BJP British Journal of Pharmacology

DOI:10.1111/bph.12294
www.brjpharmacol.org

RESEARCH PAPER

Comparison of ultra-fast 2D and 3D ligand and target descriptors for side effect prediction and network analysis in polypharmacology

Alvaro Cortés-Cabrera[1,2], Garrett M Morris[3,4], Paul W Finn[3], Antonio Morreale[1,*] and Federico Gago[2]

[1]*Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Madrid, Spain,* [2]*Departamento de Ciencias Biomédicas, Universidad de Alcalá, Madrid, Spain,* [3]*InhibOx Ltd, Oxford Centre for Innovation, Oxford, UK, and* [4]*Crysalin Ltd, Cherwell Innovation Center, Oxfordshire, UK*

Correspondence
Professor Federico Gago, Department of Biomedical Sciences, Universidad de Alcala, Alcala de Henares, E-28871 Madrid, Spain. E-mail: federico.gago@uah.es

*Present address: Repsol Technology Center, E-28923 Móstoles, Madrid, Spain.

Keywords
adverse drug reactions; chemical fingerprints; drug targets; polypharmacology; side effects

Received
10 April 2013
Revised
24 June 2013
Accepted
2 July 2013

BACKGROUND AND PURPOSE

Some existing computational methods are used to infer protein targets of small molecules and can therefore be used to find new targets for existing drugs, with the goals of re-directing the molecule towards a different therapeutic purpose or explaining off-target effects due to multiple targeting. Inherent limitations, however, arise from the fact that chemical analogy is calculated on the basis of common frameworks or scaffolds and also because target information is neglected. The method we present addresses these issues by taking into account 3D information from both the ligand and the target.

EXPERIMENTAL APPROACH

ElectroShape is an established method for ultra-fast comparison of the shapes and charge distributions of ligands that is validated here for prediction of on-target activities, off-target profiles and adverse effects of drugs and drug-like molecules taken from the DrugBank database.

KEY RESULTS

The method is shown to predict polypharmacology profiles and relate targets from two complementary viewpoints (ligand- and target-based networks).

CONCLUSIONS AND IMPLICATIONS

The open-access web tool presented here (http://ub.cbm.uam.es/chemogenomics/) allows interactive navigation in a unified 'pharmacological space' from the viewpoints of both ligands and targets. It also enables prediction of pharmacological profiles, including likely side effects, for new compounds. We hope this web interface will help many pharmacologists to become aware of this new paradigm (up to now mostly used in the realm of the so-called 'chemical biology') and encourage its use with a view to revealing 'hidden' relationships between new and existing compounds and pharmacologically relevant targets.

Abbreviations

ADR, adverse drug reactions; DUD, Directory of Useful Decoys; EVD, extreme value distribution; LBVS, ligand-based virtual screening; MDDR, MDL Drug Data Report; MMFF94, Merck Molecular Force Field; NHR, nuclear hormone receptors; PDB, Protein Data Bank; ROCS, Rapid Overlay of Chemical Structures; SEA, Similarity Ensemble Approach; SMILES, simplified molecular-input line-entry specification; WOMBAT, WOrld of Molecular BioAcTivity

Introduction

The famous 'magic bullet' term coined by Paul Ehrlich more than one hundred years ago in the field of chemotherapy (Ehrlich, 1907; Witkop, 1999) paved the way to the classical one-compound–one-target paradigm that has largely dominated drug discovery for the last 25 years or so. This reductionist concept has led to a limited appraisal of the causes underlying the side effects of commercial drugs, derived from modulation of secondary targets that can nevertheless play a fundamental role in explaining pharmacological profiles. This is particularly true for drugs acting on the CNS, which can bind to many different receptors ('magic shotgun'; Roth *et al.*, 2004), and for some multi-kinase inhibitors in oncology (Knight *et al.*, 2010). This realization suggests that a more direct approach to polypharmacology should be taken in modern drug discovery from the very early stages of screening and lead identification. A multi-target–multi-compound approach would provide a much more accurate description of the underlying pharmacology but, given the large size of both chemical and biological spaces, it is also harder to understand, hence the need for tailor-made computer programs that can handle and relate the enormous, and still increasing, amounts of bioactivity data available for both compounds and targets.

Network analysis (Hopkins, 2008; Berger and Iyengar, 2009) in systems pharmacology (Van Der Greef and McBurney, 2005) has recently emerged and promises to revolutionize the field of drug discovery. Polypharmacology, drug repurposing, target fishing and adverse effect prediction are some of the major applications made possible by this paradigm shift, which contrasts with the traditional one-ligand–one-target approach that is still in use in most high-throughput experimental and virtual screening campaigns nowadays (Ripphausen *et al.*, 2010). Traditionally, *in silico* target-fishing methods have been related to reverse docking (Cai *et al.*, 2006) using one or several compounds against multiple putative targets (Simon *et al.*, 2012). In recent years, however, ligand-based methods exploiting either fingerprints containing two- (2D) and three-dimensional (3D) chemical information or 3D shape descriptors (Bender *et al.*, 2007; Keiser *et al.*, 2007; Vidal and Mestres, 2010; Besnard *et al.*, 2012) have been employed to predict activity profiles and target–target relationships. Superpositional methods that either compare the shapes of two molecules by analytically optimizing their volume intersection (Grant *et al.*, 1996), as implemented in the program ROCS (Rapid Overlay of Chemical Structures, OpenEye Scientific Software, 2011; Rush *et al.*, 2005) or use a surface-based morphological similarity function while minimizing the overall molecular volume of the aligned structures, such as Surflex-Sim (Jain, 2000), have been shown to perform well in the task of predicting off-target activities of ligands (Ballester *et al.*, 2009; Yera *et al.*, 2011). These approaches, although successful, rely on direct pairwise comparisons and this can reduce global performance when database size grows above tens of millions of molecules (Wang *et al.*, 2009).

ElectroShape is a non-superpositional method for ultrafast comparison of ligands that expands the capabilities and improves the performance of the Ultra Shape Recognition methodology (Ballester and Richards, 2007) by incorporating the molecular charge distribution (Armstrong *et al.*, 2010). In brief, ElectroShape uses three spatial dimensions and adds partial charge as a fourth dimension to capture electrostatic information in the form of 15 descriptors that account for the first, second and third moments of the distributions of distances from five distinct points of the molecule (centroids) in a four-dimensional (4D) space. Molecular similarity is then calculated as the Manhattan distance between ElectroShape descriptors belonging to two different molecules. The following facts are key to increasing the speed of calculation and improving performance: (i) the descriptors are very small and can be pre-calculated and stored for each compound from ensembles of low-energy conformers, and (ii) the similarity calculation is non-superpositional and requires only a few mathematical operations. The value of Ultra Shape Recognition in ligand-based virtual screening (LBVS) has already been recognized (Ballester *et al.*, 2010).

In the following, we show how the ElectroShape method has been validated for off-target prediction ('target fishing'), LBVS and chemogenomic network analysis (Figure 1). The results obtained have been compared, firstly, to those produced by other well-known 2D techniques that make use of ligand and target annotations from bioactivity databases, and then to structural information from a database of ligand-binding sites in proteins. Implementation on an unsophisticated web server is also presented that can enable pharmacologists and other interested researchers to predict possible adverse effects and secondary targets for a given drug and to explore pharmacological space from the viewpoints of both ligand and target simultaneously.

Methods

Chemogenomic datasets

To explore the potential use of the ElectroShape approach (Armstrong *et al.*, 2010) for target fishing and prediction of adverse effects, and to compare it with existing 2D methods (Hert *et al.*, 2008), we first built a chemogenomic database using information available from DrugBank (Knox *et al.*, 2011). To this end, target and drug sets were downloaded from http://www.drugbank.ca/ and imported locally into a MySQL relational database. Molecules were then parsed using RDKit (Landrum, 2011) to generate canonical simplified molecular-input line-entry specification (SMILES) strings (Weininger, 1988). Drug–target associations were also imported from DrugBank using Python scripts (http://www.python.org/).

For the 2D analysis, Morgan fingerprints, which are roughly equivalent to the Extended-Connectivity Fingerprints (ECFP4) commonly used in other target-fishing applications (Hert *et al.*, 2008; Rognan and Meslamani, 2011), were generated using RDKit with a radius of two bonds and 2048 bits and inserted in the database. A simple in-house chemistry cartridge was used to allow searching within the database (Cabrera *et al.*, 2011).

To perform the 4D ElectroShape analysis (Cartesian coordinates and charges), the SMILES string representations of the drugs were converted into 3D structures using CORINA (Sadowski *et al.*, 1994; 2003) and point charges were assigned

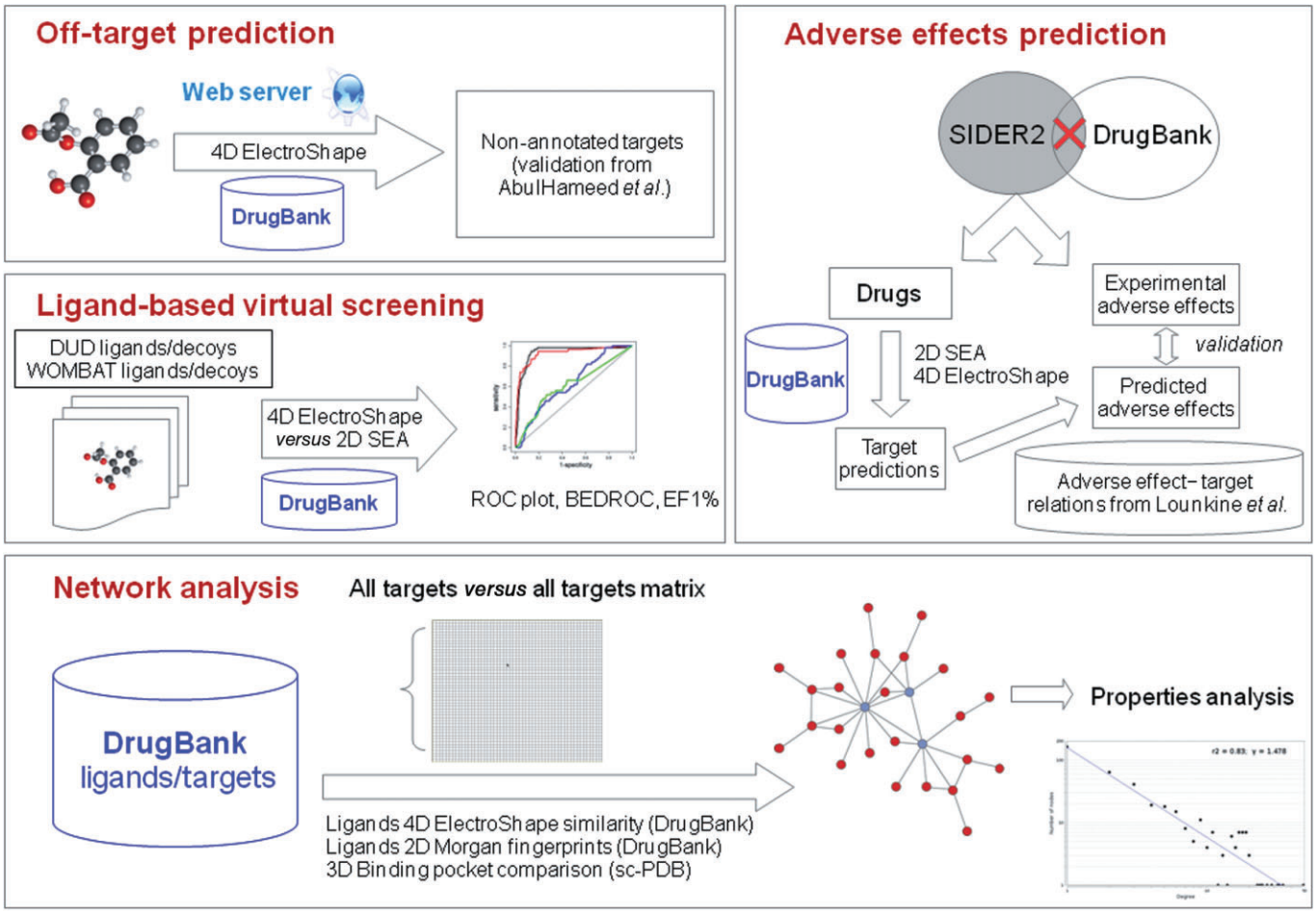

Figure 1
Scheme depicting the workflow presented in this article (see main text for details).

by OpenBabel making use of the Merck Molecular Force Field (MMFF94; Halgren, 1996). Then, a conformational analysis using ALFA (Cabrera *et al.*, 2011), a rule-based conformer generator similar to OMEGA (Hawkins *et al.*, 2010), was undertaken setting the maximum number of conformers to 200 and activating the maximum dissimilarity option to generate the most diverse ensemble possible. By selecting this number we ensure that most of the conformational space is covered for each ligand and avoid the risk of losing alternative similarity matches, a possibility that is almost a certainty when just a single low-energy conformation is taken as a representative 3D structure of ligands with a high number of rotatable bonds. Finally, the molecules were converted to Protein Data Bank (PDB) format (Berman *et al.*, 2007) using OpenBabel (O'Boyle *et al.*, 2011) and 4D descriptors were generated by ElectroShape (Armstrong *et al.*, 2010) for each molecular conformer and inserted into the database to allow simplified queries within chemical space using the cartridge.

To explore the similarity between targets in biological space, we used sc-PDB (Meslamani *et al.*, 2011), an annotated database that contains 3D coordinates for bound ligands and ligand-binding sites, as extracted from the PDB. The sc-PDB complexes were filtered so that only those with a direct match in the DrugBank set were kept.

Similarity calculation and drug–drug and target–target matrices

To calculate target similarity from 2D ligand descriptors, we used the Similarity Ensemble Approach (SEA; Keiser *et al.*, 2007) and Morgan fingerprints (Landrum, 2011) of all molecules, which are annotated to have an effect on every DrugBank target. To this end, an all-drug versus all-drug matrix was built by comparing the fingerprints using the Tanimoto coefficient (Rogers and Tanimoto, 1960). Then, the scores of all the drugs associated with a given target were summed and a final *Z*-score for each target pair was computed. To obtain the *Z*-score for the similarity between target A and target B, it is necessary to calculate the expected random average (μ) and standard deviation (σ) of the total score of any two ensembles of $n \times m$ drugs (n being the number of drugs that bind to target A and m the number of drugs that bind to target B). Drugs present in DrugBank were randomly grouped in several ensembles ranging in size from 1 to 50 molecules (which is the common range of drugs per target found in DrugBank) and the total score was computed. Then, these scores were fitted to a power law equation and used to calculate the expected values for random targets and the corresponding *Z*-scores:

$$Z\text{-score} = (\Sigma \text{ scores} - \mu)/\sigma$$

In the case of ElectroShape's 4D ligand similarity, we followed an analogous method. Similarities between molecules were computed by calculating the Manhattan distance between each pair of conformer's descriptors for both molecules and the maximum value was kept as the final score between the two ligands. Individual scores between each pair of molecules were summed and the *Z*-score was calculated using the same procedure as for 2D comparisons: generation of random sets of different sizes followed by calculation of the best power law fit for both average and standard deviation of the total sum score.

Finally, for ligand-binding site similarity, we first used the protein structure alignment algorithm TM-align (Zhang and Skolnick, 2005) to align the binding sites of the targets present in both DrugBank and sc-PDB and then a template modelling score (TM-score) was extracted as a normalized measure of the topological similarity between any two target binding sites.

Web server implementation

The ElectroShape Polypharmacology web server was implemented in Python using the Django Web framework (https://www.djangoproject.com/). It is based on a MySQL database extended with a cartridge previously developed by the authors (Cabrera *et al.*, 2011). The user only has to paste a SMILES string into the search box and then select a method for target fishing. If the traditional 2D SEA method is chosen, the SMILES string is used to generate Morgan fingerprints and perform the query to the database. If an ElectroShape-based method is requested, the server (i) converts this string into a 3D structure, (ii) explores the conformational space using the rule-based program ALFA (Cabrera *et al.*, 2011), (iii) assigns MMFF94 atom point charges using OpenBabel and (iv) calculates the ElectroShape 4D descriptors.

The network representation is based on the Javascript InfoVis toolkit and the information is extracted directly from the database through Django. Finally, 2D images for the small molecules are generated on-the-fly using the RDKit module from Django.

Prediction of side effects and off targets

Reported associations extracted from the literature (Lounkine *et al.*, 2012) between pharmacological effects (including those categorized as 'adverse effects') and certain targets were imported into a MySQL relational database. On the other hand, adverse drug reactions (ADR) and corresponding drugs' SMILES strings were downloaded from the side effect resource SIDER 2 (Kuhn *et al.*, 2010) and a subset comprising 151 drugs not present in DrugBank was created and used for validating our results. For each of these drugs, we calculated the putative binding profile and queried the database for the adverse effects associated to the predicted targets. The goodness of our predictions was assessed by evaluating how many ADR are correctly identified and/or missed out of the total number of 270 ADR contained in the MySQL database. This problem was handled as a group of 270 binary decisions for each drug: 1 = true association; 0 = no association. The true positive and negative rates, as well as the false positive and negative rates, were calculated for each drug and averaged.

For prediction of non-annotated off-targets, the dataset proposed by (AbdulHameed *et al.*, 2012) was used. SMILES strings were extracted from PubChem when possible, or generated manually.

LBVS

To perform an on-target validation, we used the Directory of Useful Decoys (DUD) (Huang *et al.*, 2006) and the 'WOrld of Molecular BioAcTivity' (WOMBAT) (Good and Oprea, 2008) subset to account for the analogue bias in the set of active compunds. Ligands and decoys were extracted from the original distributions and processed in an identical way to that followed with the DrugBank molecules in order to allow direct search and comparison because ElectroShape results are dependent on the partial charge method used. Briefly, the original files in multi-molecule structure data format were used to generate directly Morgan fingerprints and SMILES strings using RDKit. The latter were employed to generate initial 3D structures using CORINA (Sadowski *et al.*, 2003) and different conformations for each of them were obtained by using ALFA (Cabrera *et al.*, 2011).

We were able to match every target in the DUD to a corresponding DrugBank target with at least one representative ligand except for β-lactamase AmpC whose only protein-bound ligand was not transformed properly, hence results are presented for only 39 targets. As a measure of the global similarity per ligand, we used the SEA *Z*-score for 2D and 4D methods, and in the case of ElectroShape 4D, we also used the maximum similarity values among the compounds per target so that comparison with other data fusion techniques could be performed.

Network analysis

The target–target matrix was analysed and the *Z*-score distribution was fitted to an extreme value distribution (EVD) in order to calculate the expectation values (*E*-values, e.g. a probability of observing a given *Z*-score using random data) for the similarity between targets (Hert *et al.*, 2008). Using different cut-off values, we transformed those matrices into threshold networks, adding an edge between two targets or nodes if the *E*-value or a similar score was above a certain value that depends on the kind of comparison being made (2D ligand chemistry, 3D ligand-binding site or 4D ligand shape and charge distribution). For the ligand-binding sites, we used the TM-score values to build the corresponding threshold network.

The Cytoscape software and network analysis plug-in (Smoot *et al.*, 2011) was employed to study several statistical properties and to compute the union, difference and intersection of 2D ligand, 4D ligand and 3D receptor networks. Finally, to explore the usefulness of the information in the networks, a 'triviality' test was performed for each kind of network. In this test, we counted the number of edges due to a name similarity (string matching over 0.6) and percentage of common compounds between targets of more than 60%. These indicators try to measure the quantity of information that could be extracted trivially from the names of the targets or, at plain sight, from the profiles of the ligands.

Results

Chemogenomic datasets

After applying to DrugBank the protocol described above, the database contained a total of 5685 molecules and 3779 targets related by 11 559 annotated activities. In the case of the 3D database for ElectroShape, we ended up with 565 505 different conformers. This number roughly corresponds to an average of 100 conformers per molecule. Only sc-PDB ligand-binding sites belonging to targets present in DrugBank were selected and identified by their UniProt (Wu *et al.*, 2006) identification code. This procedure yielded a total of 1056 targets.

Similarity calculation and drug–drug and target–target matrices

Inspired by the work from Shoichet's lab (Hert *et al.*, 2008) that related biological targets by employing the 2D chemical descriptors of their ligands, we analogously developed a parameterization for the SEA (Keiser *et al.*, 2007) method using random groups of molecules from DrugBank. As can be seen in Table 1, after the fitting procedure, power law values are comparable to the original values obtained by Shoichet *et al.* (Hert *et al.*, 2008) for the MDL Drug Data Report (MDDR) database (MDDR, 2006). The SEA method was also re-parameterized using ElectroShape descriptors and an equivalent random set of molecules.

Finally, for the 3D receptor set arising from the intersection of DrugBank and sc-PDB databases, we obtained an all-against-all matrix for the 1056 targets using the TM-score normalized by the average number of residues of the two ligand-binding sites being compared.

Off-target validation results

Using a limited test set from DrugBank (see under Methods) that consisted of recently discovered but not yet annotated secondary targets for a certain number of compounds (Ballester *et al.*, 2009), we tested the ability of the ElectroShape chemogenomic approach to predict these annotations correctly. For three of the four compounds (Table 2), the primary activity was found to be ranked first in the target profile (except for RO-25–6981, in which case it was found in the third place), while all secondary targets were identified within the first seven predicted targets and with very close similarity values to the first scoring target.

Prediction of adverse effects

Following on recently published work (Lounkine *et al.*, 2012), we related certain targets belonging to the DrugBank set to 290 probable adverse effects. We then predicted the 2D SEA polypharmacology profiles of the drugs included in the SIDER 2 database but not in DrugBank, and we extracted the adverse effects related to those targets. For every molecule, the specificity and selectivity of the method were calculated

Table 1

Fitting comparison for DrugBank and Shoichet *et al.* sets in SEA

	Mean exponent	Mean coefficient	Pearson r^2	Standard deviation exponent	Standard deviation coefficient	Pearson r^2
DrugBank 2D Morgan fingerprints	1.01	5.81×10^{-3}	0.9991	0.534	6.03×10^{-2}	0.9977
DrugBank ElectroShape 4D	0.99	4.76×10^{-2}	0.9998	0.635	1.56×10^{-1}	0.9951
Shoichet *et al.* (Hert *et al.*, 2008)	1	4.24×10^{-4}	0.9998	0.665	4.49×10^{-3}	0.9882

Table 2

Off-target test set from (AbdulHameed *et al.*, 2012)

Compound	Primary target rank	Secondary target rank
Dimetholizine	1 (histamine H_1 receptor)	6 (α_1A adrenoceptor)
		7 (α_1D adrenoceptor)
		7 (α_1B adrenoceptor)
		7 (5-HT_1A receptor)
		1 (D_2 dopamine receptor)
Denopamine	1 (β_1 adrenoceptor)	3 (β_3 adrenoceptor)
Ifenprodil	1 (NMDA receptor)	3 (μ opioid receptor)
RO-25–6981 (NMDA receptor ligand)	3 (NMDA receptor)	4 (D_4 dopamine receptor)
		1 (noradrenaline transporter)
		1 (5-HT transporter)

and then the values were averaged, yielding 92% for specificity and 16% for selectivity. We also obtained a minimum of 20% of adverse effects predicted for 43% of the compounds. These values appear to indicate that most of the adverse effects that could be predicted are missing but the false positive rate is within a reasonable range. With ElectroShape, we obtained better values, namely, a 95% of specificity and 22% of selectivity, which highlights the improved ability of this method to discover additional targets over the general chemical scaffold comparison merely using 2D fingerprints.

LBVS

Regarding the on-target validation of the dataset (cf. Figure 1), we selected the DUD and a subset of WOMBAT to avoid, to some extent, the effect of analogues in the true binders sets. In the case of the 4D method, we also tested the SEA (4D SEA) and the maximum similarity method (4D MAX).

By-target and average results for AUC of receiver operating characteristic plots are given in Table 3. A perfectly accurate method would have an AUC of 1, while a random method would have an AUC of 0.5. It can be seen that the 2D and 4D MAX methods perform equally and that both outperform the 4D SEA. A closer look at the WOMBAT results (Table 4) reveals that a strong analogue bias is present for all methods but especially for 2D SEA, the performance of which drops more than those of the two 4D methods.

It is worth noting that the average performance of our ElectroShape method compares favourably with those obtained through other means. In particular, the 4D MAX approach compared favourably with ROCS max ComboScore (Ballester *et al.*, 2009; reported average of 0.77). Besides, the non-superpositional ElectroShape 4D method is several orders of magnitude faster than ROCS (Grant *et al.*, 1996) because (i) the molecular descriptors are pre-calculated, (ii) a direct comparison requires just a Manhattan distance calculation and (iii) ElectroShape, unlike ROCS, does not require a computationally expensive superposition of one ligand onto another to compute the optimal ligand similarity.

Network analysis

Based on the target–target similarity matrices, we built several threshold networks with different cut-off values and compared several properties. For the case of ligand-based networks, comparison between ligands yielded *Z*-scores, which are directly transformable into *E*-values through an EVD fitting procedure. In the case of structure-based networks, the direct comparison between ligand-binding sites yielded TM-score values, which were used as the threshold criteria to trace an edge between two nodes.

To examine other properties of the network, we computed the union, intersection and differences between the three kinds of networks (ligand-based 2D and 4D, and receptor-based 3D) and calculated the triviality score as the percentage of the network edges built from plain-sight knowledge deduced from ligand structures and target names.

According to the network statistical parameters (cluster coefficient, characteristic path length and node degree distribution, which accounts for the distribution of the number of neighbours per node, Figure 2), the ligand-based network could be classified as a broad-scale and small-world network. This is so because it presents a high cluster coefficient, a short characteristic path (e.g. nodes are reachable from others within a few leaps, resulting in the small-world property, Figure 2) and a power law fitting of the node degree distribution (broad scale). Remarkably, the same classification is also present in the 3D target network, whereas the properties of the PSI-BLAST-based (Altschul *et al.*, 1997) network presented by Shoichet *et al.* (Hert *et al.*, 2008) were not compatible with the categorization of broad-scale and small-world networks.

Ligand-based networks appear to be populated by very close clusters, which are themselves mostly unconnected and represent pharmacological families such as GPCR, kinases, metabolic enzymes, proteases and nuclear hormone receptors (NHR). This trend can also be observed from the 3D target network where some clusters become even clearer, that is, the NHR family because their members share very similar ligand-binding sites (Figure 3).

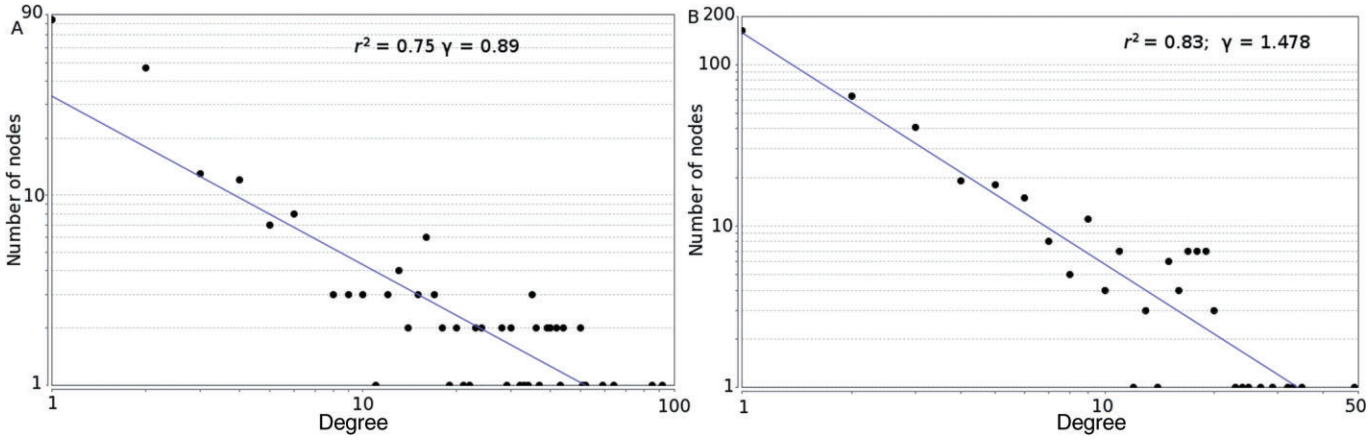

Figure 2

Node degree distribution for ElectroShape threshold network (A) and 3D binding site network (B).

Table 3

AUC results for the DUD using the three approaches tested

	4D SEA	4D MAX	2D SEA
ACE	0.59	0.71	0.81
AChE	0.57	0.61	0.64
ADA	0.59	0.64	0.91
ALR2	0.64	0.82	0.58
AmpC	–	–	0.55
AR	0.56	0.84	0.64
CDK2	0.55	0.78	0.72
COMT	0.54	0.63	0.80
COX1	0.55	0.74	0.69
COX2	0.66	0.91	0.46
DHFR	0.77	0.78	0.99
EGFR	0.77	0.73	0.95
ER_agonist	0.84	0.86	0.71
ER_antagonist	0.77	0.87	0.94
FGFR1	0.36	0.52	0.40
FXa	0.56	0.45	0.71
GART	0.93	0.92	0.77
GPb	0.70	0.81	0.86
GR	0.44	0.60	0.75
HIVPr	0.41	0.42	0.77
HIVRT	0.48	0.48	0.41
HMGA	0.86	0.96	0.97
HSP90	0.48	0.90	0.91
InhA	0.38	0.64	0.62
MR	0.69	0.84	0.88
NA	0.85	0.88	0.87
p38	0.37	0.65	0.66
PARP	0.71	0.79	0.92
PDE5	0.49	0.52	0.82
PDGFRb	0.46	0.43	0.24
PNP	0.58	0.56	0.97
PPARγ	0.80	0.79	0.93
PR	0.52	0.91	0.78
RXRα	0.87	0.82	0.97
SAHH	0.92	0.92	0.94
Src	0.70	0.66	0.41
Thrombin	0.35	0.63	0.61
TK	0.86	0.81	0.94
Trypsin	0.39	0.77	0.82
VEGFR2	0.53	0.63	0.60
Average	0.62	0.73	0.75
SD	0.17	0.15	0.19

ADA, adenosine deaminase; ALR2, aldose reductase; AmpC, AmpC β-lactamase; AR; androgen receptor; CDK2, cyclin-dependent kinase 2; DHFR, dihydrofolate reductase; EGFR, EGF receptor (kinase domain); ER_agonist, oestrogen receptor (agonist-bound conformation); ER_antagonist, oestrogen receptor (antagonist-bound conformation); FGFR1, fibroblast growth factor receptor 1 (kinase domain); FXa, factor Xa; GART, glycinamide ribonucleotide transformylase; GPb, glycogen phosphorylase β; GR, glucocorticoid receptor; HIVPr, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethylglutaryl-CoA reductase; HSP90, human heat shock protein 90; InhA, enoyl-[acyl-carrier-protein] reductase; MR, mineralocorticoid receptor; NA, neuraminidase; p38, p38 MAPK; PDGFRb, PDGF receptor β (kinase domain); PNP, purine nucleoside phosphorylase; PR, progesterone receptor; RXRα, retinoic X receptor α; SAHH, S-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase Src; TK, thymidine kinase; VEGFR2, VEGF receptor 2 (kinase domain).

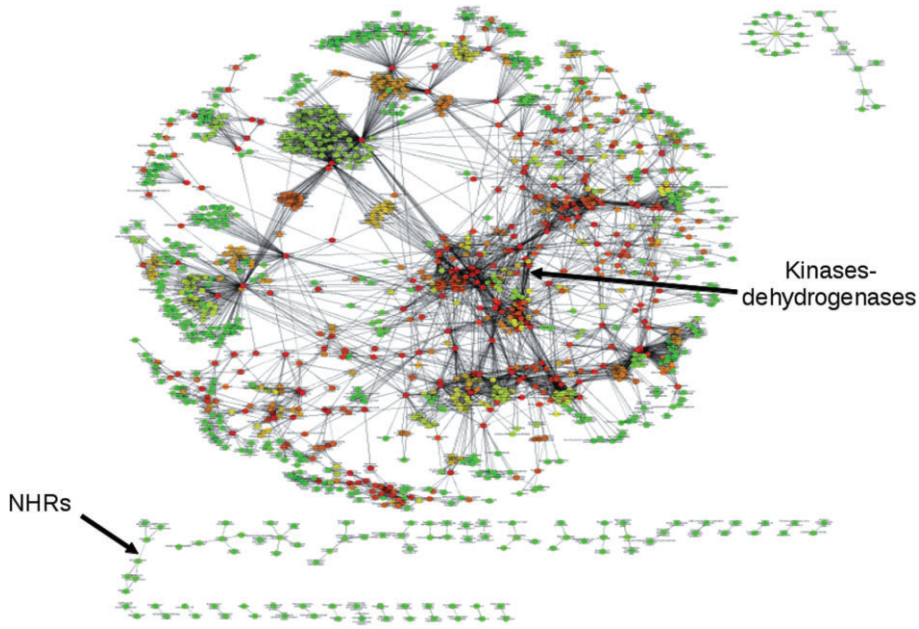

Figure 3

Small-world, broad-scale 3D receptor network with some pharmacologically relevant clusters highlighted (NHRs; kinases and dehydrogenases) and coloured by 'betweenness'.

Table 4

AUC results for the WOMBAT subset

	4D SEA	4D MAX	2D SEA
ALR2	0.56	0.67	0.42
AR	0.32	0.39	0.52
CDK2	0.51	0.67	0.69
COX2	0.58	0.81	0.43
EGFR	0.73	0.67	0.61
ER	0.65	0.73	0.66
FXa	0.50	0.43	0.77
HIVRT	0.45	0.45	0.45
p38	0.54	0.76	0.55
PDE5	0.25	0.33	0.66
PPARγ	0.83	0.81	0.72
Average	0.54	0.61	0.59
SD	0.17	0.17	0.12

ALR2, aldose reductase; AR; androgen receptor; CDK2, cyclin-dependent kinase 2; EGF receptor (kinase domain); ER, oestrogen receptor; FXa, factor Xa; HIVRT, HIV reverse transcriptase; p38, p38 MAPK.

Interestingly, the intersection of, and the difference between, the 3D target- and 4D ligand-based networks revealed that only a minimal fraction of the network is shared between them (from 0.36 to 6.80% in highly or minimally restrictive cut-off networks respectively). Low percentages may indicate that both methods yield complementary results: while the direct comparison of target ligand-binding sites could give valuable information in order to achieve some kind of target specificity, ligand-based networks could contribute with information about unexpected interactions for adverse effect prediction and polypharmacology profile optimization.

Regarding the presence of trivial information in the network, some dependence on the cut-off is observed. When the cut-off used to build the ligand-based network is low, the percentage of the network that could be considered trivial tends to increase, the score first rising due to very similar names and then, after a certain cut-off value, changing the regime to a shared-compound triviality (more than 60% of compounds in common with the other target).

In the case of 3D target networks, decreasing the cut-off has the opposite effect as it decreases the percentage of trivial edges in the network that are mostly related to a certain name similarity and independent from the number of shared compounds. This is in agreement with the origin of the network since no ligand comparisons were used to build it. In addition, the similarity in the name could be explained by the tendency of systematically naming targets that belong to the same families with similar names, which also tend to have similar functions and hence similar active (or ligand-binding) sites.

Discussion

Comparison with other methods

To our knowledge, only three other web tools are available for performing similar tasks. TarFisDock (Li *et al.*, 2006) applies the 'reverse docking' method, which consists of systemati-

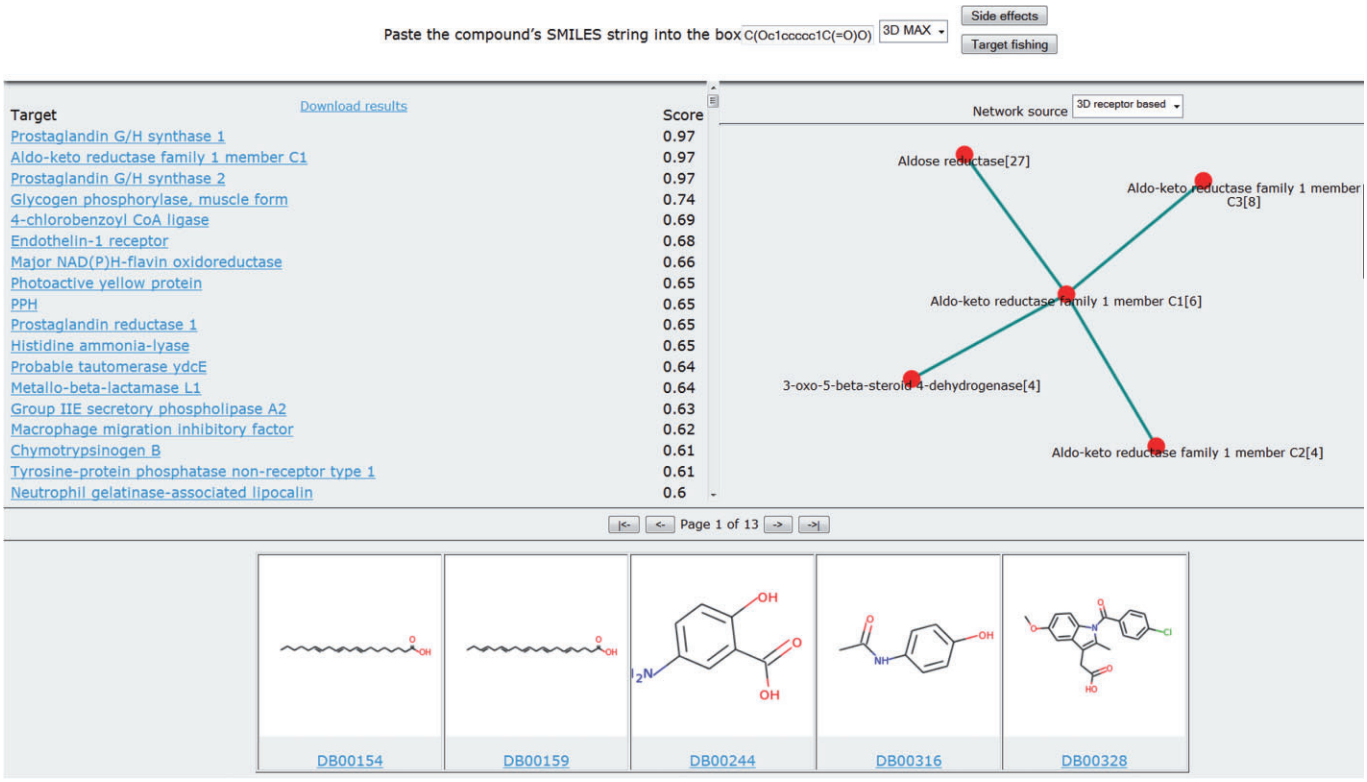

Figure 4

Screenshot of the ElectroShape Polypharmacology web server after submitting the SMILES string for acetylsalicylic acid as the query.

cally docking a ligand into hundreds of putative pharmacological target sites. This approach has the advantages of not requiring previous knowledge about true binders and not depending on functional information about the target. However, it needs a 3D model of the protein, which may be unknown or not available straightforwardly, and also suffers from the traditional problems associated with the docking methodology, namely, limited flexibility of the receptor, low speed, inaccurate scoring functions and excessive influence of the target conformation on the results. The SEA tool circumvents the docking problems by relating protein targets on the basis of the set-wise chemical similarity among all compounds that are known to interact with those targets (Keiser *et al.*, 2007). Thus, the problem is transferred to chemical space but at the expense of limiting the kind of compounds that can be predicted in a reliable manner. This method can also be used to search large chemical libraries rapidly and to build cross-target similarity maps. The third tool is STITCH 3 (Kuhn *et al.*, 2012), which allows the exploration of interactions between chemicals and targets on the basis of evidence from scientific documents. This resource currently contains interactions among 300 000 small molecules and 2.6 million proteins from 1133 organisms, and allows the visual display of interactive networks relating protein targets to which the same chemical binds.

Methodological advantages and limitations

The web tool presented here (http://ub.cbm.uam.es/chemogenomics/) implements the well-known SEA approach for target comparison using 2D ligand descriptors and also an ElectroShape comparison module that allows the estimation of 4D molecular similarities faster than other 3D methods. This is due to the pre-computation of the shape and partial charge distribution of the molecules in a format that can be stored in a database for further use, unlike the methods that need to perform *n* comparisons for every query. Our server also has a network explorer that allows the user to navigate the chemical (2D and 4D ligand) and the biological (target sites) regions of pharmacological space (Figure 4).

This methodology is obviously limited by the extent that pharmacological space is covered in current databases in terms of both compounds and targets. Therefore, only currently available information can be used to predict new targets or possible adverse effects for candidate molecules. It is also evident that chemical space has not been uniformly explored so that some parts of it (due to synthetic accessibility or other causes) can be better represented than others (or even not be represented at all).

Conclusions

In conclusion, we have presented here a new target-fishing approach that makes use of the ultra-fast LBVS ElectroShape methodology and is able to predict drug adverse effects, build polypharmacology profiles and relate targets from two complementary viewpoints (ligand- and target-based networks). The DUD and WOMBAT sets were employed for on-target

validation and the results were directly comparable to those obtained using other state-of-the-art target-fishing approaches. Off-target validation was performed using a limited set of non-annotated secondary targets for already known drugs. Finally, comparison of the predicted adverse effects with data contained in the SIDER 2 database showed good specificity and reasonable selectivity. All of these features are freely available from an undemanding and user-friendly web interface that (i) can be queried for both polypharmacology profiles and adverse effects, (ii) hyperlinks related targets in the three networks (2D, 4D ligand and 3D receptor) and (iii) displays the 2D structure of already annotated drugs.

Acknowledgements

A. C. C. gratefully acknowledges the FPU 2009-0203 grant and its foreign internship programme from the Spanish Ministry of Education. A. M. acknowledges financial support from Comunidad de Madrid through Fundación Severo Ochoa's AMAROUTO program. This research has been funded in part by the Spanish Comisión Interministerial de Ciencia y Tecnología (SAF2009-13914-C02-02) and Comunidad de Madrid (S2010-BMD-2457). The authors are grateful to Professor W. Graham Richards for encouragement and Dr Sree Vadlamudi for fruitful discussions during the elaboration of this work. We also would like to thank the three anonymous reviewers for their knowledgeable and insightful comments that helped to improve the quality of the original paper.

Conflict of interest

The authors declare no conflict of interest.

References

AbdulHameed MDM, Chaudhury S, Singh N, Sun H, Wallqvist A, Tawa G (2012). Exploring polypharmacology using a ROCS-based target fishing approach. J Chem Inf Model 52: 492–505.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI *et al*. (2010). ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. J Comput Aided Mol Des 24: 789–801.

Ballester PJ, Richards WG (2007). Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 28: 1711–1723.

Ballester PJ, Finn PW, Richards WG (2009). Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. J Mol Graph Model 27: 836–845.

Ballester PJ, Westwood I, Laurieri N, Sim E, Richards WG (2010). Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. J R Soc Interface 7: 335–342.

Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J *et al*. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem 2: 861–873.

Berger SI, Iyengar R (2009). Network analyses in systems pharmacology. Bioinformatics 25: 2466–2472.

Berman H, Henrick K, Nakamura H, Markley JL (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35 (Suppl 1): D301–D303.

Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P *et al*. (2012). Automated design of ligands to polypharmacological profiles. Nature 492: 215–220.

Cabrera AC, Gil-Redondo R, Perona A, Gago F, Morreale A (2011). VSDMIP 1.5: an automated structure-and ligand-based virtual screening platform with a PyMOL graphical user interface. J Comput Aided Mol Des 25: 813–824.

Cai J, Han C, Hu T, Zhang J, Wu D, Wang F *et al*. (2006). Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: reverse docking, enzymatic assay, and X-ray crystallography validation. Protein Sci 15: 2071–2081.

Ehrlich P (1907). On immunity with special reference to the relationship between distribution and action of antigens. J R Inst Public Health 15: 321–340.

Good AC, Oprea TI (2008). Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22: 169–178.

Grant JA, Gallardo M, Pickup B (1996). A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. J Comput Chem 17: 1653–1666.

Halgren TA (1996). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 17: 520–552.

Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model 50: 572–584.

Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK (2008). Quantifying the relationships among drug classes. J Chem Inf Model 48: 755–765.

Hopkins AL (2008). Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4: 682–690.

Huang N, Shoichet BK, Irwin JJ (2006). Benchmarking sets for molecular docking. J Med Chem 49: 6789–6801.

Jain AN (2000). Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. J Comput Aided Mol Des 14: 199–213.

Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007). Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25: 197–206.

Knight ZA, Lin H, Shokat KM (2010). Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10: 130–137.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A *et al*. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39 (Suppl 1): D1035–D1041.

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010). A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 6: 343.

Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P (2012). STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res 40 (database issue): D876–D880.

Landrum G (2011). Rdkit: Open-Source Cheminformatics. Novartis Institutes for BioMedical Research, Basel (http://www.rdkit.org/).

Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K *et al.* (2006). TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 34 (Web Server issue): W219–W224.

Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL *et al.* (2012). Large-scale prediction and testing of drug activity on side-effect targets. Nature 486: 361–367.

MDDR (2006). MDL Drug Data Report. MDL: San Leandro, CA.

Meslamani J, Rognan D, Kellenberger E (2011). sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. Bioinformatics 27: 1324–1326.

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011). Open Babel: an open chemical toolbox. J Cheminform 3: 33.

OpenEye Scientific Software (2011). ROCS. OpenEye Scientific Software: Santa Fe, NM.

Ripphausen P, Nisius B, Peltason L, Bajorath J (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. J Med Chem 53: 8461–8467.

Rogers DJ, Tanimoto TT (1960). A computer program for classifying plants. Science 132: 1115–1118.

Rognan D, Meslamani J (2011). Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. J Chem Inf Model 51: 1593–1603.

Roth BL, Sheffler DJ, Kroeze WK (2004). Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov 3: 353–359.

Rush TS 3rd, Grant JA, Mosyak L, Nicholls A (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48: 1489–1495.

Sadowski J, Gasteiger J, Klebe G (1994). Comparison of automatic three-dimensional model builders using 639 X-ray structures. J Chem Inf Comput Sci 34: 1000–1008.

Sadowski J, Schwab C, Gasteiger J (2003). CORINA, 3D Structure Generator: version 3.1, Molecular Networks GmbH, Erlangen, Germany.

Simon Z, Peragovics A, Vigh-Smeller M, Csukly G, Tombor L, Yang Z *et al.* (2012). Drug effect prediction by polypharmacology-based interaction profiling. J Chem Inf Model 52: 134–145.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432.

Van Der Greef J, McBurney RN (2005). Rescuing drug discovery: in vivo systems pathology and systems pharmacology. Nat Rev Drug Discov 4: 961–967.

Vidal D, Mestres J (2010). In silico receptorome screening of antipsychotic drugs. Mol. Inf. 29: 543–551.

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009). PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37 (Suppl 2): W623–W633.

Weininger D (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28: 31–36.

Witkop B (1999). Paul Ehrlich and his magic bullets – revisited. Proc Am Philos Soc 143: 540–557.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34 (suppl 1): D187–D191.

Yera ER, Cleves AE, Jain AN (2011). Chemical structural novelty: on-targets and off-targets. J Med Chem 54: 6771–6785.

Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33: 2302–2309.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1 Compounds used for off-target validation.
Table S2 Enrichment analysis of LBVS. BEDROC analysis ($\alpha = 20$, 80% weight for top 8%).
Appendix S1 Details on parameterization of the networks.