Chapter 3

Chemical Similarity Networks for Drug Discovery

Yu-Chen Lo and Jorge Z. Torres

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65106

Abstract

Chemical similarity networks are an emerging area of interest in medicinal chemistry, chemical biology, and systems chemoinformatics that are currently being applied to drug target prediction, drug repurposing, and drug discovery in the new paradigm of poly-pharmacology and systems biology. In this chapter, we discuss the network-based drug target identification and discovery framework called chemical similarity network analysis pull-down (CSNAP) and its applications. We highlight the utility of CSNAP in identifying novel antimitotic drugs and their targets through practical case studies.

Keywords: drug discovery, chemical similarity networks, target identification

1. Introduction

Chemical similarity is an important concept in drug discovery used to identify compounds with similar bioactivities based on structural similarity between two ligands [1, 2]. Once a lead compound has been discovered from a chemical screen, a drug designer can design a series of structural analogues with improved pharmaceutical properties. The fundamental principle behind similarity-based drug discovery is the "chemical similarity principle," which states that if two molecules share similar structures, then they will likely have similar bioactivities. While there are exceptions, correlation between chemical structure and compound activities has been well established in medicinal chemistry [3]. Consequently, determining whether two molecules are structurally similar is a prerequisite for similarity-based drug discovery. At a rudimentary level, the similarity between two ligands can be easily discerned through visual inspection by identifying common functional groups, structural motifs, or substructures. However, human intervention is often subjective and not suitable for large-scale analysis.

Thus, applying computational algorithms for unbiased chemical similarity comparison and database searching is essential for a successful drug discovery campaign.

Several computational chemical similarity search algorithms have been developed [1, 4, 5]. The most commonly used approaches use chemical substructure fingerprints. Non-hashed structural fingerprints such as MACCS keys or Obabel FP3 fingerprints detect predefined substructures or functional group patterns in a molecule by mapping common chemical motifs into binary arrays known as structural keys. To compare the chemical similarity between two molecules, each molecule is converted into a binary series of 0 and 1, indicating the absence and presence of a particular substructure. On the other hand, chemical hashed fingerprints such as Daylight fingerprints or Obabel FP2 fingerprints use path information derived from molecular graphs to compare chemical structures [4]. While path-based fingerprints usually confer higher specificity, structural fingerprints can nevertheless be useful for detecting hits with distinct chemical scaffolds. Once the chemical fingerprints have been determined in a chemical search and the molecules have been converted to appropriate data representations, the next step is to evaluate the chemical similarity using a distance metric. Common distance measures include Euclidean, Manhattan, and Mahalanobis metrics, which have been widely applied in chemoinformatics and bioinformatics applications [6]. However, in the case of binary chemical fingerprints, the simplest and most direct distance measure is the Tanimoto index. Tanimoto metrics calculate the fraction of shared bits between chemical fingerprints in the range of 0–1. Although there is no universal Tanimoto index cutoff (Tc) to determine whether two molecules are sufficiently similar, a Tc value of 0.7 is a reasonable starting point for most chemical searches. Alternatively, statistical scores such as a *Z*-score can also be calculated based on the overall Tc score distribution [7].

In addition to 2D fingerprints, 3D chemical similarity fingerprints have also been developed. 3D chemical similarity fingerprints utilize the 3D structural information of the ligands such as molecular shape, pharmacophore points, or molecular interaction fields (MIF) for structural similarity comparison. Although 3D chemical similarity comparison can often capture structural features essential for protein-ligand binding, 3D alignment algorithms often require extensive optimization procedures to maximize the overlapped volume and are computationally intensive. Alternatively, nonalignment methods based on chemical descriptors such as GETAWAY or 3D-MoRSE descriptors can also be used [8, 9]. The 3D chemical descriptor is capable of capturing 3D ligand properties from 2D information and may improve computational time. However, substantial postvalidation may be required to confirm 3D structural similarity.

2. Network-based target prediction and drug discovery

The application of chemical similarity searches for ligand bioactivity prediction has recently gained substantial interest [10]. Due to the high failure rate of many new chemical entities (NCE) in the late stage of clinical trials, understanding on- and off-target binding of a drug to predict mechanisms of action and adverse reactions has become crucial for drug discovery programs [11]. If the chemical structure of a compound is known, then it is possible to predict compound bioactivities based on the chemical similarity methods described previously. Drug targets can be inferred from bioactivity databases with annotated targets sharing the highest similarity to the target molecules. Many public bioactivity databases are freely available and can be applied to this application including ChEMBL, PubChem, DrugBank, and Binding Database to name a few [12–14].

The simplest approach for drug target inference is by a simple chemical similarity search where the target of a query compound is inferred from the annotated ligand sharing the highest chemical similarity (**Figure 1**). However, there are several limitations to this approach. First, target information for the reference molecules may be incomplete; thus, target inference from a single molecular entity can miss potential targets from molecules sharing lower chemical similarity. Likewise, pair-wise target predictions may not provide consistent predictions for a group of structurally similar ligands. Second, chemical similarity values are not effective at ranking on and off targets and do not consider the structure-activity relationship (SAR) of congeneric series. Most importantly, simple ligand-based searches cannot be applied to analyzing large numbers of ligands such as the unannotated hits from a chemical screen. To circumvent this shortcoming, we recently proposed a new network target inference approach based on chemical similarity networks called chemical similarity network analysis pull-down (CSNAP) [15].

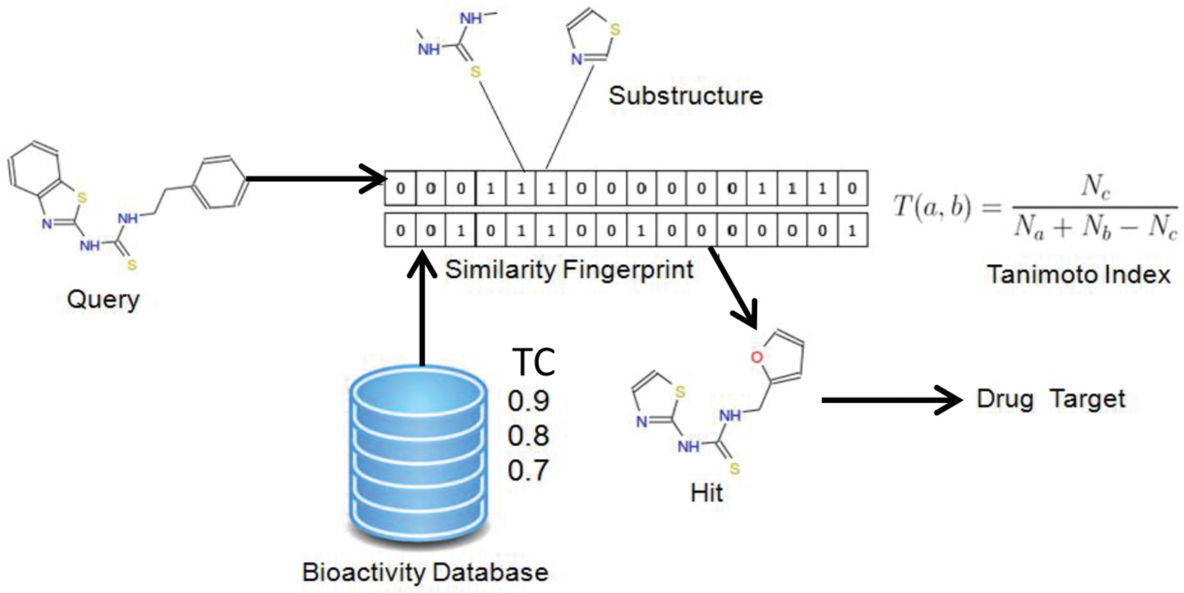


Figure 1. Chemical similarity search using 2D chemical fingerprints.

CSNAP uses a network-based algorithm to predict drug targets and does not rely on absolute chemical similarity values. It utilizes a scoring function (*S*-score) to find the consensus targets of a ligand in its nearest neighbors in a chemical similarity network, which is similar to that used to predict protein functions in a protein-protein interaction (PPI) network [16]. CSNAP is compatible with publicly available bioactivity databases, and we routinely use the ChEMBL database, which is one of the largest bioactivity databases that contains more than 1 million compounds with known target annotations. The original CSNAP algorithm applies 2D Obabel chemical similarity fingerprints (FP2, FP3, FP4, and MACCS) for target predictions. More recently, we developed CSNAP3D that combines 2D and 3D chemical search algorithms to improve the chemical search space [17]. CSNAP3D uses a fast 2D chemical similarity search using either FP2 or FP3 fingerprints to identify sets of hit molecules from large compound databases, and hit molecules are rescored using a combination of 3D similarity descriptors based on a combination of shape and pharmacophore. From our benchmark studies, we found that the CSNAP computational framework was highly accurate and was capable of analyzing large compound sets with diverse chemical structures. Consistently, the CSNAP application has been recently extended for large-scale metabolite analysis [18]. We have made the CSNAP algorithm freely available as the CSNAP web and it can be accessed at http://services.mbi.ucla.edu/CSNAP/.

3. CSNAP implementation

3.1. Chemical similarity network algorithms

Mathematically, a chemical similarity network can be considered as a graph G (V, E) where the vertex V represents compounds and the edge E represents chemical similarity and connects two compounds if they share a chemical similarity above an arbitrary threshold [19]. The CSNAP algorithm is performed in three steps: (1) chemical similarity database search, (2) chemical similarity network construction, and (3) drug target scoring and inference.

3.1.1. Chemical similarity search

Chemical similarity searching is the first step in the CSNAP algorithm (**Figure 2A**). The chemical similarity comparisons are performed using various 2D Obabel fingerprints including FP2, FP3, MACCS, and others. Query compounds prepared in SMILES or SDF formats are used as inputs to the CSNAP program. The compounds are searched sequentially against the ChEMBL database. To identify the ChEMBL compounds most similar to the query, the relative chemical similarity score is quantified by a *Z*-value relative to the distribution of the top 100 chemical similarity values. The ChEMBL compounds with a *Z*-score >2.5 are selected and serve as the annotated compounds for target inference in the next step.

3.1.2. Chemical similarity network construction

To generate chemical similarity networks, pair-wise chemical similarity values are evaluated between every pair of compounds. A network edge is established between two ligands whenever their similarity value is above a predefined threshold (>0.7) (**Figure 2B**). When large compound data sets are analyzed by the CSNAP algorithm, structurally diverse compounds are partitioned into subnetworks of distinct chemical scaffolds, known as “chemotypes.” The chemical similarity networks can be used to estimate the chemical diversity of input structures at this stage.

3.1.3. Drug target scoring and inference

CSNAP infers drug targets using consensus statistics. Specifically, drug targets in the first neighbor of the query are identified and ranked based on their target annotation frequency (**Figure 2C**). A consensus score called an *S*-score is used to rank the probability that the predicted target will interact with the query ligand (**Figure 2D**). There are several advantages of using this network-based scoring function. First, the *S*-score eliminates the possibility of missing target information from the nearest neighbor and considers all possible targets within the same congeneric series. Second, since the drug target is inferred from the target consensus and is in agreement with the observed structure-activity relationship, the robustness of the prediction increases substantially. From our benchmark studies, we showed that this network-based target inference approach improves the prediction success rate over traditional approaches that use simple chemical searches [15].

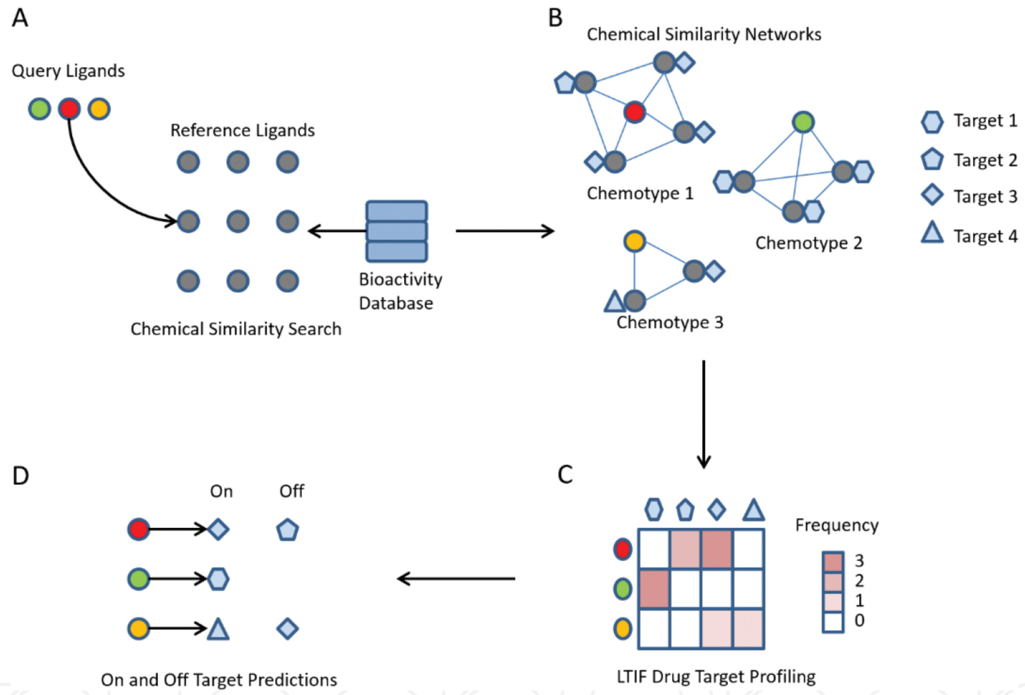


Figure 2. Chemical similarity network analysis pull-down (CSNAP) algorithm for large-scale drug target prediction. (A) Two-dimensional chemical search of three query ligands (green, red, and yellow) identified nine reference compounds from the bioactivity database. (B) Chemical similarity networks clustered compounds into subnetworks corresponding to three major chemotypes. Note that reference compounds interact with four distinct targets. (C) An *S*-score based on consensus statistics is applied to rank the most probable target based on the target annotation frequency of the first-order neighbor targets. (D) On and off targets are differentiated by ranking the predicted *S*-scores.

3.2. Case study: CSNAP web server for automated drug target prediction

To reduce concept to practice, we constructed a CSNAP web server for large-scale target prediction and drug discovery. The CSNAP web includes a front-end graphical user interface (GUI) that provides user interaction and output visualization, while target prediction is performed at the back-end by running the CSNAP algorithm.

3.2.1. CSNAP web input

The CSNAP web server accepts two ligand input formats: SDF and SMILES, which are two of the most commonly used molecular formats that handle large compound databases. In addition, a JME molecular editor is also included, which can be used to convert a chemical structure to a SMILES string on the fly (**Figure 3A**). Several chemical fingerprints are provided to perform chemical comparisons during the search and network clustering steps, including Obabel FP2, FP3, PF4, and MACCS fingerprints (**Figure 3B**). Obabel FP2 fingerprints use a path-based algorithm and are more specific than FP3, FP4, and MACCS that utilize a predefined set of substructures for chemical searches. On the other hand, when structural analogues are not available in the chemical database, FP3 can instead be used to search structurally distinct chemicals with similar bioactivities. To perform chemical searches, the chemical similarity cutoff needs to be defined. Here, CSNAP web supports a combination of absolute cutoff based on Tanimoto coefficient ($Tc > 0.7$) and relative chemical similarity cutoff based on a Z-score. From our experience, the default option using a *Z*-score cutoff of 2.5 will be optimal for most initial CSNAP predictions.

Once the query ligands and chemical search parameters have been defined, the CSNAP algorithm will search the ChEMBL compound activity database to identify structurally similar compounds for target inference (**Figure 3B**). The ChEMBL database assigns targets to a compound based on the level of target specificity (confidence score). Similarly, the compounds are also classified based on the assay type from which they are derived, including biochemical, functional, or ADMET assays. These database parameters will also need to be selected to perform the CSNAP analysis.

3.2.2. CSNAP web output

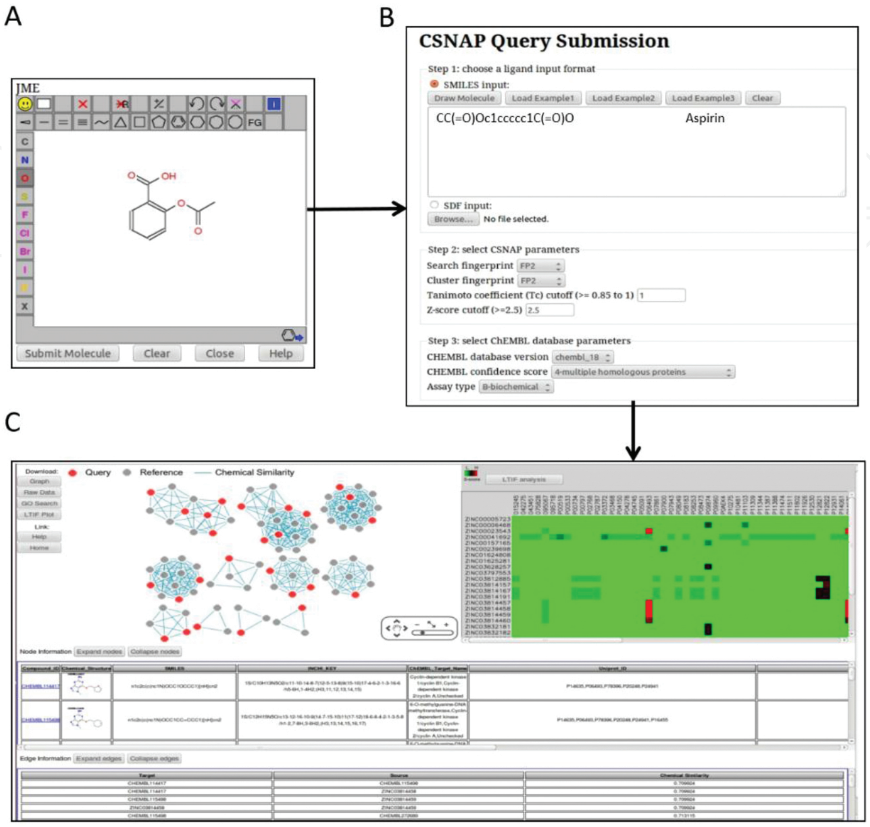
The CSNAP output page consists of three main panels: (1) chemical similarity networks, (2) chemical structure information, and (3) ligand-target interaction fingerprint (LTIF) (**Figure 3C**).

3.2.3. Chemical similarity networks

The chemical similarity networks panel displays the generated chemical networks using the CSNAP algorithm based on input ligands. The chemical similarity network connects query (red) and annotated ligands (gray) from the ChEMBL database, and the targets are inferred using consensus statistics. For large compound sets, the number of generated chemical clusters can be used to estimate the chemical diversity of the sets. To retrieve additional information regarding a specific ligand, the user can click on the node and the relevant information will be displayed in the chemical structure information panel.

3.2.4. Chemical structure information

The chemical structure information panel displays the chemical information selected from the chemical similarity network panel. The panel consists of several columns that include chemical structure information (chemical ID, chemical structure, SMILES string, InChI key) and the predicted target information (target name and UniProt ID). In the ChEMBL prediction column, the predicted targets of each compound are ranked by the *S*-score.



CSNAP Query Submission

Step 1: choose a ligand input format

● SMILES input:

CC(=O)Oc1ccccc1C(=O)O Aspirin

○ SDF input:

No file selected.

Step 2: select CSNAP parameters

Search fingerprint FP2

Cluster fingerprint FP2

Tanimoto coefficient (Tc) cutoff (>= 0.85 to 1) 1

Z-score cutoff (>=2.5) 2.5

Step 3: select ChEMBL database parameters

CHEMBL database version chembl_18

CHEMBL confidence score 4-multiple homologous proteins

Assay type B-biochemical

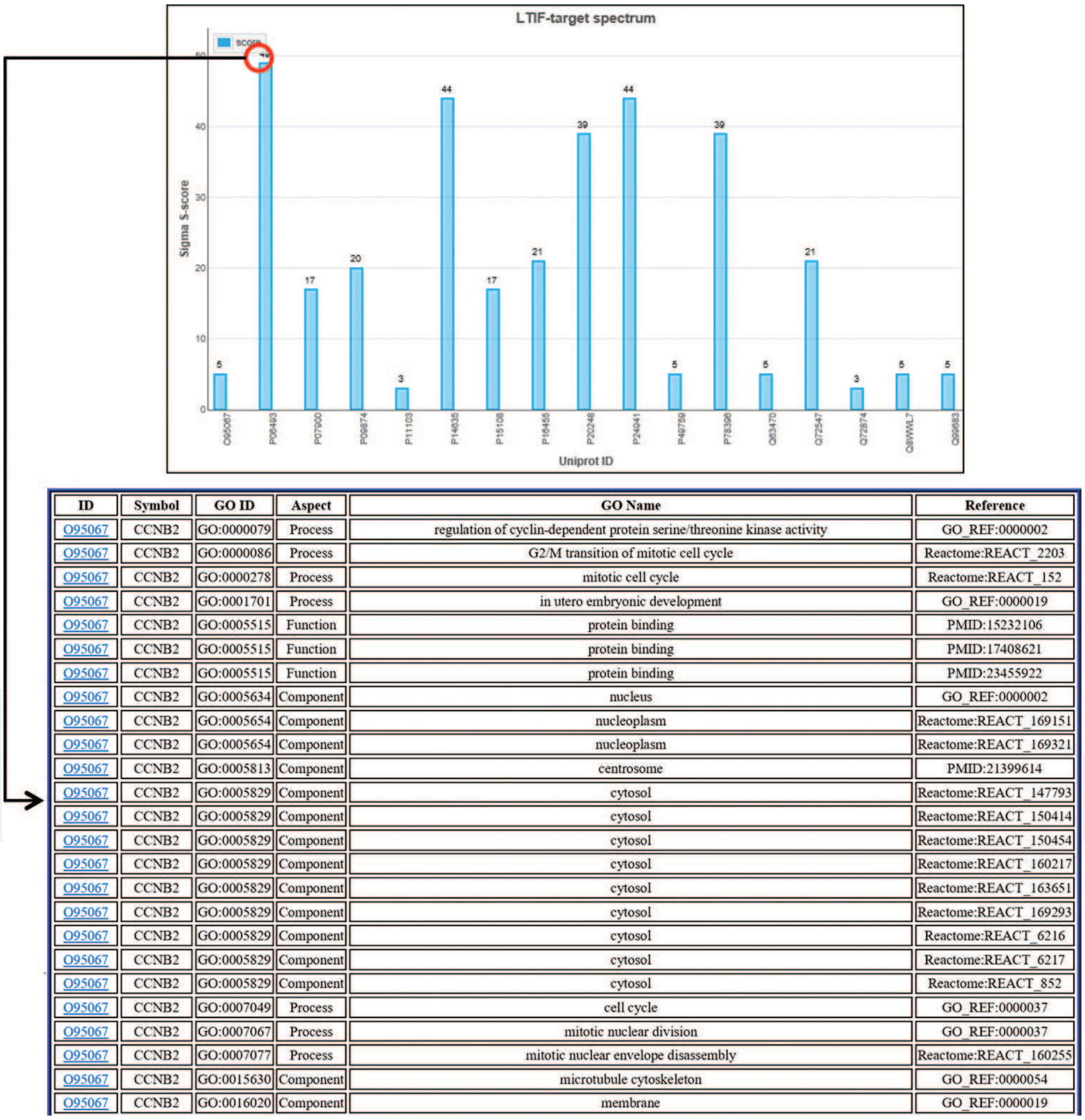
Figure 3. Drug target prediction using the CSNAP web server: (A) construct a query molecule using the JME molecular editor. (B) The molecule is converted to a SMILES string and entered into the CSNAP query submission page. (C) The CSNAP output page consists of three main panels: (1) the chemical similarity network, (2) the chemical structure information, and (3) the ligand-target interaction fingerprint (LTIF).

3.2.5. Ligand-target interaction fingerprint (LTIF)

To analyze the results from large-scale target prediction searches, the ligand-target interaction fingerprint (LTIF) is provided in the CSNAP web output. The LTIF panel displays the predicted *S*-score of each compound mapped against the predicted targets, and the color intensity of the LTIF heatmap is correlated with the *S*-score values. The LTIF can be used to infer compounds sharing similar target binding profiles, which may have similar bioactivity. By clicking on the LTIF button at the top of the LTIF panel, a separate window is created that shows the target spectrum and Gene Ontology (GO) term search derived from the LTIF analysis.

3.2.6. Target spectrum and Gene Ontology (GO) search

To further differentiate primary targets from off-targets in the LTIF, the CSNAP web also computes a target spectrum, by summing the *S*-score ($\sum S$) of all analyzed compounds for each target column (**Figure 4**). For a single compound analysis, the highest peak corresponds to the primary target. Similarly, for multi-ligand analysis, the highest peak corresponds to the most abundant target in the set. To determine the functional role of the predicted targets, the Gene Ontology (GO) search result is also provided (**Figure 4**). GO is a popular bioinformatics tool that maps genes into functions based on controlled vocabulary (GO terms) and has been widely used for pathway analysis of functional genomic data [20]. Here, CSNAP web incorporates a GO search in target predictions as a strategy for posttarget selection and validation. The GO terms can be used to further select relevant targets in a cell-based or phenotype-based screen based on the knowledge of anticipated molecular etiology including cellular components, molecular functions, and biological process. Smaller subsets of targets can then be filtered for additional experimental validation.



ID	Symbol	GO ID	Aspect	GO Name	Reference
O95067	CCNB2	GO:0000079	Process	regulation of cyclin-dependent protein serine/threonine kinase activity	GO_REF:0000002
O95067	CCNB2	GO:0000086	Process	G2/M transition of mitotic cell cycle	Reactome:REACT_2203
O95067	CCNB2	GO:0000278	Process	mitotic cell cycle	Reactome:REACT_152
O95067	CCNB2	GO:0001701	Process	in utero embryonic development	GO_REF:0000019
O95067	CCNB2	GO:0005515	Function	protein binding	PMID:15232106
O95067	CCNB2	GO:0005515	Function	protein binding	PMID:17408621
O95067	CCNB2	GO:0005515	Function	protein binding	PMID:23455922
O95067	CCNB2	GO:0005634	Component	nucleus	GO_REF:0000002
O95067	CCNB2	GO:0005654	Component	nucleoplasm	Reactome:REACT_169151
O95067	CCNB2	GO:0005654	Component	nucleoplasm	Reactome:REACT_169321
O95067	CCNB2	GO:0005813	Component	centrosome	PMID:21399614
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_147793
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_150414
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_150454
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_160217
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_163651
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_169293
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_6216
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_6217
O95067	CCNB2	GO:0005829	Component	cytosol	Reactome:REACT_852
O95067	CCNB2	GO:0007049	Process	cell cycle	GO_REF:0000037
O95067	CCNB2	GO:0007067	Process	mitotic nuclear division	GO_REF:0000037
O95067	CCNB2	GO:0007077	Process	mitotic nuclear envelope disassembly	Reactome:REACT_160255
O95067	CCNB2	GO:0015630	Component	microtubule cytoskeleton	GO_REF:0000054
O95067	CCNB2	GO:0016020	Component	membrane	GO_REF:0000019

Figure 4. Posttarget validation using Gene Ontology (GO) analysis in the CSNAP web. (Top) Target spectrum. (Bottom) GO search results.

4. Application of CSNAP for drug target prediction and discovery

4.1. CSNAP validation

The CSNAP algorithm was validated using 206 known drugs from the directory of useful decoy (DUD) set [21]. The benchmark set included 46 angiotensin-converting enzyme (ACE), 47 cyclin-dependent kinase 2 (CDK2), 23 heat-shock protein 90 (HSP90), 34 HIV reverse-transcriptase (HIVRT), 25 HMG-CoA reductase (HMGA), and 31 poly [ADP-ribose] polymerase (PARP) inhibitors. Using the default search criteria (fingerprint: FP2, Tc = 1, *Z*-score = 2.5), we evaluated the ability of the CSNAP algorithm to accurately predict the designated targets of each compound based on the *S*-score rankings. CSNAP analysis of the 206 compounds showed that the chemical similarity network clustered the drugs into distinct subnetworks, corresponding to diverse chemical scaffolds (chemotypes) (**Figure 5**). For a given subnetwork, the *S*-score was further used to predict the drug target of each compound based on their network connectivity with the reference ligands. The prediction results were then compared with those obtained by the similarity ensemble approach (SEA) [22]. The CSNAP algorithm gave an overall 80–94% true-positive prediction rate (TPR) in comparison with SEA (63–75%) based on the top 1, top 5, and top 10 ranking of target predictions. In particular, CSNAP substantially improved the target prediction rate for promiscuous ligands such as CDK2 and ACE inhibitors (92 and 96%) compared to the SEA approach (30 and 65%) (**Figure 6**).

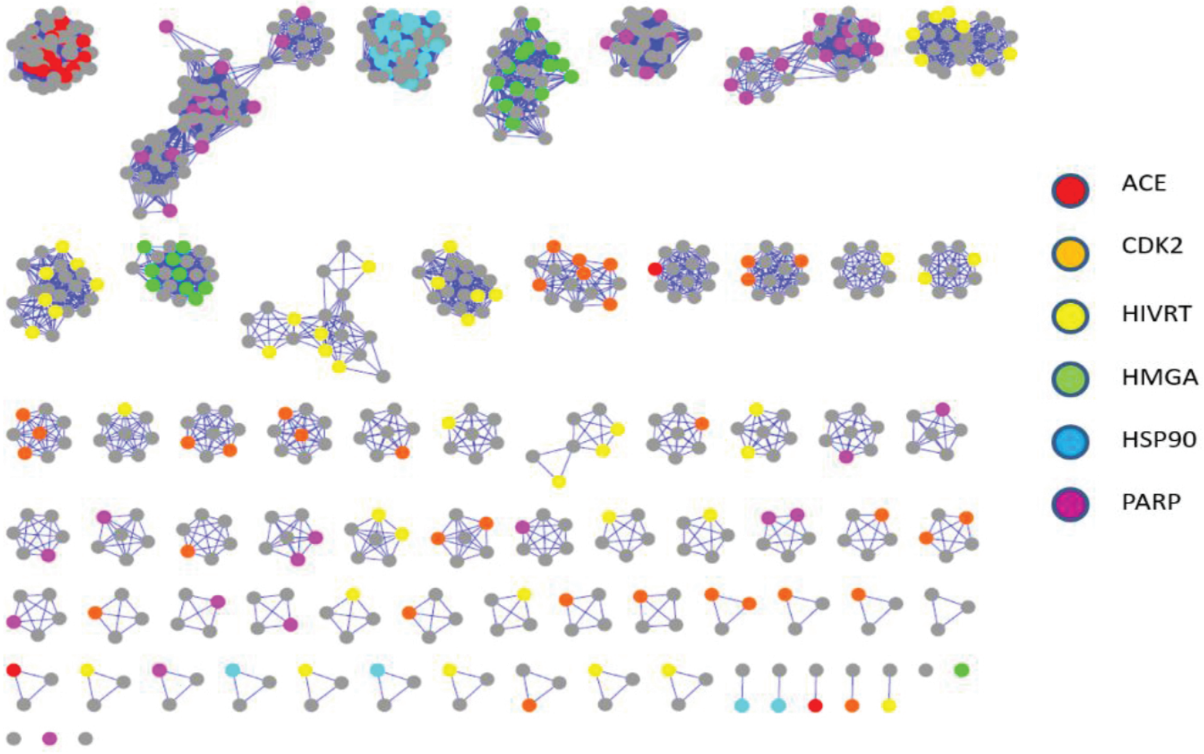


Figure 5. CSNAP2D clustering of 206 benchmark compounds consisting of six known drug classes from the directory of useful decoy (DUD) set.

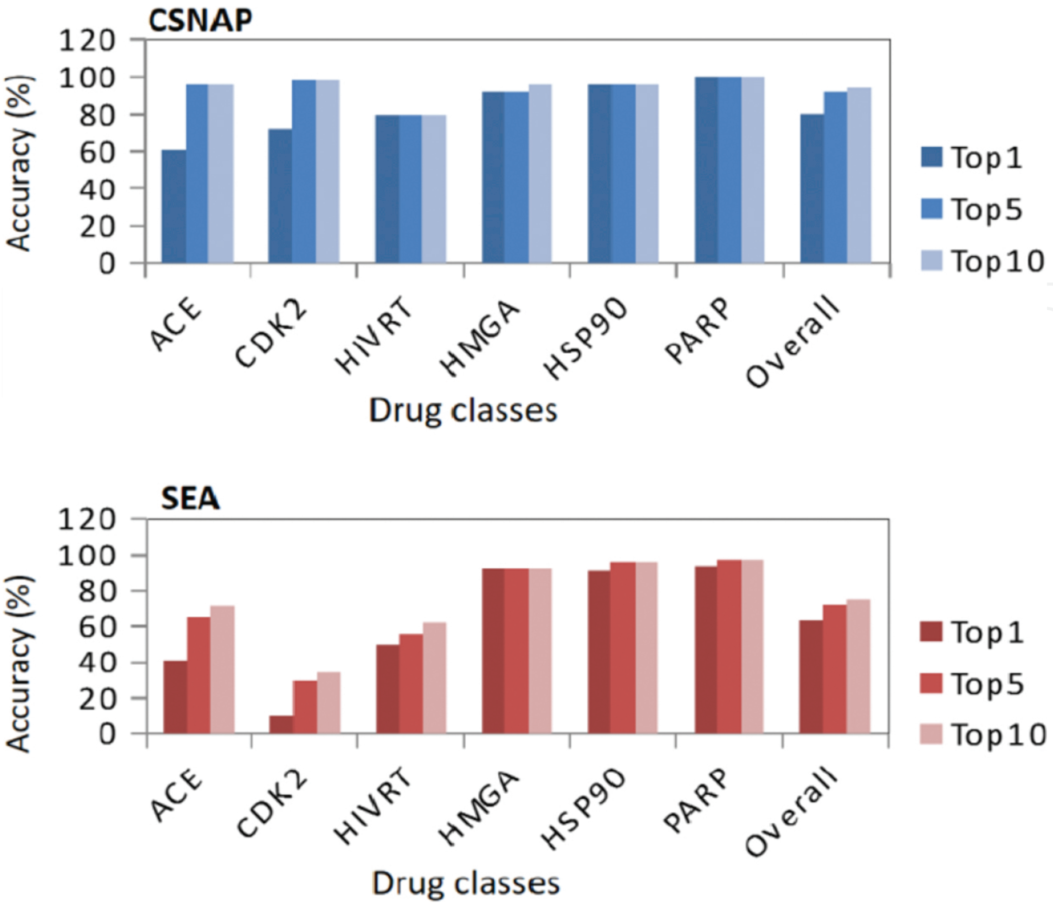


Figure 6. Performance comparison between CSNAP and SEA. CSNAP achieved improved prediction accuracy (TPR) for promiscuous drug classes like ACE, CDK2, and HIVRT inhibitors.

4.2. Target prediction of hits from an antimitotic chemical screen

We applied the CSNAP algorithm to predict the drug targets of a set of 212 compounds that were inhibitors of cell division [23]. CSNAP clustering of the mitotic compounds resulted into 85 chemical similarity subnetworks (**Figure 7A**). To identify the most common targets within the set, we applied the LTIF analysis. The target spectrum derived from the LTIF revealed four broad classes of mitotic targets including fatty acid desaturases (SCD, SCD1, and FADS2), ABL1 kinase, non-receptor-type tyrosine phosphatases (PTPN7, PTPN12, PTPN22, PTPRC, and ACP1), and beta tubulins. In particular, the target spectrum showed that beta tubulin had the largest peak height and was the most prominent protein target for the mitotic compounds. Further analysis showed that 51 compounds were associated with tubulin-targeting chemotypes. The predicted drug targets were validated by comparing siRNA-treated and drug-treated mitotic phenotypes in cell culture using immunofluorescence microscopy. In addition, *in vitro* tubulin polymerization assays were used to determine the effects of these compounds on microtubule formation. Among the 51 tested compounds, 31 compounds showed a perturbation of microtubule polymerization >25%, and thus, the CSNAP algorithm achieved a prediction accuracy of >70%.

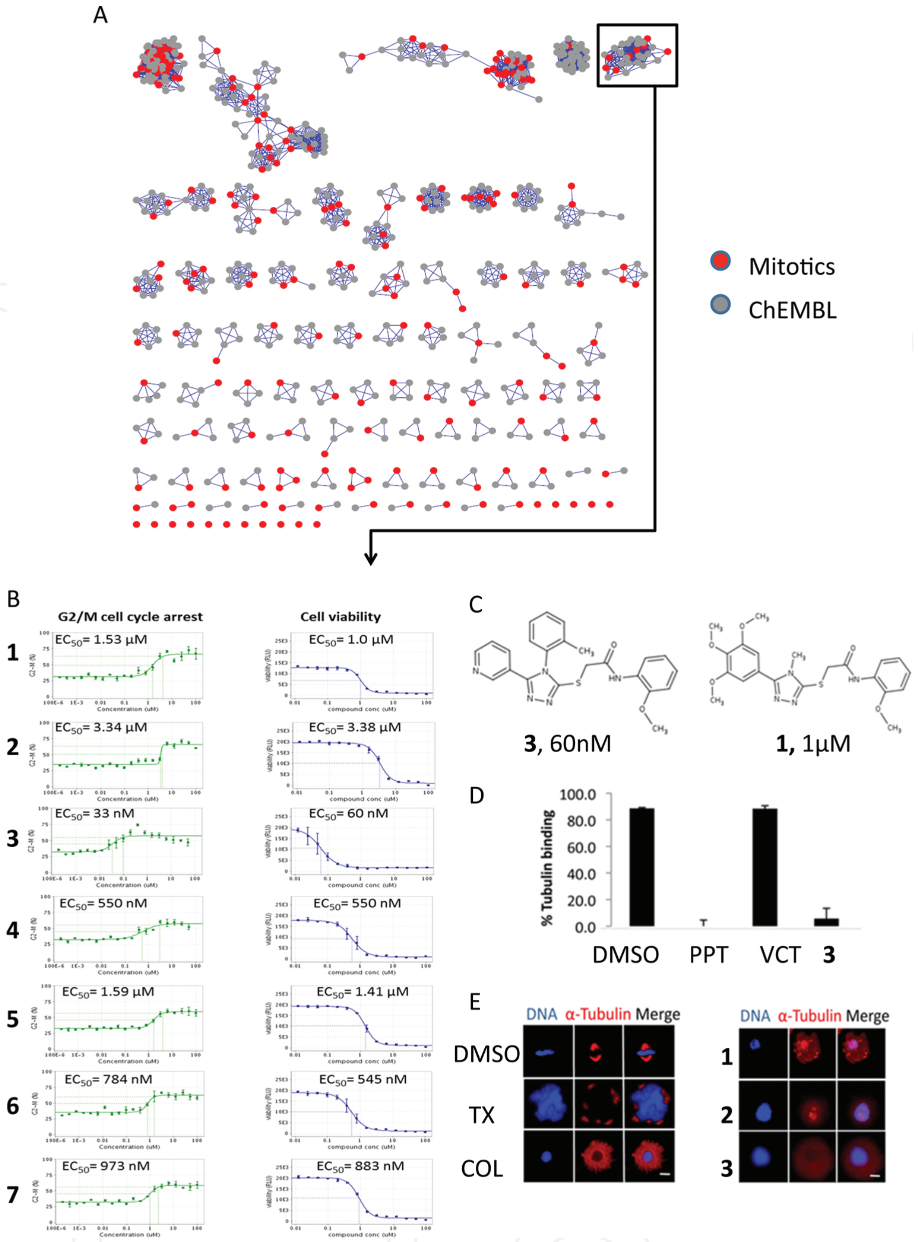


Figure 7. Discovery of novel tubulin-targeting drugs (compounds **1–7**). (A) CSNAP analysis of 212 mitotic drugs identified seven novel tubulin destabilizing compounds. (B) The compounds induced a G2/M cell cycle arrest and decreased cell viability in cell-based assays. (C) Discovery of compound **3** as the most potent compound in the series. (D) A mass spectrometry-based competition assay was used to show that compound **3** and podophyllotoxin (PPT) competed for binding to tubulin in contrast to vincristine (VCT). (E) Compound-treated (**1–3**) and colchicine-treated (COL) cells displayed a tubulin destabilizing effect in comparison with the tubulin stabilizing effect of taxol (TX).

4.3. Discovery of novel tubulin-targeting antimitotics

Using a negative selection strategy, we identified seven novel tubulin-targeting agents that were active in our tubulin polymerization assay but had not been associated with known tubulin chemotypes (**Figure 7A**). The seven compounds were analogues of phenyl-sulfanyl-thiazol-acetamide scaffolds that exhibit various degrees of tubulin destabilizing effects through a mechanism similar to that of the tubulin destabilizing agent colchicine. The most potent compound, compound **3**, in the series exhibited a cytotoxic effect in the nano-molar range (EC_{50}: G2/M = 33 nM; EC_{50}: cell death = 60 nM) when evaluated in cell viability and G2/M arrest assays (**Figure 7B, C**) [15]. The predicted mechanism was validated using a mass spectrometry-based competition assay where both the selected analogues and podophyllotoxin, a known colchicine site binder, competed for binding to tubulin in contrast to the negative control vincristine that interacted with a distant site in beta tubulin (**Figure 7D**) [24]. Likewise, both compound-treated and colchicine-treated cells displayed a tubulin destabilizing phenotype, characterized by rapid shortening of microtubule length and the disappearance of microtubule polymer mass (**Figure 7E**).

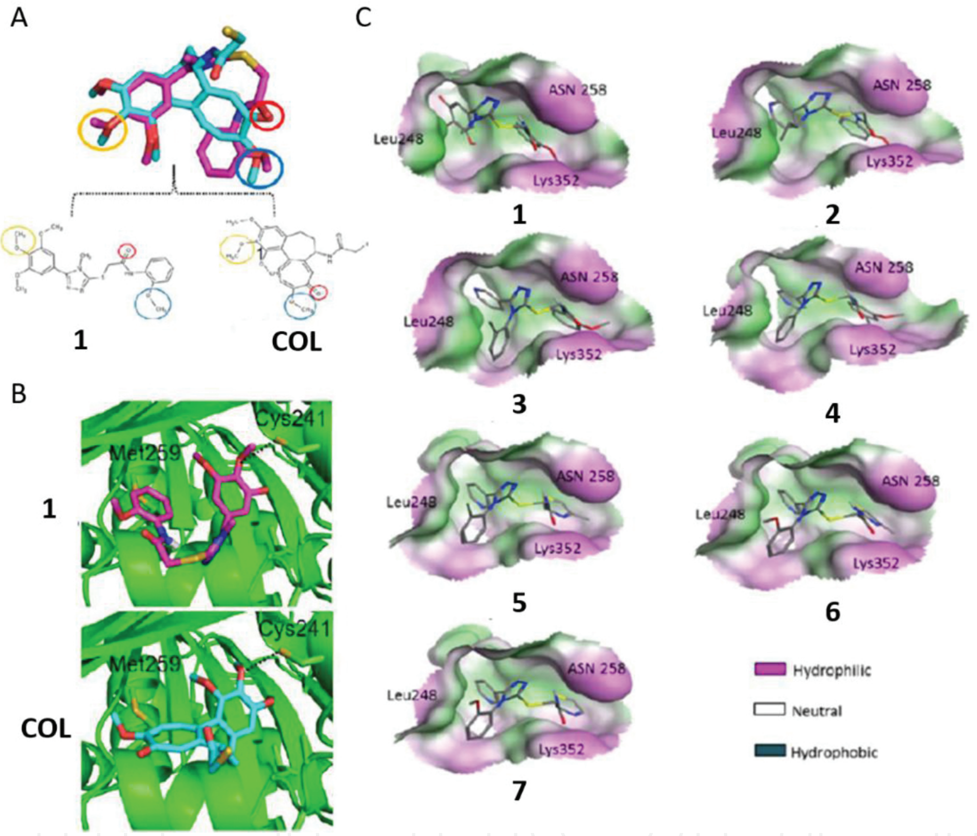


Figure 8. Binding mechanism of a novel tubulin destabilizing chemotype (compounds **1–7**). (A) Pharmacophore alignment between compound **1** and colchicine (COL) showed a consensus pharmacophore. (B) Docking of compound **1** in the colchicine site using the tubulin crystal structure (PDB code: 1SA0) revealed colchicine-like interactions with critical residues (Met259, Cys241). (C) Docking of the seven analogues into the colchicine site showed similar interactions.

4.4. Characterization of novel tubulin-targeting antimitotics

To investigate how the novel antimitotics interacted with beta tubulin, we performed structural alignments between compound **6** and colchicine and identified a consensus pharmacophore between the two molecules (**Figure 8A**). Further docking of compound **6** into the colchicine binding site also showed that both compound **6** and colchicine interacted with common residues, including the 2 and 10 methoxy groups and 9-keto group that interacted with Met259 and Cys241 of beta tubulin, respectively (**Figure 8B**). Similarly, all seven analogues docked into the same site through similar interactions. Interestingly, the elucidated binding modes could be used to explain the observed SAR. For example, the increased potency of compound **7** and **8** in comparison with **6** could be attributed to the hydrophilic interactions between the *N*-propyl and *N*-phenyl groups with Leu248 and Lys352 within the subpockets of the colchicine binding site (**Figure 8C**).

5. CSNAP3D: a 3D upgrade to the CSNAP approach

Chemical similarity searches based on 2D chemical structures have several limitations. First, compounds with distinct scaffolds can exhibit similar bioactivity due to “scaffold hopping” by interacting with a common receptor [25, 26]. Second, although 2D fingerprints based on substructure or fragment searches have the potential to detect scaffold hopping, the scaffold enrichment rate is low. Furthermore, 2D searches do not capture essential features of protein-ligand interactions in three-dimensional space. Consequently, 3D chemical searches based on the three-dimensional information of the ligands will offer additional opportunities to discover novel compounds.

5.1. 3D chemical similarity search

The most common approach to compare ligand similarity in 3D is by shape superposition, which maximizes the Gaussian volume overlap between two ligands [27]. Alternatively, ligand alignments that use molecular interaction field (MIF) or pharmacophore have also been proposed [28, 29]. These approaches take into account the shared chemical features arranged in three-dimensional space. To identify the optimal 3D chemical descriptors, we performed an unbiased screen of diverse 3D chemical descriptors based on molecular shapes or pharmacophores. Using 206 benchmark compounds from the DUD set, we tested the ability of each 3D descriptor to enrich class-specific scaffolds ranked by respective similarity scores. The lowest energy conformer of each ligand was generated using the MOE program. The results showed that 3D chemical descriptors using a combination of shape and pharmacophore features achieved the highest enrichment rate and ligand alignment accuracy compared to those based on shape or pharmacophore alone. This observation agrees with our current understanding that shape complementary and chemical matching are essential for the protein-ligand binding process.

We subsequently developed a 3D chemical similarity search method called “ShapeAlign” that utilized two open-source softwares: “Shape-it” and “Align-it” [30]. Similar to the combo score implemented in the ROCS program, the ShapeAlign algorithm also used a combination of shape and pharmacophore for 3D chemical searches. However, ShapeAlign incorporated a 2D fingerprint similarity score as an integral part of the searching process. Given a ligand with a pre-generated 3D conformation, the ShapeAlign algorithm first detects ligands from the chemical database with the highest shape matching evaluated by a shape Tanimoto index. The hit molecules are then aligned and rescored according to the degree of pharmacophore matching using the Align-it program.

5.2. Drug target prediction using CSNAP3D

We incorporated the "ShapeAlign" algorithm into our CSNAP program called "CSNAP3D" to cluster chemical structures and predict drug targets based on 3D ligand similarity. To evaluate CSNAP3D performance, we assessed the average true-positive rate (TPR) and false-positive rate (FPR) of predicting drug targets for the 206 benchmark compounds. The result showed that CSNAP3D achieved a TPR of >95% at 0.85 Tanimoto cutoff in comparison with other 2D target prediction approaches including CSNAP2D, SEA, and PASS approaches [17]. A comparison of CSNAP3D and CSNAP2D generated networks showed that diverse 2D scaffold subnetworks were clustered into smaller subsets of 3D chemical networks, suggesting that CSNAP3D could be used to identify scaffold hopping ligands not identifiable by conventional 2D methods (**Figure 9**).

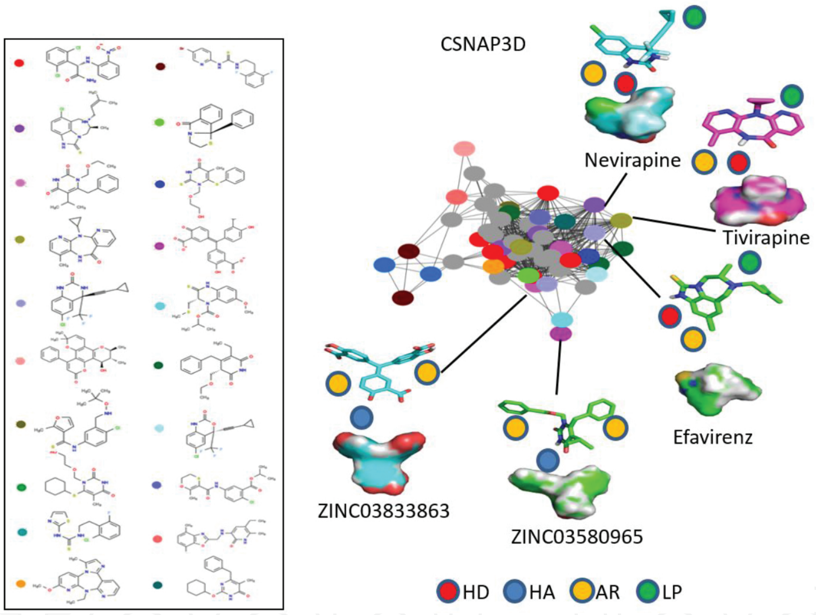


Figure 9. CSNAP3D clustered 34 distinct HIVRT NNRTI chemotypes into a shape-based chemical similarity network. The figure shows that many NNRTIs are scaffold hopping ligands to a common nucleotidyltransferase binding site. The 3D alignment between ligands was based on molecular shape and pharmacophore points (HD: hydrogen donor, HA: hydrogen acceptor, AR: aromatic, LP: lipophilic).

5.2.1. Target prediction of HIVRT inhibitors

As further validation, we presented a case study of predicting targets for a set of HIVRT inhibitors using the CSNAP3D algorithm. HIVRT inhibitors can be classified as nucleoside-based analogues (NRTIs) or non-nucleoside-based analogues (NNRTIs) [31]. In particular, NNRTIs have been difficult drug classes for computational dug target prediction due to the chemical diversity of the drug classes where many compounds are scaffold hopping ligands that bind to a common nucleotidyltransferase binding site. Although 3D ligand-based target predictions that use either the alignment or nonalignment methods have been attempted, many of these approaches yielded low predictability. Here, we applied CSNAP3D to predict the drug targets of 34 structurally diverse HIVRT NNRTIs and compared the prediction results with the CSNAP2D approach (**Figure 9**). Initial 2D chemical similarity network analysis clustered the 34 NNRTIs into 20 structurally diverse chemical similarity scaffolds. Further LTIF analysis, by mapping target prediction *S*-scores to the heatmap, showed that more than 20 compounds did not have a prediction. The NNRTIs were similarly analyzed by the CSNAP3D program using the ShapeAlign algorithm. In contrast, all the 34 ligands were clustered by CSNAP3D into a single shape-based chemical similarity network, suggesting that many NNRTIs are scaffold hopping ligands to a common binding site. LTIF analysis showed that 33 of NNRTIs were correctly predicted, thus achieving a TPR of >97%. In particular, 3D chemical similarity networks correctly identify three FDA-approved NNRTIs, namely efavirenz, nevirapine, and tivirapine whose structure alignment agreed with previous crystal structures and SAR studies (**Figure 9**). In addition, several novel scaffold hopping pairs were also identified (**Figure 9**).

5.2.2. Discovery of novel taxol scaffold hopping ligands

Taxol (paclitaxel) is a well-known anticancer natural product derived from the Yew tree, whose antiproliferative effect was first discovered in 1960s from an NCI anticancer drug screening campaign [32]. Taxol has since been found to be effective for treating a wide range of cancers including ovarian, breast, lung, bladder, prostate, melanoma, esophageal, and other solid tumors. However, the efficacy of taxol has been limited by severe side effects, toxicity, and synthetic feasibility. Thus, identification of low-weight taxol mimetics with more tolerable drug profiles is critical. While several taxol mimetics have been discovered including Synstab B and GS-164, both discovered by chemical screening, their binding mechanisms have remained undetermined [33, 34].

Here, we sought a rational approach to discover taxol mimetics using the CSNAP3D algorithm based on our existing structural knowledge of the original taxol molecule. CSNAP3D analysis of the 212 mitotic compounds from a chemical screen identified 42 potential taxol mimetics linked to 30 taxol structural conformers. Seven predicted taxol mimetics were found to be true positives with a >25% fold change in optical density when tested in tubulin polymerization assays *in vitro* and four compounds shared a consensus chemotype by co-localizing within one chemical similarity subnetwork. The structural alignment of the four selected molecules with taxol showed that they shared a similar T-shape conformation despite a simpler scaffold (**Figure 10A**). Docking studies showed that the increase in microtubule polymerization activity could be attributed to the phenyl moiety of these ligands, which was capable of forming a pi–pi stacking interaction with the critical residue His229 within the taxane site (**Figure 10B**). Three of the compounds demonstrated cytotoxic and antimitotic effects in cell culture with a potency <5 μM. Similarly, all the compounds displayed a similar tubulin stabilizing phenotype, characterized by microtubule aster formation in immunofluorescence microscopy studies (**Figure 10C**).

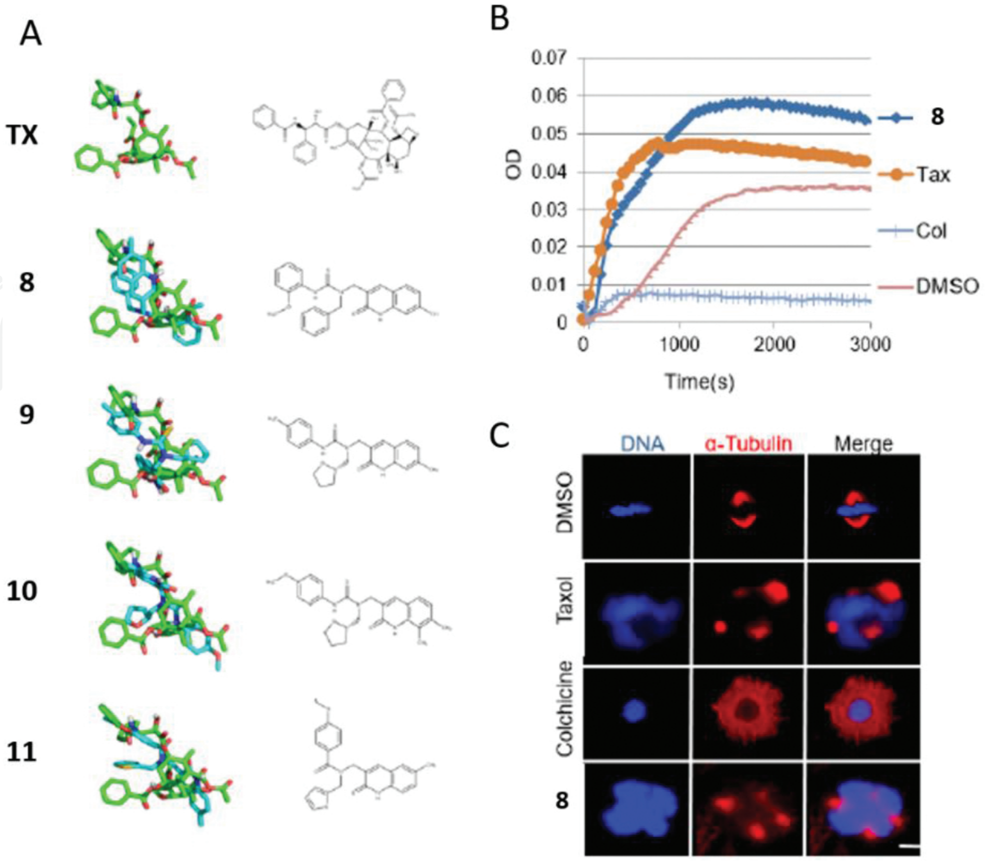


Figure 10. Structure-based discovery of taxol mimetics. (A) CSNAP3D analysis of 212 mitotic compounds from a cell-based screen identified four low molecular weight taxol (TX) mimetic analogues. (B) Compound **8 (8)** demonstrated a fast tubulin polymerization rate at 50 μM similar to taxol (Tax) at 5 μM in comparison with colchicine (Col). (C) Compound **8** displayed a tubulin stabilizing phenotype, characterized by microtubule aster formation in immunofluorescence microscopy studies.

6. Conclusions and future directions

Chemical similarity is an important concept in medicinal chemistry and drug discovery to identify similar compounds with improved bioactivities. Here, we have expanded on this concept to chemical similarity network theory, where descriptive network statistics and graph topology can be applied to large-scale analysis of chemical diversity, bioactivities, and target identification. To demonstrate the utility of this approach, we have implemented the CSNAP algorithm, which can be used for large-scale compound analysis and target predictions. Analogous to protein function prediction in PPI networks, we applied consensus statistics to identify the common targets of each query ligand. We showed that this scoring function outperforms several target prediction methods based on simple chemical similarity searches. To address the challenge of scaffold hopping, where structurally diverse ligands can potentially interact with a common receptor, we developed the CSNAP3D algorithm as a CSNAP extension. CSNAP3D searches chemical structure using the "ShapeAlign" protocol, which utilizes a combination of shape and pharmacophore descriptors. We found that CSNAP3D improves target prediction, particularly for challenging drug classes such as HIVRT NNRTIs that showed high structural diversity and are scaffold hopping ligands. Finally, we successfully applied CSNAP3D to rationally discover low molecular weight taxol mimetics, which exhibit a taxol-like anticancer mechanism and potentially possess improved transport and pharmacokinetic properties than its natural counterpart.

The current CSNAP framework can be extended in several directions. For instance, consensus scoring can be expanded by considering higher-order neighbors, which has been demonstrated to improve prediction accuracy in PPI networks. Similarly, graph theoretical approaches based on maximum network flow and other global optimization approaches can be applied for target assignments [35]. To improve posttarget validation, high throughput functional genomics data can be incorporated to aid in the identification of critical targets relevant to a disease pathway. One example is multiplayer network approaches that integrate drug, target, and annotation interaction networks to enhance target predictions and validations [36]. While CSNAP3D has substantially improved the predictability of CSNAP2D, the algorithm is limited to receptors with bound ligands and the ligand alignment is based on the lowest energy conformer. This shortcoming can be circumvented by considering multi-conformer networks that correlate ligand conformation with target specificities. Likewise, pseudo-ligands generated as the mirror image of an orphan receptor can be considered for receptor deorphanization.

In conclusion, chemical similarity networks are an emerging field in ligand-based drug discovery where the collective properties of a ligand can be easily dissected using descriptive network statistics and graph topology. Here, we presented a new network-based approach for drug discovery and target identification called chemical similarity network analysis pull-down (CSNAP) and a new CSNAP framework called CSNAP3D. The CSNAP computational framework represents a new concept in computational drug discovery with practical application in target identification and drug discovery. We anticipate that the CSNAP approach will stimulate further work in systems and network-based drug discovery that will aid in the discovery of novel drugs for the treatment of cancer and other important diseases.

Author details

Yu-Chen Lo[1] and Jorge Z. Torres[2*]

*Address all correspondence to: torres@chem.ucla.edu

1 Department of Bioengineering, Stanford University, Stanford, CA, USA

2 Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA

References

[1] Yan X, Liao C, Liu Z, Hagler AT, Gu Q, Xu J. Chemical structure similarity search for ligand-based virtual screening: methods and computational resources. Current Drug Targets. 2015, Vol. 16, 1p.

[2] Willett P. Chemoinformatics—similarity and diversity in chemical libraries. Current Opinion in Biotechnology. 2000;11(1):85–8.

[3] Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. Journal of Medicinal Chemistry. 2014;57(8):3186–204.

[4] Faulon J-L, Bender A. Handbook of chemoinformatics algorithms. Boca Raton, FL: Chapman & Hall/CRC; 2010. xii, 440 p.

[5] Gasteiger J. Handbook of chemoinformatics: from data to knowledge. Weinheim: Wiley-VCH; 2003.

[6] Lee JK. Statistical bioinformatics: a guide for life and biomedical science researchers. Hoboken, NJ: Wiley-Blackwell; 2010. xiv, 350 p., 20 p. of plates p.

[7] Baldi P, Benz RW. BLASTing small molecules--statistics and extreme statistics of chemical similarity scores. Bioinformatics. 2008;24(13):i357–65.

[8] Consonni V, Todeschini R, Pavan M, Gramatica P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. Journal of Chemical Information and Computer Science. 2002;42(3):693–705.

[9] Devinyak O, Havrylyuk D, Lesyk R. 3D-MoRSE descriptors explained. Journal of Molecular Graphics and Modelling. 2014;54:194–203.

[10] Hopkins AL. Network pharmacology. Nature Biotechnology. 2007;25(10):1110–1.

[11] Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012;486(7403):361–7.

[12] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2012;40(Database issue):D1100–7.

[13] Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug Discovery Today. 2010;15(23–24):1052–7.

[14] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research. 2006;34(Database issue):D668–72.

[15] Lo YC, Senese S, Li CM, Hu Q, Huang Y, Damoiseaux R, et al. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Computational Biology. 2015;11(3):e1004153.

[16] Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nature Biotechnology. 2000;18(12):1257–61.

[17] Lo YC, Senese S, Damoiseaux R, Torres JZ. 3D chemical similarity networks for structure-based target prediction and Scaffold hopping. ACS Chemical Biology. 2016;11(8):2244–53.

[18] Aretz I, Meierhofer D. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. International Journal of Molecular Sciences. 2016;17:632.

[19] Kolaczyk ED. Statistical analysis of network data: methods and models. New York; London: Springer; 2009. xii, 386 p.

[20] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000;25(1):25–9.

[21] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. Journal of Medicinal Chemistry. 2012;55(14):6582–94.

[22] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nature Biotechnology. 2007;25(2):197–206.

[23] Senese S, Lo YC, Huang D, Zangle TA, Gholkar AA, Robert L, et al. Chemical dissection of the cell cycle: probes for cell biology and anti-cancer drug development. Cell Death & Disease. 2014;5:e1462.

[24] Li CM, Lu Y, Ahn S, Narayanan R, Miller DD, Dalton JT. Competitive mass spectrometry binding assay for characterization of three binding sites of tubulin. Journal of Mass Spectrometry. 2010;45(10):1160–6.

[25] Schneider G, Neidhart W, Giller T, Schmid G. "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. Angewandte Chemie, International Edition in English. 1999;38(19):2894–6.

[26] Sun H, Tawa G, Wallqvist A. Classification of scaffold-hopping approaches. Drug Discovery Today. 2012;17(7–8):310–24.

[27] Yan X, Li J, Liu Z, Zheng M, Ge H, Xu J. Enhancing molecular shape comparison by weighted Gaussian functions. Journal of Chemical Information and Modeling. 2013;53(8):1967–78.

[28] Cruciani G. Molecular interaction fields: applications in drug discovery and ADME prediction. Weinheim: Wiley-VCH; 2006. xviii, 307 p.

[29] Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, et al. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Research. 2010;38(Web Server issue):W609–14.

[30] Taminau J, Thijs G, De Winter H. Pharao: pharmacophore alignment and optimization. Journal of Molecular Graphics and Modelling. 2008;27(2):161–9.

[31] Zhan P, Chen X, Li D, Fang Z, De Clercq E, Liu X. HIV-1 NNRTIs: structural diversity, pharmacophore similarity, and implications for drug design. Medicinal Research Reviews. 2013;33(Suppl 1):E1–72.

[32] Renneberg R. Biotech History: Yew trees, paclitaxel synthesis and fungi. Biotechnology Journal. 2007;2(10):1207–9.

[33] Haggarty SJ, Mayer TU, Miyamoto DT, Fathi R, King RW, Mitchison TJ, et al. Dissecting cellular processes using small molecules: identification of colchicine-like, taxol-like and other small molecules that perturb mitosis. Chemistry & Biology. 2000;7(4):275–86.

[34] Shintani Y, Tanaka T, Nozaki Y. GS-164, a small synthetic compound, stimulates tubulin polymerization by a similar mechanism to that of taxol. Cancer Chemotherapy and Pharmacology. 1997;40(6):513–20.

[35] Jungnickel D. Graphs, networks, and algorithms. Fourth edition. ed. Heidelberg; New York: Springer; 2013. xx, 675 p.

[36] Berenstein AJ, Magarinos MP, Chernomoretz A, Aguero F. A multilayer network approach for guiding drug repositioning in neglected diseases. PLoS Neglected Tropical Diseases. 2016;10(1):e0004300.