Increased Diversity of *Libraries from Libraries*: Chemoinformatic Analysis of Bis-Diazacyclic Libraries

Fabian López-Vallejo,[1] Adel Nefzi,[1] Andreas Bender,[2] John R. Owen,[3] Ian T. Nabney,[3] Richard A. Houghten,[1] and Jose L. Medina-Franco[1,*]

Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, FL 34987, USA Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom, Non-linearity and Complexity Research Group (NCRG), Aston University, Aston Triangle, Birmingham B4 7ET, United Kingdom

[*] Corresponding author. Tel.: +1 772-345-4685; fax: +1 772-345-3649. E-mail: jmedina@tpims.org

[1] *Torrey Pines Institute for Molecular Studies*

[2] *University of Cambridge*

[3] *NCRG, Aston University*

Keywords: chemical space; combinatorial chemistry; drugs; diversity-oriented synthesis; molecular diversity; visualization

Running title: Molecular Diversity of Libraries from Libraries

Abbreviations: CDF, cumulative distribution function; DOS, diversity-oriented synthesis; GpiDAPH3, graph-based three point pharmacophores; HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; LoL, library from library; MLSMR, Molecular Libraries Small Molecule Repository; MOE, Molecular Operating Environment; MW, molecular weight; LTM, Latent Trait Model; PCA, principal component analysis; PC, principal components; RB, rotatable bonds; TGD, typed graph distances; TPSA, topological polar surface area; DVMS, Data Visualization and Modeling System

ABSTRACT

Combinatorial libraries continue to play a key role in drug discovery. To increase structural diversity, several experimental methods have been developed. However, limited efforts have been performed so far to quantify the diversity of the broadly used diversity-oriented synthetic (DOS) libraries. Herein we report a comprehensive characterization of 15 bis-diazacyclic combinatorial libraries obtained through *libraries from libraries*, which is a DOS approach. Using MACCS keys, radial and different pharmacophoric fingerprints as well as six molecular properties, it was demonstrated the increased structural and property diversity of the libraries from libraries over the individual libraries. Comparison of the libraries to existing drugs, NCI Diversity and the Molecular Libraries Small Molecule Repository revealed the structural uniqueness of the combinatorial libraries (mean similarity < 0.5 for any fingerprint representation). In particular, bis-cyclic thiourea libraries were the most structurally dissimilar to drugs retaining drug-like character in property space. This study represents the first comprehensive quantification of the diversity of *libraries from libraries* providing a solid quantitative approach to compare and contrast the diversity of DOS libraries with existing drugs or any other compound collection.

INTRODUCTION

Synthetic combinatorial methods have advanced the ability to synthesize and screen large numbers of compounds because of improvements made in technology, instrumentation, and library design strategies (1). Combinatorial chemistry combined with high throughput and other screening methodologies continues to play a key role in drug discovery (1-3). A very successful synthetic method is the '*Libraries from Libraries'* (LoL) approach (4). This concept is based on the use of well-established solid-phase synthesis methods for the generation of combinatorial libraries combined with the chemical transformation of such libraries. The chemical libraries that are generated by this process have very different physical, chemical, and biological properties compared to the libraries from which they were

derived (4). As such, LoLs can be regarded as a diversity-oriented synthetic (DOS) approach (5, 6), where multiple scaffolds are generated from the same starting material. Increasing skeletal diversity is known to be a very efficient way to increase structural diversity (7). As opposed to high-throughput screening, where often a large number of compounds with different scaffolds are screened, LoL explores the bioactivity space around each scaffold of interest in much more detail, by using a large number of diversity appendages on every scaffold.

A number of small molecule libraries have been prepared in our group using the LoL approach. These libraries have been used successfully to identify novel compounds across a wide range of therapeutic applications (4, 8). Figure 1 shows the LoLs considered in this study. Each LoL includes five individual combinatorial libraries containing the same number of diversity positions, identical side chain functionalities at each diversity position, and the same number of compounds. Libraries within each LoL differ only in the chemical nature of the central scaffold.

It is well accepted that the structural diversity of LoLs improves upon the diversity of other combinatorial libraries, where the 'multidimensional diversity' regarding both scaffold and appendages is often one of the key contributing factors. However, characterizing the diversity is not an easy task, and efforts have been pursued in this regard. One example is in the work of Spandl et al.(7) in which the importance of skeletal diversity was stressed. Previously, circular fingerprints have been used to assess diversity of compound collections (9). While in this work overall good discrimination between DOS and target-oriented synthesis (TOS) libraries could be observed, the question of how to normalize for library size could not be answered completely – smaller libraries often assessed were more diverse, since larger libraries nearly necessarily contain repetitive chemical motives. Also, Rolfe et al. recently addressed the structural diversity of a number of 17 compounds obtained via a "click, click, cyclize" DOS strategy using principal component analysis (PCA) based on BCUT descriptors and principal moment of inertia. The 17 compounds covered different regions of chemical space (10).

The goal of this study was to characterize the structural diversity of 15 combinatorial libraries organized into three LoLs (Figure 1A-C). The analysis is based on two criteria: structural fingerprints and

molecular properties. Thus, the LoL concept reported previously and perceived by chemists as giving rise to diverse compounds, is now assessed quantitatively for the first time. It is demonstrated that LoL generates molecules truly diverse in both structural and molecular property space.

This work is organized into four sections. The first section describes the fingerprint-based diversity of each combinatorial library and LoL. The second section shows the cross-comparison of the three LoLs and the 15 libraries. The third section describes the structural comparison of the 15 libraries with external compound collections. Section four compares the compound datasets in terms of molecular properties. The approaches presented here are general and can be used to characterize the diversity of other LoLs or DOS libraries.

Figure 1 here

METHODS

Data sets. Each LoL contains five individual combinatorial libraries reported previously (Figure 1):

LoLA = **A1** U **A2** U **A3** U **A4** U **A5**

LoLB = **B1** U **B2** U **B3** U **B4** U **B5**

LoLC = **C1** U **C2** U **C3** U **C4** U **C5**

where LoLA-C are the three libraries from libraries considered in this study. Figure 1A shows the bis-heterocyclic LoLA that includes bis-cyclic diketopiperazines **A1**, bis-cyclic piperazines **A2**, bis-cyclic guanidines **A3**, bis-cyclic ureas **A4** and bis-cyclic thioureas **A5** (8, 11, 12). LoLB (Figure 1B) is composed of bis-cyclic diketopiperazines **B1**, bis-cyclic piperazines **B2**, bis-cyclic guanidines **B3**, bis-cyclic ureas **B4** and bis-cyclic thioureas **B5** (8, 13). LoLC (Figure 1C) includes different pentaamines and pyrrolidine bis-heterocyclic libraries, such as pyrrolidine bis-diketopiperazine **C1**, pyrrolidine bis-piperazine **C2**, pyrrolidine bis-cyclic guanidines **C3**, pentaamine **C4** and pyrrolidine bis-cyclic thiourea **C5** (14). To note, libraries with the core scaffolds of LoLC and LoLB were screened in a μ-opioid receptor binding assay (1). The most active libraries had the scaffolds of **B3** and **C3** (1) and several

compounds of **C3** had a K_i lower than 100 nM (1). Combinatorial libraries with the core scaffold of LoLC were recently screened for antitubercular activity leading to compounds with 90–100% inhibition against *M. tuberculosis* at concentrations less than 6.25μg/mL (14).

Libraries **A1**-**A5** and **B1**-**B5** have three diversity positions, and **C1**-**C5** have four diversity positions. To enumerate libraries **A1**-**A5** and **B1**-**B5**, we used ten amino acids or carboxylic acids as building blocks for each diversity position. Thus, the size of each individual combinatorial library was 10 x 10 x 10 = 1,000 compounds; hence LoLA and LoLB contained 5,000 structures each. In order to measure the effect of each core template in the diversity, the same set of ten amino acids or carboxylic acids was considered for each library. To enumerate **C1**-**C5**, we selected six amino acids and five carboxylic acids from the pool of the ten building blocks used in the libraries above. Thus, the size of each library **C1**-**C5** was 6 x 6 x 6 x 5 = 1,080 compounds, and LoLC contained 5,400 structures. A complete list of the building blocks used to enumerate the libraries is in Table S1 of the Supporting information. In order to compare the diversity across different libraries, we considered approximately the same library size; i.e., 1,000 – 1,080 compounds. In addition, it has been reported that data sets of 1,000 molecules are representative samples to study the structural diversity of combinatorial and other libraries (15, 16). The combinatorial libraries were enumerated using the QuaSAR-CombiDesign module of the Molecular Operating Environment (MOE) program, version 2009.10 (17). The collection of drugs (1,490 compounds) was obtained from DrugBank (18) as collected in the ZINC database (download July 2008) (19). The NCI diversity set (1,832 compounds with unique SMILES as computed with MOE) was obtained from ZINC (download March 2010). The MLSMR collection was obtained from PubChem (20) (347,480 compounds downloaded in May 2010) and processed with MOE by disconnecting group I metals in simple salts and keeping the largest fragment.

Comparison metrics. Compound collections were analyzed based on structural fingerprints and molecular properties. Similarity values were computed using the 2D fingerprints MACCS keys (21) (166 bits), graph-based three point pharmacophores (GpiDAPH3), typed graph distances (TGD) implemented in MOE, and radial fingerprints (equivalent to ECFP4) implemented in Canvas (22). In addition, we used

the 3D fingerprint spatial three-point pharmacophore (piDAPH3) from MOE, calculated from the structures geometrically optimized using the MMFF94x force field implemented in MOE. The Tanimoto coefficient (23, 24) was used as the similarity measure for all fingerprints.

Intra-library similarity. Pairwise similarities were computed for each combinatorial library. The distribution of similarities was analyzed by means of cumulative distribution function (CDF) curves.

Inter-library similarity. LoLs were compared to each other computing the pairwise structural similarity. To this end, we employed random samples of 300 compounds per library so that each LoL contained 1,500 members (25). The pairwise similarities were analyzed using CDF curves and multi-fusion similarity (MFS) maps. The MFS map is a method developed recently for the visual characterization and comparison of compound databases and is based on data-fusion similarity measures. The fusion data are plotted in two dimensions, where the ordinate represents the maximum-fusion values and the abscissa the mean-fusion values. Each point in the map is associated with a specific molecule in the test set, and its position is determined by the corresponding fusion values computed with respect to the molecules in the reference set (26). The MFS maps can be characterized quantitatively by the corresponding distributions of the max- and mean-fusion values (27). This approach has been employed to explore structure-activity relationships of compounds data sets (28) and to compare combinatorial libraries (16, 27, 29).

Comparison with external compound collections. The 15 libraries, with 1,000 (**A1**-**A5**, **B1**-**B5**) and 1,080 (**C1**-**C5**) molecules each, were compared with a collection of drugs, the NCI Diversity set and the Molecular Libraries Small Molecule Repository (MLSMR) using MACCS keys, GpiDAPH3, TGD and piDAPH3 fingerprints. We employed MFS maps (setting the external compound collections as the reference sets) and Latent Trait Mapping (LTM) plots. LTM is a dimensionality reduction technique that is specially designed to visualize discrete data (30). It defines a function (or mapping) from the original data space to a lower-dimension (usually 2D) visualization space. Because the mapping is non-linear, it often provides a more informative visualization than either plotting pairs of the original variables (such as

MFS maps) or linear maps (such as PCA). The price to pay for the greater insight is that there is no interpretation of the axes on the visualization plot.

Molecular properties and property space. The following properties were computed with MOE molecular weight (MW), number of rotatable bonds (RB), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), topological polar surface area (TPSA), and the octanol/water partition coefficient (SlogP). To obtain a *visual* representation of the property space (27), PCA was carried out in Spotfire 9.1.2 (31) considering the six molecular properties and plotting the first two principal components.

RESULTS AND DISCUSSION

Intra-library diversity

Figure 2 summarizes the distribution of similarities of LoLA, LoLB and LoLC and the corresponding individual libraries using MACCS keys. Figure 2A shows a comparison of the diversity of LoLA and **A1**-**A5**. The CDFs and the corresponding statistics indicate that LoLA is more diverse than each individual library. For example, compare the median and mean of the similarity distribution of LoLA (0.627 and 0.641, respectively) with the corresponding values for the individual libraries ($\geq$ 0.786 and $\geq$ 0.789, respectively). Also, compare the standard deviation of LoLA (0.110) with the standard deviation of **A1**-**A5** ($\leq$ 0.08). The CDFs also indicated that **A3** and **A5** showed slightly higher diversity than **A1**, **A2** and **A4**. Library **A1** showed the lowest diversity.

Figure 2 here

Figure 2B and 2C summarize the distribution of similarities of LoLB and LoLC, respectively, along with their corresponding individual libraries. LoLB and LoLC have higher diversity (lower similarity) than their corresponding libraries. Bis-cyclic diketopiperazine libraries (**B1** and **C1**), had lower diversity as compared to other corresponding individual libraries within the same LoL. To note, a relatively small pre-defined MACCS keys (166 bits) was sufficient to differentiate between the libraries. These results suggest that MACCS keys could focus only on the discriminant features but neglect other relevant

chemistry. However, we obtained similar conclusions using other structural representations including TGD, GpiDAPH, radial and piDAPH3 fingerprints (see below).

Inter-library diversity

Figure 3 shows the heat maps of similarity matrices of the 15 combinatorial libraries calculated with MACCS, TGD, GpiDAPH, radial and piDAPH3 fingerprints. Each map visualizes 1,500 x 1,500 = 2,250,000 pairwise comparisons and is color-coded by similarity value using a continuous scale from green (low similarity value) to red (high similarity value). The name of the individual libraries and LoLs are indicated in the figure. Each map can be divided into 15 x 15 = 225 "minor" regions or squares that correspond to the pairwise comparison of all 15 libraries. The maps can also be divided into 3 x 3 = 9 "major" regions that are associated with the cross-comparisons of LoLA, LoLB and LoLC. The minor or major regions along the main diagonals starting from the top-left correspond to the self-library comparisons. The maps help to visually inspect the similarity between individual libraries that belong to the same LoL as well as to different LoLs. In general, in this study most of the similarity values computed with radial fingerprints were close to zero. TGD, piDAPH3 and GpiDAPH3 showed comparable similarity relationships among databases although with different scales; similarity values calculated with TGD were higher than the similarities calculated with piDAPH3 and GpiDAPH3. Noteworthy, the pharmacophoric fingerprints, TGD, piDAPH3 and GpiDAPH3, were unable to distinguish the pair of libraries **A4-A5** or **B4-B5**. This is because the difference in the central scaffold of these libraries is an oxygen (**A4-B4**) or sulphur atom (**A5-B5**) that are treated as equal by TGD, piDAPH3 and GpiDAPH3. MACCS keys provided very insightful results and were able to distinguish all libraries; therefore, we will mainly focus on MACCS keys to discuss the inter-library similarity.

Figure 3 here

Along the main diagonal on the heat map based on MACCS keys (Figure 3), there are 15 black-to-red squares indicating a high similarity for the self-individual library comparisons; e.g., comparison of **A1-A1**, **A2-A2**, **A3-A3**, **A4-A4**, **B1-B1**. In contrast, the three major regions along the main diagonal (self-

LoL comparisons, LoLA-A, LoLB-B, LoLC-C), contain a number of green and black-to-green squares; in particular, LoLA-A and LoLB-B. This observation suggests a low similarity between the five individual libraries that belong to the same LoL. The visual analysis of the similarity matrix of random samples is in agreement with the increased diversity of each LoL as compared to the individual libraries (see above). A quantitative analysis of each major quadrant is discussed later in this section.

A diagonal of black-to-red squares can be identified in the major region associated with the cross-comparisons of LoLA-B. This diagonal indicates the high similarity between individual libraries that belong to different LoLs; for example, **A1**-**B1**, **A2**-**B2**, **A3**-**B3**, **A4**-**B4** and **A5**-**B5**. This is because all these pairs of libraries have a bis-cyclic heterocyclic ring in common, namely, bis-cyclic diketopiperazines (**A1**, **B1**), bis-cyclic piperazines (**A2**, **B2**), bis-cyclic guanidines (**A3**, **B3**), bis-cyclic ureas (**A4**, **B4**) and bis-cyclic thioureas (**A5**, **B5**). Similar conclusions are obtained from the cross-comparisons LoLA-C and LoLB-C. Note, however, that the **A4**-**C4** and **B4**-**C4** comparisons are green-to-black indicating lower similarity. This observation is associated with the different scaffold of **A4** and **B4** (bis-cyclic urea), and **C4** (pentaamine). To note, the black-to-red square comparing libraries **B3**-**C3** indicates high similarity between these collections. To note, both libraries have a bis-cyclic guanidine moiety in the central scaffold, and libraries with these scaffolds were the most active in a μ-opioid receptor binding assay (1). The heat map calculated using MACCS keys also provides a quick inspection of the more diverse libraries; for example, the green squares corresponding to the pair of libraries **A1**-**B3** and **A5**-**B1** indicate low similarity.

The heat map of similarity matrices is a simple and powerful tool to explore structural relationships between LoLs and the individual libraries. The visual analysis was expanded with a quantitative study of the similarities between LoLs. Figure 4 shows the CDFs and corresponding statistics of the distribution of the MACCS pairwise similarities comparing the LoLs. The CDFs for the self-LoL comparisons with 300 compounds per library are similar to the CDFs for the self-LoL comparisons with 1,000 (LoLA, LoLB) and 1,080 compounds (LoLC) per library. This result suggests that the random samples with 300 compounds per library are representative. The CDF curves indicate that each of the LoL is structurally

diverse. Figure 4 shows that LoLA is the most diverse of the three LoLs, whereas LoLC is the least diverse. This result agrees with the conclusions obtained from the heat map in Figure 3 and CDFs in Figure 2A. The CDF for the cross-LoL comparisons (Figure 4) indicate the low similarity between any pair of LoLs. The pair LoLA-B has the highest diversity, whereas the pair LoLB-C has the lowest diversity. Similar conclusions were derived from the heat map (Figure 3).

Figure 4 here

Figure 5 depicts the MFS maps comparing the LoLs with themselves (self-reference) and with other LoL (cross-comparisons) using MACCS keys. The reference sets are designated along the top and the test sets along the left-hand side of the figure. The three maps along the main diagonal are self-referential. In all maps the mean similarity ranges within similar values (~0.5 < mean similarity < ~0.75) indicating comparable diversity of the reference sets. This result is in agreement with the similarity distributions of the three LoLs (Figure 2 and 4). Concerning the self-referential MFS maps, the maximum values are high (> 0.9 and several > 0.95) indicating that each molecule in the corresponding LoL has a close nearest-neighbor (expected to belong to the same individual library). However, the corresponding mean values are lower than 0.75 indicating the higher diversity of the corresponding LoL.

Figure 5 here

The MFS maps along the left-hand side of Figure 5 show the relationship of LoLB and LoLC to LoLA (reference). LoLB has larger maximum values than LoLC indicating that LoLB has closer neighbors in LoLA. This result is difficult to deduce from the heat maps in Figure 3. The MFS maps along the center of Figure 5 show the relationship of LoLA and LoLC to LoLB (reference) suggesting a similar relationship between the two LoLs and LoLB. The MFS maps along the right-hand side of Figure 5 depict the relationship of LoLA and LoLB with LoLC indicating an overall lower similarity of LoLA (32).

Structural comparison with external data sets

Figure 6 shows the MFS maps and the corresponding CDFs of the maximum- and mean-fusion values comparing the relationship between LoLA-C (test) and drugs (reference) using MACCS keys. The CDF of the maximum-fusion values are reminiscent of the nearest-neighbors curves (33, 34). These maps suggest that there are no identical compounds between any of the combinatorial libraries and drugs; moreover, all the compounds in the combinatorial libraries have a maximum MACCS keys similarity lower than 0.9 to any of the drugs and most of the compounds have maximum similarity lower than 0.85. Note also the larger distribution of the molecules in each of the three MFS maps considering the entire LoLs as compared to the distribution of the molecules of each combinatorial library. This result further supports the increased diversity of LoLs over the individual libraries. Similar conclusions were obtained with other fingerprint representations (see below).

Figure 6 here

Figure 6A shows the MFS maps and CDF for libraries **A1**-**A5**. **A1** is towards the top right part of the MFS map suggesting a relative increased structural similarity to drugs as compared to **A2**-**A5**. This observation was further confirmed by the CDFs of the maximum- and mean-fusion values. **A5** is located towards the bottom left part of the MFS map suggesting that this library is the least similar to drugs. A similar conclusion can be derived from the CDFs. Note, however, that the CDF of the mean-fusion value cannot distinguish **A3** and **A5** since these two libraries have the same mean similarity relationship to drugs. **A2** and **A4** are, in general, the second most similar libraries to drugs after **A1** as suggested by the MFS map. To note, the corresponding CDFs of the maximum- and mean-fusion values indicate an opposite order of similarity for **A2** and **A4** with respect to drugs; the CDF of the maximum-fusion values for library **A2** is shifted towards higher values than the corresponding curve for **A4** indicating that the nearest neighbors of **A2** in DrugBank are closer that the nearest neighbors for **A4**. However, the CDF of the mean-fusion values for library **A4** is shifted towards higher values indicating that, on average, **A4** is more similar to drugs than **A2**. These observations highlight the importance of considering more than one metric for a complete assessment of the relationship between compound collections.

Figure 6B shows the MFS maps and CDFs for libraries **B1**-**B5**. According to the MFS map **B1**, a bis-cyclic diketopiperazine library (related to **A1**), is structurally more similar to drugs as compared to other libraries within LoLB. Library **B5**, a bis-cyclic thiourea (related to **A5**), is the less similar to drugs. Similar conclusions can be derived from the CDFs. The CDFs of the maximum- and mean-fusion values for **B2** and **B4**, respectively, indicate that while **B2** has closer nearest neighbors in the collections of drugs, **B4** is on average more similar to drugs. Figure 6C shows the MFS maps and CDFs for **C1**-**C5** indicating that the pyrrolidine bis-cyclic thioureas **C5** are the less structurally similar to drugs. According to the MFS map and CDF of the mean-fusion values, pyrrolidine bis-cyclic diketopiperazines **C1** are more similar to drugs as compared to other libraries within LoLC. To note, the CDF for the maximum-fusion values indicates that **C1** and **C2** have a similar nearest-neighbor relationships to drugs; however, **C1** is on average structurally more similar to drugs than **C2**.

Figure 7 depicts a *visualization* of the chemical space comparing LoLs and drugs using the LTM algorithm with MACCS keys (166 bits) as the molecular descriptors (35). The binary (0-1) data is well suited to this algorithm. Figure 7A shows LoLA, LoLB, LoLC and drugs in the same space. For clarity, Figure 7B-D show a comparison of the chemical space of drugs with each LoL, respectively, within the same coordinates as Figure 7A. The LTM plots on the left-hand side of Figure 7B-D shows the LoL as one color, whereas the chemical space on the right-hand side shows each individual combinatorial library with different colors (color scheme as in Figure 1). Visualization of the chemical space in Figure 7B-D shows that LoLA-C cover a larger area of the chemical space than the space covered by each combinatorial library. This result supports the increased structural diversity of the LoLs over the individual libraries.

Figure 7 here

Visualization of the chemical space with the LTM plots shows different relationships of **A1**-**A5**, **B1**-**B5** and **C1**-**C5** with drugs. As such, libraries with a bis-cyclic guanidine moiety in the core scaffold **A3**, **B3**, **C3** (in black), and with a bis-cyclic urea moiety **A5**, **B5** and **C5** (in yellow) are the most dissimilar to drugs. In contrast, libraries containing a diketopiperazine moiety **A1**, **B1** and **C1** (in blue) are the most

structurally similar to drugs. Measuring the mean distance of drugs to each library in the LTM space further confirmed these conclusions (Table S3 of the Supporting information). Note, however, that there are no overlaps in the structural MACCS keys space between the combinatorial libraries and drugs. These results are in agreement with the MFS maps and CDF curves discussed above (Figure 6). In contrast to MFS maps, LTM plots enable the visualization of the reference set (i.e., drugs) in chemical space. It is also noteworthy that the LTM plots show the separation between different classes much better than the MFS maps (and the PCA visualizations discussed below and shown in Figure 8). In addition, the relationships between different libraries (as measured by distance in the LTM space) are better defined, and there is more information about the groupings of compounds within each library. Further interactive exploration of these visualizations can be carried out using the Data Visualization and Modeling System (DVMS) tool (35, 36).

We also analyzed the structural relationship between LoLs and NCI diversity using the MFS maps and CDFs using MACCS keys (plots not shown). Similar to the comparison with drugs, the combinatorial libraries showed a different structural relationship to NCI diversity. It was also concluded that there are no identical molecules between NCI Diversity and any combinatorial library. Moreover, most of the compounds in any combinatorial library have maximum MACCS keys similarity of 0.80. Lower similarity values were computed with other fingerprint representations (see below).

Since the chemical space depends on the molecular representation (37), we investigated the structural relationship of the LoLs to drugs, NCI diversity and MLSMR using TGD, GpiDAPH3 and piDAPH3 fingerprints. For MLSMR, a subset of 3,000 compounds was selected at random. The maximum- and mean-fusion values (Table S4 of the Supporting information), confirmed that the combinatorial libraries are structurally different from drugs, NCI diversity and MLSMR regardless of the structural representation.

Property Diversity

Table 1 summarizes the median, mean and standard deviation of the distribution of the six molecular properties described in Methods for the 15 libraries, three LoLs and external data sets. Additional statistics of the distributions (e.g., maximum, minimum, first and third quartile and U95 and L95 values) are presented in Table S2 of the Supporting information. The three important molecular properties of size, flexibility and molecular polarity are described by MW; RB; and SlogP, TPSA, HBA and HBD, respectively. The six descriptors used here have been used to compare the property space covered by a virtual collection and reference databases (38) and other combinatorial libraries (16). According to Table 1 and Table S2 of the Supporting information, each LoL has a wider distribution of molecular properties than their corresponding individual libraries as reflected by the larger standard deviation and range for most of the properties. Since all of the combinatorial libraries within a LoL contain the same number of diversity positions, identical side chain functionalities at each diversity position, and the same number of compounds with the library (see above), the variation in the properties within a LoL is due to the central scaffold. These results further demonstrate the observation that increasing skeletal diversity is a very efficient way to increase not only the structural diversity (7), but also property diversity.

Table 1 here

According to Table 1 and S2, LoLA and the corresponding individual libraries have, in general, more HBA, HBD, and larger TPSA values than drugs. The number of rotatable bonds in any of the libraries in LoLA is quite similar to the number of RB in drugs. Also, LoLA has a distribution of SlogP values comparable to drugs; in particular **A4** and **A5**. In general, LoLB and the corresponding individual libraries have larger SlogP, MW and RB values than drugs. **B4**, **B3** and **B5** have a distribution of HBA and HBD comparable to drugs (B3 also has a similar distribution of TPSA values). LoLC has, in general, larger values of HBA, RB and MW than drugs. **C1**, **C2** and **C5** have a distribution of HBD comparable to drugs (**C1** also has similar distribution of SlogP values and **C4** a similar distribution of TPSA values).

In order to generate a *visual* representation of the property space, the six molecular properties were subjected to PCA after Z-scaling. Figure 8 depicts an *approximation* of the property space as defined by these properties. The first two principal components (PC) with eigenvalues 2.172 and 2.071, respectively,

account for 70.71% of the variance. (Components with eigenvalues less than 1.0 were not considered). Figure 8A shows LoLA, LoLB and LoLC, drugs and NCI Diversity sets in the same space. From this figure and the property distributions (Tables 1 and S2), it can be concluded that NCI diversity covers a similar region of the property space occupied by drugs. For the sake of clarity, Figures 8B-D show a comparison of the property space of drugs with each LoL, respectively, within the same coordinates as Figure 8A. The property space on the left-hand side of Figures 8B-D show the LoL is one color, whereas the property space on the right-hand side shows each individual combinatorial library with different colors (color scheme as in Figure 1). Table 2 summarizes the corresponding loadings and eigenvalues for the six PCs. For the first PC, the larger loadings correspond to MW followed by RB. For the second PC, the largest loading corresponds to HBD followed by TPSA, whereas for the third PC, the largest loading corresponds to TPSA followed by HBA. Figures 8B-D reveal that LoLA, LoLB and LoLC cover a larger area of the property space than the space covered by each individual library. This observation further demonstrates the increased property diversity of the LoLs over the individual combinatorial libraries.

Figure 8 here

Table 2 here

Figures 8B-D also show a different degree of overlap between LoLs and drugs. LoLA has a significant overlap with drugs (Figure 8B). Noteworthy, LoLA is *structurally dissimilar* to drugs (see above). In other words, the chemical structures of LoLA are different from the structures of drugs; however, the molecular properties of LoLA are similar to the properties of drugs. This observation illustrates the dependence of chemical space with the structural representation (37) and further emphasizes the importance of considering multiple criteria when comparing compound data sets (39, 40).

LoLB and LoLC cover regions of the property space sparsely populated by drugs (Figures 8C and 8D, respectively). A major difference in the distribution of LoLB, LoLC and drugs occurs along the coordinates of the first PC that is mainly associated with MW and RB. These results are in agreement with the conclusions derived from the property distributions in Tables 1 and S2. The areas in the property space with few drugs represent areas that are biologically relevant as revealed by the presence of some

drugs but may not have been sufficiently explored (16). In addition, LoLB and LoLC sample unexplored regions of the drug space. Coverage of regions unexplored by drugs has been reported for other DOS and combinatorial libraries (16, 41). Molecules in these areas, while potentially unlikely to make drugs by themselves, are valuable as chemical probes in order to better understand the structure-activity relationships associated with unknown targets (9, 16). Although the LTM (Figure 7) and PCA plots (Figure 8) represent *visual approximations* of the property space, they provide a useful idea of the molecules' distribution in the space. The non-linear nature of the LTM means that it provides greater discrimination between libraries and also within each library.

We also compared the property space of the LoLs with MLSMR using the same random subset employed in the structural study (42). A visualization of the property space is depicted in Figure S2. (The corresponding loadings and eigenvalues are summarized in Table S5 of the Supporting information). We concluded that LoLA, LoLB and LoLC not only occupy part of the property space of MLSMR but also sparse regions.

The major focus of this study was to demonstrate quantitatively the increased diversity of the combinatorial libraries generated with the LoL approach. To this end, we considered a standard set of building blocks for each LoL enumerating ~1000 compounds per library. Note, however, that the actual size of the libraries can be much larger. For example, libraries with the pyrrolidine bis-cyclic scaffolds have been synthesized and screened with 738,192 compounds per library. Similarly, libraries with the bis-cyclic scaffolds have been synthesized and screened with 45,864 compounds per library (1).

CONCLUSIONS AND PERSPECTIVES

Herein we report a comprehensive characterization of the diversity of 15 combinatorial libraries obtained with the libraries from libraries (LoLs) approach. The 15 libraries are grouped in three bis-diazacyclic LoLs. Previous studies showed that several of the central scaffolds considered in this study are biologically relevant. Structural fingerprints, including MACCS keys, GpiDAPH3, TGD, radial and piDAPH3 fingerprints, as well as molecular properties, were considered to assess the diversity. Analysis

of the structural diversity of the libraries showed that the LoLs are more diverse than the corresponding individual libraries. Distributions of pairwise structural similarities and heat maps of similarity matrices revealed that libraries within the same LoL are structurally diverse, mainly combinatorial libraries in LoLA followed by libraries in LoLB. All these results provide quantitative measures of the different chemical and physical properties of the libraries generated by the LoL approach. Fingerprints-based comparisons indicated that the 15 libraries are structurally different from drugs, NCI Diversity and MLSMR. MFS maps and LTM plots revealed that libraries containing a bis-cyclic thiourea moiety in the central scaffold (**A5**, **B5** and **C5**) are the most structurally dissimilar to drugs. In contrast, libraries with a bis-cyclic diketopiperazine moiety in the central scaffold (**A1**, **B1** and **C1**) are more structurally similar to drugs. Comparison of the libraries with drugs using molecular properties showed that combinatorial libraries have a different degree of overlap with the property space of drugs. In particular, despite the fact that LoLA is structurally dissimilar to drugs, their properties are similar.

Focused and targeted combinatorial libraries have gained an increased interest in recent years (43-45). The core scaffolds presented here represent potential starting points of focused or targeted combinatorial libraries. The most suitable core scaffold(s) for a particular target family can be explored using computer-aided target fishing approaches (46, 47) and in-silico "scaffold ranking" (1) to identify the most promising core scaffolds for a particular target or target family before synthesis.

ACKNOWLEDGEMENT

We are very grateful to Karen Gottwald for proofreading the manuscript. We also acknowledge Kyle Kryak for his assistance. This work was supported by the State of Florida, Executive Office of the Governor's Office of Tourism, Trade, and Economic Development, NIH (1R03DA025850-01A1, Nefzi), and NIH (5P41GM081261-03, Houghten). JRO is supported by a BBSRC/Pfizer CASE studentship.

REFERENCES

1. Houghten RA, Pinilla C, Giulianotti MA, Appel JR, Dooley CT, Nefzi A, Ostresh JM, Yu YP, Maggiora GM, Medina-Franco JL, Brunner D, Schneider J (2008) Strategies for the use of mixture-based synthetic combinatorial libraries: Scaffold ranking, direct testing, in vivo, and enhanced deconvolution by computational methods. J Comb Chem 10:3-19
2. Houghten RA, Pinilla C, Appel JR, Blondelle SE, Dooley CT, Eichler J, Nefzi A, Ostresh JM (1999) Mixture-based synthetic combinatorial libraries. J Med Chem 42:3743-3778
3. Kennedy JP, Williams L, Bridges TM, Daniels RN, Weaver D, Lindsley CW (2008) Application of combinatorial chemistry science on modern drug discovery. J Comb Chem 10:345-354
4. Ostresh JM, Husar GM, Blondelle SE, Dorner B, Weber PA, Houghten RA (1994) Libraries from libraries - chemical transformation of combinatorial libraries to extend the range and repertoire of chemical diversity. Proc Natl Acad Sci USA 91:11138-11142
5. Burke MD, Berger EM, Schreiber SL (2003) Generating diverse skeletons of small molecules combinatorially. Science 302:613-618
6. Martin D. Burke SLS (2004) A planning strategy for diversity-oriented synthesis. Angew Chem Int Ed 43:46-58
7. Spandl RJ, Bender A, Spring DR (2008) Diversity-oriented synthesis; a spectrum of approaches and results. Org Biomol Chem 6:1149-1158
8. Nefzi A, Ostresh JM, Yu J, Houghten RA (2004) Combinatorial chemistry: Libraries from libraries, the art of the diversity-oriented transformation of resin-bound peptides and chiral polyamides to low molecular weight acyclic and heterocyclic compounds. J Org Chem 69:3603-3609
9. Fergus S, Bender A, Spring DR (2005) Assessment of structural diversity in combinatorial synthesis. Curr Opin Chem Biol 9:304-309
10. Rolfe A, Lushington GH, Hanson PR Reagent based dos: A "Click, click, cyclize" Strategy to probe chemical space. Org Biomol Chem 8:2198-2203
11. Nefzi A, Giulianotti MA, Ong NA, Houghten RA (2000) Solid-phase synthesis of bis-2-imidazolidinethiones from resin-bound tripeptides. Org Lett 2:3349-3350

12. Acharya AN, Ostresh JM, Houghten RA (2001) Solid-phase synthesis of bis-cyclic guanidines from tripeptides. Tetrahedron 57:9911-9914
13. Nefzi A, Giulianotti MA, Houghten RA (2001) Solid-phase synthesis of bis-heterocyclic compounds from resin-bound orthogonally protected lysine. J Comb Chem 3:68-70
14. Nefzi A, Appel J, Arutyunyan S, Houghten RA (2009) Parallel synthesis of chiral pentaamines and pyrrolidine containing bis-heterocyclic libraries. Multiple scaffolds with multiple building blocks: A double diversity for the identification of new antitubercular compounds. Bioorg Med Chem Lett 19:5169-5175
15. Agrafiotis DK (2001) A constant time algorithm for estimating the diversity of large chemical libraries. J Chem Inf Comput Sci 41:159-167
16. Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. J Chem Inf Model 49:1010-1024
17. Molecular Operating Environment (MOE), version 2009.10, Chemical Computing Group Inc., Montreal, Quebec, Canada. Available at http://www.chemcomp.com (accessed October, 2010)
18. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) Drugbank: A knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:D901-D906
19. Irwin JJ, Shoichet BK (2005) ZINC - A free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177-182
20. Austin CP, Brady LS, Insel TR, Collins FS (2004) Molecular biology: NIH molecular libraries initiative. Science 306:1138-1139
21. MACCS Structural Keys: Symyx Software S. R., CA (USA)
22. Canvas, version 1.2, Schrödinger, LLC, New York, NY, 2009
23. Jaccard P (1901) Etude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat 37:547-579

24. Bender A, Glen RC (2004) Molecular similarity: A key technique in molecular informatics. Org Biomol Chem 2:3204-3218

25. The distribution of similarity values for each LoL with 1,500 compounds was nearly identical to the similarity distribution of each LoL with 5,000 (LoLA and LoLB) and 5,400 (LoLC) compounds.

26. Medina-Franco JL, Maggiora GM, Giulianotti MA, Pinilla C, Houghten RA (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. Chem Biol Drug Des 70:393-412

27. Medina-Franco JL, Martinez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. Curr Comput-Aided Drug Des 4:322-333

28. Martinez-Mayorga K, Medina-Franco JL, Giulianotti MA, Pinilla C, Dooley CT, Appel JR, Houghten RA (2008) Conformation-opioid activity relationships of bicyclic guanidines from 3D similarity analysis. Bioorg Med Chem 16:5932-5938

29. Akella LB, DeCaprio D Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325-330

30. Kabán A, Girolami M (2001) A combined latent class and trait model for the analysis and visualization of discrete data. IEEE Trans Pattern Anal Mach Intell 23:859-872

31. Spotfire, version 9.1.2, TIBCO Software, Inc., Somerville, MA. Available at http://spotfire.tibco.com (accessed October, 2010)

32. The max- and mean- fusion similarities are non-symmetrical. For example, compare the MFS maps for the relationships LoLB (test)-LoLA (reference), and LoLA (test)-LoLB (reference). If molecule “m” in a compound collection “M” is the nearest neighbor of molecule “n” in a second compound collection “N”, this does not necessarily mean that molecule “n” is the nearest neighbor of “m”. This is schematically illustrated in Figure S1 in the Supporting information

33. Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK (2009) Quantifying biogenic bias in screening libraries. Nat Chem Biol 5:479-483

34. Rosén J, Gottfries J, Muresan S, Backlund A, Oprea TI (2009) Novel chemical space exploration via natural products. J Med Chem 52:1953-1962

35. Owen JR, Nabney IT, Paolini GV Visualization of molecular fingerprints. J Chem Inf Model. submitted for publication, 2010

36. Maniyar DM, Nabney IT (2006) Visual data mining using principled projection algorithms and information visualization techniques. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, pp 643-648

37. Maggiora GM (2006) On outliers and activity cliffs-why QSAR often disappoints. J Chem Inf Model 46:1535-1535

38. 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 for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. J Chem Inf Model 47:342-353

39. Medina-Franco JL, Martínez-Mayorga K, Bender A, Marín RM, Giulianotti MA, Pinilla C, Houghten RA (2009) Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. J Chem Inf Model 49:477-491

40. Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model 49:108-119

41. Shelat AA, Guy RK (2007) The interdependence between screening methods and screening libraries. Curr Opin Chem Biol 11:244-251

42. We found that the random subset of 3,000 compounds selected from MLSMR is a representative sample of the entire database for the type of analysis conducted in this work. The distribution of molecular properties of the random subset was very similar to the distribution of the entire MLSMR collection (Table S2 in the Supporting information)

43. Shuttleworth SJ, Connors RV, Jiasheng F, Jinqian L, Lizarzaburu ME, Qiu W, Sharma R, Wanska M, Zhang AJ (2005) Design and synthesis of protein superfamily-targeted chemical libraries for lead identification and optimization. Curr Med Chem 12:1239-1281

44. Chen G, Zheng S, Luo X, Shen J, Zhu W, Liu H, Gui C, Zhang J, Zheng M, Puah CM, Chen K, Jiang H (2005) Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. J Comb Chem 7:398-406

45. Tommasi RC, Cornella, I (2006) Focused libraries: The evolution in strategy from large-diversity libraries to the focused library approach. In: Bartlett PE, Entzeroth M (ed) Exploiting chemical diversity for drug discovery. The Royal Society of Chemistry, Cambridge, pp 163-199

46. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M (2006) Bridging chemical and biological space: "Target fishing" Using 2D and 3D molecular descriptors. J Med Chem 49:6802-6810

47. Wale N, Karypis G (2009) Target fishing for chemical compounds using target-ligand activity data and ranking based methods. J Chem Inf Model 49:2190-2201

Supporting information: Non-symmetry in nearest neighbor relationships (Figure S1); property space of LoL, drugs, NCI diversity, and MLSMR (Figure S2); building blocks used to enumerate the combinatorial libraries (Table S1); molecular properties profile of combinatorial libraries and other data sets considered in the study (Table S2); mean relative Tanimoto distance of drugs to combinatorial libraries in LTM plots (Table S3); distribution of maximum and mean Tanimoto similarities of LoL to drugs, NCI diversity, and MLSMR using different structural representations (Table S4); loadings for the six principal components of the property space of drugs, NCI diversity, and MLSMR (Table S5). The scripts to generate the LTM plots are available upon request.

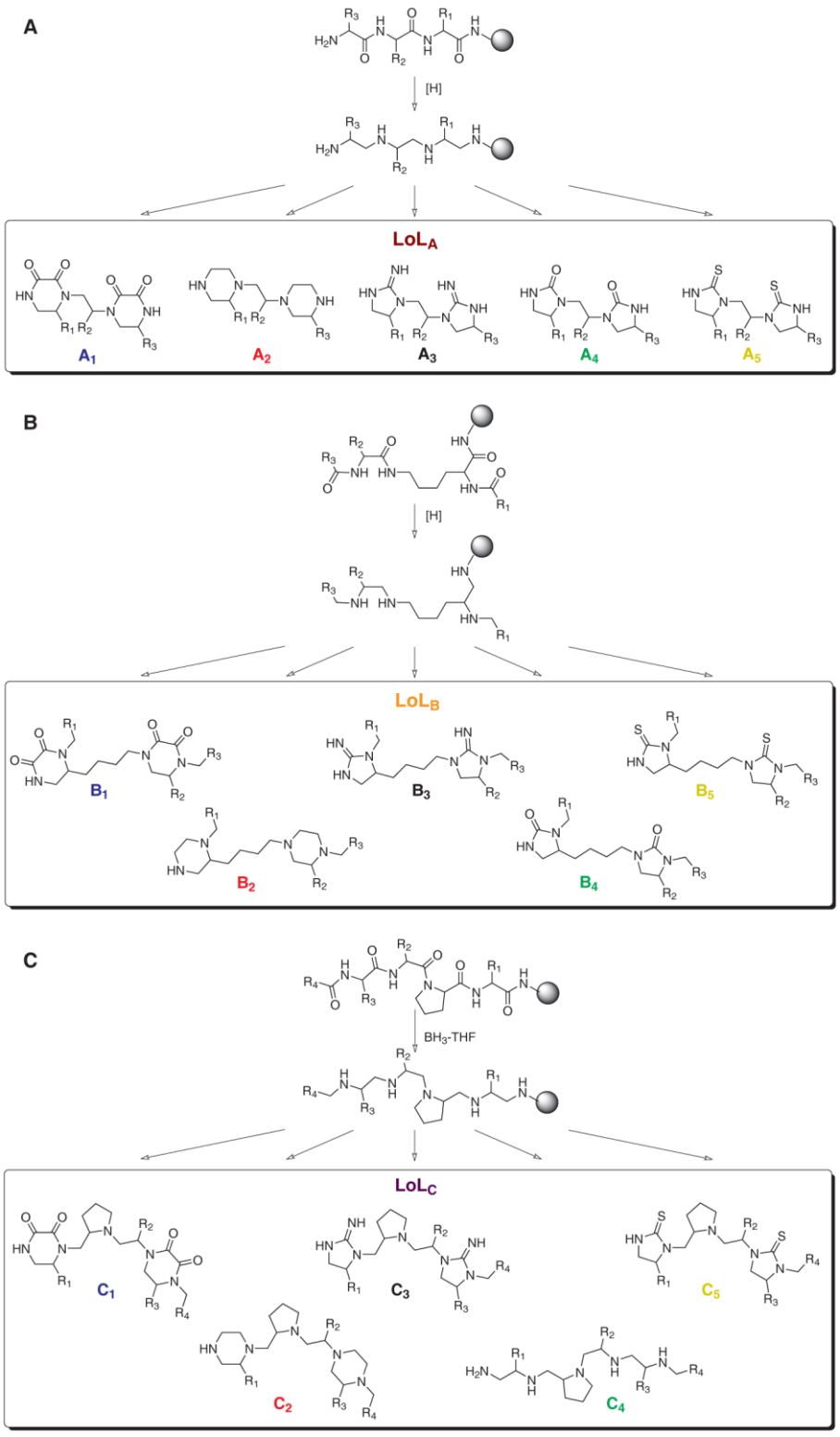
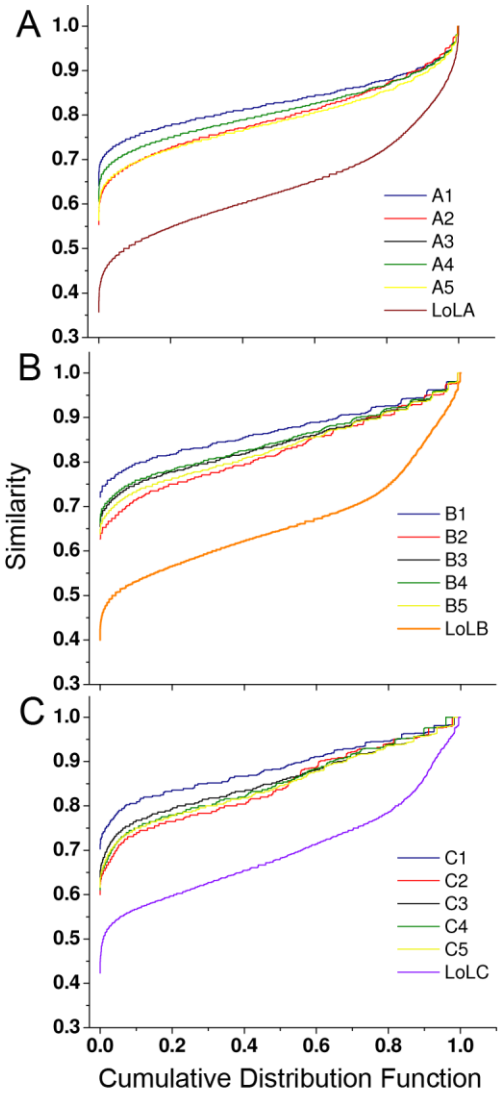


Figure 1 Fifteen combinatorial libraries organized into three libraries from libraries (LoL) analyzed in this study. Each library within a LoL shares the same number of diversity positions, identical building block side chain functionalities, and the same number of compounds within each library



	Max	Q3	Median	Q1	Min	Mean	StdDev	U95	L95
LoLA	1	0.702	0.627	0.563	0.357	0.641	0.110	0.641	0.640
A1	1	0.870	0.828	0.787	0.641	0.829	0.059	0.829	0.829
A2	1	0.854	0.792	0.740	0.554	0.796	0.080	0.797	0.796
A3	1	0.839	0.786	0.736	0.563	0.789	0.075	0.789	0.789
A4	1	0.857	0.807	0.762	0.606	0.810	0.068	0.810	0.810
A5	1	0.839	0.786	0.736	0.563	0.789	0.075	0.789	0.789

	Max	Q3	Median	Q1	Min	Mean	StdDev	U95	L95
LoLB	1	0.720	0.647	0.581	0.400	0.664	0.115	0.664	0.664
B1	1	0.915	0.873	0.828	0.721	0.872	0.059	0.872	0.872
B2	1	0.900	0.821	0.760	0.627	0.826	0.085	0.826	0.826
B3	1	0.898	0.840	0.788	0.655	0.843	0.072	0.843	0.843
B4	1	0.904	0.846	0.793	0.661	0.848	0.071	0.848	0.848
B5	1	0.896	0.830	0.774	0.638	0.834	0.078	0.834	0.834

	Max	Q3	Median	Q1	Min	Mean	StdDev	U95	L95
LoLC	1	0.765	0.681	0.611	0.423	0.698	0.113	0.698	0.698
C1	1	0.944	0.883	0.839	0.703	0.888	0.062	0.888	0.888
C2	1	0.930	0.837	0.780	0.600	0.848	0.090	0.848	0.848
C3	1	0.920	0.857	0.804	0.633	0.859	0.074	0.860	0.859
C4	1	0.930	0.851	0.792	0.610	0.855	0.084	0.855	0.855
C5	1	0.918	0.846	0.786	0.617	0.850	0.081	0.850	0.850

Figure 2 Intra-libraries similarities using MACCS keys. Figure shows the cumulative distribution function (CDF) of pair-wise similarities for each library and LoL. CDF for each library indicates the distribution of 499500 pairwise-comparisons taken from the similarity matrix. The CDF for the LoLs also contains 499500 points taken at random from the entire similarity matrix. (See also Figure 3 and text for discussion)

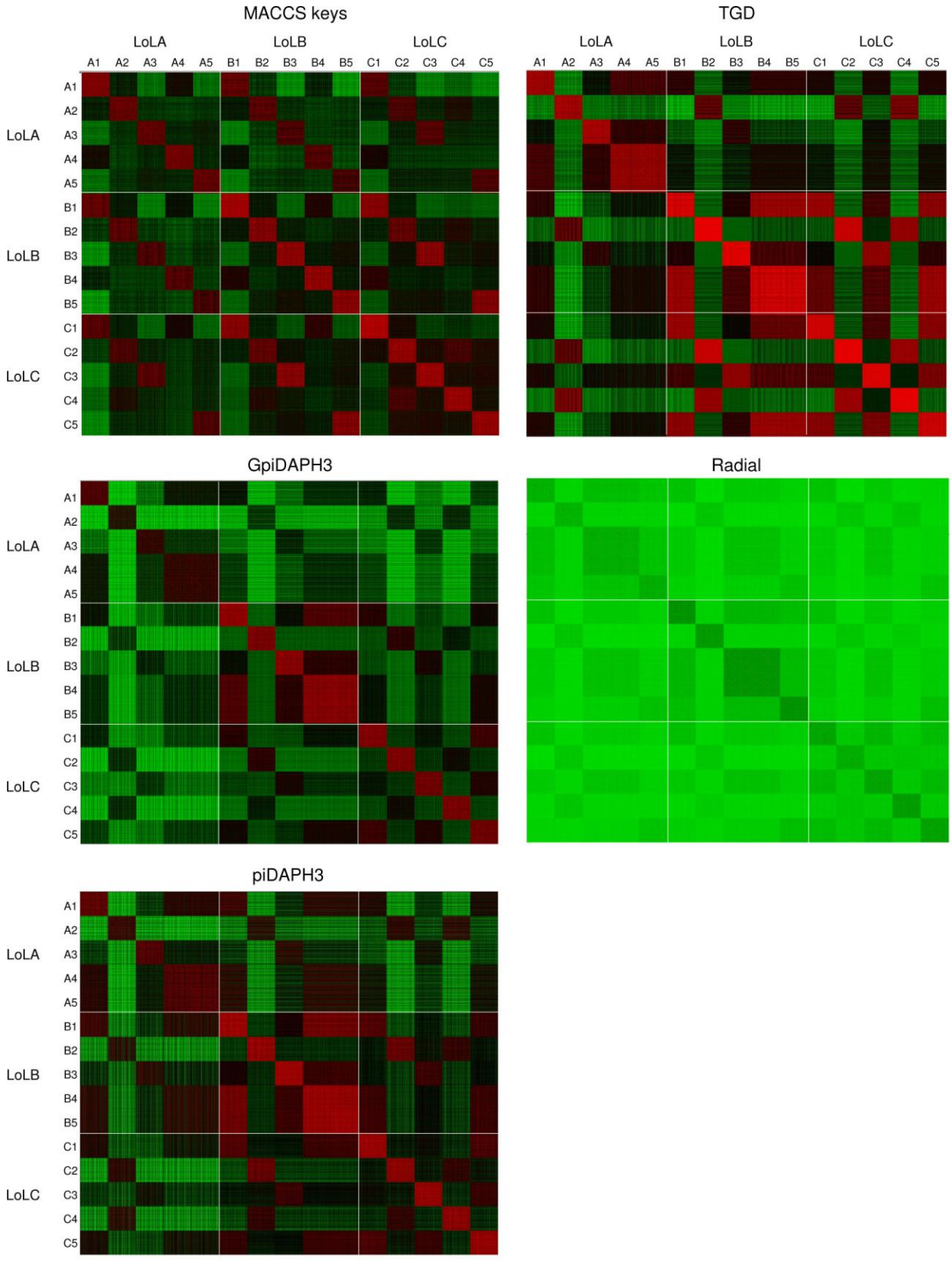
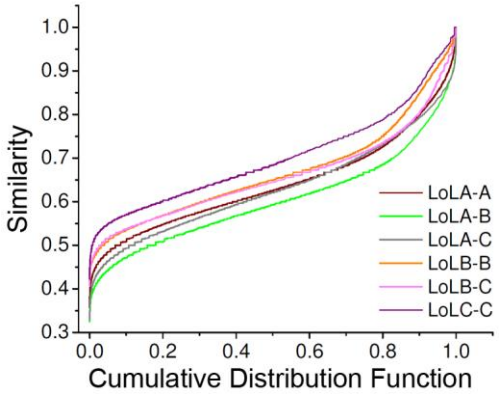


Figure 3 Heat maps of similarity matrices comparing 15 libraries (300 compounds at random per library) using different fingerprints, 2D (MACCS, TGD, GpiDAPH3 and radial) and 3D (piDAPH3). Similarity is colored using a continuous color scale from red (high similarity) to green (low similarity)



LoL	Max	Q3	Median	Q1	Min	Mean	StdDev	U95	L95
A-A	1	0.702	0.625	0.563	0.356	0.640	0.110	0.640	0.640
A-B	1	0.667	0.593	0.525	0.324	0.604	0.110	0.604	0.603
A-C	1	0.707	0.621	0.548	0.329	0.632	0.113	0.632	0.631
B-B	1	0.722	0.649	0.583	0.406	0.667	0.116	0.667	0.667
B-C	1	0.714	0.644	0.581	0.406	0.658	0.105	0.659	0.658
C-C	1	0.768	0.684	0.615	0.423	0.701	0.112	0.701	0.701

Figure 4 Cumulative distribution functions of the MACCS keys pairwise similarities between the LoLs

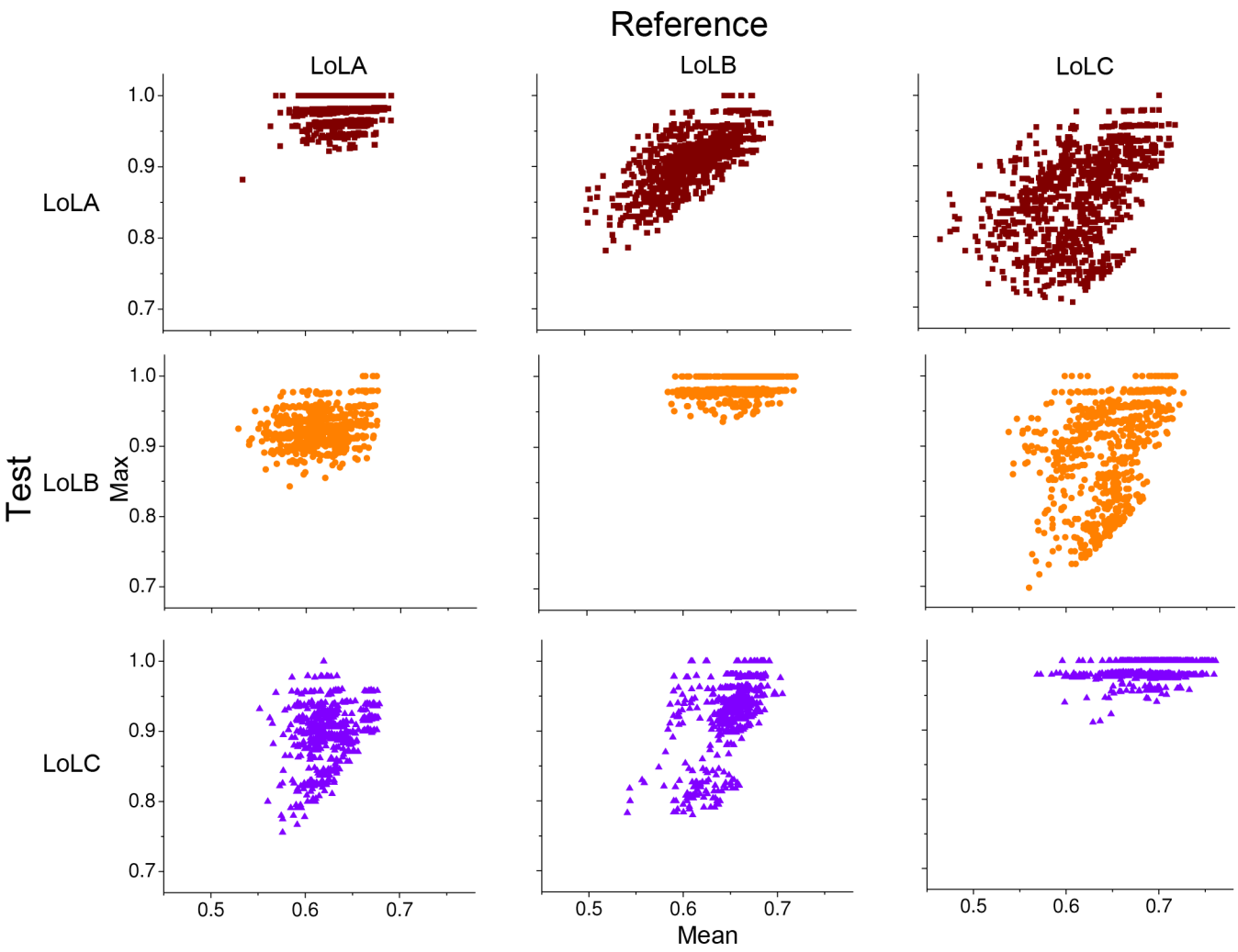


Figure 5 Multi-fusion similarity maps comparing the relationship of the LoLs. The reference LoLs are designated along the top and the test LoLs along the left-hand side of the figure. The three plots along the principal diagonal (upper left to lower right in the figure) correspond to self-referencing MFS maps. LoLA is in brown squares, LoLB in orange circles and LoLC in violet triangles

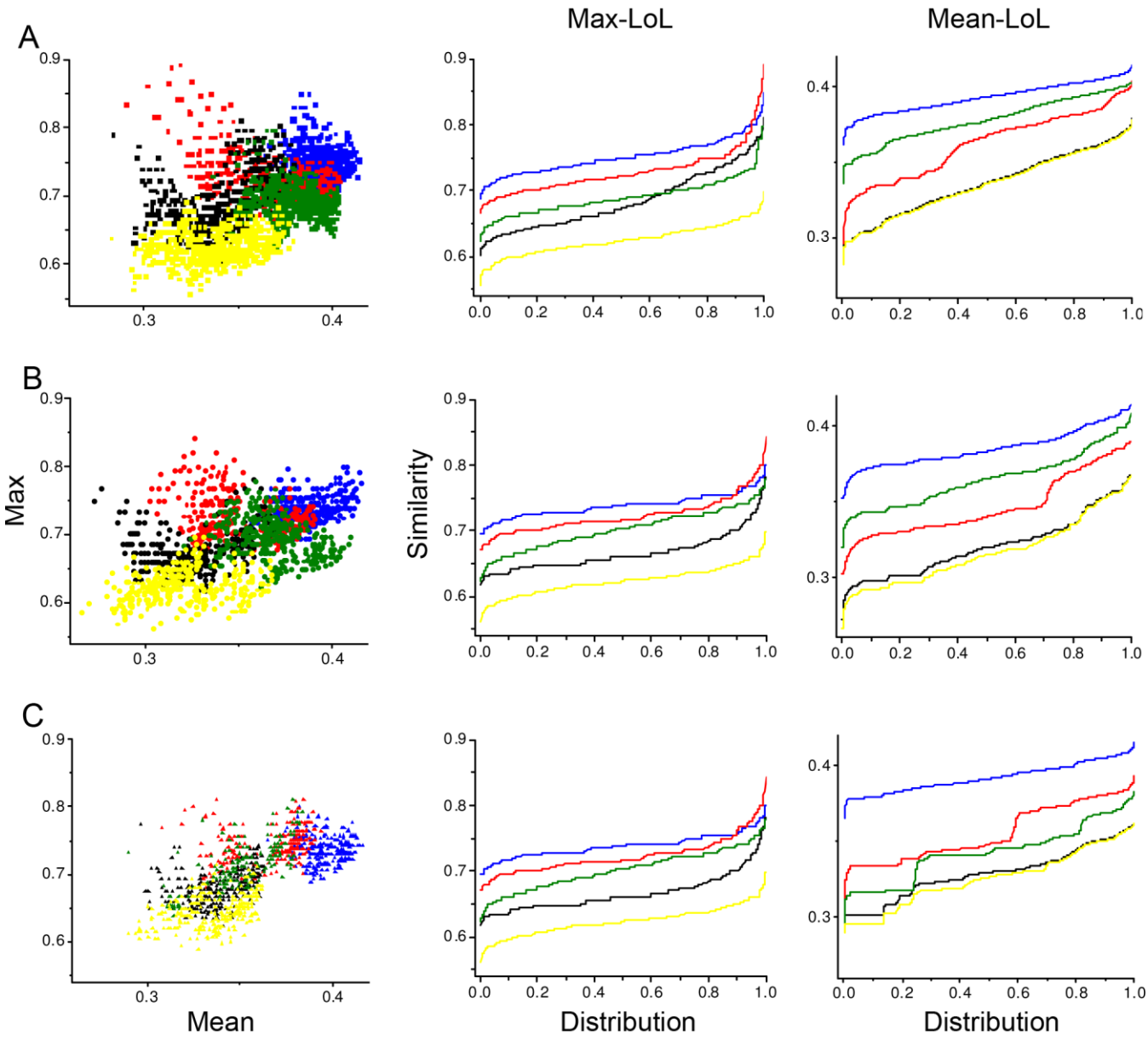


Figure 6 Relationship between a collection of drugs (reference) with the following test compounds: (A) libraries **A1**-**A5**, (B) libraries **B1**-**B5** and (C) libraries **C1**-**C5**. MFS maps and the corresponding CDF of the maximum and mean values distributions are shown

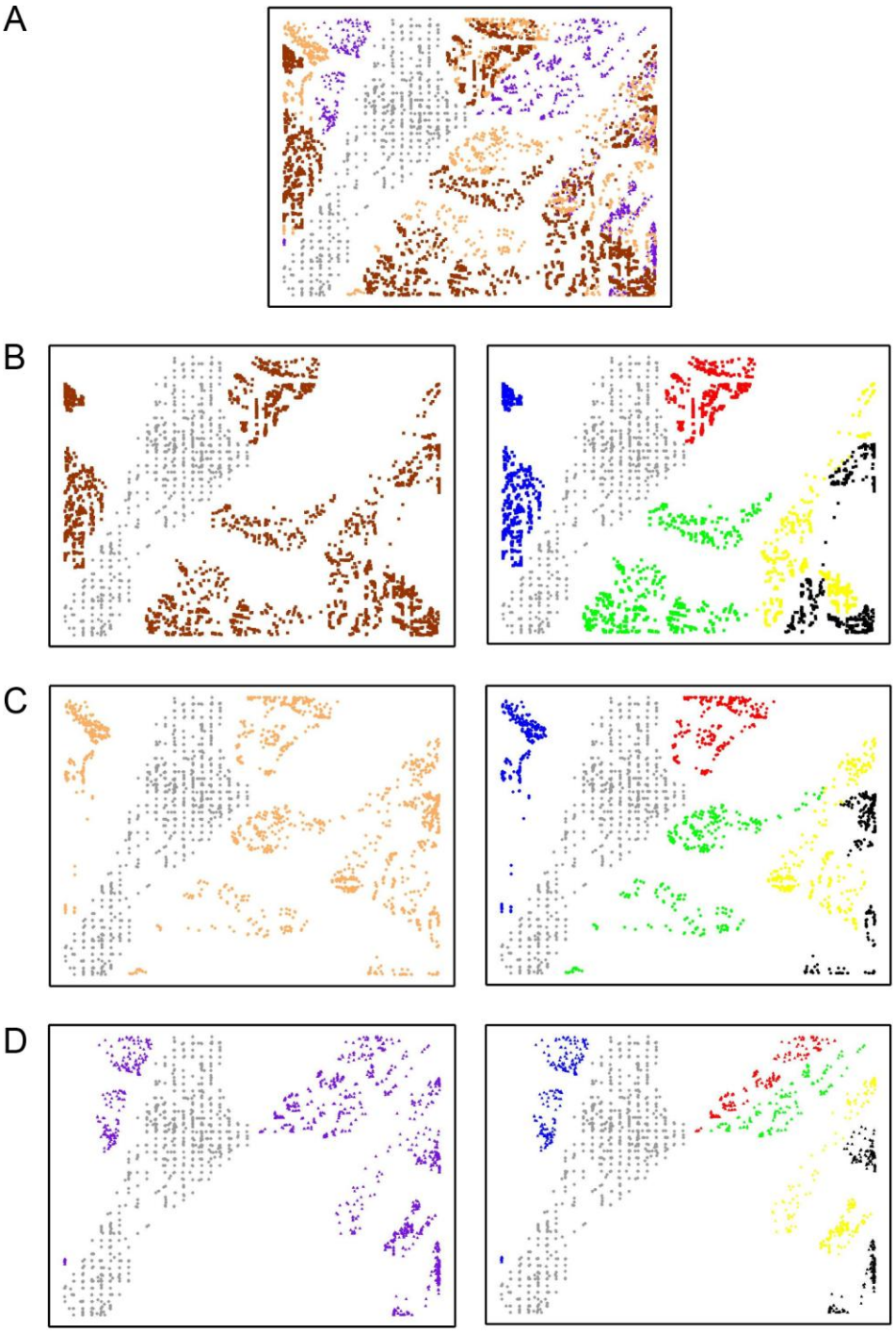


Figure 7 LTM plots of three LoLs, 15 combinatorial libraries and drugs. (A) Drugs (gray), LoLA (brown), LoLB (orange) and LoLC (purple); (B) drugs and LoLA (left), drugs and **A1**-**A5** (right); (C) drugs and LoLB (left), drugs and **B1**-**B5** (right); (D) drugs and LoLC (left), drugs and **C1**-**C5** (right). **A1**-**A5**, **B1**-**B5**, **C1**-**C5** are color- and shape-coded as in Figures 1, 5 and 6

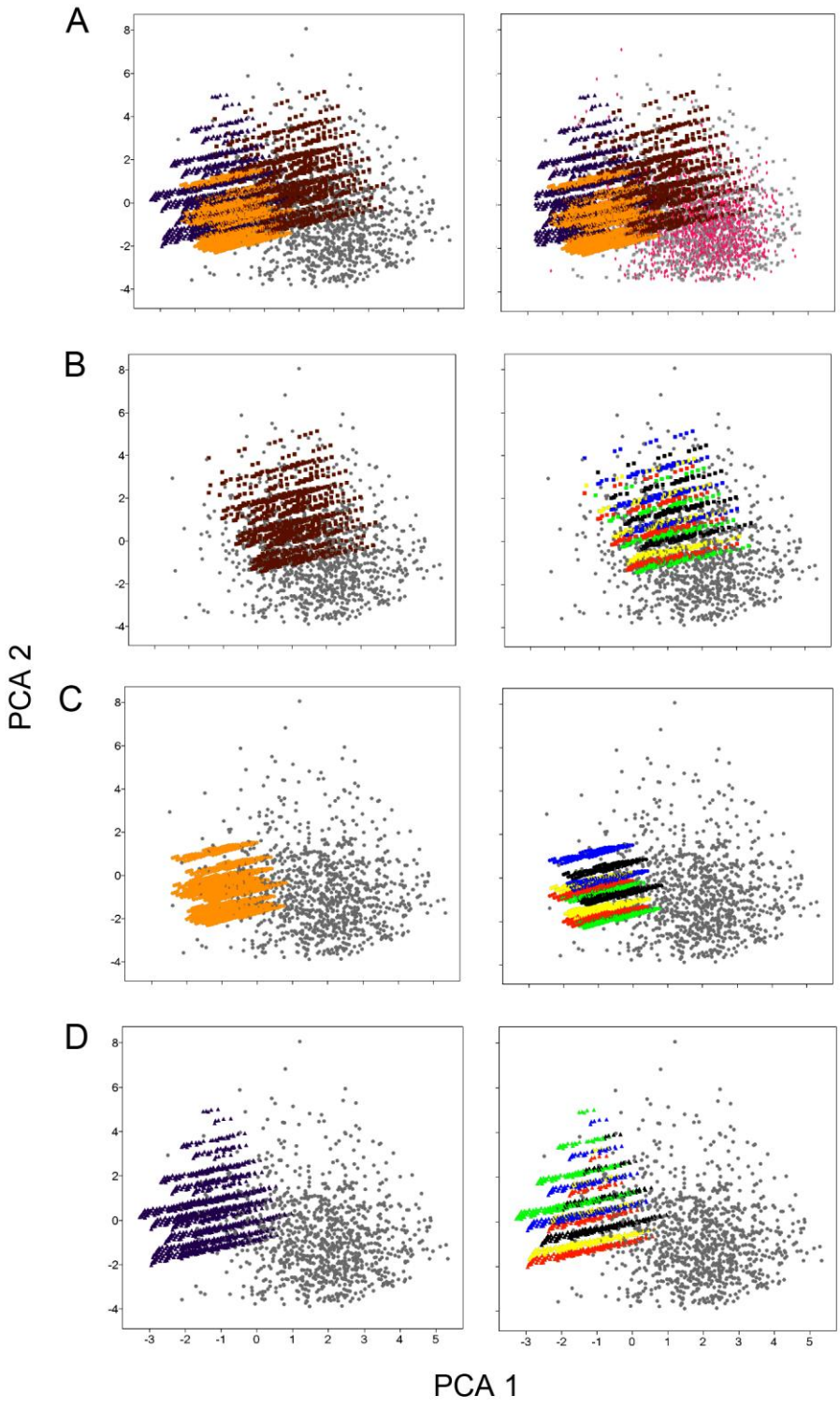


Figure 8 Property space of three LoLs, 15 combinatorial libraries, drugs (gray) and NCI diversity (pink). The first two PC account for 70.71% of the variance. The loadings are summarized in Table 2. (A) Drugs, LoLA, LoLB, LoLC and NCI Diversity; (B) drugs and LoLA (brown) and drugs and **A1**-**A5**; (C) drugs and LoLB (orange) and drugs and **B1**-**B5**; (D) drugs and LoLC (purple) and drugs and **C1**-**C5**. **A1**-**A5**, **B1**-**B5**, **C1**-**C5** are color coded as in Figures 1, 5 and 6

Table 1. Distribution of molecular properties

Library	HBA			HBD			RB		
	Median	**Mean**	**StdDev**	**Median**	**Mean**	**StdDev**	**Median**	**Mean**	**StdDev**
LoLA	4.00	3.70	1.26	3.00	3.30	1.13	6.00	6.30	1.21
A1	5.00	4.90	0.79	3.00	2.90	0.79	6.00	6.30	1.21
A2	5.00	4.90	0.79	3.00	2.90	0.79	6.00	6.30	1.21
A3	3.00	2.90	0.79	5.00	4.90	0.79	6.00	6.30	1.21
A4	3.00	2.90	0.79	3.00	2.90	0.79	6.00	6.30	1.21
A5	3.00	2.90	0.79	3.00	2.90	0.79	6.00	6.30	1.21
LoLB	3.00	3.10	1.08	1.00	1.70	0.92	11.00	11.10	1.18
B1	4.00	4.30	0.46	1.00	1.30	0.46	11.00	11.10	1.18
B2	4.00	4.30	0.46	1.00	1.30	0.46	11.00	11.10	1.18
B3	2.00	2.30	0.46	3.00	3.30	0.46	11.00	11.10	1.18
B4	2.00	2.30	0.46	1.00	1.30	0.46	11.00	11.10	1.18
B5	2.00	2.30	0.46	1.00	1.30	0.46	11.00	11.10	1.18
LoLC	5.00	4.70	1.17	2.00	2.50	1.42	11.00	11.90	2.38
C1	5.00	5.50	0.65	1.00	1.50	0.65	11.00	10.90	1.29
C2	5.00	5.50	0.65	1.00	1.50	0.65	11.00	10.90	1.29
C3	3.00	3.50	0.65	3.00	3.50	0.65	11.00	10.90	1.29
C4	5.00	5.50	0.65	4.00	4.50	0.65	16.00	15.90	1.29
C5	3.00	3.50	0.65	1.00	1.50	0.65	11.00	10.90	1.29
DrugBank	2.00	2.46	1.69	1.00	1.18	1.22	5.00	5.07	3.26
NCI Diversity	2.00	2.70	1.64	1.00	1.36	1.19	3.00	3.39	2.14
MLSMR(3000)	4.00	3.72	1.82	1.00	1.07	1.03	7.00	6.71	3.48
MLSMR(347480)	3.00	3.35	1.49	1.00	1.07	0.90	6.00	6.16	2.90

	SlogP			TPSA			MW		
	Median	**Mean**	**StdDev**	**Median**	**Mean**	**StdDev**	**Median**	**Mean**	**StdDev**
LoLA	1.13	1.02	1.70	98.47	91.61	29.42	360.55	363.30	57.56
A1	-0.58	-0.63	1.47	119.05	117.03	16.07	402.45	402.07	52.71
A2	1.31	1.26	1.47	50.77	48.75	16.07	346.52	346.14	52.71
A3	1.18	1.13	1.47	98.47	96.45	16.07	344.46	344.08	52.71
A4	1.55	1.50	1.47	84.91	82.89	16.07	346.43	346.05	52.71
A5	1.88	1.83	1.47	114.95	112.93	16.07	378.57	378.18	52.71
LoLB	3.99	3.86	1.32	76.12	70.68	26.34	448.70	450.68	45.33
B1	2.31	2.21	1.00	90.03	96.10	9.28	488.67	489.45	38.97
B2	4.21	4.11	1.00	21.75	27.82	9.28	432.74	433.52	38.97
B3	4.08	3.98	1.00	69.45	75.52	9.28	430.69	431.46	38.97
B4	4.45	4.35	1.00	55.89	61.96	9.28	432.65	433.43	38.97
B5	4.78	4.68	1.00	85.93	92.00	9.28	464.79	465.56	38.97
LoLC	2.98	2.96	1.72	85.63	79.21	27.62	487.78	487.39	60.97
C1	1.50	1.53	1.50	93.27	103.39	13.06	536.69	537.11	48.36
C2	3.39	3.43	1.50	24.99	35.11	13.06	480.76	481.18	48.36
C3	3.26	3.30	1.50	72.69	82.81	13.06	478.71	479.12	48.36
C4	2.56	2.55	1.48	65.35	75.47	13.06	425.75	426.30	47.54
C5	3.96	4.00	1.50	89.17	99.29	13.06	512.81	513.23	48.36
DrugBank	1.63	1.44	2.57	67.09	71.82	41.47	310.34	310.32	91.93
NCI Diversity	2.46	2.45	1.91	64.69	69.15	33.55	272.35	283.59	84.64
MLSMR(3000)	2.80	2.72	1.59	80.47	81.41	33.30	401.47	395.19	102.14
MLSMR(347480)	2.89	2.80	1.43	74.61	76.49	28.96	358.53	362.19	83.05

Table 2. Loadings for the six principal components of the property space

Principal component	PC1	PC2	PC3	PC4	PC5	PC6
eigenvalue	2.172	2.071	0.736	0.642	0.286	0.094
cumulative eigenvalue (%)	36.192	70.709	82.97	93.677	98.440	100
HBA	0.277	0.456	-0.659	0.273	-0.289	-0.351
HBD	0.012	0.525	0.068	-0.795	-0.255	0.154
RB	0.607	0.058	-0.038	-0.225	0.742	-0.159
SlogP	0.398	-0.488	0.211	-0.251	-0.466	-0.527
TPSA	-0.0004	0.523	0.692	0.328	0.0140	-0.375
MW	0.629	0.043	0.193	0.271	-0.289	0.639

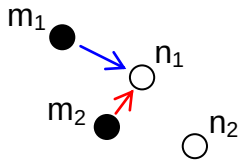
Supporting information

Increased Diversity of *Libraries from Libraries*: Chemoinformatic Analysis of Bis-Diazacyclic Libraries

Fabian López-Vallejo,[1] Adel Nefzi,[1] Andreas Bender,[2] John R. Owen,[3] Ian T. Nabney,[3] Richard A. Houghten[1] and Jose L. Medina-Franco[1,*]

[1]Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, FL 34987, USA, [2]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom, [3]Non-linearity and Complexity Research Group (NCRG), Aston University, Aston Triangle, Birmingham B4 7ET, United Kingdom

Table of Contents:



n_1 is the nearest-neighbor (NN) of m_1
however, m_1 is NOT the NN of n_1:
m_2 is the NN of n_1

Figure S1 Non-symmetry in nearest neighbor relationships. If molecule "m" in a compound collection "M" is the nearest neighbor of molecule "n" in a second compound collection "N", this does not necessarily mean that molecule "n" is the nearest neighbor of "m".

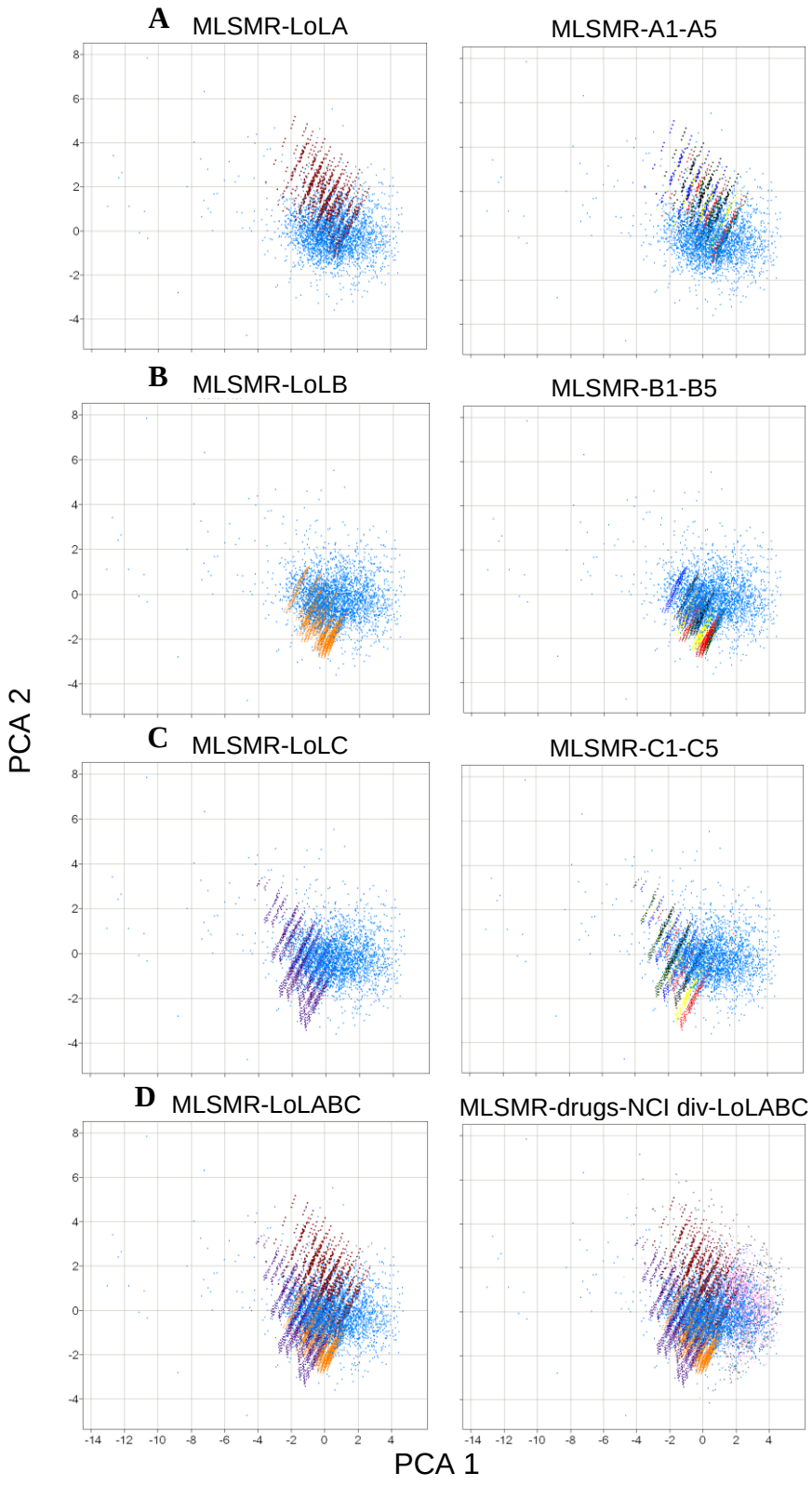


Figure S2 Property space of 3 LoLs, 15 combinatorial libraries, drugs (gray), NCI diversity (pink) and 3000 compounds from MLSMR selected at random (light blue). The first two PC account for 70.41% of the variance. The loadings are summarized in Table S5. Figures S2A, S2B and S2C show a comparison of the property space of MLSMR with each LoL. The left hand side of these figures shows the LoLs in one color and on the right hand side show, in different colors, each of the five combinatorial libraries within a LoL. The largest loadings for the first PC correspond to MW followed by RB. In contrast, the largest loadings for the second PC correspond to SlogP followed by TPSA and HBD.

Table S1. Building blocks used to enumerate the coombinatorial libraries. The name of the functionality is also indicated

LoLA	$R_1 = R_2 = R_3$ (10 amino acids)
(10^3 = 1000 compounds per library)	Methyl - Boc-L-Ala
	Hydrogen - Boc-Gly
	S-Isobutyl - Boc-L-Leu
	S-Isopropyl - Boc-L-Val
	S-4-Hydroxybenzyl - Boc-L-Tyr(BrZ)
	S-Hydroxymethyl - Boc-D-Ser(Bzl)
	(S-S)-1-Hydroxyethyl - Boc-D-Thr(Bzl)
	S-Phenyl - Boc-L-Phenylglycine
	S-Propyl - Boc-L-Norvaline
	S-Cyclohexyl - Boc-L-Cyclohexylalanine

LoLB	$R_1 = R_3$ (10 carboxilic acids)	R_2 = (10 amino acids)
(10^3 = 1000 compounds per library)	Phenethyl - phenylacetic	S-methyl - Boc-L-Ala
	Butyl - butyric	Hydrogen - Boc-Gly
	Isobutyl - isobutyric	S-isobutyl - Boc-L-Leu
	2-Methylbutyl - 2 -methylbutyric	S-isopropyl - Boc-L-Val
	3-Methylpentyl - 3-methylvaleric	S-4-hydroxybenzyl - Boc-L-Tyr(BrZ)
	4-Methyl-benzyl - p-toluic	S-hydroxymethyl - Boc-D-Ser(Bzl)
	Cyclopently-methyl - cyclopentanecarboxylic	(S-S)-1-hydroxyethyl - Boc-D-Thr(Bzl)
	Cyclohexyl-methyl - cyclohexanecarboxylic	S-phenyl - Boc-L-Phenylglycine
	(2-Methyl-cyclopropyl)-methyl - 2-methylcyclopropanecarboxylic	S-propyl - Boc-L-Norvaline
	Cyclobutyl-methyl - Cyclobutanecarboxylic	S-cyclohexyl - Boc-L-Cyclohexylalanine

LoLC	$R_1 = R_2 = R_3$ (6 amino acids)	R_4 (5 carboxilic acids)
(6 x 6 x 6 x 5 = 1080 compounds per library)	S-Methyl - Boc-L-Ala	Butyl - butyric
	Hydrogen - Boc-Gly	Isobutyl - Isobutyric
	S-isobutyl - Boc-L-Leu	2-Methylbutyl - 2-methylbutyric
	S-isopropyl - Boc-L-Val	4-Methyl-benzyl - p-toluic
	S-hydroxymethyl - Boc-D-Ser(Bzl)	Cyclopently-ethyl - cyclopentanecarboxylic
	S-phenyl - Boc-L-Phenylglycine	

Table S2. Molecular properties profile of combinatorial libraries and other data sets

Library	Column	Max	Q3	Median	Q1	Min	Mean	StdDev	U95	L95
LoLA LoLB LoLC	HBA	8.000	5.000	4.000	3.000	2.000	3.856	1.350	3.877	3.835
	HBD	7.000	3.000	2.000	1.000	1.000	2.500	1.346	2.521	2.479
	RB	20.000	12.000	10.000	7.000	3.000	9.822	2.997	9.869	9.775
	SlogP	8.024	4.115	2.793	1.309	-5.014	2.622	1.980	2.653	2.591
	TPSA	159.510	98.470	85.930	65.350	21.750	80.465	29.076	80.924	80.006
	MW	683.853	487.781	437.761	384.657	196.258	435.184	75.875	436.383	433.986
LoLA	HBA	7.000	5.000	4.000	3.000	2.000	3.700	1.261	3.735	3.665
	HBD	7.000	4.000	3.000	2.000	2.000	3.300	1.127	3.331	3.269
	RB	9.000	7.000	6.000	5.000	3.000	6.300	1.213	6.334	6.266
	SlogP	5.213	2.228	1.131	-0.137	-5.014	1.017	1.703	1.064	0.970
	TPSA	159.510	114.950	98.470	71.000	30.540	91.607	29.418	92.422	90.792
	MW	572.618	402.451	360.546	324.381	196.258	363.304	57.558	364.899	361.709
A1	HBA	7.000	5.000	5.000	4.000	4.000	4.900	0.794	4.949	4.851
	HBD	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	RB	9.000	7.000	6.000	5.000	3.000	6.300	1.213	6.375	6.225
	SlogP	2.750	0.438	-0.582	-1.657	-5.014	-0.634	1.471	-0.543	-0.725
	TPSA	159.510	119.050	119.050	98.820	98.820	117.027	16.065	118.023	116.031
	MW	572.618	436.468	402.451	366.462	254.246	402.069	52.705	405.336	398.803
A2	HBA	7.000	5.000	5.000	4.000	4.000	4.900	0.794	4.949	4.851
	HBD	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	RB	9.000	7.000	6.000	5.000	3.000	6.300	1.213	6.375	6.225
	SlogP	4.643	2.332	1.311	0.236	-3.121	1.260	1.471	1.351	1.168
	TPSA	91.230	50.770	50.770	30.540	30.540	48.747	16.065	49.743	47.751
	MW	516.686	380.536	346.519	310.530	198.314	346.137	52.705	349.404	342.871
A3	HBA	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	HBD	7.000	5.000	5.000	4.000	4.000	4.900	0.794	4.949	4.851
	RB	9.000	7.000	6.000	5.000	3.000	6.300	1.213	6.375	6.225
	SlogP	4.513	2.201	1.181	0.106	-3.251	1.129	1.471	1.220	1.038
	TPSA	138.930	98.470	98.470	78.240	78.240	96.447	16.065	97.443	95.451
	MW	514.630	378.480	344.463	308.474	196.258	344.081	52.705	347.348	340.815
A4	HBA	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	HBD	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	RB	9.000	7.000	6.000	5.000	3.000	6.300	1.213	6.375	6.225
	SlogP	4.884	2.572	1.551	0.476	-2.881	1.500	1.471	1.591	1.408
	TPSA	125.370	84.910	84.910	64.680	64.680	82.887	16.065	83.883	81.891
	MW	516.598	380.448	346.431	310.442	198.226	346.049	52.705	349.316	342.783
A5	HBA	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	HBD	5.000	3.000	3.000	2.000	2.000	2.900	0.794	2.949	2.851
	RB	9.000	7.000	6.000	5.000	3.000	6.300	1.213	6.375	6.225
	SlogP	5.213	2.902	1.881	0.806	-2.551	1.829	1.471	1.921	1.738
	TPSA	155.410	114.950	114.950	94.720	94.720	112.927	16.065	113.923	111.931
	MW	548.732	412.582	378.565	342.576	230.360	378.183	52.705	381.450	374.917
LoLB	HBA	5.000	4.000	3.000	2.000	2.000	3.100	1.082	3.130	3.070
	HBD	4.000	2.000	1.000	1.000	1.000	1.700	0.922	1.726	1.674
	RB	15.000	12.000	11.000	10.000	9.000	11.100	1.179	11.133	11.067
	SlogP	7.117	4.879	3.985	2.953	-0.173	3.863	1.319	3.899	3.826
	TPSA	110.260	90.030	76.120	55.890	21.750	70.679	26.335	71.409	69.949
	MW	596.728	480.830	448.696	418.674	336.528	450.684	45.329	451.940	449.428

Table S2. (continued)

B1	**HBA**	5.000	5.000	4.000	4.000	4.000	4.300	0.459	4.328	4.272
	HBD	2.000	2.000	1.000	1.000	1.000	1.300	0.459	1.328	1.272
	RB	15.000	12.000	11.000	10.000	9.000	11.100	1.180	11.173	11.027
	SlogP	4.653	2.950	2.313	1.479	-0.173	2.212	1.001	2.274	2.150
	TPSA	110.260	110.260	90.030	90.030	90.030	96.099	9.275	96.674	95.524
	MW	596.728	515.215	488.673	462.635	394.516	489.449	38.969	491.864	487.034
B2	**HBA**	5.000	5.000	4.000	4.000	4.000	4.300	0.459	4.328	4.272
	HBD	2.000	2.000	1.000	1.000	1.000	1.300	0.459	1.328	1.272
	RB	15.000	12.000	11.000	10.000	9.000	11.100	1.180	11.173	11.027
	SlogP	6.547	4.843	4.207	3.373	1.721	4.106	1.001	4.168	4.044
	TPSA	41.980	41.980	21.750	21.750	21.750	27.819	9.275	28.394	27.244
	MW	540.796	459.283	432.741	406.703	338.584	433.517	38.969	435.932	431.102
B3	**HBA**	3.000	3.000	2.000	2.000	2.000	2.300	0.459	2.328	2.272
	HBD	4.000	4.000	3.000	3.000	3.000	3.300	0.459	3.328	3.272
	RB	15.000	12.000	11.000	10.000	9.000	11.100	1.180	11.173	11.027
	SlogP	6.417	4.713	4.077	3.242	1.591	3.975	1.001	4.037	3.913
	TPSA	89.680	89.680	69.450	69.450	69.450	75.519	9.275	76.094	74.944
	MW	538.740	457.227	430.685	404.647	336.528	431.461	38.969	433.876	429.046
B4	**HBA**	3.000	3.000	2.000	2.000	2.000	2.300	0.459	2.328	2.272
	HBD	2.000	2.000	1.000	1.000	1.000	1.300	0.459	1.328	1.272
	RB	15.000	12.000	11.000	10.000	9.000	11.100	1.180	11.173	11.027
	SlogP	6.787	5.083	4.447	3.613	1.961	4.346	1.001	4.408	4.284
	TPSA	76.120	76.120	55.890	55.890	55.890	61.959	9.275	62.534	61.384
	MW	540.708	459.195	432.653	406.615	338.496	433.429	38.969	435.844	431.014
B5	**HBA**	3.000	3.000	2.000	2.000	2.000	2.300	0.459	2.328	2.272
	HBD	2.000	2.000	1.000	1.000	1.000	1.300	0.459	1.328	1.272
	RB	15.000	12.000	11.000	10.000	9.000	11.100	1.180	11.173	11.027
	SlogP	7.117	5.413	4.777	3.942	2.291	4.675	1.001	4.737	4.613
	TPSA	106.160	106.160	85.930	85.930	85.930	91.999	9.275	92.574	91.424
	MW	572.842	491.329	464.787	438.749	370.630	465.563	38.969	467.978	463.148
LoLC	**HBA**	8.000	6.000	5.000	4.000	3.000	4.700	1.173	4.731	4.669
	HBD	7.000	4.000	2.000	1.000	1.000	2.500	1.420	2.538	2.462
	RB	20.000	13.000	11.000	10.000	8.000	11.900	2.379	11.963	11.837
	SlogP	8.024	4.198	2.981	1.794	-2.791	2.960	1.719	3.005	2.914
	TPSA	153.960	93.270	85.630	65.350	24.990	79.209	27.615	79.946	78.472
	MW	683.853	529.862	487.781	445.740	299.507	487.388	60.966	489.014	485.762
C1	**HBA**	8.000	6.000	5.000	5.000	5.000	5.500	0.646	5.539	5.461
	HBD	4.000	2.000	1.000	1.000	1.000	1.500	0.646	1.539	1.461
	RB	15.000	12.000	11.000	10.000	8.000	10.900	1.288	10.977	10.823
	SlogP	5.561	2.551	1.496	0.438	-2.791	1.532	1.499	1.621	1.442
	TPSA	153.960	113.500	93.270	93.270	93.270	103.385	13.064	104.164	102.606
	MW	683.853	569.747	536.694	505.693	407.515	537.112	48.356	539.996	534.228
C2	**HBA**	8.000	6.000	5.000	5.000	5.000	5.500	0.646	5.539	5.461
	HBD	4.000	2.000	1.000	1.000	1.000	1.500	0.646	1.539	1.461
	RB	15.000	12.000	11.000	10.000	8.000	10.900	1.288	10.977	10.823
	SlogP	7.454	4.444	3.389	2.332	-0.898	3.425	1.499	3.515	3.336
	TPSA	85.680	45.220	24.990	24.990	24.990	35.105	13.064	35.884	34.326
	MW	627.921	513.815	480.762	449.761	351.583	481.180	48.356	484.064	478.296

Table S2. (continued)

C3	**HBA**	6.000	4.000	3.000	3.000	3.000	3.500	0.646	3.539	3.461
	HBD	6.000	4.000	3.000	3.000	3.000	3.500	0.646	3.539	3.461
	RB	15.000	12.000	11.000	10.000	8.000	10.900	1.288	10.977	10.823
	SlogP	7.324	4.314	3.259	2.201	-1.028	3.295	1.499	3.384	3.205
	TPSA	133.380	92.920	72.690	72.690	72.690	82.805	13.064	83.584	82.026
	MW	625.865	511.759	478.706	447.705	349.527	479.124	48.356	482.008	476.240
C4	**HBA**	8.000	6.000	5.000	5.000	5.000	5.500	0.646	5.539	5.461
	HBD	7.000	5.000	4.000	4.000	4.000	4.500	0.646	4.539	4.461
	RB	20.000	17.000	16.000	15.000	13.000	15.900	1.288	15.977	15.823
	SlogP	6.269	3.591	2.560	1.533	-1.693	2.552	1.479	2.640	2.464
	TPSA	126.040	85.580	65.350	65.350	65.350	75.465	13.064	76.244	74.686
	MW	575.845	459.767	425.750	391.604	299.507	426.299	47.535	429.134	423.464
C5	**HBA**	6.000	4.000	3.000	3.000	3.000	3.500	0.646	3.539	3.461
	HBD	4.000	2.000	1.000	1.000	1.000	1.500	0.646	1.539	1.461
	RB	15.000	12.000	11.000	10.000	8.000	10.900	1.288	10.977	10.823
	SlogP	8.024	5.014	3.959	2.902	-0.328	3.995	1.499	4.084	3.906
	TPSA	149.860	109.400	89.170	89.170	89.170	99.285	13.064	100.064	98.506
	MW	659.967	545.861	512.808	481.807	383.629	513.226	48.356	516.110	510.342
DrugBank	**HBA**	11.000	3.000	2.000	1.000	0.000	2.460	1.686	2.546	2.375
	HBD	8.000	2.000	1.000	0.000	0.000	1.175	1.221	1.237	1.113
	RB	24.000	7.000	5.000	3.000	0.000	5.074	3.258	5.240	4.909
	SlogP	9.908	3.033	1.634	0.192	-13.609	1.442	2.571	1.572	1.311
	TPSA	266.490	97.530	67.085	40.460	0.000	71.819	41.471	73.924	69.713
	MW	537.576	376.776	310.336	246.078	75.067	310.324	91.933	314.992	305.656
NCI	**HBA**	12.000	3.000	2.000	2.000	0.000	2.695	1.635	2.770	2.620
	HBD	8.000	2.000	1.000	0.000	0.000	1.360	1.188	1.415	1.306
	RB	19.000	5.000	3.000	2.000	0.000	3.392	2.138	3.490	3.294
	SlogP	10.392	3.614	2.464	1.188	-3.342	2.452	1.908	2.539	2.365
	TPSA	214.110	87.157	64.690	46.530	0.000	69.148	33.550	70.685	67.612
	MW	696.129	328.544	272.348	224.262	114.108	283.589	84.642	287.465	279.713
MLSMR (3000)	**HBA**	18.000	5.000	4.000	3.000	0.000	3.721	1.820	3.786	3.656
	HBD	11.000	2.000	1.000	0.000	0.000	1.065	1.029	1.101	1.028
	RB	58.000	9.000	7.000	5.000	0.000	6.712	3.484	6.836	6.587
	SlogP	10.395	3.750	2.798	1.827	-4.562	2.715	1.594	2.772	2.658
	TPSA	356.010	98.770	80.470	60.912	0.000	81.410	33.302	82.601	80.218
	MW	1304.000	463.555	401.465	328.408	75.047	395.189	102.136	398.844	391.534
MLSMR	**HBA**	38.000	4.000	3.000	2.000	0.000	3.350	1.490	3.350	3.340
	HBD	38.000	2.000	1.000	0.000	0.000	1.074	0.901	1.077	1.071
	RB	143.00	8.00	6.00	4.00	0.00	6.161	2.896	6.171	6.151
	SlogP	25.738	3.728	2.889	1.966	-25.556	2.802	1.434	2.806	2.797
	TPSA	1470.000	93.210	74.610	57.120	0.000	76.490	28.960	76.586	76.393
	MW	3362.000	414.410	358.526	306.342	30.006	362.194	83.045	362.470	361.918

Table S3. Mean relative Tanimoto distance of drugs to combinatorial libraries in LTM plots

LoL	Library	Distance	Std
LoLA	**A1**	100.00	100.00
	A2	106.57	287.25
	A3	109.57	213.63
	A4	102.78	147.69
	A5	109.70	213.32
LoLB	**B1**	100.00	100.00
	B2	108.31	206.65
	B3	110.70	162.59
	B4	103.61	144.50
	B5	111.35	177.07
LoLC	**C1**	100.00	100.00
	C2	107.60	202.91
	C3	110.21	179.62
	C4	109.93	201.73
	C5	110.76	198.90

Table S4. Distribution of maximum and mean Tanimoto similarities of LoL to drugs, NCI diversity, and MLSMR using different structural representations

Comparison with drugs (DrugBank)

TGD

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.933	0.848	0.795	0.765	0.674	0.801	0.052	0.802	0.799
LoLB	0.929	0.869	0.819	0.795	0.699	0.825	0.058	0.826	0.823
LoLC	0.920	0.891	0.880	0.804	0.739	0.856	0.048	0.857	0.854
				Mean					
LoLA	0.561	0.539	0.518	0.499	0.455	0.519	0.026	0.519	0.518
LoLB	0.537	0.519	0.502	0.489	0.460	0.502	0.020	0.503	0.501
LoLC	0.536	0.512	0.500	0.491	0.473	0.502	0.015	0.502	0.502

GpiDAPH3

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.654	0.504	0.471	0.433	0.290	0.466	0.052	0.468	0.465
LoLB	0.620	0.553	0.528	0.497	0.405	0.524	0.041	0.525	0.522
LoLC	0.603	0.504	0.478	0.452	0.340	0.478	0.044	0.479	0.477
				Mean					
LoLA	0.249	0.218	0.168	0.111	0.001	0.160	0.064	0.162	0.158
LoLB	0.234	0.214	0.203	0.148	0.088	0.180	0.042	0.181	0.179
LoLC	0.230	0.202	0.153	0.113	0.046	0.155	0.045	0.156	0.154

piDAPH3

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.759	0.638	0.579	0.531	0.273	0.581	0.075	0.583	0.579
LoLB	0.870	0.785	0.726	0.649	0.505	0.715	0.079	0.717	0.712
LoLC	0.748	0.651	0.617	0.571	0.415	0.610	0.055	0.611	0.608
				Mean					
LoLA	0.330	0.287	0.229	0.160	0.000	0.216	0.080	0.219	0.214
LoLB	0.352	0.309	0.288	0.203	0.109	0.258	0.064	0.260	0.257
LoLC	0.330	0.289	0.202	0.158	0.039	0.217	0.068	0.218	0.215

Comparison with NCI Diversity

TGD

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.926	0.879	0.803	0.775	0.696	0.818	0.060	0.820	0.816
LoLB	0.894	0.873	0.834	0.782	0.692	0.816	0.062	0.818	0.814
LoLC	0.873	0.840	0.813	0.791	0.741	0.814	0.034	0.815	0.813
				Mean					
LoLA	0.633	0.598	0.563	0.530	0.475	0.564	0.039	0.565	0.563
LoLB	0.587	0.560	0.539	0.505	0.481	0.533	0.029	0.534	0.533
LoLC	0.588	0.551	0.533	0.505	0.481	0.531	0.027	0.532	0.530

Table S4. (continued)

GpiDAPH3

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.708	0.514	0.466	0.420	0.221	0.464	0.068	0.466	0.462
LoLB	0.645	0.543	0.494	0.427	0.335	0.489	0.070	0.491	0.487
LoLC	0.543	0.448	0.415	0.388	0.306	0.419	0.043	0.420	0.418
				Mean					
LoLA	0.245	0.210	0.187	0.095	0.001	0.152	0.071	0.154	0.150
LoLB	0.221	0.185	0.170	0.099	0.040	0.142	0.054	0.143	0.140
LoLC	0.207	0.170	0.105	0.056	0.020	0.112	0.055	0.114	0.111

piDAPH3

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.778	0.625	0.569	0.521	0.000	0.574	0.074	0.576	0.572
LoLB	0.831	0.767	0.713	0.602	0.458	0.687	0.091	0.690	0.685
LoLC	0.728	0.623	0.584	0.548	0.406	0.583	0.055	0.585	0.582
				Mean					
LoLA	0.324	0.261	0.230	0.118	0.000	0.191	0.087	0.193	0.188
LoLB	0.327	0.254	0.218	0.127	0.049	0.191	0.076	0.194	0.189
LoLC	0.307	0.238	0.139	0.076	0.018	0.155	0.079	0.158	0.153

Comparison with MLSMR

MACCS

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.908	0.747	0.711	0.681	0.581	0.716	0.054	0.717	0.714
LoLB	0.911	0.750	0.711	0.683	0.590	0.724	0.061	0.726	0.722
LoLC	0.922	0.759	0.728	0.695	0.612	0.734	0.059	0.736	0.733
				Mean					
LoLA	0.469	0.435	0.403	0.374	0.320	0.403	0.035	0.404	0.402
LoLB	0.480	0.430	0.391	0.362	0.319	0.394	0.039	0.396	0.393
LoLC	0.490	0.410	0.381	0.362	0.312	0.390	0.041	0.391	0.389

TGD

Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
				Maximum					
LoLA	0.928	0.849	0.813	0.769	0.660	0.808	0.061	0.810	0.806
LoLB	0.919	0.877	0.857	0.806	0.700	0.832	0.067	0.834	0.831
LoLC	0.916	0.886	0.857	0.811	0.747	0.849	0.049	0.850	0.847
				Mean					
LoLA	0.694	0.635	0.606	0.514	0.393	0.583	0.074	0.585	0.581
LoLB	0.683	0.670	0.624	0.537	0.519	0.609	0.062	0.611	0.607
LoLC	0.697	0.648	0.612	0.519	0.481	0.593	0.070	0.594	0.591

Table S4. (continued)

GpiDAPH3									
Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
					Maximum				
LoLA	0.657	0.523	0.482	0.430	0.096	0.470	0.071	0.472	0.469
LoLB	0.608	0.547	0.522	0.496	0.387	0.519	0.038	0.520	0.518
LoLC	0.658	0.532	0.493	0.454	0.326	0.494	0.058	0.496	0.493
					Mean				
LoLA	0.294	0.253	0.220	0.091	0.000	0.178	0.088	0.180	0.175
LoLB	0.281	0.246	0.230	0.124	0.051	0.188	0.071	0.190	0.186
LoLC	0.268	0.231	0.139	0.080	0.029	0.152	0.073	0.154	0.150

piDAPH3									
Library	Max	Q3	Median	Q1	Min	Mean	Stdev	U95	L95
					Maximum				
LoLA	0.798	0.665	0.612	0.550	0.000	0.604	0.086	0.606	0.602
LoLB	0.802	0.719	0.688	0.644	0.514	0.679	0.052	0.681	0.678
LoLC	0.764	0.676	0.631	0.584	0.406	0.626	0.065	0.628	0.624
					Mean				
LoLA	0.374	0.303	0.264	0.104	0.000	0.213	0.108	0.216	0.210
LoLB	0.363	0.282	0.236	0.118	0.058	0.206	0.087	0.209	0.204
LoLC	0.342	0.274	0.145	0.079	0.010	0.171	0.096	0.174	0.169

Table S5. Loadings for the six principal components of the property space of drugs, NCI diversity and MLSMR with 3000 compounds selected at random

Principal component	PC1	PC2	PC3	PC4	PC5	PC6
eigenvalue	2.258	1.967	0.692	0.649	0.315	0.120
cumulative eigenvalue (%)	37.629	70.417	81.947	92.757	98.006	100.000
HBA	-0.483	0.195	0.386	-0.626	0.313	0.301
HBD	-0.311	0.399	-0.794	-0.044	0.302	-0.144
RB	-0.513	-0.308	-0.263	-0.099	-0.709	0.246
SlogP	-0.033	-0.650	-0.211	0.182	0.523	0.474
TPSA	-0.348	0.411	0.241	0.723	8.5e-3	0.360
MW	-0.533	-0.341	0.222	0.204	0.185	-0.689