JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

GES Polypharmacology Fingerprints: A Novel Approach for Drug Repositioning

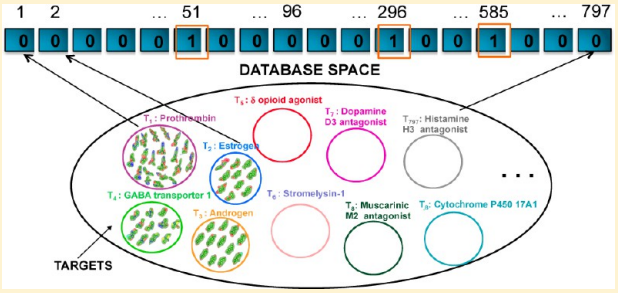
Violeta I. Pérez-Nueno,*[,†] Arnaud S. Karaboga,[†] Michel Souchet,[†] and David W. Ritchie[‡]

[†]Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique, 54600 Villers lès Nancy, France

[‡]INRIA Nancy − Grand Est, 615 rue du Jardin Botanique, 54506 Vandoeuvre lès Nancy, France

Supporting Information

ABSTRACT: Polypharmacology is now recognized as an increasingly important aspect of drug design. We previously introduced the Gaussian ensemble screening (GES) approach to predict relationships between drug classes rapidly without requiring thousands of bootstrap comparisons as in current promiscuity prediction approaches. Here we present the GES "computational polypharmacology fingerprint" (CPF), the first target fingerprint to encode drug promiscuity information. The similarity between the 3D shapes and chemical properties of ligands is calculated using PARAFIT and our HPCC programs to give a consensus shape-plus-chemistry ligand similarity score, and ligand promiscuity for a given set of targets is quantified using the GES fingerprints. To demonstrate our approach, we calculated the CPFs for a set of ligands from DrugBank that are related to some 800 targets. The performance of the approach was measured by comparing our CPF with an in-house "experimental polypharmacology fingerprint" (EPF) built using publicly available experimental data for the targets that comprise the fingerprint. Overall, the GES CPF gives very low fall-out while still giving high precision. We present examples of polypharmacology relationships predicted by our approach that have been experimentally validated. This demonstrates that our CPF approach can successfully describe drug−target relationships and can serve as a novel drug repurposing method for proposing new targets for preclinical compounds and clinical drug candidates.



INTRODUCTION

Polypharmacology describes the binding of a single ligand to multiple protein targets (a promiscuous ligand) or of multiple diverse ligands to a given target (a promiscuous target). Nowadays, the polypharmacology of promiscuous binders is recognized as an important aspect in drug design. For example, a common reason for terminating a drug development program is that the leads are found to be nonselective or promiscuous.[1] Consequently, the in silico prediction of unwanted side effects caused by the promiscuous behavior of drugs and their targets is highly relevant to the pharmaceutical industry, and considerable effort is being put into the screening of a number of suspected off-target proteins in the hope that side effects might be identified early, before the cost associated with developing a drug candidate rises steeply. On the other hand, promiscuity is not always unwelcome, and it can even be exploited for drug development. The use of old drugs for new targets has been shown to provide a promising way to reduce both the time and cost of drug development.[2] Therefore, a better global understanding of drug−disease and drug−target network relationships promises to provide effective and safe repurposing of existing drugs with known targets not only for an individual use but also for use in combination therapies for different diseases.

Although it would be desirable to be able to screen a drug against all proteins expressed by the human genome, this is currently infeasible. Therefore, several computational techniques have been developed to predict in silico the pharmacological profiles of known drugs.[3] In the past decade, in silico chemogenomic strategies have become attractive for relating protein receptors to each other quantitatively using *2D or 3D descriptor spaces*[4] (using, e.g., SEA[5] or topological descriptors[6,7]), *shape similarity* (using, e.g., ROCS,[8] SHAFTS,[9] and our own GES[10] approach), *ligand−receptor pharmacophoric descriptors* (e.g., using P−L fingerprints[11,12]), and *target similarity* (using, e.g., sequence similarity[13] or pharmacophoric binding pocket descriptors[14,15]), and *binding-pocket shapes*.[16] Some recent approaches for predicting polypharmacology combine various chemogenomic techniques.[17] Other approaches have been developed for predicting drug−target interactions using drug side-effect similarity,[18−20] machine learning approaches,[21−23] and complex network theory[24−26]

We previously introduced the 3D Gaussian ensemble screening (GES) approach[10] to predict relationships between drug classes rapidly without requiring thousands of bootstrap comparisons as is normally the case in current promiscuity prediction approaches. Here we present a GES-based "computational polypharmacology fingerprint" (CPF), the first fingerprint to encode promiscuity information. The length

Received: November 18, 2013
Published: February 4, 2014

   dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

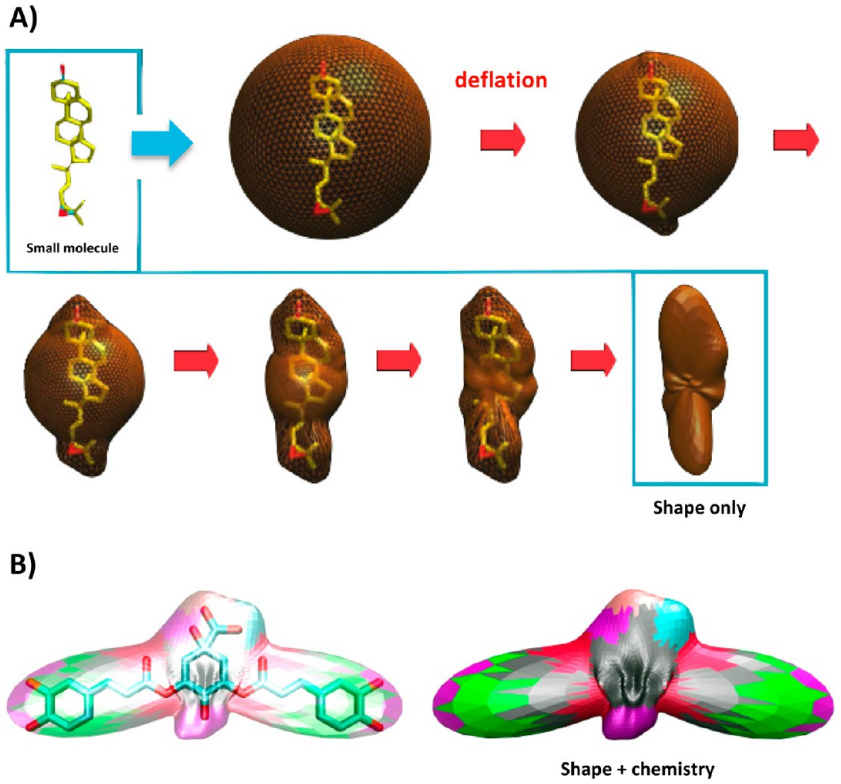


Figure 1. Spherical harmonic (SH) shape-based representations. (A) Basic principle of building an SH-based shape from a chemical structure (the "HPCC shape") using the Harmonic Pharma chemistry coefficient method. (B) Harmonic Pharma chemistry coefficient 3D molecular representation ("HPCC Combo") combining a ligand-centric pharmacophoric description ("HPCC chemistry") projected onto the SH-based shape of a ligand ("HPCC shape"). The projection of the pharmacophoric description onto the SH shape is shown by color: positive charge (red), negative charge (light blue), aromatic (green), hydrophobic (black), donor (blue), acceptor (dark red), acceptor_donor (magenta), positive_donor (tan), negative_acceptor (pink), acceptor_aromatic (yellow), donor_aromatic (dark blue), hydrophobic_aromatic (dark green), negative_acceptor_donor (cyan), and undefined (white).

of this fingerprint depends on the total number of targets for which promiscuity is being investigated. In the present work, our CPF was used to obtain a consensus promiscuity representation based on a comparison of the 3D shapes and chemistry of representative ligands that bind the targets for which the polypharmacology is studied.

Here we demonstrate the GES polypharmacology fingerprint using some 800 known drug targets from the DrugBank database. The performance of the approach was measured by comparing the present CPF with an in-house "experimental polypharmacology fingerprint" (EPF) built using publicly available experimental data for the targets that comprise the fingerprint. Statistical analyses were used to assess the quality of the method and the influence of missing data on its performance. Overall, our CPF gives very low fall-out while still giving high precision. For example, an average sensitivity of ~50% was obtained, and the CPF was able to correctly predict up to ~90% of the experimentally validated polypharmacology relationships in a best-case scenario (i.e., with no missing data).

We present examples of predicted polypharmacology relationships that have been experimentally validated. This demonstrates that GES CPFs can successfully describe drug−target relationships and provide a novel way to propose new targets for preclinical compounds and clinical drug candidates. Overall, our CPFs find interesting relationships between drug families both with and without any obvious biological similarity.

METHODS

Spherical Harmonic Shape-Based Representations. Here PARASURF is used to calculate the molecular shapes and local surface properties of all of the ligands from semiempirical quantum mechanics theory (using CEPOS Mopac[27] with the default Hamiltonian AM1) and to encode these properties as spherical harmonic (SH) expansions. PARAFIT[28] is then used to superpose the SH molecular surfaces by exploiting the special rotational properties of the SH expansions.[29−32] We also use PARAFIT to calculate the average or "consensus" shape of a group of molecules by calculating the average of their SH expansion coefficients.[33] Once an SH consensus shape has been calculated for a group of molecules, it is straightforward to use PARAFIT again to identify the center molecule (CM), that is, the real molecule whose SH surface is closest to that of the consensus shape.[34]

Additionally, the Harmonic Pharma chemistry coefficient (HPCC)[35] is also used. The "HPCC Combo" 3D molecular representation involves a ligand-centric pharmacophoric description ("HPCC chemistry") projected onto the SH-based shape of a ligand ("HPCC shape").[35] "HPCC shape" uses SH expansions to represent the molecular surface by deflating an ellipsoidal mesh around its molecular structure. The SH surface is discretized into 2562 triangle meshes. Each triangle of the SH shape is characterized by a pharmacophoric feature regarding the nearest atom type, and a chemo-type property [number of positive-charge (POS) and negative-charge (NEG) atoms, number of hydrogen-bond acceptor

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

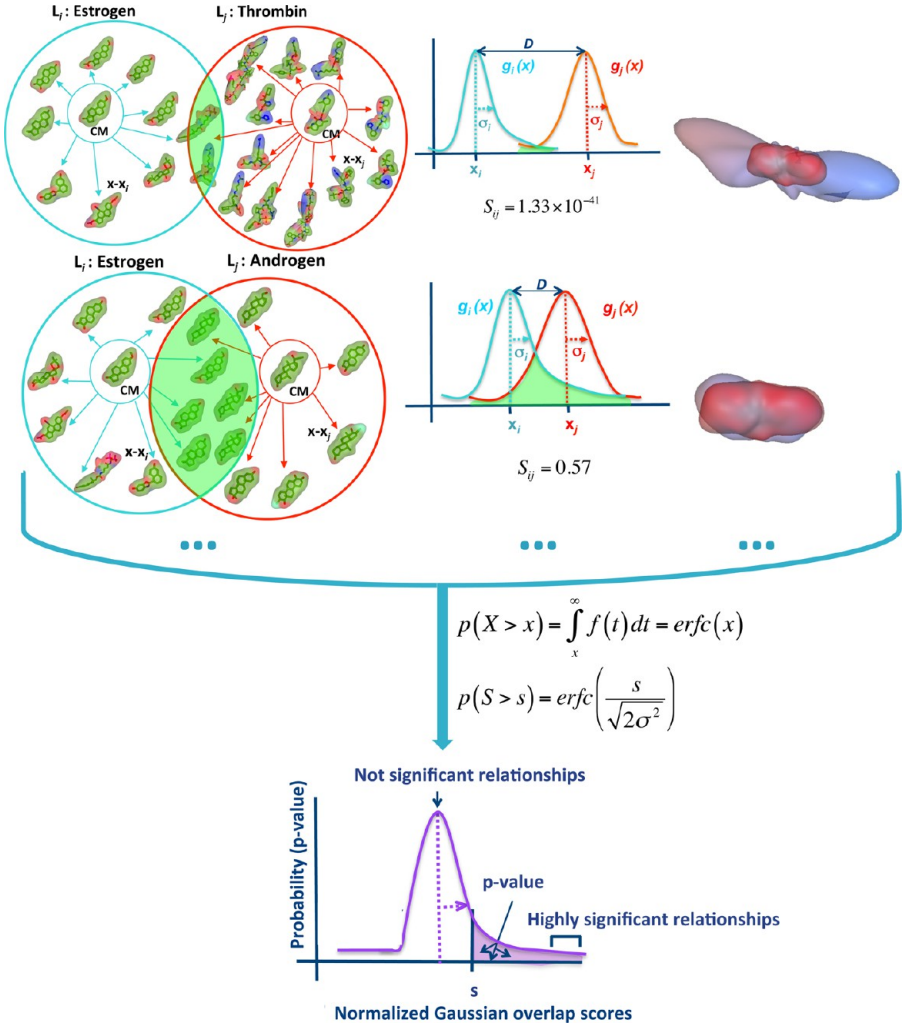


Figure 2. Gaussian ligand set comparisons and Gaussian *p*-values. The similarity between drug classes can be calculated rapidly and reliably by calculating the Gaussian overlap between pairs of such clusters. On the left are shown two examples of Gaussian overlap: a small Gaussian overlap between the estrogen and thrombin ligand sets and a large Gaussian overlap between the estrogen and androgen ligand sets. Each pairwise Gaussian overlap score is transformed into a probability value ("*p*-value") calculated analytically from the scores distribution using standard statistical techniques.

(HBA) and hydrogen-bond donor (HBD) atoms, number of aromatic (ARO) atoms, and number of hydrophobic (HYD) atoms] is assigned to it. Each triangle of molecule A is compared to the nearest triangles of molecule B and scored according to a pharmacophoric similarity matrix ("HPCC chemistry"). Thus, the similarity between molecules A and B is calculated by comparing the distributions of chemo-type values of their respective shapes. Figure 1 shows some example HPCC SH shape-based representations.

Predicting Polypharmacology Using Gaussian Ligand Sets. We define a "ligand set" as a cluster of high-affinity ligands that bind to a specific target. In the GES approach, the main idea is to represent a cluster of molecules as a Gaussian distribution with respect to a selected CM.[10] Using SH shape-based similarity scores, it is straightforward to calculate the CM of the ligand set. However, because Gaussian functions require a distance coordinate rather than a similarity score, we use PARAFIT to calculate the normalized SH *distance* ($0.0 \leq x \leq 1.0$) between the CM and each cluster member. With the assumption that these distances follow a Gaussian distribution, each cluster can be represented as a probability density function $g_i(x)$:

$$g_i(x) = \sqrt{\frac{1}{2\pi\sigma_i^2}} \cdot e^{(x-x_i)^2/2\sigma_i^2} \quad (1)$$

where $|x - x_i|$ represents the distance from the ith CM and σ_i is the standard deviation (SD) of the member distances.

We then define a desired number of ligand sets, each one representing a given target according to the possible promiscuity that is to be investigated. We can also define several ligand sets for a single target if we are dealing with a drug family with quite different ligand scaffolds. We then represent each ligand set as a Gaussian distribution, as explained above. A Hodgkin-like similarity score S_{ij} between two distributions may then be expressed as

$$S_{ij} = \frac{2\int_{-\infty}^{+\infty} g_i(x) \cdot g_j(x)\, dx}{\int_{-\infty}^{+\infty} g_i(x)^2\, dx + \int_{-\infty}^{+\infty} g_j(x)^2\, dx} \quad (2)$$

The Gaussian overlap integrals can be simplified to a closed expression involving only the Gaussian widths and the distance between the CMs using standard techniques. With $a = \sigma_i^2/2$ and $b = \sigma_j^2/2$, it can then be shown that

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

$$S_{ij} = \frac{2^{3/2} \cdot \left(\frac{a \cdot b}{a+b}\right)^{1/2} \cdot e^{-\left(\frac{a \cdot b}{a+b}\right) \cdot x_{ij}^{2}}}{a^{1/2} + b^{1/2}} \quad (3)$$

where x_{ij} is the distance between the CMs of clusters i and j. Thus, it is straightforward to calculate a matrix of similarity scores between all of the ligand set clusters (i.e., to perform "all-versus-all" cluster comparisons[10]). It is worth noting that this cluster similarity score depends only on the similarity of pairs of CMs and the SDs of their respective clusters. It does not depend on the number of members in each cluster.

Quantifying Ligand Set Scores Using Gaussian *p*-Values. In order to transform a list of cluster similarity scores into a more meaningful list of probabilities, we calculate the distribution of all pairwise Gaussian overlap scores and fit the distribution to a Gaussian probability density function $f(t)$.[10] A probability value, or "p-value", is calculated analytically from the scores distribution using standard statistical techniques. For example, for a Gaussian distribution, it can be shown that the probability of finding at random from the distribution some value X that is greater than a given value x is given by

$$p(X > x) = \int_{x}^{\infty} f(t)\, dt = \mathrm{erfc}(x) \quad (4)$$

where $f(t)$ is the standard normalized Gaussian probability density function and $\mathrm{erfc}(x)$ is the complementary error function. Hence, for a normalized distribution of scores, we obtain

$$p(S > s) = \mathrm{erfc}\left(\frac{s}{\sqrt{2\sigma^2}}\right) \quad (5)$$

where σ is the SD of the fitted Gaussian. In other words, the p-value for a given score s represents the probability of finding at random from the distribution some other score S that is greater than s. Figure 2 illustrates the Gaussian overlap between ligand sets and the calculation of Gaussian p-values.

The plots in Supplementary Figures 1 and 2 in the Supporting Information show the observed distributions of pairwise ligand set cluster scores for PARAFIT and HPCC (Supplementary Figures 1a and 2a), the fitted Gaussian functions with $\sigma = 0.059159$ and $\sigma = 0.056350$ for PARAFIT and HPCC, respectively (Supplementary Figures 1b and 2b), and the p-values calculated at each bin using the fitted distributions (Supplementary Figures 1c and 2c).

GES Polypharmacology Fingerprints. Each bit of the CPF corresponds to one target represented by a known high-affinity ligand or ligand set (what we call a "target-ligand set"). If a bit is set, the ligand for which the CPF is calculated is predicted to bind the corresponding target. Each bit position is set on or off depending on the existence or nonexistence of 3D shape and chemical similarity between the ligand for which the fingerprint is calculated and the CM of the target-ligand set of the corresponding bit position. 3D shape and chemical similarity are measured with PARAFIT and HPCC, respectively, and the possible promiscuity is quantified using GES. The p-values are transformed into CPF bit values according to a given p-value threshold. For our predictions, we consider relationships to be relevant if their p-values are less than 10^{-40}. Table 1 assigns approximate labels ("significant", "highly significant", etc.) to various ranges of p-values. As explained in the Results and Discussion, if very few or no relationships are found, we also consider relationships with $p \leq 10^{-30}$, which we call "interesting" relationships. Figure 3 shows a schematic representation of the GES polypharmacology fingerprint.

Table 1. Ranges of *p*-Values

range no.	p-value range	label
1	10^{-60} to 10^{-70}	highly significant
2	10^{-50} to 10^{-60}	very significant
3	10^{-40} to 10^{-50}	significant
4	10^{-30} to 10^{-40}	interesting
5	10^{-20} to 10^{-30}	possible
6	10^{-10} to 10^{-20}	possible
7	10^{0} to 10^{-10}	not significant

Data Preparation. We extracted all 6711 drug entries from the DrugBank database,[36] and the SH surfaces and SH shape-plus-chemistry representations of these drugs were calculated using PARASURF and HPCC. These drugs were grouped in 932 folders according to the targets they bind. After removal of peptide drugs and some entries for which PARASURF failed, 4757 ligands were distributed into 781 target folders. Two hundred conformations for each molecule were computed, and PARAFIT and HPCC representations were calculated for each conformation. We then used the CAST clustering algorithm[37] to cluster the members of each target folder using a PARAFIT Tanimoto similarity score of 0.65, as described previously.[10] This gave 797 shape clusters, of which 24 contained singleton ligands. The resulting 781 target folders mainly remained unsplit, except for 16 that had substantially different scaffold ligands. We then calculated the consensus shape and the CM for each cluster. We call each of the resulting clusters a "ligand set". The similarities between ligand sets were calculated using the Gaussian overlap score described above for both the PARAFIT and HPCC similarity measures. The results from the all-versus-all ligand set comparisons were recorded as a matrix of GES p-values. CPFs (with a length of 797 bits) were then calculated for both PARAFIT and HPCC for each of the CM of the ligand sets.

In order to analyze the performance of such CPFs and to select a good p-value threshold to set the bit values consistently, we compared the matches between our computational predictions and a hand-curated EPF built using publicly available experimental data from the PubChem (targets + enzymes), ChEMBL, and BindingDB databases for each of the 797 target-ligand set CMs that comprise the fingerprint. More specifically, an EPF bit was set whenever a ligand and its corresponding target were found to be related either (i) in PubChem according to the ligand "Biomolecular Interactions and Pathways" when the field "Outcome" has the value "Active", (ii) in BindingDB when it is marked as "Active", or (iii) in ChEMBL when the field "Organism" is "*Homo sapiens*" and the field "Activity comment" is one of "active", "inhibitor", "agonist", "antagonist", "partial agonist", "unspecific", "inhibitor [30% of Control]", "inhibitor", or "substrate". We excluded relationships with an "Activity comment" field value of "toxic", "not active", "inactive", "not determined", or "inconclusive". Compounds with activity values of <100 μM were considered as "Active" (IC50, EC50, Potency, AC50, and K_i values were used). The experimental information downloaded for the DrugBank compounds is provided in the Supplementary_Data folder in the Supporting Information.

It is worth mentioning that some care is required when searching for relationships in the experimental databases. Different databases use different ways to quantify the activity

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

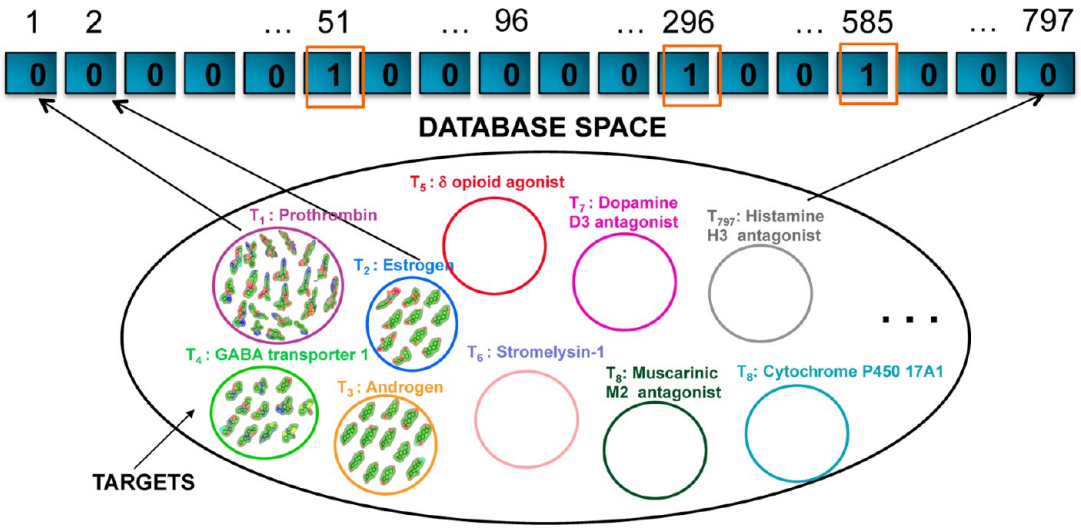


Figure 3. Schematic representation of the GES polypharmacology fingerprint. This figure illustrates a GES polypharmacology fingerprint calculated for ∼800 targets that have DrugBank ligands. The number above each fingerprint bit position is a numerical identifier assigned to a particular target. The compound represented by this example polypharmacology fingerprint is a promiscuous ligand that binds three different targets, namely, targets 51, 296, and 585.

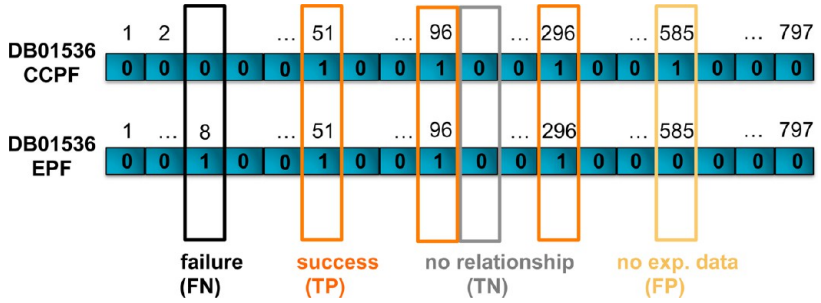


Figure 4. Computing confusion matrix statistics for the GES polypharmacology fingerprint. Shown are schematic representation of the CCPF and EPF for the DrugBank compound "DB01536". As in Figure 3, the number above each fingerprint bit is a numerical target identifier. When a computational fingerprint bit is zero but the corresponding experimental bit is set, the computational method was not able to predict the experimentally validated target relationship, so this represents a false negative (FN). When both the computational and experimental bits are set, the computational method was able to predict the experimentally validated target relationship, so this represents a true positive (TP). Conversely, when both bits are zero, neither the computational nor the experimental data account for a target relationship, so this represents a true negative (TN). Finally, when the computational bit is set but the corresponding experimental bit is zero, the computational method predicts a target relationship but a relationship is not experimentally validated, so this represents a false positive (FP). This last case can arise from either a genuine FP prediction or a correct computational prediction that currently lacks experimental evidence to support it.

of a ligand (such as the labels mentioned above) and often use different activity units even within the same database. For this reason, internal thresholds regarding the degree of activity (low, medium, or highly active) according to the "Activity value" were not imposed. Similar care is required when searching the "Target name" field. All synonyms must be checked and, if necessary, corrected manually (e.g., by looking at the "BioAssay" explanation field).

CPFs were first calculated for PARAFIT and HPCC individually using different *p*-value thresholds to set the bit values. Confusion matrix statistics (see Table 2) were calculated to select the *p*-value thresholds for which the PARAFIT and HPCC CPFs agree most closely with the hand-curated EPF. A combination of the PARAFIT and HPCC methods using the best *p*-value thresholds was then used to define a "combined CPF" (CCPF).

Figure 4 shows a schematic representation of the CCPF and EPF for the DrugBank compound "DB01536" and their comparison using confusion matrix statistics. As shown in the figure, the CCPF calculated for this ligand predicts that it can bind to the targets 6-deoxyerythronolide B hydroxylase, aldo-keto reductase family 1 member C3, estradiol 17-β-dehydrogenase 1, and prolactin receptor (corresponding to the bit positions 51, 96, 296, and 585, respectively). If we compare this prediction with the EPF, we can see that this ligand is experimentally related to the targets 3-β-hydroxysteroid dehydrogenase/Δ 5→4 isomerase type II, 6-deoxyerythronolide B hydroxylase, aldo-keto reductase family 1 member C3, and estradiol 17-β-dehydrogenase (bit positions 8, 51, 96, and 296). To calculate the confusion matrix, we consider cases where both the computational and experimental bits are set to be true positive (TP) predictions (shown in orange in the example). Conversely, when both bits are zero, we take it as a true negative (TN) prediction (shown in gray in the example). On the other hand, when the computational bit is zero but the corresponding experimental bit is set, this represents a false negative (FN) prediction (shown in black). Finally, a false positive (FP) prediction occurs when the computational bit is set but the corresponding experimental bit is zero. It is worth mentioning that this last case can arise from either a genuine FP

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

prediction or a correct computational prediction that currently lacks experimental evidence to support it.

RESULTS AND DISCUSSION

GES Polypharmacology Fingerprint Threshold. Confusion matrix statistics were calculated for both HPCC and PARAFIT to define the *p*-value threshold that best relates a CPF to the corresponding EPF. Definitions of the statistical measures used here are provided in Table 2. Table 3 compares the CCPFs for HPCC and PARAFIT using the *p*-value threshold ranges listed in Table 1 against the EPF obtained from the PubChem targets, PubChem enzymes, ChEMBL, and BindingDB databases, which we subsequently call the "full EPF" (FEPF). It can be seen that for HPCC, *p*-value ranges 2 and 3 give the highest sensitivity with the lowest fall-out and false discovery rate (FDR). After *p*-value range 3, the FDR rises steeply without much improvement in the sensitivity. PARAFIT shows a similar trend in which *p*-value ranges 1 and 2 give the highest sensitivity with the lowest fall-out and FDR. Beyond *p*-value range 2, the FDR increases rapidly without a significant improvement in the sensitivity. Moreover, in all of these cases we obtain a high Matthews correlation coefficient (MCC). The MCC gives a balanced measure of accuracy even when the numbers of TPs and TNs differ greatly, as is the case here. Hence, in order to obtain the lowest fall-out with the highest sensitivity, it seems appropriate to apply *p*-value range 2 or 3 when using HPCC and *p*-value range 1 or 2 when using PARAFIT. We note here that it is interesting to retain some of the FPs corresponding to the "possible polypharmacology to explore" category because examples in this band could be particularly fruitful to consider in more detail.

Comparing CPFs and EPFs. Combinations of the PARAFIT and HPCC scores using the *p*-value thresholds defined above were used to define a "combined CPF" (CCPF). According to the desired degree of permissiveness in retrieving FPs, the CCPF is divided into three categories: (i) "tolerant" (using HPCC *p*-value range 3 + PARAFIT *p*-value range 2), which allows a high FDR with a high sensitivity; (ii) "compromise" (using HPCC range 3 + PARAFIT range 1), which keeps a high sensitivity with a lower FDR; and (iii) "restrictive" (using HPCC range 2 + PARAFIT range 1), which gives the lowest FDR but with somewhat lower sensitivity.

Table 4 shows an analysis of the performance of the CCPF according to these categories. It can be seen that the "compromise" and "restrictive" fingerprints give the lowest FDR with high sensitivity. With the assumption that it could be useful to test some FP predictions experimentally, one possible scenario would be to select a combination of methods and *p*-value thresholds that returns many predictions in this category without including too many real FPs. In order to be conservative, we therefore use the "restrictive" fingerprint, with which the predicted relationships between targets should have a high probability to be true [low fall-out and FDR (0.02% and 4.23%, respectively), high MCC (0.7), and high precision (95.8% in the example in Table 4)], although some relationships will not be found (46.1% sensitivity). On the other hand, if we wish to find novel associated biological target(s) and modes of action (MoAs) for safe clinical or approved drugs or to identify novel biological targets and MoAs for the early characterization of preclinical compounds, our aim is to find high-confidence positive predictions, and it is not a great inconvenience if the method does not retrieve all possible

Table 2. Statistical Measures Used for the Performance Analysis of the GES Polypharmacology Fingerprint

statistical measure	definition
sensitivity = TP/(TP + FN)	True positive rate (TPR) or recall. The proportion of experimentally validated polypharmacology relationships that are correctly predicted as such.
fall-out = FP/(FP + TN)	False positive rate (FPR) or 1-specificity. The number of incorrect positive results occurring among all negative samples available during the test. It can be viewed as the probability that an FP is retrieved by the query.
precision = TP/(TP + FP)	Positive predictive value (PPV). The proportion of predicted polypharmacology relationships that are real polypharmacology relationships according to experimental data.
F1 score = 2·TP/(2·TP + FP + FN)	Harmonic mean of precision and recall. Measure of a test's accuracy. The F1 score can be interpreted as a weighted average of the precision and recall, and it reaches its best value at 1 and its worst value at 0.
Matthews correlation coefficient $(MCC) = (TP \cdot TN - FP \cdot FN)/(PNP'N')^{1/2}$ where $P = TP + FN$, $N = FP + TN$, $P' = TP + FP$, and $N' = FN + TN$	Correlation coefficient between the observed and predicted binary classifications. MCC ranges from −1 to +1: a value of +1 represents a perfect prediction; a value of 0 indicates a prediction no better than random prediction; and a value of −1 indicates total disagreement between prediction and observation. It is generally regarded as a balanced measure that can be used even if the classes are of very different sizes.
accuracy = (TP + TN)/[(TP + FN) + (FP + TN)]	Degree of closeness of the polypharmacology prediction to the real experimental value. Not a useful measure when the two classes are of very different sizes.
specificity = TN/(FP + TN)	True negative rate (TNR). The proportion of no-polypharmacology-relationship results according to experimental information (no data + no relationships) that are correctly identified/predicted as such.
negative predictive value (NPV) = TN/(TN + FN)	The proportion of predicted no-polypharmacology-relationship results that are real no-polypharmacology-relationship results according to experimental data.
false discovery rate (FDR) = FP/(FP + TP)	Possible "no experimental data" or FP in the polypharmacology prediction. Takes into account the expected proportion of incorrectly rejected null hypotheses (no relationship). Gives an idea of the possible polypharmacology to explore.

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

Table 3. Analysis of the Performance of the CPFs for the HPCC and PARAFIT Methods with Different *p*-Value Ranges Using the FEPF Obtained from the Compilation of the Experimental Data in the PubChem Targets, PubChem Enzymes, ChEMBL, and BindingDB Databases

	HPCC				PARAFIT			
statistical measure	*p*-value range 1	*p*-value range 2	*p*-value range 3	*p*-value range 4	*p*-value range 1	*p*-value range 2	*p*-value range 3	*p*-value range 4
TPR (%)	45.000	**45.800**	**47.300**	50.100	**42.000**	**47.100**	51.400	55.700
FPR (%)	0.007	**0.010**	**0.070**	0.300	**0.006**	**0.100**	0.900	3.000
precision (%)	98.300	96.300	85.600	58.300	98.300	74.200	34.000	16.000
F1 score	0.600	0.600	0.600	0.500	0.600	0.600	0.400	0.200
MCC	0.700	**0.700**	**0.600**	0.500	**0.600**	**0.600**	0.400	0.300
accuracy (%)	99.500	99.500	99.500	99.300	99.500	99.400	98.700	97.100
specificity (%)	100.000	100.000	99.900	99.700	100.000	99.900	99.100	97.500
NPV (%)	99.500	99.500	99.600	99.600	99.500	99.500	99.600	99.600
FDR (%)	1.700	**3.660**	**14.400**	41.700	**1.690**	**25.800**	66.000	84.000

Table 4. Analysis of the Performance of the CCPF with Different *p*-Value Ranges Using the FEPF; The CCPF Performance Is Divided into Three Cases According to the Degree of Permissiveness in Obtaining Higher Numbers of FPs (Possible Polypharmacology To Explore)

statistical measure	"tolerant" CCPF[a]	"compromise" CCPF[b]	"restrictive" CCPF[c]
TPR (%)	50.00	**47.60**	**46.10**
FPR (%)	0.20	**0.07**	**0.02**
precision (%)	69.60	**85.40**	**95.80**
F1 score	0.60	0.60	0.60
MCC	0.60	**0.60**	**0.70**
accuracy (%)	99.40	99.50	99.50
specificity (%)	99.80	99.90	100.00
NPV (%)	99.60	99.60	99.50
FDR (%)	30.40	**14.60**	**4.23**

[a]HPCC *p*-value range 3 + PARAFIT *p*-value range 2. [b]HPCC *p*-value range 3 + PARAFIT *p*-value range 1. [c]HPCC *p*-value range 2 + PARAFIT *p*-value range 1.

target relationships. If the "restrictive" fingerprint gives very few or no target relationships, it is reasonable to use the "compromise" fingerprint to explore more exhaustively all of the possibilities.

If we compare the "restrictive" CCPF (Table 4) with HPCC *p*-value range 2 alone (Table 3), HPCC *p*-value range 2 retrieves fewer FPs (3.66% vs 4.23% for the "restrictive" CCPF) and higher precision (96.2% vs 95.8% for the "restrictive" CCPF) with almost no fall-out (0.01% vs 0.02% for the "restrictive" CCPF) and just slightly less sensitivity (45.8% vs 46.1% for the "restrictive" CCPF). Hence, it seems that adding PARAFIT *p*-value range 1 to HPCC *p*-value range 2 does not substantially improve a CCPF. However, the "restrictive" CCPF could find further polypharmacology relationships because it includes a different target-ligand set similarity comparison method that might retrieve somewhat different results.

The example in Table 4 compares the CCPF with the FEPF. The same pattern can be observed if we compare the CCPF with the EPFs obtained from the individual databases: PubChem targets (Supplementary Table 1), PubChem enzymes (Supplementary Table 2), PubChem targets + PubChem enzymes (Supplementary Table 3), ChEMBL (Supplementary Table 4), and BindingDB (Supplementary Table 5). It can be seen that the statistical measures can vary when dealing with experimental information from small databases such as PubChem enzymes, ChEMBL, and BindingDB. For example, the sensitivity increases for these databases because they have smaller all-versus-all matrices, and we can predict all that is in the experimental database (i.e., sensitivities of 80%, 76.7%, and 66.6% for the "restrictive" CCPF compared with the PubChem enzymes EPF, the ChEMBL EPF, and the BindingDB EPF, respectively). On the other hand, the FDR increases for these databases because there is much missing experimental data in the all-versus-all matrix that cannot be compared with the prediction (i.e., FDRs of 68.8%, 69.2%, and 56.6% for the "restrictive" CCPF compared with the PubChem enzymes EPF, the ChEMBL EPF, and the BindingDB EPF, respectively).

Dealing with Missing Data. Overall, our method returns very few FPs (and consequently has very high precision and very low FDR and fall-out). In regard to sensitivity, it gives a moderate value of ~50%, that is, the method does not find all of the experimentally determined TPs (there are a total of 5400 experimentally determined TPs in a total of 635 209 relationships). However, this level of sensitivity is similar to that obtained using other polypharmacology approaches. For example, Lounkine et al.[38] used the SEA[5] to predict the activities of 656 marketed drugs on 73 unintended side-effect targets, and they reported that approximately half of the predictions were confirmed either from other proprietary databases or by new experimental assays (the affinities for these new off-targets ranged from 1 nM to 30 μM). Meslamani et al.[17] reported a mean FPR of 50%, which is also similar to that reported by SEA.

It is reasonable to suppose that the sensitivity of our CPF approach could be improved if we could take into account the fact that the data in the experimental databases is far from complete. In order to test this assumption, we performed an experiment that involved "correcting" the FPRs obtained for the small ChEMBL, PubChem enzymes, and BindingDB databases by using activity data from the larger and more complete PubChem targets database. This "FP correction" involves reassigning an FP calculated for a small database as a TP whenever any member of the corresponding ligand set is annotated as "active" in the larger PubChem targets database. This procedure has the effect of increasing the calculated sensitivity of the method when working with smaller databases. The results are shown in Tables 5, 6, and 7. For example, comparison of the CPF made using HPCC *p*-value range 2 with the FEPF gives approximately the same number of FPs as the comparisons of this CPF with the EPFs for the ChEMBL, PubChem enzymes, and BindingDB databases, but this CPF

Table 5. Statistical Measures for the "Compromise" and "Restrictive" CCPF Methods and the Best Individual CPF Method (HPCC *p*-Value Range 2) Compared with the FP-Corrected BindingDB Database

statistical measure	"compromise" CCPF[a]	"restrictive" CCPF[b]	CPF for HPCC *p*-value range 2
TPR (%)	**82.30**	**81.50**	**81.30**
FPR (%)	**0.10**	**0.00**	**0.00**
precision (%)	**85.30**	**95.70**	**96.30**
F1 score	0.80	0.90	0.90
MCC	**0.80**	**0.90**	**0.90**
accuracy (%)	99.80	99.90	99.90
specificity (%)	99.90	100.00	100.00
NPV (%)	99.90	99.90	99.90
FDR (%)	**14.70**	**4.30**	**3.70**
% of FPs in BindingDB found as TPs in PubChem targets	76.30	92.50	93.40

[a]HPCC *p*-value range 3 + PARAFIT *p*-value range 1. [b]HPCC *p*-value range 2 + PARAFIT *p*-value range 1.

Table 6. Statistical Measures for the "Compromise" and "Restrictive" CCPF Methods and the Best Individual CPF Method (HPCC *p*-Value Range 2) Compared with the FP-Corrected ChEMBL Database

statistical measure	"compromise" CCPF[a]	"restrictive" CCPF[b]	CPF for HPCC *p*-value range 2
TPR (%)	**91.40**	**91.10**	**91.00**
FPR (%)	**0.10**	**0.00**	**0.00**
precision (%)	**85.30**	**95.70**	**96.30**
F1 score	0.90	0.90	0.90
MCC	**0.90**	**0.90**	**0.90**
accuracy (%)	99.90	99.90	99.90
specificity (%)	99.90	100.00	100.00
NPV (%)	100.00	100.00	100.00
FDR (%)	**14.70**	**4.30**	**3.70**
% of FPs in ChEMBL found as TPs in PubChem targets	79.90	93.70	94.60

[a]HPCC *p*-value range 3 + PARAFIT *p*-value range 1. [b]HPCC *p*-value range 2 + PARAFIT *p*-value range 1.

Table 7. Statistical Measures for the "Compromise" and "Restrictive" CCPF Methods and the Best Individual CPF Method (HPCC *p*-Value Range 2) Compared with the FP-Corrected PubChem Enzymes Database

statistical measure	"compromise" CCPF[a]	"restrictive" CCPF[b]	CPF for HPCC *p*-value range 2
TPR (%)	**92.70**	**92.50**	**92.40**
FPR (%)	**0.10**	**0.00**	**0.00**
precision (%)	**85.30**	**95.70**	**96.30**
F1 score	0.90	0.90	0.90
MCC	**0.90**	**0.90**	**0.90**
accuracy (%)	99.90	100.00	100.00
specificity (%)	99.90	100.00	100.00
NPV (%)	100.00	100.00	100.00
FDR (%)	**14.70**	**4.30**	**3.70**
% of FPs in PubChem enzymes found as TPs in PubChem targets	79.80	93.80	94.60

[a]HPCC *p*-value range 3 + PARAFIT *p*-value range 1. [b]HPCC *p*-value range 2 + PARAFIT *p*-value range 1.

gives a significantly higher sensitivity when the small databases are FP-corrected using PubChem targets. The FDR for comparison of HPCC *p*-value range 2 with the FEPF is 3.66% (Table 3), and the FDR is also 3.7% for comparisons of HPCC *p*-value range 2 with the FP-corrected BindingDB (Table 5), ChEMBL (Table 6), and PubChem enzymes (Table 7) databases. However, the sensitivity increases greatly from 45.8% for comparison of HPCC *p*-value range 2 with the FEPF (Table 3) to 81.3%, 91%, and 92.4% for comparisons of this CPF with the FP-corrected BindingDB (Table 5), ChEMBL (Table 6), and PubChem enzymes (Table 7) databases, respectively.

Regarding the possible existence of polypharmacology in DrugBank, Hu and Bajorath[39] calculated the average number of targets for compounds from ChEMBL, PubChem, and DrugBank and the corresponding statistics for promiscuous compounds. According to their analysis, the promiscuous drugs in DrugBank have the highest average number of targets per compound (6.9 for approved drugs and 4.7 for experimental drugs), while in ChEMBL and PubChem the average numbers of targets per promiscuous active compound are 2.9 and 3.7, respectively. For both approved and experimental DrugBank drugs and different activity measurements, the probability of a compound being active against two or more targets or more than five targets is reported. The probability of promiscuity of approved drugs from DrugBank is ∼84%, and the probability to interact with more than five targets is still ∼37%. For experimental drugs, the corresponding probabilities are lower (only ∼24% and ∼3%, respectively). These percentages are also higher than the probabilities of promiscuity in the smaller databases. For a ChEMBL compound with available IC_{50} and K_i measurements, the current probabilities of activity against two or more targets are ∼25% and ∼38%, respectively. However, for activity against more than five targets, the probabilities fall to only ∼1%. Similar observations are made for confirmed PubChem screening hits, for which the probabilities of activity against two or more targets and more than five targets are ∼51% and ∼8%, respectively.

Mestres et al.[40] show similar results, and they report that the average numbers of two targets per drug derived from DrugBank and three targets per drug when supplemented with Wombat increased to six targets per drug when they used all of the experimental drug−target interaction data available at that time, and this value may go up to 13 targets per drug when the prediction of an in silico target profiling method is taken into account.

Using our CCPF with the FEPF gives an FDR of 14.6% for the "compromise" CCPF, which is similar to the rate reported by Mestres et al. (projected polypharmacology using in silico methods), and an FDR of 4.23% for the "restrictive" CCPF, which is similar to the rate reported by Hu and Bajorath (i.e., "4.7 average polypharmacology" value for DrugBank and 3.4% probability of an experimental DrugBank drug to be active against more than five targets). Moreover, if we look at the results for the "restrictive" CCPF versus the EPFs obtained for the experimental PubChem targets and PubChem targets + enzymes databases (i.e., databases with less "missing data"), the FDRs obtained are 4.31% and 4.27%, respectively (see Supplementary Tables 1 and 3). Hu and Bajorath report an averaged promiscuity of 3.7 targets per compound for PubChem and a probability of 7.6% for an experimental DrugBank drug to be active against more than five targets. Hence, this indicates that the GES polypharmacology approach

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

retrieves very few confirmed FPs and that the FDR captures rather well ligand sets in the "possible polypharmacology to explore" category.

According to the possible existence of polypharmacology in DrugBank mentioned above, and in order to take into account missing data when analyzing the performance of GES fingerprints, we used a receiver operating characteristic (ROC) plot (Figure 5). Several methods have been proposed

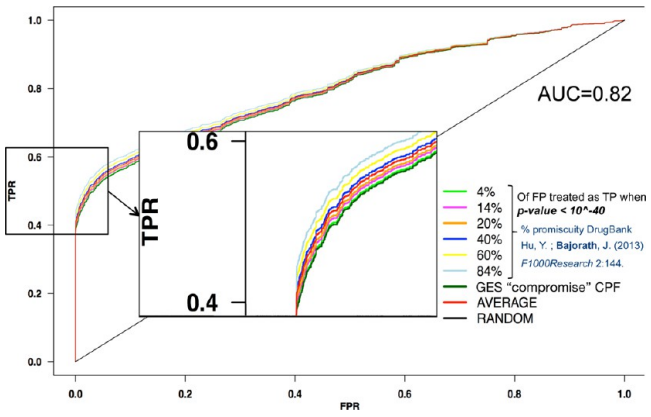


Figure 5. ROC plot representing 316 410 target relationships calculated using the GES "compromise" CPF for ~5000 DrugBank ligands belonging to ~800 targets (dark-green curve). Also shown are average curves when randomly 4% (minimum probability of promiscuity for DrugBank drugs as reported by Hu and Bajorath), 14% (FDR found with the "compromise" CPF), and 20%, 40%, 60%, or 84% (which correspond to the probabilities of promiscuity for DrugBank drugs reported by Hu and Bajorath) of FPs were treated as TPs, always with a p-value of $<10^{-40}$ ("compromise" CPF), in order to take into account the "missing data" when analyzing the performance of GES CPF. An average of all the curves (red) is also shown.

in the literature to deal with missing data (e.g., see ref 41 and references therein), and some of these use ROC analyses.[42−44] However, as far as we know, the present work is the first to use an ROC analysis in the context of polypharmacology. In order to do this, a rank-by-rank consensus was applied to the GES CPF relationships calculated for all 4757 DrugBank compounds belonging to 781 targets using HPCC and PARAFIT to obtain 316 410 consensus-ranked relationships. Then the relationships found for which there is experimental confirmation were treated as "actives" (i.e., TPs), and all of the other relationships were treated as "inactives". The ROC plots were calculated as TPR (y axis) against FPR (x axis) for various p-values ("compromise" CPF curve in Figure 5), and average curves were calculated when various percentages of FPs were introduced (always such that $p < 10^{-40}$, corresponding to the "compromise" CPF). The percentages of FPs used here were 4% (which corresponds to the lowest probability of promiscuity for DrugBank drugs reported by Hu and Bajorath), 14% (FDR found with the "compromise" CPF), and 20%, 40%, 60%, and 84% (which again correspond to DrugBank drug promiscuity probabilities reported by Hu and Bajorath). The average of all the curves was also calculated. The area under the curve (AUC) for the average of all curves was 0.82, while the AUC for the GES "compromise" CPF was 0.80 and that for the 84% curve was 0.85. It is worth noting that the difference between the AUC for the best-case scenario when 84% of the FPs are FP-corrected to TPs (84% curve) and that for the actual curve (GES "compromise" CPF curve) is not substantial given the low number of FPs retrieved by our approach.

Best- and Worst-Case Scenarios. Tables 8 and 9 show two examples of the effect that missing data can have on any conclusions that can be drawn from a polypharmacology prediction. Table 8 shows statistical measures using the "restrictive" CCPF applied to several DrugBank targets in the best-case scenario, namely, targets having much experimental data and several computationally predicted relationships. For these targets, we obtain the highest sensitivity (95.8%) and precision (100%) and lowest fall-out (0%). It is worth noting that all of these targets belong to the same family (i.e., they are all already related). Table 9 shows statistical measures using the "restrictive" CCPF applied to several DrugBank targets in the worst-case scenario, namely, targets with little experimental data but many computationally predicted relationships. For these targets, we obtain the lowest precision (ranging from 12.5% for tripartite-motif-containing protein 13 up to 40% for high-affinity cAMP-specific and IBMX-insensitive 3′,5′-cyclic phosphodiesterase 8A), the highest fall-out (ranging from 0.9% for tripartite-motif-containing protein 13 up to 0.3% for aldo-keto reductase family 1 member C2), and the highest FDR (ranging from 87.5% for tripartite-motif-containing protein 13 up to 60% for high-affinity cAMP-specific and IBMX-insensitive 3′,5′-cyclic phosphodiesterase 8A).

Examples of Experimentally Validated Polypharmacology Predictions. Table 10 lists those targets predicted to be highly related for which the relationships are corroborated by experimental data. Some typical examples of SH shape superpositions of the CMs of these strongly related targets are also shown. The data in this table were extracted from the most significantly related targets found for the "restrictive" CCPF

Table 8. Statistical Measures Using the "Restrictive" CCPF Applied to Several DrugBank Targets in the Best-Case Scenario, with Targets Having the Most Experimental Data and Also the Most Computationally Predicted Relationships

statistical method	T_00054_C1 72 kDa type IV collagenase	T_00220_C1 collagenase 3	T_00485_C1 interstitial collagenase	T_00505_C1 macrophage metalloelastase	T_00512_C1 matrilysin	T_00526_C1 matrix metallo-proteinase-9	T_00588_C1 neutrophil collagenase	T_00796_C1 stromelysin-1
TPR (%)	**95.8**	**95.8**	**95.8**	**95.8**	**95.8**	**95.8**	**95.8**	**95.8**
FPR (%)	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
precision (%)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
F1 score	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MCC	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
accuracy (%)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
specificity (%)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
NPV (%)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
FDR (%)	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

Table 9. Statistical Measures Using the "Restrictive" CCPF Applied to Several DrugBank Targets in the Worst-Case Scenario, with Targets Having the Least Amount of Experimental Data and the Most Computationally Predicted Relationships

statistical method	T_00094_C1 aldo-keto reductase family 1 member C2	T_00219_C1 cocaine- and amphetamine-regulated transcript protein	T_00234_C1 cytochrome P450 17A1	T_00326_C1 estrogen sulfotransferase	T_00429_C1 high-affinity cAMP-specific and IBMX-insensitive 3′,5′-cyclic phosphodiesterase 8A	T_00430_C1 high-affinity cAMP-specific and IBMX-insensitive 3′,5′-cyclic phosphodiesterase 8B	T_00841_C1 tripartite-motif-containing protein 13
TPR (%)	**100.0**	**20.0**	**25.0**	**50.0**	**20.0**	**20.0**	**20.0**
FPR (%)	**0.3**	**0.3**	**0.8**	**0.5**	**0.4**	**0.4**	**0.9**
precision (%)	**33.3**	**33.3**	**14.3**	**20.0**	**40.0**	**40.0**	**12.5**
F1 score	0.5	0.3	0.2	0.3	0.3	0.3	0.2
MCC	**0.6**	**0.3**	**0.2**	**0.3**	**0.3**	**0.3**	**0.2**
accuracy (%)	99.7	99.2	98.9	99.4	98.6	98.6	98.6
specificity (%)	99.7	99.7	99.2	99.5	99.6	99.6	99.1
NPV (%)	100.0	99.5	99.6	99.9	99.0	99.0	99.5
FDR (%)	**66.7**	**66.7**	**85.7**	**80.0**	**60.0**	**60.0**	**87.5**

(PARAFIT p-value range 1 + HPCC p-value range 2; the p-values after rank-by-rank consensus are listed) for which there exists experimental confirmation in databases such as BindingDB, DrugBank, ChEMBL, PubChem, KiDb, ChemProt, or Wombat. As mentioned in Methods, we consider relationships to be relevant when the p-value is less than 10^{-40} (i.e., "highly significant", "very significant", and "significant" relationships). If very few or no relationships are found, we also explore "interesting" relationships with $p < 10^{-30}$.

The results in Table 10 show that the GES CPF is effective in identifying known relationships between drug families, that is, between drugs from the same family known to have related activity classes, such as 5-hydroxytryptamines (i.e., 5-hydroxytryptamine 2B receptor and 5-hydroxytryptamine 7 receptor; $p = 10^{-40}$), mitogen-activated protein kinases (i.e., mitogen-activated protein kinase 1 and mitogen-activated protein kinase 3; $p = 10^{-59}$), carbonic anhydrases (i.e., carbonic anhydrase 1 and carbonic anhydrase 4; $p = 10^{-47}$), γ-aminobutyric acid (GABA) receptors (i.e., γ-aminobutyric acid receptor subunit γ-3 and γ-aminobutyric acid receptor subunit π; $p = 10^{-70}$), retinoic acid receptors (retinoic acid receptor RXR-α and retinoic acid receptor RXR-γ; $p = 10^{-68}$), and the collagenase family (i.e., 72 kDa type IV collagenase and collagenase 3; $p = 10^{-70}$). On the other hand, it is much more difficult to predict a compound's polypharmacology for drug targets that share no discernible similarity in amino acid sequence, protein structure, or endogenous ligands. GES can also relate drug targets from different families having no obvious biological similarity, such as peroxisome proliferator-activated receptor (PPAR)-δ with prostacyclin receptor ($p = 10^{-41}$), 3-hydroxy-3-methylglutaryl-coenzyme A reductase with integrin β2 ($p = 10^{-65}$), and reverse transcriptase and histamine H3 receptor. This last relationship is not presented in Table 10 because it was found as a "probable" relationship ($p = 10^{-20}$). If this putative relationship is real, it might explain why anti-HIV reverse transcriptase medications sometimes cause painful skin rashes as a side effect.

If we examine further particular examples in Table 10, we can see that the GES polypharmacology results are in accordance with some target relationships that are well-known or are under investigation in the current peer-reviewed literature. For example, GES finds a significant relationship between the glutamate [NMDA] receptor subunit ε3 and opioid receptor σ1 ($p = 10^{-41}$). Many studies have shown that σ1 receptors are able to modulate several neurotransmitter systems. It has been reported that σ1 receptors can potentiate glutamatergic neurotransmission, enhance cholinergic neurotransmission, enhance serotonergic neurotransmission, negatively modulate the GABA-ergic system, diminish noradrenaline release, and modulate dopaminergic neurotransmission.[45]

Serotonin receptors modulate the release of many neurotransmitters, including glutamate, GABA, dopamine, epinephrine/norepinephrine, and acetylcholine, as well as many hormones, including oxytocin, prolactin, vasopressin, cortisol, corticotropin, and substance P, among others. Here, GES finds a significant relationship between the 5-hydroxytryptamine 2A receptor and the D(1A) dopamine receptor ($p = 10^{-38}$), and studies have shown that the 5-HT1A and 5-HT2A receptors mediate the changes in cortical dopaminergic transmission induced by atypical antipsychotic drugs.[46]

Some other examples of related targets belonging to different target families include cytochrome P450 17A1 and androgen receptor ($p = 10^{-47}$), and it is known that cytochrome P450 17A1 catalyzes the biosynthesis of androgens in humans.[47] There is also a relationship between cystic fibrosis transmembrane conductance regulator and solute carrier family 12 member 2 ($p = 10^{-50}$), and previous studies have shown that solute carrier family 12 member 2 operates together with cystic fibrosis transmembrane conductance regulator to produce depolarizing GABA/glycine-mediated synaptic events.[48] Finally, long-chain-fatty-acid-CoA ligase 4 is related to PPAR-γ ($p = 10^{-51}$), and there is evidence in the literature to show that rosiglitazone, an insulin-sensitizing agent, exerts beneficial effects on atherosclerosis. Additionally, rosiglitazone is known to affect other targets as well as PPAR-γ, such as directly inhibiting recombinant long-chain acyl-CoA synthetase (ACSL)-4 activity.[49]

Advantages of the GES Polypharmacology Fingerprint. As demonstrated here, the GES CPF can correctly identify targets that have previously been selected as targets for preclinical compounds or approved drugs. Because of its analytical representation, GES CPF is faster than previous polypharmacology approaches, allowing thousands of comparisons to be calculated in a just a few minutes. Moreover, the GES approach could also be used with other molecular properties, similarity measures, and center molecules. Here, the GES approach was applied using ligand shapes, but it could also be used in the ligand−target and target−target spaces.

We have shown an example of the GES CPF calculated for some 800 targets, but we stress that the CPF does not have a fixed length and is applicable to any number of targets. Furthermore, the GES cluster similarity score depends only on the similarity of pairs of CMs and the SDs of their respective

Table 10. Targets Predicted to Be Highly Related[a]

Primary biological target	New predicted target	Activity for the predicted target (μM)	Scoring (P-value)
Calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1B	Voltage-dependent T-type calcium channel subunit alpha-1H	pIC50 5.15 [8]	2.65E-59
Hydroxyapatite	V-type proton ATPase catalytic subunit A	NA [3]	8.40E-37
Cocaine- and amphetamine-regulated transcript protein	Amine oxidase [flavin-containing] B	Ki 280 [4]	3.68E-35
Voltage-dependent T-type calcium channel subunit alpha-1H	Sodium/potassium-transporting ATPase alpha-1 chain	NA [3]	6.29E-47
Cystic fibrosis transmembrane conductance regulator	Solute carrier family 12 member 2	NA [3]	7.42E-50
Cytochrome c	Interleukin-1 beta	NA [7]	1.84E-70
Cysteinyl leukotriene receptor 2	fMet-Leu-Phe receptor	NA [7]	1.84E-70
Dr hemagglutinin structural subunit	Complement decay-accelerating factor	NA [7]	1.84E-70
Carbonic anhydrase-related protein 10	Sodium channel subunit beta-1	NA [7]	1.84E-70
Coagulation factor IX	Vitamin K-dependent protein Z	NA [7]	1.84E-70
Carbonic anhydrase-related protein 11	Sodium channel subunit beta-4	NA [7]	1.84E-70
Sodium channel subunit beta-4	Carbonic anhydrase 6	Ki 0.089 [4]	1.84E-70
Chloride channel protein ClC-Ka	Phospholipase A2	NA [7]	1.84E-70
Chloramphenicol acetyltransferase	Elongation factor G	NA [7]	1.84E-70
Peroxisome proliferator-activated receptor delta	Prostacyclin receptor	NA [3]	1.27E-41
BC269730_2	Cytochrome P450 17A1	IC50 53.2 [4]	8.96E-31
Alpha-1D adrenergic receptor	Cytochrome P450 17A1	NA [4]	7.06E-30
Prostaglandin E2 receptor EP4 subtype	Prostaglandin E2 receptor, EP2 subtype	NA [3]	3.20E-31
Caspase-3	Interleukin-1 beta	NA [7]	1.84E-70
DNA polymerase epsilon subunit 2	Ribonucleoside-diphosphate reductase subunit M2 B	NA [7]	1.84E-70
Gamma-aminobutyric acid receptor subunit gamma-3	Gamma-aminobutyric acid receptor subunit pi	NA [7]	1.84E-70
Retinoic acid receptor RXR-alpha	Retinoic acid receptor RXR-gamma	NA [7]	1.03-68
ATP-sensitive inward rectifier potassium channel 8	Gamma-aminobutyric-acid receptor subunit alpha-6	NA [7]	4.27E-53
Acetylcholinesterase	Neuronal acetylcholine receptor subunit alpha-2	NA [7]	1.21E-45
Apoptosis regulator Bcl-2	Tubulin beta-1 chain	ID50 ratio 0.50 [4]	2.15E-42
5-hydroxytryptamine 2B receptor	5-hydroxytryptamine 7 receptor	Ki 0.02691 [1]	1.47E-40
Adenosine A2a receptor	cAMP-specific 3',5'-cyclic phosphodiesterase 4C	NA [3]	2.83E-43
Peroxisome proliferator-activated receptor delta	Prostacyclin receptor	NA [7]	1.27E-41
3-hydroxy-3-methylglutaryl-coenzyme A reductase	Integrin beta-2	NA [7]	4.26E-65
Long-chain-fatty-acid--CoA ligase 4	Peroxisome proliferator-activated receptor gamma	NA [7]	1.24E-51
5-hydroxytryptamine 2A receptor	D(1A) dopamine receptor	Ki 10 [6], IC50 0.943 [4]	1.51E-38
Potassium voltage-gated channel subfamily KQT member 1	Voltage-dependent P/Q-type calcium channel subunit alpha-1A	NA [3]	1.08E-48
Cytochrome P450 17A1	Androgen receptor	IC50 0.037 [4]	3.79E-47

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

Table 10. continued

Primary biological target	New predicted target	Activity for the predicted target (μM)	Scoring (P-value)
Glutamate [NMDA] receptor subunit epsilon-3	Opioid receptor, sigma 1	IC50 1.1 [1]	6.69E-41
D(1A) dopamine receptor	5-hydroxytryptamine 2A receptor	Ki 0.73 [4]	1.51E-38
cAMP-specific 3',5'-cyclic phosphodiesterase 4D	Gamma-aminobutyric-acid receptor subunit beta-2	IC50 2 [2]	8.53E-25
Beta-2 adrenergic receptor	5-hydroxytryptamine 1A receptor	Ki 51 [6]	2.80E-38
RET proto-oncogene	High affinity nerve growth factor receptor	Kd 10 [4]	1.43E-42
Calcium-transporting ATPase type 2C member 1	Glutamate receptor 1	NA [3]	9.85E-70
ATP synthase delta chain, mitochondrial	Glutamate receptor 1	NA [4]	9.85E-70
72 kDa type IV collagenase	Collagenase 3	IC50 0.0012 [4]	3.18E-70
	Interstitial collagenase	IC50 0.0015 [4]	3.18E-70
	Matrix metalloproteinase-9	IC50 0.0016 [4]	2.09E-70
	Neutrophil collagenase	IC50 0.002 [4]	3.18E-70
	Stromelysin-1	IC50 0.0044 [4]	3.18E-70
	Macrophage metalloelastase	NA [5]	4.02E-61
	Matrilysin	NA [5]	3.18E-70
DNA polymerase epsilon catalytic subunit A	Ribonucleoside-diphosphate reductase M2 subunit	NA [7]	1.84E-70
D-lactate dehydrogenase	Succinate dehydrogenase [ubiquinone] cytochrome b small subunit, mitochondrial	NA [7]	1.84E-70
RNA-directed RNA polymerase catalytic subunit	Cytosolic purine 5'-nucleotidase	NA [7]	1.84E-70
5-hydroxytryptamine 2B receptor	5-hydroxytryptamine 7 receptor	Ki 0.02691 [1]	1.47E-40
Mitogen-activated protein kinase 1	Mitogen-activated protein kinase 3	IC50 3.3 [4]	7.35E-59
Carbonic anhydrase 1	Carbonic anhydrase 4	Ki 0.449 [4]	5.02E-47

[1] BindingDB
[2] Etazolate in clinical trials (http://www.scbt.com/datasheet-201186-Etazolate-Hydrochloride.html)
[3] DrugBank
[4] ChemBL
[5] PubChem
[6] KiDb
[7] ChemProt
[8] Wombat

[a]Red boxes highlight examples of strongly predicted relationships between drug targets that are known to have related activity classes (e.g. 5-hydroxytryptamines, mitogen-activated protein kinases, carbonic anhydrases, or collagenase) and also targets that do not have any obvious biological similarity (e.g. PPAR-δ with prostacyclin receptor or 3-hydroxy-3-methylglutaryl-coenzyme A reductase with integrin β2). "NA" denotes cases where there is experimental evidence to corroborate a predicted relationship but activity values are not available.

clusters. It does not depend on the number of members of each cluster, which is another advantage compared with other chemogenomic approaches. GES can also be used to represent multiple ligand conformations and to analyze virtual screening hit lists. Using clusters of conformations helps to circumvent the difficult problem of how to select the best ligand conformation to use. This is an important feature of GES, given the strong dependence of 3D methods on the particular conformation used.

Finally, in regard to performance, GES has comparable sensitivity to current polypharmacology approaches but clearly has lower fall-out and FDR. As we noted before,[10] it cannot be expected that different polypharmacology methods should give identical results. Thus, it is not appropriate to compare absolute scores or p-values directly. Nonetheless, we previously compared[10] the relative ranks obtained for the compounds in each ligand set using both 3D GES and SEA.[5] Those results show that GES finds in the first ranking positions the same one or two related ligand sets as the SEA algorithm. In view of the fact that GES uses 3D SH shape representations and groups ligands using smooth analytic functions whereas SEA uses combinatorial sampling of 2D Daylight fingerprints, the overall similarity between these results is rather remarkable. Indeed, further comparison of the GES promiscuity predictions with those of Keiser et al.[5] shows that, like SEA, GES also predicts that serotonin reuptake inhibitors might inhibit uptake of serotonin into presynaptic neurons and therefore that these inhibitors could also act as β-blockers (which bind to β-adrenergic receptors in, e.g., blood vessels and heart muscle). These predictions were confirmed in vitro by Keiser et al. However, there are also several targets that GES relates with significant p-values but SEA does not, such as AMPA and

dx.doi.org/10.1021/ci4006723 | J. Chem. Inf. Model. 2014, 54, 720−734

phosphatidylinositol, chymotrypsin and carbacephem, 5 HTF 1 agonist and mGlu1, and adrenergic β1 agonist and squalene epoxidase inhibitor. In particular, GES relates GABA A and 5HT1F agonist with a significant p-value while SEA does not, yet there is experimental evidence of this relationship.[50] Thus, some targets identified by the 2D fingerprints used in SEA might not be identified by the 3D shape representation used in GES. On the other hand, when shape plays an important role in the comparison of the target-ligand sets but the ligands differ in two dimensions, some targets can be identified by GES and not by SEA.

CONCLUSION

We have presented our polypharmacology approach using a set of 5000 DrugBank ligands related to some 800 targets. The estimated rate of off-target interactions obtained with GES CPF agrees with previously reported data. The FDR calculated with the GES CPF is in accordance with the average promiscuity for DrugBank-approved and experimental drugs calculated previously by Hu and Bajorath[39] and Mestres et al.[40] The GES method is suitable for predicting and quantifying (using p-values) polypharmacology with a larger set. CPF is applicable to any number of targets, and the GES cluster similarity score does not depend on the number of members of each cluster. Furthermore, by using clusters of conformations, GES helps to circumvent the difficult problem of how to select the best ligand conformation to use in current virtual screening protocols.

Measuring the performance of the GES CPF requires handling of missing data. In order to analyze the effects of the missing data in the polypharmacology prediction, we used a ROC analysis in the context of polypharmacology and performed an "FP correction", which involved reassigning an FP calculated for a small-database ligand as a TP whenever any member of the corresponding ligand set was annotated as "active" in a larger database.

The method is effective in identifying relationships between drug families that are both expected (i.e., between families known to have related activity classes, e.g., 5-hydroxytryptamines, mitogen-activated protein kinases, carbonic anhydrases, or the collagenase family) and unexpected (i.e., between families that do not have any obvious biological similarity, e.g., PPAR-δ with prostacyclin receptor or 3-hydroxy-3-methylglutaryl-coenzyme A reductase with integrin β2). Although it cannot be expected that different polypharmacology methods should give identical results, the overall similarity between the results obtained using 3D GES CPF and SEA is rather remarkable.

Overall, the GES CPF is clearly strong in finding high-confidence positive predictions. The method retrieves very few FPs and therefore gives very high precision and very low FDR and fall-out. The experimentally validated examples demonstrate that GES CPFs are efficient for studying drug−target relationships and provide a novel and powerful approach for proposing new targets for preclinical compounds and clinical drug candidates. We therefore expect that within the scope of a drug development project, GES CPFs could be used for (a) early characterization of preclinical compounds and (b) finding new therapeutic indications for safe clinical or approved drugs. In the first scenario, there are clear application cases such as the need to characterize thoroughly the pharmacological profiles of preclinical compounds, to strengthen the company's early-stage pipeline, or to diversify potential indications. In these cases, GES CPFs could be applied at the beginning of a preclinical project, when investigating therapeutic effects and adverse side effects, or when deciphering the MoA of preclinical compounds. In the second scenario, there are also clear application cases such as reducing the time and cost to market (because existing drugs have known pharmacokinetic and safety profiles) or the rapid evaluation of a new use in phase II clinical trials. In these cases, GES CPFs could be used if a clinical trial of a safe drug fails due to lack of efficacy or if there is a need to add value to existing compounds that are already in development or on the market.

ASSOCIATED CONTENT

Supporting Information

The experimental information downloaded for the DrugBank compounds is provided in the Supplementary_Data folder, which contains the data from ChEMBL (data_ChEMBL.txt), PubChem (data_DrugBank_smiles_target_action.txt, data_DrugBank_smiles_enzyme_action.txt, data_Correspondance_DrugBank_Pubchem_ID.txt, Biological_results_PubChem folder), and BindingDB (data_BindingDB.txt). The plots of the observed distributions of pairwise ligand set cluster scores for both PARAFIT and HPCC are provided as supplementary figures. The comparison of the CCPF with the EPFs obtained for each of the individual experimental databases (PubChem targets, PubChem enzymes, PubChem targets + PubChem enzymes, ChEMBL, and BindingDB) are provided as supplementary tables. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: +33-354 958 604. Fax: +33-383 593 046. E-mail: pereznueno@harmonicpharma.com.

Notes

The authors declare the following competing financial interest(s): David W. Ritchie is a shareholder in Harmonic Pharma.

ACKNOWLEDGMENTS

The authors thank Cepos InSilico Ltd. for providing an Academic License for PARASURF and also ChemAxon for the license for the Marvin and JChem toolkits.

REFERENCES

(1) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2007**, *2*, 874−880.

(2) Chong, C. R.; Sullivan, D. J. New uses for old drugs. *Nature* **2007**, *448*, 645−646.

(3) Carrieri, A.; Pérez-Nueno, V. I.; Lentini, G.; Ritchie, D. W. Recent trends and future prospects in computational GPCR drug discovery: From virtual screening to polypharmacology. *Curr. Top. Med. Chem.* **2013**, *13*, 1069−1097.

(4) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "Target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802−6810.

(5) 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.

(6) Gregori-Puigjane, E.; Mestres, J. SHED: Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615−1622.
(7) Vidal, D.; Mestres, J. In Silico Receptorome Screening of Antipsychotic Drugs. *Mol. Inf.* **2010**, *29*, 543−551.
(8) AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring polypharmacology using a ROCS-based target fishing approach. *J. Chem. Inf. Model.* **2012**, *52*, 492−505.
(9) Liu, X.; Jiang, H.; Li, H. SHAFTS: A hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 2372−2385.
(10) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. W. Detecting drug promiscuity using Gaussian Ensemble Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1948−1961.
(11) Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H. O.; Rognan, D. Protein−ligand-based pharmacophores: Generation and utility assessment in computational ligand profiling. *J. Chem. Inf. Model.* **2012**, *52*, 943−955.
(12) Weill, N.; Rognan, D. Development and validation of a novel protein−ligand fingerprint to mine chemogenomic space: Application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049−1062.
(13) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.
(14) Milletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: From kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418−1431.
(15) Weskamp, N.; Hüllermeier, E.; Klebe, G. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins* **2009**, *76*, 317−330.
(16) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. W. Predicting drug promiscuity using spherical harmonic (SH) shape-based similarity comparisons. *Open Conf. Proc. J.* **2011**, *2*, 113−129.
(17) Meslamani, J.; Bhajun, R.; Martz, F.; Rognan, D. Computational profiling of bioactive compounds using a target-dependent composite workflow. *J. Chem. Inf. Model.* **2013**, *53*, 2322−2333.
(18) Simon, Z.; Peragovics, A.; Vigh-Smeller, M.; Csukly, G.; Tombor, L.; Yang, Z.; Zahoránszky-Kohalmi, G.; Végner, L.; Jelinek, B.; Hári, P.; Hetényi, C.; Bitter, I.; Czobor, P.; Málnási-Csizmadia, A. Drug effect prediction by polypharmacology-based interaction profiling. *J. Chem. Inf. Model.* **2012**, *52*, 134−145.
(19) Simon, Z.; Vigh-Smeller, M.; Peragovics, A.; Csukly, G.; Zahoránszky-Kohalmi, G.; Rauscher, A. A.; Jelinek, B.; Hári, P.; Bitter, I.; Málnási-Csizmadia, A.; Czobor, P. Relating the shape of protein binding sites to binding affinity profiles: Is there an association? *BMC Struct. Biol.* **2010**, *10*, 1−13.
(20) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263−266.
(21) Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313−2325.
(22) Niijima, S.; Yabuuchi, H.; Okuno, Y. Cross-target view to feature selection: Identification of molecular interaction features in ligand-target space. *J. Chem. Inf. Model.* **2011**, *51*, 15−24.
(23) Takigawa, I.; Tsuda, K.; Mamitsuka, H. Mining significant substructure pairs for interpreting polypharmacology in drug−target network. *PLoS One* **2011**, *6*, No. e16999.
(24) Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug−target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **2012**, *8*, No. e1002503.
(25) Berger, S. I.; Iyengar, R. Network analyses in systems pharmacology. *Bioinformatics* **2009**, *25*, 2466−2472.
(26) Pujol, A.; Mosca, R.; Farrés, J.; Aloy, P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.* **2010**, *31*, 115−123.
(27) *CEPOS Mopac*; CEPOS InSilico Ltd.: Erlangen, Germany, 2009; http://www.ceposinsilico.de/ (accessed Oct 3, 2013).
(28) Ritchie, D. W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* **1999**, *20*, 383−395.
(29) Lin, J.; Clark, T. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Chem. Inf. Model.* **2005**, *45*, 1010−1016.
(30) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Clark, T.; Ritchie, D. W. Using spherical harmonic surface property representations for ligand-based virtual screening. *Mol. Inf.* **2010**, *30*, 151−159.
(31) Pérez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixidó, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand−receptor docking. *J. Chem. Inf. Model.* **2008**, *48*, 509−533.
(32) Ritchie, D. W.; Kemp, G. J. L. Protein Docking Using Spherical Polar Fourier Correlations. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 178−194.
(33) Pérez-Nueno, V. I.; Ritchie, D. W.; Borrell, J. I.; Teixidó, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape based virtual screening approach: Further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146−2165.
(34) Ritchie, D. W.; Pérez-Nueno, V. I. In *Scaffold Hopping in Medicinal Chemistry*; Brown, N., Ed.; Methods and Principles in Medicinal Chemistry, Vol. *58*; Wiley-VCH: Weinheim, Germany, 2013; Chapter 12.
(35) Karaboga, A. S.; Petronin, F.; Marchetti, G.; Souchet, M.; Maigret, B. Benchmarking of HPCC: A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments. *J. Mol. Graphics Modell.* **2013**, *41*, 20−30.
(36) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36* (Suppl. 1), D901−D906.
(37) Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **1999**, *6*, 281−297.
(38) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−368.
(39) Hu, Y.; Bajorath, J. High-resolution view of compound promiscuity. *F1000Research* **2013**, *2*, 144.
(40) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. The topology of drug−target interaction networks: Implicit dependence on drug properties and target families. *Mol. BioSyst.* **2009**, *5*, 1051−1057.
(41) Little, R.; Rubin, D. B. *Statistical Analysis with Missing Data*, 2nd ed.; John Wiley & Sons: New York, 2002.
(42) Liu, X.; Zhao, Y. Semi-empirical likelihood inference for the ROC curve with missing data. *J. Stat. Plann. Inference* **2012**, *142*, 3123−3133.
(43) Martínez-Camblor, P. Area under the ROC curve comparison in the presence of missing data. *J. Korean Stat. Soc.* **2013**, *42*, 431−442.
(44) Xia, J.; Broadhurst, D. I.; Wilson, M.; Wishart, D. S. *Metabolomics* **2013**, *9*, 280−299.
(45) Cobos, E. J.; Entrena, J. M.; Nieto, F. R.; Cendán, C. M.; Del Pozo, E. *Curr. Neuropharmacol.* **2008**, *6*, 344−366.
(46) Ichikawa, J.; Ishii, H.; Bonaccorso, S.; Fowler, W. L.; O'Laughlin, I. A.; Meltzer, H. Y. 5-HT2A and D-2 receptor blockade increases cortical DA release via 5-HT1A receptor activation: A possible mechanism of atypical antipsychotic-induced cortical dopamine release. *J. Neurochem.* **2001**, *76*, 1521−1531.
(47) DeVore, N. M.; Scott, E. E. Structures of cytochrome P450 17A1 with prostate cancer drugs abiraterone and TOK-001. *Nature* **2012**, *482*, 116−119.
(48) Ostroumov, A.; Simonetti, M.; Nistri, A. Cystic fibrosis transmembrane conductance regulator modulates synaptic chloride

homeostasis in motoneurons of the rat spinal cord during neonatal development. *Dev. Neurobiol.* **2011**, *71*, 253−68.
(49) Askari, B.; Kanter, J. E.; Sherrid, A. M.; Golej, D. L.; Bender, A. T.; Liu, J.; Hsueh, W. A.; Beavo, J. A.; Coleman, R. A.; Bornfeldt, K. E. Rosiglitazone inhibits acyl-CoA synthetase activity and fatty acid partitioning to diacylglycerol and triacylglycerol via a peroxisome proliferator-activated receptor γ-independent mechanism in human arterial smooth muscle cells and macrophages. *Diabetes* **2007**, *56*, 1143−1152.
(50) Phebus, L. A.; Johnson, K. W.; Zgombick, J. M.; Gilbert, P. J.; Van Belle, K.; Mancuso, V.; Nelson, D. L.; Calligaro, D. O.; Kiefer, A. D., Jr.; Branchek, T. A.; Flaugh, M. E. Characterization of LY344864 as a pharmacological tool to study 5-HT1F receptors: Binding affinities, brain penetration and activity in the neurogenic dural inflammation model of migraine. *Life Sci.* **1997**, *61*, 2117−2126.