How to Valorize Biodiversity? Let's Go Hashing, Extracting, Filtering, Mining, Fishing

Authors Quoc Tuan Do[1], José L. Medina-Franco[2], Thomas Scior[3], Philippe Bernard[1]

Affiliations
[1] Greenpharma S. A. S., Orléans, France
[2] Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Mexico City, Mexico
[3] Department of Pharmacy, Benemérita Universidad Autónoma de Puebla, Puebla, México

Key words
- chemogenomics
- data mining
- inverse docking
- profiling
- reverse pharmacognosy
- similarity searching

Abstract

Nature was and still is a prolific source of inspiration in pharmacy, cosmetics, and agro-food industries for the discovery of bioactive products. Informatics is now present in most human activities. Research in natural products is no exception. *In silico* tools may help in numerous cases when studying natural substances: in pharmacognosy, to store and structure the large and increasing number of data, and to facilitate or accelerate the analysis of natural products in regards to traditional uses of natural resources; in drug discovery, to rationally design libraries for screening natural compound mimetics and identification of biological activities for natural products. Here we review different aspects of *in silico* approaches applied to the research and development of bioactive substances and give examples of using nature-inspiring power and ultimately valorize biodiversity.

Abbreviations

ADMET:	absorption, distribution, metabolism, excretion, toxicity
ANN:	artificial neural networks
BBB:	blood-brain barrier
COX:	cyclooxygenase
1D, 2D, 3D:	one-, two-, three-dimensional
DNMT:	DNA methyltransferase
EPS:	electrostatic potentials
FAK:	focal adhesion kinase
FEMA:	Flavor and Extract Manufacturers Association
GRAS:	generally recognized as safe
HERG K+:	human ether-a-go-go-related gene potassium
HDAC1:	histone deacetylase-1
HTS:	high-throughput screening
MOE:	molecular operating environment
NABATIVI:	Novel Approaches to Bacterial Target Identification Validation and Inhibition
PCA:	principle component analysis
PD:	pharmacodynamic
PESD:	properties encoded shape distributions
PK:	pharmacokinetic
PLA2:	phospholipase A2
PPAR:	peroxisome proliferator-activated receptors
QSAR:	quantitative structure-activity relationship
R&D:	research and development
SAR:	structure-activity relationships
SPID:	structure-promiscuity index difference
SVM:	support vector machine
UFSR:	ultrafast shape recognition
vHTS:	virtual HTS
VLS:	virtual library screening
ZINC:	free database of commercially available compounds for virtual screening

received July 15, 2014
revised October 21, 2014
accepted January 9, 2015

Bibliography
DOI http://dx.doi.org/10.1055/s-0034-1396314
Published online February 25, 2015
Planta Med 2015; 81: 436–449

ISSN 0032-0943

Correspondence
Dr. Quoc Tuan Do
Greenpharma S. A. S.
3 allée du Titane
45100 Orléans
France
Phone: + 33 2 38 25 99 80
quoctuan.do@greenpharma.com

Introduction

"'Biological diversity' or 'biodiversity' means the variability among living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are a part; this includes diversity within species, between species and of ecosystems" [1]. Biodiversity is endangered by human activities, and its decline in some regions is exacerbated by climate changes. The extent of such modifications of the environment depends on complex criteria including geographical, environmental, political, and societal conditions [2, 3].

This makes global protection policies, though necessary, very difficult to implement. Why and how to protect biodiversity? Some encouraging attempts have been made by regional actors to stimulate industry, nonprofit, and academic research in chemical and life sciences [4], national governments for biodiversity protection and societal issues [5] or, at a continental level such as the European FP7 project Marex composed of nine European Union countries and four developing countries, to explore the possible industrialization of bioactive substances from marine resources [6]. Ethical and utilitarian arguments are the common points from the numerous existing examples. Sustainable developments can be envisaged to valorize biodiversity (i.e., to estimate economic value, to highlight its value and/or to increase its value) in a variety of domains such as biofuels to replace fossil energy [7], materials, e.g., batteries made with emodin derivatives [8], biomimetics, i.e., nature as a source of inspiration to design new materials, processes, etc… [9], cosmetics [10], functional foods [11], or pharmacy [12].

Historically, natural products, e.g., plants, have been a source of food and medicine; as a matter of fact, in ancient civilizations, the two things are "interchangeable" according to Hippocrates. Therefore, a lot of knowledge was accumulated as evidenced by traditional Chinese medicine's so-called "Yellow Emperor's Inner Classic" or Dioscorides' "*De Materia Medica*", to cite a few. Nowadays, with the evolution of analytical techniques [13] and the fast development and advances of computers, the amount of data about natural products has grown drastically [14–16]. This allows for the emergence of new strategies to valorize the natural products, such as reverse pharmacognosy [17,18], by working with natural flavor molecules [19,20], by relating traditional medicine concepts to modern Western medicine pathologies [21], or by using ancestral knowledge as a starting point for scientific investigations [22]. In this review, we focus on the contributions of natural products in drug discovery aided by *in silico* techniques with a particular emphasis on the most frequently used approaches such as database mining, systematic screening using similarity searching and molecular docking, and inverse docking techniques.

Drug Profiling during Research and Development

The term "drug profiling" is commonly used by academic groups, the pharmaceutical industry, and other institutions with drug research centers to define the experimental – and sometimes computational – measurements of physicochemical and pharmacokinetics properties, and the biological activities of their new drug candidates during R&D processes [23].

Pharmaceutical profiling

Pharmaceutical profiling provides opportunities to deprioritize or eliminate undesirable molecules with unsuitable characteristics during the early stages of drug discovery. This practice has a great impact on the costs of R&D since unnecessarily passing along the R&D pipeline a plethora of non-promising pharmaceutical agents is becoming even more labor intensive and resource demanding with each new R&D stage reached [23]. During the error-prone attrition phase of discovery, when sorting out and reducing the amount of drug candidates, many research sites are establishing in-house drug candidate property guidelines based on scientifically sound concepts. In addition, thanks to their high-speed and low-budget nature, computational (*in silico*) tools have since been applied to complement or – on occasions – even substitute certain laboratory assays [23,24]. Those tasks where they have been proven to be really good at encompass the predicting of physicochemical properties and, to a lesser extend, pharmacokinetics for ADMET modeling. The latter simulates at a molecular level and numerically describes biological processes of drug absorption, body distribution, biotransformation (metabolism), excretion, elimination as well as toxic behavior. Nowadays, *in vitro* biological screening is the preferred tool for PK profiling [24]. It is undeniable in daily practice that a trade-off does exist between the fast, neat, and clean computational methods to calculate properties at the expense of data reliability and the by far more expensive and time-consuming techniques of experimental measurements in high-tech laboratories [24].

Experimental and *in silico* profiling

Many observed parameters can also be estimated with computer-based software [25,26] (**Table 2** in [27]). "Wet" HTS of compounds can be imitated by vHTS to identify promising candidates for further lead optimization and gives feedback about the identified single drug target (**Table 1** in [27]). If a vHTS does not exist, it can also be carried out against a pharmacophore model (substructures' interaction of the ligands) [28]. Computer programs used to predict the substrate selectivity and the regioselectivity (structures of metabolites, sites of metabolism on the substrates) are presented in the literature [29]. Recently, an alternative strategy to single target-based screening has been proposed by Fang using phenotypic profiling [30]. He combined the examination of the biological endpoints (drug effects) on a specific phenotypic behavior in cells, tissues, or whole animals. The advantage is that drug candidates can show their overall disease-modifying action based on simultaneously hitting several hitherto unknown biomolecular targets in the cells [30]. Targeting more than one target (multitarget paradigm) has received attention as a feasible approach in the literature [31]. The latter must only be identified when the drug candidates are selected as hits. The logical workflow can be summarized as: (1) selection of disease with associated phenotypic endpoints (controlled symptoms); (2) phenotype profiling and endpoint(s) screening (by HTS); (3) intracellular, biomolecular target identification upon hitting; (4) compound library expansion to enrich it with more promising candidates; (5) *in silico* studies like vHTS (in parallel with HTS) and computational similarity analysis based on the chemical structures of the early hits for lead structure prioritization, ligand docking to target structures, lead compound optimization, VLS [27], QSAR [28] as well as docking studies to search for similar substances for compound library expansion; (6) drug safety profiles and tox screens; (7) preclinical studies; and, finally, (8) clinical trials (adapted from Fig. 1 in [30]).

Biopharmaceutical profiling

Historically, drug profiling focused on PD as the pharmacological endpoint with means to describe the molecular mode of action. Typical efforts embraced *in vitro* ligand binding assays, and ligand protein crystallography [24]. All too often the promising substance did not reveal its poor biopharmaceutical (formulation incompatibilities) or PK behavior (ADMET) in the initial stages of development but rather at the very end of the long road with the fatal consequences of losing time and money, or even worse with the complete loss of the candidate as a new drug in the pipeline [23]. Another paradigm has been changing during the last decade or so when shifting from late stage profiling to early stage

Table 1 Empirical descriptors or patterns for a typical biopharmaceutical profile.

Descriptors/parameters/patterns or features	References
MW < 500	[33]
logP < 5	[33]
Hydrogen bond donors (HBD) < 5	[33]
Hydrogen bond acceptors (HBA) < 10	[33]
Number of rotatable bonds (nrb) < 10	[34–36]
Solubility logS at pH 6.5 > 10 mg/L	[32]
Topological polar surface area (TPSA) < 140 Å^2	[37, 38]
Aromatic rings < 4	[39]
GI tract and BBB permeability decreases with a lower log D < 0	[40]
Water solubility and renal excretion increases with a lower log D < 0	[40]
Water solubility and membrane permeability are "drug-like" in a range of log D between 0 to 3 units (0 < log D < 3)	[40]
Water solubility is increased by polar groups, hydrogen bonding, or dissociation into ions or permanent ionization (cations, anions)	[40]
Potency increases with higher log D > 5	[40]
Hepatic biotransformation by CYP450 increases with a higher log D > 5	[40]
Water solubility, oral absorption, and bioavailability tend to decrease with a higher log D > 5	[40]
Water solubility increases, and lipophilicity and membrane permeability (by passive diffusion of a given drug) diminish	[40]
Esters and other prodrug solutions increase lipophilicity if the acidic drug is too hydrophilic	[40]

Table 2 Natural product database.

Database name	Accessibility	Data types	Advantages	Drawbacks
AfroDB [60]	Freely accessible from the supplementary information of [60]	1000 Compounds; physico-chemical data and ADMET properties	Comprehensive predicted data	No data on plants
Chem NetBase [62]	Searches are free; results browsing under license http://dnp.chemnetbase.com	170 000 Natural compounds	Very comprehensive; frequently updated	Lack of organism data; commercial database
Dr Duke's database [63]	Freely accessible http://www.ars-grin.gov/duke	7500 Molecules, 2000 organisms, 2200 traditional uses; biological activities	Many ways to query the database; huge amount of data	Lack of molecule data (structures, etc.); not updated since 1998
GPDB [10]	Greenpharma internal search	140 000 Compounds, 160 000 organisms, 4360 targets, 10 000 activities, 1000 traditional uses	Rich query system; structural searches; numerous links between data	Lack of data, but very frequently updated
KNApSAcK [64]	Freely accessible http://kanaya.naist.jp/knapsack_jsp/top.html	51 000 Molecules, 22 000 organisms, 110 000 metabolite/species pairs	Frequently updated; query to database can be implemented in software	No structural search
Napr alert [65]	Searches are free, but pay per view for results report http://www.napralert.org	200 000 Publications annotated; organisms; molecules; biological activities; ethnopharmacological data	Very comprehensive; frequently updated	Lack of molecule data; commercial database; lack of flexibility in results presentation
Pfaf [66]	Freely accessible http://www.pfaf.org	7000 Plants; traditional uses; medical and edible quality scores	Seldom used and original plants; highly suited to RPG	No molecule data
Supernatural [67]	Freely accessible http://bioinformatics.charite.de/supernatural/	46 000 Natural compounds; molecule characteristics; supplier data	Similarity searches	No organism data
TCM-ID [68]	Freely accessible http://bidd.nus.edu.sg/group/TCMsite/Default.aspx	12 000 Compounds; 1100 plants; 1200 TCM formula	Interesting relation with TCM-molecules	Cannot be exported
UNPD [69]	Free accessible http://pkuxxj.pku.edu.cn/UNPD	200 000 Compounds	Largest noncommercial and freely available database for natural products	No data on organisms

intervention. The ultimate goal of pharmaceutical profiling is to predict potential drawbacks concerning critical issues of PD and PK as well as the development of trial dosage forms or final delivery systems as early as possible and to evaluate drug usage and security risks in general [23].

Medicinal chemistry textbooks contain some popular rules of thumb like the empiric replacement patterns for chemical groups known as biostereo-isomerism, the traffic light scheme for Lobel's "Oral PhysChem Score" [32] or "Lipinski's rule of five" [33], which can be embedded in drug-likeness screens or biopharmaceutical profiling efforts (◉ **Table 1**).

Computer-based profiling

Recently, a theoretical study demonstrated the toxicological characterization of a series of chemicals with cheminformatics. To this end, the cytotoxicity profile was estimated on the basis of structural molecular fragments to identify several moieties that can be regarded as bearing cell toxicity (cytotoxicophores) [41]. The detection of fragments with proven or alleged toxic properties, so-called toxicophores, can be carried out on the Web-based server Ochem, (https://ochem.eu/) via a link to ToxAlerts. The server also provides QSAR modeling. Another helpful, almost all-in-one solution is Vega ZZ for 3D model generation, biomolecules, manual docking, empirical molecular mechanics force field calculations or semiempirical quantum mechanics, to list only a few features [42]. The use is free for public universities and not-for-profit research institutes. As a general rule for the software novice, computed values can be used with confidence if the compound lies within the applicability domain (scope) of a program [43,44]. Software algorithms looking up databases or parameter sets when applying empirical equations are more susceptible than first principle *ab initio* methods. The latter are not foolproof either and can fail, too, if the underlying theory does not reflect natural processes. In general, conventional small organic molecules are more likely to be in the applicability or "calibration" range. Other structures fall short of expectations because they possess noncanonical electronic constellations, like carbamoyl, azid, nitro, sulfon, and metal organic groups or they are hydrazones, thioesters, etc. Sometimes, electronic, mesomeric effects depend on the conformation (between bridged aryl rings), and pK_a predictions tend to fail. Recent approaches successfully applied classification models for drug profiling in combination with public databases (PubChem [27,45], AntiMarin database [46]). The success was documented for some modeled activities that could be found in the literature and thereby confirmed [47].

Newman and Cragg have reviewed the contributions of natural products as sources of new drugs for three decades from 1981 to 2010 [12,48,49]. They are still important because they provide the final entity or are starting points in drug discovery (mimetics, derivatives, botanicals, etc.), particularly in oncology and infection domains. To date, only one *de novo* drug obtained from combinatorial chemistry has been approved during the reviewed period. What makes natural products so successful? What lessons can the medicinal chemist learn from natural products and their properties?

In the pharmaceutical industry the attrition rate remains very high [50], particularly at the later stage of expensive clinical trials. Therefore a "fail early, fail cheap" paradigm represents an attractive strategy. Many investigators have tried to capture the essence of existing drugs to extrapolate physicochemical criteria with the ease of implementation along the drug discovery workflow. ◉ **Table 1** lists examples of the different empirical descriptors derived from statistical mining of drug databases. Thanks to their numerical nature they can be used as prefilters in virtual library screenings, QSAR studies [28,51], or in compound selection in general. However, one must be cautious about their use. For instance, Lipinski's rule of five of "drug likeness" (oral delivery and passive absorption mechanisms) [33], though widely used, is not applicable to natural products and does screen out many drugs derived from natural products. Keller et al. [52] hypothesized that natural compounds may have evolved over millennia to take advantage of active transport or gained specific conformations suited to passive transport. For Kellenberger et al., the reason of natural drugs may be the similarity of interactions of natural products with biosynthetic enzymes and therapeutic targets [53]. As a large number of drugs are derived from natural sources, we see an overlap of the drug chemical space and the natural compound space [54]. Indeed, several authors have advocated mimicking certain physicochemical profiles of natural compounds to synthesize compounds that are more diverse and biologically relevant [54–57], e.g., a reduced number of nitrogen atoms or aromatic rings, the presence of nonaromatic, polycyclic core structures, etc.

HTS assays can be perturbed by certain chemical features, generating a false positive. Rishton surveyed such reactive functional groups in [58]; among them some can be frequently found in active natural compounds and drugs, such as aldehydes, aliphatic esters and ketones, epoxides, 1,2-dicarbonyl compounds (tanshinones), Michael acceptors (chalcones), peroxides (artemisin derivatives), and disulfides (glutathione disulfides). Some natural compounds may therefore be filtered out because they are not suitable for HTS.

Data Mining

▼

The scientific literature search, storing, and exploitation are imperative and have come to terms with data mining in the modern ages of electronic information technologies. Information that is produced and kept in-house (corporate data sources) is not publicly available. Proprietary data can be used by costumers on a commercial basis, while other sources lay open on the Internet (free web services). Helpful web sites to assist the profiling phases are scientific journals, patents, and bioinformatics services dealing with genomics, proteomics, and metabolomics. The literature survey compiles relevant data on comparative or disease-associated genetics, pharmacogenetics, pharmacodynamics, pharmacokinetics, metabolic and cellular signaling pathways, *in vitro* and *ex-vivo* (cell based) pathophysiological models, etc. [59]. Databases on traditional usage of natural substances are also booming with descriptions from the organisms to the molecules with folk uses, biological data, and predicted properties.

Ethnopharmacology data offers valuable "clinical" observations that can guide the drug discovery process (see ◉ **Table 2** for examples of database). Bernard et al. [19] gathered a list of plants used in several populations of Latin America in cases of insect stings or snakebites to find new anti-inflammatory agents targeting PLA2. Extracts of plants used by several populations for these ailments were tested in priority on PLA2. By searching for compounds that were common to the active extracts in their in-house database, the authors could narrow the possible candidates and perform docking on PLA2 to retrieve betulin and betulinic acid as potent inhibitors of PLA2. This prediction was further validated by *in vitro* binding tests. This work demonstrated the usefulness of computer-based analysis of ancestral knowledge to guide and accelerate the modern drug discovery process. Moreover, it also enabled experiments to prove some intuitive relations between folk medicine concepts (insect stings) and modern medicine pathology (inflammation). Rollinger et al. [21] have demonstrated the efficiency on combining *in silico* techniques with ethnopharmacological knowledge. Molecules from plants listed in Dioscorides' "*De Materia Medica*" as having "anti-inflammatory properties" were screened on structure-based pharmacophore models of COX 1 and 2. The hit rate using this procedure was about 100% higher compared to the same virtual screening

but on molecules from databases comprising marketed and development drug substances or natural compounds.

More recently, an interesting initiative by Ntie-Kang et al. [60] offered access to a database of more than 1000 natural compounds isolated from African medicinal plants, called AfroDB. The authors calculated numerous descriptors of drug-, lead-, and fragment-likeness and ADMET. They predicted the following ADMET properties: bioavailability, BBB penetration, dermal penetration, plasma-protein binding, metabolism, and blockage of the HERG K + channel. These parameters were also made available to the research community to help with compound selection, comparison, and virtual screening. The p-ANAPL library [61] containing most of the compounds of AfroDB was supplied upon request to the authors for *in vitro* validation. This type of initiative will no doubt encourage drug discovery from African plants and collaborations to valorize these resources.

The wealth of structural data on natural compounds allows investigators to compare drugs with natural compounds, and vice versa, to derive SAR and, subsequently, to deduce putative biological activities. Similarity searching is a fertile approach in drug discovery.

Similarity Searching

▼

Similarity-based screening or similarity searching is a typical ligand-based approach that can be conducted without prior knowledge of the 3D structure of the target. This approach is based on the notion that similar compounds have similar activity [70]. Remarkable exceptions to this concept are the "activity cliffs", i.e., similar compounds with an unexpectedly high activity difference. The interested reader is referred to reviews that address in detail the role of activity cliffs in medicinal chemistry and elaborate on the computational approaches to identify them [71,72].

Similar to other computational screening efforts, similarity searching should be part of an iterative process that involves the prediction, experimental testing of selected compounds, and design of new chemical data sets based on the structure of the experimental hits. Also, if enough information of the system is available, e.g., 3D coordinates of the target, similarity searching should be combined with other ligand-based and/or structure-based methods. The selection of a particular approach or set of methods depends on the aim of the project, the information of the system, and the computational resources available. Moreover, one needs to consider the inherent limitations of each step involved and the associated computational cost.

In natural products research, the combination of computational approaches has been emphasized by Yue et al., who have recently discussed progress on the target profiling of natural products using experimental (genomics and proteomics) and computational approaches [73]. In that review, Yue et al. emphasized the convenience of integrating various methods, such as inverse docking (docking compounds across different targets), mapping ligand-target profiling space, and network analysis.

Similarity searching can be combined with other current major strategies in drug discovery such as drug repurposing. A recent example of this successful synergy is the similarity searching of a database of approved drugs that led to the identification of olsalazine, an anti-inflammatory drug approved for the treatment of ulcerative colitis, as a novel DNA hypomethylating agent [74].

Comprehensive reviews of virtual screening that cover methods, successful applications, pitfalls, and workarounds are published elsewhere [75–78]. Advances in the progress in the virtual screening of NPs have also been presented [79–85].

Any similarity searching involves several essential components, which are briefly outlined below.

A) One or more query or reference molecules that are compared against a molecular data set. The reference molecule is typically a chemical structure that can be represented in 2D or 3D. In general, in similarity searching, a notable advantage of 2D over 3D approaches is computational speed since most 2D methods (with the exception of those using chemical graphs) do not require costly structure alignments. In contrast, many but not all 3D methods require such alignments [86]. Moreover, 3D approaches have to deal with the conformational flexibility of the molecules, which, in many instances, give rise to multiple low-energy conformers. Diverse solutions have been proposed to alleviate this problem [87]. Currently, most 3D similarity searching studies use a single low-energy conformer (usually the global minimum or other representative 3D conformation). This, in any case, raises the question if such a conformation is biologically significant [88].

The performance of 2D and 3D similarity approaches has been compared directly in a number of applications, including virtual screening [89–92]. Since 3D similarity searching should incorporate, at least in principle, more accurate features than 2D methods, it would be expected that the results obtained from 3D methods should be more reliable than those obtained by 2D methods. However, in many instances, 2D approaches have outperformed 3D approaches, although it has been noted that this superiority is somewhat case-dependent [92].

Depending on both the data set and the biological activity, it is feasible that one or more reference compounds are associated with activity cliffs. In other words, they may be an "activity cliff generator" (defined as a molecular structure that has a high probability of forming an activity cliff with molecules tested in the same biological assay) [93]. Since, as discussed above, activity cliffs are exceptions to the similarity principle and lead to misleading results in similarity searching, it has been proposed that activity cliff generators be identified and removed from the data sets before selecting the reference compounds. In addition, the removal of activity cliff generators has been proposed as a general approach to be employed before developing predictive models, such as those obtained with traditional QSAR or other machine learning algorithms based on the similarity property principle [94].

B) Another element in similarity searching is the compound database. Compound databases have been reviewed elsewhere, including collections of natural products in the public domain [95]. A current trend in screening libraries for drug discovery is to balance chemical novelty with confined chemical space [96]. In this context, natural product databases (and natural product derivatives) are excellent sources for virtual screening as they expand the currently known medicinal chemistry space [97]. The "expansion" is associated in part with molecular complexity. This feature makes natural product databases attractive to identify compounds with a high selectivity towards molecular targets (including a target family) and can be ideal resources to identify "master key" compounds that selectively bind to a series of targets in order to yield a desired clinical effect [31]. Examples of specific and appealing regions in chemical space covered by databases of natural products include peptides and macrocycles [96].

C) A third and critical component in similarity searching is chemical representation, which is at the core of virtually any chemoin-

Table 3 Representative and recent studies using similarity searching to uncover bioactive compounds in natural products and related compounds.

Study	Similarity searching method used	Ref.
Sequential virtual screening of ZINC natural compounds identifies five compounds as PPAR-γ partial agonists.	Electrostatic and fingerprint-based similarity analysis combined with ADMET and structure-based filtering.	[103]
Sequential docking-based virtual screening followed by similarity searching to select promising inhibitors of DNMT1 in two natural products collections.	Fingerprint-based similarity searching using MACCS keys and the similarity coefficients Tanimoto and Tanimoto-substructure.	[104]
Searching of GRAS compounds to uncover compounds similar to approved antidepressants. Identification of nonanoic acid and 2-decenoic acid (similar to valproic acid) as inhibitors of HDAC1.	Fingerprint-based similarity searching with MACCS keys/Tanimoto.	[19]
Structural comparison of the FEMA GRAS list with analgesics and with compounds used as satiety agents.	Comparison based on physicochemical properties and seven structural representations obtained from three different software programs.	[105]
Similarity searching to identify compounds in a compiled database of phytochemicals with activity against a protein involved in the colon cancer pathway or a colon cancer drug target.	Text mining in PubMed abstracts led to the collection of more than 20 000 diverse chemical structures present in the human diet. Authors systematically explore their numerous targets using chemoinformatics methods.	[106]

formatics application. However, chemical representation is not an easy task because similarity is a subjective concept. It is largely known that chemical space (including similarity searching) depends heavily on molecular representation. It has been shown that if one uses different representations in similarity searching, the hit compounds (the most similar molecules to the query) will likely be different [98]. In actual applications of similarity searching, and molecular similarity analysis in general, a number of different types of representations are used. The information contained in the representations is usually in the form of molecular or chemical features called descriptors that are obtained from the structural and chemical properties of molecules. Descriptors are nominally classified as 1D, 2D, or 3D. 1D descriptors are commonly related to whole molecule properties such as molecular weight, logP, solubility, number of hydrogen bond donors, number of rotatable bonds, etc. 2D descriptors are associated with the topological structure of molecules as typically depicted in chemists' drawings. This type of representation shows the atoms, the bonds connecting them, and in some cases includes stereochemical features, but they do not explicitly depict the 3D structures of molecules. 3D descriptors, as their name implies, are associated with the 3D structures of molecules [88]. Todeschini and Consonni have assembled a comprehensive list of the descriptors used in chemical informatic applications [99].

Despite the fact that many descriptors are available, it is highly unlikely that a single representation and set of descriptors will capture all of the many different aspects of molecular and chemical information [88]. Therefore, in order to reduce the impact of the dependence of chemical representation in similarity searching, it has been proposed to use several methods and then combine the solutions. This is called "data fusion", and the group of Willet is a pioneer in this field [100]. A recent exhaustive study conducted by Holliday et al. [101] provides strong evidence that suggests that fusion-based approaches to similarity searching yield improved results over single-search-based similarity methods. Following a similar approach, the use of several molecular representations and then the combination of such representations has been implemented in different areas of chemoinformatics, including activity landscape modeling. In the latter, the term "consensus activity cliffs" have been proposed [102].

D) A fourth component of similarity-based virtual screening is a similarity measure which, in turn, depends on three elements: (1) the representation used to encode the desired molecular and chemical information, (2) whether and how much information is weighted, and (3) the similarity function, also called the similarity coefficient, that maps the set of ordered pairs of representations onto the unit interval of the real line [88].

Using the components of similarity searching outlined above, different groups have been using similarity searching alone or in combination with other computational approaches to uncover bioactive compounds from natural products. Examples of recent investigations are summarized in **Table 3** and described in the following paragraphs.

Guasch et al. used a combination of computational methods to identify five PPAR-γ or PPARG [107] partial agonists from a compound collection with more than 89 000 natural products and natural product derivatives from ZINC [108]. The authors of that work implemented a sequential or cascade virtual screening approach using a set of ADMET filters, structure-based pharmacophore screening, molecular docking, electrostatic, and fingerprint-based similarity analysis. A total of ten compounds with different chemical scaffolds were selected for experimental validation using *in vitro* assays. All five compounds were confirmed as PPAR-γ partial agonists [108].

Also in a combined approach, Medina-Franco and Yoo implemented a sequential computational screening of five compound libraries to identify candidate compounds for testing as potential inhibitors of DNMT1. The reference molecule was a known DNMT inhibitor recently identified from HTS whose chemical structure was made publicly available in PubChem. The compound databases screened included two collections of natural products, a DNMT-focused library, a general screening collection, and a set of approved drugs. Similarity searching was performed using the widely used MACCS keys (166 bits) as implemented in MOE. The molecular similarity was computed using two measures, Tanimoto and Tanimoto-substructure. Of note is that Tanimoto-substructure takes into account the putative different sizes of the query molecule and the compounds in the databases screened. Compounds selected from similarity searching were subject to docking with a crystallographic structure of human DNMT1 using a validated docking protocol. At least 108 molecules with promising DNMT1 inhibitory activity were identified. The chemical structures of the computational hits were disclosed to encourage the research community working on epigenetics to experimentally test the enzymatic and demethylating activity *in vivo* [104].

Feng et al. [109] used chemoinformatics analysis based on Lipinski's rule-of-five, ChemGPS-NP [110] principal component analysis, and chemical clustering to compare a set of antitrypanosomal marine natural products with approved drugs to prioritize products with a similar profile as the reference drugs.

GRAS compounds are largely comprised of natural products. A recent and notable application of similarity searching of GRAS compounds for bioactive compounds is represented by the work of Martinez-Mayorga et al. In that work, the authors searched for similar structures to approved antidepressant drugs in the food flavoring components in the FEMA GRAS list [19]. The virtual screening was conducted using fingerprint-based similarity searching with the MACCS keys and the Tanimoto coefficient. Hit compounds in the FEMA GRAS list were chosen as the most similar compounds (ranked with the highest similarity values) to any of the 32 approved antidepressant drugs. Selected compounds represented the "nearest neighbors" of the approved antidepressants. Valproic acid was the most similar antidepressant to GRAS molecules. Based on the knowledge that the inhibition of HDAC1 could be related to the efficacy of valproic acid in the treatment of bipolar disorder, Martinez-Mayorga et al. screened the GRAS compounds most similar to valproic acid for HDAC1 inhibition. The GRAS compounds nonanoic acid and 2-decenoic acid inhibited HDAC1 at a micromolar level with a potency comparable to that of valproic acid. Of note is that the GRAS chemicals were not expected to have strong enzymatic inhibitory effects at the concentrations typically employed in flavor formulations designed for use in foods and beverages. However, as shown in that work, GRAS chemicals were able to bind to a relevant therapeutic target. That study also served as a proof-of-principle of the feasibility of exploring the FEMA GRAS flavoring list using computational methods as a potential source of biologically active molecules. In addition, the study demonstrated that similarity searching followed by experimental evaluation could be used for rapid identification of GRAS chemicals with potential bioactivity [19].

In two subsequent and separate studies, Martinez-Mayorga et al. employed structural similarity to compare the FEMA GRAS list with analgesics and with compounds used as satiety agents [105]. The list of analgesics used as query molecules contained ten structurally diverse molecules currently used in clinics. A total of eight satiety agents were identified in the literature, which were used as reference compounds for similarity searching. The satiety agents included those currently used in clinics, as well as those still in clinical trials. In both studies, reference compounds were compared with the FEMA GRAS list using a total of seven structural representations obtained from three different software programs, MOE, ChemAxon, and PowerMV. Compounds identified by different programs and representations were chosen as consensus compounds for additional studies. Then, a chemical space was constructed based on physicochemical properties. Nearest neighbors were identified based on Euclidian distances, considering all the dimensions (properties). Based on the comparison of structural features and physicochemical properties, two FEMA GRAS compounds were selected as being similar to the reference analgesics. In the second study, a total of nine FEMA GRAS molecules were identified as being similar to those used as reference satiety agents. For compounds having a known mode of action, *in vitro* studies using the identified GRAS chemicals could help determine whether or not they may have a satiety or analgesic effect in humans. However, it must be considered that in the large majority of cases biological effects result from complex and multiple interactions in the body [105].

As previously discussed in this review, phytochemicals derived from edible plants are notable sources of bioactive molecules. In a recent study, Jensen et al. [106] performed a high-throughput analysis of phytochemicals in order to reveal associations between diet and health benefits using text mining and chemoinformatic methods. The first step of that work was the retrieval of associations between the terms plants and phytochemicals from 21 million abstracts in PubMed/MEDLINE during the period 1998–2012. This information was merged with the Chinese Natural Product Database and the Ayurveda data set, which was also curated by the authors. The final data set included nearly 37 000 phytochemicals. A major outcome of that study is the structured and standardized database of phytochemicals associated with medicinal plants. The authors pointed out that their approach facilitates the identification of novel bioactive compounds from natural sources, and the repurposing of medicinal plants for diseases other than those for which they are traditionally used, with the added benefit that the information collected can help elucidate a mechanism of action [106]. As a case study, Jensen et al. conducted structural similarity searching in order to find molecules in their compiled database of phytochemicals with activity against a protein involved in the colon cancer pathway or a colon cancer drug target. The reference compounds were those reported in ChEMBL. A set of molecules from this study not only showed reported health benefits against colon cancer, but activity was also verified against colon cancer protein targets [106].

Polypharmacology and Chemogenomics in Natural Products Research

The increasing awareness that a drug may have its clinical effect through the interaction of multiple targets (called "polypharmacology") is changing the drug discovery paradigm from a single target to a multi-target approach [31]. This change is enriching chemogenomics data sets that capture ligand-target relationships [111]. As a consequence, a number of computational and experimental approaches are being developed to generate, store, analyze, mine, and visualize target-ligand interactions that define chemogenomic spaces [112–114].

Using the literature reports, the identification of the pharmacological evaluation of compounds (in particular with novel chemical structures) isolated from natural sources is frequent. The pharmacological evaluation usually includes a handful of biological endpoints. In light of the generation of chemogenomics data sets, natural products are being evaluated systematically across a large number of biological endpoints, and the screening data is being released to the public. A representative example of a chemogenomics data set that contains natural products is the large microarray data released by Clemons et al. [115]. In that work, the authors evaluated the binding specificity of 2477 natural products (which were part of a larger collection with 15 000 compounds) across 100 sequence-unrelated proteins. The authors released the results of the screening to the public domain (the interested reader has access to the screening data along with the chemical structures in the paper of Clemons et al. [115]). The microarray data set has been analyzed with chemoinformatic approaches with the goal of elucidating the SAR; in particular to uncover structural characteristics related to the selectivity or promiscuity of the molecules using fingerprint or substructure rep-

resentations [116–118]. For instance, Yongye and Medina-Franco developed the SPID metric to quantify and uncover specific structural changes that have a significant impact on the number of proteins to which a compound binds [116]. In a subsequent publication, Dimova et al. reported an analysis of the same data set using matched molecular pairs [119] to identify single-site substitutions that are associated with large magnitude differences in apparent compound promiscuity. The results of Dimova et al. further confirmed the results of Yongye and Medina-Franco previously published in that promiscuity can be induced by small chemical substitutions.

Docking

The concept of one disease/one target was a milestone in modern molecular medicine because it enabled the simplification of complex *in vivo* symptoms and related them to simple *in vitro* models. Though this paradigm is shifting to multitargets [31] as our knowledge progresses, this reductionist approach did prove successful in many diseases. To better understand the molecular mechanism of action of molecules on their biological targets, several methods were developed to determine the 3D structure of these proteins, e.g., X-ray, NMR, and electron microscopy. During the past decades, the number of solved protein crystallographic structures grew exponentially, and now tops at 94 000 structures (statistics from the PDB homepage [120]). At the same time, computer power has also increased dramatically. Molecular modelling software could then be developed to exploit these types of data. The first docking software was DOCK [121]. Docking refers to methods that predict the orientation of a molecule bound to another. The stabilities or the affinities of the resulting complex are estimated by a mathematical or scoring function [122]. Many different strategies and algorithms for docking exist, e.g., AutoDock [123], FlexX [124], Glide [125], Gold [126], and Surflex [127], to predict the positioning of molecules into the protein-binding site. Authors have also studied scoring methods to improve the hit rate. Many are related to their cognate docking software (refer to the review by Li et al. [128] describing 20 scoring functions). Because different docking software and scoring functions have different strengths and weaknesses, several authors tested the combination of docking and scoring methods to find the optimal procedure [129], while others proposed to use consensus scoring to accommodate the weaknesses [130–133]. The scoring of the predicted poses from docking will be performed by several scoring functions, not only by one. Predictions well scored by multiple scoring functions will be better ranked. Interested readers can refer to [134] for a review of several consensus-scoring methods. As the scoring is dependent on the pose predictions, authors have also worked on improving this step by using consensus docking. It consists in retaining the poses predicted by a majority of docking software [135–137]. In an ideal case, the software can be selected because the natural substance is structurally related to either a ligand or a receptor, or even both, which belong to the software's calibration set resulting in a higher confidence that the computed solutions are trustworthy [138]. Sometimes docking problems arise when the target receptor is a constitutively inactive mutant or exists in unliganded states (inactive vs. active); it could also be under allosteric control (conformational modulation) [139].

Natural products remain a large source of active products and also an inspirational source for medicinal chemists; most of the resources, particularly from the microorganisms, are underexploited [12]. Structure-based techniques constitute a possible way to find new applications to these natural products. The majority of drugs in oncology and biocide products are derived from natural products. It is not surprising that many docking studies with natural products fall in these therapeutic domains.

Thiyagarajan et al. [140] targeted FAK by docking a library of 109 natural products. Four selected candidates showed activity of C6 glioma and N18 neuroblastoma cell lines by promoting apoptosis. Medina-Franco and Yoo [104] screened by combining structure-based pharmacophore filtering and docking on DNMT with a library composed of natural products, approved drugs, a DNMT-focused library, and general screening compounds. One hundred and eight potential hits were disclosed to the scientific community for experimental validation. Hussain et al. [141] adopted a docking strategy coupled to a 3D-QSAR to predict the activity of the analogues of aplyronine A that bind to actin. Their models may be helpful in designing more efficient and tolerable antitumor agents.

Docking may be used to assess the binding mode of natural products and subsequently guide the design of more potent candidates. For instance, the comparative docking of forskolin (activator) and labd-13(E)-ene-8a,15-diol diterpene (inhibitor) into the active site of adenylyl cyclase revealed important features in the binding mode of the activator and the inhibitor, allowing for the design of potential cytotoxic and cytostatic agents against cancer cells [142].

Due to antibiotic multiresistant bacteria, finding a new class of antibiotics with a new mode of action has become of paramount importance. This can be evidenced by numerous public fundings at national or international levels. A list of the multimillion Euros projects financed by the EU can be found in [143]. It is noteworthy that the NABATIVI project [144] succeeded in finding a peptidomimetic product with a new mode of action targeting a membrane receptor [145]. This product is in clinical trial phase II. Docking techniques were extensively applied not only to discover new antibiotics but also antiviral, antifungal, or antiprotozoan products. One strategy that is successful so far is to target essential bacterial genes, whose inhibition will kill the microorganisms. An example of a structure-based screening on an essential gene such as the filamenting temperature-sensitive mutant Z (FtsZ) provides promising leads [146]. Other authors also performed docking studies to demonstrate the bactericidal potentiality of xanthone derivatives [147]. An interesting use of docking was exemplified by Harris et al. [148]. They performed docking on the bacterial essential enzyme peptidyl-tRNA hydrolase to identify possible active compounds and guide their activity-directed isolation to discover antibacterial molecules from an ethanol bark extract of *Syzygium johnsonii*. For examples of drug discovery from natural compounds using docking to target viruses, fungi, and protozoan parasites, the reader is invited to consult the following respective works: [149] reviewed several *in silico* approaches to tackle urgent threats caused by new viruses or their variants (HIV, SARS, etc.) and how helpful computational techniques were to disclose the antiviral properties of natural products; docking studies helped to hypothesize the mechanism of action of antifungal pyranocoumarin derivatives in [150,151]; the authors performed docking studies with geldanamycin targeting the HSP90 homolog proteins of pathogenic protozoans *Plasmodium falciparum, Leishmania donovani, Trypanosoma brucei*, and *Entamoeba histolytica*. This work allowed for designing

Table 4 Databases useful for target fishing.

Database name	Accessibility	Data types	Advantages	Drawbacks
PDB [155]	Freely accessible http://www.rcsb.org/pdb	94 000 Protein structures with unique PDB code	Reference database; standard PDB format	Lack of data about biological activities
BRENDA [156]	Freely accessible http://www.brenda-enzymes.org	4800 Enzymes; ligands; organisms; biological activities	Very comprehensive database	Only enzymes
TTD [157]	Freely accessible http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp	1900 Targets; 5000 ligands; biological pathways and activities; patents	Very useful for reverse pharmacognosy; frequently updated	Relatively small amount of data
PDTD [158]	Freely accessible http://www.dddc.ac.cn/pdtd/index.php	1200 Aelected protein structures; biological activities; cross-linked with other databases	Link to TarFisDock, an inverse screening platform	Relatively small amount of data
Sc-PDB [159]	Freely accessible http://bioinfo-pharma.u-strasbg.fr/scPDB	3D structures selected from PDB	Useful to enrich a target database for inverse screening	Only a subset of PDB
Drug Bank [160]	Freely accessible http://www.drugbank.ca	2500 Proteins; 4800 drugs; pathways	Important part of FDA-approved drugs and proteins	Lack of data about biological activities
ChEM BL [160]	Freely accessible https://www.ebi.ac.uk/chembldb/	1.4 Million compounds; 10 000 targets; 13 millions activities	Very comprehensive	Results may be complex to analyze

selective analogues of protozoan HSP90 with a reduced affinity to the human homologue.

Finally, some investigators applied docking on several targets to identify molecules with synergistic effects on a particular biological pathway, e.g., modulation of testosterone [10]. Bernard et al. identified honokiol as a dose-dependent inhibitor of aromatase and 5-alpha-reductase 1; the inhibition of both enzymes mitigates the decrease of the testosterone level in aging men. Noteworthy is that honokiol is not active on the 5-alpha-reductase 2.

Target Fishing

The researchers quoted above not only applied docking to screening but also for identifying putative interacting protein partners (or "target fishing"), hence the mode of action of natural compounds. In each case, the authors have to hypothesize the possible target based on an "educated guess" or hints from the scientific literature. A docking study is performed with the active molecules to the selected protein target, and the score of the complex is evaluated. According to this score, the authors will then judge the plausibility of that ligand-protein interaction. An obvious caveat of such an approach resides in the picking of the targets, which will miss targets that are not evident or targets not yet known to be related to the biological effects.

To circumvent this difficulty and explore systematically possible interactions of a molecule with proteins, inverse docking was first introduced by Chen and Zhi in 2007 [152]. It consists in docking a molecule to a set of 3D protein structures. Therefore, inverse docking is in need of a docking program (see previous section) or a more specific tool in combination with a database of 3D protein structures (see **Table 4** for a list of possible databases). Docking software generally lacks the ability to correctly rank possible ligands in one site. This represents a serious limitation. Several authors developed corrected scoring functions to work around this limitation [153] and demonstrated the feasibility of this technique [154]. Vigers and Rizzi [153] showed that their new scoring function could assess the selectivity of compounds among a family of proteins, such as kinases, and selectivity among proteins of unrelated families.

Inverse docking plays a key role in the concept of "reverse pharmacognosy" introduced by Do and Bernard [161] and extended by Blondeau et al. [17]. Pharmacognosy starts with natural sources (e.g., extracts of plants and microorganisms) and thanks to activity-guided fractionation, identifies the molecule(s) responsible for a biological activity. Conversely, reverse pharmacognosy begins with a natural molecule and, thanks to inverse docking, identifies putative targets of interest. The predictions are then validated with related *in vitro* assays. Thanks to a database linking molecules and the organisms producing them, we can identify new applications for plants, for example, with the mode of action at the molecular level (ligand-protein interactions). With this approach, Do et al. could identify protein interacting partners for epsilon-viniferin from *Vitis vinifera*, which inhibits phosphodiesterase 3 and 4 [162], and for meranzin from *Limnocitrus littoralis*, which blocks COX 1 and 2, and activates PPAR-γ [163]. Thus, extracts from these two plants at an adequate concentration of the active molecules may be used in indications involving the described proteins.

Other scenarios of inverse docking were described for pharmacological profiling of natural products [164], either to understand the mode of action as well as repurpose molecules, e.g., tanshinone II a [165], or to evaluate the toxicity profile [166].

Only a few software programs have been developed for inverse docking, but the field is gaining more and more attention, as we can notice through the development of tools based on an existing docking engine or on a specific software: Invdock [152], iRAISE [167], Mdock [168], Selnergy [161] based on the Surflex programme, Tarfisdock [169] based on the DOCK programme, and TarSearch-X [170]. Inverse docking is not yet mature technology but should mutually fertilize other approaches, e.g., chemogenomics and bioinformatics.

It should be mentioned that inverse docking is one out of the several techniques available to conduct target fishing. Other common approaches such as data mining and similarity searching (see above) are extensively used to explore putative targets of bioactive compounds. In similarity searching, targets are represented by their ligands and query molecules are compared with the known ligands. Based on the concept of SAR, similar molecular structures will certainly have similar biological activities.

Thus, by finding similar ligands to query structures, one can relate the query compounds to the ligands' targets. Databases such as DrugBank, PubChem, and ChEMBL [171] are key to have as many as possible interaction pairs of ligands and targets. Machine learning techniques (e.g., Support Vector Machine, Neural Networks [172, 173]) are also popular to identify the relationship of molecules and possible targets. These systems are usually trained with a training set of known pairs of ligands/proteins based on descriptors, then validated with an external validation set (known pairs of ligands/proteins not used to build the models). We will evoke in the next paragraph the different types of descriptors. Structure-based and other computational approaches for target fishing are reviewed elsewhere [80, 174].

Pharmacophoric and Other Descriptors in Virtual Screening

A crucial point for the success of virtual screening is the design of the filter layer that constitutes the similarity patterns to retrain the potential candidates and discard the reminder. Many virtual screening programs have special graphical or scripting methods to write such filter definitions. Sometimes they consist of physicochemical properties, for example, "filter out all compounds with pK_a greater than … and/or without aryl rings … and/or with a nonpolar surface larger than". On occasion, they also describe a substructure of the scaffold common to all or almost all expected hit compounds. The underlying assumption is twofold: (1) the existence of a pharmacophore, "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" [78, 175] and (2) similar chemicals have similar biological activities [176]. Ligand-based virtual screening for structures with similar pharmacophoric patterns has become a successful method to identify potential drug candidates. Some of the latter find their role as lead compounds for lead expansion, lead hopping, and scaffold hopping in the desired therapeutic area. When 3D structures or homology models of the target protein are available, protein-based screening can be carried out. The pharmacophore defines spatial requirements like interatomic distances, angles, or the location of particular properties, and ionic sites as well as other descriptors that depend on the spatial coordinates. Such 3D information renders the screening query (filter) more precise but also more error-prone. Hence, not unexpectedly, researchers noticed that virtual screening based on 2D fingerprints (filter concepts based on atoms and bonds and their connectivity but without spatial coordinates) could be more successful than 3D pharmacophore. The authors recommended the combination of 2D and 3D descriptors [176]. Sometimes the sheer number of conformations under which compounds are collected is so overwhelming that the docking, screening, or simply identifying relationships between compounds based on their shape similarities risks overthrowing the computer resources at hand. Thus, screening for whole molecules, their side chain substituents, or their central scaffolds by conformationally-independent topomer similarities becomes a useful strategy. The partitioning of solutes in liquids along with surface defining descriptors, like nonpolar surface area, solvent accessibility, etc., is commonly applied in ADMET prediction models or in studies of membrane crossing, transport into cell compartments, or diffusion kinetics. The partial charges of the compounds can be calculated and projected as isocontour lines in the space surrounding the molecule of interest. TARIS is an approach based on such molecular EPS. The classes and types of descriptors are far too many, thereby lying beyond the scope of this review [177] (refer to topics in cheminformatics, e.g., GETAWAY, 3D MoRSE, MS-WHIM, FEPOPS circular fingerprints, MACCS keys, or graph-based multi-point pharmacophore as well as the so-called ROCS shape descriptors [178, 179]). ROC profiles (receiver operating characteristics) show a sort of hit enrichment in the final solution list against other compounds, e.g., decoys for testing (benchmarking), docking, and screening simulations. They have highly similar structures but are biologically inactive. A well-performing method should discard them from the hit list [180]. Although the number of descriptors used in a study may end up in the thousands, the right choice remains a challenging task of its own kind. Apparently, descriptors fail in reflecting in exactly which item the molecules resemble each other. PCA, a sort of statistical factorial analysis, simplifies the level of data complexity to a minimum set of orthogonal (independent) diagram axes (factors or components). PCA eventually sheds some light and explains some of the failures and downside when molecules are screened [181].

When needing to save time and running costs, HTS can be elegantly simulated *in silico* by screening virtual libraries (vHTS) [182]. To this end, vHTS descriptors have been developed, which do not need a lengthy superposition of data set molecules (for comparison) like PESD or UFSR. In addition, other techniques have been developed, for instance, 2D and higher dimensional QSAR, SVM, rule-based methods, or ANN [28]. ANN outperforms rule-based pharmacophore screens in those cases when decision taking in a straightforward manner is not behaving well, or tenets are ill-designed or believed to be just "better than no rule at all". Such rule-based pharmacophore screens generally make use of binary variables (simple "Yes/No" criteria) or integer values (the Lipinski's rule-of-five, more than five hydrogen bonds, etc.). The architecture of ANNs is based on a multilayer of criteria with individual weight put on them by training, i.e., probabilities of their contribution to yield the right answer which is *a priori* known during ANN training. What the ANN learned (as a black box to the user) through the training set of known cases is then applied to the test set. Although very complex phenomena can be handled, wrong answers emerge, mostly in cases when the molecule has unforeseeable characteristics.

Conclusions and Perspectives

Thanks to the large amount of information accumulated on natural product research, *in silico* techniques related to chemoinformatics, database mining, and molecular modeling facilitate the use of this information to further valorize natural products as a source and/or inspiration of drugs. *In silico* approaches enable the characterization of their physicochemical profile, analysis of chemical diversity, coverage of chemical space, and uncovering of trends in their SAR. The outcome of such analyses is valuable to guide medicinal chemistry efforts to optimize their properties or inspire the synthesis of novel scaffolds. Molecular modeling approaches, either ligand based or structure based, coupled with experimental methods, constitute techniques of choice to identify putative biological properties for natural products in a systematic manner and thus find ways to valorize them. To this end, numerous authors have applied computational structure similarity techniques to the GRAS list compounds [19] to repurpose them

as potential functional foods or use reverse pharmacognosy to find new uses for the molecules and their sources [18]. Although in this review we only examine health-related aspects of natural product utility, many applications can be found in numerous domains, such as material science and energy engineering among others. With new insights in microorganism biomes, the possibilities offered by Nature become even more tremendous [12], and preserving biodiversity has never been so crucial even at the restrictive economical point of view. It is anticipated that *in silico* approaches will continue to be part of the research to study and further potentiate the use of biodiversity.

Acknowledgements

We thank Dr. Karina Martinez-Mayorga for her insightful discussions, stimulating ideas, and fruitful conversations.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

1 Convention on biological diversity. Montreal: Secretariat of the Convention on Biological Diversity; 2012
2 *Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO.* The biodiversity of species and their rates of extinction, distribution, and protection. Science 2014; 344: 1246752
3 *Rao M, Htun S, Platt SG, Tizard R, Poole C, Myint T, Watson JE.* Biodiversity conservation in a changing climate: a review of threats and implications for conservation planning in Myanmar. Ambio 2013; 42: 789–804
4 *Camp D, Newman S, Pham NB, Quinn RJ.* Nature Bank and the Queensland Compound Library: unique international resources at the Eskitis Institute for Drug Discovery. Comb Chem High Throughput Screen 2014; 17: 201–209
5 http://www.ffem.fr/lang/en/accueil/activites-ffem/biodiversite_protection. Accessed September 26, 2013
6 http://www.marex.fi/. Accessed September 26, 2013
7 *Burton RA, Fincher GB.* Plant cell wall engineering: applications in biofuel production and improved human health. Curr Opin Biotechnol 2014; 26: 79–84
8 *Huskinson B, Marshak MP, Suh C, Er S, Gerhardt MR, Galvin CJ, Chen X, Aspuru-Guzik A, Gordon RG, Aziz MJ.* A metal-free organic-inorganic aqueous flow battery. Nature 2014; 505: 195–198
9 *Bhushan B.* Biomimetics: lessons from nature–an overview. Philos Trans A Math Phys Eng Sci 2009; 367: 1445–1486
10 *Bernard P, Scior T, Do QT.* Modulating testosterone pathway: a new strategy to tackle male skin aging? Clin Interv Aging 2012; 7: 351–361
11 *Graziose R, Lila MA, Raskin I.* Merging traditional Chinese medicine with modern drug discovery technologies to find novel drugs and functional foods. Curr Drug Discov Technol 2010; 7: 2–12
12 *Newman DJ, Cragg GM.* Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod 2012; 75: 311–335
13 *Berkov S, Mutafova B, Christen P.* Molecular biodiversity and recent analytical developments: a marriage of convenience. Biotechnol Adv 2014; 32: 1102–1110
14 *Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D.* Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int 2014; 2014: e134023
15 *Eid S, Zalewski A, Smieško M, Ernst B, Vedani A.* A Molecular-Modeling Toolbox Aimed at Bridging the Gap between Medicinal Chemistry and Computational Sciences. Int J Mol Sci 2013; 14: 684–700
16 *Medina-Franco JL, Martínez-Mayorga K, Peppard TL, Del Rio A.* Chemoinformatic analysis of GRAS (Generally Recognized as Safe) flavor chemicals and natural products. PLoS One 2012; 7: e50798
17 *Blondeau S, Do QT, Scior T, Bernard P, Morin-Allory L.* Reverse pharmacognosy: another way to harness the generosity of nature. Curr Pharm Des 2010; 16: 1682–1696
18 *Do QT, Driscoll M, Slitt A, Seeram N, Peppard TL, Bernard P.* Reverse pharmacognosy: a tool to accelerate the discovery of new bioactive food ingredients. In: Martinez-Mayorga K, Medina-Franco JL, editors. Food informatics: applications of chemical information to food chemistry. Heidelberg: Springer; 2014: 111–130
19 *Martinez-Mayorga K, Peppard TL, López-Vallejo F, Yongye AB, Medina-Franco JL.* Systematic mining of Generally Recognized as Safe (GRAS) flavor chemicals for bioactive compounds. J Agric Food Chem 2013; 61: 7507–7514
20 *Audouze K, Tromelin A, Le Bon AM, Belloir C, Petersen RK, Kristiansen K, Brunak S, Taboureau O.* Identification of odorant-receptor interactions by global mapping of the human odorome. PLoS One 2014; 9: e93037
21 *Bernard P, Scior T, Didier B, Hibert M, Berthon JY.* Ethnopharmacology and bioinformatic combination for leads discovery: application to phospholipase A(2) inhibitors. Phytochemistry 2001; 58: 865–874
22 *Rollinger JM, Haupt S, Stuppner H, Langer T.* Combining ethnopharmacology and virtual screening for lead structure discovery: COX-inhibitors as application example. J Chem Inf Comput Sci 2004; 44: 480–488
23 *Kerns EH.* High throughput physicochemical profiling for drug discovery. J Pharm Sci 2001; 90: 1838–1858
24 *Avdeef A, Testa B.* Physicochemical profiling in drug research: a brief survey of the state-of-the-art of experimental techniques. Cell Mol Life Sci 2002; 59: 1681–1689
25 *Sinko JS.* Drug selection in early drug development: screening for acceptable pharmacokinetic properties using combined *in vitro* and computational approaches. Curr Opin Drug Discov Devel 1999; 2: 42–48
26 *Cartmell J, Krstajic D, Leahy DE.* Competitive Workflow: novel software architecture for automating drug design. Curr Opin Drug Discov Devel 2007; 10: 347–352
27 *Scior T, Bernard P, Medina-Franco JL, Maggiora GM.* Large compound databases for structure-activity relationships studies in drug discovery. Mini Rev Med Chem 2007; 7: 851–860
28 *Scior T, Medina-Franco JL, Do QT, Martínez-Mayorga K, Yunes Rojas JA, Bernard P.* How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem 2009; 16: 4297–4313
29 *Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, Bender A, Glen RC.* Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. J Chem Inf Model 2012; 52: 617–648
30 *Fang Y.* Label-free drug discovery. Front Pharmacol 2014; 5: 1–8
31 *Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA.* Shifting from the single to the multitarget paradigm in drug discovery. Drug Discov Today 2013; 18: 495–501
32 *Lobell M, Hendrix M, Hinzen B, Keldenich J, Meier H, Schmeck C, Schohe-Loop R, Wunberg T, Hillisch A. In silico* ADMET traffic lights as a tool for the prioritization of HTS hits. ChemMedChem 2006; 1: 1229–1236
33 *Lipinski CA, Lombardo F, Dominy BW, Feeney PJ.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001; 46: 3–26
34 *Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD.* Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 2002; 45: 2615–2623
35 *Blake JF.* Examination of the computed molecular properties of compounds selected for clinical development. Biotechniques 2003; 34: S16–S20
36 *Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD.* A comparison of physiochemical property profiles of development and marketed oral drugs. J Med Chem 2003; 46: 1250–1256
37 *Palm K, Stenberg P, Luthman K, Artursson P.* Polar molecular surface properties predict the intestinal absorption of drugs in humans. Pharm Res 1997; 14: 568–571
38 *Clark DE.* Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. J Pharm Sci 1999; 88: 807–814
39 *Ritchie TJ, Macdonald SJ.* The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? Drug Discov Today 2009; 14: 1011–1020
40 *Di L, Kerns EH.* Profiling drug-like properties in discovery research. Curr Opin Chem Biol 2003; 7: 402–408
41 *Cruz-Monteagudo M, Cordeiro MN.* Chemoinformatics profiling of ionic liquids–uncovering structure-cytotoxicity relationships with network-like similarity graphs. Toxicol Sci 2014; 138: 191–204

42 *Pedretti A, Villa L, Vistoli G.* VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. J Mol Graph Model 2002; 21: 47–49
43 *Carrió P, Pinto M, Ecker G, Sanz F, Pastor M.* Applicability Domain ANalysis (ADAN): a robust method for assessing the reliability of drug property predictions. J Chem Inf Model 2014; 54: 1500–1511
44 *Norinder U, Carlsson L, Boyer S, Eklund M.* Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J Chem Inf Model 2014; 54: 1596–1603
45 *Moura-Barbosa AJ, Del Rio A.* Freely accessible databases of commercial compounds for high-throughput virtual screenings. Curr Top Med Chem 2012; 12: 866–877
46 *Blunt JW, Munro MHG, Laatsch H.* AntiMarin database. Christchurch: University of Canterbury. Göttingen: University of Göttingen; 2007
47 *Pereira F, Latino DA, Gaudêncio SP.* A chemoinformatics approach to the discovery of lead-like molecules from marine and microbial sources en route to antitumor and antibiotic drugs. Mar Drugs 2014; 12: 757–778
48 *Newman DJ, Cragg GM.* Natural products as sources of new drugs over the last 25 years. J Nat Prod 2007; 70: 461–477
49 *Newman DJ, Cragg GM, Snader KM.* Natural products as sources of new drugs over the period 1981–2002. J Nat Prod 2003; 66: 1022–1037
50 *DiMasi JA.* Pharmaceutical R&D performance by firm size: approval success rates and economic returns. Am J Ther 2014; 21: 26–34
51 *Bergström CA, Holm R, Jørgensen SA, Andersson SBA, Artursson P, Beato S, Borde A, Box K, Brewster M, Dressman J, Feng KI, Halbert G, Kostewicz E, McAllister M, Muenster U, Thinnes J, Taylor R, Mullertz A.* Early pharmaceutical profiling to predict oral drug absorption: current status and unmet needs. Eur J Pharm Sci 2014; 57: 173–199
52 *Keller TH, Pichota A, Yin Z.* A practical view of 'druggability'. Curr Opin Chem Biol 2006; 10: 357–361
53 *Kellenberger E, Hofmann A, Quinn RJ.* Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. Nat Prod Rep 2011; 28: 1483–1492
54 *Feher M, Schmidt JM.* Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. J Chem Inf Comput Sci 2003; 43: 218–227
55 *Ortholand JY, Ganesan A.* Natural products and combinatorial chemistry: back to the future. Curr Opin Chem Biol 2004; 8: 271–280
56 *Messer R, Fuhrer CA, Häner R.* Natural product-like libraries based on non-aromatic, polycyclic motifs. Curr Opin Chem Biol 2005; 9: 259–265
57 *Camp D, Davis RA, Campitelli M, Ebdon J, Quinn RJ.* Drug-like properties: guiding principles for the design of natural product libraries. J Nat Prod 2012; 75: 72–81
58 *Rishton GM.* Reactive compounds and *in vitro* false positives in HTS. Drug Discov Today 1997; 2: 382–384
59 *Hughes JP, Rees S, Kalindjian SB, Philpott KL.* Principles of early drug discovery. Br J Pharmacol 2011; 162: 1239–1249
60 *Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, Mbah JA, Mbaze LM, Sippl W, Efange SM.* AfroDb: a select highly potent and diverse natural product library from African medicinal plants. PLoS One 2013; 8: e78085
61 *Ntie-Kang F, Amoa Onguéné P, Fotso GW, Andrae-Marobela K, Bezabih M, Ndom JC, Ngadjui BT, Ogundaini AO, Abegaz BM, Meva'a LM.* Virtualizing the p-ANAPL library: a step towards drug discovery from African medicinal plants. PLoS One 2014; 9: e90655
62 *Buckingham J.* Dictionary of natural products. London: Chapman & Hall; 1994
63 *Duke J.* Dr. Duke's phytochemical and ethnobotanical databases. ARS, USDA. Available at http://www.ars-grin.gov/duke (Oct 2009). Accessed September 26, 2013
64 *Nakamura Y, Afendi FM, Parvin AK, Ono N, Tanaka K, Hirai Morita A, Sato T, Sugiura T, Altaf-Ul-Amin M, Kanaya S.* KNApSAcK Metabolite Activity Database for retrieving the relationships between metabolites and biological activities. Plant Cell Physiol 2014; 55: e7
65 *Loub WD, Farnsworth NR, Soejarto DD, Quinn ML.* NAPRALERT: computer handling of natural product research data. J Chem Inf Comput Sci 1985; 25: 99–103
66 Plant for a future database. Available at http://www.pfaf.org Oct 2009. Accessed September 26, 2013
67 *Dunkel M, Fullbeck M, Neumann S, Preissner R.* SuperNatural: a searchable database of available natural compounds. Nucleic Acids Res 2006; 34: 678–683
68 *Chen X, Zhou H, Liu YB, Wang JF, Li H, Ung CY, Han LY, Cao ZW, Chen YZ.* Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. Br J Pharmacol 2006; 149: 1092–1103
69 *Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X.* Use of natural products as chemical library for drug discovery and network pharmacology. PLoS One 2013; 8: e62839
70 *Bender A, Glen RC.* Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2004; 2: 3204–3218
71 *Stumpfe D, Hu Y, Dimova D, Bajorath J.* Recent progress in understanding activity cliffs and their utility in medicinal chemistry. J Med Chem 2014; 57: 18–28
72 *Medina-Franco JL.* Scanning structure-activity relationships with structure-activity similarity and related maps: from consensus activity cliffs to selectivity switches. J Chem Inf Model 2012; 52: 2485–2493
73 *Yue R, Shan L, Yang X, Zhang W.* Approaches to target profiling of natural products. Curr Med Chem 2012; 19: 3841–3855
74 *Méndez-Lucio O, Tran J, Medina-Franco JL, Meurice N, Muller M.* Towards drug repurposing in epigenetics: olsalazine as a novel hypomethylating compound active in a cellular context. ChemMedChem 2014; 9: 560–565
75 *Villoutreix BO, Eudes R, Miteva MA.* Structure-based virtual ligand screening: recent success stories. Comb Chem High Throughput Screen 2009; 12: 1000–1016
76 *Ripphausen P, Nisius B, Bajorath J.* State-of-the-art in ligand-based virtual screening. Drug Discov Today 2011; 16: 372–376
77 *López-Vallejo F, Caulfield T, Martínez-Mayorga K, Giulianotti MA, Nefzi A, Houghten RA, Medina-Franco JL.* Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. Comb Chem High Throughput Screen 2011; 14: 475–487
78 *Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK.* Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 2012; 52: 867–881
79 *Harvey AL, Clark RL, Mackay SP, Johnston BF.* Current strategies for drug discovery through natural products. Expert Opin Drug Discov 2010; 5: 559–568
80 *Medina-Franco JL.* Advances in computational approaches for drug discovery based on natural products. Rev Latinoam Quim 2013; 41: 95–110
81 *Schuster D, Wolber G.* Identification of bioactive natural products by pharmacophore-based virtual screening. Curr Pharm Des 2010; 16: 1666–1681
82 *Ehrman TM, Barlow DJ, Hylands PJ.* Phytochemical informatics and virtual screening of herbs used in Chinese medicine. Curr Pharm Des 2010; 16: 1785–1798
83 *Shen JH, Xu XY, Cheng F, Liu H, Luo XM, Shen JK, Chen KX, Zhao WM, Shen X, Jiang HL.* Virtual screening on natural products for discovering active compounds and target information. Curr Med Chem 2003; 10: 2327–2342
84 *Ma DL, Chan DSH, Leung CH.* Molecular docking for virtual screening of natural product databases. Chem Sci 2011; 2: 1656–1665
85 *Geldenhuys WJ, Bishayee A, Darvesh AS, Carroll RT.* Natural products of dietary origin as lead compounds in virtual screening and drug design. Curr Pharm Biotechnol 2012; 13: 117–124
86 *Lemmen C, Lengauer T.* Computational methods for the structural alignment of molecules. J Comput Aided Mol Des 2000; 14: 215–232
87 *Yongye AB, Bender A, Martinez-Mayorga K.* Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble. J Comput Aided Mol Des 2010; 24: 675–686
88 *Medina-Franco JL, Maggiora GM.* Molecular similarity analysis. In: Bajorath J, editor. Chemoinformatics for drug discovery. New York: John Wiley & Sons, Inc.; 2014: 343–399
89 *Ebalunode JO, Zheng WF.* Unconventional 2D shape similarity method affords comparable enrichment as a 3D shape method in virtual screening experiments. J Chem Inf Model 2009; 49: 1313–1320
90 *Hu GP, Kuang GL, Xiao W, Li WH, Liu GX, Tang Y.* Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. J Chem Inf Model 2012; 52: 1103–1113
91 *Kalászi A, Szisz D, Imre G, Polgár T.* Screen3D: a novel fully flexible high-throughput shape-similarity search method. J Chem Inf Model 2014; 54: 1036–1049
92 *Zhang Q, Muegge I.* Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. J Med Chem 2006; 49: 1536–1548

93 *Mendez-Lucio O, Perez-Villanueva J, Castillo R, Medina-Franco JL.* Identifying activity cliff generators of PPAR ligands using SAS maps. Mol Inf 2012; 31: 837–846
94 *Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F.* Activity cliffs in drug discovery: Dr. Jekyll or Mr. Hyde? Drug Discov Today 2014; 19: 1069–1080
95 *Yongye AB, Waddell J, Medina-Franco JL.* Molecular scaffold analysis of natural products databases in the public domain. Chem Biol Drug Des 2012; 80: 717–724
96 *Medina-Franco JL, Martinez-Mayorga K, Meurice N.* Balancing novelty with confined chemical space in modern drug discovery. Expert Opin Drug Discov 2014; 9: 151–165
97 *López-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL.* Expanding the medicinally relevant chemical space with compound libraries. Drug Discov Today 2012; 17: 718–726
98 *Bender A.* How similar are those molecules after all? Use two descriptors and you will have three different answers. Expert Opin Drug Discov 2010; 5: 1141–1151
99 *Todeschini R, Consonni V.* Molecular descriptors for chemoinformatics, 2nd edition. New York: Wiley-VCH; 2009
100 *Willett P.* Combination of similarity rankings using data fusion. J Chem Inf Model 2013; 53: 1–10
101 *Holliday JD, Kanoulas E, Malim N, Willett P.* Multiple search methods for similarity-based virtual screening: analysis of search overlap and precision. J Cheminform 2011; 3: 29
102 *Yongye A, Byler K, Santos R, Martínez-Mayorga K, Maggiora GM, Medina-Franco JL.* Consensus models of activity landscapes with multiple chemical, conformer and property representations. J Chem Inf Model 2011; 51: 1259–1270
103 *Guasch L, Sala E, Castell-Auvi A, Cedo L, Liedl KR, Wolber G, Muehlbacher M, Mulero M, Pinent M, Ardevol A, Valls C, Pujadas G, Garcia-Vallve S.* Identification of PPARgamma partial agonists of natural origin (I): development of a virtual screening procedure and *in vitro* validation. PLoS One 2012; 7: e50816
104 *Medina-Franco JL, Yoo J.* Docking of a novel DNA methyltransferase inhibitor identified from high-throughput screening: insights to unveil inhibitors in chemical databases. Mol Div 2013; 17: 337–344
105 *Martinez-Mayorga K, Peppard TL, Ramirez-Hernandez AI, Terrazas-Alvarez DE, Medina-Franco JL.* Chemoinformatics analysis and structural similarity studies of food-related databases. In: Martinez-Mayorga K, Medina-Franco JL, editors. FoodInformatics: applications of chemical information to food chemistry. Heidelberg: Springer; 2014: 97–110
106 *Jensen K, Panagiotou G, Kouskoumvekaki I.* Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. PLoS Comput Biol 2014; 10: e1003432
107 *Pochetti G, Godio C, Mitro N, Caruso D, Galmozzi A, Scurati S, Loiodice F, Fracchiolla G, Tortorella P, Laghezza A, Lavecchia A, Novellino E, Mazza F, Crestani M.* Insights into the mechanism of partial agonism: crystal structures of the peroxisome proliferator-activated receptor gamma ligand-binding domain in the complex with two enantiomeric ligands. J Biol Chem 2007; 282: 17314–17324
108 *Al-Najjar BO, Wahab HA, Tengku Muhammad TS, Shu-Chien AC, Ahmad Noruddin NA, Taha MO.* Discovery of new nanomolar peroxisome proliferator-activated receptor γ activators via elaborate ligand-based modeling. Eur J Med Chem 2011; 46: 2513–2529
109 *Feng Y, Campitelli M, Davis RA, Quinn RJ.* Chemoinformatic analysis as a tool for prioritization of trypanocidal marine derived lead compounds. Mar Drugs 2014; 12: 1169–1184
110 *Rosén J, Lövgren A, Kogej T, Muresan S, Gottfries J, Backlund A.* ChemGPS-NP(Web): chemical space navigation online. J Comput Aided Mol Des 2009; 23: 253–259
111 *Rognan D.* Towards the next generation of computational chemogenomics tools. Mol Inf 2013; 32: 1029–1034
112 *Medina-Franco JL, Aguayo-Ortiz R.* Progress in the visualization and mining of chemical and target spaces. Mol Inf 2013; 32: 942–953
113 *Bajorath J.* A perspective on computational chemogenomics. Mol Inf 2013; 32: 1025–1028
114 *Kjærulff SK, Wich L, Kringelum J, Jacobsen UP, Kouskoumvekaki I, Audouze K, Lund O, Brunak S, Oprea TI, Taboureau O.* ChemProt-2.0: visual navigation in a disease chemical biology database. Nucleic Acids Res 2013; 4: D464–D469
115 *Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, Koehler AN, Schreiber SL.* Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. Proc Natl Acad Sci U S A 2010; 107: 18787–18792
116 *Yongye AB, Medina-Franco JL.* Data mining of protein-binding profiling data identifies structural modifications that distinguish selective and promiscuous compounds. J Chem Inf Model 2012; 52: 2454–2461
117 *Dimova D, Hu Y, Bajorath J.* Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. J Med Chem 2012; 55: 10220–10228
118 *Yongye AB, Medina-Franco JL.* Toward an efficient approach to identify molecular scaffolds possessing selective or promiscuous compounds. Chem Biol Drug Des 2013; 82: 367–375
119 *Dossetter AG, Griffen EJ, Leach AG.* Matched molecular pair analysis in drug discovery. Drug Discov Today 2013; 18: 724–731
120 http://pdb.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html. Accessed June 22, 2014
121 *Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE.* A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982; 161: 269–288
122 *Lengauer T, Rarey M.* Computational methods for biomolecular docking. Curr Opin Struct Biol 1996; 6: 402–406
123 *Goodsell DS, Morris GM, Olson AJ.* Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 1996; 9: 1–5
124 *Rarey M, Kramer B, Lengauer T, Klebe G.* A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996; 261: 470–489
125 *Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004; 47: 1739–1749
126 *Jones G, Willett P, Glen RC, Leach AR, Taylor R.* Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997; 267: 727–748
127 *Jain AN.* Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 2003; 46: 499–511
128 *Li Y, Han L, Liu Z, Wang R.* Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. J Chem Inf Model 2014; 54: 1717–1736
129 *Bissantz C, Folkers G, Rognan D.* Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000; 43: 4759–4767
130 *Bar-Haim S, Aharon A, Ben-Moshe T, Marantz Y, Senderowitz H.* SeleX-CS: a new consensus scoring algorithm for hit discovery and lead optimization. J Chem Inf Model 2009; 49: 623–633
131 *Teramoto R, Fukunishi H.* Structure-based virtual screening with supervised consensus scoring: evaluation of pose prediction and enrichment factors. J Chem Inf Model 2008; 48: 747–754
132 *Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB.* Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002; 20: 281–295
133 *Charifson PS, Corkery JJ, Murcko MA, Walters WP.* Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999; 42: 5100–5109
134 *Feher M.* Consensus scoring for protein-ligand interactions. Drug Discov Today 2006; 11: 421–428
135 *Plewczynski D, Łaźniewski M, von Grotthuss M, Rychlewski L, Ginalski K.* VoteDock: consensus docking method for prediction of protein-ligand interactions. J Comput Chem 2011; 32: 568–581
136 *Houston DR, Walkinshaw MD.* Consensus docking: improving the reliability of docking in a virtual screening context. J Chem Inf Model 2013; 53: 384–390
137 *Paul N, Rognan D.* ConsDock: A new program for the consensus analysis of protein-ligand interactions. Proteins 2002; 47: 521–533
138 *Scior T, Verhoff M, Gutierrez-Aztatzi I, Ammon HP, Laufer S, Werz O.* Interference of boswellic acids with the ligand binding domain of the glucocorticoid receptor. J Chem Inf Model 2014; 54: 978–986
139 *Peeters M, Li Q, Elands R, van Westen GJ, Lenselink EB, Müller CE, IJzerman AP.* Domains for activation and inactivation in G protein-coupled receptors–a mutational analysis of constitutive activity of the adenosine A2B receptor. Biochem Pharmacol 2014; 92: 348–357
140 *Thiyagarajan V, Lin SH, Chia YC, Weng CF.* A novel inhibitor, 16-hydroxy-cleroda-3,13-dien-16,15-olide, blocks the autophosphoryla-

tion site of focal adhesion kinase (Y397) by molecular docking. Biochim Biophys Acta 2013; 1830: 4091–4101
141 *Hussain A, Melville JL, Hirst JD.* Molecular docking and QSAR of aplyronine A and analogues: potent inhibitors of actin. J Comput Aided Mol Des 2010; 24: 1–15
142 *Koukoulitsa C, Zervou M, Demetzos C, Mavromoustakos T.* Comparative docking studies of labdane-type diterpenes with forskolin at the active site of adenylyl cyclase. Bioorg Med Chem 2008; 16: 8237–8243
143 http://ec.europa.eu/research/health/infectious-diseases/antimicrobial-drug-resistance/projectsfp7_en.html. Accessed June 22, 2014
144 http://www.nabativi.org/. Accessed September 26, 2013
145 *Srinivas N, Jetter P, Ueberbacher BJ, Werneburg M, Zerbe K, Steinmann J, Van der Meijden B, Bernardini F, Lederer A, Dias RL, Misson PE, Henze H, Zumbrunn J, Gombert FO, Obrecht D, Hunziker P, Schauer S, Ziegler U, Käch A, Eberl L, Riedel K, DeMarco SJ, Robinson JA.* Peptidomimetic antibiotics target outer-membrane biogenesis in *Pseudomonas aeruginosa*. Science 2010; 327: 1010–1013
146 *Chan FY, Sun N, Neves MA, Lam PC, Chung WH, Wong LK, Chow HY, Ma DL, Chan PH, Leung YC, Chan TH, Abagyan R, Wong KY.* Identification of a new class of FtsZ inhibitors by structure-based design and *in vitro* screening. J Chem Inf Model 2013; 53: 2131–2140
147 *Huang KJ, Lin SH, Lin MR, Ku H, Szkaradek N, Marona H, Hsu A, Shiuan D.* Xanthone derivatives could be potential antibiotics: virtual screening for the inhibitors of enzyme I of bacterial phosphoenolpyruvate-dependent phosphotransferase system. J Antibiot (Tokyo) 2013; 66: 453–458
148 *Harris SM, McFeeters H, Ogungbe IV, Cruz-Vera LR, Setzer WN, Jackes BR, McFeeters RL.* Peptidyl-tRNA hydrolase screening combined with molecular docking reveals the antibiotic potential of *Syzygium johnsonii* bark extract. Nat Prod Commun 2011; 6: 1421–1424
149 *Kirchmair J, Distinto S, Liedl KR, Markt P, Rollinger JM, Schuster D, Spitzer GM, Wolber G.* Development of anti-viral agents using molecular modeling and virtual screening techniques. Infect Disord Drug Targets 2011; 11: 64–93
150 *Srinivasan S, Sarada DV.* Antifungal activity of phenyl derivative of pyranocoumarin from *Psoralea corylifolia* L. seeds by inhibition of acetylation activity of trichothecene 3-o-acetyltransferase (Tri101). J Biomed Biotechnol 2012; 2012: e310850
151 *Singh C, Atri N.* Chemo-informatic design of antibiotic geldenamycin analogs to target stress proteins HSP90 of pathogenic protozoan parasites. Bioinformation 2013; 9: 329–333
152 *Chen YZ, Zhi DG.* Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 2001; 43: 217–226
153 *Vigers GP, Rizzi JP.* Multiple active site corrections for docking and virtual screening. J Med Chem 2004; 47: 80–89
154 *Paul N, Kellenberger E, Bret G, Müller P, Rognan D.* Recovering the true targets of specific ligands by virtual screening of the protein data bank. Proteins 2004; 54: 671–680
155 *Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE.* The Protein Data Bank. Nucleic Acids Res 2000; 28: 235–242
156 *Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D.* BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004; 32: D431–D433
157 *Chen X, Ji ZL, Chen YZ.* TTD: therapeutic target database. Nucleic Acids Res 2002; 30: 412–415
158 *Gao Z, Li H, Zhang H, Liu X, Kang L, Luo X, Zhu W, Chen K, Wang X, Jiang H.* PDTD: a web-accessible protein database for drug target identification. BMC Bioinf 2008; 9: e104
159 *Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D.* sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. J Chem Inf Model 2006; 46: 717–727
160 *Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008; 36: 901–906
161 *Do QT, Bernard P.* Pharmacognosy and reverse pharmacognosy: a new concept for accelerating natural drug discovery. IDrugs 2004; 7: 1017–1027
162 *Do QT, Renimel I, Andre P, Lugnier C, Muller CD, Bernard P.* Reverse pharmacognosy: application of selnergy, a new tool for lead discovery. The example of epsilon-viniferin. Curr Drug Discov Technol 2005; 2: 161–167
163 *Do QT, Lamy C, Renimel I, Sauvan N, André P, Himbert F, Morin-Allory L, Bernard P.* Reverse pharmacognosy: identifying biological properties for plants by means of their molecule constituents: application to meranzin. Planta Med 2007; 73: 1235–1240
164 *Rollinger JM.* Accessing target information by virtual parallel screening–the impact on natural product research. Phytochem Lett 2009; 2: 53–58
165 *Chen SJ.* A potential target of Tanshinone IIA for acute promyelocytic leukemia revealed by inverse docking and drug repurposing. Asian Pac J Cancer Prev 2014; 15: 4301–4305
166 *Chen YZ, Ung CY.* Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. J Mol Graph Model 2001; 20: 199–218
167 *Schomburg KT, Bietz S, Briem H, Henzler AM, Urbaczek S, Rarey M.* Facing the challenges of structure-based target prediction by inverse virtual screening. J Chem Inf Model 2014; 54: 1676–1686
168 *Grinter SZ, Liang Y, Huang SY, Hyder SM, Zou X.* An inverse docking approach for identifying new potential anti-cancer targets. J Mol Graph Model 2011; 29: 795–799
169 *Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, Wang X, Jiang H.* TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 2006; 34: W219–W224
170 *Hui-fang L, Qing S, Jian Z, Wei F.* Evaluation of various inverse docking schemes in multiple targets identification. J Mol Graph Model 2010; 29: 326–330
171 *Nicola G, Liu T, Gilson MK.* Public domain databases for medicinal chemistry. J Med Chem 2012; 55: 6987–7002
172 *Wale N, Karypis G.* Target fishing for chemical compounds using target-ligand activity data and ranking based methods. J Chem Inf Model 2009; 49: 2190–2201
173 *Pandini A, Fraccalvieri D, Bonati L.* Artificial neural networks for efficient clustering of conformational ensembles and their potential for medicinal chemistry. Curr Top Med Chem 2013; 13: 642–651
174 *Rognan D.* Structure-based approaches to target fishing and ligand profiling. Mol Inf 2010; 29: 176–187
175 *Wermuth G, Ganellin CR, Lindberg P, Mitscher LA.* Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl Chem 1998; 70: 1129–1143
176 *Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M.* Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. J Med Chem 2006; 49: 6802–6810
177 *Todeschini R, Lasagni M, Marengo E.* New molecular descriptors for 2D and 3D structures. Theory. J Chemometr 1994; 8: 263–272
178 *Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A.* MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. J Comput Aided Mol Des 1997; 11: 79–92
179 *Medina-Franco JL, Martínez-Mayorga K, Bender A, Marín RM, Giulianotti MA, Pinilla C, Houghten RA.* Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. J Chem Inf Model 2009; 49: 477–491
180 *Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM.* Evaluation and optimization of virtual screening workflows with DEKOIS 2.0 – a public library of challenging docking benchmark sets. J Chem Inf Model 2013; 53: 1447–1462
181 *Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW.* How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model 2009; 49: 108–119
182 *Scior T, Bernard P, Medina-Franco JL, Maggiora GM.* Large compound databases for structure-activity relationships studies in drug discovery. Mini Rev Med Chem 2007; 7: 851–860