Scope of 3D shape-based approaches in predicting the macromolecular targets of structurally complex small molecules including natural products and macrocyclic ligands

Ya Chen,1 Neann Mathai2 and Johannes Kirchmair1,2,3*

1 Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

2 Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

3 Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

ABSTRACT

A plethora of similarity-based, network-based, machine learning and docking approaches for predicting the macromolecular targets of small molecules are available today and recognized as valuable tools for providing guidance in early drug discovery. With the increasing maturity of
target prediction methods, researchers have started to explore ways to expand their scope to more challenging molecules such as structurally complex natural products and macrocyclic small molecules. In this work we systematically explore the capacity of an alignment-based approach to identify the targets of structurally complex small molecules (including large and flexible natural products and macrocyclic compounds) based on the similarity of their 3D molecular shape to non-complex molecules (i.e. more conventional, "drug-like", synthetic compounds). For this analysis, query sets of ten representative, structurally complex molecules were compiled for each of 35 pharmaceutically relevant proteins. Subsequently, ROCS, a leading shape-based screening engine, was utilized to generate rank-ordered lists of the potential targets of the 35x10 queries according to the similarity of their 3D molecular shape with that of compounds from a knowledge base of 272 640 non-complex small molecules active on a total of 3642 different proteins. Four of the scores implemented in ROCS were explored for target ranking, with the TanimotoCombo score consistently outperforming all others. The score successfully recovered the targets of 29% and 40% of the 350 queries among the top-5 and top-20 positions, respectively. For 29 out of the 35 investigated targets (83%), the method correctly assigned the first rank (out of 3642) to the target of interest for at least one of the ten queries. The shape-based target prediction approach showed remarkable robustness, with good success rates obtained even for compounds that are clearly distinct from any of the ligands present in the knowledge base. However, complex natural products and macrocyclic compounds proved to be challenging even with this approach, although cases of complete failure were recorded only for a small number of targets.
INTRODUCTION

The last decade has seen a boost in the development of in silico approaches for the prediction of the macromolecular targets of small molecules.\textsuperscript{1-3} Progress has been fueled by, among other factors, (i) the increasing amount of chemical and biological data available in the public domain, (ii) the strategic shift from the “one drug-one target” paradigm that had dominated small-molecule drug discovery for decades to the concept of polypharmacology,\textsuperscript{4} and (iii) advances in computational power and algorithms. Despite the rapid development however, it is challenging to obtain a realistic understanding of the performance of target prediction methods.\textsuperscript{5}

In silico approaches for target prediction can be classified into similarity-based, network-based, machine learning and docking methods. A large proportion of models reported in the scientific literature are available as free public web services or commercial tools.\textsuperscript{6} Most models utilize information from the largest public resources of chemical and biological data, PubChem\textsuperscript{7} and the ChEMBL database.\textsuperscript{8} PubChem currently contains more than 102 million compounds and 268 million bioactivity data points,\textsuperscript{9} and the latest release of the ChEMBL database contains close to 2 million compounds, with more than 15.5 million measured activities.\textsuperscript{10}

With the increasing coverage and reliability of the models, researchers have started to develop strategies for predicting the likely targets of more challenging compounds such as natural products,\textsuperscript{11,12} for which there is a notorious lack of available measured data,\textsuperscript{13} and macrocyclic compounds, characterized by a large number of conformational degrees of freedom in combination with distinct torsional angle preferences.\textsuperscript{14-16} For example, Reker et al.\textsuperscript{17} dissected the macrocyclic anti-tumor agent archazolid A and used pharmacophoric descriptions of these fragments to relate them to small molecules with known bioactivities. Several then unknown
targets of archazolid A that were predicted by this approach have subsequently been confirmed in biological tests. More recently, Cockroft et al.\textsuperscript{12} have reported on the development of a stacked ensemble approach which, despite being trained on data for synthetic compounds, is able to predict the macromolecular targets of natural products with good accuracy.

In silico methods based on the comparison of the 3D molecular shapes of aligned molecules are predestined for use in target prediction because of their ability to recognize similarity among structurally dissimilar compounds, as long as their molecular shapes (or at least parts of their molecular shapes) are preserved. Most shape-based methods take the distribution of chemical features (“color”) into account, which contributes substantially to their performance.\textsuperscript{18} They form the basis of several target prediction approaches\textsuperscript{19–21} and are also attractive tools for virtual screening and scaffold hopping.\textsuperscript{18,22,23}

Here, we systematically investigate the capacity of a leading 3D shape-based approach to identify the macromolecular targets of structurally complex small molecules (CSMs) on the basis of their molecular similarity with non-CSMs. In the context of small-molecule drug discovery, 3D shape-based screening, and this study alike, non-CSMs are compounds that medicinal chemists would identify as typical drug-like small molecules of low structural complexity. In contrast, CSMs represent less conventional compounds, characterized by higher molecular weight (larger size), larger numbers of conformational degrees of freedom and/or higher 3D shape complexity (Figure 1). CSMs include complex natural products and macrocyclic compounds, many of which are of high relevance to drug discovery but typically lack experimental data. Therefore, if it is found in this study that computational approaches based on 3D shape-based alignment are indeed capable of deriving the likely macromolecular targets of CSMs based on data measured for more conventional small molecules, this could open new
avenues to support drug discovery efforts in less densely populated, and hence more innovative, areas of the relevant chemical space.

Figure 1. Examples of CSMs and non-CSMs. Represented on the left are the three most diverse CSMs (used as queries in this study) identified for the HIV-1 protease, paired box protein Pax-8 and mu opioid receptor, and on the right the five most diverse non-CSMs (representing the knowledge base compounds). More details on the automated compound selection procedure are reported in the Methods section.
METHODS

Data extraction from ChEMBL

The ChEMBL database\cite{8,24} was processed following a protocol inspired by the work of Bosc et al.\cite{25} First, any data records matching the following criteria were extracted from ChEMBL:

1. Bioactivity record includes a molecular structure (canonical_smiles is not null)
2. Reported bioactivity is measured on a single protein or a protein complex (i.e. confidence_score 7 or 9)
3. data_validity_comment is null OR “manually validated”
4. duplicate_flag is “0”
5. activity_comment is not “inconclusive” OR “unspecified” (capitalization ignored)
6. standard_type is “Kd” OR “Potency” OR “AC50” OR “IC50” OR “Ki” OR “EC50”
7. NOT (standard_value is null AND pchembl_value is null AND activity_comment is not “active” (capitalization ignored))
8. NOT (standard_relation “>”, “>=”, or “>>” AND standard_value less than 20 000)

This data extraction procedure resulted in a total of 1 452 655 data records. A small number of these data records (2157) had concentrations applied to bioactivity measurements reported in µg.mL$^{-1}$ as opposed to nM; these values were converted into nM. Next, median activity values were calculated for unique compounds (defined by unique SMILES). Any compound with a median activity smaller than or equal to 10 000 nM was defined as active and all other compounds
discarded. This resulted in a total of 481,194 molecules, corresponding to 786,817 bioactivity records.

**Processing of molecular structures**

The molecular structures extracted from ChEMBL as SMILES were imported into MOE [26] (parsing failed for one molecule) and prepared using MOE’s Wash function. Processing included the removal of the minor components of salts, neutralization, and the addition of hydrogen atoms. Any molecules with a molecular weight in the range of 150 to 1,500 Da were kept. The molecules were then labeled “CSM” or “non-CSM” according to the following definition: CSMs are (i) macrocycles (i.e. molecules with at least one ring formed by at least 12 atoms) with 30 to 55 heavy atoms and (ii) non-macrocycles with 45 to 55 heavy atoms. Non-CSMs are non-macrocycles with 15 to 30 heavy atoms. Compounds consisting of more than 55 heavy atoms were discarded (because of the excessive size of their conformational space), as were very small molecules (less than 15 heavy atoms) and CSMs with at least one undefined chiral atom (to ensure that stereochemistry is unambiguously defined for all queries).

Next, conformers were generated with OMEGA [27,28]. OMEGA features a "default" or "classic" mode, which handles molecules with rings formed by up to nine atoms, and a macrocycle mode, which handles molecules with larger ring systems. Accordingly, OMEGA's default mode was employed to generate conformer ensembles with a maximum of 400 conformers per enantiomer for all non-CSMs without any rings formed by more than nine atoms (the flipper option, which enumerates the stereochemical configurations of undefined chiral atoms, was enabled). OMEGA’s macrocycle mode was employed to generate conformer ensembles (maximum of 400 per enantiomer) for any molecule with rings formed by more than nine atoms.
For any complex small molecule, a single conformer was generated with OMEGA’s default or macrocycle modes, applying the same ring size cutoffs as above. The number of CSMs and non-CSMs contained in this preprocessed data set are reported in Table 1.

**Table 1. Composition of the Processed Data set.**

|                     | Number of compounds | Number of bioactivity records | Number of targets |
|---------------------|---------------------|-----------------------------|------------------|
| Complex small molecules (CSMs) | macrocycles        | 2780                         | 4618            | 474\(^1\) |
|                     | non-macrocycles     | 10,870                       | 16,640          | 1164\(^1\) |
| Non-complex small molecules (non-CSMs) | non-macrocycles | 272,640                      | 460,047         | 3642    |

\(^1\) Corresponding to a total of 1318 unique targets.

**Compilation of a test set for target prediction**

A test set was compiled which covers the 35 proteins with the highest number of CSMs (between 108 and 730) in the processed ChEMBL data set (Table 2). Note that the delta opioid receptor from *Mus musculus*, the mu opioid receptor from *Cavia porcellus* and the mu opioid receptor
from *Homo sapiens* were excluded from this selection because larger and/or more balanced sets of CSMs were available for the respective receptors from *Rattus norvegicus*. In addition, the protein "protease" from human immunodeficiency virus 1 (CHEMBL2366517) was excluded because of the availability of a more comprehensive set of data on the protein "human immunodeficiency virus type 1 protease" (CHEMBL243).

For each of the 35 selected targets, the ten most diverse CSMs were determined with MOE’s function for the generation of diverse subsets (using MACCS fingerprints in combination with the Tanimoto coefficient). These 350 (35x10) compounds were used as queries for screening with ROCS.\textsuperscript{29,30} Note that the number of unique CSMs is 321 as a minority of the selected CSMs are active on more than one of the selected 35 proteins.

**Target prediction**

The 350 queries were screened with ROCS against the knowledge base consisting of 272 640 non-CSMs (active on a total of 3642 proteins). The proteins were ranked according to the maximum similarity between a CSM query and all non-CSMs recorded for a protein in the ChEMBL database. The similarity was quantified separately by each of four similarity metrics implemented in ROCS: ShapeTanimoto, TanimotoCombo, RefTverskyCombo and FitTverskyCombo score. As suggested by their names, metrics are either based on the Tanimoto or the Tversky coefficient. The Tanimoto coefficient quantifies the similarity of two molecules, \( f \) and \( g \), based on their self-volume overlaps (\( I_f \) and \( I_g \)) and the volume overlap between the two molecules (\( O_{f,g} \)):

\[
T_{\text{animoto}}_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}}
\]
The Tversky coefficient can be asymmetric (depending on the \textit{alpha} and \textit{beta} parameters chosen), hence allowing to emphasize either substructure or superstructure matching:

\[
Tversky_{f,g} = \frac{o_{f,g}}{a_{f} + \beta_{g}}
\]

The ShapeTanimoto score ranges from 0 to 1, with a value of 1 indicating a perfect fit of molecular shapes. Importantly, the ShapeTanimoto score only considers the fit of shapes for the volume overlap, whereas the three "combo" scores additionally take the type and distribution of chemical features into account. The "combo" scores typically range from 0 to 2, with equal weights applied to the shape and color components.

The RefTverskyCombo score assigns an \textit{alpha} value of 0.95 to the CSM query molecule as the main self-overlap term, meaning, in the context of this study, that it emphasizes the matching of the CSM (which by design of the data sets is the superstructure). The FitTverskyCombo score, on the contrary, assigns a \textit{beta} value of 0.95 to the fit (knowledge base) molecule, emphasizing the match of the non-CSM (substructure). Note that the RefTverskyCombo and FitTverskyCombo scores can have values greater than 2, because the overlap of two compounds can be larger than a molecule's self-overlap.

ROCS was run with factory settings with the following exceptions: both "-besthits" and "-maxhits" were set to "0" in order to cause ROCS to retain all results. The "-rankby" option was set to an appropriate value in order to have the results ranked by the four similarity metrics. For experiments using ShapeTanimoto score, the "-shapeonly" function was enabled in order to cause ROCS to align molecules by taking only molecular shape into account (and not color). Targets assigned identical scores were also assigned identical ranks.
RESULTS AND DISCUSSION

The capacity of shape-based approaches to predict the macromolecular targets of CSMs was explored with ROCS, a leading and widely applied 3D shape-based screening engine. As explained in detail in the Methods section, for the purpose of this study CSMs are compounds that are either macrocyclic and consisting of at least 30 heavy atoms, or non-macrocyclic and consisting of at least 45 heavy atoms. In contrast, non-CSMs are any compounds which consist of 15 to 30 heavy atoms.

ROCS was tested on the 35 proteins with the highest number of CSMs in the processed ChEMBL data set (Table 2; see Methods for details). For each of the 35 proteins, the ten most diverse CSMs recorded for that target were chosen for the evaluation (see Methods). These 350 selected CSMs (35 targets x 10 CSMs) served as queries for screening of a knowledge base of 272,640 non-CSMs (covering 3642 different proteins) with ROCS (Figure 2).

Figure 2. Schematic overview of the general approach.
Table 2. Overview of Targets Selected for Testing the Performance of the 3D Shape-Focused Target Prediction Approach.

| Target id     | Target name                                      | Target abbreviation | Organism                  | No. CSMs \(^1\) | No. non-CSMs \(^2\) |
|---------------|--------------------------------------------------|---------------------|---------------------------|-----------------|---------------------|
| CHEMBL243     | Human immunodeficiency virus type 1 protease     | HIV-1 protease      | Human immunodeficiency    | 703             | 185                 |
| CHEMBL340     | Cytochrome P450 3A4                              | CP3A4               | Homo sapiens              | 423             | 2945                |
| CHEMBL2362980 | Paired box protein Pax-8                         | PAX8                | Homo sapiens              | 390             | 465                 |
| CHEMBL270     | Mu opioid receptor                               | MOR                 | Rattus norvegicus         | 337             | 299                 |
| CHEMBL4616    | Ghrelin receptor                                 | GHSR                | Homo sapiens              | 299             | 127                 |
| CHEMBL2001    | Purinergic receptor P2Y12                         | P2Y12               | Homo sapiens              | 290             | 70                  |
| CHEMBL4822    | Beta-secretase 1                                  | BACE1               | Homo sapiens              | 289             | 1634                |
| CHEMBL3717    | Hepatocyte growth factor receptor                | HGFR                | Homo sapiens              | 274             | 800                 |
| CHEMBL269     | Delta opioid receptor                            | DOR                 | Rattus norvegicus         | 271             | 166                 |
| CHEMBL3948    | Angiotensin II type 1a (AT-1a) receptor          | AGTR1               | Oryctolagus cuniculus     | 266             | 43                  |
| CHEMBL4860    | Apoptosis regulator Bcl-2                        | BCL2                | Homo sapiens              | 266             | 84                  |
| CHEMBL ID  | Protein Name                                      | Gene Symbol | Species       | Fold Change | Log EC50 |
|------------|--------------------------------------------------|-------------|---------------|-------------|----------|
| CHEMBL203  | Epidermal growth factor receptor erbB1            | EGFR        | Homo sapiens  | 233         | 1451     |
| CHEMBL259  | Melanocortin receptor 4                           | MC4R        | Homo sapiens  | 233         | 85       |
| CHEMBL325  | Histone deacetylase 1                             | HDAC1       | Homo sapiens  | 192         | 1453     |
| CHEMBL1957 | Insulin-like growth factor I receptor              | IGF1R       | Homo sapiens  | 177         | 514      |
| CHEMBL2820 | Coagulation factor XI                             | F11         | Homo sapiens  | 173         | 15       |
| CHEMBL1824 | Receptor protein-tyrosine kinase erbB-2           | HER2        | Homo sapiens  | 171         | 288      |
| CHEMBL5023 | p53-binding protein Mdm-2                         | MDM2        | Homo sapiens  | 156         | 183      |
| CHEMBL5658 | Prostaglandin E synthase                          | PGES        | Homo sapiens  | 153         | 288      |
| CHEMBL1981 | Insulin receptor                                  | INSR        | Homo sapiens  | 147         | 43       |
| CHEMBL5251 | Tyrosine-protein kinase BTK                       | BTK         | Homo sapiens  | 147         | 83       |
| CHEMBL286  | Renin                                            | REN         | Homo sapiens  | 144         | 84       |
| CHEMBL4414 | Plasmepsin 2                                      | PM2         | Plasmodium falciparum | 144 | 15 |
| CHEMBL220  | Acetylcholinesterase                              | AChE        | Homo sapiens  | 130         | 1083     |
| CHEMBL2327 | Neurokinin 2 receptor                             | NK2R        | Homo sapiens  | 129         | 45       |
| CHEMBL2954 | Cathepsin S                                      | CTSS        | Homo sapiens  | 123         | 424      |
| CHEMBL ID  | Protein Name                          | Gene Symbol | Species          | Column1 | Column2 |
|------------|--------------------------------------|-------------|------------------|---------|---------|
| CHEMBL4662 | Proteasome Macropain subunit MB1     | MB1         | Homo sapiens     | 121     | 73      |
| CHEMBL4625 | Apoptosis regulator Bel-X            | BCLX        | Homo sapiens     | 120     | 73      |
| CHEMBL3952 | Kappa opioid receptor                | KOR         | Cavia porcellus  | 119     | 825     |
| CHEMBL240  | HERG                                 | HERG        | Homo sapiens     | 117     | 2260    |
| CHEMBL244  | Coagulation factor X                 | F10         | Homo sapiens     | 115     | 277     |
| CHEMBL3572 | Cholesteryl ester transfer protein   | CETP        | Homo sapiens     | 114     | 26      |
| CHEMBL1865 | Histone deacetylase 6                | HDAC6       | Homo sapiens     | 112     | 1070    |
| CHEMBL1829 | Histone deacetylase 3                | HDAC3       | Homo sapiens     | 111     | 406     |
| CHEMBL3706 | ADAM17                               | ADAM17      | Homo sapiens     | 108     | 256     |

1 Number of ligands that are complex small molecules.

2 Number of ligands that are non-complex small molecules.
Performance of shape-based screening with different similarity metrics

ROCS features two different alignment modes: a default mode, which takes into account both molecular shape and color, and the shape-only mode, which considers molecular shape only. Both of these alignment modes were assessed in this study with different scores implemented in ROCS in the following setups: the default alignment mode in combination with the TanimotoCombo, RefTverskyCombo and FitTverskyCombo scores, and the ShapeTanimoto score in combination with ROCS' shape-only mode.

Among the four investigated scores, the TanimotoCombo score clearly outperformed all other scores in ranking the targets of CSMs among the top positions of 3642 proteins (Table 3 and Figure 3a; note for the figure that steeper curves indicate worse performance and that the y-axis is on a logarithmic scale). With the TanimotoCombo score, the target of interest (i.e. the target assigned to this particular query) was ranked among the top-5 positions for 102 (29%) of the 350 CSM queries (note that the automated query selection procedure resulted in the selection of 23 CSMs which are active on more than one of the 35 targets; accordingly, these CSMs represent more than one query). The success rate increases to 40% when considering the top-20 ranks, and to 46% when considering the 40 top-ranked proteins (which corresponds to roughly 1% of the total list of proteins represented by the knowledge base).

Compared to the TanimotoCombo score, the success rates obtained by the ShapeTanimoto, RefTverskyCombo and FitTverskyCombo scores were roughly 20 percentage points lower. The RefTverskyCombo score tended to have higher success rates than the ShapeTanimoto and FitTverskyCombo scores when considering a greater number of ranks (top-40, top-80 and top-200).
Figure 3. Percentage of queries for which the target of interest (out of 3642 proteins) was assigned ranks better than or equal to the ranks indicated on the y-axis, for (a) all queries, (b) non-macrocyclic queries and (c) macrocyclic queries. Note that steeper curves indicate worse performance and that the y-axis is on a logarithmic scale.

The scatter plots in Figure 4 visualize the relationship between scores and ranks obtained for the individual CSM-target pairs. As seen in the figure, a wide range of scores correspond to targets that were appropriately ranked within the top ranks. The correlation between the rank of the target of interest and the score for this CSM-target pair is stronger for the three "combo" scores than for the ShapeTanimoto score. For the ShapeTanimoto score in particular, high scores often do not correspond to high rankings (lack of a tail towards the bottom right corner of the plot, as opposed to the plots for the other scores), indicating a lack of specificity of the approach that is expected to be related to the neglect of chemistry. Figure 5 visualizes the distributions of ShapeTanimoto vs. TanimotoCombo scores (scaled to have identical ranges). It shows that high scores are more frequently observed for the ShapeTanimoto than the TanimotoCombo score, which is plausible because good overlaps of molecular shapes are more likely when chemical features (color) are not taken into account. Because of the emphasis on the coverage of the knowledge base molecule (the substructure), the FitTverskyCombo score is higher than the other
scores for most CSM-target pairs. Clearly, this shows that values obtained with different scores should not be directly compared. Moreover, the scores obtained for individual query-target combinations should not be used as a measure of the likelihood of a compound to be active on that target. In other words, the predictions provide an indication of the likelihood of a protein being a target only relative to all other possible targets.

Table 3. Success Rates for Predicting the Targets of interest of the Queries with Different Scoring Functions.

| Rank       | TanimotoComb | ShapeTanimoto | RefTverskyCombo | FitTverskyCombo |
|------------|--------------|---------------|----------------|----------------|
|            | o score      | score         | score          | score          |
| All/macrocyclic/non-macrocyclic complex small molecules (CSMs) [%] |
| Top-5      | 29/20/31     | 9/2/11        | 10/7/11        | 8/7/9          |
| Top-10     | 36/25/38     | 15/5/17       | 11/8/12        | 10/7/11        |
| Top-20     | 40/28/43     | 21/8/23       | 22/15/23       | 14/8/15        |
| Top-40     | 46/31/49     | 25/8/28       | 35/20/38       | 20/10/22       |
| (~1%)      |              |               |                |                |
| Top-80     | 53/43/55     | 35/20/38      | 46/28/50       | 30/20/32       |
| Top-200    | 63/64/63     | 51/39/53      | 60/61/60       | 47/49/46       |
Figure 4. Relationship between the (a) TanimotoCombo, (b) RefTverskyCombo, (c) FitTverskyCombo and (d) ShapeTanimoto scores, and the ranks obtained for the targets of interest of the 350 CSM queries. Note that there is one instance where the FitTverskyCombo score is greater than 2.0 (see Methods for explanation).
Figure 5. Density distributions of the ShapeTanimoto and TanimotoCombo scores over all lists of scores obtained for all 350 queries. The TanimotoCombo score values were scaled in order to have the same range as the ShapeTanimoto score.

A further way of analyzing success rates is on a per-target basis, evaluating the results for query sets (the ten queries) rather than individual queries. For 29 of the 35 targets (83%), the TanimotoCombo score assigned the top rank to the target of interest for at least one of the ten queries (Figure 6). For the ShapeTanimoto, RefTverskyCombo and FitTverskyCombo scores, this was only the case for 40%, 49% and 26% of the 35 proteins, respectively. Additional details are provided in Table 4.
Figure 6. Ranks assigned with the TanimotoCombo score to the target of interest for the 350 CSM queries. Note that the y-axis is on a logarithmic scale. The numbers reported at the bottom of the graph indicate the number of CSM queries for which the target of interest was assigned the rank of 1 (indicating perfect prediction); the dashed line indicates the rank of 10.

Only for four out of 35 targets, the TanimotoCombo score failed to rank the target of interest among the top-10 positions with any of the ten queries: the paired box protein Pax-8 (*Homo sapiens*), plasmepsin 2 (*Plasmodium falciparum*), neurokinin 2 receptor (*Homo sapiens*), and cholesteryl ester transfer protein (*Homo sapiens*).
Table 4. Best and Median Target Ranks Obtained by Different Scores for the Query Sets Consisting of Ten CSMs Each.

| Protein\(^1\) | Target rank with score | TanimotoCombo | RefTverskyCombo | FitTverskyCombo | ShapeTanimoto |
|---------------|------------------------|---------------|-----------------|-----------------|---------------|
|               | best | median | best | median | best | median | best | median |
| HIV-1 protease | 1.0 | 116.0 | 1.0 | 135.0 | 2.0 | 381.5 | 7.0 | 356.0 |
| CP3A4         | 1.0 | 42.0 | 2.0 | 56.5 | 1.0 | 29.5 | 9.0 | 64.0 |
| PAX8          | 32.0 | 294.0 | 83.0 | 315.0 | 80.0 | 216.0 | 126.0 | 253.0 |
| MOR           | 1.0 | 1.0 | 16.0 | 19.5 | 12.0 | 88.0 | 1.0 | 34.0 |
| GHSR          | 1.0 | 260.0 | 1.0 | 213.5 | 11.0 | 794.0 | 4.0 | 349.0 |
| P2Y12         | 1.0 | 1.5 | 1.0 | 24.0 | 1.0 | 67.0 | 1.0 | 185.5 |
| BACE1         | 1.0 | 162.0 | 16.0 | 320.0 | 32.0 | 304.5 | 54.0 | 197.0 |
| HGFR          | 1.0 | 87.5 | 1.0 | 84.5 | 6.0 | 162.5 | 1.0 | 59.0 |
| DOR           | 1.0 | 4.5 | 16.0 | 22.0 | 37.0 | 152.5 | 1.0 | 59.0 |
| AGTR1         | 1.0 | 2.0 | 1.0 | 2.0 | 3.0 | 89.5 | 2.0 | 20.5 |
| BCL2          | 1.0 | 236.5 | 16.0 | 188.5 | 153.0 | 705.0 | 1.0 | 280.5 |
| EGFR          | 1.0 | 4.5 | 1.0 | 18.0 | 1.0 | 69.5 | 1.0 | 59.0 |
| MC4R          | 1.0 | 233.0 | 28.0 | 475.5 | 25.0 | 274.0 | 1.0 | 289.5 |
| HDAC1         | 1.0 | 21.5 | 1.0 | 63.0 | 1.0 | 96.0 | 1.0 | 78.5 |
| IGF1R         | 1.0 | 25.0 | 1.0 | 29.0 | 1.0 | 310.0 | 1.0 | 126.5 |
| Target | F11  | HER2 | MDM2 | PGES | INSR | PM2  | AChE  | NK2R | CTSS | MB1 | BCLX | KOR | HERG | F10  | CETP  | HDAC6 | HDAC3 | ADAM17 |
|--------|------|------|------|------|------|------|-------|------|------|-----|------|-----|------|------|-------|-------|-------|--------|
|        | 1.0  | 2.0  | 1.0  | 1.0  | 5.0  | 420.0| 1.0   | 96.0 | 1.0  | 1.0 | 1.0  | 1.0 | 1.0  | 1.0  | 53.0  | 1.0   | 1.0   | 1.0    |
|        | 774.0| 93.5 | 240.5| 6.0  | 131.5| 1308.5| 3.0   | 712.0| 18.5 | 132.5| 912.5| 156.0| 12.5 | 28.5 | 625.0 | 39.5  | 55.5  | 102.5  |
|        | 1.0  | 3.0  | 2.0  | 1.0  | 8.0  | 534.0| 1.0   | 305.0| 1.0  | 8.0  | 1.0  | 16.0 | 1.0  | 1.0  | 1063.0| 16.0  | 3.0   | 1.0    |
|        | 901.0| 159.5| 326.0| 41.0 | 191.5| 1257.0| 47.5  | 908.5| 64.0 | 116.5| 724.0| 79.0 | 49.0 | 74.5 | 1772.0| 84.5  | 3.0   | 1.0    |
|        | 139.0| 16.0 | 3.0  | 3.0  | 27.0 | 636.0| 1.0   | 83.0 | 1.0  | 17.0 | 6.0  | 37.0 | 1.0  | 10.0 | 1722.0| 5.0   | 5.0   | 4.0    |
|        | 1765.0| 291.5| 235.0| 285.5| 745.5| 1225.0| 29.5  | 372.5| 88.0 | 136.0| 1220.0| 367.5| 81.5 | 420.5| 443.5  | 89.5  | 105.0 | 229.0  |
|        | 1.0  | 1.0  | 1.0  | 1.0  | 5.0  | 440.0| 17.0  | 287.0| 4.0  | 5.0  | 2.0  | 7.0  | 1.0  | 1.0  | 484.0  | 11.0  | 4.0   | 2.0    |
|        | 462.5| 49.0 | 143.5| 8.0  | 798.5| 1452.0| 41.0  | 921.5| 99.0 | 529.5| 474.0| 310.0| 62.0 | 58.5 | 484.0  | 166.0 | 166.0 | 222.0  |

1 For the explanation of all target acronyms see Table 2.
For the paired box protein Pax-8, the highest rank obtained with any of the ten queries was 32 (TanimotoCombo score). One of the reasons for failure is the fact that most of the CSMs active on this target are very different from the bioactive non-CSMs in terms of chemistry. They are characterized by long and flexible scaffolds; a minority are macrocyclic (indicated in Figure 6).

In the case of plasmepsin 2, the best rank obtained was just 420 (TanimotoCombo score). This target is characterized by a highly flexible ligand binding site to which small molecules are known to bind in several distinct modes. The fact that there were only 15 non-CSMs recorded for that target may contribute to the difficulties in recognizing CSMs active on this protein (note however that coagulation factor XI was correctly identified as the target of two out of the ten CSMs and ranked among the top-3 positions even though the target is represented by only 15 non-CSMs in the knowledge base).

For the neurokinin 2 receptor, the best rank obtained with any of the ten CSMs was 96 (TanimotoCombo score). The reasons for failure appear to be similar to those for Pax-8: Most of the CSMs have a substantial number of rotatable bonds; a minority are macrocyclic.

For the cholesteryl ester transfer protein, the best rank obtained with any of the ten CSMs was 53 (TanimotoCombo score). The CSM queries of the cholesteryl ester transfer protein are characterized by three to four similarly sized branches originating from a central carbon or nitrogen atom. The structures of most CSM queries are clearly distinct from those of the ligands represented in the knowledge base.

Overall, the results obtained on a per-target basis indicate that the value of the method can be substantially higher in cases where clusters of related compounds are explored rather than a singleton. A further fact that can be distilled from Figure 6 is that there is no correlation between
the success rates for a target and the number of non-CSM representing that target in the knowledge base.

**Performance on macrocyclic and non-macro cyclic complex small molecules**

Sixty-one of the 350 CSMs are macrocyclic, covering 19 out of the 35 targets studied in this work. Our results show that the task of target prediction is more challenging for macrocyclic compounds than for non-macro cyclic ones (Figure 3b and c). For the TanimotoCombo score, the top-5, top-10, top-20 and top-40 success rates for non-macro cyclic CSMs were 31%, 38%, 43% and 49%, whereas for macrocyclic CSMs they were just 20%, 25%, 28% and 31%, respectively. Besides the low molecular similarity of macrocyclic compounds with the non-CSMs of the knowledge base, a major reason for the lower success rates observed for macrocyclic compounds are the complexities involved in representing the 3D conformations of these queries, related to a high number of conformational degrees of freedom and torsional properties that are distinct from non-macro cyclic compounds.

**Cases where at least one score worked well while others failed**

There are several examples of CSMs for which their targets were ranked at high positions with one score while other scores failed. We identified nine CSMs (three thereof being macrocyclic compounds) for which their targets were assigned ranks of ten or better by at least one score, while other score(s) assigned ranks of 450 or worse (Table 5). In seven out of the nine cases, the TanimotoCombo score performed well while others failed (Figure 7a, b); in two cases the ShapeTanimoto score outperformed the other scores (Figure 7c, d). For the examples reported in Table 5, it can be seen that the alignments produced by the three combo scores are generally more consistent in terms of chemistry (in particular with regard to the orientation of chemical
features) than the alignments produced by the ShapeTanimoto score. However, the FitTverskyCombo score failed to identify the target of interest for many CSMs due to its emphasis on matching the knowledge base molecule (substructure). More specifically, the FitTverskyCombo score assigns high scores to many small knowledge base molecules, thus giving rise to a high false-positive prediction rate. In contrast, the ShapeTanimoto score often failed because of its disregard of chemistry, which is reflected by alignments that lack the matching of chemical features.
Table 5. Examples of CSMs for which Their Targets were Successfully Identified by One at Least One Score While Others Failed.

| Query<sup>1</sup> | Target<sup>2</sup> | Rank by score | Alignments of CSM queries and reference compounds obtaining the highest rank<sup>3</sup> | Alignments of CSM queries and reference compounds obtaining the lowest rank<sup>3</sup> |
|------------------|------------------|----------------|-------------------------------------------------|-------------------------------------------------|
| CHEMBL 3699200* | F11              | 1 208 1 1649   | TanimotoCombo score CHEMBL3393362; CHEMBL3355664 | ShapeTanimoto score CHEMBL3355686 |
| CHEMBL 3676156* | F11              | 3 549 27 815   | TanimotoCombo score CHEMBL3393362; CHEMBL3355664 | ShapeTanimoto score CHEMBL3355685 |
| CHEMBL  |  Protein | TanimotoCombo score | ShapeTanimoto score |
|---------|----------|---------------------|---------------------|
| CHEMBL53424 | BACE1 | 2 | 32 |
| CHEMBL1760861 | | 578 | |
| CHEMBL1795908 | REN | 5 | 79 |
| CHEMBL508748 | | 561 | |
| CHEMBL | CETP  | 53  | 93  | 1351 | 6     | Shape/Tanimoto score | Ref/TverskyCombo score  |
|--------|-------|-----|-----|------|-------|----------------------|------------------------|
| 3683924|       |     |     |      |       | CHEMBL340397         | CHEMBL340397           |

| CHEMBL | MOR   | 41  | 461 | 85   | 3     | Shape/Tanimoto score | Fit/TverskyCombo score  |
|--------|-------|-----|-----|------|-------|----------------------|------------------------|
| 445869*|       |     |     |      |       | CHEMBL171763         | CHEMBL2048969           |
1 Queries marked with a "*" are macrocyclic compounds.

2 F11, coagulation factor XI; BACE1, beta-secretase 1; REN, renin; AGTR1, angiotensin II type 1a (AT-1a) receptor; PGES, prostaglandin E synthase; CETP, cholesteryl ester transfer protein; MOR, mu opioid receptor.

3 ChEMBL IDs reported are those that obtained the highest/lowest rank for the target of interest of the individual CSM queries, according to the scoring function indicated in the respective table cells. Alignments shown are those that obtained the highest rank for a CSM query. In cases where multiple alignments obtained identical scores (and ranks), only one alignment is shown.
Figure 7. Ranks assigned to the targets of interest of the 350 CSM queries by the (a) TanimotoCombo, (b) RefTverskyCombo, (c) FitTverskyCombo and (d) ShapeTanimoto scores. The nine compounds for which one score produced good results while others failed are highlighted in blue.

Performance as a function of molecular similarity

The performance of similarity-based approaches depends on how well the query is represented by the data stored in the knowledge base. We quantified the distance between the individual queries and their nearest ligand recorded in the knowledge base by the Tanimoto coefficient
derived from 2D Morgan2 fingerprints. As shown in Figure 8, ROCS (in combination with the TanimotoCombo score) ranked 42% of all CSMs with a maximum Tanimoto coefficient between 0.2 and 0.3 among the top-10 positions, and 75% of all CSMs with a coefficient between 0.3 and 0.4. This robustness is remarkable, as molecular structures with a Morgan2 fingerprint-based Tanimoto coefficient below 0.4 are clearly distinct in most cases.

Figure 8. Fraction of CSM queries for which the target of interest was ranked among the top-\(k\) positions ("success rate") as a function of the 2D molecular similarity between the CSM query and the most similar ligand (non-CSM) recorded in the knowledge base. The graph shows the performance of the TanimotoCombo score. The Tanimoto coefficients based on Morgan2 fingerprints are binned with a bin size of 0.1. Success rates for queries with a Tanimoto coefficient greater than 0.7 are not reported because of the limited number of examples.
Among the 350 queries investigated in this work, we identified eleven compounds (six of them are macrocyclic compounds) for which their target was ranked within the top-10 positions out of 3642 targets, despite being structurally dissimilar from any ligands (non-CSMs) recorded in the knowledge base (Tanimoto coefficient lower than 0.18). As shown in Table 6, most of the alignments produced by ROCS for the eleven compounds are not only plausible and sensible from a chemistry point of view, but also visually easily interpretable, thanks to the hard Gaussians used by ROCS for chemical features (color), which cause a lock-in of the alignment on hydrogen bond donors and acceptors.

We did not observe any cases of CSMs for which their targets were not ranked early in the hit list and at least one known ligand shared a high degree of 2D similarity with the query (note that the number of CSMs in this category was small).
Table 6. Examples of CSMs for Which Their Targets were Successfully Identified Despite Being Dissimilar from Any Reference Compound.

| Query    | Closest Reference compound | 3D alignment | 2D Tanimoto similarity | Combo Combo score | Target rank | Target with TanimotoCombo |
|----------|---------------------------|-------------|------------------------|-------------------|-------------|-------------------------|
| CHEMBL584549* | CHEMBL493517 | ![Image](image1) | ![Image](image2) | 0.08 | 0.71 | 7 | HERG |
| CHEMBL2170017* | CHEMBL3356937 | ![Image](image3) | ![Image](image4) | 0.12 | 0.73 | 1 | HDAC1 |
| CHEMBL3621333* | CHEMBL3415568 | 0.12 | 0.77 | 1 | AChE |
| CHEMBL508629 | CHEMBL225421 | 0.13 | 1.09 | 1 | PGES |
| CHEMBL1783518 | CHEMBL413793 | 0.13 | 1.05 | 5 | AChE |
| CHEMBL124139 | CHEMBL104253 | 0.15 | 0.78 | 5 | HIV-1 protease |
| CHEMBL503270 | CHEMBL455681 | 0.15 | 0.83 | 1   | HERG |
|--------------|--------------|------|------|-----|------|
| ![Image](image1) | ![Image](image2) | ![Image](image3) | ![Image](image4) | ![Image](image5) |

| CHEMBL1917826* | CHEMBL1819169 | 0.16 | 1.11 | 1   | AChE |
|----------------|--------------|------|------|-----|------|
| ![Image](image6) | ![Image](image7) | ![Image](image8) | ![Image](image9) | ![Image](image10) |

| CHEMBL3676156* | CHEMBL3393362;CHE | 0.17 | 1.07 | 3   | F11  |
|----------------|-------------------|------|------|-----|------|
| ![Image](image11) | ![Image](image12) | ![Image](image13) | ![Image](image14) | ![Image](image15) |
| CHEMBL524997* | CHEMBL317520 | 0.18 | 1.30 | 1 | HERG |
|---------------|--------------|------|------|---|------|
| ![Image](image1.png) | ![Image](image2.png) | ![Image](image3.png) |

| CHEMBL243062 | CHEMBL3402709 | 0.18 | 0.79 | 8 | AChE |
|--------------|--------------|------|------|---|-----|
| ![Image](image4.png) | ![Image](image5.png) | ![Image](image6.png) |

1 Queries marked with a "*' are macrocyclic compounds.

2 2D molecular similarity between the CSM query and the closest ligand recorded in the knowledge base (measured as Tanimoto coefficient based on Morgan2 fingerprints).

3 HDAC1, histone deacetylase 1; AChE, acetylcholinesterase; PGES, prostaglandin E synthase; HIV-1 protease, human immunodeficiency virus type 1 protease; F11, coagulation factor XI.
Performance as a function of common substructures

In analogy to the trends observed above, target rankings are expected to improve with the size of the maximum common substructure (MCS) shared between the CSM query and the closest related non-CSM in the knowledge base (as determined by ROCS). The results presented in Figure 9 confirm this assumption: For the TanimotoCombo score, the median ranking of the targets of interest was 5 for CSMs sharing a MCS of at least 20 heavy atoms with the closest ligand (non-CSM) recorded in the knowledge base whereas the median target rank was just 110 for CSMs with an MCS of 15 to 19 heavy atoms. The median target ranks obtained by the RefTverskyCombo, FitTverskyCombo and ShapeTanimoto scores were substantially lower (worse): 25, 87 and 42 for CSMs sharing an MCS of at least 20 heavy atoms, and 287.5, 327 and 256 for CSMs with an MCS of 15 to 19 heavy atoms, respectively. We repeated this analysis using the percentage of heavy atoms rather than absolute numbers covered by the MCSs, and observed the same trends (data not shown).
Figure 9. Ranks obtained for the targets of interest as a function of the size of the MCS shared between the CSM queries and most similar ligand (non-CSM) recorded for the respective target, for the (a) TanimotoCombo, (b) RefTverskyCombo, (c) FitTverskyCombo and (d) ShapeTanimoto scores. The lines are merely a guide for the eye and indicate the median values of the target rankings in relation to the size of the MCS.

Performance on natural products

By overlapping the queries with a data set of 201,761 natural products compiled as part of the work reported in ref 32, we determined that at least eight out of the 321 (unique) CSMs are natural products (which is a surprisingly low portion of natural products). Using NP-Scout, a
machine learning approach for identifying natural products in large molecular libraries, we identified an additional 30 CSMs with a high likelihood (probability > 0.70) of being natural products (these compounds are natural product-like). The 38 natural products and natural product-like compounds cover a total of 25 different targets; 14 of the queries are macrocyclic.

Using the TanimotoCombo score, ROCS ranked the targets of interest of the natural products among the top-10 positions for only eight out of 48 queries (17%; the 48 queries result from the 38 unique natural products and natural product-like compounds). This success rate is considerably lower than the ones averaged over all 350 queries (36%), all 289 non-macro cyclic queries (38%) and all macrocyclic queries (25%), indicating that the prediction of the targets of complex natural products is more challenging than of complex synthetic molecules. A main reason for the low prediction success rates is the fact that the similarity of complex natural products and natural product-like compounds, and the non-CSM of the knowledge base is generally low.

CONCLUSIONS

Under real-life conditions, an in silico method capable of ranking the targets of CSMs among the top-20 out of more than 3600 (covered) targets may well be a valuable tool for prioritizing research efforts in early drug discovery because (i) it is likely that researchers will be able to rule out many of the proteins wrongly predicted as targets based on their expert knowledge on a compound of interest and (ii) in many cases, clusters of related compounds rather than a singleton will be explored, which allows predictions for multiple queries to be combined in order to improve the signal-to-noise ratio of the target prediction approach. Given these usage scenarios, we have shown that chances are good for molecular shape-based methods to identify
the macromolecular targets of small molecules, even if they are of high molecular complexity and only distantly related to any known bioactive compounds. More specifically, we found that ROCS, in combination with the TanimotoCombo score, ranked the targets of 29% of the 350 investigated CSM queries among the top-5 ranks of hit lists of more than 3600 proteins (40% if the top-20 ranks are considered). For 29 of the 35 proteins (83%), the target of interest was ranked at the top position with at least one of the ten queries. The robustness of ROCS in terms of recognizing distant similarity among pairs of molecules is impressive, and a further important advantage of this particular screening engine is that it produces alignments that are easily interpretable thanks to the lock-in effect caused by the use of hard Gaussians for chemical features. This enables chemists to quickly judge the plausibility and reliability of predictions. Researchers will also be able, in many cases, to refute likely false predictions and to understand when queries are outside the applicability domain. Taking all observations into consideration, we believe that shape-based approaches are predestined for use in target prediction for CSMs and molecules for which data on structurally related compounds are scarce, although also these approaches are challenged by macrocyclic compounds and structurally complex natural products. With the increasing amount of bioactivity data, the reach and value of these and related methods will continue to improve.

AUTHOR INFORMATION

Corresponding Author

*J. Kirchmair. E-mail: johannes.kirchmair@univie.ac.at. Tel.: +43-1-4277-55104
NOTES

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

The authors thank Christina de Bruyn Kops from the Universität Hamburg and Christoph Bauer from the University of Bergen for valuable discussions, and OpenEye for providing an academic license for the use of OMEGA and ROCS.

FUNDING

YC is supported by the China Scholarship Council (201606010345). NM and JK are supported by the Trond Mohn Foundation (BFS2017TMT01).

ABBREVIATIONS

CSM, complex small molecule

MCS, maximum common substructure

REFERENCES

(1) Sydow, D.; Burggraaff, L.; Szengel, A.; van Vlijmen, H. W. T.; IJzerman, A. P.; van
Westen, G. J. P.; Volkamer, A. Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 1728–1742.

(2) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for in silico Target Fishing. *Methods* **2015**, *71*, 98–103.

(3) Agamah, F. E.; Mazandu, G. K.; Hassan, R.; Bope, C. D.; Thomford, N. E.; Ghansah, A.; Chimusa, E. R. Computational/in silico Methods in Drug Target and Lead Prediction. *Brief. Bioinform.* **2019**. https://doi.org/10.1093/bib/bbz103.

(4) Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov.* **2017**, *16*, 531–543.

(5) Mathai, N.; Chen, Y.; Kirchmair, J. Validation Strategies for Target Prediction Methods. *Brief. Bioinform.* **2019**. https://doi.org/10.1093/bib/bbz026.

(6) Sam, E.; Athri, P. Web-Based Drug Repurposing Tools: A Survey. *Brief. Bioinform.* **2019**, *20*, 299–316.

(7) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109.

(8) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutow, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, S.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.

(9) PubChem. PubChem. https://pubchem.ncbi.nlm.nih.gov/ (accessed Feb 4, 2020).

(10) ChEMBL Database. https://www.ebi.ac.uk/chembl/ (accessed Feb 4, 2020).

(11) Moumbock, A. F. A.; Li, J.; Mishra, P.; Gao, M.; Günther, S. Current Computational
Methods for Predicting Protein Interactions of Natural Products. *Comput. Struct. Biotechnol. J.* 2019, 17, 1367–1376.

(12) Cockroft, N. T.; Cheng, X.; Fuchs, J. R. STarFish: A Stacked Ensemble Target Fishing Approach and Its Application to Natural Products. *J. Chem. Inf. Model.* 2019, 59, 4906–4920.

(13) Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* 2017, 57, 2099–2111.

(14) Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* 2011, 54, 1961–2004.

(15) You, L.; An, R.; Liang, K.; Cui, B.; Wang, X. Macro cyclic Compounds: Emerging Opportunities for Current Drug Discovery. *Curr. Pharm. Des.* 2016, 22, 4086–4093.

(16) Mallinson, J.; Collins, I. Macrocycles in New Drug Discovery. *Future Med. Chem.* 2012, 4, 1409–1438.

(17) Reker, D.; Perna, A. M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Mönch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; Müller, R.; Schubert-Zsilavecz, M.; Werz, O.; Schneider, G. Revealing the Macromolecular Targets of Complex Natural Products. *Nat. Chem.* 2014, 6, 1072–1078.

(18) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* 2009, 49, 678–692.

(19) Gfeller, D.; Michielin, O.; Zoete, V. Shaping the Interaction Landscape of Bioactive Molecules. *Bioinformatics* 2013, 29, 3073–3079.

(20) Potshangbam, A. M.; Polavarapu, R.; Rathore, R. S.; Naresh, D.; Prabhu, N. P.;
Potshangbam, N.; Kumar, P.; Vindal, V. MedPServer: A Database for Identification of Therapeutic Targets and Novel Leads Pertaining to Natural Products. *Chem. Biol. Drug Des.* **2019**, *93*, 438–446.

(21) Gong, J.; Cai, C.; Liu, X.; Ku, X.; Jiang, H.; Gao, D.; Li, H. ChemMapper: A Versatile Web Server for Exploring Pharmacology and Chemical Structure Association Based on Molecular 3D Similarity Method. *Bioinformatics* **2013**, *29*, 1827–1829.

(22) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. “Scaffold-Hopping” by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed Engl.* **1999**, *38*, 2894–2896.

(23) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front Chem* **2018**, *6*, 315.

(24) ChEMBL Database version 25. [https://www.ebi.ac.uk/chembl/](https://www.ebi.ac.uk/chembl/) (accessed May 15, 2019).

(25) Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large Scale Comparison of QSAR and Conformal Prediction Methods and Their Applications in Drug Discovery. *J. Cheminform.* **2019**, *11*, 4.

(26) Chemical Computing Group: Montreal, QC, Canada. *Molecular Operating Enviroment (MOE)*.

(27) OpenEye Scientific Software. OMEGA 3.1.1.2: OpenEye Scientific Software, Santa Fe, NM. [https://www.eyesopen.com](https://www.eyesopen.com) (accessed Nov 13, 2019).

(28) Hawkins, P. C. D.; Geoffrey Skillman, A.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*. 2010, pp 572–584. [https://doi.org/10.1021/ci100031x](https://doi.org/10.1021/ci100031x).

(29) ROCS 3.3.1.2: OpenEye Scientific Software, Santa Fe, NM. [https://www.eyesopen.com](https://www.eyesopen.com)
(30) Hawkins, P. C. D.; Geoffrey Skillman, A.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *Journal of Medicinal Chemistry* 2007, 50, 74–82.

(31) Rasina, D.; Otikovs, M.; Leitans, J.; Recacha, R.; Borysov, O. V.; Kanepe-Lapsa, I.; Domraceva, I.; Pantelejevs, T.; Tars, K.; Blackman, M. J.; Jaudzems, K.; Jirgensons, A. Fragment-Based Discovery of 2-Aminoquinazolin-4(3H)-Ones As Novel Class Nonpeptidomimetic Inhibitors of the Plasmepsins I, II, and IV. *J. Med. Chem.* 2016, 59, 374–387.

(32) Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* 2019, 9. https://doi.org/10.3390/biom9020043.
