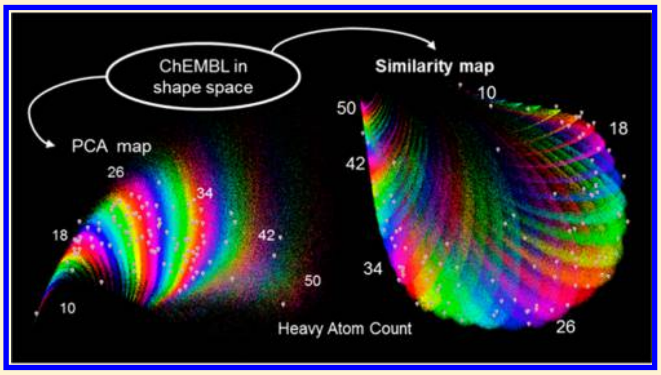
Journal of
CHEMICAL INFORMATION
AND MODELING

Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces

Mahendra Awale and Jean-Louis Reymond*

Department of Chemistry and Biochemistry, National Center of Competence in Research NCCR TransCure, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland

S *Supporting Information*



ABSTRACT: An Internet portal accessible at www.gdb.unibe.ch has been set up to automatically generate color-coded similarity maps of the ChEMBL database in relation to up to two sets of active compounds taken from the enhanced Directory of Useful Decoys (eDUD), a random set of molecules, or up to two sets of user-defined reference molecules. These maps visualize the relationships between the selected compounds and ChEMBL in six different high dimensional chemical spaces, namely MQN (42-D molecular quantum numbers), SMIfp (34-D SMILES fingerprint), APfp (20-D shape fingerprint), Xfp (55-D pharmacophore fingerprint), Sfp (1024-bit substructure fingerprint), and ECfp4 (1024-bit extended connectivity fingerprint). The maps are supplied in form of Java based desktop applications called “similarity mapplets” allowing interactive content browsing and linked to a “Multifingerprint Browser for ChEMBL” (also accessible directly at www.gdb.unibe.ch) to perform nearest neighbor searches. One can obtain six similarity mapplets of ChEMBL relative to random reference compounds, 606 similarity mapplets relative to single eDUD active sets, 30 300 similarity mapplets relative to pairs of eDUD active sets, and any number of similarity mapplets relative to user-defined reference sets to help visualize the structural diversity of compound series in drug optimization projects and their relationship to other known bioactive compounds.

■ INTRODUCTION

The advent of high-throughput experimentation and big data informatics has led to an explosion in the number of molecules under investigation in a typical drug discovery project.[1,2] Furthermore, the evidence of polypharmacology and the necessity to perform multiparameter optimization to mitigate off-target effects requires the constant tracking of the relationship between any focused compound series and all other know bioactive molecules.[3−6] Methods to visualize the structural diversity of bioactive compound series and their relationship to other known molecules have therefore become essential tools to assist drug discovery.[7−38]

Recently we reported Java based desktop applications called “mapplets” which visualize large compound databases in form of color-coded interactive images linking image points to chemical structures and their source database.[39−41] These mapplets are representations of the (PC1, PC2)-plane resulting from principal component analysis (PCA) of the databases placed in the chemical spaces of MQN (molecular quantum numbers, 42 integer value descriptors of molecular structure),[42,43] and SMIfp (SMILES fingerprint, counts of 34 characters frequently used in SMILES).[40] In contrast to other map generation methods recently applied to represent compound databases such as self-organizing maps[27] or generative topographic mapping[32,33] which are computationally intensive, PCA computation can be parallelized and is therefore scalable and applicable to very large databases. We therefore readily obtained interactive PCA maps of a variety of databases from thousands (DrugBank, Fragrances)[44,45] to many millions (ChEMBL, ZINC, PubChem, the generated databases GDB-11, GDB-13, and GDB-17)[46,47] of molecules.

Although rapidly computed, the PCA-maps of the above mapplets only provided global database overviews and were limited to simple fingerprints such as MQN or SMIfp for which the (PC1, PC2)-plane covers a sufficient percentage of data variability (>70%) for visualization. By contrast the practice of drug discovery requires not only global overviews but also targeted representations of chemical space to visualize the local structural diversity of specific compound series considering detailed analyses of the molecules such as those provided by binary substructure (Sfp) and extended connectivity fingerprints (ECfp4).[48,49] Herein we report an Internet portal that addresses this need by generating interactive mapplets representing ChEMBL in the local chemical space of user-defined compound series by similarity mapping. This method consists in computing similarity values to a set of reference molecules, followed by projection of the resulting similarity

Received: April 2, 2015
Published: July 24, 2015

Table 1. Fingerprints Used in This Study

fingerprint	feature perceived	description	ref
MQN	composition	42-dimensional scalar fingerprint, counts 42 MQNs counting atom types, bond types, polar groups and topologies	42, 43
SMIfp	composition	34-dimensional scalar fingerprint, counts 34 characters appearing in the SMILES notation of molecules	40
APfp	shape	20-dimensional scalar fingerprint, each dimension counts the number of atom pairs at one particular topological distance between 1 and 20 bonds, normalized to HAC	58
Xfp	pharmacophore	55-dimensional scalar fingerprint, category extended version of APfp counting the number of category atom pairs at one particular topological distance between 0 and 10 bonds, normalized to the number of category atoms, for categories: hydrophobic atoms, H-bond donor atoms, H-bond acceptor atoms, sp2 hybridized atoms, and HBA/HBD cross-pairs	58
Sfp	substructure	1024-dimensional binary fingerprint, perceives the presence of substructures	48
ECfp4	substructure	1024-dimensional binary fingerprint, perceives the presence of extended connectivity elements up to 4 bonds around each atom	49

space into two dimensions.[25,50−53] The method has for example been used to visualize taxonomic data by similarity mapping of 16S RNA sequences[54,55] and to visualize the structural diversity of small sets of small molecule drugs using the Tanimoto coefficient of 2D MACCS key fingerprints as similarity measure.[12,26]

The similarity mapplet portal produces color-coded interactive similarity maps of the 1.2 million molecules up to 50 heavy atoms in ChEMBL,[56] in relation to up to two active sets taken from the enhanced directory of useful decoys (eDUD),[57] a random set of molecules, or user-defined reference molecules that can be uploaded as SMILES lists. The similarity maps can be obtained from six different high dimensional chemical spaces, namely the above-mentioned MQN and SMIfp, the atom pair fingerprint APfp (20-D shape fingerprint) and its category extended version Xfp (55-D pharmacophore fingerprint) which perceive 3D-features of molecules comparably well to 3D-methods,[58,59] and the well-known binary Sfp (1024-bit substructure fingerprint) and ECfp4 (1024-bit extended connectivity fingerprint) which we recently showed to allow rapid similarity searches (Table 1).[48,49,60] To the best of our knowledge such an automatic interactive map generation tool is unprecedented and should be generally useful to assist drug optimization projects.

RESULTS AND DISCUSSION

Similarity Map Calculation. The calculation of similarity maps is discussed here for the case of using 100 randomly selected molecules as references. In the first step, the property spaces MQN, SMIfp, APfp, Xfp, Sfp, and ECfp4 are transformed into similarity spaces by computing similarities to the reference compounds. The Tanimoto coefficient T_{fp} is used as similarity measure for the binary fingerprints Sfp and ECfp4, while the city-block distance (CBD) derived similarity value $X/(\mathrm{CBD}_{fp} + X)$ is used for the scalar fingerprints MQN, SMIfp, APfp, and Xfp, with X representing the median CBD_{fp} (city-block distance in the chemical space fp) separating compound pairs in ChEMBL (Figure 1A/B).[61] PCA is then performed to produce the similarity map in form of the (PC1, PC2)-plane, which is represented as a 1000 × 1000 pixel image.

For MQN, SMIfp and APfp the variance covered by the random reference similarity map reaches over 70%, which is comparable to the variance covered in the (PC1, PC2)-plane resulting from direct PCA of the parent chemical space. This shows that the definition of a 100-dimensional similarity space does not increase the dimensionality of the parent chemical space. Note that for the case of MQN the variance covered by the first PCs and the resulting 2D-map are comparable when using between 50 and 1000 reference molecules, with 100−200 references appearing as a practical number covering sufficient chemical variability within a reasonable amount of computation and generated data. Choosing reference compounds by clustering rather than at random does not significantly influence the variance covered or the appearance of the similarity maps. The similarity maps also cover a significant variance of the similarity space in the case of the higher dimensional fingerprints Xfp, Sfp, and ECfp4 for which the (PC1, PC2)-plane obtained by direct PCA only covers very low data variance of the original space,[62] showing that the similarity space derived from these high-dimensional fingerprints has reduced dimensionality (Figure 1C/D).

For the six fingerprints used for the similarity mapplets there are almost as many fingerprint bit value combination as ChEMBL molecules, a consequence of the relatively small size of the ChEMBL database. Indeed only a small fraction of fingerprint bins contain multiple molecules in the case of Xfp (5%), MQN and SMIfp (15%), and APfp (20%) (Figure 1E). At the level of the various similarity maps molecules are distributed in pixels following a typical power law distribution with approximately 50% of ChEMBL compounds occupying the 20% most occupied pixels of the maps (Figure 1F). This distribution is comparable to the distribution observed previously in direct PCA maps[39] and reflects a certain degree of superimposition resulting from projection of the high dimensional chemical space into two dimensions, in particular for high density areas of the map where relatively dissimilar compounds partly end up in the same pixel due to a folding effect.

Similarity Mapplet Internet Portal. The similarity mapplet Internet portal is a web interface to generate interactive Java based mapplet applications. The portal allows users to select one of the six chemical spaces and up to two sets of reference molecules taken either from the list of 101 active sets in the eDUD, a random set of molecules, or up to two user-defined list of molecules that can be uploaded as SMILES lists (Figures S1A). The server computes the similarity map (in usually less than 1 h) and sends a link to download the corresponding similarity mapplet to the e-mail address given by the user. One can thus obtain six random reference similarity mapplets, 606 single active set and 30 300 dual active set similarity mapplets of ChEMBL relative eDUD active sets, and any number of additional similarity mapplets with user-defined reference sets. The mapplets are built on the principle of our recently reported Mapplet application,[39] which is a Java application presenting maps of chemical spaces with each pixel color-coded according to various properties; here the highest similarity value in that pixel to any of the reference molecules (Figures S1B). This most similar molecule in each

DOI: 10.1021/acs.jcim.5b00182

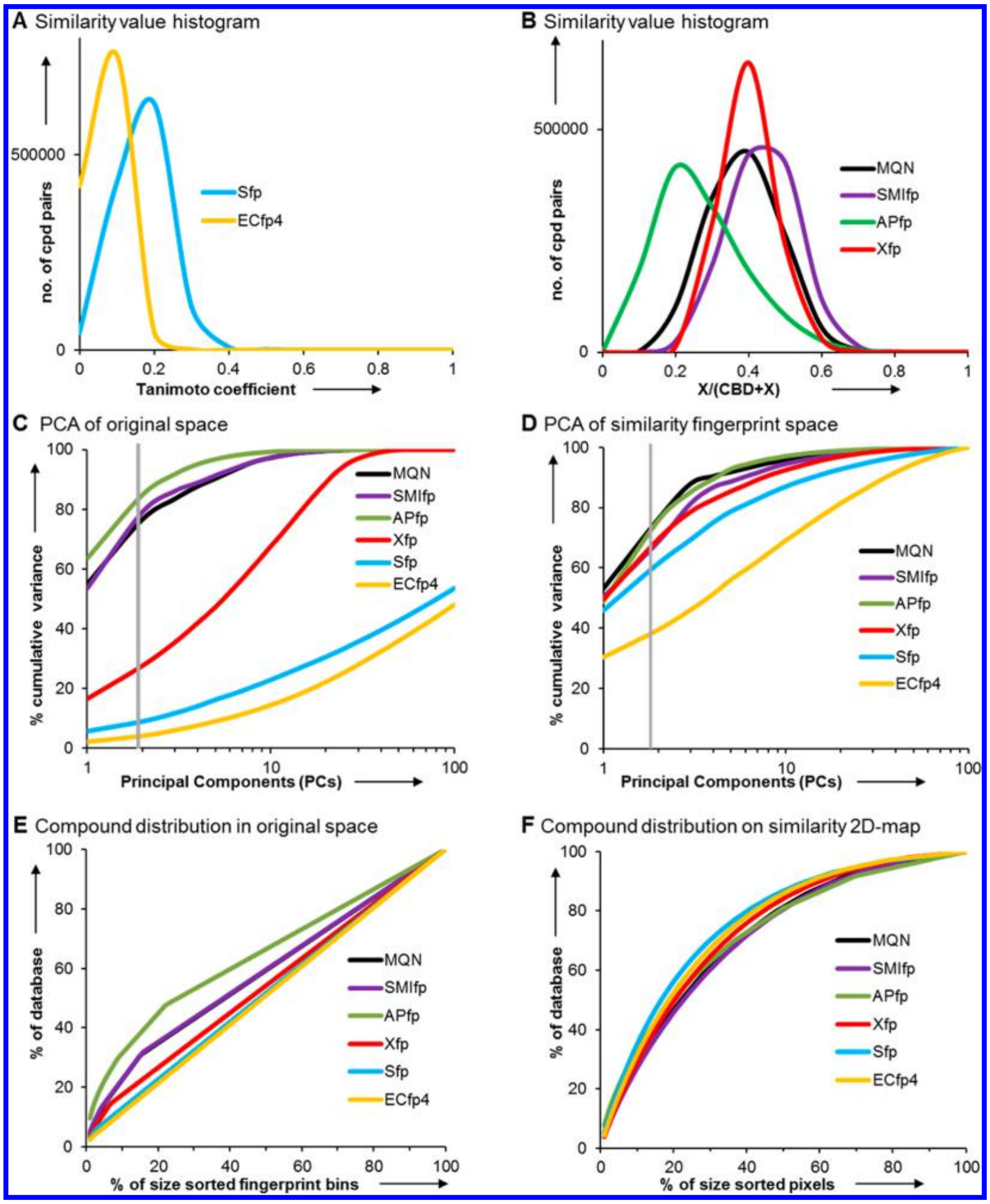


Figure 1. Similarity map generation. **A**. Tanimoto histogram of ChEMBL with Sfp and ECfp4. **B**. $(X/(\mathrm{CBD}_{\mathrm{fp}} + X))$ histogram of ChEMBL for MQN, SMIfp, APfp, and Xfp. X is the median city-block distance $(\mathrm{CBD}_{\mathrm{fp}})$ separating compound pairs of ChEMBL in the corresponding chemical space. **C**. Variance covered by direct PCA of ChEMBL in the different chemical spaces. **D**. Variance covered by similarity PCA of ChEMBL relative to 100 random molecules as references. The cumulative variance covered by increasing number of PCs is shown. The vertical gray line indicates the variance covered by the (PC1,PC2)-plane. **E**. Distribution of ChEMBL molecules in the different fingerprint values (bins) sorted by bin occupancy. **F**. Distribution of ChEMBL molecules in the different pixels of the random reference similarity maps shown in Figure 2B/C, sorted by decreasing occupancy. The 1 224 769 molecules up to 50 heavy atoms in ChEMBL version 18 were considered, and 1.2 million randomly chosen compound pairs were used for the similarity histograms.

pixel appears in a side window when the mouse pointer passes over that pixel and can be linked either to the compound in the ChEMBL website, or to a multifingerprint browser for ChEMBL. This multifingerprint browser is also directly accessible at www.gdb.unibe.ch and allows the user to perform nearest neighbor similarity searches in the entire ChEMBL by CBD in any of the six high-dimensional chemical spaces used for the similarity maps (Figures S1C/D).

Similarity Mapplets with Random Compounds As References. The similarity mapplets obtained using random reference compounds provide a global database overview. The similarity maps of these random reference similarity mapplets can be compared with the maps obtained by direct PCA. In contrast to the direct PCA maps of MQN, SMIfp, and APfp featuring regular stripe features (Figure 2A), the similarity maps in the random reference similarity mapplets have a cometlike shape with reference molecules scattered on the image (Figure 2B). As exemplified for single descriptor cases for each map in Figure 2, the global organization of these similarity maps follows the logic of the parent high dimensional chemical space,

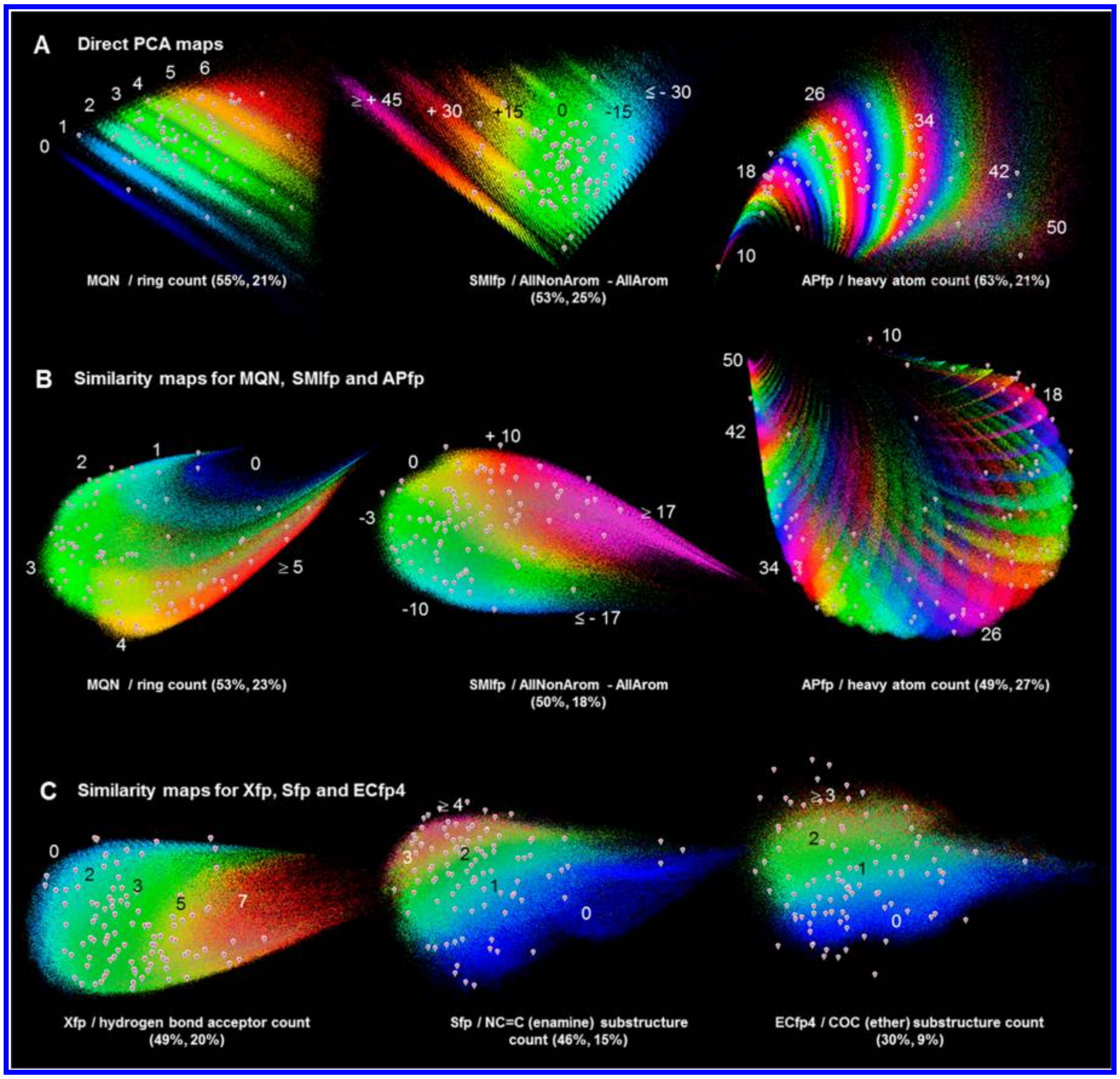


Figure 2. Maps of ChEMBL in the (PC1, PC2)-plane obtained by PCA of the different chemical spaces. The variance covered by PC1 and PC2 is given in parentheses. **A.** Direct PCA of MQN, SMIfp, and APfp spaces. **B.** Similarity maps obtained by PCA of the similarity spaces. **C.** Similarity maps for Xfp, Sfp, and ECfp4. All similarity maps were obtained relative to 100 randomly chosen reference molecules, shown as balloons on the images (reference points are also shown on direct PCA maps for comparison but are not used for calculation). The pixels (1000 × 1000 pixels for each map) are colored according to the average value of the indicated property in the pixel in the HSL color space from low value to high value in the range blue−cyan−green−yellow−red−magenta, with five successive rotations through the color scale in the case of the APfp space maps colored by HAC. The saturation to gray is used to indicate the standard deviation of the pixel average value. The 1 224 769 molecules up to 50 heavy atoms in ChEMBL version 18 were used.

with categorized areas resembling those in the direct PCA maps but with a very different graphical layout spreading the central area of the direct PCA map. Although this is not very visible in the figure, the similarity maps are more compact than the corresponding direct PCA images because they do not have diffuse peripheral pixels.

The similarity maps of the random reference similarity mapplets for Xfp, Sfp, and ECfp4 are also cometlike but with somewhat more diffuse edges. Molecules in these maps are distributed according to the counts of structural features particularly well perceived by the fingerprints, such as counts of HBA for Xfp, and counts of specific substructures for Sfp and ECfp4 (Figure 2C). In the case of the Sfp and ECfp4 map a detailed study shows that the overall map organization is independent of the choice of the reference molecules (Figure S3/S4).

Activity Focused Similarity Mapplets. Similarity mapplets using bioactive compounds against the same target as references provide interesting targeted insights into compound series. The Similarity Mapplet web portal offers mapplets for each of the 101 different active sets from the eDUD database, considering all molecules in each active set (between 50 and 500 molecules) as references.[57] The similarity maps in these 606 similarity mapplets cover over 70% data variability except for the maps derived from ECfp4 for which variance coverage by the first two PCs is in the range 50%−70% (Figure S4). The larger coverage of variance in activity focused similarity maps compared to random reference similarity maps reflects the limited structural diversity of the active sets used as references.

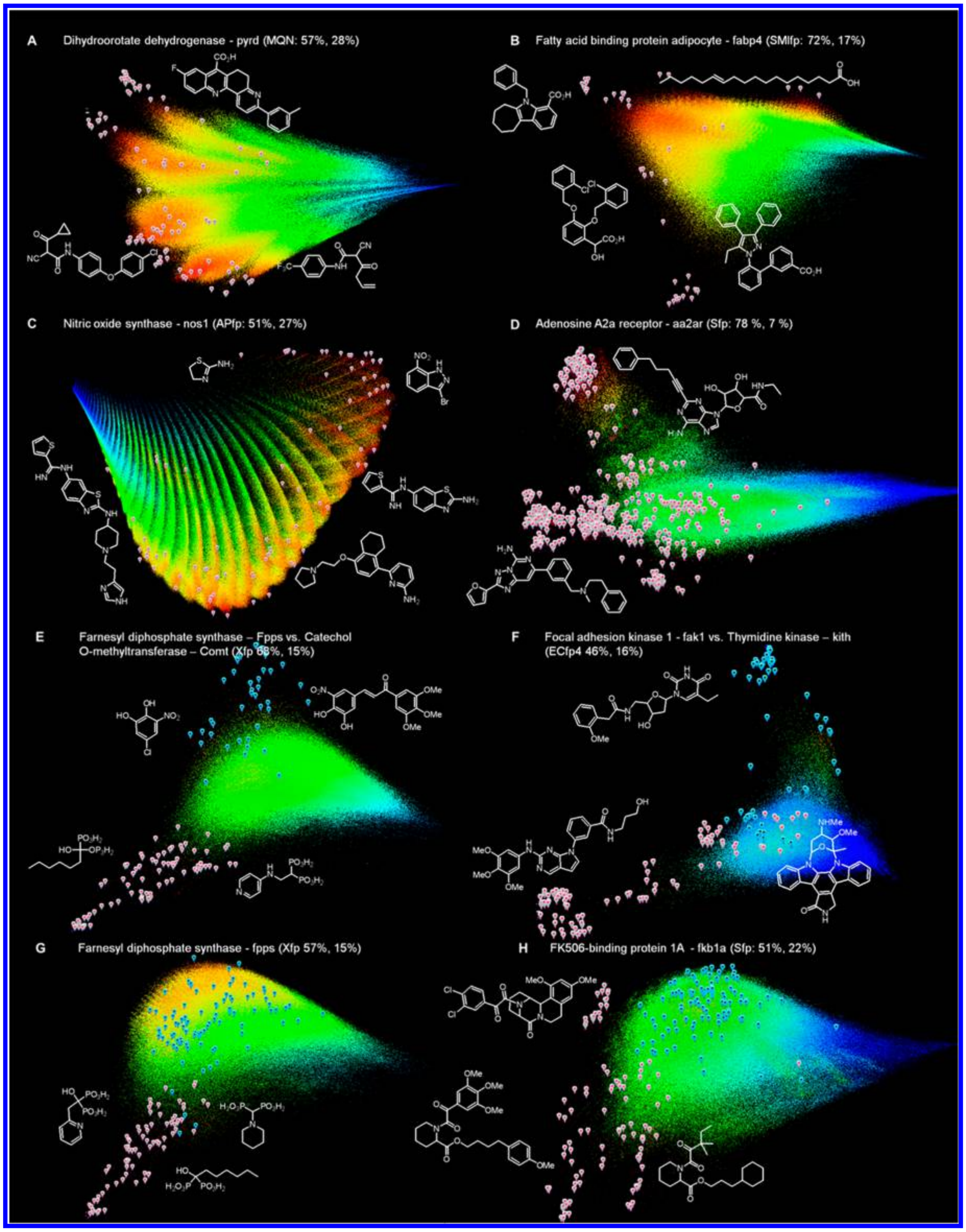


Figure 3. Similarity maps in activity focused ChEMBL similarity mapplets. **A–D**. Inhibitors against one eDUD target (all reported actives) as references. **E–F.** Inhibitors against two different eDUD targets (all reported actives) as references. **G–H**. Inhibitors against one target (up to 100 actives) and the corresponding decoys (100 decoys) from eDUD. The variance covered by PC1 and PC2 is given in parentheses. The pixels are colored according to the best similarity score of ChEMBL compounds in the pixel to any of the inhibitors in the set in scale blue ≤ 0.2, cyan = 0.36, green = 0.52, yellow = 0.68, red = 0.84, magenta = 1. References are marked as white or blue balloons. Structures of selected references are shown on the map close to where they can be found. The maps can be inspected with the similarity Mapplets given as examples at the similarity mapplet web interface accessible at www.gdb.unibe.ch.

Indeed similarity values to a set of similar molecules are correlated, implying that the corresponding similarity space has reduced dimensionality compared to a similarity space produced with random compounds as references.

In contrast to random reference similarity maps where reference molecules are scattered on the image, the reference active sets in the activity focused similarity maps are organized in different similarity clusters generally located on the circumference of the images (Figure 3A–D). ChEMBL molecules are distributed according to their similarities to these reference clusters with the most similar molecules appearing closest to each reference as indicated by the color scale from blue (low similarity) to magenta (identity). Reference molecules are not always surrounded by magenta or red pixels (indicating high similarity) due to the fact that ChEMBL does not always contain compounds with high similarity values to active reference molecules, in particular for Sfp and ECfp4 where high similarity values are rather rare (Figures 1A and 3D). As for random reference similarity maps the activity focused maps follow the intrinsic logic of each fingerprint, comprising the number of rings for MQN, the number of aromatic atoms for SMIfp, molecule size for APfp and specific substructural similarity for Sfp and ECfp4. The organization of the maps is illustrated by examples of active set molecules shown close to their respective cluster on the images in Figure 3 and can be appreciated in more detail by inspecting the interactive Mapplets of each image available for download as examples at the similarity mapplet Internet portal.

Similarity Mapplets with Pairs of Active Sets As References. Interesting similarity mapplets are also obtained when using pairs of bioactive sets as references. These dual set similarity mapplets provide an insight into the specific structural differences between the sets. For example the Xfp similarity mapplet generated from farnesyl diphosphate synthase inhibitors and catechol *O*-methyl transferase inhibitors nicely separates the acyclic bis-phophonates specific for the first target versus the nitrocatechols specific for the second target (Figure 3E). In the case of focal adhesion kinase inhibitors versus thymidine kinase inhibitors, the ECfp4 similarity mapplet separates the extended adenine analogs inhibiting the first target from the thymidine analogs specific for the second, while the notoriously non selective kinase inhibitor staurosporine appear in the central part of the image (Figure 3F). Any number of such dual set similarity mapplets can be generated by the similarity mapplet Web site by choosing from the list of 101 eDUD active sets or by uploading up to two user-defined SMILES lists.

Similarity Mapplets with Active and Decoys As References. Similarity mapplets generated by combining one of the eDUD active sets with its corresponding decoys as references provide a graphical illustration of the separation between actives and decoys that is possible when performing enrichment studies with the various fingerprints. The reference active molecules are generally grouped on the map periphery, while decoy molecules are scattered on the central high density area of the map. Such active + decoys similarity mapplets are exemplified here for farnesyl diphosphate synthase inhibitors and Xfp (Figure 3G) and for FK506 binding protein 1A and Sfp (Figure 3H).

Accuracy of the Similarity Maps. The various color coded images and the grouping of molecule types in Figures 2 and 3 illustrate that the global organization of the similarity maps used for the similarity mapplet is comparable to their parent high dimensional chemical space. A closer analysis of map quality is presented here in form of a recovery study measuring how well nearest neighbor relationships are preserved between the original chemical space (measured using the city-block distance CBD) and the 2D-maps (measured in euclidean distance in the 2D-plane), thus focusing on the issue of map accuracy at close range. For each of 150 000 randomly selected compounds in ChEMBL the ROC (receiver operator characteristic) curve is computed for recovering its 15 nearest neighbors (among these 150 000 ChEMBL compounds) in the original high dimensional chemical space by a proximity search on each of the 2D-maps presented in Figures 2 and 3. The results are analyzed for each pixel of the maps in terms of average area under the curve (AUC) and average enrichment factor at 0.1% ($EF_{0.1}$) (Figure S5).

For a 2D-map representing a high dimensional space with high fidelity one would expect high AUC (>90%) and high $EF_{0.1}$ (>900) for every pixel of the map. AUC values are indeed very high in the case of MQN, SMIfp, and APfp for both direct PCA and similarity maps, with the majority of pixels having average AUC values larger than 90% (Figure S5A/E). AUC values are independent of pixel density for MQN and SMIfp, but increase with pixel density in the case of APfp which gave the highest AUC values in both direct PCA maps and similarity maps (Figure S5B/F). By contrast Sfp and ECfp4 similarity maps are far from ideal in terms of AUC, with values distributed across the entire range (cyan and yellow lines in Figure S5A/E). In these two cases pixels with high AUC values have low compound density, indicating better map quality in the less densely populated areas of the maps (cyan and yellow lines in Figure S5B/F). For $EF_{0.1}$ results are generally poorer. The majority of pixels have low $EF_{0.1}$ values and only relatively few pixels have very high $EF_{0.1}$ values (Figure S5C/G). As for AUC values the better performance in terms of $EF_{0.1}$ is obtained with the APfp derived maps and the worst performance with the ECfp4 maps. Most importantly, high $EF_{0.1}$ values occur predominantly in low density pixels (Figure S5D/H).

Visual inspection of the maps in Figures 2 and 3 color-coded by AUC, $EF_{0.1}$, and pixel density values shows that the low density, high fidelity regions of the maps are generally located at the map periphery (Figures S6–S11). For the case of the activity focused similarity maps these high fidelity regions correspond to the position of the reference active molecules. Thus, activity focused similarity maps provide not only a graphical layout distributing active compounds according to their fingerprint similarities, but also a reliable insight into the availability of close analogs in ChEMBL. In fact the activity focused similarity maps use the largest fraction of the 2D-map for a high fidelity representation of ChEMBL in the specific area of the chemical space relevant to the particular set of active molecules, leaving the larger but less relevant part of the database in a dense, poorly resolved but relatively compact portion of the map.

Note that the partial inaccuracy of the similarity maps can be partly compensated for during browsing with the different similarity mapplets by linking to the multifingerprint browser for ChEMBL, which allows one to identify the nearest neighbors of any selected molecule in the original high-dimensional chemical space.

■ CONCLUSION

In summary the interactive similarity mapplets produced by the similarity mapplet web portal offer an efficient and intuitive visualization of the relationship between ChEMBL and various sets of user-defined compounds in a variety of contexts, providing global overviews when using random compounds as reference, and focused insights into the structural diversity of compounds series when using one or two active sets or an active set and its decoys as reference molecules. The similarity mapplets are linked to a multifingerprint browser for ChEMBL, which is also directly accessible at www.gdb.unibe.ch, to perform nearest neighbor searches in any of the six high dimensional chemical spaces, allowing to search for closest analogs of any selected molecule. While many of the structural differences between different compound classes may seem obvious to a specialist in one particular target area, the similarity mapplets help to communicate this knowledge to a broader audience, or to familiarize oneself with a new series of compounds and rapidly understand the relationship between any given compound series and the ChEMBL database in an intuitive interactive manner. The Similarity Mapplet web portal should be generally applicable to visualize drug optimization projects and their relationship to other known bioactives.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00182.

> Methods for similarity maps calculation. Layout of the similarity mapplet web portal (Figure S1). Similarity maps of ChEMBL in Sfp and ECfp4 space color-coded according to various substructure counts (Figure S3/S4). Variance covered by eDUD focused similarity maps (Figure S4). Quality assessment of similarity maps (Figures S5−S11) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported financially by the University of Berne, the Swiss National Science Foundation and the NCCR TransCure.

■ REFERENCES

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

(2) Duffy, B. C.; Zhu, L.; Decornez, H.; Kitchen, D. B. Early Phase Drug Discovery: Cheminformatics and Computational Techniques in Identifying Lead Series. *Bioorg. Med. Chem.* **2012**, *20*, 5324−5342.

(3) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874−7887.

(4) Perez-Nueno, V. I.; Ritchie, D. W. Identifying and Characterizing Promiscuous Targets: Implications for Virtual Screening. *Expert Opin. Drug Discovery* **2012**, *7*, 1−17.

(5) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X. P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **2012**, *492*, 215−220.

(6) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Multi-Objective Molecular De Novo Design by Adaptive Fragment Prioritization. *Angew. Chem., Int. Ed.* **2014**, *53*, 4244−4248.

(7) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 339−353.

(8) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(9) Takahashi, Y.; Konji, M.; Fujishima, S. Molspace: A Computer Desktop Tool for Visualization of Massive Molecular Data. *J. Mol. Graphics Modell.* **2003**, *21*, 333−339.

(10) Haggarty, S. J.; Clemons Pa Fau - Wong, J. C.; Wong Jc Fau - Schreiber, S. L.; Schreiber, S. L. Mapping Chemical Space Using Molecular Descriptors and Chemical Genetics: Deacetylase Inhibitors. *Comb. Chem. High Throughput Screening* **2004**, *7*, 669−676.

(11) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225−233.

(12) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-Based Data-Fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chem. Biol. Drug Des.* **2007**, *70*, 393−412.

(13) Medina-Franco, J. L.; Martinez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322−333.

(14) Medina-Franco, J. L.; Martinez-Mayorga, K.; Bender, A.; Marin, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477−491.

(15) Rosen, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. Novel Chemical Space Exploration Via Natural Products. *J. Med. Chem.* **2009**, *52*, 1953−1962.

(16) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010−1024.

(17) Ivanenkov, Y. A.; Savchuk, N. P.; Ekins, S.; Balakin, K. V. Computational Mapping Tools for Drug Discovery. *Drug Discovery Today* **2009**, *14*, 767−775.

(18) Akella, L. B.; DeCaprio, D. Cheminformatics Approaches to Analyze Diversity in Compound Screening Libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325−330.

(19) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

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

(21) Le Guilloux, V.; Colliandre, L.; Bourg, S. p.; Guénegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. Visual Characterization and Diversity Quantification of Chemical Libraries: 1. Creation of Delimited Reference Chemical Subspaces. *J. Chem. Inf. Model.* **2011**, *51*, 1762−1774.

(22) Reutlinger, M.; Guba, W.; Martin, R. E.; Alanine, A. I.; Hoffmann, T.; Klenner, A.; Hiss, J. A.; Schneider, P.; Schneider, G. Neighborhood-Preserving Visualization of Adaptive Structure-Activity Landscapes: Application to Drug Discovery. *Angew. Chem., Int. Ed.* **2011**, *50*, 11633−11636.

(23) Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; López-Vallejo, F. Visualization of Molecular Fingerprints. *J. Chem. Inf. Model.* **2011**, *51*, 1552−1563.

(24) Medina-Franco, J. L.; Yongye, A. B.; Pérez-Villanueva, J.; Houghten, R. A.; Martínez-Mayorga, K. Multitarget Structure−Activity Relationships Characterized by Activity-Difference Maps and Consensus Similarity Measure. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.

(25) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. *Methods Mol. Biol.* **2011**, *672*, 39−100.

(26) Yoo, J.; Medina-Franco, J. Chemoinformatic Approaches for Inhibitors of DNA Methyltransferases: Comprehensive Characterization of Screening Libraries. *Comput. Mol. Biosci.* **2011**, *1*, 7−16.

(27) Digles, D.; Ecker, G. F. Self-Organizing Maps for in Silico Screening and Data Visualization. *Mol. Inf.* **2011**, *30*, 838−846.

(28) Gutlein, M.; Karwath, A.; Kramer, S. Ches-Mapper - Chemical Space Mapping and Visualization in 3D. *J. Cheminf.* **2012**, *4*, 7.

(29) Ertl, P.; Rohde, B. The Molecule Cloud - Compact Visualization of Large Collections of Molecules. *J. Cheminf.* **2012**, *4*, 12.

(30) Lachance, H.; Wetzel, S.; Kumar, K.; Waldmann, H. Charting, Navigating, and Populating Natural Product Chemical Space for Drug Discovery. *J. Med. Chem.* **2012**, *55*, 5989−6001.

(31) Medina-Franco, J. L.; Aguayo-Ortiz, R. Progress in the Visualization and Mining of Chemical and Target Spaces. *Mol. Inf.* **2013**, *32*, 942−953.

(32) Deng, Z.-L.; Du, C.-X.; Li, X.; Hu, B.; Kuang, Z.-K.; Wang, R.; Feng, S.-Y.; Zhang, H.-Y.; Kong, D.-X. Exploring the Biologically Relevant Chemical Space for Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53*, 2820−2828.

(33) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2014**, *55*, 84−94.

(34) Hoksza, D.; Skoda, P.; Vorsilak, M.; Svozil, D. Molpher: A Software Framework for Systematic Chemical Space Exploration. *J. Cheminf.* **2014**, *6*, 7.

(35) Miyao, T.; Reker, D.; Schneider, P.; Funatsu, K.; Schneider, G. Chemography of Natural Product Space. *Planta Med.* **2015**, *81*, 429.

(36) Rodrigues, T.; Hauser, N.; Reker, D.; Reutlinger, M.; Wunderlin, T.; Hamon, J.; Koch, G.; Schneider, G. Multidimensional De Novo Design Reveals 5-HT2b Receptor-Selective Ligands. *Angew. Chem., Int. Ed.* **2015**, *54*, 1551−1555.

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

(38) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. Datawarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460−473.

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

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

(41) Reymond, J. L.; Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, *3*, 649−657.

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

(43) van Deursen, R.; Blum, L. C.; Reymond, J. L. A Searchable Map of Pubchem. *J. Chem. Inf. Model.* **2010**, *50*, 1924−1934.

(44) Awale, M.; Reymond, J. L. Cluster Analysis of the Drugbank Chemical Space Using Molecular Quantum Numbers. *Bioorg. Med. Chem.* **2012**, *20*, 5372−5378.

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

(46) Blum, L. C.; van Deursen, R.; Reymond, J. L. Visualisation and Subsets of the Chemical Universe Database GDB-13 for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637−647.

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

(48) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Model.* **1992**, *32*, 515−521.

(49) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(50) Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular Similarity Matrices and Quantitative Structure-Activity Relationships: A Case Study with Methodological Implications. *J. Med. Chem.* **1995**, *38*, 629−635.

(51) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from Seal Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(52) Klein, C.; Kaiser, D.; Kopp, S.; Chiba, P.; Ecker, G. F. Similarity Based SAR (SIBAR) as Tool for Early Adme Profiling. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 785−793.

(53) Raghavendra, A. S.; Maggiora, G. M. Molecular Basis Sets - a General Similarity-Based Approach for Representing Chemical Spaces. *J. Chem. Inf. Model.* **2007**, *47*, 1328−1240.

(54) Garrity, G. M.; Lilburn, T. G. Mapping Taxonomic Space: An Overview of the Road Map to the Second Edition of Bergey's Manual of Systematic Bacteriology. *WFCC Newsl.* **2002**, *35*, 5−15.

(55) Lilburn, T. G.; Garrity, G. M. Exploring Prokaryotic Taxonomy. *Int. J. Syst. Evol. Microbiol.* **2004**, *54*, 7−13.

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

(57) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(58) Awale, M.; Reymond, J. L. Atom Pair 2D-Fingerprints Perceive 3D-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of Zinc and GDB-17. *J. Chem. Inf. Model.* **2014**, *54*, 1892−1897.

(59) Awale, M.; Jin, X.; Reymond, J. L. Stereoselective Virtual Screening of the Zinc Database Using Atom Pair 3D-Fingerprints. *J. Cheminf.* **2015**, *7*, 3.

(60) Awale, M.; Reymond, J. L. A Multi-Fingerprint Browser for the Zinc Database. *Nucleic Acids Res.* **2014**, *42*, W234−W239.

(61) Khalifa, A. A.; Haranczyk, M.; Holliday, J. Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection. *J. Chem. Inf. Model.* **2009**, *49*, 1193−1201.

(62) Martin, E.; Cao, E. Euclidean Chemical Spaces from Molecular Fingerprints: Hamming Distance and Hempel's Ravens. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 387−395.