6). These filters discard compounds that contain non-planar graphs, a variety of unstable chemical or readily hydrolysable moieties (e.g., strained allenes, cumulenes, hemiacetals, aminals, orthoesters, and carbamic acids), or tetravalent carbon centers with non-tetrahedral geometry.

Graph planarity was assessed using Boyer and Myrvold's publicly available graph embedding algorithm (7).

To screen for non-tetrahedral carbon centers, 3-dimensional conformations were generated using OpenEye Omega software (4), and the lowest-energy conformation was then screened. Compounds were discarded if the software failed to generate a 3D conformation.

Stability and synthetic accessibility filters. As SMU-RUL compounds are generally much larger than those in GDB13, additional synthetic accessibility filters were implemented. Several of these filters are based on a smallest set of smallest rings (SSSR) decomposition of each molecule (8). Compounds were discarded if they contained:

1. more than 7 SSSR rings;
2. more than one SSSR ring of 8 or more atoms;
3. rings containing more than one bridge;
4. unsaturations in bridges (except in bicyclooctene);
5. ring systems which cannot be uniquely decomposed into rings and bridges;
6. non-planar sp^2 systems (defined as systems X=C(Y)Z in which atom X is more than 0.15 Å outside the plane defined by atoms C, Y and Z);
7. non-linear sp systems (systems A≡BC in which the angle $\angle ABC<178.5^{\circ}$);
8. terminal sulfur atoms (except in thiourea and rhodanine);
9. single bonds between sulfur and oxygen;
10. single bonds between sulfur and nitrogen (except in aromatic sulfonamides);
11. more than 3 stereogenic carbons; or
12. multiple triple bonds separated by less than 8 bonds.

Druglike filters. Compounds were removed if they contained:

1. no nitrogen or oxygen atoms;
2. more than 5 halogens;
3. more than 2 aldehydes;
4. more than one methylidene group;
5. more than 2 unconjugated double bonds;
6. more than 3 basic amines;
7. XLogP>7.0;
8. more than 10 hydrogen bond acceptors;
9. more than 5 hydrogen bond donors;
10. more than 11 rotatable bonds;
11. enol ethers, acyl-halides, anhydrides, beta-heterosubstituted carbonyls, perhalo-ketones, unsubstituted hexane chains, or halopyrimidines;
12. a heteroatom adjacent to (but not part of) an sp system;
13. cyclic sulfur bonded to fused or bridgehead nitrogens;
14. heteroaromatic halogens;
15. non-aromatic halogens (except in aromatic trihalomethyl groups);
16. double bonds to terminal atoms (except in sulfones); or
17. chains of 3 or 4 aromatic nitrogens (except in triazoles and tetrazoles).

S2.1. Public library screens

The PubChem Compound and ZINC drug and natural product libraries contain compounds that violate the specific SMU definition used here. Compounds from these libraries were therefore not considered if they had:

1. MW < 150 Da or MW > 500 Da;
2. atoms other than H, C, N, O, S, F, or Cl;
3. XLogP>7.0;
4. more than 10 hydrogen-bond acceptors;
5. more than 5 hydrogen-bond donors;
6. more than 11 rotatable bonds;
7. more than 5 halogens; or
8. no heteroatoms.

These criteria are a subset of those used to screen the SMU-RUL generated by ACSESS. 56% of the 3.6×10^7 Pubchem compounds, 77% of the 7240 ZINC drug compounds, and 85% of the 2.0×10^5 ZINC natural products were retained.

S3. GDB13 benchmarks

To benchmark ACSESS's performance, two ACSESS-generated representative universal libraries of the GDB13 chemical space were compared to comparable sets generated from the complete enumerated GDB13 library of 970,000,000 compounds.

Benchmark 1: GDB13 chemical space. ACSESS was employed to create a 10,000-member representative subset of GDB13. To generate a representative subset of the

GDB13 chemical space, compounds were limited to 13 heavy atoms, and synthetic feasibility filters similar to those employed in the construction of GDB13 were employed (see section S2). The following protocol was used:

1. Four separate 10,000-compound libraries were generated over 100 generations, with 5,000 mutants and 1,000 crossovers generated in each generation; all GDB13 synthetic filters EXCEPT for 3D geometry screening were employed after generating mutants. A maximally diverse set of 10,000 compounds ("library **A**") was then selected from these results.
2. Library **A** was used to seed four new libraries, which again were evolved over 100 generations. A second maximally diverse library, "library **B**," was then selected from these results.
3. Again, library **B** was used to seed four new libraries, which were evolved over 100 generations, but with the inclusion of 3D geometry screening. A maximally diverse set of these results was selected as the final GDB-RUL.

The evolution of the libraries' diversities is shown in Figure S1A. The final diversity of the GDB-RUL was 8.07. For comparison, a maximally diverse library of 10,000 compounds was selected from the fully enumerated GDB13 library. The library selected from GDB13 had a diversity of 7.38, which is in fact less than that of the GDB-RUL library. We attribute this difference to slight variations in the implementation of the synthetic filter set. However, as shown in Figure S1B-D, the GDB-RUL chemical space closely reproduces that of the full GDB13.

Benchmark 2: RUL of GDB13 structural analogs of mercaptopurine. To benchmark ACSESS's ability to map property spaces, a representative universal library of mercaptopurine-like compounds in GDB13 was generated. Mercaptopurine (Figure S1E inset), or 6-MP, is a widely used immunosuppressive drug containing only 10 heavy atoms.

This RUL is defined both by the GDB13 database constraints and a threshold similarity to 6-MP. Similarity was measured as the Tanimoto coefficient of molecular fingerprints based on the PubChem fingerprint, and molecules were included in the library if their Tanimoto coefficient with mercaptopurine was greater than 0.70.

ACSESS was seeded with the GDB13-RUL (described above), and the algorithm ran for 1,000 generations; a total of approximately 7×10^6 compounds were screened. The cutoff value for Tanimoto similarity was linearly ramped from 0.2 to 0.7 over the first 750 generations. The algorithm constructed an RUL of 400 compounds, which occupy approximately the same chemical space as the 32,000 mercaptopurine-like compounds in GDB13 (Figure S1E). In addition to the RUL, over 9,500 compounds meeting both the similarity criteria and the GDB13 screening criteria were identified.

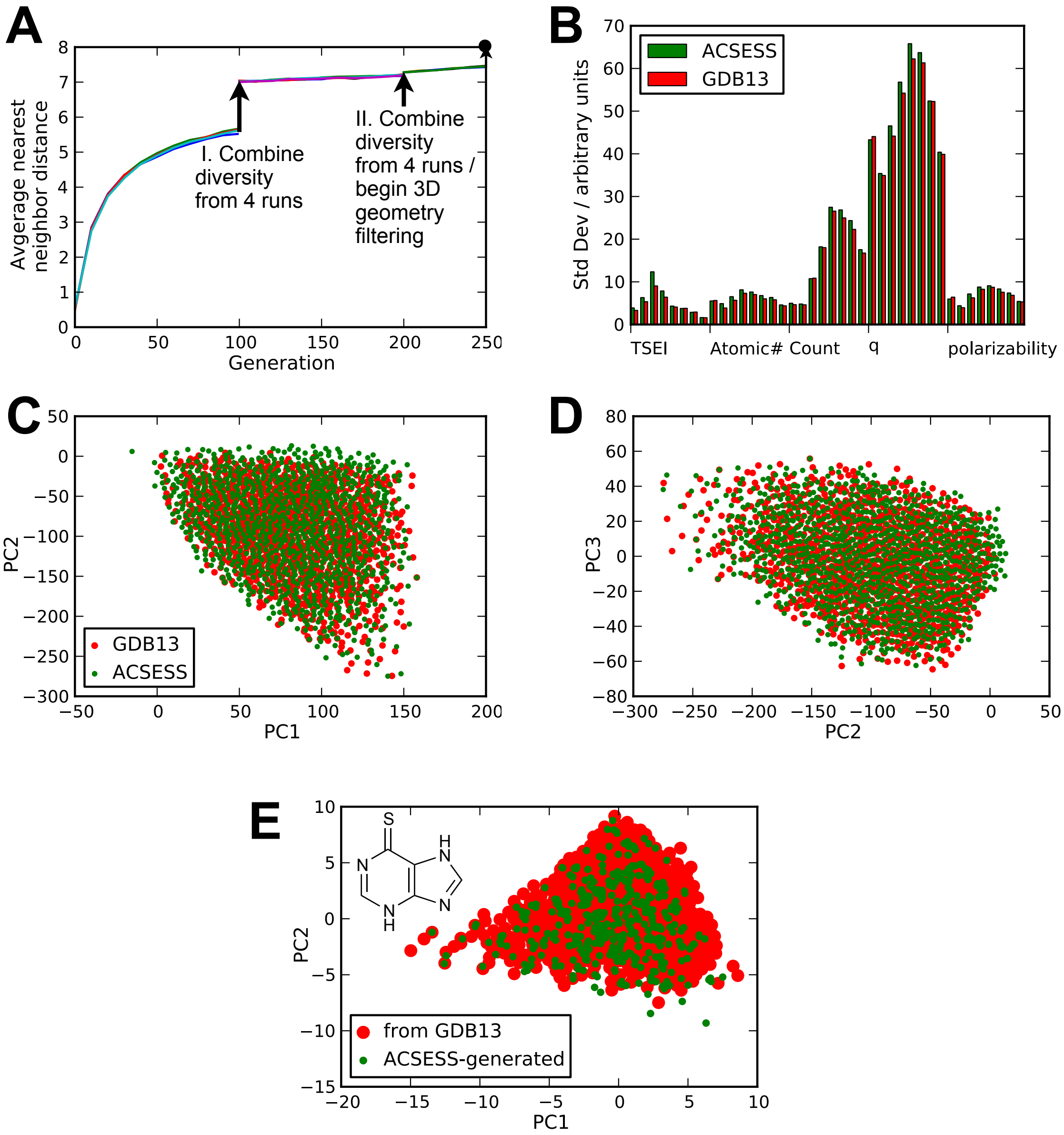


Figure S1. GDB13-RUL. A, nearest-neighbor diversity of ACSESS library over 250 generations. For computational efficiency, 3D geometries were generated and checked for steric clashes only during the final 50 iterations. **B**, standard deviations of each of the 40 autocorrelation descriptors in the final ACSESS library (green) and a maximally diverse set of compounds selected from GDB13 (red). The 8 autocorrelation components for each atomic property (topological steric index or "TSEI", atomic number, atom count, partial charge or "q", and atomic polarizability) are grouped and arranged from left to right. **C,D**, projections of ACSESS (green) and GDB13 (red) compounds into GDB13 principal components. The ACSESS compounds are uniformly distributed over the GDB13 chemical space. **E,** PCA comparison of mercaptopurine (inset) analogs generated by ACSESS (green) and found in GDB13 (red).

S4. SMU mapping

S4.1. Initial mapping

An initial SMU-RUL was constructed using the maximin diversity algorithm and Moreau-Broto autocorrelation descriptors. PCA analysis of this set was then used to construct the grid for the secondary diversity calculation, as described in the main text.

Beginning with an initial seed selection of cyclohexane and benzene, 1,200 mutants and 1,400 crossover compounds were generated at each generation over 5,000 generations. In each generation, a maximally diverse subset of 2,000 compounds was chosen to seed the following generation. Convergence of the library's diversity and descriptor distributions and the results of the PCA analysis are shown in Figure S2 and Table S1.

S4.2. SMU-RUL maps

Additional SOM property maps are shown in Figure S3, and generated 3-dimensional structures of the compounds shown in Chart 1 are shown in Figure S4.

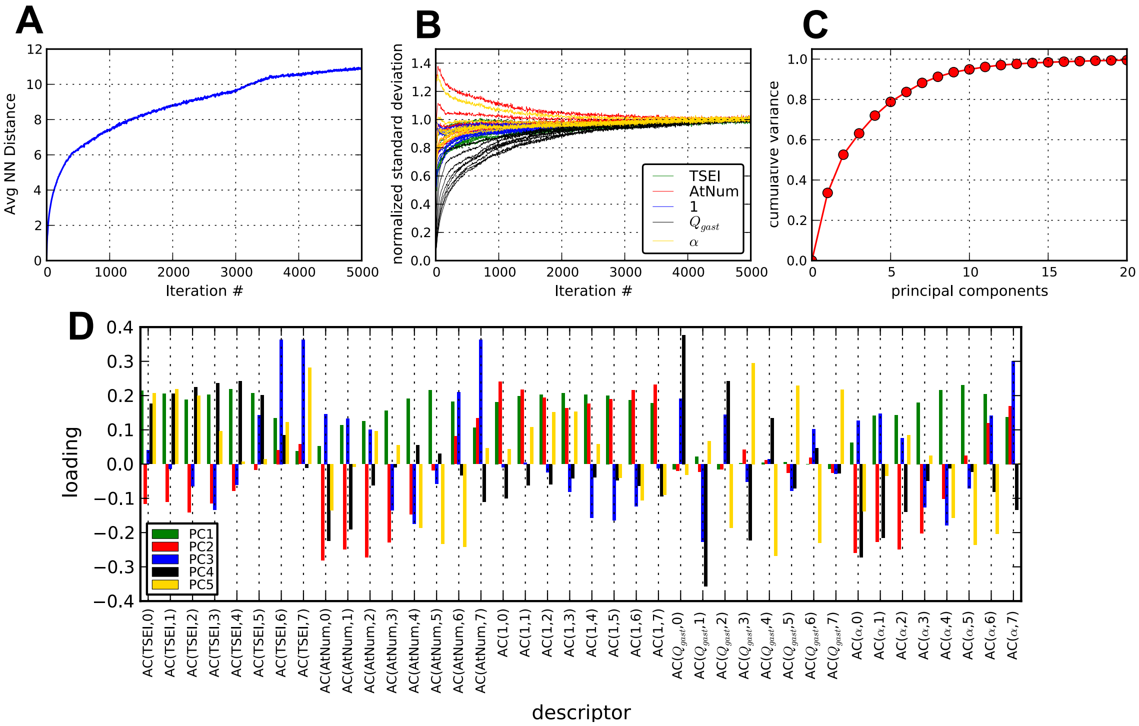


Figure S2. Construction and PCA of a 2,000-structure SMU-RUL. A, convergence of diversity over 5,000 generations. **B**, standard deviations of each descriptor over 5,000 generations. Each property type (topological steric effect index, green; atomic number, red; unity, blue; Gasteiger partial charge, black; and atomic polarizability, yellow) corresponds to 8 individual descriptors, corresponding to correlation lengths from 0-7. Descriptor standard deviations are shown as a proportion of their final values. **C**, cumulative proportion of total variance explained by the first 20 principal components derived from PCA. PCA was performed on the autocorrelation vectors of the 2,000 structures. Each of these descriptors was first mean-centered, then normalized to unit variance. **D**, PCA loadings for the first five principal components, where AC(*p*,*d*) gives the loading for the *normalized* autocorrelation descriptor corresponding to atomic property *p* with correlation length *d*. The properties used here are topological steric index (TSEI), atomic number (AtNum), unity (1), Gasteiger charge (Q_{gast}), and atomic polarizability (α). The full loadings for the unnormalized descriptors are provided in Table S1.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Variance	13.759	7.468	4.189	3.523	2.709	1.969	1.748	1.219	0.905	0.548
AC(TSEI,0)	2.410	-1.255	-0.692	-2.076	2.250	0.757	0.143	0.021	1.350	0.479
AC(TSEI,1)	3.616	-1.845	-0.188	-3.996	3.729	1.368	0.451	0.369	2.612	1.056
AC(TSEI,2)	6.639	-4.905	1.369	-9.099	6.795	2.398	1.238	1.301	6.092	1.821
AC(TSEI,3)	9.697	-5.157	4.883	-12.794	3.851	1.050	2.426	2.543	1.228	0.823
AC(TSEI,4)	10.830	-3.610	1.429	-12.543	-0.764	-3.056	3.998	6.288	5.781	0.594
AC(TSEI,5)	9.159	-0.693	-7.306	-7.903	-0.111	-4.996	6.449	6.387	9.197	1.127
AC(TSEI,6)	5.180	1.432	-13.436	-1.082	4.281	1.476	3.100	2.999	4.145	0.102
AC(TSEI,7)	1.577	1.825	-11.538	1.884	9.397	3.700	-4.484	-2.035	-3.780	-4.233
AC(AtNum,0)	0.488	-2.612	-1.002	2.200	-1.046	-0.546	-0.186	-0.017	-0.970	-0.165
AC(AtNum,1)	0.859	-2.001	-0.778	1.555	0.033	-0.074	-0.079	0.178	-0.409	-1.099
AC(AtNum,2)	2.034	-4.577	-1.482	1.100	1.681	0.809	-0.631	-1.248	1.063	1.134
AC(AtNum,3)	4.009	-6.060	3.753	-0.296	1.530	2.485	-0.481	0.095	-7.396	-4.829
AC(AtNum,4)	6.978	-5.498	6.248	-2.529	-6.777	-1.240	-2.972	2.114	-7.773	-5.436
AC(AtNum,5)	9.002	-0.766	2.599	-1.230	-9.793	-10.132	5.745	3.822	0.283	0.576
AC(AtNum,6)	7.802	3.299	-8.001	3.313	-9.876	-4.091	7.556	5.606	3.281	3.623
AC(AtNum,7)	4.207	4.683	-12.144	5.995	1.939	-1.594	-2.359	2.788	-3.912	-1.689
AC(1,0)	2.397	3.035	0.271	1.319	0.687	0.653	-1.207	-1.226	-1.358	-1.171
AC(1,1)	2.577	2.715	0.011	0.800	1.458	0.873	-1.078	-1.128	-0.404	-0.867
AC(1,2)	5.536	5.111	0.757	1.413	4.246	2.504	-2.378	-2.784	1.373	-1.126
AC(1,3)	9.016	6.792	3.624	1.246	6.661	5.740	-3.988	-7.116	1.259	-0.084
AC(1,4)	15.708	13.399	12.860	1.723	4.626	7.629	-6.732	-11.586	-8.722	2.959
AC(1,5)	22.079	20.760	19.861	3.432	-4.197	3.130	-5.225	-8.415	-14.073	8.882
AC(1,6)	28.558	32.761	21.304	8.657	-15.281	-4.944	-5.640	-4.744	-9.688	6.713
AC(1,7)	25.127	31.898	4.456	14.070	-12.419	-9.842	0.118	7.977	3.752	-9.134
AC(Q_{gast},0)	-0.504	-0.622	-21.169	-29.498	-3.222	-6.625	6.348	13.372	-39.283	-25.535
AC(Q_{gast},1)	0.702	-2.234	20.507	23.205	5.364	8.685	-0.574	16.088	29.609	-10.155
AC(Q_{gast},2)	-0.962	-1.787	-22.065	-24.055	-24.076	-16.071	-22.850	-70.058	4.347	36.912
AC(Q_{gast},3)	-0.251	7.995	18.394	35.925	58.486	6.560	83.095	35.525	-53.729	43.044
AC(Q_{gast},4)	2.310	2.103	-12.205	-28.042	-68.970	65.859	-122.861	80.083	37.697	-51.622
AC(Q_{gast},5)	-0.604	-5.638	26.766	9.550	63.120	-161.692	28.146	-74.981	37.759	-99.397
AC(Q_{gast},6)	2.198	3.021	-32.370	-3.344	-71.112	157.313	110.944	-44.149	-18.173	57.306
AC(Q_{gast},7)	-4.232	-4.481	4.847	3.001	58.304	-66.520	-108.711	89.240	-52.771	145.984
AC(α,0)	0.687	-2.848	-0.902	3.055	-1.239	-0.462	-0.490	-0.663	-0.208	0.966
AC(α,1)	1.175	-1.950	-0.935	1.873	-0.142	0.051	-0.380	-0.599	-0.045	-0.160
AC(α,2)	2.655	-4.834	-1.016	2.641	1.829	1.430	-1.189	-2.136	2.693	2.442
AC(α,3)	5.154	-5.951	4.041	0.829	0.889	3.140	-1.879	-3.083	-5.694	-2.287
AC(α,4)	8.707	-4.157	7.480	-0.192	-6.055	-1.024	-1.735	-1.538	-7.272	0.733
AC(α,5)	11.176	1.259	4.071	1.049	-11.314	-7.459	3.969	2.440	0.343	6.829
AC(α,6)	10.164	5.759	-5.816	5.388	-9.753	-6.652	4.311	5.301	5.133	4.417
AC(α,7)	6.271	7.067	-11.650	7.671	0.333	-1.661	-1.716	1.800	-0.243	-4.466

Table S1. Principal component analysis of the SMU. Variance and loadings associated with each of the first 10 PCs. Loadings are for the 40 *unnormalized* autocorrelation descriptors, where AC(*p*,*d*) is the autocorrelation descriptor corresponding to atomic property *p* with correlation length *d*. The properties used here are topological steric index (TSEI), atomic number (AtNum), unity (1), Gasteiger charge (Q_{gast}), and atomic polarizability (α).

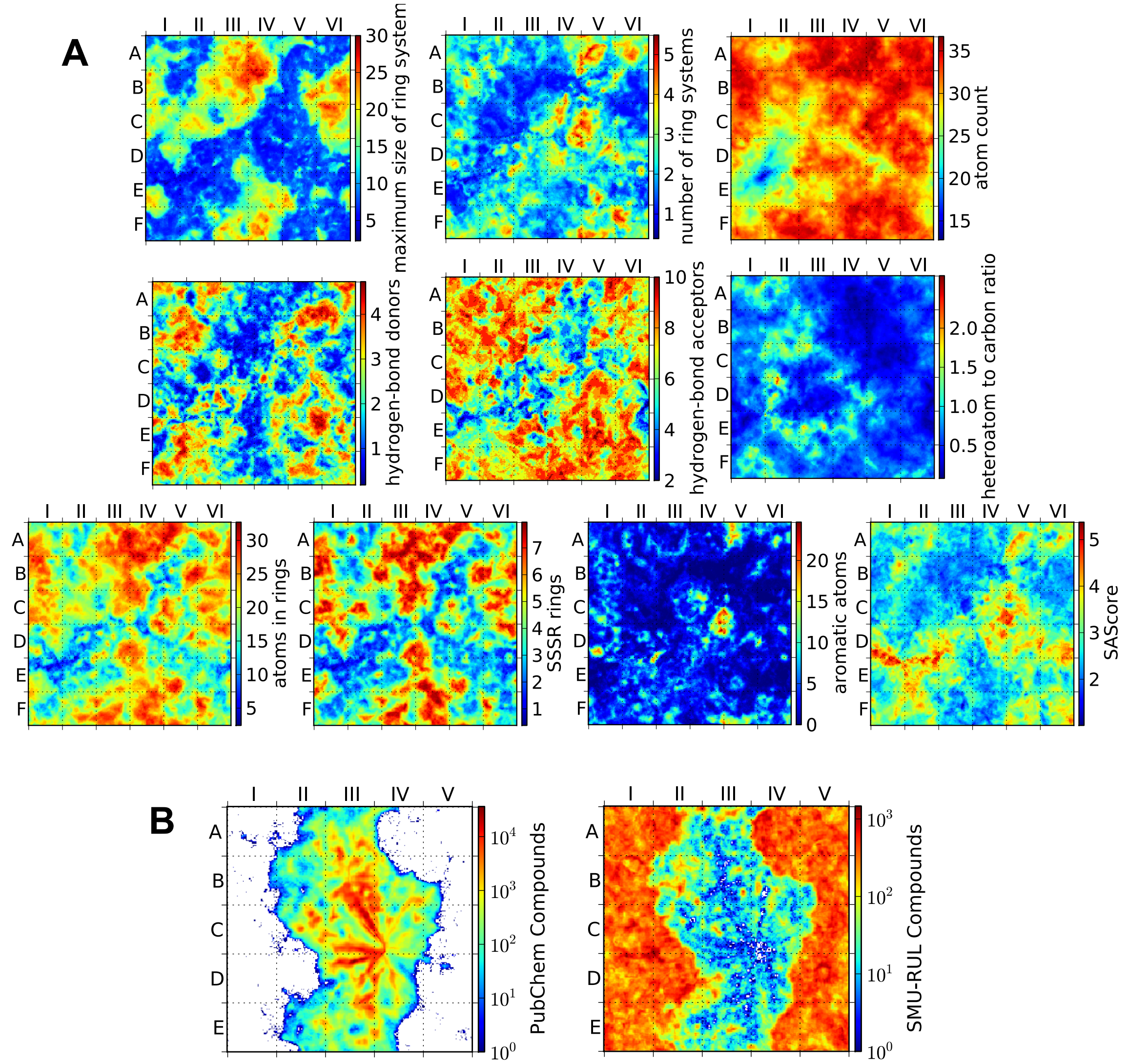


Figure S3. Self-organizing maps. A, average property values for each neuron on SMU-RUL SOM. **B,** number of PubChem (left) and SMU-RUL compounds (right) occupying each neuron on a 200×200 SOM that was trained using maximally diverse subsets of 180,000 compounds from each library.

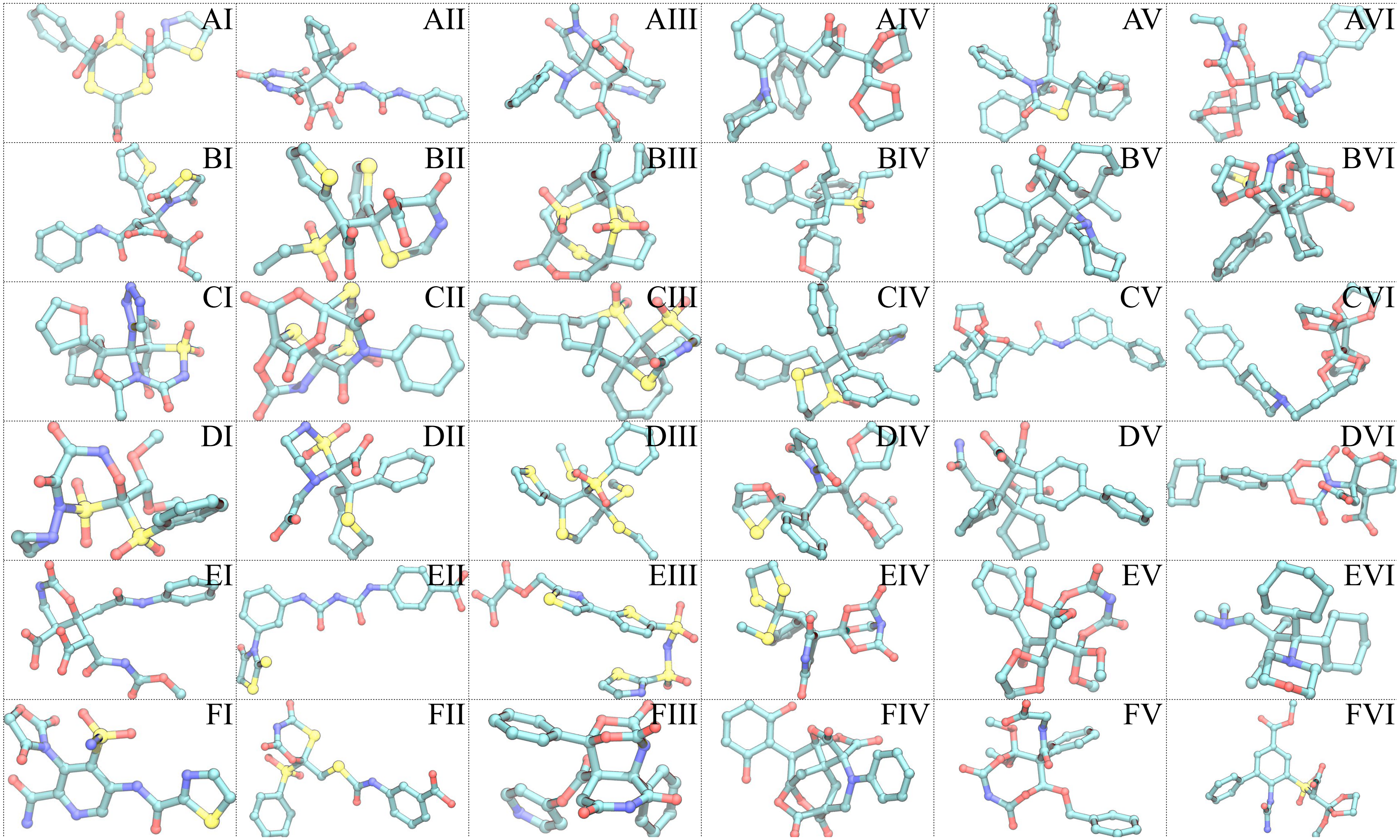


Figure S4. Generated 3D structures of SMU-RUL examples in Chart 1.

S5. Benzodiazepine isomer library

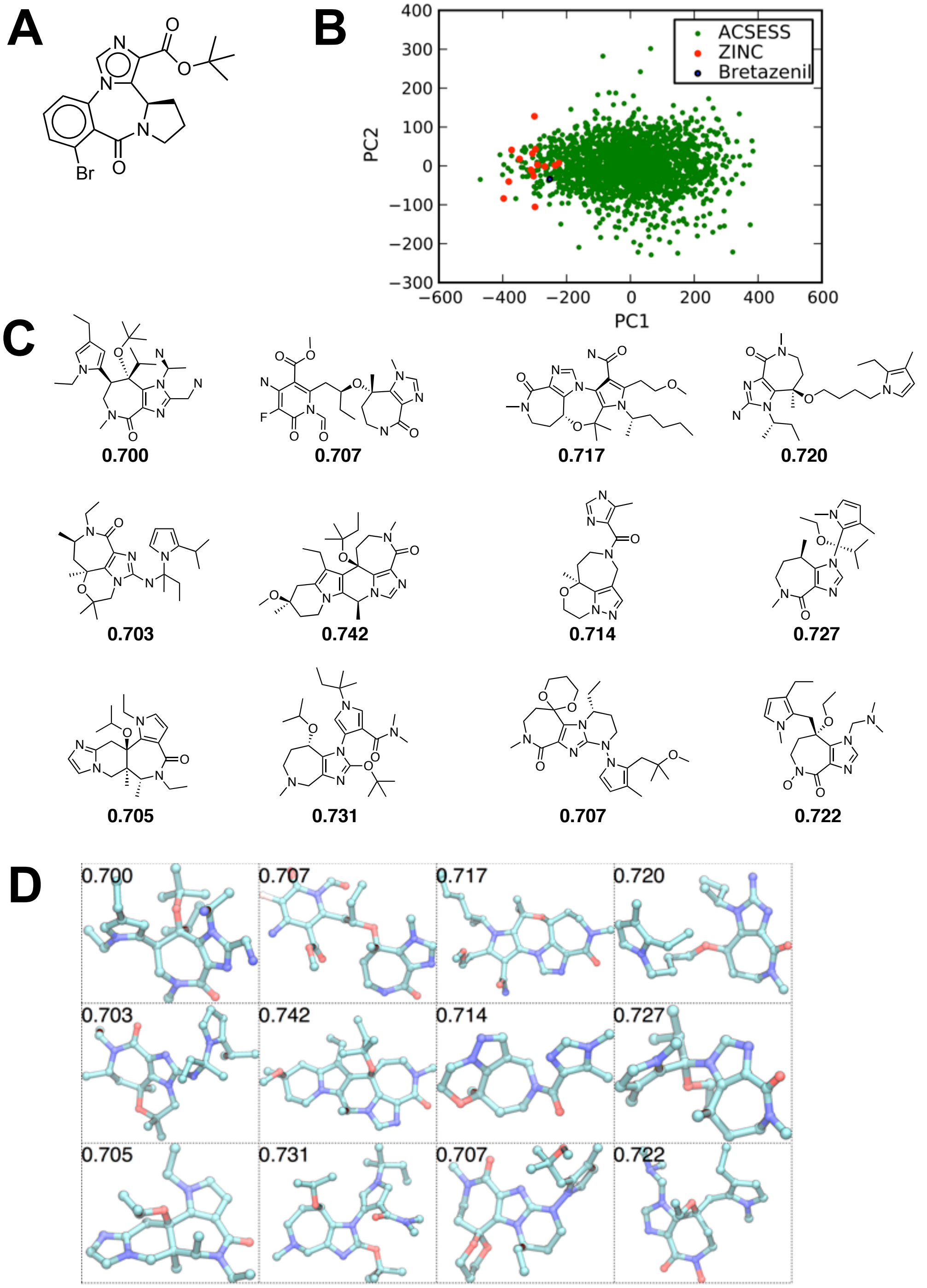


Figure S5. Representative universal library of small molecule bretazenil isomers. **A,** bretazenil. **B**, PCA analysis of the generated RUL and comparison to currently known compounds. **C**, examples of compounds from the generated RUL and their Tanimoto similarities to the target compound. **D,** 3D structures of compounds from **C**.

ACSESS was used to map the space of all SMU structural isomers of the benzodiazepine bretazenil (Figure S5A), a 418 Da broad-spectrum $GABA_A$ agonist. Structural isomers are defined here as compounds with 0.7 or higher Tanimoto similarity to the target compound.

ACSESS was seeded with the SMU-RUL library, and generated the representative universal bretazenil isomer library using a protocol similar to that used in mercaptopurine isomer generation (see above). A maximally diverse library of 2,000 compounds was generated (database S2).

In total, 450,000 compounds were screened, and a total of 12,944 unique isomers of bretazenil were identified (database S3). None of these compounds were found in existing chemical libraries. PCA analysis shows that the generated library covers a much larger chemical space than existing analogs (Figure S5B). Several compounds (and their Tanimoto similarities to the target compound) are shown for reference in Figure S5C-D.

S6. Completeness of genetic operators.

The set of genetic operators employed is “complete,” in the sense that it can be used to mutate any chemical compound into any other compound. The following proof demonstrates the existence of such a mutation path (although in practice, far more efficient mutation paths exist): given any arbitrary molecule, all bonds in rings can be broken with the bond-breaking operator. Terminal atoms in the resulting acyclic structure can be repeatedly deleted with the terminal-atom-deletion operator until only a single atom remains. Thus, any arbitrary molecule may be transformed into a single atom with this set of operators. Because each operator is reversible, the reverse is also true; a single atom may be mutated into any arbitrary molecule. Thus, any chemical structure may be first mutated into a single atom, and then into any other arbitrary structure. Although only seven of these operators are required for completeness (bond creation and annihilation, saturation and unsaturation, terminal atom addition and removal, and element mutation), an extra five genetic operators, as well as “crossover” between compounds, are included to increase the efficiency of chemical space exploration.

S7. External databases

Database S1. SMU-RUL. The SMU-RUL (8.9×10^6 structures) in SMILES format.

Database S2. Bretazenil analog RUL. Generated RUL of bretazenil analogs in SMU chemical space, in SMILES format.

Database S3. All bretazenil analogs. All bretazenil analogs generated during construction of database S2 in SMILES format. This set is a superset of the structures in database S2.

Database S4. SDF-format coordinates for structures shown in Figures S4-S5.

Database S5. ACSESS source code. The ACSESS source may be downloaded as part of the supporting information, and may be used and modified for all non-commercial use. Note that while the source code is made freely available, it requires the commercially available OpenEye chemoinformatic python toolkits. The following packages are required:

1. Python 2.7
2. OpenEye python toolkits (commercially available from http://www.eyesopen.com/; version 20110909 or later required) (4)
3. NumPy (version 1.5.1 or later) and Scipy (0.9 or later)

This computer program is able to generate representative universal libraries for several types of chemical space, including those based on a common scaffold or collection of scaffolds, and/or Tanimoto similarity to a target compound. Detailed documentation and several usage examples may be found in the root directory of the source distribution.

References

1. Deursen Rv & Reymond J-L (2007) Chemical Space Travel. *ChemMedChem* 2:636-640.
2. Farnum MA, DesJarlais RL, & Agrafiotis DK (2003) Molecular Diversity. *Handbook of Chemoinformatics: From Data to Knowledge*, Ed Gasteiger J (Wiley-VCH, Weinheim), Vol 4, pp 1640-1686.
3. Xue L, Stahura FL, & Bajorath J (2004) Cell-Based Partitioning. *Methods in Molecular Biology*, Ed Bajorath J (Humana Press), Vol 275, pp 279-289.
4. OEChem 1.7.7, OpenEye Scientific Software, Inc. http://www.eyesopen.com/ (2012).
5. Blum LC & Reymond JL (2009) 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J Am Chem Soc* 131(25):8732-8733..
6. Fink T & Reymond J-L (2007) 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. *J Chem Inf Model* 47:342-353.
7. Boyer JM & Myrvold WJ (2004) On the Cutting Edge: Simplified O(n) Planarity by Edge Addition. *J Graph Algor App* 8(3):241-273.
8. Downs GM, Gillet VJ, Holliday JD, & Lynch MF (1989) Theoretical Aspects of Ring Perception and Development of the Extended Set of Smallest Rings Concept. *J Chem Inf Comp Sci* 29(3):187-206.