

*Research Article*

## Virtual target screening to rapidly identify potential protein targets of natural products in drug discovery

Yuri Pevzner <sup>1</sup>, Daniel N. Santiago <sup>1</sup>, Jacqueline L. von Salm <sup>1,5</sup>, Rainer S. Metcalf <sup>1</sup>, Kenyon G. Daniel <sup>3,4</sup>, Laurent Calcul <sup>5</sup>, H. Lee Woodcock <sup>1,2</sup>, Bill J. Baker <sup>1,5</sup>, Wayne C. Guida <sup>1,2,3,\*</sup>, Wesley H. Brooks <sup>1,\*</sup>

<sup>1</sup> Department of Chemistry, University of South Florida, Tampa, FL 33620, USA

<sup>2</sup> Center for Molecular Diversity in Drug Design, Discovery and Delivery, University of South Florida, Tampa, FL 33620, USA

<sup>3</sup> Virtual Screening and Molecular Modeling Core, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA

<sup>4</sup> Department of Cell Biology, Microbiology, and Molecular Biology, University of South Florida, Tampa, FL 33620, USA

<sup>5</sup> The Center for Drug Discovery and Innovation (CDDI), University of South Florida, Tampa, FL 33620, USA

\* **Correspondence:** E-mail: wesleybrooks@usf.edu; wayne.guida@moffitt.org

**Abstract:** Inherent biological viability and diversity of natural products make them a potentially rich source for new therapeutics. However, identification of bioactive compounds with desired therapeutic effects and identification of their protein targets is a laborious, expensive process. Extracts from organism samples may show desired activity in phenotypic assays but specific bioactive compounds must be isolated through further separation methods and protein targets must be identified by more specific phenotypic and *in vitro* experimental assays. Still, questions remain as to whether all relevant protein targets for a compound have been identified. The desire is to understand breadth of purposing for the compound to maximize its use and intellectual property, and to avoid further development of compounds with insurmountable adverse effects. Previously we developed a Virtual Target Screening system that computationally screens one or more compounds against a collection of virtual protein structures. By scoring each compound-protein interaction, we can compare against averaged scores of synthetic drug-like compounds to determine if a particular protein would be a potential target of a compound of interest. Here we provide examples of natural products screened through our system as we assess advantages and shortcomings of our current system in regards to natural product drug discovery.

**Keywords:** natural products; virtual target screening; drug discovery

---

**Abbreviation list:**

ADME = absorption, distribution, metabolism, excretion;

CDDI = Center for Drug Discovery and Innovation;

HTS = High-throughput screening;

MMOA = Molecular mechanism of action;

MOI = Molecule of interest;

NCI = National Cancer Institute;

NME = New molecular entity;

VS = Virtual screening;

VTS = Virtual Target Screening.

---

## 1. Introduction

Natural products are receiving renewed interest as potential sources for therapeutic agents since the number of new drugs discovered has dropped in recent years in spite of the growth in available synthetic compound libraries for drug discovery campaigns. Advanced drug discovery techniques developed in recent years to support high-throughput screening (HTS) that primarily used synthetic compounds can benefit new drug discovery efforts that focus on natural products. We are particularly interested in applying new computational methods we are developing towards drug discovery in natural products.

At the University of South Florida (USF) in Tampa, there are numerous drug discovery projects underway, targeting primarily cancers and infectious diseases. The Center for Drug Discovery and Innovation (CDDI) provides shared resources to USF researchers [1]. Among these resources the CDDI provides natural product libraries of extracts, sub-fractions and individual compounds for screening in drug discovery campaigns. The CDDI is continually adding to and replenishing its collection of natural products, including marine natural products (extracts, sub-fractions and compounds) from new sources retrieved in annual expeditions in Antarctic waters [2]. As compounds are identified from these sources, we want to quickly identify potential targets for the compounds. This includes primary therapeutic targets but also potential alternate targets for repurposing, or unintended targets with potential detrimental interactions that could lead to adverse reactions.

Previously we developed a Virtual Target Screening (VTS) system at H. Lee Moffitt Cancer Center & Research Institute on the USF campus. VTS can assess a molecule of interest (MOI) for potential interactions with other proteins by screening the MOI against a large collection of virtual protein structures [3]. VTS was originally developed towards cancer-related drug discovery and contains primarily human proteins. Also, VTS was developed, calibrated and used primarily with synthetic small molecules. Our current VTS system has been of use in cancer-related drug discovery projects in which VTS has brought to the attention of investigators potential off-target interactions of their lead MOIs. Here we describe testing of natural product libraries from the NCI and CDDI through VTS to assess benefits and shortcomings of our current VTS system with regards to natural products and with regards to drug discovery targeting infectious diseases (i.e. including more viral and bacterial proteins). The aim of this exercise is to improve and expand our VTS system to make it an effective tool with both natural products and synthetic compounds in drug discovery efforts for all disease types.

## 2. Natural products versus synthetic compounds in drug discovery

Humankind has always explored natural products to meet needs such as medicines, health supplements, dyes, inks, perfumes, pest repellents and many other uses. In regards to medicines, this exploration and utilization has been rewarding but, in many cases, the specific active compound, its unique characteristics (ex. chirality) and/or synthetic approaches were not known and, therefore, lacked optimization for the most effective sourcing, purification, formulation and dosing. Rational design of analogs to improve a compound was often not possible. The modern advances in synthetic chemistry and analysis allowed for more control in compound identification, design and development and, using high throughput synthesis and combinatorial chemistry, the number of available synthetic compounds for drug discovery has grown rapidly. For example, the ZINC database currently has virtual structures for over 35 million commercially available compounds ready for virtual screening and purchase of samples for experimental validation of virtual “hits” [4]. A “hit” is a compound that shows the desired effect based on computational scoring of its interactions with the virtual protein target, but the “hit” still needs experimental confirmation of the effect. The abundance of synthetic compounds has facilitated development of virtual and experimental high-throughput screening (HTS) in drug discovery; however, the yield of new drugs, based on new molecular entities (NMEs) discovered among synthetic compounds, has declined. Though the synthetic compounds should provide a diverse sampling of chemical space, synthetic collections lack relevant biodiversity and bioactivity due to such problems as having insufficient chiral representation, or lacking the size and complexity of the bioactive natural products with which they must compete *in vivo*. Some of this results from a lack of proper synthetic approaches and understanding to recreate a diverse representation of bioactive motifs, such as difficulties in covalent incorporation of bromide or chloride into organic compounds [5]. Synthetic compounds have been developed for the most part from conceptions and abilities of researchers to meet perceived biological needs whereas natural products have arisen under direct influence of highly demanding bioenvironmental pressures giving natural products, in general, an edge in filling new uses in real biological functions, i.e. biological validation.

Another potential difficulty with synthetic compound “hits” is that they are often selected based on target-specific screening but fail to advance due to poor phenotypic effects in more complex experimental assays (e.g. cell-based assays) in which there can be unanticipated targets with different molecular mechanisms of action (MMOA) for the compound. A synthetic compound may have poor ADME (absorption, distribution, metabolism, excretion) characteristics that are not appreciated until later in drug development. Promising natural product compounds are typically found through a phenotypic approach in which sample extracts are tested in phenotypic assays and then the specific compounds in their bioactive state in the extract are determined through further separation and analysis (e.g. NMR or LC-MS). Dereplication is also performed to ascertain if it is, in fact, a new unexplored compound. Phenotypic screening, as typically used with natural products, has proven to be more productive than the target-based screening predominantly done with synthetic compound libraries. A recent review of new drugs between 1999 and 2008 found that there were 259 agents (183 small-molecule drugs, 20 imaging agents, and 56 therapeutic biologics) approved by the FDA, of which 75 were first-in-class drugs and 50 of these are small molecules [6]. Based on the authors’ analysis of published information accompanying the small molecule drugs, 28 were found through phenotypic screening, 17 were found through target-specific screening and there was insufficient

information for the other 5. The authors concluded that, phenotypic screening had been more successful in finding NMEs even though target-based screening had been used much more often in drug discovery during that period. Due to their biological origin and validation, natural product libraries should typically be enriched with phenotypically active compounds relative to most synthetic compound libraries.

Lower than expected yield of new drugs from target-based approaches and synthetic compound libraries has brought natural products back into fashion for drug discovery campaigns [7–13]. There have been improvements in applicable techniques, such as miniaturized high-throughput phenotypic assays to rapidly assess extracts and compounds [6]. Advances in extraction automation and fractionation techniques allow improvements in speed and quantity for testing extracts and compounds at sufficiently high concentrations to better determine phenotypic effect [14]. Individual compound structures can now be determined from only a few micro-grams of a compound obtained by analytical HPLC purification of samples and run through “nanoscale structure elucidation” using highly sensitive mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy [15]. The elucidated structures can then be screened rapidly against chemoinformatics databases in a dereplication step to determine uniqueness of a compound.

Natural product collections can have overall advantages in bioavailability, biodiversity, biological validation and enrichment towards phenotypic activities. Natural products address biologically relevant chemical space whereas synthetic compound libraries are not as well focused on biological relevance. In addition, due to the variety of biosynthetic pathways, some of which are yet to be discovered, working with natural product materials, such as extracts, could lead to discovery of enzymatic activities that yield some of the more complex natural products that synthetic chemists cannot yet replicate. These discoveries can be elucidated to identify genes and proteins responsible for the activities, and then they can be exploited to ramp up production of natural products that were previously difficult to obtain, thereby improving on sourcing problems. These new discoveries can also be used in rational design starting with natural products or synthetic compounds to create compounds not seen previously (i.e., NMEs) [16,17].

### 3. Virtual screening

Virtual screening (VS) has proven to be a useful, cost-saving computational means of rapidly identifying compounds that have potential desired interactions, such as inhibition, with a protein target of interest. In typical VS, virtual compound structures from a library are individually docked and scored in a binding site of a virtual protein structure, such as the active site of an enzyme or an allosteric site. The compounds are then ranked by their scores and the top of the ordered list is considered to be enriched in compounds with potential for the desired interactions. Of course there are many caveats to successful VS, such as: having a properly prepared virtual protein structure; having docking/scoring algorithms that can rapidly and accurately generate and interpret docked poses of the compound with the protein; and having a means of experimentally validating the virtual “hits” to confirm that they do, in fact, show interaction with the protein of interest. Another caveat is to use libraries of compounds that have drug-like characteristics (ex. low molecular weight; solubility; potential for some ionic and/or hydrogen bonding). “Lipinski’s rule of 5” (a.k.a. “ro5”, 500 Daltons or smaller; cLogP or calculated log P of 5 or less for solubility; 5 or less potential hydrogen bonds) is often used to describe these characteristics and is applied with some leniency in

building virtual compound libraries [18]. The concept of drug-likeness and its application in drug discovery has been evolving, as discussed in a recent review [19]. Since some of the most successful drugs currently on the market do not strictly comply with ro5, improvements have been developed for assessment of a compound, such as a quantitative estimate of drug-likeness (QED) approach which can assess a compound's potential even if it does not strictly meet one or two of the ro5 guidelines, but balances deviations in any one category against the compound's overall distribution of properties [20].

There are many published successes with VS in which lead compounds were first identified by VS runs followed by experimental confirmation and then proceeded further into development towards becoming drug candidates, as described in a recent review [21]. One example is discovery of an inhibitor selective for CK2 kinase using a homology model of human CK2a and screening over 400,000 virtual compounds from libraries at Novartis Pharma [22]. The CK2 kinase inhibitor found by Vangrevelinghe et al., (5-oxo-5,6-dihydroindolo[1,2-a]quinazolin-7-yl) acetic acid, was discovered using the DOCK 4.0 virtual screening application [23]. The inhibitor was very potent with an  $IC_{50}$  of 80 nM. As an example of our work with VS, we used Schrodinger's Glide 2.7 (Grid-based Ligand Docking with Energetics) [24] to virtually screen the NCI Diversity Set I, which contains 1,990 compounds [25], against S-adenosylmethionine decarboxylase (AdoMetDC), a key enzyme in polyamine synthesis and an important target in cancers [26]. The NCI Diversity Sets (currently NCI Diversity Set IV) are often used in initial screens in oncological drug discovery campaigns to test new targets since powder samples are available for free along with published data for each compound from the NCI for experimental confirmation. For the virtual protein structure, we used a crystal structure of human AdoMetDC solved at 2.24 Å resolution by Tolbert et al. available from the Protein Data Bank as entry PDBID 1I7M [27,28]. In our VS against AdoMetDC we identified one compound, NSC 354961, a 9-amino-acridine compound, which ranked #14 among the NCI Diversity Set compounds but NSC 354961 had the best experimental confirmation with an  $IC_{50}$  of 12  $\mu$ M. It served as a starting structure for further development.

VS can be used in natural product drug discovery but it runs into problems when investigators are trying to find disruptors of protein-protein interactions because the binding areas are broader, protein dynamics on a larger scale become more important and larger more complex compounds are deemed necessary similar to the multiple amino acid residues from one protein that interact with the protein partner. Natural products collections can contain larger molecules but larger molecules can entail more rotatable bonds leading to more difficulty in sampling the range of poses. Gerwick and Moore, in their recent review of natural product drug discovery, call for further development and application of chemoinformatics and *in silico* screening (i.e. VS) to drug discovery with natural products [5]. They believe it presents exciting possibilities for the field.

#### 4. Virtual target screening

Virtual Target Screening (VTS) uses the NCI Diversity Set I as a reference for scoring MOI-protein interactions. Whereas VS docks multiple compounds against a single protein target to identify compound "hits", the basic function of VTS is to dock a single MOI against multiple proteins to identify protein "hits". A protein "hit" in VTS is a protein that has potentially important interactions with the MOI. VTS can bring to attention those proteins that may interact with the MOI so that further research may determine its selectivity. VTS can potentially identify adverse

interactions, allosteric interactions and possible targets for repurposing. VTS can also be used to test a focused library built around an MOI so that analogs of the MOI can be compared for their improvements in avoiding detrimental interactions and perhaps being more specific towards the intended targets of the original MOI.

Although systems such as this are in their infancy, they can greatly reduce the time and expense of experimental analysis, allowing drug discovery efforts to focus on the most promising compounds and avoid compounds that may be too promiscuous or have adverse interactions that cannot be removed readily by further rational design work.

We have published our work on VTS previously, discussing its development and validation against kinase targets [3]. Other groups have developed systems to screen individual compounds across protein collections and those have been discussed in recent reviews [29–32]. These systems have a variety of names, such as reverse pharmacognosy, target fishing, inverse docking, virtual counter screening and drug repositioning and they have differing approaches to determining potential MOI-protein interactions, such as comparing the MOI to a known inhibitor of a protein or comparing the MOI to spatial requirements of a protein's active site. Difficulties arise when there is no known inhibitor for comparison or the only known inhibitor is particularly weak or is a covalent inhibitor. Spatial definitions (i.e. “shapes”) of binding sites can have deficiencies if they are based on a static protein and limited to a few known small molecule interactions.

In our VTS system, we used a calibration step for each protein to define standards an MOI should meet in order for that MOI-protein to be interpreted as a VTS “hit”. The NCI Diversity Set I is docked against the protein and statistics for each protein structure is calculated from the Glide Scores (GScores). The averages for the top 20, top 200 and top 50% of NCI compound docking scores are used as benchmarks. Since the NCI Diversity Set I has almost 2,000 compounds, we consider these averages to approximate the top 0.5%, 5%, and 25% marks, respectively. The use of these averages reduces the impact of skewing since extremely strong interactions by one or a few NCI compounds with a protein might set the bar too high for an MOI to register an impact as a true inhibitor even though the MOI may have relevant interactions with the protein that should be noted for their potential as adverse interactions.

Adapting VTS for routine use in the CDDI with natural products and assuring sufficient coverage of protein targets relevant in infectious diseases requires critical analysis of current shortcomings as well as demonstration to new users of the potential of VTS in infectious disease drug discovery. This is also an opportunity to assess the needs for improvement in VTS's overall performance, capacity and quality. The purpose of this exercise in screening NCI and CDDI natural products is to demonstrate the usefulness of VTS in identifying potential off-target interactions with human proteins and to determine those issues that must be addressed in our current VTS system to optimize it for natural product drug discovery.

## 5. Materials and methods

### 5.1. Virtual compound libraries

The NCI Diversity Set I, consisting of 1,990 virtual compounds, was obtained from the National Cancer Institute's (NCI) Developmental Therapeutics Program (DTP) and used as described previously for our VTS system [3,25].

The NCI natural Product Set II includes 120 compounds [33]. Due to cost and time considerations, we selected 67 compounds that ranged from 314 to 692 Daltons in a loose application of “Lipinski’s rule of 5”. These compounds were screened in VTS. The NCI DTP website also provides links to published experimental data on the compounds in their virtual libraries [34]. This allows for comparison and validation of VTS “hits” against known experimentally-determined interactions of the compounds when the data is available.

The CDDI library provided to us contained virtual structures for 160 natural product compounds, many originating from marine sources [1]. These are from the larger collections at the CDDI, which include over 2,500 extracts and 950 characterized bioactive compounds. Again, due to cost and time consideration, we selected 87 of these compounds for VTS testing with a range of 350–675 Daltons.

### 5.2. Ligand preparation

Prior to screening all compounds were prepared using the LigPrep utility in Schrodinger's Maestro software [35,36]. LigPrep generates 3D structures of all tautomers, ionic states and stereoisomers of input compounds. In the case of NCI compounds the structure files have indicated chiralities of some stereo centers. Therefore, when preparing the ligands, only the unspecified centers were allowed to sample different chiralities. For CDDI compounds however, to reduce computational time and the amount of isomers, all the chiralities were inferred from the 3D structure. As a result, LigPrep generated a total of 1133 structures from the 67 original NCI compounds and, because no chiralities were varied in the case of CDDI set, a total of only 103 structures resulted from LigPrep for the 87 original CDDI compounds.

### 5.3. Virtual target screening

Our VTS system was developed in a Linux environment and consists of a script (written in Perl) that repeatedly executes Glide via a command line call. Input files include: 1) the prepared protein structures (prepared using Schrödinger’s Protein Preparation Wizard application [35]); 2) the list of average scores for the proteins determined when they were calibrated by docking the NCI diversity set; and 3) the file containing one or more MOIs. A typical VTS run is described above and in our previous publication [3]. We have developed a user-friendly interface using CHARMMing [37] to invoke VTS windows that allow: 1) uploading or drawing of MOI structures; 2) selection of proteins, either all proteins or subsets selected by criteria, such as species and/or protein type (ex. kinases); 3) initiate and monitor a job with ongoing updates of estimated finish time; and 4) review the output. Currently, the CHARMMing based VTS is only available internally at the University of South Florida and H. Lee Moffitt Cancer Center. Future development of VTS will include a URL available for a broader user group. The VTS protein collection currently is 1,418 structures represented by 1,567 grids (some having multiple sites) for prepared and calibrated proteins. Predominantly it is human proteins (~71%) since VTS was developed initially towards oncological drug discovery projects but the collection also contains some bacterial (~11%), viral (~5%) and fungal (~1%) proteins as well as murine, bovine and other species. The proteins were originally selected from the Protein Data Bank and primarily wild-type entries were selected [28].

We selected primarily human proteins as the targets for this VTS exercise, which yielded 1,059

target grids representing 1,011 unique crystal structures. For this exercise with a large number of MOIs to process in VTS, we estimated there would be over one million docking and scoring calculations. Therefore, the compounds were divided into 25 separate files, each containing subsets of 30–60 structures. This was done to effectively parallelize the VTS calculations. Each of these subsets was then uploaded through the VTS web interface's MOI upload section. A screening job was then launched from the VTS web interface for each of the 25 subsets to be screened against all 1,059 target grids. On average each job took approximately 30 hours to complete. When analyzing the results of the screening jobs, a compound-protein was deemed to be a "hit" if the score outperformed the average score for the top 20 NCI diversity set compounds for that protein structure. The "hits" were compared against the literature for consistency with published experimental data. The PubChem database was used to look-up experimental activities for the compounds that hit one or more targets [38,39].

#### 5.4. Hardware

VTS was originally developed and run on a Dell Precision 490 workstation running Fedora 8 Linux with dual Xeon 3.06 GHz processors, 4 GB RAM and a 250 GB hard drive. This was sufficient when running one or a few MOIs but, for the larger sets of MOIs for the 67 NCI natural products and the 87 CDDI natural products, we incorporated execution on a local cluster we have developed in the Virtual Screening & Molecular Modeling Core at H. Lee Moffitt Cancer Center & Research Institute. This cluster is comprised of 6 racks each containing 8 Intel-based dual core desktop computers connected via a NETGEAR Pro Sage 16 router (model GS716T).

#### 5.5. Graphics & modeling

Visualization of docked MOIs in protein targets was performed with Schrödinger's Maestro, a graphical user interface that allows modeling and facilitates use of many of Schrödinger's applications with the models, such as MacroModel and Glide [24,35,40]. Refined depictions used for publication figures were developed using Schrödinger's PyMol [41].

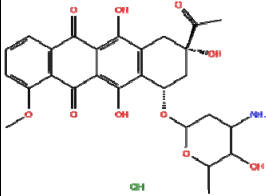
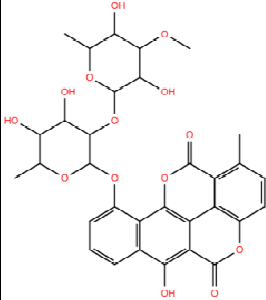
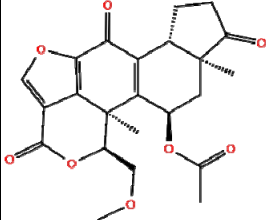
## 6. Results

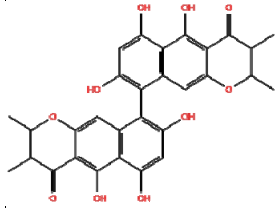
### 6.1. VTS on NCI natural products

Upon compilation of the results of all the jobs, a total of 13,278 hits were identified for the NCI Natural Products Set II compounds tested. For 4 of the screened NCI molecules VTS identified a total of 16 targets against which screened compounds were found active in one or more experimental assays (Table 1). The most "active" MOI Daunorubicin (Figure 1) hit a total of 6 unique proteins. The most interesting of those hits can be considered the Lck tyrosine kinase protein. Daunorubicin has hit 5 different structures that represent this protein. Moreover in at least two reported assays Daunorubicin had indeed shown activity against this particular target. It may be speculated that a compound hitting multiple structures of the same protein may indicate likely activity against that protein. In addition, results of the NCI natural products screen show preference of some compounds towards one or more



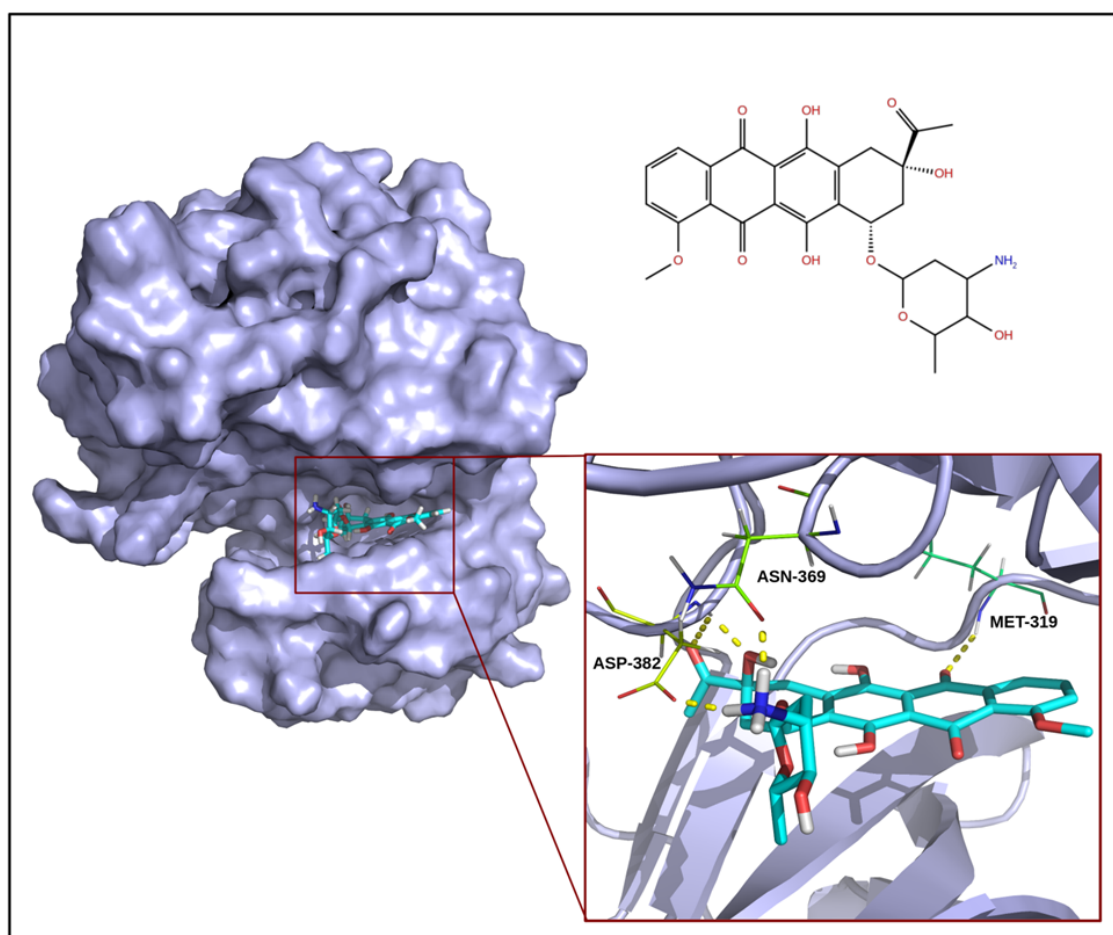
**Table 1. Top NCI natural products screen hits. Compounds with known activity against the targets that have been identified as hits in the VTS screen. Column 1 contains the NSC number of the compound, column 2 contains the common name of the compound, column 3 contains 2D sketch of the compound, column 4 contains the name of the protein targeted by the compound, column 5 shows the number of unique structures representing the target identified as a hit for the given MOI, column 6 is the number of assays that have reported activity of the MOI against this target.**

NSC	Compound Name	Structure	Target	# of unique structures identified by VTS	# of assays reporting activity
82151	Daunorubicin		Lck tyrosine kinase	5	2
			Estrogen Receptor Alpha	2	8
			Epidermal Growth Factor Receptor	3	1
			Thyroid hormone receptor	1	4
			Mitogen Activated Protein Kinase 14	1	1
5159	Chartreusin		Fyn tyrosine kinase	1	1
			Urokinase-type plasminogen activator*	2	1
			Heat Shock Protein HSP 90-alpha**	1	1
221019	Wortmannin		Cyclin-Dependent Kinase 5	1	1
			Cyclin-Dependent Kinase 2	6	1
			Serine/threonine kinase PIM1	2	1
			Src tyrosine kinase	2	1
			Lck tyrosine kinase	1	2

345647	Chaetochromin		Cyclin-Dependent Kinase 5	1	1
			3-phosphoinositide-dependent protein kinase 1	1	1
			Glutathione-S-transferase Omega 1	1	2

\* - VTS screening included and hit the human version of the protein while the experimental data is on mouse.

\*\* - VTS screening included and hit the human version of the protein while the experimental data is on a structurally similar Plasmodium falciparum.



**Figure 1. Daunorubicin docked into Lck. Daunorubicin's docked pose according to the VTS docking procedure. Inset shows the local hydrogen bond interactions in yellow dashed lines as well as the residues with which these interactions are formed in stick representation. The Lck structure was 1QPJ.pdb [52].**

protein families as can be seen in the case of Wortmannin and a kinase family (Table 1). It should be noted however that kinases are disproportionately well represented (25%) in our protein database due to their importance in cancer. Still an insight like this can be used to gauge the potential for more specific targeting as well as possible promiscuity of an MOI.

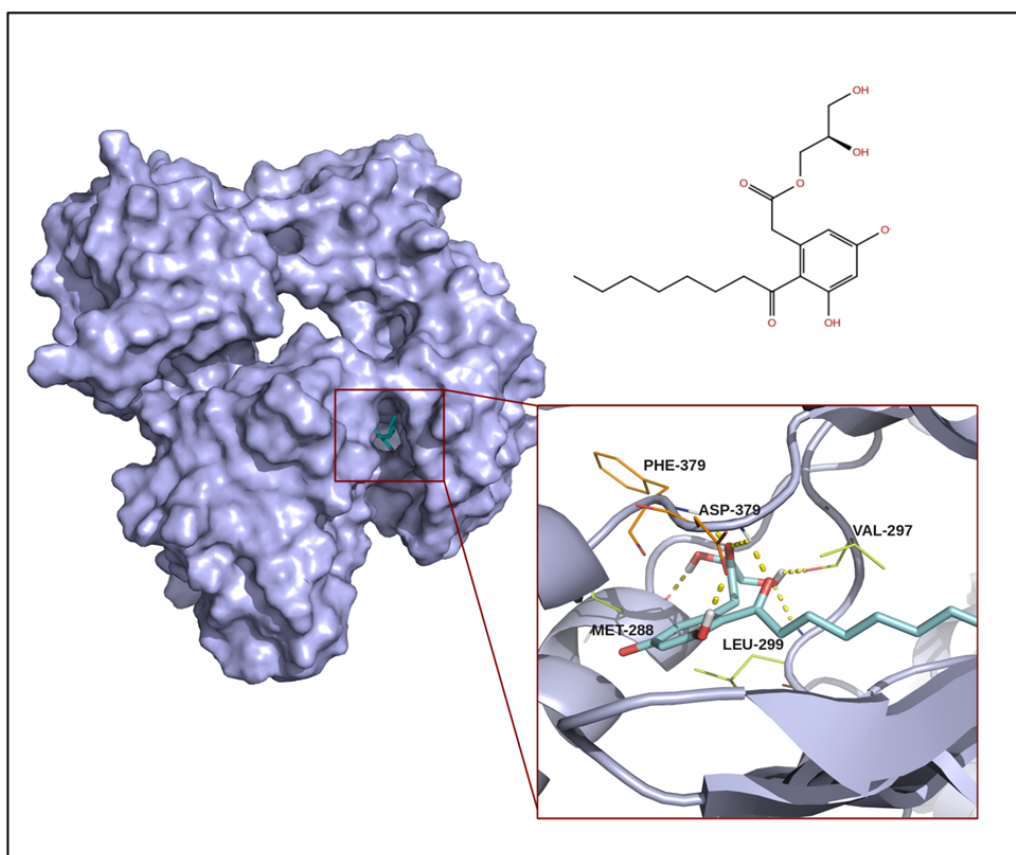
## 6.2. VTS on CDDI natural products

The screening of CDDI compounds against our database has also yielded interesting results. Although assay information is not as readily available for these compounds, the results of the VTS can still show interesting trends and insights valuable for drug discovery. Table 2 shows the top 40 hits where the MOI has outperformed the top performers of the calibration set by the largest margin,

**Table 2. Top CDDI set hits. Top 40 hits based on the largest margin by which an MOI outperformed the top 20 calibration set average. First column contains the name of the target protein with co-crystallized ligand identifier in square brackets. Second column indicates the percentage by which a given MOI outsourced the average docking score of the top 20 calibration set molecules for that target. Third column contains the rounded molecular weight of a docked MOI and serves as its identifier. The table is sorted by the molecular weight.**

Target	Outperformed calibration by (%)	Molecular Weight (compound ID)
[R11]Coagulation factor x	28	287
[L1G]HCK	28	368
[NHB]Histone deacetylase 8	30	372
[NHB]Histone deacetylase 8	24	372
[NHB]Histone deacetylase 8	36	374
[AO5]Methionine aminopeptidase 2	35	374
[GIP]Protein (lactoylglutathione lyase)	35	374
[NHB]Histone deacetylase 8	29	374
[GIP]Protein (lactoylglutathione lyase)	30	374
[AO2]Methionine aminopeptidase 2	23	374
[NHB]Histone deacetylase 8	26	437
[NHB]Histone deacetylase 8	23	530
[HYF]Orphan nuclear receptor pxx	35	587
[HYF]Orphan nuclear receptor pxx	33	587
[QPP]CAMK1D	32	587
[SWF]CYP2C9	26	587
[DEO]Macrophage metalloelastase	20	587
[XLD]Coagulation factor x heavy chain	30	587
[RAP]Fkbp25	36	587
[SDK]Cathepsin K	30	587
[POS]Cathepsin K	30	587
[INA]Cathepsin K	29	587
[POS]Cathepsin K	28	587

[BOG]p38-alpha(MAPK14)	25	587
[2CA]Cathepsin k	27	587
[FMM]Epidermal growth factor receptor	23	625
[L1G]HCK	39	629
[471]Peroxisome proliferator activated receptor	42	629
[U66]Protein farnesyltransferase alpha subunit	31	629
[CIU]Epoxide hydrolase 2- cytoplasmic	29	629
[POS]Cathepsin K	33	629
[SDK]Cathepsin K	31	629
[HYF]Orphan nuclear receptor pxx	25	629
[155]Urokinase-type plasminogen activator	27	629
[SAH]Histamine Methyltransferase (Ile105Var)	22	629
[AIJ]Estrogen receptor	21	629
[BNE]Protein farnesyltransferase alpha subunit	21	629
[BOG]p38-alpha(MAPK14)	23	629
[LA1]Integrin alpha-L	31	663
[LA1]Integrin alpha-L	19	674



**Figure 2.** 368 docked into HCK. 368's docked pose according to the VTS docking procedure. Inset shows the local hydrogen bond interactions in yellow dashed lines as well as the residues with which these interactions are formed in stick representation. The HCK structure was 2COI.pdb [53].

thus making these compounds more likely to be tighter binders to the indicated targets. It can be seen that there are a number of molecules, as identified by their unique molecular weight that hit several targets, as is especially the case with compounds 374, 587 and 629. Once again, this may be an early indicator of potentially promiscuous compounds. At the other end of the spectrum are compounds such as 287, 368 (Figure 2), 437, 530 and 674. These compounds only hit one target. Especially interesting are those compounds with lower molecular weight as it is easier for a larger compound to score high based on the increased number of favorable interactions it can form with the protein. If a smaller sized yet still a drug-like compound such as 368 significantly outperforms a calibration set for a single particular target, it may be worthy of a further inquiry as a potential binder to that protein

## 7. Discussion

We have used 67 compounds from the NCI Natural Products Set II to test our VTS system for its ability to virtually identify potential protein targets, i.e. “hits” that can be compared to published data for those compounds. We have also used 87 compounds from the CDDI natural products to assess their potential interactions with protein targets when they were screened through VTS. The work has also brought to light improvements that can be made to the current VTS system to make it more applicable to our new areas of application in drug discovery and development: infectious diseases and natural products, particularly marine natural products. There is an unmet need for large-scale drug discovery efforts with marine natural products [42,43]. There is also opportunity for these large-scale efforts with marine natural products and natural products in general to be used more in academic settings, such as the CDDI [44]. Drug discovery efforts with marine natural products have been successful with 13 products in clinical trials in 2010 and so these efforts should be expanded to take advantage of new natural products from new marine sources [45]. In order to scale-up these efforts and make them more productive, particularly as natural product libraries grow, innovation is needed [46]. We believe that VTS is an innovative, cost-effective approach to address the needs for rapid identification and appraisal of natural products. Besides our previous paper [3] in which we compared our VTS results against experimental data focusing on kinases, there are other recent reports in which other groups have compared results from computational drug repurposing applications against experimental data demonstrating the usefulness of the tools [47,48].

This exercise demonstrates that VTS can identify potential promiscuity of natural product compounds that may deter further development of an MOI if the promiscuity is experimentally confirmed for some of the “hit” proteins. VTS can also give ideas of other targets that may not have been considered previously for the MOI. However, since our VTS protein collection is limited, we cannot yet consider the VTS runs as comprehensive searches. There may be more proteins that would arise as “hits” for these compounds if those proteins had been included. Other systems are now available online for screening compounds against collections of protein structures, such as HitPick [49], ChemMapper [50], and Mantra [51]. A comparison of VTS and other target finding applications was beyond the scope of this study since our purpose was to analyze VTS with regards to natural products and infectious disease targets. However, we did test two of the top scoring compounds (wortmannin and daunorubicin) from the NCI Natural Products II and got similar hits of kinase proteins as were found by HitPick. A thorough study comparing the various target finding applications is certainly warranted now that more applications are published but it should be comprehensive with regards to ligands (small synthetic molecules and natural products). We believe

the VTS approach has benefits in that it can be used towards different sites on a protein, such as allosteric sites, for which there may not be any previously known ligands. We can use Schrödinger's SiteMap as a means of identifying potential binding sites [54,55]. The drug-like molecules of the NCI Diversity set can quickly calibrate new sites even when there is no previous data. Many other applications are dependent on previously known ligands for specific binding sites in order to make their assessment of an MOI's interactions.

VTS also can be used to compare an MOI against known inhibitors of VTS "hit" proteins and to compare the MOI against its own analogs to test focusing by functional groups as rational design is applied to the MOI to generate improved analogs. Some of the analogs may show more specificity to the intended target and generate fewer VTS "hits", suggesting that the analog is an improvement over the original MOI. With regards to a known inhibitor, if the MOI shows more specificity (fewer VTS "hits") than the inhibitor, this can increase interest for the MOI.

The main purpose of VTS is to identify new targets for an MOI, whether they are new targets for repurposing or targets that may indicate possible adverse interactions. For VTS to be effective towards this purpose, we need to increase our collection of proteins. Since VTS was originally developed for supporting oncological drug discovery projects, the emphasis was for inclusion of primarily human proteins. For testing compounds in VTS for infectious disease projects, we need a great increase in protein structures from viruses, bacteria and other microorganism that are relevant to those diseases. We need to grow the VTS protein collection as much as possible using as many structures from the Protein Data Bank, but we still want to focus on those structures that have high resolution ( $\sim 2.0$  Å or less) and primarily wild-type proteins. We realize now that, even towards oncological drug discovery, proteins from microorganisms are important in VTS since often the beneficial microbes in the patient's gut microbiota can be adversely affected by drug therapy. And of course there need to be more proteins represented for the problematic microorganisms in infectious diseases.

We intend also to include more proteins involved in protein-protein interactions. The target sites may be more difficult to isolate for Glide grids and so multiple copies of the protein structure with grids at different sites could be used. We can use Schrödinger's SiteMap utility to identify potential binding sites as part of the protein preparation and center the Glide grids on those sites [54]. Disruptors of protein-protein interactions can conceivably be larger than typical small molecule substrates of enzyme active sites and so we would want to expand VTS to deal with larger MOIs. This would require an additional approach to calibrating the prepared proteins. We can envision using a designed virtual di-, tri- or tetra-peptide or peptidomimetic library for an additional set of averages to use in calibration and VTS analysis. These averages of larger molecules would be of use for larger MOIs, which can occur with natural products. Another possible improvement is to incorporate into VTS an assessment of ligand binding efficiencies. We can improve overall interpretation of VTS dockings by taking into account the number of non-hydrogen atoms to avoid inflation of the scores for larger ligands. The basic approach would be to divide the GScore by the number of non-hydrogen atoms. In addition, approaches being developed to assess drug-likeness need to be incorporated, such as the QED approach [20]. Another issue to address in improving VTS is the quality of the prepared proteins. MacroModel has been used in VS and VTS for achieving a relaxed state, more *in vivo*-like, for the protein compared to the original crystal structure. However, we might think of this as local relaxation of protein domains, whereas a thorough molecular dynamics (MD) relaxation of the protein, in which it is put in a virtual environment that simulates the

individual water molecules and ions, can relax the protein to a global energy minimum in solution that may be even more pertinent to creating the protein in a realistic state for screening. In the last few years MD applications, such as Schrödinger's Desmond, have become available so that we can now implement MD into our protein preparation steps [56]. This implies that we should reprocess our existing protein collection with MD as well as apply MD to new proteins. MD requires more powerfully processing power and memory, which we can now access with clusters making these improvements feasible.

VTS and VS are developed primarily towards screening for non-covalent interactions of ligands and proteins. High-throughput virtual screening of potential covalent inhibitors has not been developed to our knowledge. This would be an interesting addition but does present many difficulties, even to develop a limited application. It would require a list of potential reactant groups, identifying key reactive residues in a protein that could be in play, and an algorithm to compare distances, charges and intermediates. It would seem to preclude the actions of cofactors and other agents. So covalent screening, at least with current tools, would be too challenging. However, it is interesting to contemplate such tools and their potential benefits in drug discovery.

## 8. Conclusion

This work has shown the potential for VTS with natural products to identify targets that may represent new purposes for the MOI, possible promiscuity of an MOI, or possible adverse interactions that warrant further investigation. Our VTS system can serve these purposes but we have used this current work to identify issues to improve in order to maximize the use and benefits of VTS for infectious disease drug discovery projects, particularly with newly acquired marine natural products.

## Conflict of Interest

All authors declare no conflict of interest in this paper.

## References

1. The Florida Center of Excellence in Drug Discovery and Innovation (CDDI) is described at: <http://www.research.usf.edu/cddi/drugdiscovery/resources.asp>.
2. Huang YM, Amsler AO, McClintock JB, et al. (2007) Patterns of gammarid amphipod abundance and species composition associated with dominant subtidal macroalgae along the western Antarctic Peninsula. *Polar Biol* 30: 1417-1430.
3. Santiago DN, Pevzner Y, Durand AA, et al. (2012) Virtual Target Screening: Validation Using Kinase Inhibitors. *J Chem Inf Model* 52: 2192-2203.
4. Irwin JJ, Sterling T, Mysinger MM, et al. (2012) ZINC: A Free Tool to Discover Chemistry for Biology. *J Chem Info Model* 52:1757-1768.
5. Gerwick WH, Moore BS (2012) Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* 19:85-98.
6. Swinney DC, Anthony J (2011) How were new medicines discovered? *Nat Rev Drug Discov* 10: 507-519.

7. Newman DJ, Cragg GM (2012) Natural Products as Sources of New Drugs over the 30 Years from 1981 to 2010. *J Nat Prod* 75: 311-335.
8. Mishra BB, Tiwari VK (2011) Natural products: An evolving role in future drug discovery. *Eur J Med Chem* 46: 4769-4807.
9. Bachmann BO, Van Lanen SG, Baltz RH (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Indust Microbiol Biotech* 41: 175-184.
10. Williams P, Sorribasa A, Howes MR (2011) Natural products as a source of Alzheimer's drug leads. *Nat Prod Rep* 28: 48-77.
11. Kuete V, Alibert-Franco S, Eyong KO, et al. (2011) Antibacterial activity of some natural products against bacteria expressing a multidrug-resistant phenotype. *Int J Antimicrob Agents* 37: 156-161.
12. Cragg GM, Newman DJ (2013) Natural products: A continuing source of novel drug leads. *Biochim Biophys Acta* 1830: 3670-3695.
13. Dayan FE, Owens DK, Duke SO (2012) Rationale for a natural products approach to herbicide discovery. *Pest Mgmt Sci* 68: 519-528.
14. Bugni TS, Harper MK, McCulloch MWB, et al. (2008) Fractionated marine invertebrate extract libraries for drug discovery. *Molecules* 13: 1372-1383.
15. Molinski TF (2010) NMR of natural products at the "nanomole-scale". *Nat Prod Rep* 27: 321-329.
16. Nicolaou KC, Snyder SA (2005) Chasing molecules that were never there: misassigned natural products and the role of chemical synthesis in modern structure elucidation. *Angew. Chem Int Ed Engl* 44: 1012-1044.
17. Penn K, Jenkins C, Nett M, et al. (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *ISME J* 3: 1193-1203.
18. Lipinski CA, Lombardo F, Dominy BW, et al. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46: 3-26.
19. Ursu O, Rayan A, Goldblum A, et al. (2011) Understanding drug-likeness. *WIREs: Comp Mol Sci* 1:760-781.
20. Bickerton GR, Paoliuni GV, Besnard J, et al. (2012) Quantifying the chemical beauty of drugs. *Nat Chem* 4: 90-98.
21. Matter H, Sottriffer C (2011) Applications and success stories in virtual screening. In: *Virtual screening: principles, challenges, and practical guidelines*. Wiley-VCH Verlag GmbH & Co. Ch. 12: 319-358.
22. Vangrevelinghe E, Zimmermann K, Schoepfer J, et al. (2003) Discovery of a Potent and Selective Protein Kinase CK2 Inhibitor by High-Throughput Docking. *J Med Chem* 46: 2656-2662.
23. Ewing TA, Makino S, Skillman AG, et al. (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comp Mol Des* 15: 411-428.
24. Friesner RA, Banks JL, Murphy RB, et al. (2004) Glide: New approach for rapid, accurate docking and scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* 47: 1739-1749.



25. NCI Diversity Set was provided by the Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 6130 Executive Blvd., Room 8020, Rockville, MD 20852. Details of NCI collections in the DTP diversity collections are available at: [http://dtp.nci.nih.gov/branches/dscb/div2\\_explanation.html](http://dtp.nci.nih.gov/branches/dscb/div2_explanation.html).
26. Brooks WH, McCloskey DE, Daniel KE, et al. (2007) In Silico Chemical Library Screening and Experimental Validation of a Novel 9-Aminoacridine Based Lead-Inhibitor of Human S-Adenosylmethionine Decarboxylase. *J Chem Info Model* 47: 1897-1905.
27. Tolbert WD, Ekstrom JL, Mathews II, et al. (2001) The structural basis for substrate specificity and inhibition of human S-adenosylmethionine decarboxylase. *Biochemistry* 40: 9484-9494.
28. Berman HM, Westbrook J, Feng Z, et al. (2000) The Protein Data Bank. *Nucl Acids Res* 28: 235-242.
29. Hui-Fang L, Qing S, Jian Z, et al. (2010) Evaluation of various inverse docking schemes in multiple targets identification. *J Mol Graph Model* 29: 326-330.
30. Grinter SZ, Liang Y, Huang SY, et al. (2011) An inverse docking approach for identifying new potential anti-cancer targets. *J Mol Graph Model* 29: 795-799.
31. Do QT, Lamy C, Renimel I, et al. (2007) Reverse Pharmacognosy: Identifying Biological Properties for Plants by Means of their Molecule Constituents: Application to Meranzin. *Planta Med* 73: 1235-1240.
32. Li YY, An J, Jones SJM (2006) A Large-Scale Computational Approach to Drug Repositioning. *Genome Info* 17: 239-247.
33. NCI Natural Products was provided by the Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 6130 Executive Blvd., Room 8020, Rockville, ME 20852. Details of the DTP natural products repository are available at: <http://dtp.nci.nih.gov/branches/npb/repository.html>.
34. Online data searching is available through the NCI DTP website "data search" link: [http://dtp.nci.nih.gov/docs/dtp\\_search.html](http://dtp.nci.nih.gov/docs/dtp_search.html).
35. Suite 2012: Maestro, version 9.3, Schrödinger, LLC, New York, NY, 2012.
36. Suite 2012: LigPrep, version 2.5, Schrödinger, LLC, New York, NY, 2012.
37. Miller BT, Singh RP, Klauda JB, et al. (2008) CHARMMing: a new, flexible web portal for CHARMM. *J Chem Info Model* 48: 1920-1929.
38. Wang Y, Xiao J, Suzek TO, et al. (2012) PubChem's BioAssay Database. *Nucl Acids Res* 40: D400-12.
39. Bolton E, Wang Y, Thiessen PA, et al. (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 in Annual Reports in Computational Chemistry, Volume 4, ACS, Washington, DC.
40. Suite 2012: MacroModel, version 9.9, Schrödinger, LLC, New York, NY, 2012.
41. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.
42. Imhoff JF, Labes A, Wiese J (2011) Bio-mining the microbial treasures of the ocean: New natural products. *Biotech Adv* 29: 468-482.
43. Liu X, Ashforth E, Ren B, et al. (2010) Bioprospecting microbial natural product libraries from the marine environment for drug discovery. *J Antibiotics* 63: 415-422.

44. Bauer RA, Wurst JM, Tan DS (2010) Expanding the range of “druggable” targets with natural product-based libraries: an academic perspective. *Curr Op Chem Biol* 14: 308-314.
45. Mayer AMS, Glaser KB, Cuevas C, et al. (2010) The odyssey of marine pharmaceuticals: a current pipeline perspective. *Trends in Pharm Sci* 31: 255-265.
46. Bennani YL (2011) Drug discovery in the next decade: innovation needed ASAP. *Drug Disc Today* 16: 779-792
47. Keiser MJ, Setola V, Irwin JJ, et al. (2009) Predicting new molecular targets for known drugs. *Nature* 462: 175-182.
48. Reker D, Rodrigues T, Schneider P, et al. (2014) Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci USA* 111: 4067-4072.
49. Liu X, Vogt I, Hague T, et al. (2013) HitPick: A web server for hit identification and target prediction of chemical screenings. *Bioinformatics* 29: 1910-1912.
50. Gong J, Cai C, Liu X, et al. (2013) ChemMapper: A versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics* 29: 1827-1829.
51. Iorio F, Bosotti R, Scachen E, et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Nat Acad Sic USA* 107: 14621-14626.
52. Zhu X, Kim JL, Newcomb JR, et al. (1999) Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. *Struct Fold Des* 7: 651-661.
53. Goto M, Miyahara I, Hirotsu K, et al. (2005) Structural determinants for branched-chain aminotransferase isozyme-specific inhibition by the anticonvulsant drug gabapentin. *J Biol Chem* 280: 37246-37256.
54. SiteMap, version 2.7; Schrödinger, LLC: New York, 2012.
55. Halgren T (2009) Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* 49:377-389.
56. Schrödinger Release 2014-1: Desmond Molecular Dynamics System, version 3.7, D. E. Shaw Research, New York, NY, 2014.

© 2014, Wayne C. Guida and Wesley H. Brooks, et al., licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)