frontiers in Genetics

REVIEW
published: 30 April 2019
doi: 10.3389/fgene.2019.00368

Informatics and Computational Methods in Natural Product Drug Discovery: A Review and Perspectives

Joseph D. Romano [1,2,3,4] *and Nicholas P. Tatonetti* [1,2,3,4]*

[1] Department of Biomedical Informatics, Columbia University, New York, NY, United States, [2] Department of Systems Biology, Columbia University, New York, NY, United States, [3] Department of Medicine, Columbia University, New York, NY, United States, [4] Data Science Institute, Columbia University, New York, NY, United States

The discovery of new pharmaceutical drugs is one of the preeminent tasks—scientifically, economically, and socially—in biomedical research. Advances in informatics and computational biology have increased productivity at many stages of the drug discovery pipeline. Nevertheless, drug discovery has slowed, largely due to the reliance on small molecules as the primary source of novel hypotheses. Natural products (such as plant metabolites, animal toxins, and immunological components) comprise a vast and diverse source of bioactive compounds, some of which are supported by thousands of years of traditional medicine, and are largely disjoint from the set of small molecules used commonly for discovery. However, natural products possess unique characteristics that distinguish them from traditional small molecule drug candidates, requiring new methods and approaches for assessing their therapeutic potential. In this review, we investigate a number of state-of-the-art techniques in bioinformatics, cheminformatics, and knowledge engineering for data-driven drug discovery from natural products. We focus on methods that aim to bridge the gap between traditional small-molecule drug candidates and different classes of natural products. We also explore the current informatics knowledge gaps and other barriers that need to be overcome to fully leverage these compounds for drug discovery. Finally, we conclude with a "road map" of research priorities that seeks to realize this goal.

Keywords: drug discovery, methods, cheminformatics, bioinformatics, ontologies, translation, natural products

OPEN ACCESS

Edited by:
Dana C. Crawford, Case Western Reserve University, United States

Reviewed by:
Brett K. Beaulieu-Jones, Harvard Medical School, United States
Harry Hochheiser, University of Pittsburgh, United States
Ellen L. Palmer, Case Western Reserve University, United States

***Correspondence:**
Nicholas P. Tatonetti
npt2105@cumc.columbia.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *12 December 2018*
Accepted: *05 April 2019*
Published: *30 April 2019*

Citation:
Romano JD and Tatonetti NP (2019) Informatics and Computational Methods in Natural Product Drug Discovery: A Review and Perspectives. Front. Genet. 10:368. doi: 10.3389/fgene.2019.00368

1. INTRODUCTION

Drug discovery is the process by which new pharmaceutical drugs are identified, and along with drug development (validating, testing, and marketing a new drug), it comprises one of the most substantial activities in pharmaceutical science. A 2018 analysis showed that roughly 20% of the US National Institutes of Health (NIH) budget for the years 2010–2016 funded the discovery and development of 210 new molecular entities (Cleary et al., 2018). Since the advent of modern medical science, most systematic drug discovery has focused on small molecule candidates—for example, over 86% of the drugs (both approved and experimental) in the DrugBank database are comprised of small molecules (Wishart et al., 2017). This is due to many reasons, including relative ease of synthesis, generally high chemical stability, and more straightforward characterization of

reactivity (Drews, 2000). The pervasiveness of small molecules in drug discovery is even reflected in Lipinski's "rule of five," which defines a set of common best-practice guidelines for filtering potential orally-active drug candidates: "Good" compounds should have a molecular mass <500, no more than five hydrogen bond donors, and no more than 10 hydrogen bond acceptors, among other principles (Lipinski, 2004).

In recent decades, the ubiquity of computers and computational methods in science has extended to drug discovery (Sliwoski et al., 2014). Cheminformatics, for example, is the application of computer science to understanding and characterizing molecular attributes and chemical behavior of specific compounds. These methods have generated massive libraries of small molecules to screen against specific therapeutic processes (Blaney and Martin, 1997). Once candidates are identified, other cheminformatics methods can be used to generate libraries of compounds structurally and chemically similar to the identified "hits," in order to optimize stability, toxicity, and kinetics. Complementarily, bioinformatics techniques can be used to discover how candidate drugs cause therapeutic activity within the human body, which can include predicting interactions between drugs and proteins, analysis of impact on biological pathways and functions, and elucidating genomic variants that can alter drug response (Drews, 2000).

Despite these technological advances in drug discovery, the approval of new therapeutic drugs has slowed considerably in recent years. For example, between 1996 and 2007, the number of new molecular entities approved by the US FDA has fallen from 53 to 17 per year—the same rate as over 50 years ago (FitzGerald, 2008; Munos, 2009). This seems to be due to many factors, including the following:

1. The "lowest hanging fruits" in terms of small molecule drug candidates have been extensively investigated, and computational challenges hinder extension of traditional methods to more complex structures. Researchers refer to "rediscovering the sweet spot" in the discovery process (Brown and Superti-Furga, 2003), and have devoted a great amount of effort to producing new, targeted screening libraries that leverage anticipated characteristics of lead compounds (Welsch et al., 2010; Cheng et al., 2012).
2. Many remaining diseases of top clinical priority have highly complex etiologies, and are accordingly difficult to associate with potential drug targets (Ramsay et al., 2018).
3. Model organisms may not provide adequate templates for testing treatments of more complex diseases, due to inter-species variations that are crucial to therapeutic action (Hunter, 2008; Ehret et al., 2017).

A natural way to address the first two challenges is to focus on new classes of potential drugs outside of small molecules. Natural products (NPs) may serve this need by returning to the sources of therapeutic compounds that have treated illness for thousands of years (Dias et al., 2012). Although rigorous pharmaceutical science is young in comparison to the historical use of NP drugs, many cutting-edge advances have emerged with the promise of "modernizing" this field (Harvey, 2008). Along with a renewed interest for NP drugs within the biomedical research community, this has already resulted in substantial developments in the pharmaceutical industry—a comprehensive enumeration by Newman and Cragg shows that 41% (646/1562) of all new drug approvals between 1981 and 2014 are NPs or derived from NPs (Newman and Cragg, 2016). Several recent reviews provide excellent summaries of NP drugs and the broad spectrum of techniques that have been used both for their identification and characterization (Katz and Baltz, 2016; Rodrigues et al., 2016), particularly from the perspective of bench research techniques and state-of-the-art developments in biotechnology. Considering the aforementioned trends in new computational methods and advances in classical informatics for translational applications of these methods, these reviews can be complemented by a dedicated discussion restricted to *in silico* approaches for NP drug discovery.

Another trend in drug discovery enabled by informatics and computational methods is an increasing shift toward a data driven drug discovery (Tatonetti et al., 2012; Lusher et al., 2014). Traditionally, drug discovery has been performed as follows: basic scientists first find a target structure in the human body related to a disease or illness, followed by screening for "lead" compounds that show affinity for the target. Subsequently, the list of candidates is narrowed down (using some of the methods described in this review) to find the most promising leads, which then go through the development process to assess safety and efficacy in model organisms and, eventually, humans. A detailed description of these steps can be found in other reviews (Hughes et al., 2011). Failure at any stage in this workflow can—and usually does—necessitate starting over from the beginning, contributing to the estimated cost of 2.6 billion USD to bring a new drug to market (Avorn, 2015). Data-driven drug discovery turns the process on its head, by using data mining on large data repositories of candidate compounds and disease knowledge to generate novel therapeutic hypotheses systematically rather than hoping for a single therapeutic hypothesis to deliver actionable results. Aside from avoiding systematic biases present in the hypothesis-driven model, this additionally helps to improve the return rate on subsequent manual experimentation and validation of lead compounds, ultimately lowering costs and increasing productivity (Jorgensen, 2004). Data-driven drug discovery leverages new data types that were previously inaccessible, and relies heavily upon computers and informatics techniques to produce increasingly accurate results (Butte and Ito, 2012).

In this review, we first discuss various major classes of natural products based both on source organism and their biological functions. In addition, we provide examples of specific members of those classes with demonstrated therapeutic potential. We then explore several major disciplines based upon informatics and computational methods—cheminformatics, bioinformatics, and semantic (or "knowledge-based") informatics—and their associated methods that can be used specifically for NP drug discovery. These methods are summarized graphically in **Figure 1**. Finally, we conclude with a recap of the major gaps currently facing the field of computational NP drug discovery, and suggest actions for the future that could help to resolve these problems.

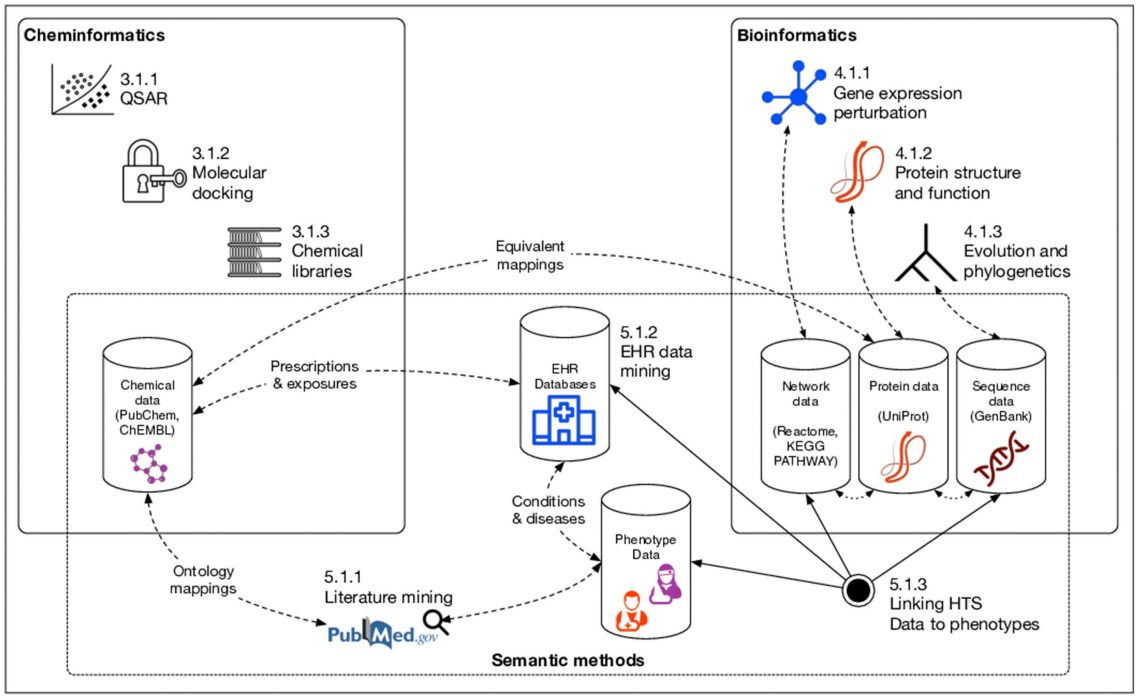

FIGURE 1 | Informatics methods for natural product drug discovery covered in this review. Numbers preceding methods correspond to section/subsection numbers in the manuscript describing the method. Dashed lines indicate inferred links between various data resources.

2. CLASSES OF THERAPEUTIC NATURAL PRODUCTS

There is no definitive consensus on what groups of substances comprise "natural products," with some authors restricting them to small molecule secondary metabolites (Nature Publishing Group, 2007), and others more broadly stating that an NP is any chemical substance produced by a living organism (National Center for Complementary and Integrative Health, 2017). For the purpose of this review, we adopt the latter of these two definitions: that natural products include all classes of chemical substances that are produced or recruited by living organisms, and have the ability to be isolated and reused by humans. This definition includes an incredibly diverse range of compound types; therefore, it is crucial to understand the different subgroups of NPs, along with their characteristics. These classes of NPs frequently overlap and have vaguely defined boundaries, but they are nevertheless useful for understanding the methods that can be applied to them.

2.1. Phytochemicals

Phytochemicals—chemicals synthesized by plants—encompass a broad range of NPs, including members of many of the other classes described later in this section. Phytochemicals can be toxic, they can provide important dietary nutrients (such as amino acids, antioxidants, and dietary fiber), or they can be inert in humans. For most research purposes, however, phytochemicals are limited to primary and secondary metabolites in plants, which can be generally divided into phenolic acids, stilbenes, and flavonoids (which, themselves, can be further subdivided into more specific subclasses), all of which are small molecules (rather than macromolecules, which tend to be prevalent in many of the other classes we discuss) (Harborne, 1999). These chemicals have been the source of many traditional and modern medicines, famous examples of which include the analgesic acetylsalicylic acid (aspirin), the heart medication digoxin, and the chemotherapy drug paclitaxel (Molyneux et al., 2007).

2.2. Fungal Metabolites

Fungal metabolites serve a relatively similar role to plant metabolites, so much so that they share some of the same subclasses (perhaps most notably the flavonoid compounds). Like plant metabolites, fungal metabolites can treat a wide variety of diseases and conditions, but they are perhaps most famous as a source of many successful antibiotics. Other areas of successful application include antimalarials (antiamoebin), immunosuppressants (ciclosporin), statins (mevastatin, lovastatin), and more (Thomford et al., 2018).

2.3. Toxins

Toxins are substances that can potentially harm or kill. They include poisons and venoms, and are (by definition) produced by living organisms. Poisons are toxins that cause harmful effects when swallowed, inhaled, or absorbed through the surface of the skin, while venoms are toxins that cause harm when actively injected via a sting or a bite.

Poisons are produced by members of many major clades of organisms, including plants, fungi, bacteria, and most groups of animals. Natural poisons are usually used for defensive purposes, although some species have adapted them for more complex roles (Klaassen and Watkins, 1996). They can include members of all classes of molecules, and although many tend to consist of relatively small molecular structures, macromolecules, such as proteins, large carbohydrates, and lipids can be poisonous as well. NP poisons include many chemotherapy drugs, particularly when their toxic effects act more selectively on cancer cells than healthy cells. Some examples include paclitaxel (from *Taxus brevifolia*) and vinblastine (from *Catharanthus roseus*) (Thomford et al., 2018).

Venoms are complex mixtures of chemicals produced by animals for either defensive or offensive purposes (or, sometimes, both in the same species). An individual species' venom can include hundreds of unique chemical compounds, many of which are proteins that act on specific molecular targets. Venoms are highly evolutionarily optimized to fit organisms' biological niches (Daltry et al., 1996), but due to interspecies homology, the effects of individual venom components have led to numerous therapeutic applications, including FDA-approved treatments for hypertension, diabetes, neuropathic pain, and more (Lewis and Garcia, 2003). Like poisons, venoms have also demonstrated potent anti-cancer effects, and their high target specificity has made them of particular interest for applications of precision medicine, particularly for rare or aggressive cancer types (Romano and Tatonetti, 2016; Yang et al., 2018).

2.4. Antibodies

Components of the immune system—particularly antibodies—have long been attractive for drug discovery and design. Their primary function is recognition and inactivation of pathogens, including bacteria and viruses, but biotechnologists have repurposed them for many "unintended" uses, including the targeted treatment of various diseases. One approach, known as immunotherapy, involves the design and application of monoclonal antibodies that bind specifically to certain cells or proteins related to the disease of interest. Naturally, these are often autoimmune diseases, such as rheumatoid arthritis (Seo et al., 2004) and allergies (Jutel et al., 1995), but they have also been applied to diverse diseases, such as viral infections (Letvin and Walker, 2003) and multiple sclerosis (Hohlfeld, 1997). Recently, substantial attention has been given to immunotherapy treatments for cancer, exemplified by the 2018 Nobel Prize in Medicine being awarded for research in this area (Ishida et al., 1992; Leach et al., 1996; Rosenberg et al., 2004). The second approach involves using antibodies as delivery agents for therapeutic compounds, which is also being explored extensively for cancer, due to its capacity to mitigate off-target effects (Awwad and Angkawinitwong, 2018). Interestingly, this delivery method has attracted specific attention for the delivery of chemotherapeutics that are, themselves, NPs (Mann, 2002).

It should be noted that—in spite of the substantial accomplishments described above—antibodies have failed to deliver on several therapeutic applications that originally held promise, often for characteristics that are inherent to antibodies in general. One example involves the treatment of Alzheimer Disease (AD) using monoclonal antibodies. Antibody-based treatments for AD performed strongly in mouse models (Bard et al., 2000) and in early-phase clinical trials (Hock et al., 2003), but in phase-2 trials and beyond, they have failed to deliver (Tayeb et al., 2013). Multiple theories have been posed, but the two leading hypotheses for failure have been that (1) antibodies are limited in their ability to cross the blood-brain barrier, and (2) certain degenerative diseases require early treatment for antibodies to be effective, far before patients begin to show symptoms (Sperling et al., 2011). Other failures in antibody therapy are related to the activity of antibodies themselves—drugs like theralizumab (designed to treat leukemia and rheumatoid arthritis) failed in human trials due to inciting a life-threatening "cytokine storm" in all healthy volunteers (Eastwood et al., 2010). Nonetheless, much research on new antibody therapies is being conducted to treat the same diseases associated with these early failures (Sevigny et al., 2016).

2.5. NPs With Limited Therapeutic Potential

The classes of NPs described above cover substantial breadth. However, to provide a more complete image of drug discovery in terms of NPs, it is also important to consider classes with only limited—or at least presently unknown—therapeutic potential. For the purposes of this review, we focus on whether a compound is reactive enough in living systems to potentially perturb that system. If it is, then there exists an opportunity to exploit the perturbations for potentially therapeutic outcomes. The largest group of NPs that falls short in this regard is those with purely structural purposes, including materials like wood, biopolymers, and excretions like spider silk, which suggests that the drug discovery methods discussed in subsequent sections of this review are unlikely to generate many new lead compounds.

Nonetheless, biology is rife with exceptions to every rule, and even these groups of NPs have occasionally yielded compounds with therapeutic use. Wood creosote has been used for centuries as a treatment for diarrhea, and is currently marketed in Japan under the trade name Seirogan (Hiramoto et al., 2012). Biopolymers have not resulted in drugs themselves, but have been used many times to successfully deliver drugs within living systems (Nitta and Numata, 2013). Even spider silk has shown potential in drug delivery (Spiess et al., 2010), and has been bioengineered to have antibiotic properties (Harvey et al., 2017). For this reason, we hesitate to say that any class of NPs has no therapeutic potential. In a practical sense, these observations are most useful in a cost-benefit analysis scenario, when it is necessary to balance research budget with scientific risk, highlighted by Dickson and Gagnon as one of the major factors influencing the total output of the pharmaceutical industry (Dickson and Gagnon, 2004).

3. CHEMINFORMATICS METHODS

Cheminformatics methods can generally be classified according to the types of characteristics they exploit: either direct measures of chemical activity (e.g., chemical constants, reactive groups, or ADME measurements), or indirect measures (e.g., structural

motifs, compound class membership, or other higher-order observations). These techniques can be further subdivided; for example, structural comparisons can be applied either before or after promising chemical activity is known (which we refer to here as prospective and retrospective structure mining, respectively). Prospective structure mining is conducted in a supervised manner, where known chemical activity of well-characterized compounds is compared to the structures of query compounds to predict the therapeutic potential of the queries. Retrospective structure mining, on the other hand, is more analogous to unsupervised learning techniques, where other screening techniques first identify a compound of interest (referred to as a "hit"), and then seek to expand the number of candidate compounds by searching for structures that are similar to the hit compound.

3.1. Cheminformatics and Natural Products

Many traditional cheminformatics methods are challenging to adapt to certain classes of NPs, particularly when the NPs consist of large chemical structures (like venoms, antibodies, or other protein-based NP drug candidates). For example, generating combinatorial libraries of large polypeptides is currently intractable, due to the massive search space. However, additional characteristics that are unique to these classes of NPs enable either simplifying assumptions to be made or the invention of entirely new approaches for predicting bioactivity (Huang et al., 2016). Here, we divide cheminformatics into 3 major categories of methods that have been used to success with NPs, providing discussion of the caveats that must be considered for NPs in particular.

3.1.1. Natural Product QSAR Analysis

Quantitative Structure Activity Relationship (QSAR) analysis is a widely used—if often ambiguously defined—technique in cheminformatics for predicting a response variable given a set of structural, chemical, and or physical input variables (known as *molecular descriptors*). Generally, the goal is to learn a function of the form

$$\hat{y} = f(\mathbf{x}) + \epsilon$$

where $\mathbf{x} = (x_1, \ldots, x_N)$ is the vector of N input variables, $\hat{y}$ is the estimated response (continuous in the case of regression, and integer-valued in the case of classification), and ϵ is an error term. f can be any appropriate model; common choices include logistic regression, support vector machines, random forest, artificial neural networks, and others. Recently, deep learning has shown to be particularly effective for predicting a wide variety of responses, including solubility, probe-likeness, and others (Korotcov et al., 2017). A number of free and commercial software implementations of QSAR are available for a variety of use cases (Benfenati et al., 2011; Tosco and Balle, 2011), and approaches for adapting generic statistical and machine learning models for QSAR are readily available (Lavecchia, 2015).

QSAR has been applied fairly widely to different classes of NPs, where specific classes tend to dictate the chosen molecular descriptors. Typical choices for non-NP applications include symbolic (1- or 2-D) descriptors, 3-D spatial organization, higher-order (e.g., time-dependent or ligand-bound) conformational characteristics (Polanski, 2009), experimental measurements (partition coefficient, polarizability, refractivity, etc.), and many others. For a detailed review of these and similar descriptors, see Cherkasov et al. (2014). Additional characteristics that can be used for small-molecule NPs include categorical ("one-hot") variables indicating class membership (e.g., alkaloid, terpenoid), species of origin (or more general taxonomic clades), and other biological features. Macromolecular NPs are substantially more restricted in terms of the types of descriptors that can be used effectively. Generally, 3-D conformational descriptors and binding data function best for these NPs, and yield good results (Mladenović et al., 2017; Dhiman et al., 2018). QSAR has performed adequately for predicting binding affinity of antibodies to proteins—Mandrika et al. describe a model consisting of 26 physicochemical descriptors (covering hydrophobicity, polarity, electronegativity, etc.) at each amino acid position in a library consisting of single chain monoclonal antibodies (Mandrika et al., 2007). While this model has not yet been applied to NP drug discovery, it seems to be a feasible way forward.

3.1.2. Molecular Docking and Dynamics

QSAR is a useful statistical method for predicting potentially therapeutic interactions, but it is often desirable to directly model the chemical or physical interaction that is being investigated. Molecular docking is an approach that seeks to predict if and how two compounds (usually a target and a ligand) physically interact. This is usually performed in two steps: (1) searching for potential conformational fits, and (2) scoring those fits. Molecular dynamics is a particular simulation technique that can be applied to docking, and is popular in drug development. From a high level, molecular dynamics performs a computational simulation of the atoms and molecules (often including solvents) present in a putative reaction, and allows the molecules to interact for a period of time. The technical details and algorithms for docking and dynamics are well-summarized elsewhere (Karplus and McCammon, 2002; Pagadala et al., 2017)—we will instead focus on broad caveats, issues, and innovations in applying these to NPs.

The class of NP compound tends to dictate the role (target vs. ligand) that the compound plays in docking simulations. Typically, small molecule NPs and relatively short polypeptides (e.g., peptide toxins and venom components) act as ligands, while larger proteins and protein complexes act as targets (although exceptions are common). This distinction is important, especially when the goal is screening many candidate compounds: usually, the target is held fixed, while the ligand can be drawn from libraries of many compounds. Therefore, it is computationally feasible to perform docking of many small molecule compounds when a specific molecular target is already known (Khan et al., 2009; Lee et al., 2011; Ma et al., 2011). Conversely, if a macromolecular NP is suspected of interacting with endogenous small-molecule metabolites (e.g., in human cancer cells), docking simulations can be used to mine *which* metabolites could bind to the NP (Pithayanukul et al., 2009). If both a target and a ligand are already predicted by other means (e.g.,

QSAR or other methods described in this review), docking is commonly used as a secondary validation method. In spite of their large molecular weight, antibodies are relatively easy to screen in large numbers via docking, due to their specific structural and binding constraints that can substantially reduce computational complexity of simulations (Walls and Sternberg, 1992; Abagyan and Totrov, 2001).

Molecular dynamics is an important technique for characterizing physical interactions of putative drugs with their targets, but due to computational challenges it cannot be used with current technologies in a data-driven manner to screen very large numbers of NPs against similarly large numbers of potential targets simultaneously (Salmaso and Moro, 2018). However, it has proven incredibly valuable in uncovering specific therapeutic mechanisms of NPs (venom proteins in particular). An early and influential example of this came in 1995, when Albrand et al. combined molecular dynamics with NMR to explain how Toxin FS2 (from Black Mamba venom) blocks L-type calcium channels, causing potent cardiotoxic effects (Albrand et al., 1995). Additionally, there are noteworthy success stories that have emerged from screening relatively small NP databases against specific drug targets: The compound ellagic acid—which has shown both antiproliferative and antioxidant properties—was identified by Moro et al. by screening a proprietary database of 2,000 NPs against the oncoprotein casein kinase 2 (Cozza et al., 2006). Similarly, Fu et al. identified Jadomycin B—another molecule with anticancer effects—by screening 15,000 microbial small molecule metabolites against the oncoprotein Aurora-B kinase (Fu et al., 2008). These examples illustrate the feasibility of molecular dynamic studies for discovering new therapeutic NPs, and suggest that overcoming associated computational challenges will enable their widespread application in diverse and data-driven contexts.

3.1.3. Computational Mutagenesis and Library Construction

One of the most common techniques for identifying drug candidates is to generate massive libraries of compounds that can be screened in parallel, with the understanding that only a very small fraction will result in "hits" (potential therapeutic activity). There are many ways such libraries are generated, many of which fall under the umbrella term of combinatorial chemistry (i.e., enumerating chemical structures using combinatorics) (Terrett et al., 1995). NPs provide some advantages over traditional (non-NP) classes of candidate compounds, namely that such "libraries" already exist in nature. General purpose online databases of chemical compounds (such as PubChem and ChEMBL) (Li et al., 2010; Gaulton et al., 2016) contain many NPs that are annotated by compound class, while other, more specific databases (such as ArachnoServer, VenomKB, and the Dictionary of Marine Natural Products) provide even more granular annotations for aggregating NP libraries with various characteristics of interest (Pineda et al., 2017; Romano et al., 2018).

Computational mutagenesis is a related class of techniques that has shown efficacy in certain classes of NPs. This method involves specifying a template (e.g., a certain antibody with putative therapeutic activity that requires optimization), and then sequentially mutating locations in the template's structure to generate a library of candidate compounds. These libraries can then be screened *in silico* (e.g., using molecular docking simulations as described in section 3.1.2) to find structures that can be engineered in the lab. Antibodies, in particular, are particularly well-suited to computational mutagenesis, by modifying amino acids in binding regions (Sivasubramanian et al., 2009; Wollacott et al., 2019). The feasibility of mutagenesis techniques in the context of NP drug discovery was demonstrated by Chen et al., who generated a library of analogs of the 7-residue NP peptide HUN-7293 to optimize its inhibitory effects on cell-adhesion (Chen et al., 2002).

It should be noted that one of the advantages of working with NPs is the potential of avoiding library screening entirely, under the assumption that nature has optimized it for biological activity. This point is expanded on in section 4.1.3.

4. BIOINFORMATICS METHODS

Bioinformatics methods for drug discovery include anything related to the *biological* function of potential drug candidates, including sequence-based characteristics, interactions with body structures (metabolites, proteins, cells, tissues, etc.), pathway perturbations, and toxicity, among others. Multi-omics and high-throughput sequencing are also major areas within bioinformatics. Most subdisciplines of bioinformatics can be applied in some way to the drug discovery process (Wishart et al., 2017; Thomford et al., 2018).

4.1. Bioinformatics and Natural Products

In the case of NPs, researchers are able to make use of an entire range of techniques related to the organisms that produce the compounds. In particular, phylogenetics and evolution provide many routes for various drug discovery activities. Closely related organisms often produce similar proteins and metabolites, so when one natural compound with promising activity has an unsuitable therapeutic index for human use, libraries of similar compounds can be easily constructed by searching in organisms within the same genus. However, these techniques must be applied with caution: members of some groups of natural compounds (such as venom proteins) are heavily optimized to fit a very particular biological niche, so even members of the same species may have entirely unique metabolic profiles with respect to compounds of interest. One prominent example of this was found in the rattlesnake species *Crotalus oreganus helleri*, where members of the species living on different sides of a mountain range produced entirely separate venom profiles (Sunagar et al., 2014).

4.1.1. Gene Expression Perturbation

The rise of multi-omics approaches to uncovering mechanisms of disease has led to multitudes of ways to assess the effect that putative drugs have on cells. In particular, gene expression perturbation—quantified using RNA-sequencing and transcriptomics—has led to a number of innovative breakthroughs in drug discovery for diseases associated with gene disregulation, including cancers and various other diseases

with complex genetic etiologies (Sirota et al., 2011; Subramanian et al., 2017). Along with environmental exposures, structural abnormalities, and other influencing factors, these diseases often can be attributed in part to abnormalities in gene expression, including the systems-level effects of expression perturbation in the larger context of cell signaling and metabolic networks (Nica and Dermitzakis, 2008; Cookson et al., 2009). More accurately, differential expression can be treated as a phenotypic signal that arises from underlying disease etiology. Accordingly, drugs and drug candidates that effectively invert such deleterious effects are potential therapies for these diseases.

This technique is particularly well-adapted for use in NP drug discovery, as vast numbers of compounds from all classes of NPs are specifically optimized to have roles in cell signaling or metabolic networks, and are already known to be relatively biologically stable (Lewis and Garcia, 2003). Compounds used in Traditional Chinese Medicine (TCM) have been particularly well-utilized in this area. In a 2014 study, researchers uncovered likely mechanisms by which the TCM compound berberine exhibits anti-cancer activity, using publicly-available expression data for berberine-perturbed human cells taken from the Connectivity Map (CMap) project (Lee et al., 2014). Another important recent example by Lv et al. provides differential gene expression profiles in response to 102 different TCM compounds, presented as a framework from which to base future systematic research on the activities of TCMs (Lv et al., 2017).

A separate but related approach involves analysis of differential expression in the organisms producing the NPs (rather than the organisms that NPs act upon). An investigation by Amos et al. discovered previously unknown NPs—as well as putative mechanisms describing their functionality—by comparing transcriptome profiles of different bacterial species in the genus *Salinispora* (Amos et al., 2017), underscoring the diversity of emerging multi-omics techniques that can be employed within NP drug discovery.

4.1.2. Modeling Protein Structure and Function

Although the size and complexity of proteins is often prohibitive to structure-based analyses designed for small molecules, other drug discovery approaches leverage the unique characteristics of proteins and other macromolecules to perform discovery in ways that are otherwise impossible. Since many classes of NPs are comprised of proteins, these techniques can often be adapted to NP drug discovery with relative ease.

Some methods use supervised machine learning algorithms trained on protein structures (and motifs) with known activity to predict activity in new, uncharacterized proteins—this is essentially traditional QSAR designed to work on proteins. The FEATURE framework (Halperin et al., 2008) does this using 3-dimensional spatial orientation of atoms to predict activity at numerous "microenvironments" within a larger macromolecule, and is therefore generalizable to diverse proteins with conserved functional activity. Other research teams have designed similar frameworks based on other machine learning models, including deep learning models like convolutional neural networks (Torng and Altman, 2017; Thomas et al., 2018). For further details on learning protein function from structure, we refer the reader to Pérez et al. (2018).

Still other protein functional modeling approaches rely on input variables that behave like "abstractions" of raw molecular characteristics, including amino acid or DNA structure (along with sequence alignment algorithms) (Vyas et al., 2012), ontology annotations (see section 5 for more details) (Mutowo et al., 2016), and biomarker response (Frank and Hargreaves, 2003).

4.1.3. Using Evolution to Discover Drug Candidates

The fact that NPs are derived from living organisms implies that they either serve a specific purpose in the context of that organism, or they are a byproduct of an important process (Stone and Williams, 1992). Therefore, we can use evolution and taxonomy as tools for both discovering new compounds and their effects, as well as for generating libraries of similar natural products (Maplestone et al., 1992).

The simplest—and most common—use of phylogenetics in natural product drug discovery revolves around the axiom that *closely related species produce similar NPs*. This can be used to predict the structures of NPs, given structures for similar NPs in related species are already known (Ziemert and Jensen, 2012). Following a pattern akin to QSAR modeling (described in section 3.1.1), phylogenetics can also be repurposed to predict other characteristics of closely related NPs, including molecule classes, toxicity, stability, and others, where instead of using molecular descriptors as observed features of the NP, you instead use evolutionary characteristics to build a predictive model. A noteworthy example is given by Malhotra et al., who used discriminant function analysis (DFA) to classify and predict functions of over 250 phospholipase A_2 proteins from viperid snakes, where aligned amino acid sequences alone were used to construct the input features for the DFA model (Malhotra et al., 2013).

Other uses of evolution in drug discovery employ phylogenomics to discover associations across more distantly related species (e.g., between humans and microbes). This includes efforts to catalog the entire breadth of various classes of natural products to create comprehensive NP class libraries (see section 3.1.3 for more details) (Rønsted et al., 2012). In 2016, Rudolf et al. showed that comparative genomics in diverse microbial species could identify 87 distinct gene clusters across 78 bacterial species corresponding to a class of putative NP anticancer drugs known as enediynes (Rudolf et al., 2016). By finding instances of NP coevolution in distantly related species, studies have uncovered compounds that play keystone roles in metabolic processes, leading to therapeutic solutions in analogous processes in humans. A noteworthy and sophisticated example is shown in the CSMNA method (Zhang et al., 2016), which is based on the hypothesis that similarities between human and plant metabolic networks can be used to guide phytochemical drug discovery. The authors validate their drug discovery algorithm by showing that similarities between the plant Halliwell-Asada (HA) cycle and the human Nrf2-ARE pathway underlie antioxidant activity of HA cycle molecules on proteins in the Nrf2-ARE pathway.

Some caveats need to be kept in mind when using evolutionary approaches. Certain classes of NPs are under evolutionary pressures that complicate phylogenetic analysis. Venom proteins, in particular, can be highly divergent even among species within the same genus (Calvete et al., 2014), a phenomenon attributed to the high metabolic cost of venom production, and the highly targeted nature of many venom proteins to specific prey species.

5. SEMANTIC (KNOWLEDGE-BASED) METHODS

Cheminformatics and bioinformatics are both major subdivisions of biomedical informatics, and comprise two of the primary disciplines involved in translational research and drug discovery. We now turn our focus to a set of methods that emerged from semiotics, linguistics, and library science, but have been adapted to serve broad functions in computer science and artificial intelligence—known as knowledge-based or semantic (i.e., relating to human-interpretable meaning) methods. In general, these are methods involving the application of various knowledge representations, such as ontologies and structured terminologies. Some activities within this group include rule-based natural language processing, certain types of clinical data mining, knowledge extraction, semantic data normalization, and others. Especially in the context of drug discovery, knowledge-based methods are frequently applied in coordination with bioinformatics and/or cheminformatics methods, and serve as one of the main approaches to combining and unifying findings and intermediate results spread across separate research activities.

Perhaps the most well-utilized resource in knowledge-based approaches to drug discovery is the Gene Ontology (Ashburner et al., 2000), which classifies conceptual biological entities into 3 groups: molecular functions, cellular components, and biological processes (each of which is important in various stages of the drug discovery process). Researchers have created multitudes of data resources to assist in drug discovery, and many of these are mapped to the Gene Ontology to assist with *in silico* aggregation and preliminary validation of putative hypotheses. Some of these linked resources include DrugBank (Wishart et al., 2017), UniProtKB/Swiss-Prot (and associated annotation programs like ToxProt) (Jungo et al., 2012), and ChEMBL (Gaulton et al., 2016), all of which catalog compounds that may confer some therapeutic effect.

Still other tools have been created to map unstructured data relevant to drug discovery (such as journal article abstracts in PubMed) to more structured representations. MetaMap, SemRep, and Semantic Medline from the National Library of Medicine, as well as the NCBO Annotator from the National Center for Biomedical Ontology identify ontology and terminology terms within free text (usually pulled from journal articles) at various levels of abstraction. These tools have been used to successfully perform ontological inference across multiple levels of evidence for many discovery tasks, including drug discovery. For further details, we refer the reader to the original paper describing Swanson's Fish Oil-Raynaud's Syndrome hypothesis (Swanson, 1986), which explains how structured knowledge and graph algorithms can be used to discover informative associations fragmented across otherwise unrelated publications (Cameron et al., 2013).

Other levels of knowledge representation (e.g., not formally controlled at the concept level) also have important roles in drug discovery; tools like OMIM can be used to map newly discovered drug-gene associations to diseases that are modulated by that gene or set of genes. For comprehensive listings of the various ontologies, knowledge representations, and similar tools with proven roles in drug discovery, we refer the reader to a number of existing reviews (Gardner, 2005; Vazquez-Naya et al., 2010; Thomford et al., 2018).

5.1. Semantic Methods and Natural Products

While the number of ontologies and similar resources relevant to drug discovery are vast, advanced applications of these resources are relatively scarce. This trend is even more striking in regards to NP drug discovery. As of now, most therapeutic associations between NPs and disease are discovered serendipitously rather than through systematic, rigorous applications, although earlier sections of this review describe notable exceptions to this trend. In light of the fact that advanced use of semantic methods is rare in NP drug discovery, we will additionally consider applications of ontologies and terminologies used for drug discovery that *could* be applied to NPs, based on current knowledge.

5.1.1. Literature Mining

Literature mining—the process of performing text mining on scientific literature databases—is one of the most common usages of semantic biomedical knowledge resources. The MEDLINE/PubMed database contains over 26 million biomedical text citations, many thousands of which contain knowledge related to NPs, and possibly describing characteristics of those NPs that provide direct or indirect evidence of therapeutic activity. There are generally two ways to automatically extract such knowledge from biomedical publications: (1) Using existing ontology/terminology annotations, or (2) using natural language processing (NLP) techniques that discover such annotations.

Medical Subject Headings (MeSH) are one terminology resource designed to structure the content of PubMed articles, and are applied manually by expert annotators at the US National Library of Medicine (NLM) to new articles shortly after indexing in PubMed (Lipscomb, 2000). MeSH terms cover a diverse range of biomedical concepts, arranged in a hierarchical fashion, and cover various classes of NPs. MeSH can be used to aggregate PubMed articles describing certain types of NPs, and can be refined using additional terms (e.g., "`Drug Discovery`") or qualifiers (e.g., "`/therapeutic use`"). MeSH terms can link journal entities to structured external databases by either using cross-mappings [including via the NLM's Unified Medical Language System (UMLS)] or annotations in external databases directly to MeSH terms (Ruau et al., 2011). MeSH terms have been used to summarize

components of plant genomes (Beissinger and Morota, 2017), demonstrating potential paths forward in discovering novel NPs (rather than using the terms to gather knowledge about known NPs).

A limited number of databases provide access to curated sets of articles describing NPs. VenomKB provides articles annotated to venom components as well as literature predictions describing the putative therapeutic effects of those components and mappings to other external databases (Romano and Tatonetti, 2015). Similarly, NPASS presents chemical characteristics of a broader range of NPs and provides references to PubMed entries describing manually-curated biological activity measurements in a range of organisms (including humans) (Zeng et al., 2017). Other databases, including MarinLit and NAPRALERT, provide commercial and paid access to curated NP literature data.

5.1.2. Electronic Health Record Mining

Similarly to literature mining, we can apply knowledge retrieval techniques to observational data sources. As far as drug discovery is concerned, observational data provides a method for assessing the effects compounds have on humans in the absence of rigorously controlled clinical research studies. This style of data analysis offers several major advantages over clinical trials, including avoidance of exposing new patients to potentially harmful treatments, and mitigating certain types of bias associated with eligibility and patient selection. Observational data can often produce larger cohorts than clinical trials. Various sources of observational data can be utilized for drug discovery, but here we will focus on electronic health records (EHRs), due to their prevalence and proven utility for many translational research tasks. Although privacy concerns, data fragmentation, and standardization have traditionally hampered access to EHR data—particularly for research teams without clinical expertise or affiliation with a large academic medical center—rapidly growing efforts, such as Observational Health Data Sciences and Informatics (OHDSI) (Hripcsak et al., 2015) and the Electronic Medical Records and Genomics (eMERGE) network (McCarty et al., 2011) are breaking these barriers in ways that will increase access to data covering the breadth of the translational spectrum.

EHR data are complex, multimodal, and subject to many unique biases and ethical/legal constraints (Weiskopf and Weng, 2013). In addition to free text (recorded by health care providers), a number of structured data types are also present (including claims data, medication orders, laboratory measurements, patient demographics, and others). As of now, no major applications of EHR data mining to NP drug discovery have been reported, but a number of related areas provide hints as to its feasibility. A review by Yao et al. highlights 3 specific ways that EHRs can aid drug discovery: (1) Finding relationships between diseases for the purposes of drug repurposing, (2) evaluating the usage patterns and safety of drugs and/or drug candidates, and (3) discovering phenotype–genotype associations that can lead to the discovery of new drug targets for specific diseases (Yao et al., 2011). Relevant caveats of each of these can be discussed from the perspective of NP drug discovery, including specific advantages and disadvantages that NPs provide when compared to non-NP drugs and drug candidates.

Drug repositioning involves taking an existing drug and using it to treat a different disease than what it is currently intended for Ashburn and Thor (2004). EHRs have been used for a number of drug repositioning approaches. The most common repositioning strategy involves discovering similarities between diseases, and then using those similarities to imply new treatments. This is based on the assumption that diseases with similar etiologies will produce similar signals in the EHR, and that similar etiologies may imply similar treatments. An important example by Rzhetsky et al. showed unexpected similarity between bipolar disorder and breast cancer (Rzhetsky et al., 2007). Recently, it has been demonstrated that the breast cancer drug tamoxifen may be useful for treating the symptoms of bipolar disorder (Kulkarni et al., 2006).

EHR data can also be used to assess the safety of drugs (or putative drugs), by determining whether exposure to the drug increases risk of adverse effects (Schuemie et al., 2012; Tatonetti et al., 2012). This is easiest for approved drugs that have coded representations in the EHR software (e.g., those with ATC codes or similar—experimental and unapproved drugs generally do not have a structured representation in EHR databases), but natural language processing can identify experimental and putative drugs with reasonable efficacy (Björne et al., 2013). This suggests that NP drug-candidate safety surveillance could be performed on free-text notes in the EHR, especially when treated as environmental exposures rather than physician-prescribed interventions. The feasibility of this approach was demonstrated by Zhang et al., who showed that herbal and natural supplements (which are usually considered NPs) could be identified in medication lists using natural language processing, and quantified the gap between structured drug representations and these compounds (Zhang et al., 2015). Two of the main gaps in need of resolution to realize this goal include specifying a standardized nomenclature for NPs (Dewick, 2002), and identifying where (geographically) hospital patients may be exposed to the NPs being investigated.

Discovering new drug targets is not strictly the same thing as drug discovery, but it does provide an essential starting point for identifying new drug leads. Recent decades have seen a steady decline in the discovery of new targets, and previous reviews on the topic have called for new and innovative strategies to address this issue (Lindsay, 2005; Spedding, 2006). Using EHR data and clinical biobanks to conduct Genome Wide Association Studies (GWASs) and Phenome Wide Association Studies (PheWASs) are touted as solutions (Yao et al., 2011), by providing associative links between diseases and specific genetic loci, which can then be used as targets for new precision drug therapies (McCarty and Wilke, 2010; Wilke et al., 2011). NPs, in particular, come into play when considering their unique abilities to target certain genes and gene products that are poorly targeted by small molecules. Both monoclonal antibodies and protein-based therapeutics are known for their ability to target individual cell types, especially useful in cancers with specific genetic signatures (Adams and Weiner, 2005; Cox et al., 2016). GWAS and PheWAS are relatively new compared to the drug discovery and development

timeline, but we will likely see many NP drugs emerging from clinical trials that used EHR- and biobank-enabled analyses for target discovery in the coming decades (Thomford et al., 2018).

5.1.3. Linking HTS Data to Putative Disease Treatments

Until now, we have discussed ways that ontologies and terminologies can be used to retrieve and structure knowledge, but another important role semantic techniques play in biomedicine is integrating disparate data sources in ways that otherwise require massive amounts of manual interpretation and annotation to apply at scale. This is important for many reasons, including experimental validation, increasing statistical power and inferential capacity, and even to discover new knowledge entirely. A particular application that has experienced rapid growth and major methodological advancements in drug discovery is linking new types of high-throughput sequencing (HTS) data to clinically-meaningful associations. Previously mentioned techniques, such as gene expression perturbation (section 4.1.1) yield results consisting of signals that have biological meaning, but no explicit connection to clinical phenotypes. Important early examples of data-driven drug discovery from gene expression formed therapeutic associations between cimetidine and lung adenocarcinoma (Sirota et al., 2011), as well as topiramate and inflammatory bowel disease (Dudley et al., 2011), but these examples required manual curation of many phenotype-linked expression profiles from which discovery could be performed. Knowledge representations provide a method for making these connections automatically, when correctly leveraged.

Successful knowledge integration of this type requires links to be formed between (a.) sets of genes (or, more specifically, groups of probe sets) and metabolic pathways, as well as (b.) links between pathways and phenotypes. A number of well-established and richly annotated gene-pathway databases (including Reactome and KEGG) (Fabregat et al., 2015; Kanehisa et al., 2016) already exist, and are used widely by the biomedical research community. Resources linking pathways to phenotypes are considerably less prevalent (and less complete), due largely to a limitation of available, relevant data, but ongoing efforts in the translational bioinformatics community are changing this. Integrating differences in gene expression and phenotypic response at the cell- and tissue-level with pathway data has shown particular promise in this area (Hao and Tatonetti, 2016; Hao et al., 2018). A recent review by Oellrich et al. outlines emerging and established tools for computational phenotyping (Oellrich et al., 2015).

Similar studies are, however, nearly absent from the realm of NP drug discovery. The unique characteristics of different NP classes (especially those described earlier in this review) can facilitate the phenotyping process. Metabolomics data provides clues as to NPs' original functions in their source organisms, which can often be extended to their effects when applied to humans (Xie et al., 2008; Yan et al., 2015; Zhang et al., 2016). Phylogenomics can highlight similarities between the genetic epidemiologies of complex diseases in humans vs. model organisms, possibly suggesting species from which to mine compounds that can treat these diseases (Romano et al., 2015). Even the predator/prey adaptations of NP-producing species can suggest the biological function of NPs (de la Vega and Possani, 2005; Miller et al., 2016); the discovery that the cone snail *Conus geographus* hunts fish by releasing insulin into the surrounding water (resulting in rapid hypoglycemic shock in the prey) led to the identification of a powerful insulin-receptor-binding motif that has shown considerable promise for future treatments of diabetes (Menting et al., 2016). Some recent studies focusing on discovery from TCM data show promise: Cui et al., for example, created a TCM chemical structure database that they screened against acetylcholinesterase (ACE) inhibitors, both via docking simulations with the known structure of ACE, as well as similarity to existing ACE inhibitors retrieved from BindingDB (Cui et al., 2015). Conceivably, ontology resources could be used to adapt these methods into an automated approach for screening many drug classes with little to no manual curation.

Linking HTS data to disease phenotypes is only one application of semantic knowledge resources that could be a boon for NP drug discovery. There are many other conceivable uses for linking evidence between clinical datasets, drug terminologies, literature-mined associations, and organismal biodiversity data, any of which could lead to potentially valuable discoveries and improved evidence for unproven hypotheses.

6. GAPS AND OPPORTUNITIES

6.1. Comparing the Use of Informatics Disciplines in NP Drug Discovery

Computers have revolutionized the way medicine and biomedical research are conducted, and the same applies to drug discovery. In doing so, it is critical to consider all of the ways in which computers can assist the discovery process in order to maximize the return on research efforts. In terms of *natural product* drug discovery, this review reveals that while some branches of informatics are being utilized extensively, other methods have not been fully explored. By summarizing nine representative groups of informatics methods (see **Figure 1** and **Table 1**), we highlight these disparities and, by extension, areas of opportunity for future research.

Pharmacologists and the pharmaceutical industry have championed the use of advanced cheminformatics techniques in concert with cutting-edge biotechnology innovations. Although NP drug discovery has always been a hallmark activity in pharmacology, pharmaceutical researchers have only applied these cheminformatic techniques to NPs rather recently. Both QSAR (section 3.1.1) and docking simulations (section 3.1.2) are standard practice for studying the therapeutic potential and mechanisms of NPs. There is also a fair number of NP library studies (section 3.1.3) that have been used to success—especially when focused on antibodies (Hoogenboom, 2005)—leading to the discovery of drugs, such as adalimumab (Jespers et al., 1994), ecallantide (Markland et al., 1996), and others (Nixon et al., 2014). As computing power improves, it is likely that we will see similar attention be paid to more challenging NP classes, such as venom peptides and other macromolecular compounds.

TABLE 1 | Summary of popular computational drug discovery methods described in this review and their applicability to NP drug discovery, stratified by the major branches of informatics discussed in this review.

Informatics branch	Method	Use with NPs
Cheminformatics	QSAR analysis (section 3.1.1)	Multiple
	Molecular docking (section 3.1.2)	Multiple
	Computational library design (section 3.1.3)	Multiple
Bioinformatics	Gene expression perturbation (section 4.1.1)	Little to none
	Protein structure/function modeling (section 4.1.2)	Multiple
	Phylogenetic approaches (section 4.1.3)	Multiple
Semantic methods	Literature mining (section 5.1.1)	Limited
	EHR mining (section 5.1.2)	None
	Linking HTS data to effects (section 5.1.3)	Little to none

Bioinformatics demonstrates a similar trend, albeit somewhat earlier in its development (with regards to NP drug discovery) than cheminformatics. The bioinformatics methods covered in this review are intriguing in that each is a technique originally intended for uses other than drug discovery. Differential gene expression analysis (section 4.1.1) was originally used to explore differences between cell lines and disease states rather than the effects of drug perturbation, although the conceptual jump in applying expression analysis to drug discovery is arguably an obvious one. However, due to this technique's relatively recent emergence, few examples using NPs (as opposed to non-NP small molecule candidates) currently exist in the literature, none of which are truly data-driven (i.e., agnostic to both specific diseases and specific NP drug candidates). Nonetheless, analyses targeted toward specific diseases compared against the Connectivity Map dataset have resulted in two substantial discoveries based on plant metabolites: Celastrol as a treatment for acute myeloid leukemia (Hassane et al., 2008), and gedunin as a treatment for prostate cancer (Hieronymus et al., 2006). Therefore, the preliminary groundwork for truly data-driven drug discovery for NPs via perturbational differential expression analysis has already been established. For further examples of the successes of the Connectivity Map approach to data-driven drug discovery overall, we direct readers to a previous review by Musa et al. (2017). Phylogenetics (section 4.1.3)—one of the earlier uses for computers in biology—has become known for its diverse areas of application, including drug discovery. Since NPs come from organisms that can be studied in a phylogenetic context, bioinformaticians have realized just how valuable of a tool this can be for NP drug discovery, and a number of completed and ongoing research initiatives capitalize on this.

Semantic methods have been used much less frequently for drug discovery than the other branches of informatics, and even less so for NPs. Only a few sparse examples of literature mining applications (section 5.1.1) exist for NP drug discovery. A few studies show that ontologies and similar methods that link experimental evidence to HTS data and structured knowledge representations (section 5.1.3) could easily be adapted to perform preliminary validation for expensive and time-consuming manual experimentation to prove therapeutic activity in NPs, but the actual use of these methods for this purpose is also virtually non-existent. EHRs and other clinical data resources are in a similar situation—as far as we can tell, there are currently no published examples of clinical data mining (section 5.1.2) being used to discover therapeutic associations from NPs.

6.2. Data Needs for NP Drug Discovery

Throughout this review, we have touched upon computational and informatics methods with varying data needs, and have naturally mentioned several data resources that are dedicated to (or have strong relevance to) NP drug discovery. Just as certain discovery methods are enabled by characteristics specific to NPs, certain data types and dimensions are as well. This includes taxonomic/evolutionary data (Cordell, 2000; Larsen et al., 2005), primary (i.e., "intended") targets and functions of NPs in nature (Bernardoni et al., 2014), the crude composition of NPs (often leading to synergistic effects, analogous to drug combination therapies) (Borkow et al., 1993; Casewell et al., 2013), and others specific to particular classes of NPs. A more comprehensive description of NP databases is presented in a review by Xie et al. (2015), but here we will cover some of them in brief as they pertaining to specific data needs.

The diversity and complexity of data types relevant for NP drug discovery research poses challenges in storing, representing, and exchanging these data. An immediate consequence is that many NP databases are limited to a narrow range of closely related NPs, which results in data fragmentation for the sake of completeness (Williams, 1997). ConoServer (Kaas et al., 2011) and ArachnoServer (Pineda et al., 2017) are two NP databases with rich and highly descriptive data, but each only applies to toxins produced by a single clade of species. One partial solution to this problem is to form dedicated efforts within larger, more general purpose databases that are dedicated to improving the representation of NPs, which is the approach taken by the Tox-Prot manual annotation program within UniProtKB/Swiss-Prot (Jungo et al., 2012). However, this does not completely resolve the greater issue of being able to leverage all important data types that are unique to certain classes of NPs. One other advantage that larger database efforts have over smaller, specialized NP databases is the presence of APIs and other tools that enable computational access. Many of the specialized databases do offer the ability to download data in bulk, but these can be incomplete and out-of-date. Furthermore, APIs can assist in making databases interoperable—an integrated network of specialized and well-annotated databases that can exchange semantic knowledge solves the issue of adequately representing granular characteristics while providing many of the benefits of larger data repositories.

Fragmentation of NP databases has also led to issues in maintaining those databases in the event of funding inconsistencies and institutional career changes—an issue that is at least partially safeguarded against when data resources are maintained by larger teams with more robust

operating budgets. Three examples of now-defunct NP databases are the Traditional Chinese Medicine Systems Pharmacology (TCMSP) database (Ru et al., 2014), the Animal Toxin Database (ATDB) (He et al., 2007), and the SuperNatural database (Dunkel et al., 2006). Smaller NP databases can also suffer from issues like having unwieldy and non-descriptive URLs, such as that for the Tea Metabolome Database (found at http://pcsb.ahau.edu.cn:8080/TCDB/f) (Yue et al., 2014). Furthermore, if ownership of such a database changes, or if the principle investigator moves to a new institution, the URL would likely break, creating issues in finding the database when reading the manuscript that describes it—a phenomenon sometimes referred to as "link rot" (Markwell and Brooks, 2003).

Taking into account these and related issues, a wealth of opportunity is available for informatics researchers and data scientists to improve the quality, quantity, and interconnectedness of NP databases and knowledge representations. In the following section, we will reiterate these and other areas of importance for the near future, as elucidated over the course of this review.

6.3. A Road Map for the Future of Natural Product Drug Discovery

In spite of the disparities outlined above, renewed interest in bioontologies, semantic knowledge integration, and data-driven approaches to drug discovery suggests that this could be in the early stages of change. This review brings to light several concrete ways that the research industry could address existing issues and encourage the development of new innovations for NP drug discovery:

1. **Creating new ontology resources**: Structured semantic knowledge resources for NPs and NP drug discovery are scarce. Most databases are either overly general or overly specific, and therefore cannot capitalize on many of the unique characteristics demonstrated by entire classes of NPs. Resources with the appropriate ontological commitment are necessary to support the integration of the methods we have described—specifically, new standards compliant ontologies and tools for performing inference over (and between) these ontologies. To increase impact, these new ontologies should aim to span the translational divide, linking concepts that join fundamental biological characteristics of NPs to the clinically meaningful effects those NPs exert on the human body. Alternatively, the design of tools and frameworks that link more specialized ontologies (e.g., covering only taxonomy of NPs, or molecular targets of NPs) that together bridge this gap could be used to accomplish the same goal.
2. **Generating public HTS data for NPs**: Although the biomedical community is experiencing a deluge of multi-omic HTS data, the vast majority of non-human species are underrepresented or completely absent in public repositories. Unless more resources are devoted to publishing multi-omics data for species of interest to NP drug discovery, many of the discovery methods we have discussed will remain out-of-reach to most researchers.
3. **Utilizing clinical data**: New collaborative efforts such as OHDSI and eMERGE enable greater access to real clinical data that can be used for both discovery and evaluation of new drugs. As coverage of NPs improves in semantic knowledge resources, the ability to perform inference on NPs using observational data will improve as well.

AUTHOR CONTRIBUTIONS

JR and NT conceived of, wrote, and edited the content of this review.

FUNDING

This work was supported by a grant awarded by the National Institute for General Medical Sciences (R01 GM107145; PI: NT).

REFERENCES

Abagyan, R., and Totrov, M. (2001). High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* 5, 375–382. doi: 10.1016/S1367-5931(00)00217-9

Adams, G. P., and Weiner, L. M. (2005). Monoclonal antibody therapy of cancer. *Nat. Biotechnol.* 23:1147. doi: 10.1038/nbt1137

Albrand, J. P., Blackledge, M. J., Pascaud, F., Hollecker, M., and Marion, D. (1995). NMR and restrained molecular dynamics study of the three-dimensional solution structure of toxin fs2, a specific blocker of the l-type calcium channel, isolated from black mamba venom. *Biochemistry.* 34, 5923–5937. doi: 10.1021/bi00017a022

Amos, G. C., Awakawa, T., Tuttle, R. N., Letzel, A.-C., Kim, M. C., Kudo, Y., et al. (2017). Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc. Natl. Acad. Sci. U.S.A.* 114, E11121–E11130. doi: 10.1073/pnas.1714381115

Ashburn, T. T., and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3:673. doi: 10.1038/nrd1468

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25–29. doi: 10.1038/75556

Avorn, J. (2015). The $2.6 billion pill–methodologic and policy considerations. *N. Engl. J. Med.* 372, 1877–1879. doi: 10.1056/NEJMp1500848

Awwad, S., and Angkawinitwong, U. (2018). Overview of antibody drug delivery. *Pharmaceutics.* 10:83. doi: 10.3390/pharmaceutics10030083

Bard, F., Cannon, C., Barbour, R., Burke, R. L., Games, D., Grajeda, H., et al. (2000). Peripherally administered antibodies against amyloid β-peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nat. Med.* 6:916. doi: 10.1038/78682

Beissinger, T. M., and Morota, G. (2017). Medical subject heading (mesh) annotations illuminate maize genetics and evolution. *Plant Methods.* 13:8. doi: 10.1186/s13007-017-0159-5

Benfenati, E., Toropov, A. A., Toropova, A. P., Manganaro, A., and Gonella Diaza, R. (2011). Coral software: Qsar for anticancer agents. *Chem. Biol. Drug Des.* 77, 471–476. doi: 10.1111/j.1747-0285.2011.01117.x

Bernardoni, J. L., Sousa, L. F., Wermelinger, L. S., Lopes, A. S., Prezoto, B. C., Serrano, S. M., et al. (2014). Functional variability of snake venom metalloproteinases: adaptive advantages in targeting different prey and implications for human envenomation. *PLoS ONE.* 9:e109651. doi: 10.1371/journal.pone.0109651

Björne, J., Kaewphan, S., and Salakoski, T. (2013). "Uturku: drug named entity recognition and drug-drug interaction extraction using svm classification and

domain knowledge," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Atlanta, GA), Vol. 2, 651–659.

Blaney, J. M., and Martin, E. J. (1997). Computational approaches for combinatorial library design and molecular diversity analysis. *Curr. Opin. Chem. Biol.* 1, 54–59.

Borkow, G., Gutiérrez, J., and Ovadia, M. (1993). Isolation and characterization of synergistic hemorrhagins from the venom of the snake bothrops asper. *Toxicon.* 31, 1137–1150.

Brown, D., and Superti-Furga, G. (2003). Rediscovering the sweet spot in drug discovery. *Drug Discov. Today.* 8, 1067–1077. doi: 10.1016/S1359-6446(03)02902-7

Butte, A. J., and Ito, S. (2012). Translational bioinformatics: data-driven drug discovery and development. *Clin. Pharmacol. Ther.* 91, 949–952. doi: 10.1038/clpt.2012.55

Calvete, J. J., Sanz, L., Pla, D., Lomonte, B., and Gutiérrez, J. M. (2014). Omics meets biology: application to the design and preclinical assessment of antivenoms. *Toxins.* 6, 3388–3405. doi: 10.3390/toxins6123388

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., et al. (2013). A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *J. Biomed. Inform.* 46, 238–251. doi: 10.1016/j.jbi.2012.09.004

Casewell, N. R., Wüster, W., Vonk, F. J., Harrison, R. A., and Fry, B. G. (2013). Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol. Evol.* 28, 219–229. doi: 10.1016/j.tree.2012.10.020

Chen, Y., Bilban, M., Foster, C. A., and Boger, D. L. (2002). Solution-phase parallel synthesis of a pharmacophore library of hun-7293 analogues: a general chemical mutagenesis approach to defining structure-function properties of naturally occurring cyclic (depsi) peptides. *J. Am. Chem. Soc.* 124, 5431–5440. doi: 10.1021/ja020166v

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* 14, 133–141. doi: 10.1208/s12248-012-9322-0

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR modeling: where have you been? where are you going to? *J. Med. Chem.* 57, 4977–5010. doi: 10.1021/jm4004285

Cleary, E. G., Beierlein, J. M., Khanuja, N. S., McNamee, L. M., and Ledley, F. D. (2018). Contribution of nih funding to new drug approvals 2010–2016. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2329–2334. doi: 10.1073/pnas.1715368115

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194. doi: 10.1038/nrg2537

Cordell, G. A. (2000). Biodiversity and drug discovery – a symbiotic relationship. *Phytochemistry.* 55, 463–480. doi: 10.1016/S0031-9422(00)00230-2

Cox, N., Kintzing, J. R., Smith, M., Grant, G. A., and Cochran, J. R. (2016). Integrin-targeting knottin peptide–drug conjugates are potent inhibitors of tumor cell proliferation. *Angew. Chem. Int. Ed.* 55, 9894–9897. doi: 10.1002/anie.201603488

Cozza, G., Bonvini, P., Zorzi, E., Poletto, G., Pagano, M. A., Sarno, S., et al. (2006). Identification of ellagic acid as potent inhibitor of protein kinase ck2: a successful example of a virtual screening application. *J. Med. Chem.* 49, 2363–2366. doi: 10.1021/jm060112m

Cui, L., Wang, Y., Liu, Z., Chen, H., Wang, H., Zhou, X., et al. (2015). Discovering new acetylcholinesterase inhibitors by mining the buzhongyiqi decoction recipe data. *J. Chem. Inf. Model.* 55, 2455–2463. doi: 10.1021/acs.jcim.5b00449

Daltry, J. C., Wüster, W., and Thorpe, R. S. (1996). Diet and snake venom evolution. *Nature.* 379:537.

de la Vega, R. C. R., and Possani, L. D. (2005). Overview of scorpion toxins specific for Na^+ channels and related peptides: biodiversity, structure–function relationships and evolution. *Toxicon* 46, 831–844. doi: 10.1016/j.toxicon.2005.09.006

Dewick, P. M. (2002). *Medicinal Natural Products: A Biosynthetic Approach*. West Sussex, UK: John Wiley & Sons.

Dhiman, P., Malik, N., and Khatkar, A. (2018). 3D-QSAR and *in-silico* studies of natural products and related derivatives as monoamine oxidase inhibitors. *Curr. Neuropharmacol.* 16, 881–900. doi: 10.2174/1570159X15666171128143650

Dias, D. A., Urban, S., and Roessner, U. (2012). A historical overview of natural products in drug discovery. *Metabolites.* 2, 303–336. doi: 10.3390/metabo2020303

Dickson, M., and Gagnon, J. P. (2004). Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discov.* 3, 417–429. doi: 10.1038/nrd1382

Drews, J. (2000). Drug discovery: a historical perspective. *Science* 287, 1960–1964. doi: 10.1126/science.287.5460.1960

Dudley, J. T., Sirota, M., Shenoy, M., Pai, R. K., Roedder, S., Chiang, A. P., et al. (2011). Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 3:96ra76. doi: 10.1126/scitranslmed.3002648

Dunkel, M., Fullbeck, M., Neumann, S., and Preissner, R. (2006). SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res.* 34(Suppl._1), D678–D683. doi: 10.1093/nar/gkj132

Eastwood, D., Findlay, L., Poole, S., Bird, C., Wadhwa, M., Moore, M., et al. (2010). Monoclonal antibody TGN1412 trial failure explained by species differences in CD28 expression on $CD4^+$ effector memory t-cells. *Br. J. Pharmacol.* 161, 512–526. doi: 10.1111/j.1476-5381.2010.00922.x

Ehret, T., Torelli, F., Klotz, C., Pedersen, A. B., and Seeber, F. (2017). Translational rodent models for research on parasitic protozoa–a review of confounders and possibilities. *Front. Cell. Infect. Microbiol.* 7:238. doi: 10.3389/fcimb.2017.00238

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., et al. (2015). The reactome pathway knowledgebase. *Nucleic Acids Res.* 44, D481–D487. doi: 10.1093/nar/gkv1351

FitzGerald, G. A. (2008). Drugs, industry, and academia. *Science* 320:1563. doi: 10.1126/science.1161006

Frank, R., and Hargreaves, R. (2003). Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov.* 2, 566–580. doi: 10.1038/nrd1130

Fu, D. H., Jiang, W., Zheng, J. T., Zhao, G. Y., Li, Y., Yi, H., et al. (2008). Jadomycin b, an aurora-b kinase inhibitor discovered through virtual screening. *Mol. Cancer Ther.* 7, 2386–2393. doi: 10.1158/1535-7163.MCT-08-0035

Gardner, S. P. (2005). Ontologies in drug discovery. *Drug Discov. Today Technol.* 2, 235–240. doi: 10.1016/j.ddtec.2005.08.004

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2016). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954. doi: 10.1093/nar/gkw1074

Halperin, I., Glazer, D. S., Wu, S., and Altman, R. B. (2008). The feature framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics* 9:S2. doi: 10.1186/1471-2164-9-S2-S2

Hao, Y., Quinnies, K., Realubit, R., Karan, C., and Tatonetti, N. P. (2018). Tissue-specific analysis of pharmacological pathways. *CPT Pharmacometrics Syst. Pharmacol.* 7, 453–463. doi: 10.1002/psp4.12305

Hao, Y., and Tatonetti, N. P. (2016). Predicting g protein-coupled receptor downstream signaling by tissue expression. *Bioinformatics* 32, 3435–3443. doi: 10.1093/bioinformatics/btw510

Harborne, J. B. (1999). Classes and functions of secondary products from plants. *Chem. Plants.* 1–25.

Harvey, A. L. (2008). Natural products in drug discovery. *Drug Discov. Today* 13, 894–901. doi: 10.1016/j.drudis.2008.07.004

Harvey, D., Bardelang, P., Goodacre, S. L., Cockayne, A., and Thomas, N. R. (2017). Antibiotic spider silk: site-specific functionalization of recombinant spider silk using "click" chemistry. *Adv. Mater.* 29:1604245. doi: 10.1002/adma.201604245

Hassane, D. C., Guzman, M. L., Corbett, C., Li, X., Abboud, R., Young, F., et al. (2008). Discovery of agents that eradicate leukemia stem cells using an *in silico* screen of public gene expression data. *Blood* 111, 5654–5662. doi: 10.1182/blood-2007-11-126003

He, Q. Y., He, Q. Z., Deng, X. C., Yao, L., Meng, E., Liu, Z. H., et al. (2007). ATDB: a uni-database platform for animal toxins. *Nucleic Acids Res.* 36(Suppl._1), D293–D297. doi: 10.1093/nar/gkm832

Hieronymus, H., Lamb, J., Ross, K. N., Peng, X. P., Clement, C., Rodina, A., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* 10, 321–330. doi: 10.1016/j.ccr.2006.09.005

Hiramoto, K., Yamate, Y., Kobayashi, H., Ishii, M., Miura, T., Sato, E. F., et al. (2012). Effect of the smell of seirogan, a wood creosote, on dermal and intestinal mucosal immunity and allergic inflammation. *J. Clin. Biochem. Nutr.* 51, 91–95. doi: 10.3164/jcbn.11-82

Hock, C., Konietzko, U., Streffer, J. R., Tracy, J., Signorell, A., Müller-Tillmanns, B., et al. (2003). Antibodies against β-amyloid slow cognitive decline in Alzheimer's disease. *Neuron.* 38, 547–554. doi: 10.1016/S0896-6273(03)00294-0

Hohlfeld, R. (1997). Biotechnological agents for the immunotherapy of multiple sclerosis. principles, problems and perspectives. *Brain.* 120, 865–916.

Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* 23:1105–1116. doi: 10.1038/nbt1126

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., et al. (2015). Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Stud. Health Technol. Inform.* 216:574. doi: 10.3233/978-1-61499-564-7-574

Huang, P. S., Boyken, S. E., and Baker, D. (2016). The coming of age of *de novo* protein design. *Nature.* 537:320. doi: 10.1038/nature19946

Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *Br. J. Pharmacol.* 162, 1239–1249. doi: 10.1111/j.1476-5381.2010.01127.x

Hunter, P. (2008). The paradox of model organisms: the use of model organisms in research will continue despite their shortcomings. *EMBO Rep.* 9, 717–720. doi: 10.1038/embor.2008.142

Ishida, Y., Agata, Y., Shibahara, K., and Honjo, T. (1992). Induced expression of pd-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO J.* 11, 3887–3895.

Jespers, L. S., Roberts, A., Mahler, S. M., Winter, G., and Hoogenboom, H. R. (1994). Guiding the selection of human antibodies from phage display repertoires to a single epitope of an antigen. *Biotechnology.* 12:899.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science* 303, 1813–1818. doi: 10.1126/science.1096361

Jungo, F., Bougueleret, L., Xenarios, I., and Poux, S. (2012). The uniprotkb/swiss-prot tox-prot program: a central hub of integrated venom protein data. *Toxicon.* 60, 551–557. doi: 10.1016/j.toxicon.2012.03.010

Jutel, M., Pichler, W. J., Skrbic, D., Urwyler, A., Dahinden, C., and Müller, U. (1995). Bee venom immunotherapy results in decrease of IL-4 and IL-5 and increase of IFN-gamma secretion in specific allergen-stimulated T cell cultures. *J. Immunol.* 154, 4187–4194.

Kaas, Q., Yu, R., Jin, A. H., Dutertre, S., and Craik, D. J. (2011). Conoserver: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.* 40, D325–D330. doi: 10.1093/nar/gkr886

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* 9:646. doi: 10.1021/ar020082r

Katz, L., and Baltz, R. H. (2016). Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* 43, 155–176. doi: 10.1007/s10295-015-1723-5

Khan, M. T. H., Orhan, I., Şenol, F., Kartal, M., Şener, B., Dvorská, M., et al. (2009). Cholinesterase inhibitory activities of some flavonoid derivatives and chosen xanthone and their molecular docking studies. *Chem. Biol. Interact.* 181, 383–389. doi: 10.1016/j.cbi.2009.06.024

Klaassen, C. D., and Watkins, J. B. (1996). *Casarett and Doull's Toxicology: The Basic Science of Poisons*, Vol. 5. New York, NY: McGraw-Hill New York.

Korotcov, A., Tkachenko, V., Russo, D. P., and Ekins, S. (2017). Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol. Pharm.* 14, 4462–4475. doi: 10.1021/acs.molpharmaceut.7b00578

Kulkarni, J., Garland, K. A., Scaffidi, A., Headey, B., Anderson, R., de Castella, A., et al. (2006). A pilot study of hormone modulation as a new treatment for mania in women with bipolar affective disorder. *Psychoneuroendocrinology* 31, 543–547. doi: 10.1016/j.psyneuen.2005.11.001

Larsen, T. O., Smedsgaard, J., Nielsen, K. F., Hansen, M. E., and Frisvad, J. C. (2005). Phenotypic taxonomy and metabolite profiling in microbial drug discovery. *Nat. Prod. Rep.* 22, 672–695. doi: 10.1039/b404943h

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20, 318–331. doi: 10.1016/j.drudis.2014.10.012

Leach, D. R., Krummel, M. F., and Allison, J. P. (1996). Enhancement of antitumor immunity by CTLA-4 blockade. *Science* 271, 1734–1736.

Lee, K.-H., Lo, H.-L., Tang, W.-C., Hsiao, H. H.-Y., and Yang, P.-M. (2014). A gene expression signature-based approach reveals the mechanisms of action of the chinese herbal medicine berberine. *Sci. Rep.* 4:6394. doi: 10.1038/srep06394

Lee, K. W., Bode, A. M., and Dong, Z. (2011). Molecular targets of phytochemicals for cancer prevention. *Nat. Rev. Cancer* 11, 211–218. doi: 10.1038/nrc3017

Letvin, N. L., and Walker, B. D. (2003). Immunopathogenesis and immunotherapy in aids virus infections. *Nat. Med.* 9, 861–866. doi: 10.1038/nm0703-861

Lewis, R. J., and Garcia, M. L. (2003). Therapeutic potential of venom peptides. *Nat. Rev. Drug Discov.* 2, 790–802. doi: 10.1038/nrd1197

Li, Q., Cheng, T., Wang, Y., and Bryant, S. H. (2010). Pubchem as a public resource for drug discovery. *Drug Discov. Today.* 15, 1052–1057. doi: 10.1016/j.drudis.2010.10.003

Lindsay, M. A. (2005). Finding new drug targets in the 21st century. *Drug Discov. Today.* 10, 1683–1687. doi: 10.1016/S1359-6446(05)03670-6

Lipinski, C. A. (2004). Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1, 337–341. doi: 10.1016/j.ddtec.2004.11.007

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bull. Med. Lib. Assoc.* 88:265.

Lusher, S. J., McGuire, R., van Schaik, R. C., Nicholson, C. D., and de Vlieg, J. (2014). Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today.* 19, 859–868. doi: 10.1016/j.drudis.2013.12.004

Lv, C., Wu, X., Wang, X., Su, J., Zeng, H., Zhao, J., et al. (2017). The gene expression profiles in response to 102 traditional chinese medicine (TCM) components: a general template for research on TCMs. *Sci. Rep.* 7:352. doi: 10.1038/s41598-017-00535-8

Ma, D.-L., Chan, D. S.-H., and Leung, C.-H. (2011). Molecular docking for virtual screening of natural product databases. *Chem. Sci.* 2, 1656–1665. doi: 10.1039/C1SC00152C

Malhotra, A., Creer, S., Harris, J. B., Stöcklin, R., Favreau, P., and Thorpe, R. S. (2013). Predicting function from sequence in a large multifunctional toxin family. *Toxicon.* 72, 113–125. doi: 10.1016/j.toxicon.2013.06.019

Mandrika, I., Prusis, P., Yahorava, S., Tars, K., and Wikberg, J. E. (2007). QSAR of multiple mutated antibodies. *J. Mol. Recognit.* 20, 97–102. doi: 10.1002/jmr.817

Mann, J. (2002). Natural products in cancer chemotherapy: past, present and future. *Nat. Rev. Cancer* 2, 143–148. doi: 10.1038/nrc723

Maplestone, R. A., Stone, M. J., and Williams, D. H. (1992). The evolutionary role of secondary metabolites—a review. *Gene.* 115, 151–157.

Markland, W., Ley, A. C., and Ladner, R. C. (1996). Iterative optimization of high-affinity protease inhibitors using phage display. 2. Plasma kallikrein and thrombin. *Biochemistry.* 35, 8058–8067.

Markwell, J., and Brooks, D. W. (2003). "Link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology. *Biochem. Mol. Biol. Educ.* 31, 69–72. doi: 10.1002/bmb.2003.494031010165

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., et al. (2011). The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics.* 4:13. doi: 10.1186/1755-8794-4-13

McCarty, C. A., and Wilke, R. A. (2010). Biobanking and pharmacogenomics. *Pharmacogenomics.* 11, 637–641. doi: 10.2217/pgs.10.13

Menting, J. G., Gajewiak, J., MacRaild, C. A., Chou, D. H.-C., Disotuar, M. M., Smith, N. A., et al. (2016). A minimized human insulin-receptor-binding motif revealed in a conus geographus venom insulin. *Nat. Struct. Mol. Biol.* 23, 916–920. doi: 10.1038/nsmb.3292

Miller, D. W., Jones, A. D., Goldston, J. S., Rowe, M. P., and Rowe, A. H. (2016). Sex differences in defensive behavior and venom of the striped bark scorpion centruroides vittatus (Scorpiones: Buthidae). *Integr. Comp. Biol.* 56, 1022–1031. doi: 10.1093/icb/icw098

Mladenović, M., Patsilinakos, A., Pirolli, A., Sabatino, M., and Ragno, R. (2017). Understanding the molecular determinant of reversible human monoamine oxidase B inhibitors containing 2H-chromen-2-one core: structure-based and ligand-based derived three-dimensional quantitative structure–activity relationships predictive models. *J. Chem. Inf. Model* 57, 787–814. doi: 10.1021/acs.jcim.6b00608

Molyneux, R. J., Lee, S. T., Gardner, D. R., Panter, K. E., and James, L. F. (2007). Phytochemicals: the good, the bad and the ugly? *Phytochemistry* 68, 2973–2985. doi: 10.1016/j.phytochem.2007.09.004

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* 8:959. doi: 10.1038/nrd2961

Musa, A., Ghoraie, L. S., Zhang, S. D., Glazko, G., Yli-Harja, O., Dehmer, M., et al. (2017). A review of connectivity map and computational approaches in pharmacogenomics. *Brief. Bioinform.* 19, 506–523. doi: 10.1093/bib/bbw112

Mutowo, P., Bento, A. P., Dedman, N., Gaulton, A., Hersey, A., Lomax, J., et al. (2016). A drug target slim: using gene ontology and gene ontology annotations to navigate protein-ligand target space in ChEMBL. *J. Biomed. Semantics* 7:59. doi: 10.1186/s13326-016-0102-0

National Center for Complementary and Integrative Health (2017). *Natural Products Research—Information for Researchers*.

Nature Publishing Group (2007). All natural. *Nat. Chem. Biol.* 3:351. doi: 10.1038/nchembio0707-351

Newman, D. J., and Cragg, G. M. (2016). Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* 79, 629–661. doi: 10.1021/acs.jnatprod.5b01055

Nica, A. C., and Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.* 17, R129–R134. doi: 10.1093/hmg/ddn285

Nitta, S. K., and Numata, K. (2013). Biopolymer-based nanoparticles for drug/gene delivery and tissue engineering. *Int. J. Mol. Sci.* 14, 1629–1654. doi: 10.3390/ijms1401162

Nixon, A. E., Sexton, D. J., and Ladner, R. C. (2014). Drugs derived from phage display: from candidate identification to clinical practice. *MAbs.* 6, 73–85. doi: 10.4161/mabs.27240

Oellrich, A., Collier, N., Groza, T., Rebholz-Schuhmann, D., Shah, N., Bodenreider, O., et al. (2015). The digital revolution in phenotyping. *Brief. Bioinform.* 17, 819–830. doi: 10.1093/bib/bbv083

Pagadala, N. S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. *Biophys. Rev.* 9, 91–102. doi: 10.1007/s12551-016-0247-1

Pérez, A., Martínez-Rosell, G., and De Fabritiis, G. (2018). Simulations meet machine learning in structural biology. *Curr. Opin. Struct. Biol.* 49, 139–144. doi: 10.1016/j.sbi.2018.02.004

Pineda, S. S., Chaumeil, P.-A., Kunert, A., Kaas, Q., Thang, M. W., Le, L., et al. (2017). Arachnoserver 3.0: an online resource for automated discovery, analysis and annotation of spider toxins. *Bioinformatics* 34, 1074–1076. doi: 10.1093/bioinformatics/btx661

Pithayanukul, P., Leanpolchareanchai, J., and Saparpakorn, P. (2009). Molecular docking studies and anti-snake venom metalloproteinase activity of thai mango seed kernel extract. *Molecules* 14, 3198–3213. doi: 10.3390/molecules14093198

Polanski, J. (2009). Receptor dependent multidimensional qsar for modeling drug-receptor interactions. *Curr. Med. Chem.* 16, 3243–3257. doi: 10.2174/092986709788803286

Ramsay, R. R., Popovic-Nikolic, M. R., Nikolic, K., Uliassi, E., and Bolognesi, M. L. (2018). A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* 7:3. doi: 10.1186/s40169-017-0181-2

Rodrigues, T., Reker, D., Schneider, P., and Schneider, G. (2016). Counting on natural products for drug design. *Nat. Chem.* 8, 531–541. doi: 10.1038/nchem.2479

Romano, J., Nwankwo, V., and Tatonetti, N. (2018). Venomkb v2.0: a knowledge repository for computational toxinology. *bioRxiv [preprint]. preprint: 295204.* doi: 10.1101/295204

Romano, J. D., and Tatonetti, N. P. (2015). VenomKB, a new knowledge base for facilitating the validation of putative venom therapies. *Sci. Data* 2:150065. doi: 10.1038/sdata.2015.65

Romano, J. D., and Tatonetti, N. P. (2016). Using a novel ontology to inform the discovery of therapeutic peptides from animal venoms. *AMIA Summits Transl. Sci. Proc.* 2016:209.

Romano, J. D., Tharp, W. G., and Sarkar, I. N. (2015). Adapting simultaneous analysis phylogenomic techniques to study complex disease gene relationships. *J. Biomed. Inform.* 54, 10–38. doi: 10.1016/j.jbi.2015.01.002

Rønsted, N., Symonds, M. R., Birkholm, T., Christensen, S. B., Meerow, A. W., Molander, M., et al. (2012). Can phylogeny predict chemical diversity and potential medicinal activity of plants? a case study of amaryllidaceae. *BMC Evol. Biol.* 12:182. doi: 10.1186/1471-2148-12-182

Rosenberg, S. A., Yang, J. C., and Restifo, N. P. (2004). Cancer immunotherapy: moving beyond current vaccines. *Nat. Med.* 10, 909–915. doi: 10.1038/nm1100

Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., et al. (2014). TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminform.* 6:13. doi: 10.1186/1758-2946-6-13

Ruau, D., Mbagwu, M., Dudley, J. T., Krishnan, V., and Butte, A. J. (2011). Comparison of automated and human assignment of mesh terms on publicly-available molecular datasets. *J. Biomed. Inform.* 44, S39–S43. doi: 10.1016/j.jbi.2011.03.007

Rudolf, J. D., Yan, X., and Shen, B. (2016). Genome neighborhood network reveals insights into enediyne biosynthesis and facilitates prediction and prioritization for discovery. *J. Ind. Microbiol. Biotechnol.* 43, 261–276. doi: 10.1007/s10295-015-1671-0

Rzhetsky, A., Wajngurt, D., Park, N., and Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci.* 104, 11694–11699. doi: 10.1073/pnas.0704820104

Salmaso, V., and Moro, S. (2018). Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: an overview. *Front. Pharmacol.* 9:923. doi: 10.3389/fphar.2018.00923

Schuemie, M. J., Coloma, P. M., Straatman, H., Herings, R. M., Trifirò, G., Matthews, J. N., et al. (2012). Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Med. Care.* 50, 890–897. doi: 10.1097/MLR.0b013e31825f63bf

Seo, S. K., Choi, J. H., Kim, Y. H., Kang, W. J., Park, H. Y., Suh, J. H., et al. (2004). 4-1BB-mediated immunotherapy of rheumatoid arthritis. *Nat. Med.* 10, 1088–1094. doi: 10.1038/nm1107

Sevigny, J., Chiao, P., Bussière, T., Weinreb, P. H., Williams, L., Maier, M., et al. (2016). The antibody aducanumab reduces Aβ plaques in Alzheimer's disease. *Nature* 537, 50–56. doi: 10.1038/nature19323

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77. doi: 10.1126/scitranslmed.3001318

Sivasubramanian, A., Sircar, A., Chaudhury, S., and Gray, J. J. (2009). Toward high-resolution homology modeling of antibody fv regions and application to antibody–antigen docking. *Proteins* 74, 497–514. doi: 10.1002/prot.22309

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395. doi: 10.1124/pr.112.007336

Spedding, M. (2006). New directions for drug discovery. *Dialogues Clin. Neurosci.* 8:295.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Spiess, K., Lammel, A., and Scheibel, T. (2010). Recombinant spider silk proteins for applications in biomaterials. *Macromol. Biosci.* 10, 998–1007. doi: 10.1002/mabi.201000071

Stone, M. J., and Williams, D. H. (1992). On the evolution of functional secondary metabolites (natural products). *Mol. Microbiol.* 6, 29–34.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452. doi: 10.1016/j.cell.2017.10.049

Sunagar, K., Undheim, E. A., Scheib, H., Gren, E. C., Cochran, C., Person, C. E., et al. (2014). Intraspecific venom variation in the medically significant southern pacific rattlesnake (crotalus oreganus helleri): biodiscovery, clinical and evolutionary implications. *J. Proteomics* 99, 68–83. doi: 10.1016/j.jprot.2014.01.013

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* 30, 7–18.

Tatonetti, N. P., Patrick, P. Y., Daneshjou, R., and Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* 4:125ra31. doi: 10.1126/scitranslmed.3003377

Tayeb, H. O., Murray, E. D., Price, B. H., and Tarazi, F. I. (2013). Bapineuzumab and solanezumab for Alzheimer's disease: is the 'amyloid

cascade hypothesis' still alive? *Expert Opin. Biol. Ther.* 13, 1075–1084. doi: 10.1517/14712598.2013.789856

Terrett, N. K., Gardner, M., Gordon, D. W., Kobylecki, R. J., and Steele, J. (1995). Combinatorial synthesis—the design of compound libraries and their application to drug discovery. *Tetrahedron.* 51, 8135–8173.

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., et al. (2018). Tensor field networks: rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv [Preprint], arXiv:1802.08219*.

Thomford, N. E., Senthebane, D. A., Rowe, A., Munro, D., Seele, P., Maroyi, A., et al. (2018). Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int. J. Mol. Sci.* 19:1578. doi: 10.3390/ijms19061578

Torng, W., and Altman, R. B. (2017). 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinform.* 18:302. doi: 10.1186/s12859-017-1702-0

Tosco, P., and Balle, T. (2011). Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. *J. Mol. Model.* 17, 201–208. doi: 10.1007/s00894-010-0684-x

Vazquez-Naya, J. M., Martinez-Romero, M., Porto-Pazos, A. B., Novoa, F., Valladares-Ayerbes, M., Pereira, J., et al. (2010). Ontologies of drug discovery and design for neurology, cardiology and oncology. *Curr. Pharm. Des.* 16, 2724–2736. doi: 10.2174/138161210792389199

Vyas, V., Ukawala, R., Ghate, M., and Chintha, C. (2012). Homology modeling a fast tool for drug discovery: current perspectives. *Indian J. Pharm. Sci.* 74, 1–17. doi: 10.4103/0250-474X.102537

Walls, P. H., and Sternberg, M. J. (1992). New algorithm to model protein-protein recognition based on surface complementarity: applications to antibody-antigen docking. *J. Mol. Biol.* 228, 277–297.

Weiskopf, N. G., and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc* 20, 144–151. doi: 10.1136/amiajnl-2011-000681

Welsch, M. E., Snyder, S. A., and Stockwell, B. R. (2010). Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* 14, 347–361. doi: 10.1016/j.cbpa.2010.02.018

Wilke, R. A., Xu, H., Denny, J., Roden, D., Krauss, R., McCarty, C., et al. (2011). The emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.* 89, 379–386. doi: 10.1038/clpt.2010.260

Williams, N. (1997). How to get databases talking the same language. *Science.* 275, 301–302.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wollacott, A. M., Robinson, L. N., Ramakrishnan, B., Tissire, H., Viswanathan, K., Shriver, Z., et al. (2019). Structural prediction of antibody-april complexes by computational docking constrained by antigen saturation mutagenesis library data. *J. Mol. Recognit.* e2778. doi: 10.1002/jmr.2778

Xie, G., Plumb, R., Su, M., Xu, Z., Zhao, A., Qiu, M., et al. (2008). Ultra-performance LC/TOF MS analysis of medicinal panax herbs for metabolomic research. *J. Sep. Sci.* 31, 1015–1026. doi: 10.1002/jssc.200700650

Xie, T., Song, S., Li, S., Ouyang, L., Xia, L., and Huang, J. (2015). Review of natural product databases. *Cell Prolif.* 48, 398–404. doi: 10.1111/cpr.12190

Yan, T., Fu, Q., Wang, J., and Ma, S. (2015). UPLC-MS/MS determination of ephedrine, methylephedrine, amygdalin and glycyrrhizic acid in beagle plasma and its application to a pharmacokinetic study after oral administration of ma huang tang. *Drug Test. Anal.* 7, 158–163. doi: 10.1002/dta.1635

Yang, Y., Mis, M. A., Estacion, M., Dib-Hajj, S. D., and Waxman, S. G. (2018). NaV 1.7 as a pharmacogenomic target for pain: Moving toward precision medicine. *Trends Pharmacol. Sci.* 39, 258–275. doi: 10.1016/j.tips.2017.11.010

Yao, L., Zhang, Y., Li, Y., Sanseau, P., and Agarwal, P. (2011). Electronic health records: Implications for drug discovery. *Drug Discov. Today* 16, 594–599. doi: 10.1016/j.drudis.2011.05.009

Yue, Y., Chu, G. X., Liu, X. S., Tang, X., Wang, W., Liu, G. J., et al. (2014). TMDB: a literature-curated database for small molecular compounds found from tea. *BMC Plant Biol.* 14:243. doi: 10.1186/s12870-014-0243-1

Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., et al. (2017). NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* 46, D1217–D1222. doi: 10.1093/nar/gkx1026

Zhang, B., Fu, Y., Huang, C., Zheng, C., Wu, Z., Zhang, W., et al. (2016). New strategy for drug discovery by large-scale association analysis of molecular networks of different species. *Sci. Rep.* 6:21872. doi: 10.1038/srep21872

Zhang, R., Manohar, N., Arsoniadis, E., Wang, Y., Adam, T. J., Pakhomov, S. V., et al. (2015). Evaluating term coverage of herbal and dietary supplements in electronic health records. *AMIA Annu. Symp. Proc.* 2015:1361.

Ziemert, N., and Jensen, P. R. (2012). Phylogenetic approaches to natural product structure prediction. *Methods Enzymol.* 517, 161–182. doi: 10.1016/B978-0-12-404634-4.00008-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.