Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Investigating Pharmacological Similarity by Charting Chemical Space

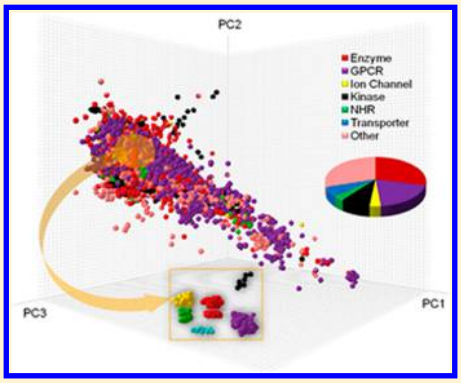
Rosa Buonfiglio,[*,†] Ola Engkvist,[†] Péter Várkonyi,[†] Astrid Henz,[‡] Elisabet Vikeved,[‡] Anders Backlund,[‡] and Thierry Kogej[†]

[†]Chemistry Innovation Centre, Discovery Sciences, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden

[‡]Division of Pharmacognosy, Department of Medicinal Chemistry, Uppsala University, BMC box 574, S-751 23 Uppsala, Sweden

Ⓢ *Supporting Information*

ABSTRACT: In this study, biologically relevant areas of the chemical space were analyzed using ChemGPS-NP. This application enables comparing groups of ligands within a multidimensional space based on principle components derived from physicochemical descriptors. Also, 3D visualization of the ChemGPS-NP global map can be used to conveniently evaluate bioactive compound similarity and visually distinguish between different types or groups of compounds. To further establish ChemGPS-NP as a method to accurately represent the chemical space, a comparison with structure-based fingerprint has been performed. Interesting complementarities between the two descriptions of molecules were observed. It has been shown that the accuracy of describing molecules with physicochemical descriptors like in ChemGPS-NP is similar to the accuracy of structural fingerprints in retrieving bioactive molecules. Lastly, pharmacological similarity of structurally diverse compounds has been investigated in ChemGPS-NP space. These results further strengthen the case of using ChemGPS-NP as a tool to explore and visualize chemical space.



■ INTRODUCTION

At an early stage, one of the main objectives of a drug discovery campaign is the identification of a novel, potent, and selective molecule modulating a therapeutically relevant target—often a protein. In the case that the target protein is known, this is in many cases still achieved by screening a (usually) large number of compounds or a smaller number of fragments in a biochemical assay directly related to the target of interest. Alternatively, in a cell- or organism-based phenotypic screening campaign, many proteins are simultaneously screened by a set of compounds, and the read-out consists in detecting drug candidate effect on the disease-associated phenotype.[1,2] While the identification of responsible proteins during the target deconvolution process[3] is usually time-consuming and tedious, such phenotypic screenings could simultaneously lead to findings of disease relevant target proteins and the active ligands that modulate them. To facilitate the protein identification, compounds with known activity, so-called *tool compounds*, are often desired as these are specifically designed to modulate isolated targets or signaling pathways. In this respect, similarity-based methods may serve to search for new drug candidates in the neighborhood of such tool compounds. These approaches often compare molecules via molecular fingerprints which are mainly based on specific sets of structural or pharmacophoric features. While describing structural features of the compounds, these approaches may fail in associating structurally diverse compounds with similar biological profiles. For this reason, analyses based on fingerprint comparison might be complemented through visualization tools which provide intuitive representation of molecular similarity. In general, visualization techniques are required to obtain information from large quantities of data.[4] For instance, they represent useful tools to project large compound collections in a low dimensional space, amenable to visual inspection and intuitive analysis by the human brain.[5] Molecular representation and data reduction techniques are two major factors that have strong influence on the interpretability of visual representations of chemical space.[6,7] Two common visualization techniques are Principal Component Analysis (PCA) and Kohonen networks. Other techniques to explore structure−activity relationships, conduct structural analysis and further applications in drug discovery, are extensively reported in the literature.[4,5,8,9]

In this context, ChemGPS-NP might be helpful for searching for new drug candidates in the neighborhood of known bioactive molecules as (i) it provides an additional way to capture compound similarity by comparing them on the basis of their physicochemical properties in contrast to the structure-based fingerprint and (ii) it is intrinsically constructed to provide a three-dimensional visualization of the absolute location of compounds or group of compounds within an eight-dimensional chemical space.[10−12] Also, this tool is developed and tuned for handling broad chemical diversity encountered in natural products.

Received: June 11, 2015
Published: October 20, 2015

DOI: 10.1021/acs.jcim.5b00375

This model is based on PCA of a set of 35 molecular descriptors calculated on a structurally diverse set of 1779 natural products. The PCA process allows a reduction of the dimensionality to eight principal components (PCs) described in Table 1. Briefly, PC1 correlates primarily to size, shape, and polarizability, PC2, to aromaticity, PC3, to lipophilicity, polarity, and H-bond capacity, and PC4, to flexibility and rigidity. These molecular descriptors and the derived PCs are found to efficiently describe the molecular recognition between a ligand and its target protein.

Table 1. Summary of the Most Important Features in the Eight Dimensions of ChemGPS-NP, Based on Their Loading Scores in the Different Principal Components[a]

PC	physicochemical features
1	size, shape, polarizability
2	aromaticity, conjugation related properties
3	lipophilicity, polarity, H-bond donor capacity
4	flexibility, rigidity
5	electronegativity, number of nitrogens, halogens and amides
6	number of rings, rotational bonds, amids and hydroxyl groups
7	number of double-bonds, oxygen and nitrogen atoms
8	aromatic and aliphatic hydroxyl groups, molecular saturation, lai

[a] For a full list, please refer to the paper by Larsson et al., 2007.[10]

ChemGPS-NP was developed to explore regions of chemical space populated by natural products, and thus from an evolutionary perspective most likely to enclose compounds exhibiting biological activities. The selection of understandable physicochemical properties and the PCA based reduction to eight dimensions enable to easily chart and navigate this chemical property space. This tool can not only be employed to visualize ligands or groups of ligands within the physicochemical property space, in addition the distance over all eight dimensions between compounds in ChemGPS-NP can be exploited as a measure of their similarity. As a corollary, molecules close in the ChemGPS-NP space present similar physicochemical profiles.

To date, several different applications of ChemGPS-NP have been reported in literature. For instance, Rosén et al. have investigated known anticancer agents and differentiated between diverse mechanisms of action according to the positioning in distinct areas of the chemical space. They also suggested the use of ChemGPS-NP to predict the anticancer mechanism of action of novel cytotoxic compounds.[13] Similarly, Lee and co-workers have predicted topoisomerase II inhibitor activity of a series of phenanthrene derivatives by analyzing their positioning in the ChemGPS-NP space.[14] These works emphasize the potential of using this tool in predicting the pharmacological profile by exploiting the biological activities of the surrounding molecules. On the other hand, ChemGPS-NP has been used to compare the coverage of biologically relevant chemical space by bioactive compounds, drugs and natural products.[15] A similar study has been performed on sets of complex molecular structures demonstrating inherent differences between natural products of marine or terrestrial origin, and synthetic compounds.[16] ChemGPS-NP has been also recently reported in literature, in combination with Lipinski's rule-of-five[17] and chemical clustering for prioritization of trypanocidal marine derived lead compounds.[18]

In the present study, a systematic analysis was undertaken to assess the performance of ChemGPS-NP in retrieving the bioactivity of candidate compounds by comparison with reference sets from the ChEMBL collection.[19] This will contribute in demonstrating the efficiency of ChemGPS-NP with respect to predicting compound activities for a large range of biological profiles. From another perspective, we studied the relationship between target proteins by exploiting chemical similarity of their modulators. Previously, Keiser et al. addressed the same question with the Similarity Ensemble Approach (SEA).[20] The distinctive features of the current study reside in the usefulness of the ChemGPS-NP visualization to prospect for the bioactive chemical space, as well as the combination of the qualitative evaluation of the positioning within the chemical space and the quantitative validation through similarity calculation.

In order to assess the potential of ChemGPS-NP, a comparative study with one of the most commonly used structure-based fingerprint methods (ECFP_4) has been done.[21] Strengths and weaknesses of both methods in finding biological correlations based on similarity of physicochemical properties or chemical features are further discussed. Eventually, a case study with compound sets from GoStar database is presented, aimed at investigating pharmacological similarity of structurally diverse molecules in ChemGPS-NP.[22]

■ METHODS

Data Set. In the present study, the ChEMBL screening collection[19] was selected as the source of bioactive compounds. A cutoff of 10 μM in activity was used to define compounds as active. In case of the molecules having multiple activity values on the same target, the measurement corresponding to the highest activity was selected. In this procedure, assay annotation was disregarded due to the large number of compounds and activity data. For reasons of consistency, we merely considered modulators of human targets of which the Entrez Gene IDentifiers (EGID) were available. Human EGID information was extracted from NCBI database[23] and an internally curated target class annotation.[24] In addition, only targets associated with at least 25 active compounds were considered. Based on this latter criterion, 783 targets were discarded, enabling to set the stage for a robust validation study by preventing low confidence observations for targets consisting of only a single or few active compounds. Additionally, compounds with less than 10 carbon atoms (both aromatic and aliphatic) were filtered out in order to focus on bioactive compounds spanning lead-like, drug-like and larger molecules and simultaneously excluding structural fragments also included in the ChEMBL library. This atom count cutoff was employed in congruence with published AstraZeneca filters.[25] As a final result, 909 targets were included in the study. The distribution of data set size (Figure 1) shows that most of the targets include a few hundreds of active compounds, whereas 11% have more than a thousand molecules (a full list of targets is reported in Table S1). The target classes of the bioactive molecules are here represented as "Enzymes" such as hydrolases, lipases, isomerases, "G-Protein Coupled Receptors" (GPCR), "Ion Channels", "Kinases", "Nuclear Hormon Receptors" (NHR), and "Transporters", along with an unspecified class (referred here to as "Other") including the remaining targets lacking target class annotation. The distribution of the 909 targets across the target classes is reported in Table 2.

The overall sets consist of 514 257 bioactivity data points. Not surprisingly, overlap between some targets was observed, as some molecules are reported to be active on different targets,

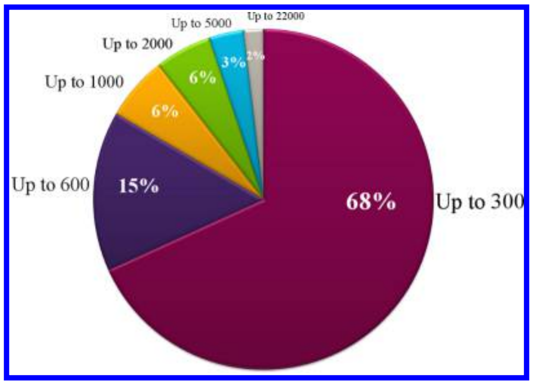


Figure 1. Distribution of the number of active compounds retrieved for the 909 targets.

thus exhibiting polypharmacology. In this study, we decided to keep these multiple-acting modulators along with single-target compounds, because they provide a realistic description of the biologically relevant areas in the chemical space. Disregarding the multiple bioactivity data of some compounds, the final 909 targets include 296 793 unique structures. Details about the distribution of the targets, including associated target classes and number of compounds, are described in Table 2.

Retrieving ChemGPS-NP Coordinates. The basic principles of the ChemGPS-NP space that were introduced above have been extensively described in the literature.[10−12] Hereafter follows a brief description of the process by which we retrieved the ChemGPS-NP coordinates.

As a first step, molecular structures of the 296793 compiled ChEMBL compounds, expressed as SMILES, were standardized.[26] In particular, salts were removed, tautomers were consistently selected and aromaticity was assigned. The 35 physicochemical descriptors forming the basis for ChemGPS-NP were calculated with Dragon 6.0,[27] and a multivariate prediction was carried out by using SIMCA P+ 11.5 software,[28] to derive the eight ChemGPS-NP coordinates which define the position of the molecules in the chemical space. Prior to PCA, all data were centered and scaled to unit variance.

Similarity Metrics in ChemGPS-NP. Compounds having similar physicochemical properties reside closely in the chemical property space described by the ChemGPS-NP coordinates. Thus, a metrics which defines the closeness of compounds in ChemGPS-NP could quantitatively provide a measure of molecular similarity. In this respect, different methods have been suggested. Two examples are the Tanimoto Index (TI) based on the ChemGPS coordinates used as fingerprint[29] and the distance in activity-centered chemical space (DACCS) approach.[30] The latter consists of the calculation of a Euclidean-like distance to produce a distance-based ranking of database compounds as a measure of similarity.

In this study, the Euclidean Distance (ED) based on the eight ChemGPS-NP coordinates has been used to measure the relative distance between molecules:

$$\mathrm{ED} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

where n is equal to 8, and p_i and q_i are the coordinates for molecule p and q, respectively.

Low Euclidean distances reveal proximity between molecules in the ChemGPS-NP map. However, such compounds do not necessarily correspond to similar chemical structures. In that respect, some examples of ChEMBL compound pairs close in ChemGPS-NP (Euclidean distance below 1.00) and, at the same time, structurally diverse according to ECFP_4 fingerprints, are reported in Figure 2. This trend is due to the fact that the closeness in ChemGPS-NP map depends upon similar 2D physicochemical descriptors. Nevertheless, the same descriptors might be attributable to diverse chemical features. Simply, structural features are implicitly taken into account in ChemGPS-NP similarity search, in contrast to fingerprint methods.

Example of ChemGPS-NP Space Illustration: ChEMBL Target Classes Mapping. In Figures 3 and 4, we show the 514 257 bioactive compounds from ChEMBL, colored according to the target class annotation. For the sake of clarity, 3D plots based on the three first coordinates, or 2D plots with the combination of the first four PCs are reported, as they cover 77% of the data variance.[10] The 3D plot reveals a dense region between negative and low positive values across PC1, PC2, and PC3 axes, and few molecules in sparse islands. The most populated section is occupied by molecules following Lipinski's rule-of-five[17] and could then be referred to as the drug-like region. The highly similar distribution of the bioactive compounds in this zone confirms that drug-likeness filters are generally followed. Nevertheless, bioactive compounds populate also other regions of the ChemGPS-NP space characterized by large molecular weight (MW), higher polarity, and flexibility. Some examples are reported in Table S2.[31−33] The 2D plots corresponding to the projection of the first four dimensions, PC2 vs PC1 and PC4 vs PC3, represent an alternative way to analyze the distribution of the entire collection (Figure 4). For instance, coverage of the Enzyme and GPCR groups is largely comparable, with the exception

Table 2. Distribution of Targets and Compounds Across the Different Target Classes

	number of targets	number of compounds	number of unique compounds
Enzyme	298	148875	104658
GPCR	165	100275	73834
Ion Channel	52	16664	12038
Kinase	176	44552	30756
NHR	30	16650	11801
Transporter	36	32684	29330
Other	152	154557	101718
total	**909**	**514257** (296793 unique componds)	364135[a]

[a]The sum from non-redundant compounds within each target class diverges from the overall unique compounds reported in bracket, due to the presence of compounds with multiple activities for targets from different classes.

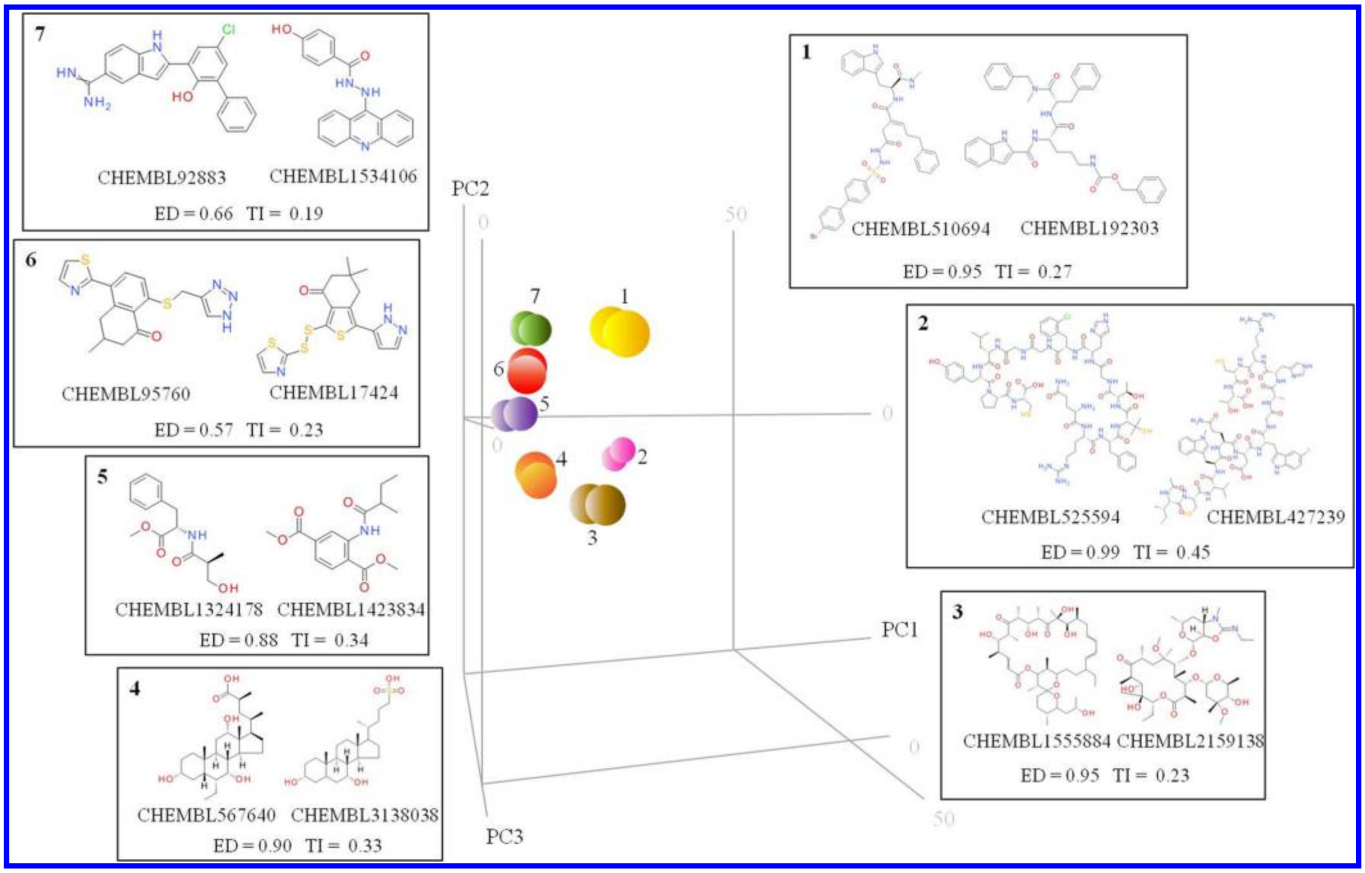


Figure 2. 3D score plot of seven ChEMBL compound pairs residing closely in the ChemGPS-NP map. The first three principal components are shown as the *X*, *Y*, and *Z* axes, respectively. In the boxes, the chemical structures are reported, along with the Euclidean Distance (ED) based on the ChemGPS-NP coordinates and the Tanimoto Index (TI) based on the ECFP_4 fingerprints.

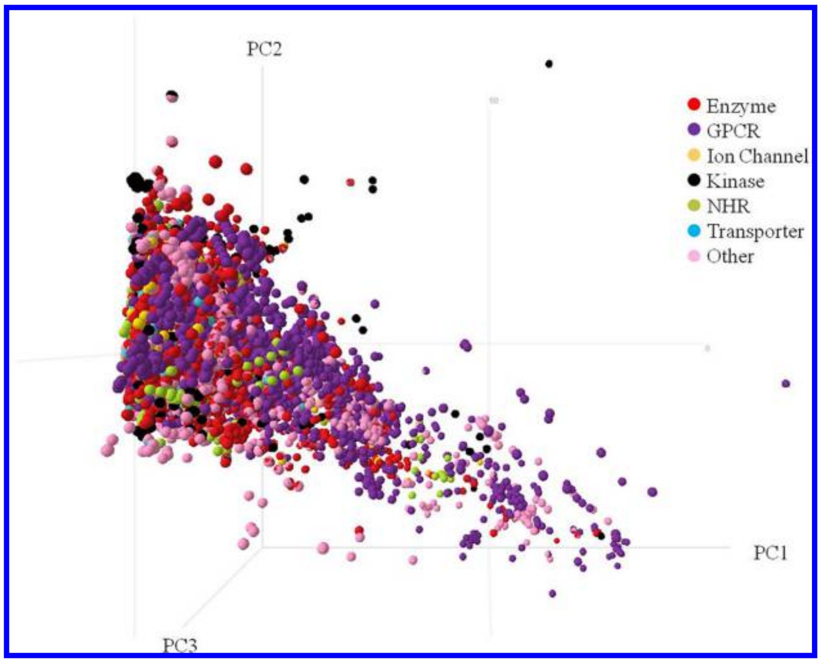


Figure 3. 3D score plot of the ChEMBL bioactive compounds colored according to the target classes: Enzyme in red, GPCR in purple, Ion Channel in yellow, Kinase in black, NHR in green, Transporter in cyan, and Other in pink. The first three principal components of he ChemGPS-NP chemical property space are shown as the *X*, *Y*, and *Z* axes, respectively.

that a larger number of GPCR bioactive compounds populate the positive direction of PC1 and PC4, and the negative direction of PC2 and PC3. These values indicate that the GPCR ligands present specific physicochemical properties (e.g., larger molecular size, higher polarity and flexibility, and a low degree of aromaticity). Many of the compounds included in the

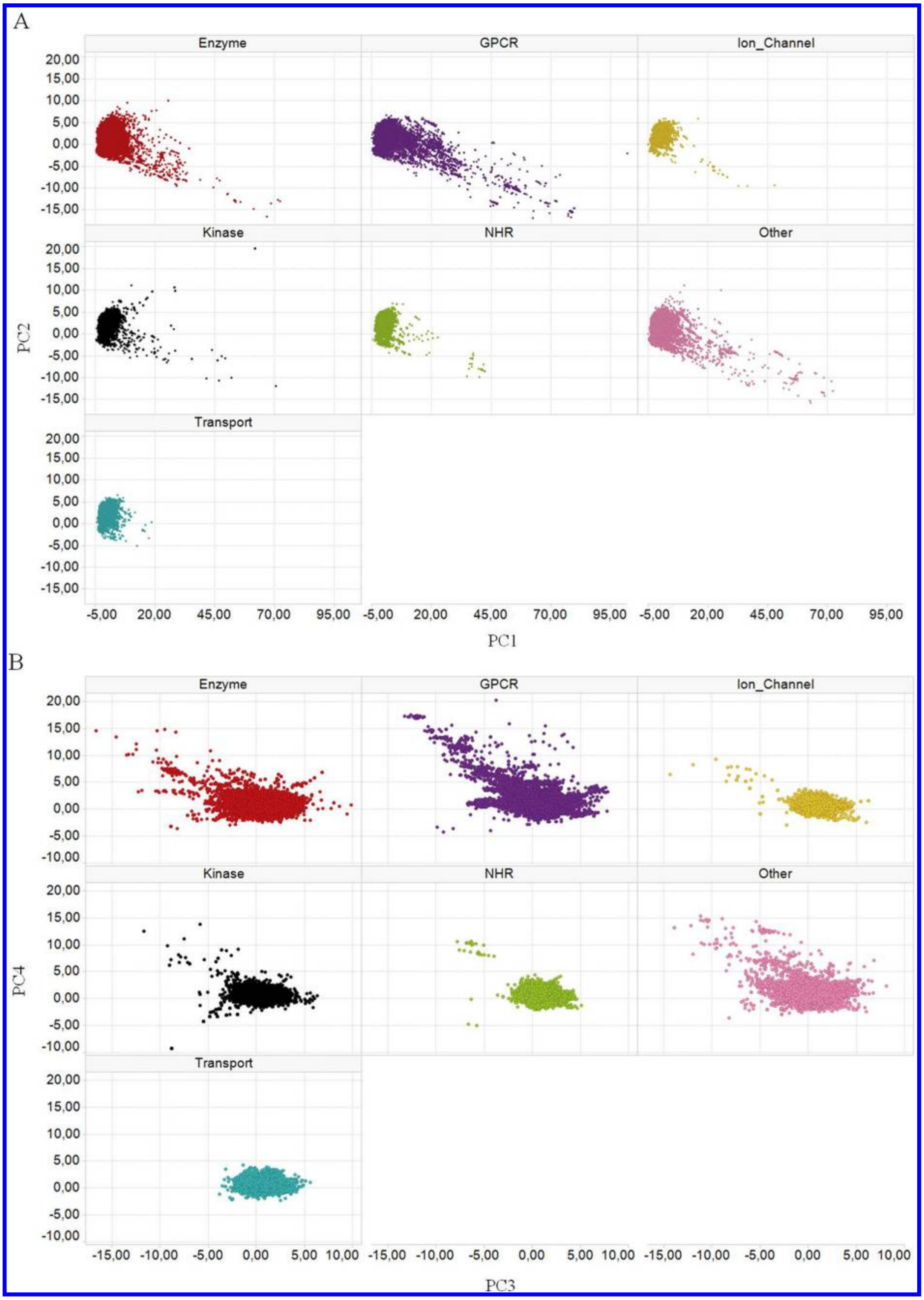


Figure 4. 2D score plots (PC2 vs PC1 in A, and PC4 vs PC3 in B) of the target classes: Enzyme in red, GPCR in purple, Ion Channel in yellow, Kinase in black, NHR in green, Transporter in cyan, and Other in pink.

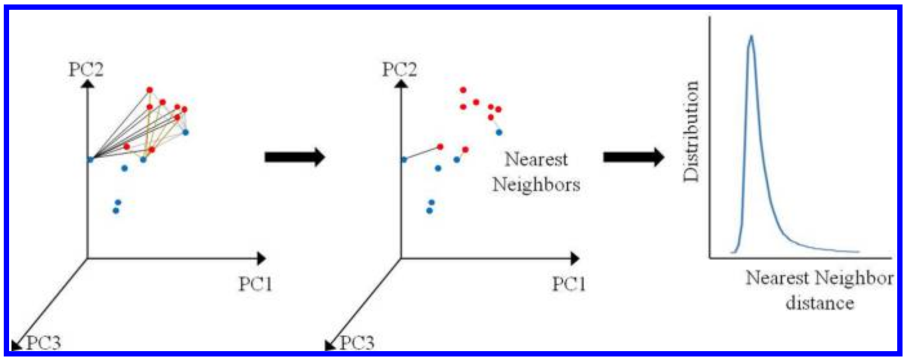


Figure 5. Workflow for the average Euclidean distance (ED) calculation. A training set and a test set are represented as red and blue dots, respectively. For sake of simplicity, the nearest neighbors of only three test set compounds are shown in the central plot.

target classes Ion Channel, Kinase, NHR, and Transporter are found in the most densely populated drug-like region and only a negligible fraction is scattered in the unpopulated area. The latter includes large compounds (e.g., oligosaccharide glycosides, peptides, tannin derivatives) that modulate targets like heparanase (Enzyme), gonadotropin-releasing hormone receptor (GPCR), different isoforms of protein kinase C (Kinase), and androgen receptor (NHR).[34−37] The targets annotated as Other have a higher percentage of compounds distributed in this low-density area. However, the lack of ontology annotation prevented a better understanding of this groups and, consequently, further analyses, hereafter, are referred to the remaining classes.

Thus, ChemGPS-NP enables to visualize target classes or, at a deeper level, single targets in the chemical space, although many times a clear distinction between two targets is not possible due to overlap in the physicochemical space. In fact, most of them are located in the highly dense area, due to both (i) the limited range of physicochemical properties observed in the bioactive compounds, mainly derived from drug discovery projects, and (ii) the parametrization of the ChemGPS-NP which is calibrated to map the largest chemical space as possible (e.g., including molecules with more extreme physicochemical properties and chemical diversity often encountered in natural products).[10,11]

ChemGPS-NP as a Tool for Biologically Annotated Compounds—Validation Study. Many active compounds located in the so-called "dense region" share similar location with compounds that could present completely different biological profiles. While this could open the way for the use of ChemGPS-NP as promising tool for ligand target identification, it could also lead to misleading similarities. To solve this issue, in a first step, we decided to compare sets of active compounds instead of single molecules. This allowed for taking into account the diversity of molecules that modulate the same target and, at the same time, investigate protein relationships based on the chemistry of their ligands. Thus, in order to include the overall sets in the analysis, we decided to use an average distance approach by taking into account all the compounds within each target.

To evaluate the accuracy of the ChemGPS-NP method to map the targets according to the similarity of their modulators, a cross validation procedure was performed. More specifically, all the targets were randomly divided into a training set containing 80% of the compounds and a test set with the remaining 20% of the compounds. During the validation process, the distances between all the training and test sets were calculated by averaging all Euclidean distances between the test set compounds and their nearest neighbors (NN) in the training sets, that were pinpointed through the shortest Euclidean distance. Figure 5 shows a simplified scheme of this approach. Here, the training sets represent the reference compound sets, whereas the test sets simulate new candidates with unknown pharmacological profile to be profiled based on the calculated similarity score against all the targets. For this reason, we computed the distance of the test sets against all the training sets. The entire procedure was repeated three times, in order to mitigate any chance correlation. As results were not significantly affected, this proved to be sufficient, so only the observation from one experiment is reported hereafter. A square matrix containing 826 281 records was obtained by comparing all 909 test sets to the training sets (Figure S1A). The diagonal dots correspond to training-test set pairs sharing the same activity (thereafter referred to as "pharmacologically related" or "EGID related" target pairs). Note that, in the following discussion, target pairs, set pairs and combinations all refer to training-test set pairs. According to the similarity property principle,[38,39] the test sets and their nearest training sets being EGID related, are supposed to reside closely in the chemical space. Therefore, the diagonal dots are desired to be found as the closest combinations, resulting in the lowest Euclidean distance and the best rank among the other possible set pairs. In this context, a similarity threshold turned out to be necessary for rapid interpretation of similar training-test sets across all the combinations. In fact, although the visualization enables to generally delineate close sets, a reference distance could provide more straightforward results. To this aim, an accepted Euclidean distance cutoff in the high-dimensional ChemGPS-NP space has not been published so far. In order to rationalize the definition of an ED cutoff, we computed the distribution of the group Euclidean distances of (i) the diagonal pairs and (ii) 10 000 random combinations extracted from the square matrix. Thus, a threshold of 1.00 was suggested as robust criterion to classify similar and dissimilar target pairs. In fact, this cutoff clearly separated pharmacologically related from random combinations (blue and red curves in Figure S2A, respectively). Namely, the pharmacologically related pairs mostly resided in the leftmost part of the plot, with comparably low averaged Euclidean distances (ED < 1.00); whereas larger values were obtained for pharmacologically unrelated sets.

Comparison between ChemGPS-NP and Fingerprint-Based Similarity. As we mentioned previously, 2D fingerprint-based methods are currently among the most commonly used approaches for querying the bioactive space in search for new active ligands. To investigate the efficiency and the potentiality of ChemGPS-NP as an additional physicochemical based approach, a comparative study with ECFP_4 was performed. Originally, ECFP_4 was designed to capture molecular features relevant to molecular activity and, in extension, came in use for tasks such as similarity searching, clustering and virtual screening.[21] In this study, ECFP_4 were computed for all compounds in our data set, and similarity between molecules was assessed with the Tanimoto Index.[40] Inversely to the ChemGPS-NP case, different Tanimoto similarity threshold have been reported in literature for ECFP_4.[21,41,42] However, a calculation similar to the one to define the ED threshold has been carried out to suggest a suitable TI cutoff for the data set used in this study (Figure S2B). This TI enables to estimate the similarity between set pairs presenting (or not) the same pharmacological profiles. On the basis on this assessment, we can refer to highly similar set pairs if the Tanimoto Index is greater than 0.70, while a TI below 0.60 is associated with a larger number of pharmacological unrelated pairs.

As previously implemented in the case of the ChemGPS-NP method, the distance between each test and training set was computed as the average of TIs between all ligands in each test sets and all their nearest neighbors for each training set. Note that, in opposite to the ChemGPS-NP derived Euclidean distance, the nearest neighbor is associated with the *highest* TI value. Furthermore, the ED and the TI spanned different ranges. In fact, while the former can range to very large values as previously discussed, the Tanimoto Index is restricted to values between 0.00 and 1.00. For this reason, a distance-based ranking was found more appropriate to compare the data. Additionally, two ways of comparing the ranking from the two methods were followed: (i) the difference of the ranks ($\text{rank}_{\text{ChemGPS}} - \text{rank}_{\text{FP}}$) and (ii) sum of the ranks subtracted by two ($[\text{rank}_{\text{ChemGPS}} + \text{rank}_{\text{FP}}] - 2$). Further details on these approaches are discussed below.

Test Case—Pharmacological Similarity of Structurally Diverse Target Sets. Pharmacological similarities between related targets consisting of congeneric structures with comparable physicochemical properties are expected to be observed in both ChemGPS-NP and fingerprint methods. Also in the opposite scenario of structurally diverse targets, ChemGPS-NP similarity search could retrieve biological correlations. In order to test this assumption, an external validation was carried out. ChEMBL training and test sets for each target were grouped together resulting in 909 reference sets, while bioactive compounds from GoStar database were selected as external test cases. Initially, GoStar and ChEMBL bioactive molecules modulating the same target were compared by means of ECFP_4 fingerprints. To ensure structural diversity between these two data sets, GoStar molecules presenting TIs lower than 0.5 were kept for the external validation sets. In total, 465 target sets consisting of 209 674 unique compounds with high structural diversity as compared with ChEMBL data set were obtained from this selection process. Subsequently, a comparison of all retrieved 465 GoStar targets to the ChEMBL sets was performed. The averaged Euclidean distance was computed as described above and, in total, 422 685 combinations were obtained.

■ RESULTS AND DISCUSSION

The pairwise similarity comparison between all training and test sets based on ChemGPS-NP and ECFP_4 are reported as square matrixes (see Figure S1A and S1B). For each test set (row of the matrix), a distance-based ranking was created by sorting the similarity in ascending or descending order depending on the method used. Subsequently, a rank value of "1" corresponds to the closest training set while a value of "909" is the farthest and the least similar training set to a given test set. Moreover, the cells on the matrix diagonal correspond to the training and test set for the same target. Therefore, an assessment of the resulting rank of such diagonal combinations was provided.

In the ChemGPS-NP case, 378 out of 909 data points along the diagonal have the best rank. In other words, the nearest training set to 378 of the test sets belong to the same target according to their ED values. A similar result was obtained from the ECFP_4 study, where 429 set pairs from the same target have the highest ranking. The Venn diagram in Figure 6 highlights a substantial overlap of the training-test sets matching the same target between the ChemGPS-NP and the ECFP_4 methods. In fact, 356 diagonal combinations are best ranked in both approaches, whereas 73 and 22 dots are unique for the ECFP_4 and ChemGPS-NP validation study, respectively. The results are also reported in Figure 7, where the rank sum and difference of the diagonal combinations from the two approaches are displayed in a scatter plot. Increasing values of the rank sum along the *X*-axis correspond to low ranked combinations for both or at least one method; whereas low numbers indicate correlations among the top ranked sets. Simultaneously, the *Y*-axis describes the consensus or discrepancy of the ranks from the two methods. A *Y* value of zero means that both methods give the same ranking for a given pair, the higher, positive, values indicate a better similarity identified through the fingerprints, whereas lower, negative, values implicate a better performance of the ChemGPS-NP approach. Therefore, the diagonal combinations retrieved among the top ranked groups in both methods fall in the leftmost part of the *X*-axis and present *Y* values approaching zero. The use of the sum of the ranks subtracted by two was preferred over a simple addition. In this way, the EGID related pairs found as the best ranking combinations (rank equal to one) in both approaches could be easily retrieved at the origin of the axes. Hence, the 356 common best ranked combinations are represented by zero values of both coordinates (cyan dot in Figure 7).

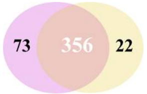


Figure 6. Venn diagram with the best ranked targets retrieved from the ChemGPS-NP approach (in yellow) and fingerprint method (in purple). The overlap is reported as the color combination of the yellow and purple.

In order to understand the strength and weakness of the two methods, here we discuss some cases in which ChemGPS-NP and ECFP_4 give the same rank and others where one of the methods uniquely retrieves the pharmacologically related pairs as the best combination.

Among the GPCR targets, the training and test set consisting of modulators of the κ Opioid receptor (OPRK) was perfectly retrieved from both approaches, with significantly high similarity. More precisely, the averaged ED is lower than 0.35 and the averaged TI is close to 0.82 and can be then associated with a high similarity for ECFP_4. The training and test set from this target include similar alkaloid derivatives, peptides,

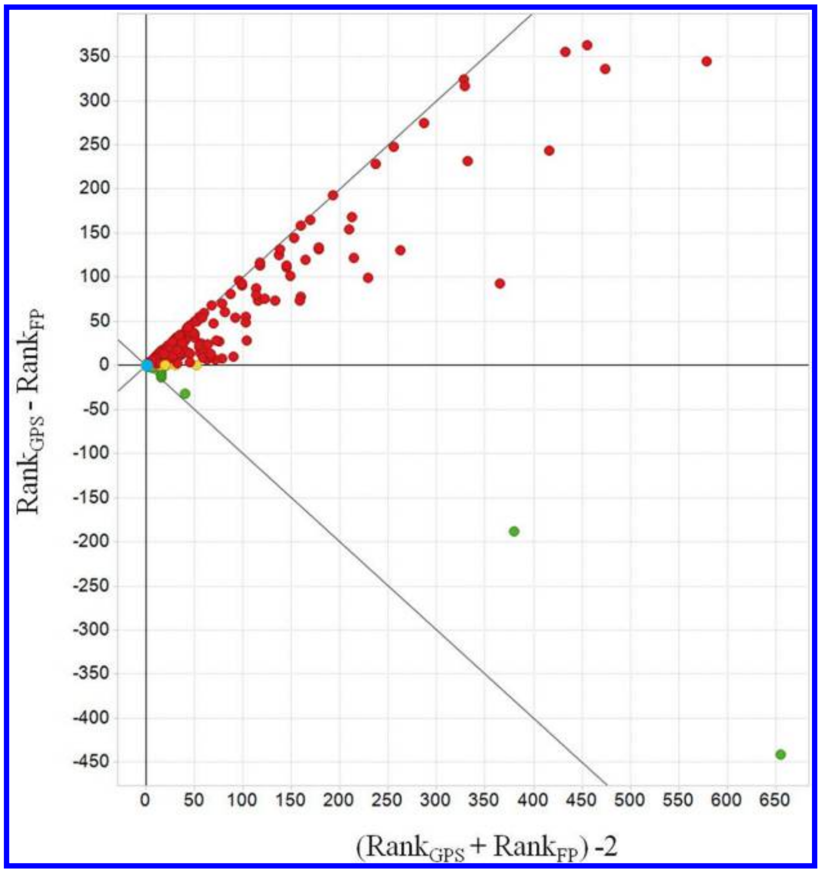


Figure 7. Scatter plot with rank difference and sum in *Y*-axis and *X*-axis, respectively. Red dots represent differences larger than one corresponding to a better performance of fingerprint approach compared to ChemGPS-NP. The opposite (negative difference) is reported in green. The rank consensus (zero value of difference) is showed in yellow. The cyan dot represents the pharmacologically related pairs retrieved as the best ranking combinations (rank equal to 1) for both the ChemGPS-NP and the ECFP_4 approach.

and other modulators, so as to bestow high similarity according to both physicochemical properties and structural features. Similar observations with consensus and high similarity simultaneously obtained were found, for instance, for the BACE1 (Enzyme), P2RX7 (Ion Channel), and PPARG (NHR) targets. All these targets consist of hundreds of active molecules and are mainly located in the drug-like region of the ChemGPS-NP map with high compound density, as reported in Figure 8A (left) for PPARG set, while a smaller number of compounds corresponding to cyclic and linear peptidic OPRK modulators[43,44] and BACE1 inhibitors[45,46] occupy other parts of the map.

With respect to the 22 best ranked diagonal pairs uniquely found with the ChemGPS-NP method, the corresponding EDs fall within the similarity range (smaller than 1.00) to distinguish similar from dissimilar target pairs. These 22 diagonal combinations were all retrieved among the second or third rank in the fingerprint approach, nearly resulting in a consensus. Thus, these combinations forming a bisector with negative slope (Figure S3A) result in a small difference and sum according to the rank, and are plotted in proximity of the origin of the axes. Interestingly, some training–test pairs result in a TI smaller than 0.70 corresponding to a lower structural similarities compared to the other target pairs analyzed in this study. For instance, the orexin receptor 1 (OX1R) test set was found close to the corresponding training set in the ChemGPS-NP map, and presented a TI = 0.66 in the ECFP_4 approach. The OX1R modulators include small molecules and polypeptides. These chemical classes are clearly distinguishable in the ChemGPS-NP map as shown in Figure 8A (central plot), with a high density area populated by drug-like molecules, a small group of modulators with intermediate size, and another small group with large compounds isolated in the bottom-right corner of the 3D plot. The TIs between the drug-like compounds of the OX1R training and test set turn out to be low, eventually affecting the final group distance. Besides, these values are comparable to the TIs between small molecules and some peptides. This result could be expected, as the differences in molecular complexity and size are known to bias the evaluation of fingerprint similarity.[47] In fact, subtle structural differences have a significant impact on fingerprints in the low molecular complexity space. This example highlights the advantage to combine these two methods based on different similarity criteria, in particular in the event that small molecules are under investigation. In this instance, pharmacological correlations based on ligand similarity can be still anticipated by using ChemGPS-NP method. Other cases of low similarity in fingerprint study concern signal transducer and activator of transcription 3 (STAT3), ubiquitin-conjugating enzyme E2 N (UBE2N) and nuclear receptor corepressor 2 (NCOR2) test sets which result in averaged TIs ranging from 0.55 to 0.64. Analogous to the previous example, such values depend on the chemical heterogeneity of the modulators of each target. Anticancer STAT3 inhibitors span from small aromatic molecules, quinolinone derivatives to phosphopeptides including compounds with unique scaffolds. It reflects on low TIs for

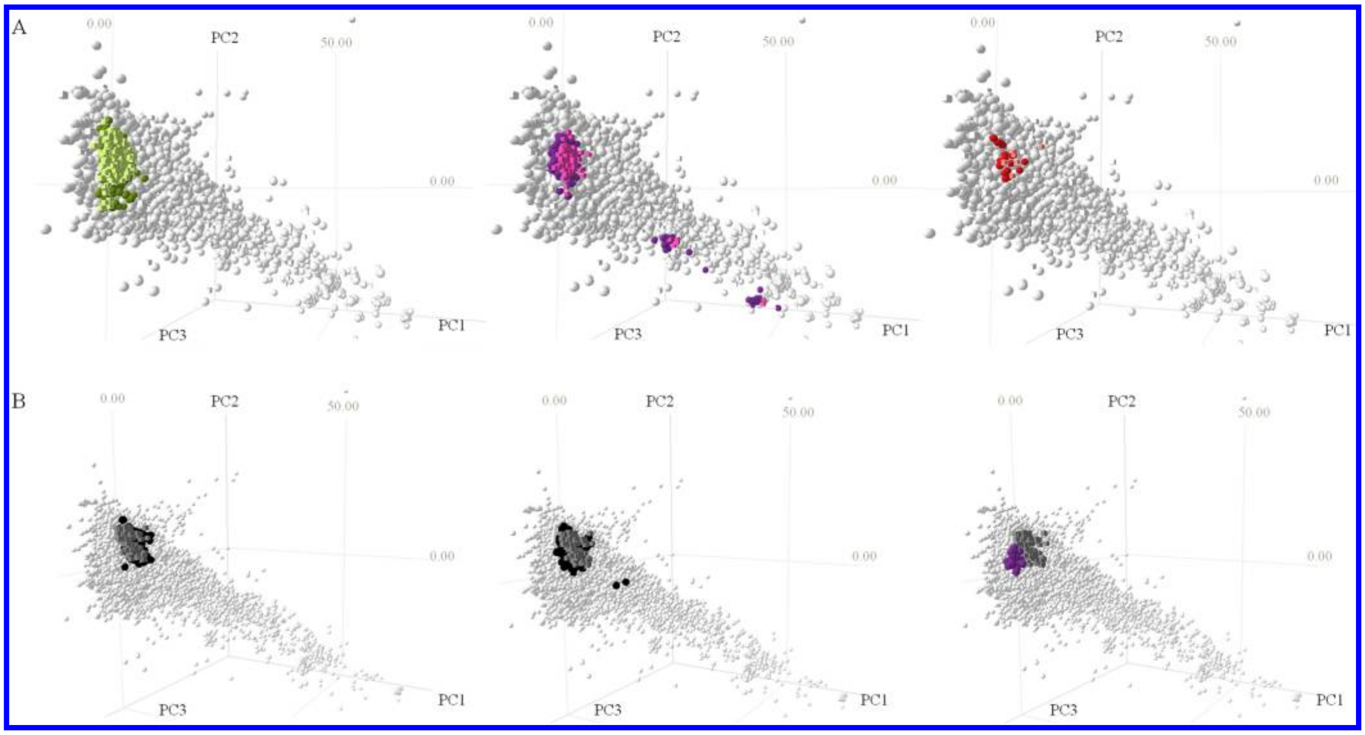


Figure 8. 3D visualization in ChemGPS-NP space. The first three principal components are shown as the *X*-, *Y*-, and *Z*-axes, respectively. (A) Training and test sets from the same target. Green, purple, and red colors correspond to PPARG, OX1R, and H-PGDS target sets, with light and dark tones for training and test sets, respectively. (B) AURKB test set (dark gray) is projected with the closest training set from the same target (black in the left panel), the second closest training set AURKA from the same target class (black in the central plot), and a distant training set from GPCR target class (purple in the right plot).

many nearest neighbor pairs between the training and test set. The same observation can be expanded to UBE2N and NCOR2 sets. Some examples of nearest neighbors between these test and training sets as obtained from fingerprint and ChemGPS-NP are reported in Table S3. Overall, the ChemGPS-NP method can provide good predictions where chemical structures are not frequent in a target set. In fact, structurally diverse compounds modulating similar targets can reside, in any case, in adjacent areas of the physicochemical property space and, consequently be picked as nearest neighbors, so as to result in better predictions compared to fingerprint method.

With regard to the 73 cases of top ranked training-test pairs that were identified correctly with the ECFP_4 fingerprint, while the same pairs were not best ranked with the ChemGPS-NP method, some general observations can be made. Some of these (11 combinations) have a TI below 0.60, meaning that within a virtual screening experiment the pharmacological similarity between these training and test compounds could have been missed. In parallel, most of these 11 pairs include a low number of compounds making target prediction increasingly difficult. In contrary to what was observed for the 22 targets that were best ranked with ChemGPS-NP and found among the second or third rank in the fingerprint approach, the 73 diagonal combinations show variable ranking in the ChemGPS-NP results. In fact, in 42 of the cases the target of the training set is not retrieved among the top five most similar targets of the test set. This trend is clearly observed in Figure S3B showing that a significant amount of the diagonal combinations have a low ranking in the ChemGPS-NP method (high *X* values). Further analysis reveals that the poor performance of the ChemGPS-NP approach in these cases appears to arise due to specific conditions. First, the poorly retrieved set pairs consist of compounds with very different physicochemical profiles and are therefore broadly distributed in the ChemGPS-NP space. It follows that some molecules in the test set are very distant from those in the training set even though they are active on the same target. Consequently, these isolated compounds impact the final ED leading to low similarity. Moreover, the training and test sets for a target located in the highly populated region of the ChemGPS-NP can lead to a worse ranking than one might expect, as these sets share a part of the chemical space where many other bioactive compounds are found that modulate other targets. Thus, there is a high probability that the NN compounds belong to different targets. The hematopoietic prostaglandin D synthase (H-PGDS) set is an example in which these conditions are met. In fact, the test set consists of 6 H-PGDS inhibitors with antiallergic or anti-inflammatory effect, including the cibacron blue[48−50] which occupies an isolated position in the ChemGPS-NP map compared to the other H-PGDS inhibitors located in the dense area (right plot in Figure 8A). Due to the outlier cibacron blue, the H-PGDS training-test set pair has a low rank. On the contrary, the correct set pair is best ranked in the fingerprint method, despite a TI = 0.61.

In summary, this analysis shows the scenarios in which ChemGPS-NP and fingerprint approaches can contribute to proper investigation of pharmacological similarities, based on comparability of molecular physicochemical properties or structural features. Within this validation study, it appears that both methods can rank adequately training-test pairs belonging to the same target where the sets are uniformly populated and structurally homogeneous. The presence of isolated compounds or *singletons* is challenging for both

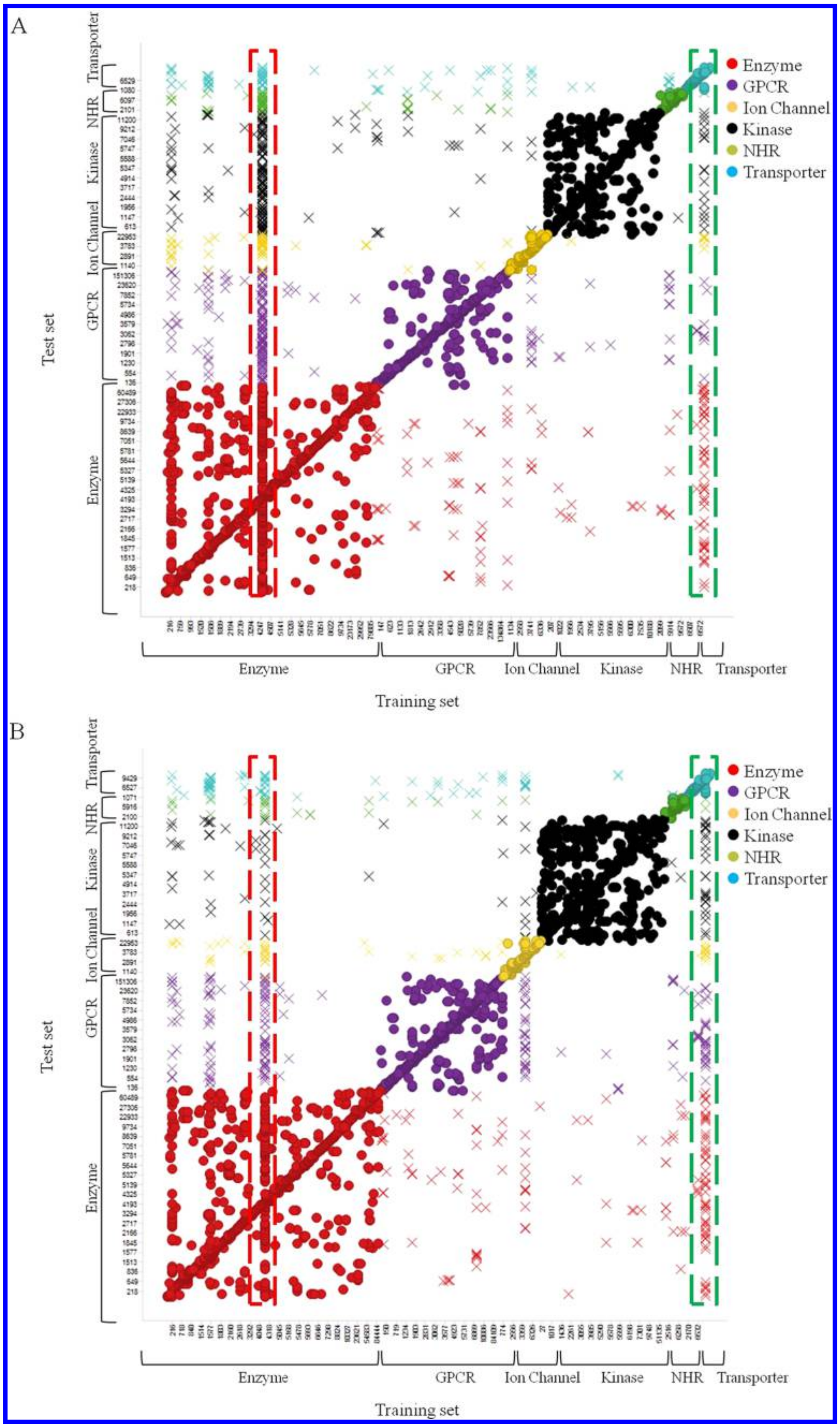


Figure 9. Square matrixes, as obtained through the average of the Euclidean distance approach (A) and the Tanimoto Index approach (B), after selecting the top five ranked training sets (column) for each test set (row). The dots are colored based on the target classes. Dots are represented as circles or crosses according as training and test sets belong to the same target classes or not. The red and green dashed boxes correspond to histone-lysine *N*-methyltransferase 2A (KMT2A) and Ras-related protein Rab-9A (RAB9A) training sets, respectively, which were retrieved among the top five ranked results of several test sets as discussed.

ChemGPS-NP and fingerprint approaches. In these cases, low averaged EDs can be obtained, although the fingerprint method turns out to give better ranking results. In addition, ChemGPS-NP provides a straightforward visualization, which enables to easily distinguish targets when their ligands occupy distinct areas of the chemical space.

So far, we have mainly focused our observation on the diagonal combinations of the similarity matrixes with the highest ranking. However, in the early stage of a target identification and validation process, scientists often have to follow a series of target hypothesis in parallel. This means that methods that could rank pharmacologically similar molecules at a sufficiently high ranking position can turn to be as useful as methods giving the highest ranking to those compounds.

In the following section, we investigate the performance of ChemGPS-NP in associating test and training sets corresponding to the same target by analyzing the top five similar pairs. While arbitrarily chosen, five target hypotheses are reasonably in line with what people can usually consider in a real life project to further validate experimentally. In addition, we discuss how ChemGPS-NP can in some cases elucidate the similarity of sets of compounds within the target classes.

We extended the diagonal selection to the top five ranked training sets for each line of the square matrix and analyzed the correlations in terms of target and protein classes. The square matrixes with five training sets for each test set are reported in Figure 9A and B from ChemGPS-NP and fingerprint results, respectively. As it can clearly be seen, the diagonal of both ChemGPS-NP and fingerprint matrixes become more populated, with 674 and 783 (out of 909) pharmacologically related combinations, respectively. The fingerprint approach still performs better in the number of retrieved combinations, but the relative TIs indicate low similarity to a greater extent compared to the ED results. The consensus between the methods is high as 667 diagonal pairs are simultaneously retrieved by both approaches, whereas 7 and 116 are corresponding to unique results from ChemGPS-NP and fingerprint study, respectively.

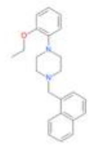
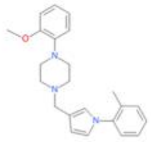
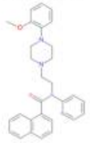
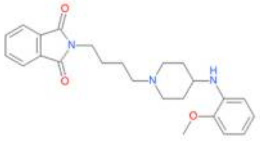
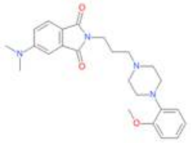
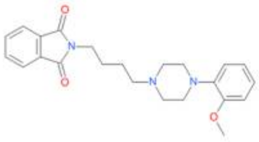
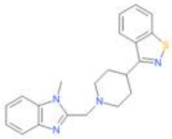
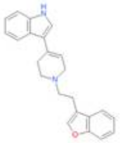
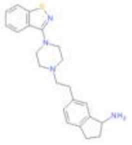
Interestingly, a large number of top five training-test pairs are associated with the same target class and are represented as circle dots in Figure 9. In other words, the Enzyme test sets are mainly located in proximity of the Enzyme training sets and the same trend is observed for the other classes. On the other side, the training and test sets modulating different classes of proteins are shown as crosses in the matrix and are clearly less frequent than the target class related combinations. Interestingly, this analysis reflects the presence of sets with compounds modulating multiple proteins, due to either their lack of selectivity or their polypharmacological profile. Also, there are conserved features within the ligand binding regions within a target class as has been previously studied, however, it is reassuring that the same results are also seen in the ChemGPS-NP study. The target class depicted here as Other is disregarded, as it obscures a more comprehensive understanding of the pharmacological similarities obtained from the two approaches.

From the analysis of the top five predicted combinations, we identified cases in which the correct target was retrieved as the nearest neighbor of the test set prior to other sets from the same target class. One example is the protein kinase C alpha type (KPCA) test set which resides near the corresponding training set in both the approaches, followed by KPCD, KPCB, KPCE, and KPCG.[51] It can be noted that all of these targets belong to the CMGC family and cluster in the same branch of the protein kinase dendrogram, according to the amino-acid sequence similarity analyses of their catalytic domains.[52] The same distance-based rank is observed when both the physicochemical properties in ChemGPS-NP and the structural features in fingerprint study are taken into account. Analogous similarities mirroring the kinase classification were obtained. The Aurora Kinase B (AURKB) test set retrieves AURKB and AURKA reference sets as the closest neighbors. In Figure 8B, the nearest neighbors of AURKB test set are illustrated along with a GPCR training set, in order to demonstrate the visualization capability of ChemGPS-NP model. In detail, AURKB test set extends along AURKB and AURKA training sets, while overlap is not observed against the atypical chemokine receptor 3 (ACKR3) training set shown in purple. Another example of pharmacological similarity in the chemical space concerns c-Jun N-terminal protein kinase 1 (JNK1) and JNK2 reference groups which lie close to the JNK1 test set. In the GPCR family, we also found strong pharmacological correlations. Returning to the OPRK example, all the opioid receptor sets (μ, δ, κ, and X) emerge among the top five ranked combinations suggesting a high degree of structural conservation between the opioid receptors, which can also be confirmed from comparison of their amino acid sequences. Accurate predictions are also observed for NHR targets, such as the steroid hormone receptors using both ChemGPS-NP and ECFP_4 approaches. The glucocorticoid and progesterone reference compounds represent the nearest neighbors of the glucocorticoid test modulators, and inverted order was found for the progesterone set.

We also obtained cases where the set pairs from the same target class turned out to be more similar than the diagonal combination. Such observations are expected to occur in the case of compounds active on multiple targets. In fact, if training and test sets of two different targets include the same modulators, the latter are picked as nearest compounds (with ED equal to 0.00 or TI equal to 1.00). This impacts the final group distance and, consequently, results in a better rank for these target pairs than for training-test set pairs of the same target. An example is represented by the carbonic anhydrase (CAH) superfamily, which is a subclass of metalloenzymes.[53] The most pharmacologically relevant class of CAH inhibitors includes sulfonamides and their isosteres (sulfamates, sulfamides). Many derivatives are used as antiglaucoma agents, anticonvulsants, diuretics or antiobesity drugs, and as anticancer agents/diagnostic tools for imaging tumors.[54] The CAH set used in this study consists of modulators of 12 isoforms (CA I–IV, Va and Vb, VI–VII, IX, XII, XIII, and XIV), some of which are reported with multiple annotations. For instance, about 1400 modulators are active on both CAH I and II sets, and CAH I and IX share about 1100 compounds. The top ranked combinations of CAH I test set include CAH II and IX training sets prior to those from the same EGID group. Moreover, the next nearest set is represented by the CAH XII set sharing 870 modulators. The same distance rank was obtained in the two approaches.

Thus, expanded selection of the neighbors of a test set of interest enables to investigate the correlations with training sets within the same target class and increase the chance to correctly identify pharmacologically related compound sets. However, in addition to these examples with consensus within the target classes, we also retrieved combinations consisting of unrelated sets in both the ChemGPS-NP and fingerprint space (crosses in

Table 3. Three Examples of the Nearest Neighbors between Compounds in the DRD2 Test Set and the 5-HT1_A Training Set[a]

Query in DRD2 test set	Nearest neighbor in ChemGPS-NP method	Nearest neighbor in ECFP_4 method
CHEMBL2420781 	CHEMBL177064 **ED** = 0.48 **TI** = 0.43 	CHEMBL2164348 **ED** = 2.36 **TI** = 0.46 
CHEMBL392791 	CHEMBL574667 **ED** = 0.41 **TI** = 0.47 	CHEMBL8618 **ED** = 0.43 **TI** = 0.64 
CHEMBL596527 	CHEMBL602244 **ED** = 0.63 **TI** = 0.19 	CHEMBL404000 **ED** = 1.14 **TI** = 0.27 

[a] ED and TI are shown for each pair. The compounds in the left column are the nearest neighbors selected through the ChemGPS-NP approach. For these compounds, the ED is reported and, for the sake of completeness, the corresponding TI is also given. In the right column, the opposite is shown. Biological annotations of these compounds are described in the literature.[61−66]

Figure 9). Some of these combinations consist of active ligands on multiple target proteins. One prominent example concerns the pair consisting of dopamine receptor 2 (DRD2) and potassium voltage-gated channel subfamily H member 2 (KCNH2 or hERG), forming the test and training set, respectively. It is widely known that hERG binding is one of the causes of long QT syndrome (LQTS), an effect which is in turn linked to life-threatening ventricular arrhythmias (*Torsades de Pointes*), in the clinic recognized as a very dangerous side effect of these drugs.[55] In order to reduce the risk, the hERG blocking liability of chemical entities is usually investigated and assessed along a drug discovery program.

Some dopaminergic antipsychotics (e.g., pimozide, haloperidol, chlorprothixene, etc.) fit with one of the most accepted pharmacophoric models for hERG blockers, which is characterized by a central positive ionizable feature linked to aromatic or hydrophobic groups.[56−59] In this respect, the design of new DRD2 inhibitors, frequently used as typical and atypical antipsychotics, is completed by hERG liability tests so as to minimize the risk of cardiac side effects. The presence of such dual-acting modulators then explains the correlation obtained through the similarity study in ChemGPS-NP and fingerprint space.

Another interesting observation arises from the next closest training sets of DRD2 test set. Apart from the hERG group, the 5-HT1_A set was found among the top five ranked combinations in both the ED and TI study. This set is known to be involved in cognitive disorders similarly to DRD2 ligands.[60] Also, the endogenous neurotransmitters of these two targets (dopamine and serotonine) are both monoamines, with an amino group that is connected to an aromatic ring by a two-carbon chain. In addition to being partially populated by the same modulators, these sets also include compounds with unique but similar physicochemical properties and, consequently, are located in the same region of the ChemGPS-NP map and with comparable distributions. Table 3 displays a series of nearest neighbors between the DRD2 test set and 5-HT1_A training set.[61−66] Notably, these compound pairs present low TI but reside closely in the ChemGPS-NP space. Hence, we took advantage of the visual inspection in the ChemGPS-NP plots, additionally confirmed by calculated ED, to introduce new target hypotheses of potential therapeutic interest. In other words, compounds which modulate the dopaminergic receptor could be proposed as lead candidates of 5-HT1_A based on their close vicinity in the ChemGPS-NP map. In first approximation, this observation would have been more difficult to achieve merely by a study of structural similarity in fingerprint space. If, on one hand, fingerprint similarity search is preferred as the most conservative method in retrieving similar chemical structures, then simultaneously the structural heterogeneity of the NN compounds in the ChemGPS-NP can enrich the screening campaign with new scaffolds, in the spirit of charting new borders of the chemical space. As a precaution, one would consider to limit these new hypotheses to nearest neighbors concerning modulators within the same target class or target proteins of similar therapeutic interest, such as the dopaminergic and serotonergic receptors. Alternatively, target hypotheses from different protein families could be provided and validated by means of biological assays. Similarity methods such as SEA[20] based on Daylight fingerprints have provided unexpected ligand relationships between proteins from different families. In this study, ChemGPS-NP gives the possibility to introduce new target hypotheses according to a different similarity criterion. Extending this observation in the context of drug discovery, similarity search in ChemGPS-NP might open the way for the investigation of known modulators on new target proteins involved in similar pathological processes and not.

In addition to a few cases in which unrelated training and test sets were retrieved among the top five ranked combinations for specific reasons, the remaining pairs from different target classes were the results of limitations in the capacity of either ChemGPS-NP or fingerprints to reveal the expected patterns. One such example concerns two training sets that were found among the top five ranked results of several test sets. They are reported in Figure 9 as red and green dashed boxes, and correspond to histone-lysine *N*-methyltransferase 2A (KMT2A) and Ras-related protein Rab-9A (RAB9A) from the Enzyme and Transporter classes, respectively. The reason why these two training sets were commonly retrieved relates to their size and inherent features. Both these sets included over 15 000 compounds each.[67−69] Also, the visualization in ChemGPS-NP map shows that they are mainly distributed in the most densely populated region of the ChemGPS-NP chemical property space which corresponds to small drug-like modulators, so as to be found similar to several test sets from the same and other target classes. This shows inherent limitation to similarity-based methods in explaining pharmacological similarity, especially in the event of promiscuity in the analyzed set.

In order to strengthen the applicability of ChemGPS-NP similarity search, a further validation study was carried out using a new pool of compounds with high structural diversity. They were selected from the GoStar database resulting in 465 target sets (see the Methods section above). When compared to the 909 structurally diverse targets from ChEMBL, 97 diagonal pairs were retrieved with an averaged ED below the proposed 1.00 similarity cutoff. In other words, pharmacologically related sets consisting of compounds with low chemical similarity according to ECFP_4 fingerprint could be retrieved in the ChemGPS-NP space with reasonably short Euclidean distances. Such well predicted diagonal pairs consist of highly populated target sets which are mainly distributed in the drug-like area. As examples, PPARG test set from GoStar was found close to the corresponding ChEMBL training set, despite the low structural similarity. The same trend was obtained for PPARA target, estrogen receptor 1 and 2 (ESR1 and ESR2), and also for neuropeptide S receptor (NPSR1) and Beta-3 adrenergic receptor (ADRB3) both from the GPCR class. A careful analysis of the nearest neighbor compounds between ChEMBL training sets and GoStar test sets of these targets highlights once more that close compounds in the ChemGPS-NP space are not necessarily similar in terms of chemical structures. Therefore, this external study emphasizes the added value of ChemGPS-NP in similarity search practice, by virtue of a similarity criterion which is more generic in respect to fingerprint-based approaches.

In the case of both structural and physicochemical diversity in a compound set, the ChemGPS-NP method, from inherent limitations, results in poor pharmacological predictions as observed for the remaining diagonal pairs with high averaged Euclidean distances.

■ CONCLUSIONS

The chemical space is extremely vast and estimated to be populated by 10^{23}−10^{60} compounds.[70] Techniques to reduce dimensionality and tools to visualize this complex and mindboggling expanse might simplify the navigation and enable us to extract useful information for drug discovery. In this paper, we made use of ChemGPS-NP to chart and navigate the biologically relevant chemical space defined by the ChEMBL collection. Initially we took advantage of the straightforward visualization provided by ChemGPS-NP to distinguish different distribution according to distinct target classes. A dense region mostly populated by drug-like active compounds was easily identified including molecules from all the target classes. Moreover, it was easily recognizable compared to a sparsely populated region with "peptide-like" compounds. The low population density of some areas might also arise from uncommon (or by evolution yet unexplored) combinations of physicochemical properties described across the ChemGPS-NP axes.

Ideally, the biologically active volume of chemical space would be formed by separate clusters of compounds, each one associated with a different receptor.[71] In consideration of the similar property principle, we can project new compounds in the chemical space and select surrounding molecules from these clusters to predict their biological profile. In this study, we partitioned target classes in 909 targets and assessed ChemGPS-NP for the ability to associate similar physicochemical properties to similar biological activities, in perspective of future applications for drug discovery. In order to make this validation study more robust, we also evaluated the structural similarity based on ECFP_4 of the same collection, as an alternative or complement to the ChemGPS-NP descriptors. Training and test sets for each target protein set were then compared each other and their similarity was quantified through ED metrics in ChemGPS-NP and TI in fingerprint approach. Lastly, an external validation was carried out exploiting the GoStar database, aimed to investigate the ability of ChemGPS-NP similarity search to retrieve pharmacologically related targets from a set consisting of structurally diverse compounds.

Overall, the similarity of the targets based on physicochemical properties or structural features reflected the pharmacological correlations at target class level. Also, a large number of the training and test sets from the same target were found as nearest neighbors in both ChemGPS-NP and fingerprint space. An even more robust result was obtained when the closest five nearest neighbors for each test set were analyzed. In this case, significant correlations among targets of the same class were obtained. Poor similarity was observed for a few cases due to methodological limitations. In some cases, the physicochemical descriptors implemented in ChemGPS-NP cannot efficiently describe the key structural features needed to distinguish distinct targets. Similarity could be also misinterpreted with fingerprint approach in case of small modulators, as described for the OX1R example, because of the accepted bias of the fingerprint approach caused by molecular complexity. In case of highly populated targets with unique scaffolds, ChemGPS-NP can provide better results due to a different similarity criterion which is not strictly and uniquely related to structural features, as discussed for STAT3 target. With regard to the external validation, we also provided examples in which physicochemical similarity can be an advantageous criterion to depict pharmacologically related targets when structural correlations are missing. Although in this study demonstrated only for a limited number of targets, it appears a useful alternative or complementary method in similarity search campaigns, in particular in combination with its capabilities of providing basis for a clear visual inspection.

In conclusion, this study showed the potentiality of using different approaches to investigate pharmacological correlations through ligand similarity. In addition, the visual inspection through ChemGPS-NP can easily guide to depict similar targets. Both methods based on physicochemical properties or structural features can properly identify biological correlations and be helpful in target identification for drug discovery campaigns.

■ ASSOCIATED CONTENT

S Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00375.

> List of targets used in this study, Table S1. Examples of isolated compounds in the ChemGPS-NP map, Table S2. Examples of nearest neighbors from ChemGPS-NP and ECFP_4 methods, Table S3. Square matrixes of the group distances from the ChemGPS-NP (A) and fingerprint (B) approach, Figure S1. Distribution curves used to define the ED similarity threshold (A) and TI threshold (B), Figure S2. Scatter plot with rank difference and rank sum between ChemGPS-NP and ECFP_4 results, Figure S3 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rosa.buonfiglio@astrazeneca.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 606895−MedPlant ITN. The authors wish to thank Prof. Johan Gottfries for the stimulating discussions during the early stages of this work.

■ REFERENCES

(1) Stockwell, B. R. Chemical Genetics: Ligand-Based Discovery of Gene Function. *Nat. Rev. Genet.* **2000**, *1*, 116−125.

(2) Stockwell, B. R. Exploring Biology with Small Organic Molecules. *Nature* **2004**, *432*, 846−854.

(3) Terstappen, G. C.; Schlupen, C.; Raggiaschi, R.; Gaviraghi, G. Target Deconvolution Strategies in Drug Discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 891−903.

(4) Maniyar, D. M.; Nabney, I. T.; Williams, B. S.; Sewing, A. Data Visualization During the Early Stages of Drug Discovery. *J. Chem. Inf. Model.* **2006**, *46*, 1806−1818.

(5) 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.

(6) Maggiora, G. M. On Outliers and Activity Cliffs–Why Qsar Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(7) 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.

(8) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. Sar Maps: A New Sar Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926−5937.

(9) 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.

(10) Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. Chemgps-Np: Tuned for Navigation in Biologically Relevant Chemical Space. *J. Nat. Prod.* **2007**, *70*, 789−794.

(11) Rosen, J.; Lovgren, A.; Kogej, T.; Muresan, S.; Gottfries, J.; Backlund, A. Chemgps-Np(Web): Chemical Space Navigation Online. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 253−259.

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

(13) Rosén, J.; Rickardson, L.; Backlund, A.; Gullbo, J.; Bohlin, L.; Larsson, R.; Gottfries, J. Chemgps-Np Mapping of Chemical Compounds for Prediction of Anticancer Mode of Action. *QSAR Comb. Sci.* **2009**, *28*, 436−446.

(14) Lee, C. L.; Lin, Y. T.; Chang, F. R.; Chen, G. Y.; Backlund, A.; Yang, J. C.; Chen, S. L.; Wu, Y. C. Synthesis and Biological Evaluation of Phenanthrenes as Cytotoxic Agents with Pharmacophore Modeling and Chemgps-Np Prediction as Topo Ii Inhibitors. *PLoS One* **2012**, *7*, e37897.

(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) Muigg, P.; Rosén, J.; Bohlin, L.; Backlund, A. In Silico Comparison of Marine, Terrestrial and Synthetic Compounds Using Chemgps-Np for Navigating Chemical Space. *Phytochem. Rev.* **2013**, *12*, 449−457.

(17) Lipinski, C. A. Lead- and Drug-Like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technol.* **2004**, *1*, 337−341.

(18) Feng, Y.; Campitelli, M.; Davis, R. A.; Quinn, R. J. Chemoinformatic Analysis as a Tool for Prioritization of Trypanocidal Marine Derived Lead Compounds. *Mar. Drugs* **2014**, *12*, 1169−1184.

(19) 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*, D1100−1107.

(20) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.

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

(22) Jagarlapudi, S. A.; Kishan, K. V. Systems for Knowledge-Based Discovery. *Chemogenomics, Methods Mol. Biol.* **2009**, *575*, 159−172.

(23) Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T. Entrez Gene: Gene-Centered Information at Ncbi. *Nucleic Acids Res.* **2011**, *39*, D52−57.

(24) Eriksson, M.; Nilsson, I.; Kogej, T.; Southan, C.; Johansson, M.; Tyrchan, C.; Muresan, S.; Blomberg, N.; Bjareland, M. Sarconnect: A Tool to Interrogate the Connectivity between Proteins, Chemical Structures and Activity Data. *Mol. Inf.* **2012**, *31*, 555−568.

(25) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discovery* **2013**, *12*, 948−962.

(26) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. Making Every Sar Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discovery Today* **2011**, *16*, 1019−1030.

(27) *Dragon (Software for Molecular Descriptor Calculation)*, version 6.0; Talete Srl, 2012; http://www.talete.mi.it/.

(28) *Simca-P+ 11.5*; Umetrics AB: Umeå, Sweden, 2008.

(29) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint Based Similarity Searches for Targeted Class Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201−1213.

(30) Godden, J. W.; Bajorath, J. A Distance Function for Retrieval of Active Molecules from Complex Chemical Space Representations. *J. Chem. Inf. Model.* **2006**, *46*, 1094−1097.

(31) Brabez, N.; Saunders, K.; Nguyen, K. L.; Jayasundera, T. B.; Weber, C.; Lynch, R. M.; Chassaing, G.; Lavielle, S.; Hruby, V. J. Multivalent Interactions: Synthesis and Evaluation of Melanotropin Multimers - Tools for Melanoma Targeting. *ACS Med. Chem. Lett.* **2013**, *4*, 98−102.

(32) Gilligan, P. J.; He, L.; Clarke, T.; Tivitmahaisoon, P.; Lelas, S.; Li, Y. W.; Heman, K.; Fitzgerald, L.; Miller, K.; Zhang, G.; Marshall, A.; Krause, C.; McElroy, J.; Ward, K.; Shen, H.; Wong, H.; Grossman, S.; Nemeth, G.; Zaczek, R.; Arneric, S. P.; Hartig, P.; Robertson, D. W.; Trainor, G. 8-(4-Methoxyphenyl)Pyrazolo[1,5-a]-1,3,5-Triazines: Selective and Centrally Active Corticotropin-Releasing Factor Receptor-1 (Crf1) Antagonists. *J. Med. Chem.* **2009**, *52*, 3073−3083.

(33) Huynh, A. S.; Chung, W. J.; Cho, H. I.; Moberg, V. E.; Celis, E.; Morse, D. L.; Vagner, J. Novel Toll-Like Receptor 2 Ligands for Targeted Pancreatic Cancer Imaging and Immunotherapy. *J. Med. Chem.* **2012**, *55*, 9751−9762.

(34) Johnstone, K. D.; Karoli, T.; Liu, L.; Dredge, K.; Copeman, E.; Li, C. P.; Davis, K.; Hammond, E.; Bytheway, I.; Kostewicz, E.; Chiu, F. C.; Shackleford, D. M.; Charman, S. A.; Charman, W. N.; Harenberg, J.; Gonda, T. J.; Ferro, V. Synthesis and Biological Evaluation of Polysulfated Oligosaccharide Glycosides as Inhibitors of Angiogenesis and Tumor Growth. *J. Med. Chem.* **2010**, *53*, 1686−1699.

(35) Samant, M. P.; Gulyas, J.; Hong, D. J.; Croston, G.; Rivier, C.; Rivier, J. Synthesis, in Vivo and in Vitro Biological Activity of Novel Azaline B Analogs. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2894−2897.

(36) Kashiwada, Y.; Nonaka, G.-i.; Nishioka, I.; Ballas, L. M.; Jiang, J. B.; Janzen, W. P.; Lee, K.-H. Tannins as Selective Inhibitors of Protein Kinase C. *Bioorg. Med. Chem. Lett.* **1992**, *2*, 239−244.

(37) Seoane, M. D.; Petkau-Milroy, K.; Vaz, B.; Mocklinghoff, S.; Folkertsma, S.; Milroy, L.-G.; Brunsveld, L. Structure-Activity Relationship Studies of Miniproteins Targeting the Androgen Receptor-Coactivator Interaction. *MedChemComm* **2013**, *4*, 187−192.

(38) Johnson, M. A.; Maggiora, G. M. In *Concepts and Applications of Molecular Similarity*; John Wiley and Sons, Ltd.: New York, 1990.

(39) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(40) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.

(41) Drwal, M. N.; Banerjee, P.; Dunkel, M.; Wettig, M. R.; Preissner, R. Protox: A Web Server for the in Silico Prediction of Rodent Oral Toxicity. *Nucleic Acids Res.* **2014**, *42*, W53−58.

(42) Sharma, R.; Lawrenson, A. S.; Fisher, N. E.; Warman, A. J.; Shone, A. E.; Hill, A.; Mbekeani, A.; Pidathala, C.; Amewu, R. K.; Leung, S.; Gibbons, P.; Hong, D. W.; Stocks, P.; Nixon, G. L.; Chadwick, J.; Shearer, J.; Gowers, I.; Cronk, D.; Parel, S. P.; O'Neill, P. M.; Ward, S. A.; Biagini, G. A.; Berry, N. G. Identification of Novel Antimalarial Chemotypes Via Chemoinformatic Compound Selection Methods for a High-Throughput Screening Program against the Novel Malarial Target, Pfndh2: Increasing Hit Rate Via Virtual Screening Methods. *J. Med. Chem.* **2012**, *55*, 3144−3154.

(43) Pickett, J. E.; Nagakura, K.; Pasternak, A. R.; Grinnell, S. G.; Majumdar, S.; Lewis, J. S.; Pasternak, G. W. Sandmeyer Reaction Repurposed for the Site-Selective, Non-Oxidizing Radioiodination of Fully-Deprotected Peptides: Studies on the Endogenous Opioid Peptide Alpha-Neoendorphin. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 4347−4350.

(44) Urbano, M.; Guerrero, M.; Rosen, H.; Roberts, E. Antagonists of the Kappa Opioid Receptor. *Bioorg. Med. Chem. Lett.* **2014**, *24*, 2021−2032.

(45) Hamada, Y.; Abdel-Rahman, H.; Yamani, A.; Nguyen, J. T.; Stochaj, M.; Hidaka, K.; Kimura, T.; Hayashi, Y.; Saito, K.; Ishiura, S.; Kiso, Y. Bace1 Inhibitors: Optimization by Replacing the P1′ Residue with Non-Acidic Moiety. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1649−1653.

(46) Hamada, Y.; Tagad, H. D.; Nishimura, Y.; Ishiura, S.; Kiso, Y. Tripeptidic Bace1 Inhibitors Devised by in-Silico Conformational Structure-Based Design. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 1130−1135.

(47) Wang, Y.; Bajorath, J. Balancing the Influence of Molecular Complexity on Fingerprint Similarity Searching. *J. Chem. Inf. Model.* **2008**, *48*, 75−84.

(48) Trujillo, J. I.; Kiefer, J. R.; Huang, W.; Day, J. E.; Moon, J.; Jerome, G. M.; Bono, C. P.; Kornmeier, C. M.; Williams, M. L.; Kuhn, C.; Rennie, G. R.; Wynn, T. A.; Carron, C. P.; Thorarensen, A. Investigation of the Binding Pocket of Human Hematopoietic Prostaglandin (Pg) D2 Synthase (Hh-Pgds): A Tale of Two Waters. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 3795−3799.

(49) Christ, A. N.; Labzin, L.; Bourne, G. T.; Fukunishi, H.; Weber, J. E.; Sweet, M. J.; Smythe, M. L.; Flanagan, J. U. Development and Characterization of New Inhibitors of the Human and Mouse Hematopoietic Prostaglandin D(2) Synthases. *J. Med. Chem.* **2010**, *53*, 5536−5548.

(50) Weber, J. E.; Oakley, A. J.; Christ, A. N.; Clark, A. G.; Hayes, J. D.; Hall, R.; Hume, D. A.; Board, P. G.; Smythe, M. L.; Flanagan, J. U. Identification and Characterisation of New Inhibitors for the Human Hematopoietic Prostaglandin D2 Synthase. *Eur. J. Med. Chem.* **2010**, *45*, 447−454.

(51) Mellor, H.; Parker, P. J. The Extended Protein Kinase C Superfamily. *Biochem. J.* **1998**, *332* (2), 281−292.

(52) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.

(53) Lindskog, S. Structure and Mechanism of Carbonic Anhydrase. *Pharmacol. Ther.* **1997**, *74*, 1−20.

(54) Winum, J. Y.; Carta, F.; Ward, C.; Mullen, P.; Harrison, D.; Langdon, S. P.; Cecchi, A.; Scozzafava, A.; Kunkler, I.; Supuran, C. T. Ureido-Substituted Sulfamates Show Potent Carbonic Anhydrase Ix Inhibitory and Antiproliferative Activities against Breast Cancer Cell Lines. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4681−4685.

(55) Sanguinetti, M. C.; Tristani-Firouzi, M. Herg Potassium Channels and Cardiac Arrhythmia. *Nature* **2006**, *440*, 463−469.

(56) Aronov, A. M.; Goldman, B. B. A Model for Identifying Herg K + Channel Blockers. *Bioorg. Med. Chem.* **2004**, *12*, 2307−2315.

(57) Cavalli, A.; Buonfiglio, R.; Ianni, C.; Masetti, M.; Ceccarini, L.; Caves, R.; Chang, M. W.; Mitcheson, J. S.; Roberti, M.; Recanatini, M. Computational Design and Discovery of "Minimally Structured" Herg Blockers. *J. Med. Chem.* **2012**, *55*, 4010−4014.

(58) Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. Three-Dimensional Quantitative Structure-Activity Relationship for Inhibition of Human Ether-a-Go-Go-Related Gene Potassium Channel. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427−434.

(59) Pearlstein, R. A.; Vaz, R. J.; Kang, J.; Chen, X. L.; Preobrazhenskaya, M.; Shchekotikhin, A. E.; Korolev, A. M.; Lysenkova, L. N.; Miroshnikova, O. V.; Hendrix, J.; Rampe, D. Characterization of Herg Potassium Channel Inhibition Using Comsia 3d Qsar and Homology Modeling Approaches. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829−1835.

(60) Sumiyoshi, T.; Kunugi, H.; Nakagome, K. Serotonin and Dopamine Receptors in Motivational and Cognitive Disturbances of Schizophrenia. *Front. Neurosci.* **2014**, *8*, 395.

(61) Abdelfattah, M. A.; Lehmann, J.; Abadi, A. H. Discovery of Highly Potent and Selective D4 Ligands by Interactive Sar Study. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 5077−5081.

(62) D'Alessandro, P. L.; Corti, C.; Roth, A.; Ugolini, A.; Sava, A.; Montanari, D.; Bianchi, F.; Garland, S. L.; Powney, B.; Koppe, E. L.; Rocheville, M.; Osborne, G.; Perez, P.; de la Fuente, J.; De Los Frailes, M.; Smith, P. W.; Branch, C.; Nash, D.; Watson, S. P. The Identification of Structurally Novel, Selective, Orally Bioavailable Positive Modulators of Mglur2. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 759−762.

(63) Lacivita, E.; Leopoldo, M.; Masotti, A. C.; Inglese, C.; Berardi, F.; Perrone, R.; Ganguly, S.; Jafurulla, M.; Chattopadhyay, A. Synthesis and Characterization of Environment-Sensitive Fluorescent Ligands for Human 5-Ht1a Receptors with 1-Arylpiperazine Structure. *J. Med. Chem.* **2009**, *52*, 7892−7896.

(64) Sagnes, C.; Fournet, G.; Satala, G.; Bojarski, A. J.; Joseph, B. New 1-Arylindoles Based Serotonin 5-Ht7 Antagonists. Synthesis and Binding Evaluation Studies. *Eur. J. Med. Chem.* **2014**, *75*, 159−168.

(65) Sasse, B. C.; Mach, U. R.; Leppaenen, J.; Calmels, T.; Stark, H. Hybrid Approach for the Design of Highly Affine and Selective Dopamine D(3) Receptor Ligands Using Privileged Scaffolds of Biogenic Amine Gpcr Ligands. *Bioorg. Med. Chem.* **2007**, *15*, 7258−7273.

(66) Venkatesan, A. M.; Dos Santos, O.; Ellingboe, J.; Evrard, D. A.; Harrison, B. L.; Smith, D. L.; Scerni, R.; Hornby, G. A.; Schechter, L. E.; Andree, T. H. Novel Benzofuran Derivatives with Dual 5-Ht1a Receptor and Serotonin Transporter Affinity. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 824−827.

(67) Sweis, R. F.; Pliushchev, M.; Brown, P. J.; Guo, J.; Li, F.; Maag, D.; Petros, A. M.; Soni, N. B.; Tse, C.; Vedadi, M.; Michaelides, M. R.; Chiang, G. G.; Pappano, W. N. Discovery and Development of Potent and Selective Inhibitors of Histone Methyltransferase G9a. *ACS Med. Chem. Lett.* **2014**, *5*, 205−209.

(68) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. Pubchem Bioassay: 2014 Update. *Nucleic Acids Res.* **2014**, *42*, D1075−1082.

(69) Verma, S. K.; Tian, X.; LaFrance, L. V.; Duquenne, C.; Suarez, D. P.; Newlander, K. A.; Romeril, S. P.; Burgess, J. L.; Grant, S. W.; Brackley, J. A.; Graves, A. P.; Scherzer, D. A.; Shu, A.; Thompson, C.; Ott, H. M.; Aller, G. S.; Machutta, C. A.; Diaz, E.; Jiang, Y.; Johnson, N. W.; Knight, S. D.; Kruger, R. G.; McCabe, M. T.; Dhanak, D.; Tummino, P. J.; Creasy, C. L.; Miller, W. H. Identification of Potent, Selective, Cell-Active Inhibitors of the Histone Lysine Methyltransferase Ezh2. *ACS Med. Chem. Lett.* **2012**, *3*, 1091−1096.

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

(71) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432*, 855−861.