BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 3073–3079
doi:10.1093/bioinformatics/btt540

Systems biology

Advance Access publication September 17, 2013

Shaping the interaction landscape of bioactive molecules

David Gfeller[1], Olivier Michielin[1,2,3,*] and Vincent Zoete[1,*]

[1]Swiss Institute of Bioinformatics (SIB), Quartier Sorge, Bâtiment Génopode, CH-1015 Lausanne, Switzerland, [2]Ludwig Institute for Cancer Research and [3] Pluridisciplinary Center for Clinical Oncology, Centre Hospitalier Universitaire Vaudois, CH-1015 Lausanne, Switzerland

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Most bioactive molecules perform their action by interacting with proteins or other macromolecules. However, for a significant fraction of them, the primary target remains unknown. In addition, the majority of bioactive molecules have more than one target, many of which are poorly characterized. Computational predictions of bioactive molecule targets based on similarity with known ligands are powerful to narrow down the number of potential targets and to rationalize side effects of known molecules.

Results: Using a reference set of 224 412 molecules active on 1700 human proteins, we show that accurate target prediction can be achieved by combining different measures of chemical similarity based on both chemical structure and molecular shape. Our results indicate that the combined approach is especially efficient when no ligand with the same scaffold or from the same chemical series has yet been discovered. We also observe that different combinations of similarity measures are optimal for different molecular properties, such as the number of heavy atoms. This further highlights the importance of considering different classes of similarity measures between new molecules and known ligands to accurately predict their targets.

Contact: olivier.michielin@unil.ch or vincent.zoete@unil.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 2, 2013; revised on September 4, 2013; accepted on September 12, 2013

1 INTRODUCTION

A large number of small molecules ranging from metabolites to signaling molecules to drugs display strong bioactivity in different living systems. This activity is often mediated by physical interactions with proteins or other macromolecules. Therefore, information about the targets of bioactive molecules is crucial to understand, predict and interfere with their activity. In particular, it can be used (i) to predict unfavorable side effects due to off-target interactions and thus potentially decrease the attrition rate in clinical trials due to toxicity (Kola and Landis, 2004; Lounkine *et al.*, 2012), or (ii) to predict a new target for an approved drug and reposition it for another disease (Ashburn and Thor, 2004; Keiser *et al.*, 2009; Novac, 2013).

Experimental identification of bioactive molecule targets has received much attention (Ziegler *et al.*, 2013). This has led to technological developments of large facilities enabling researchers to screen a given molecule against arrays of targets, such as kinases. Chemogenomic strategies have also been introduced to identify the targets of bioactive molecules in model organisms such as yeast or bacteria (Smith *et al.*, 2010). Several databases have been developed by various groups to provide access to these data, such as ChEMBL (Gaulton *et al.*, 2012), DrugBank (Knox *et al.*, 2011), PubChem (Bolton *et al.*, 2008) or ZINC (Irwin *et al.*, 2012). These databases contain unprecedentedly large datasets of interactions between proteins and small molecules. For instance, only for human protein ligands, the ChEMBL database (Gaulton *et al.*, 2012) contains close to 350 000 reported direct interactions (i.e. annotated as binding) with activity $<10\,\mu$M involving >200 000 small molecules.

However, a significant fraction of bioactive molecules still do not have any known target. This is especially true for compounds tested uniquely in functional assays. For instance, 17.4% of the compounds in ChEMBL with reported functional activity in human cells do not have direct target information (see Methods Section 2.1). Moreover, even for well-studied molecules, our knowledge of their targets is far from complete. For instance, *N*,*N*-dimethyltryptamine was initially described as a ligand of sigma-1 receptor (Fontanilla *et al.*, 2009). Later on, it was shown to also bind hydroxytryptamine receptors (Keiser *et al.*, 2009). More generally, one can expect that even a significant fraction of Food and Drug Administration (FDA)-approved drugs have at least some unknown target.

Computational predictions of bioactive molecule targets are helpful to narrow down the set of potential targets to be tested and to predict off-target effects of known molecules or drugs (Keiser *et al.*, 2009; Kuhn *et al.*, 2013; Lounkine *et al.*, 2012). A widely used strategy consists in identifying proteins with known ligands similar to a query molecule (i.e. the so-called ‘ligand-based’ approach). Standard approaches use similarity measures between molecules based on chemical fingerprints (Dunkel *et al.*, 2008; Keiser *et al.*, 2009; Wang *et al.*, 2013). Historically, fingerprint similarity measures have been developed to classify molecules into families. Thus, they are powerful to identify molecules derived from the same chemical series.

More recently, other measures of similarity have been introduced (Ballester and Richards, 2007; Perez-Nueno *et al.*, 2012; Rahman *et al.*, 2009). For instance, it is known that the 3D shape of molecules plays an important role when binding to a target. Therefore, the ligands of a protein often display similarity in their shape. This similarity can be quantified by methods based on structural alignment (Gong *et al.*, 2013; Liu *et al.*, 2011; Sastry *et al.*, 2011) or shape recognition (Ballester and Richards, 2007; Ballester *et al.*, 2009; Wirth and Sauer, 2011),

[*]To whom correspondence should be addressed.

such as Ultrafast Shape Recognition. The latter technique was recently expanded to consider partial charges distribution (Armstrong *et al.*, 2010) and atomic lipophilicity (Armstrong *et al.*, 2011). A very attractive feature of shape comparison is the ability to detect similarities between molecules with different chemical structures, a property often referred to as 'scaffold hopping' (Renner and Schneider, 2006). Despite this advantage, previous studies suggest that fingerprint similarity gives higher performance, both in virtual screening (Venkatraman *et al.*, 2010) and in target predictions (Nettles *et al.*, 2006). Yet, it is unclear whether this is due to limitations of shape comparison (e.g. limited conformational sampling, shortcomings of shape descriptors) or to biases in benchmarking datasets (e.g. overrepresentation of ligands from the same chemical series).

Other approaches to predict targets of bioactive molecules have been proposed based on higher-level features such as side effects (Campillos *et al.*, 2008), transcriptional responses (Iorio *et al.*, 2010; Iskar *et al.*, 2013) and text-mining (Li *et al.*, 2009). Although these approaches are powerful to identify similarities between molecules without chemical similarity, they require additional experimental data and therefore are limited to molecules for which such data are available. For proteins with known structures, binding site similarity can be used to identify proteins that could accommodate similar ligands (Haupt and Schroeder, 2011). Moreover, in this case, docking algorithms (Grosdidier *et al.*, 2011; Morris *et al.*, 2009; Zoete *et al.*, 2009) can help predict whether a molecule can bind to a target (Li *et al.*, 2006; Wang *et al.*, 2012).

Here, we develop an original method of bioactive molecule target predictions that combines chemical and shape similarity (see Fig. 1). We perform extensive validation of the method and show that combining these two similarity measures leads to improved performance. This is especially true when predicting the targets of a molecule in the absence of other ligands originating from the same chemical series. Thus, our results suggest that the lower performance of shape similarity measures that was previously reported, is, at least partly, due to biases in available benchmarks (Cleves and Jain, 2008). Nevertheless, fingerprint-based methods still perform well even after stringent removal of all molecules that are trivially similar (e.g. with the same scaffold). We also observe that different combinations of similarity measures are optimal for different molecular properties.

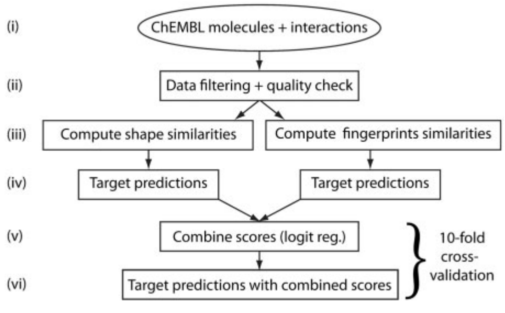
(i)
ChEMBL molecules + interactions
(ii)
Data filtering + quality check
(iii)
Compute shape similarities
Compute fingerprints similarities
(iv)
Target predictions
Target predictions
(v)
Combine scores (logit reg.)
(vi)
Target predictions with combined scores
10-fold cross-validation

Fig. 1. General workflow. (i) Molecules interacting with human targets are retrieved from ChEMBL. (ii) Filters are applied to remove large molecules and ambiguous interactions. (iii) The different similarities are computed for all pairs of molecules. (iv) Targets are predicted based on the most similar ligands excluding comparison with the query molecule, ligands with the same scaffold or tested in the same assays. (v) Regression coefficients are learned on the training set. (vi) Targets are predicted for molecules on the testing sets (10-fold cross-validation) and average AUC values over these molecules are computed. Steps (v) and (vi) apply only for the combined approach

2 METHODS

2.1 Datasets

Release 15 of the ChEMBL database (Gaulton *et al.*, 2012) was used throughout this work. Interactions were selected according to the following criteria: they should (i) involve human proteins, (ii) be annotated as direct binding ('assay_type' = 'B') with an activity (K_i, K_d, IC_{50} or EC_{50}) <10 μM, (iii) involve molecules consisting of <80 heavy atoms and (iv) involve targets that are single proteins or protein complexes (e.g. excluding targets corresponding to protein families and assays with a confidence level <4). We further discarded ambiguous interactions that had reported activity values both below and above 10 μM in different assays. This was done to address the observed uncertainty of many protein–small molecule interaction datasets (Kramer *et al.*, 2012). This results in a set of 347 889 interactions involving 1700 human proteins (1627) or protein complexes (73) and 224 412 molecules. As an additional benchmark, we also retrieved all ChEMBL molecules interacting with human proteins only with activities between 10 μM and 100 μM (i.e. none of them are part of the previous set of molecules). This consists of 79 682 molecules involved in 94 672 interactions (see Section 3.4). For all molecules, SMILES were retrieved from ChEMBL using the parent form.

To compute the fraction of molecules with functional activity but without direct target, we retrieved all molecules involved in assays with assay_type = 'F' in human using the same threshold of 10 μM (340 256 molecules in total). In all, 59 311 of them (17.4%) do not have direct targets in ChEMBL based on the two criteria: (i) no binding data or only binding activity >1000 μM, and (ii) target_type equal to 'ORGANISM', 'CELL-LINE', 'TISSUE' or 'ADMET'.

When determining whether two molecules have been tested in the same assay, all ChEMBL assays involving a human target were considered.

2.2 Ligand similarity measures

Shape similarity is defined as in the study by Armstrong *et al.* (2011). Each atom of a molecule is mapped to a 5D space, where the three first coordinates are determined by the 3D conformation of the molecule. The two remaining dimensions encode the atomic partial charges and the atomic lipophilicity (AlogP). Shape comparison is carried out by comparing the three first moments of the distribution of 5D-distances of every atom of the molecule to six different centroid positions (Armstrong *et al.*, 2011). Gasteiger atomic partial charges and contributions to molecular lipophilicity (AlogP) have been determined with the ChemAxon cxcalc tool (version 5.3.1) and OpenBabel (version 2.2.0), respectively. All other parameters correspond to the ones in the original publications (Armstrong *et al.*, 2010, 2011). A maximum of 20 different conformations have been generated for each molecule with ChemAxon cxcalc tool, after converting the SMILES from ChEMBL in 3D using the Chemaxon molconvert tool, and protonating it at pH = 7.4, using OpenBabel. For this, OpenBabel calculates which species (protonated or unprotonated) is preponderant at pH = 7.4 with the Henderson–Hasselbalch equation, and adjusts the number of protons and the charge of the molecule accordingly. Each conformation of a molecule was compared with each conformation of another one (resulting in 400 comparisons) and the highest similarity value was chosen. To avoid storing all similarity values between all pairs of molecules ($>5 \times 10^{10}$ numbers), similarity values <0.65 have been set to 0.65 and are not stored explicitly. The fingerprints similarity measure used in this work is based on the FP2 fingerprints implemented

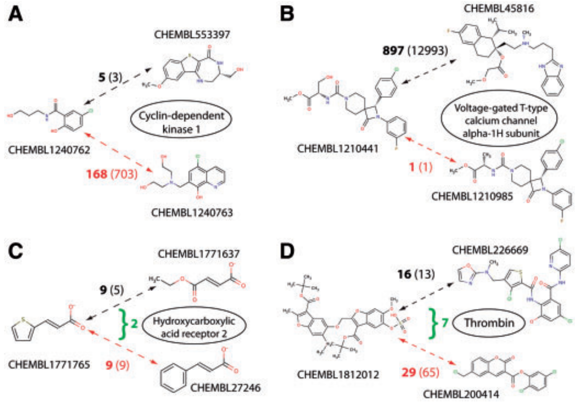

Fig. 2. Target predictions for four molecules with exactly one target in ChEMBL. Query molecules are shown on the left. Actual target names are displayed in black circles. Ligands displaying the highest similarity according to their shape or their fingerprints are displayed above (black arrow) or below (red arrow), respectively. Bold numbers correspond to the target rank using each similarity measure. Numbers in parenthesis show the rank of the ligands when computing similarity with all other molecules in ChEMBL. Green numbers show the target ranks obtained with the combined approach. A: shape similarity performs best. B: fingerprint similarity performs best. C and D: combining the two similarity values give the best predictions

in OpenBabel (version 2.2.0). Fingerprints are calculated directly from the molecular SMILES. Similarity between two molecules is measured as the Tanimoto coefficient. Similarity values lower than a threshold (here 0.25) have been set to 0.25 to allow us storing only a sparse version of the full similarity matrix. Figure 2 shows examples of ligands with shape and fingerprint similarity.

2.3 Scaffold identification

Molecular scaffolds were uniquely identified with the Strip-it^{TM} software (version 1.0.1) from Silicos-it. Here we used the OPREA-1 definition of molecular scaffolds (Pollock *et al.*, 2008), where any cycle is represented as an n-membered ring ($n = 3, 4, \ldots$), linkers between cycles as single bonds and all side chains are removed (Fig. 3). Based on visual inspection, this approach provided the most reasonable definition of scaffolds. Nevertheless, some important limitations arise, e.g. if aromatic rings are used as side chains within a series of molecules (Fig. 3).

2.4 Combining different similarity measures

To combine different similarity values, we used a multiple logistic regression, as implemented in *R* (bayesglm). The features used as inputs correspond to the similarity values based on shape and fingerprint similarities with either the most similar molecule (K = 1) or the K most similar ligands of each target (K = 5). For K = 1, the logistic regression is given by the following equation: $f(s_1, s_2) = (1{+}\exp[-a_0{-}a_1s_1{-}a_2s_2])^{-1}$, where s_1 stands for the shape similarity, s_2 for the fingerprint similarity and a_0, a_1 and a_2 are parameters learned by the model. In particular, a_i ($i = 1, 2$) is proportional to the slope of $f(s_1, s_2)$ along the i^{th} axis at the inflexion point. To enable comparison between coefficients a_1 and a_2, similarity values have been normalized between 0 and 1. Neural networks as implemented in *R* (nnet package, with one hidden layer consisting of two nodes) were also tested when using similarity with the K = 5 most similar ligands as input.

For the cross-validation, molecules were split in 10 different groups. Training of the model was done iteratively on nine groups and tested on the remaining one. A well-known issue when training and benchmarking binary classifiers of interactions is to have reliable negative data. Here we used negative data generated as follows. First, we retrieved all molecules in ChEMBL with binding activity >500 μM on some target (1799 interactions involving 1214 molecules). As these numbers are too low to train a model, we then randomly selected non-interacting ligand–protein pairs so as to have 10 times more non-interacting pairs than interacting pairs. In this selection, non-interacting pairs involving a protein with a paralog (BLAST E-value $<10^{-5}$) reported to interact with the query molecule were excluded from the list of negatives, as they could result from incompleteness in the ChEMBL database. Moreover, we further excluded all interacting pairs that have activity values between 10 and 100 μM, or for which a paralog of the protein has activity between 10 and 100 μM with the query molecule.

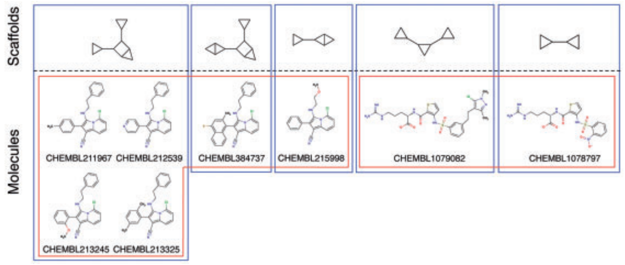

Fig. 3. Subset of neuropilin-1 (CHEMBL5174) ligands. Blue boxes show the molecules with the same scaffolds as defined with OPREA-1 (see Methods). Scaffolds are displayed above the dashed line. Red boxes show the molecules tested in the same assays. All other neuropilin-1 ligands would fall in the first column

The performance was measured with standard receiver operating characteristic (ROC) curves and area under the ROC curve (AUC). AUC values displayed in this work always correspond to the average of AUC values over all molecules. As our dataset is large and the number of model parameters small, we typically used a subset of molecules consisting of molecules with annotated non-interactions (>500 μM, see earlier in the text) to which 1000 randomly selected ligands were added. We observe in Supplementary Figure S1 that this approach is sufficient to compare different similarity measures. Similarly, when training models on different sets of molecules (e.g. based on properties such as the number of heavy atoms), all molecules with negative data were first considered and up to 1000 randomly selected molecules with the desired properties were added.

3 RESULTS

Identifying proteins with ligands similar to bioactive molecules is a strong indication of possible direct interactions that could predict or explain observed bioactivity. Here, we use the ChEMBL database to collect known ligands of human proteins and we explore different similarity measures, based on chemical structure (Fingerprints) and molecular shape (Electroshape) (Armstrong *et al.*, 2011), as well as combinations of these (see Methods). Target prediction is carried out following the workflow of Fig. 1. A query molecule is compared with all other molecules in ChEMBL to assign a score to each target corresponding to the similarity value with the most similar ligand. When combining different measures of similarity, we use logistic regressions to combine the similarity values obtained with each similarity measure (see Methods).

To illustrate the different strategies investigated in this work, in Figure 2, we show examples of ligands whose target is best predicted by shape similarity (Fig. 2A), fingerprint similarity

(Fig. 2B) as well as combination of both (Fig. 2C and D). In (Fig. 2A), the query molecule (CHEMBL1240762) has only one reported target (CDK1, $IC_{50} = 2.5\,\mu M$), and no other ligand of this protein is a simple analog of this molecule. Therefore, predictions based on fingerprint similarity rank this target at position 168 among all 1700 possible targets. Shape similarity instead identifies a ligand of CDK1 that displays similar overall shape and partial charges distribution, and based on the similarity with this molecule, CDK1 is ranked at the fifth position among all targets. Figure 2B shows a frequently encountered case, in which a clear analog (CHEMBL1210985) of the query molecule is present among the ligands of its target. In this case, fingerprint similarity is more appropriate. In particular, the missing hydroxyl groups in the most similar ligand detected by the fingerprint similarity (CHEMBL1210985) result in lower shape similarity that prevents accurate predictions. This example highlights some of the possible weaknesses of shape comparison and already suggests that combining different similarity measures is useful to optimally harness the information present among the ligands of a target. As examples of the combined strategy, we show in Figure 2C and D, typical cases where combining the two kinds of similarity results in better predictions. Here, both shape and fingerprint similarity values give reasonable predictions, with the cognate target being ranked between the 9th and 29th position. However, combining the two kinds of similarities results in significantly better performance, indicating how the presence of different ligands similar to the query molecule according to both similarity measures gives even higher confidence in the predictions.

More generally, Table 1, first row, shows the average cross-validation AUC values over all tested ligands. We can observe that fingerprint similarity performs better than shape similarity and the combined approach leads to the highest AUC, although all three approaches give very high AUC values. However, as discussed in the next section, these high AUC values mainly result from the habitual design of chemically related molecules to target the same protein. In particular, cases of close analogs among the ligands of a target (see example in Fig. 2B) appear very often in ChEMBL.

3.1 Addressing redundancy in benchmarking datasets

A major issue in benchmarking ligand-based target prediction methods comes from the strong biases observed in sets of protein–ligand interactions (Cleves and Jain, 2008; Rohrer and Baumann, 2008; Yera *et al.*, 2011). The standard strategy used in drug design or medicinal chemistry projects is to start from a lead compound and replace one by one the different fragments of this compound (Cleves and Jain, 2008). Therefore, a set of ligands targeting a protein consists often of highly similar molecules belonging to the same series (see Fig. 3). As a result, ligand-based approaches trained on such datasets tend to give over-optimistic predictions. For practical applications, it is important not to be restricted to this kind of direct similarity, as new bioactive molecules for which targets are to be predicted may typically not be part of existing series. Similarly, when attempting to predict off-target effects of existing drugs, it is unlikely that molecules from the exact same series have already been tested on the new targets (Yera *et al.*, 2011). These biases have been identified in smaller datasets, for instance, by observing a higher chemical similarity between molecules having the same primary targets compared with those having the same secondary targets (Cleves and Jain, 2008).

Table 1. AUC values for different similarity measures and different comparison schemes

Comparison	Shape	Fingerprints	Combined
All comparison	0.985	0.993	0.994
Distinct scaffolds	0.943	0.971	0.979
Distinct scaffolds + distinct assays	0.894	0.921	0.932

To address these issues, we first used information about the scaffold of molecules, as molecules of the same series often display a conserved scaffold (Pollock *et al.*, 2008; Schuffenhauer *et al.*, 2007). We reasoned that, when making predictions for a given compound, molecules with the same scaffold may typically not be present among the ligands of its targets. Therefore, we prevented comparison between molecules with the same scaffolds that are ligands of the actual targets in our cross-validation study (see Methods). Table 1, second row, shows the obtained AUC values. As expected, AUC values are lower than before. This further confirms that the very high AUC values reported in the first row of Table 1 are merely a consequence of the high level of redundancy among ligands in our dataset, and are only relevant if the query molecule is part of a well-known chemical series.

Although clustering molecules with respect to their scaffold helps identifying those coming from the same series, we still observe several cases of highly similar molecules that clearly come from the same series but have different perceived scaffolds (see example in Fig. 3). This is a well-known limitation of using scaffolds to unveil chemical series. However, to the best of our knowledge, there is no unified computational method that can unravel the chemical series among any set of molecules. Fortunately, in our case, we can use historical information present in ChEMBL, and especially assay numbers that can provide further clues about chemical series. Thus, as a second filter, we also prevented comparisons between molecules that had been tested in the same assay (see Methods). While this approach cannot fully exclude comparisons between molecules from the same series, visual inspection of many specific targets indicates that it already filters out many cases (see Fig. 3). The AUC values obtained when preventing comparison between molecules with the same scaffolds or tested in the same assay are shown in Table 1, third row. The same trend as before can be observed: the more we prevent comparisons between molecules from the same series, the lower the AUC. Importantly, we can observe that combining the two methods leads to significant improvements. This clearly suggests that the two approaches are complementary and combining them is useful to optimally capture the information present among the ligands of a target.

3.2 Influence of ligand properties

Different similarity measures may show different performance depending on the characteristics of the molecules. Here we

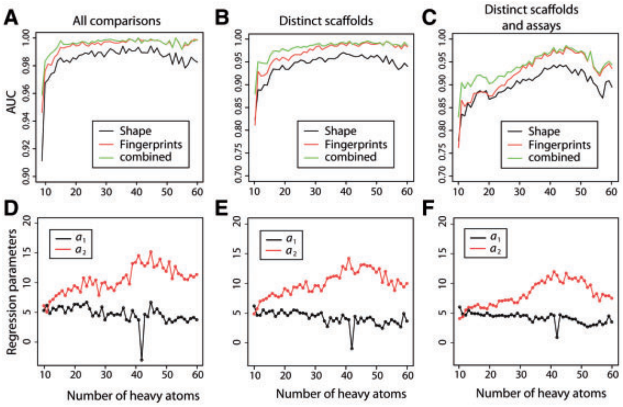

Fig. 4. AUC values for different molecule sizes. (**A**) The data obtained with the full ChEMBL dataset, (**B**) when restricting comparison to molecules with different scaffolds and (**C**) when restricting comparison to molecules with different scaffolds and not tested in the same assays. Note the different y-axis scales. (**D–F**) regression coefficients for the combined approach

investigate how the number of heavy atoms, the absolute total charge and the lipophilicity (logP) affect our predictions. We first cluster all 224 412 molecules investigated in this work according to their number of heavy atoms and trained separate linear regression models for each group. Figure 4A–C shows average cross-validation AUC values as a function of the number of heavy atoms for the three comparison schemes studied here. We observe that predicting targets of smaller molecules is more challenging. Moreover, our results indicate that combining different similarity measures is especially useful for molecules with ≤30 atoms, which corresponds approximately to half of the bioactive molecules targeting human proteins in ChEMBL (Fig. 5A), and to the majority of FDA-approved drugs (Fig. 5B). To gain further insights into the influence of the ligand size, we show in Figure 4D–F, the evolution of the two logistic regression coefficients a_1 and a_2 (see Methods). The higher these coefficients, the more discriminative the similarity value is for the predictions. We can observe that the two coefficients take similar values for small molecules, but coefficient a_2 (associated with fingerprint similarity) increases, whereas coefficient a_1 (associated with shape similarity) gently decreases with molecule sizes, further suggesting that shape similarity mainly helps discriminating interacting from non-interacting pairs for smaller size molecules. In one case (number of heavy atoms = 42), the fingerprint similarity was enough to drive the predictions, and a_1 takes values close to 0 or negative.

We also analyzed the effect of the absolute total charge and the lipophilicity (logP). Apart from being important properties of many ligands, we note that Electroshape similarity was developed to include partial charges and lipophilicity in the shape description (Armstrong *et al.*, 2011). Therefore, one could expect, for instance, that this similarity measure will perform better on predicting targets of charged molecules for which the presence of a charged group plays an important role. Overall, we do not observe a clear influence of these two molecular properties on our predictions. When grouping ligands according to the charge, the performances are not markedly different and optimal combinations of the two similarity measures are similar for different absolute charges (see Supplementary Fig. S2). When considering ligands with different lipophilicity, we can observe different behaviors for different logP values (Supplementary Fig. S3). However, these differences are largely explained by the correlation ($r = 0.45$) between the logP and the number of heavy atoms for the molecules present in our dataset (Supplementary Figs S4 and S5, and Supplementary Method S1). This suggests that lipophilicity has only a minor impact on our predictions. Overall, it appears that building separate models (here logistic regressions) for different numbers of heavy atoms is useful and informative, but not for different absolute total charges or logP values.

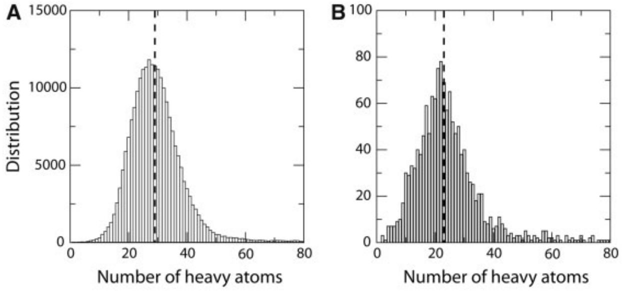

Fig. 5. (**A**) Distribution of the number of heavy atoms of ChEMBL molecules and (**B**) FDA-approved drugs. Dashed bars indicate the median

3.3 Similarity with many ligands

So far, we have based our predictions on the ligand of each target that displays the highest similarity with the query molecule, neglecting information about other ligands. However, one could imagine that being similar to several ligands could give higher confidence in the predictions. To explore this aspect, we trained a regression model using as features the similarities with the K = 5 most similar ligands of each target. If less than K ligands are known for a target, the minimal similarity value was used for all missing values. Results in Figure 6A–C indicate that only very little, if any, improvement is obtained. In this case, the number of parameters is larger (5 for each kind of similarity and 10 when combining them), suggesting that more elaborate machine learning models may prove beneficial. We also trained a neural network (see Methods), but did not observe any improvement either (see Supplementary Fig. S6).

3.4 Predicting lower affinity targets

To complement the previous cross-validation studies, we analyzed the set of molecules in ChEMBL that only have reported binding data between 10 and 100 μM. These molecules are not part of the training set and can be used to independently test our method. Figure 6D–F shows the AUC values obtained for different numbers of heavy atoms, using the regression coefficients of Figure 4F for the combined approach. AUC values are lower than in cross-validation studies as expected, as molecules with lower affinity display less similarity with other ligands of their cognate target. Importantly, the general trend that combining different kinds of similarity measures yields more accurate

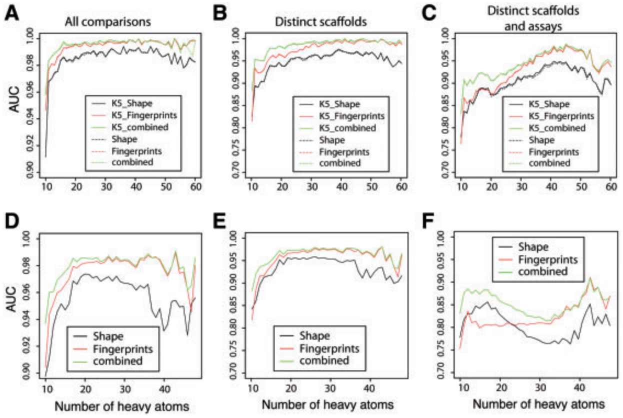

Fig. 6. (**A–C**) AUC values obtained when training the logistic regression with the similarity values of the K = 5 most similar ligands. Dashed curves correspond to AUC obtained with the most similar ligand (K = 1). (**D–F**) AUC values for target predictions of molecules interacting with human targets with activities between 10 μM and 100 μM. The same comparison schemes as in Figure 4 are used. Note the different x-axis and y-axis scales

predictions is confirmed, even in the absence of comparison restrictions between molecules with the same scaffold and/or tested in the same assays. These results suggest that our method is able to unveil weak affinity targets, which is useful for the detection of secondary targets, and for the identification of molecules that could then be optimized to improve the activity and efficiency once their targets are validated.

4 DISCUSSION

Accurate predictions of bioactive molecule targets are powerful to unravel the mechanisms of action of new molecules, guide experimental testing or predict off-target effects of known drugs. Several ligand-based strategies have been recently designed that use different kinds of similarity measures between small molecules (Armstrong *et al.*, 2011; Ballester and Richards, 2007; Campillos *et al.*, 2008; Keiser *et al.*, 2007; Wang *et al.*, 2013). Here, to have the broadest possible coverage, we focus on similarity measures that only use information from the 2D and 3D structures of the small molecule (i.e. without requiring data such as side effects or transcriptional response), and we tested our approach on the ChEMBL dataset of human protein ligands, which is one of the largest datasets of protein–small molecule interactions.

Our results indicate in general that fingerprint similarity performs better than shape similarity. However, as it was also observed in previous studies on smaller datasets (Cleves and Jain, 2008; Nettles *et al.*, 2006; Yera *et al.*, 2011), we show that this is partly due to the presence of structural analogs resulting from biases toward chemically similar molecules when developing compounds targeting a given protein. The differences between the two similarity measures become smaller when preventing comparisons between molecules with the same scaffold or tested in the same assays, and this trend is particularly strong for molecules with <30 heavy atoms. In those cases, combining the two methods gives significantly better performance. Considering that even our stringent approach to identify chemical series may not fully cover all cases (e.g. if two teams have developed molecules starting from the same lead compound, the resulting similar molecules will often not appear in the same assay), it clearly suggests that including molecular shape descriptors into target prediction approaches is useful for achieving the best accuracy.

Our work indicates that prediction accuracy changes with the size of molecules. Different mechanisms can explain this observation. First, for large molecules, the number of distinct conformers is typically high, and therefore, conformational sampling used in shape comparison may be more difficult. Second, a small change (e.g. replacing a carbon by a nitrogen in an aromatic ring) in a molecule with only a few heavy atoms will impact a large fraction of the fingerprints describing the molecule, whereas the same change in a large molecule will only impact a small fraction of the fingerprints. Therefore, fingerprint comparison may be slightly more sensitive to small changes for molecules with a lower number of heavy atoms. As a consequence, models trained for small molecules give almost equal weight to shape and fingerprint similarity, whereas models used for larger molecules rely more on fingerprint similarity values. We also stress that, while the heavy atom distribution of molecules in ChEMBL is peaked ~30 (median at 29), 75% of FDA-approved drugs have <30 heavy atoms (median at 23, see Fig. 5). Therefore, our combined approach is likely to be especially appropriate for target predictions of drug-like molecules.

Our results also show that similarity with the most similar ligand of a target is driving the predictions, and integrating similarity with other (less similar) ligands does not improve the performance in all our benchmarks. Although this came a bit as a surprise (one could imagine that being similar to many ligands would strengthen the confidence in the predictions), it appears that a similarity larger than a given value T with several ligands is almost always accompanied by a similarity T'>T with at least one ligand.

Overall, our comparison of the two kinds of similarity measures suggests that chemical similarity is more appropriate if other molecules with similar scaffolds are present among the ligands of a target. If not, and this might correspond to more realistic situations of new scaffolds being developed, combining shape comparison with chemical similarity gives the best performance. In particular, it has the potential of identifying similarities between molecules that do not display similar chemical structures, thereby raising the possibility of scaffold hopping. On the more technical side, our results indicate that one should first carefully study the biases in protein ligand datasets when comparing small molecule target prediction approaches. In particular, preventing comparison between molecules with the same automatically determined scaffold is often not sufficient to remove biases due to the presence of analogs. We also note that similarity measures tend to perform differently on molecules of different sizes, suggesting that different models should be applied depending on the nature of the query molecules. Although the current method has been applied only on molecules with activity in human, it could be expanded to different organisms, such as other vertebrates, yeast, fungi or plants, where large datasets of protein–small molecule interactions are also available and active molecules have interesting pharmaceutical or agricultural applications.

ACKNOWLEDGEMENTS

The authors thank Matthias Wirth and Antoine Daina for useful discussions. Computations were performed at the Vital-IT Center for high-performance computing of the SIB.

Conflict of interest: none declared

REFERENCES

Armstrong,M.S. *et al.* (2011) Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J. Comput. Aided Mol. Des.*, **25**, 785–790.

Armstrong,M.S. *et al.* (2010) ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided Mol. Des.*, **24**, 789–801.

Ashburn,T.T. and Thor,K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, **3**, 673–683.

Ballester,P.J. *et al.* (2009) Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J. Mol. Graph Model*, **27**, 836–845.

Ballester,P.J. and Richards,W.G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.*, **28**, 1711–1723.

Bolton,E. *et al.* (2008) PubChem: integrated platform of small molecules and biological activities. In: *Annual Reports in Computational Chemistry*. American Chemical Society, Washington, DC.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Cleves,A.E. and Jain,A.N. (2008) Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput. Aided Mol. Des.*, **22**, 147–159.

Dunkel,M. *et al.* (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.

Fontanilla,D. *et al.* (2009) The hallucinogen N,N-dimethyltryptamine (DMT) is an endogenous sigma-1 receptor regulator. *Science*, **323**, 934–937.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Gong,J. *et al.* (2013) ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics*, **29**, 1827–1829.

Grosdidier,A. *et al.* (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.*, **39**, W270–W277.

Haupt,V.J. and Schroeder,M. (2011) Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief Bioinform.*, **12**, 312–326.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci. USA*, **107**, 14621–14626.

Irwin,J.J. *et al.* (2012) ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model*, **52**, 1757–1768.

Iskar,M. *et al.* (2013) Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol. Syst. Biol.*, **9**, 662.

Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.

Keiser,M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Kola,I. and Landis,J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.*, **3**, 711–715.

Kramer,C. *et al.* (2012) The experimental uncertainty of heterogeneous public K(i) data. *J. Med. Chem.*, **55**, 5165–5173.

Kuhn,M. *et al.* (2013) Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.*, **9**, 663.

Li,H. *et al.* (2006) TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.*, **34**, W219–W224.

Li,J. *et al.* (2009) Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput. Biol.*, **5**, e1000450.

Liu,X. *et al.* (2011) SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J. Chem. Inf. Model*, **51**, 2372–2385.

Lounkine,E. *et al.* (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, **486**, 361–367.

Morris,G.M. *et al.* (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.*, **30**, 2785–2791.

Nettles,J.H. *et al.* (2006) Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.*, **49**, 6802–6810.

Novac,N. (2013) Challenges and opportunities of drug repositioning. *Trends Pharmacol. Sci.*, **34**, 267–272.

Perez-Nueno,V.I. *et al.* (2012) Detecting drug promiscuity using Gaussian ensemble screening. *J. Chem. Inf. Model*, **52**, 1948–1961.

Pollock,S.N. *et al.* (2008) Scaffold topologies. 1. Exhaustive enumeration up to eight rings. *J. Chem. Inf. Model*, **48**, 1304–1310.

Rahman,S.A. *et al.* (2009) Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.*, **1**, 12.

Renner,S. and Schneider,G. (2006) Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem*, **1**, 181–185.

Rohrer,S.G. and Baumann,K. (2008) Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model*, **48**, 704–718.

Sastry,G.M. *et al.* (2011) Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J. Chem. Inf. Model*, **51**, 2455–2466.

Schuffenhauer,A. *et al.* (2007) The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model*, **47**, 47–58.

Smith,A.M. *et al.* (2010) A survey of yeast genomic assays for drug and target discovery. *Pharmacol. Ther.*, **127**, 156–164.

Venkatraman,V. *et al.* (2010) Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model*, **50**, 2079–2093.

Wang,J.C. *et al.* (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic Acids Res.*, **40**, W393–W399.

Wang,L. *et al.* (2013) TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J.*, **15**, 395–406.

Wirth,M. and Sauer,W.H. (2011) Bioactive molecules: perfectly shaped for their target. *Mol. Inform.*, **30**, 677–688.

Yera,E.R. *et al.* (2011) Chemical structural novelty: on-targets and off-targets. *J. Med. Chem.*, **54**, 6771–6785.

Ziegler,S. *et al.* (2013) Target identification for small bioactive molecules: finding the needle in the haystack. *Angew. Chem. Int. Ed. Engl.*, **52**, 2744–2792.

Zoete,V. *et al.* (2009) Docking, virtual high throughput screening and in silico fragment-based drug design. *J. Cell. Mol. Med.*, **13**, 238–248.