 doi:10.20944/preprints201811.0561.v1

"Drug-likeness" *versus* "natural product-likeness"

Fidele Ntie-Kang[1,2,3]*, Kennedy D. Nyongbela[1]*, Godfred A. Ayimele[1], Suhaib Shekfeh[4]

[1]University of Buea, Pharmacochemistry Research Group, Chemistry Department, P. O. Box 63 Buea, Buea, Cameroon

[2]Martin-Luther-Universitat Halle-Wittenberg, Department of Pharmaceutical Chemistry Wolfgang-Langenbeck Str. 4, 06120 Halle (Saale), Germany

[3]University of Chemistry and Technology Prague, Department of Informatics and Chemistry, Technická 5 166 28 Prague 6, Dejvice, Czech Republic

[4]MTS, Medicinal Chemistry, Wolframstr. 3, 86161 Augsburg, Germany

Corresponding authors: ntiekfidele@gmail.com (FNK); knyongbela@gmail.com (KDN)

Abstract. We discuss further details on the concepts of "drug-likeness", "lead-likeness", and "natural product-likeness". The discussion will first focus on natural products as drugs, then a discussion of previous studies in which the complexities of the scaffolds and chemical space of naturally occurring compounds have been compared with synthetic, semi-synthetic compounds and FDA-approved drugs. This is followed by guiding principles for designing "drug-like" natural product libraries for lead compound discovery purposes. We end up by presenting a tool for measuring "natural product-likeness" of compounds and a brief presentation of machine learning approaches and a binary quantitative structure-activity relationship (QSAR) for classifying drugs from non-drugs and natural compounds from non-natural ones, respectively.

Keywords: cheminformatics, drugs, drug-likeness, drug discovery, natural products.

1 Introduction

In a previous chapter, focused on the definition and classification of natural products (NPs), NPs were defined as substances isolated from plants, micro-organisms, insects, mammals, etc. [1]. Since the term NPs is not inclusive of products of primary metabolism or those found in all living cells, e.g. proteins, nucleic acids, carbohydrates, and compounds that are substrates for biological transporters. The definition of NPs is rather restricted to products of secondary metabolism. Thus, NPs could also be referred to as secondary metabolites (SMs). Consequently, the terms NP, SM or naturally-occurring metabolite or compound (NOC) will be used interchangeably throughout the text. NPs and NP mimics comprise an important source for human therapeutics and are estimated to compass a significant market share among the approved drugs [2]. Throughout human history, nature's chemical library has been proven to be a rich resource for many biologically active medicinal leads and drugs. However, major changes in trends of the drug discovery programs have occurred during the last four decades. Thus, drug discovery programs started to focus on target-based methods; after the emergence of in vitro assays and the development of large combinatorial chemical libraries. With the increasing need to have access to large libraries or chemical library collections for screening, which was not clearly possible with classical NP extraction and purification methods, these last changes enabled the shift from the natural products-based discovery programs to the high-throughput screening (HTS) technology as the main strategy for target-based drug discovery programs [3].

Following the availability of huge chemical screening libraries from combinatorial synthesis and valuable biological activities data collected from HTS, it became clear that medicinal chemists needed some criteria to distinguish between biologically active compounds and drugs. Christopher Lipinski was able to analyze a wealth

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

of data that had accumulated from the HTS and failed drug discovery programs which had been stored in the World Drug Index (WDI) during the 1980s and 1990s. He then suggested that the concept of "drug-likeness" was linked to oral bioavailability, hence to the famous "rule of 5" (RO5). Oral bioavailability is the ability of a drug to be administered orally in an efficient manner, a concept often linked to the absorption, distribution, metabolism, excretion and toxicity (ADME/T) of drug molecules. This is the ability of the drug to cross the intestinal walls, go through general blood circulation, reach it's intended target site and eventually stay at the target site in sufficient time to carry out its pharmacological function, then be eliminated efficiently so as not to accumulate into amounts that are unsafe (toxic) to the body. The RO5 comprises a simple set of 4 physical-chemical property ranges, which give the biologically active compound higher probability to be orally bioavailable and promising favorable pharmacokinetic (ADME/T) profile [4]. The RO5 include:

1) molecular weight (MW) less than 500 Da
2) computed logarithm of on *n*-octanol/water partition coefficient (clogP) less than 5
3) number of hydrogen bond acceptors (HBA, defined as the number of N and O atoms) less than 10
4) number of hydrogen bond donors (HBD, defined as the sum of OH and NH groups) less than 5

The RO5 derives its appellation from the fact that these numbers are all multiples of 5. A further proposal by Oprea and colleagues suggested more stringent rules to identify what is called lead-likeness [5, 6]. Compounds to be classified as "leads" were by definition:

1) less complex compounds with less chemical features,

2) display good biological activity with good ADMET profile and

3) amenable for chemical optimization to improve the biological activity or enhance the pharmacokinetic properties.

Hence, Oprea and coworkers were able to distinguished the lead-like chemical space from the drug-like space by stating the following lead-likeness conditions, otherwise known as Oprea lead-likeness filters:

1) MW: maximum 450
2) clogP: between -3,5 and +4,5
3) HBA: maximum 8
4) HBD: maximum 5

The statistical analysis of three compound classes; natural products, molecules from combinatorial synthesis, and drug molecules, displayed significant differences between the combinatorial synthetic libraries and NP libraries [7]. In another chapter in this book, Saldívar-González et al. discuss some major compound databases of NPs and cheminformatics strategies that have been used to characterize the chemical space of natural products. The authors analyzed NPs from different sources and their relationships with other compounds are also discussed using novel chemical descriptors and data mining approaches that are emerging to characterize the chemical space of naturally occurring compounds [8]. In this chapter, our discussion will first focus on NPs as drugs, then a discussion of previous studies comparing NPs and drugs, a brief discussion on NPs, principles of designing NP libraries, the concept of "natural product-likeness" and finally tools and used for the prediction of "drug-likeness" and NP-likeness.

2 Natural products as drugs

2.1 The proportion of natural products in catalogues of drugs

NPs play important roles in drug discovery, providing scaffolds as starting points for hit/lead discovery [9, 10]. Several known drugs, e.g. the anticancer compounds (**1** to **5**, Fig. 1), are from natural sources [11]. We must note that NPs continue to play a role as drugs [2], as biological probes, and as study targets for synthetic and analytical chemists [12]. About half of all approved drugs between 1981 and 2010 were shown to be NP-based [13]. This study also showed that, of all approved drugs, NPs constituted 6% (unaltered), 26% (NP derivatives), 32% (NP mimics) or from NP pharmacophores, 73% of small molecule antibacterials and 50% of anticancer drugs (including taxol, vinblastine, vincristine, topotecan, etc.), Table 1. This implies that if structural features provided by nature

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

are successfully incorporated into synthetic drugs (SDs), this would increase the chemical diversity available for small-molecule drug discovery [2]. Howbeit, the reasons for the decline of interest by the pharmaceutical industry during the last two decades include the time factor involved in the search for NP lead compounds to the labor intensiveness of the whole process [14]. This has now been rendered much easier within industrial settings by streamlined screening procedures and enhanced organism sourcing mechanisms [15].

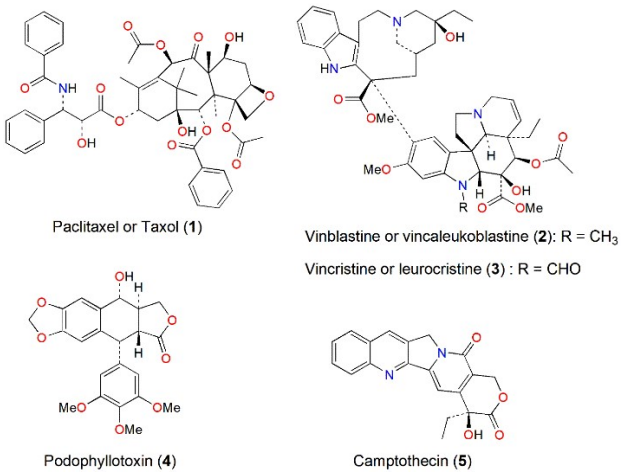

Figure 1: Selected naturally occurring NP anticancer drug leads.

2.2 The future of natural product drug discovery

Despite their evolving role in drug discovery [16, 17], a recent chemoinformatic study involving a dataset of all published microbial and marine-derived compounds since the 1940s (comprising 40,229 NPs) showed that most NPs being published today bear close similarity to previously published structures, with a plateau being observed since the mid-1990s [18], Figure 2. The authors observed a general trend that the rate of discovery of new NPs had flattened out since the 1990s, structures with novel scaffolds had become scarce (Figure 3). In the mentioned study, two compounds were considered to be dissimilar by taking a Tanimoto cutoff of $T_c < 0.4$. This study had, thus, suggested that the range of scaffolds readily accessible from nature is limited, i.e. scientists were close to having described all of the chemical space covered by NPs, even though appreciable numbers of NPs with no structural precedents continue to be discovered.

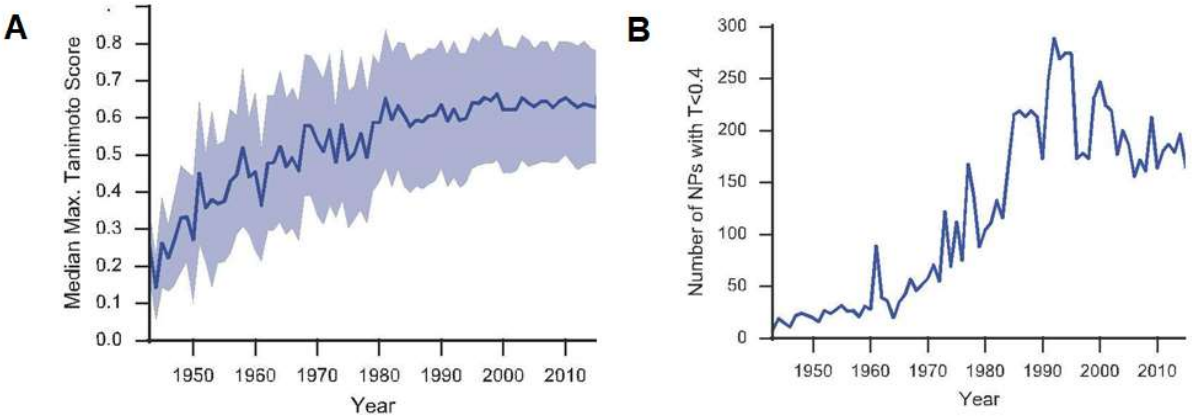

Figure 2: Presentation of structural diversity (A) by plotting the median maximum Tanimoto scores as a function of time. The median average deviation shown as shaded blue region. (B) by plotting the absolute number of low similarity compounds ($T_c < 0.4$) per year [18]. Figures reproduced by permission.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

A reproduction of the same study using another dataset of 32,380 NPs, showed the same trend [19]. However, a similar analysis on a dataset of randomly selected compounds from the ZINC database having overall lower structural similarity, the authors of the latter study further proved that such trends may be a feature of any growing database of chemical structures, rather than reflecting trends specific to NP discovery. Besides, a Kolmogorov–Smirnov test conducted on the dataset of 40,229 NPs, with $P = 6.2\times10^{-14}$, showed that since 1990, the rate of structurally novel compound discovery has dramatically outpaced random expectation [19]. This implies that NPs discovered within the last three decades have been characterized by unprecedented chemical diversity, suggesting that the dream of continuously discovering new chemical structures from nature remains positive.

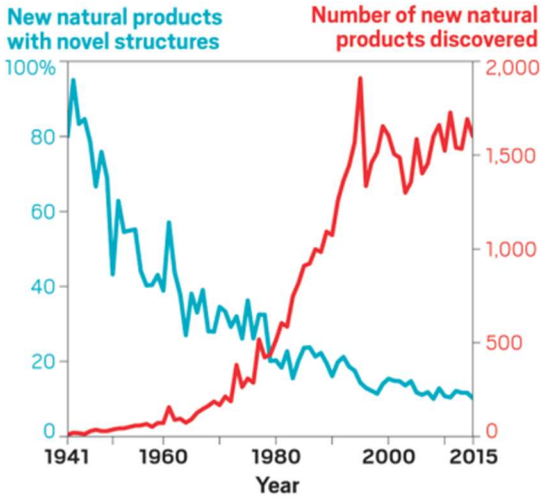

Figure 3: Presentation of structural diversity by plotting the number of compounds published per year and rate of novel compounds isolated as a percentage of total natural products isolated [18]. Figure reproduced by permission.

3 Natural *versus* synthetic drugs

3.1 The uniqueness and potential of natural products for drug discovery

NPs are unique, when compared with SDs in that they often contain more complex scaffolds and chiral centers, with more O-atoms and aromatic groups [20], Table 1. In addition, a study involving a comparison of SDs *versus* NPs showed that drugs derived from NP-based structures display greater chemical diversity and occupy wider regions of chemical space [2]. This is because drugs which are synthesized based on NP pharmacophores often exhibit lower hydrophobicity and greater stereochemical content when compared with drugs which are completely of synthetic origin. Natural products mostly are more potent with higher binding affinities to a specific biological receptor. Consequently, their biological activities are often more selective than the synthetic compounds. A property distribution of three investigated datasets consisting of 3,287 NPs, 10,968 drug molecules and 13,506 randomly selected combinatorially-derived lead candidates, respectively, led to the analysis of the number of chiral centers, rotatable bonds, aromatic rings, complex ring systems, degrees of saturation, as well as the ratios of different heteroatoms (O, N, etc.) [7]. This study showed that the main structural differences between NPs and combinatorially-derived libraries arise from properties introduced during the synthetic process in order to render combinatorial synthesis more efficient. Moreover, it was shown that, since drug molecules originate from both natural and synthetic sources, they occupy a joint area of chemical space spread between NPs and combinatorially-derived compounds.

Although NPs are often said not to satisfy all criteria of the RO5, a large proportion of NP libraries provide very good leads for drug development. For example, 60% of the 126,140 unique compounds in the DNP were found as 'drug-like', complying with the RO5 [21]. Moreover, other investigations revealed that only 10% of analyzed NP libraries violated two or more of Lipinski's RO5 [17, 22]. Attempts to quantify biosynthetic bias in screening

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

libraries showed that 83% (12 977) of core ring scaffolds present in NPs are missed in the combinatorial databases [23, 24], and the inclusion of these missed NP fragments inside the screening libraries would improve the hit rates [23]. In order to bring the drug-like space of synthesized chemical closer to the properties of natural products, a new measure called natural product-likeness score was proposed by Ertl et al. (section 7) [25].

Table 1: Comparative summary between natural products and synthetic drugs. Data derived from [2].

Property	Natural Products	Synthetic Drugs
Samples	Limited quantities (time consuming due to cumbersome extraction processes)	Readily available from combinatorial libraries
Drug-likeness	Weaker bioavailability (generally poorer ADME/T profiles)	More bioavailable
Chemistry	Complex scaffolds, more stereogenic centres	Less O-atoms, less aromatics, etc.
Proportion of marketed drugs	- 6% (unaltered), -26% (NP derivatives), -32% (NP mimics) or from NP pharmacophores -73% of small molecule antibacterials -50% of anticancer drugs (taxol, vinblastine, vincristine, and topotecan, etc.)	

3.2 The complexity and diversity of natural product scaffolds

NPs are, generally, compounds with large, diverse and structurally complex scaffolds (see previous chapter [1]). This is because during their often complex biosynthesis processes. The NPs contained in the DNP were previously according to their origins using a classification tree approach [26], with the aim of analyzing systems of rings that are typical according to the source. The high selectivity of natural products is attributed to their higher degree of complexity, higher number of stereogenic centers, more polar functional groups, and different ratios of atom types, e.g. N, O, S, and halogens [7]. This study shed the light on the remarkable diversity of natural products occupying different region of the chemical space with distinct ranges of the physical-chemical properties. The complexity and diversity of NPs has been illustrated by use of the tool ChemGPS-NP [27], which was designed for handling the chemical diversity encountered in natural products research, in contrast to previously designed chemical global positioning system (ChemGPS) [28], which focused on the much more restricted drug-like chemical space. The uniqueness of the ChemGPS-NP tool is that, as contrasted to ChemGPS is that a better representation of biologically relevant chemical space is achieved by including complex structural examples from the creative chemistry of naturally-occurring bioactive molecules. Rules for plotting the chemical space maps include aspects of size, shape, lipophilicity, polarity, polarizability, flexibility, rigidity, and hydrogen bond capacity. In ChemGPS-NP the chemical space map coordinates are t-scores derived from principal component analysis (PCA) [29]. This is achieved through a carefully selected subset of 35 descriptors that evaluate rules on a total set of 1779 chosen satellite and core structures [27]. In Figure 4, we illustrate the complexity of NPs by the diversities of the three most important principal component values or t-scores (t_1, t_2 and t_3) [27].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

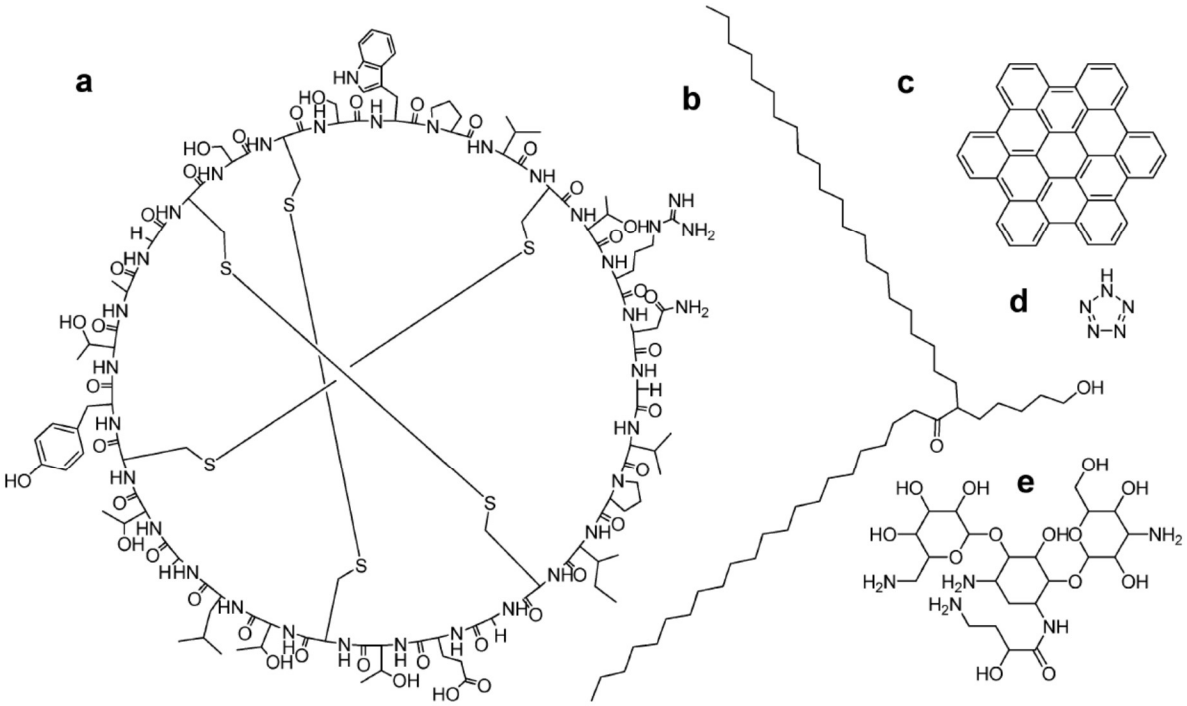

Figure 4: Five selected structures and their ChemGPS-NP coordinates for the three most important principal components (a) varv F: $t_1 = 46.4$, $t_2 = -10.2$, $t_3 = -5.84$; (b) 23-(5-Hydroxypentyl)-22-pentatetracontanone: $t_1 = 4.83$, $t_2 = -4.52$, $t_3 = 7.58$; (c) Hexabenzocoronene: $t_1 = 4.33$, $t_2 = 12.0$, $t_3 = 4.79$; (d) Pentazole: $t_1 = -4.82$, $t_2 = -2.32$, $t_3 = -2.98$; (e) Amikacin: $t_1 = 5.01$, $t_2 = -3.85$, $t_3 = -6.89$. Figure reproduced by permission.

4 Navigating the natural product chemical space

Several investigations of the three-dimensional (3D) chemical space, occupied by compounds of synthetic and natural origins, using principal component analysis (PCA) have been published [2, 7, 27, 30-35]. It was generally observed that, when compared with Food and Drug Administration (FDA)-approved drugs and SDs, the distribution of NPs in chemical space cover regions that lack representation in synthetic medicinal chemistry compounds (Figure 5), thus showing that NPs have a much wider coverage of chemical space. In the following sub-paragraphs, we examine a few case studies in more detail.

4.1 The Universal Natural Products Database (UNPD) *versus* FDA-approved drugs

Figure 4A shows an example of the visualization of the chemical space of NPs according to the origin of the compounds from the Universal Natural Products Database (UNPD), shown in green, when compared with a dataset of drugs approved by the Food and Drug Administration (FDA), USA, shown in black [30]. In this study, Gu and colleagues collected a total 197,201 NPs, by including data structures from the Reaxys database [36], the Chinese Natural Product Database (CNPD) [37], the Traditional Chinese Medicines Database (TCMD) [38], and the Chinese traditional medicinal herbs database (CHDD) [39]. The authors then used PCA to explore their chemical space, by superposing with that of FDA-approved drugs. This study showed that the NPs occupied a much large portion of overlap between NPs and FDA-approved drugs in the chemical space, indicating that the investigated NPs had large quantity of potential lead compounds not yet approved by the FDA, thus NPs have a vast chemical diversity when compared with known drugs. Besides, the authors explored the network properties of NP-target networks and found that their polypharmacology was greatly enriched to those compounds with large degree and high betweenness centrality. Although a vast number of the NPs included in the UNPD had no biological activities, by docking all the derived 3D structures towards the 332 target proteins of the FDA-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

approved drugs, it was shown, based on a docking score-weighted prediction model, that NPs have good drug-like properties and could interact with multiple cellular target proteins.

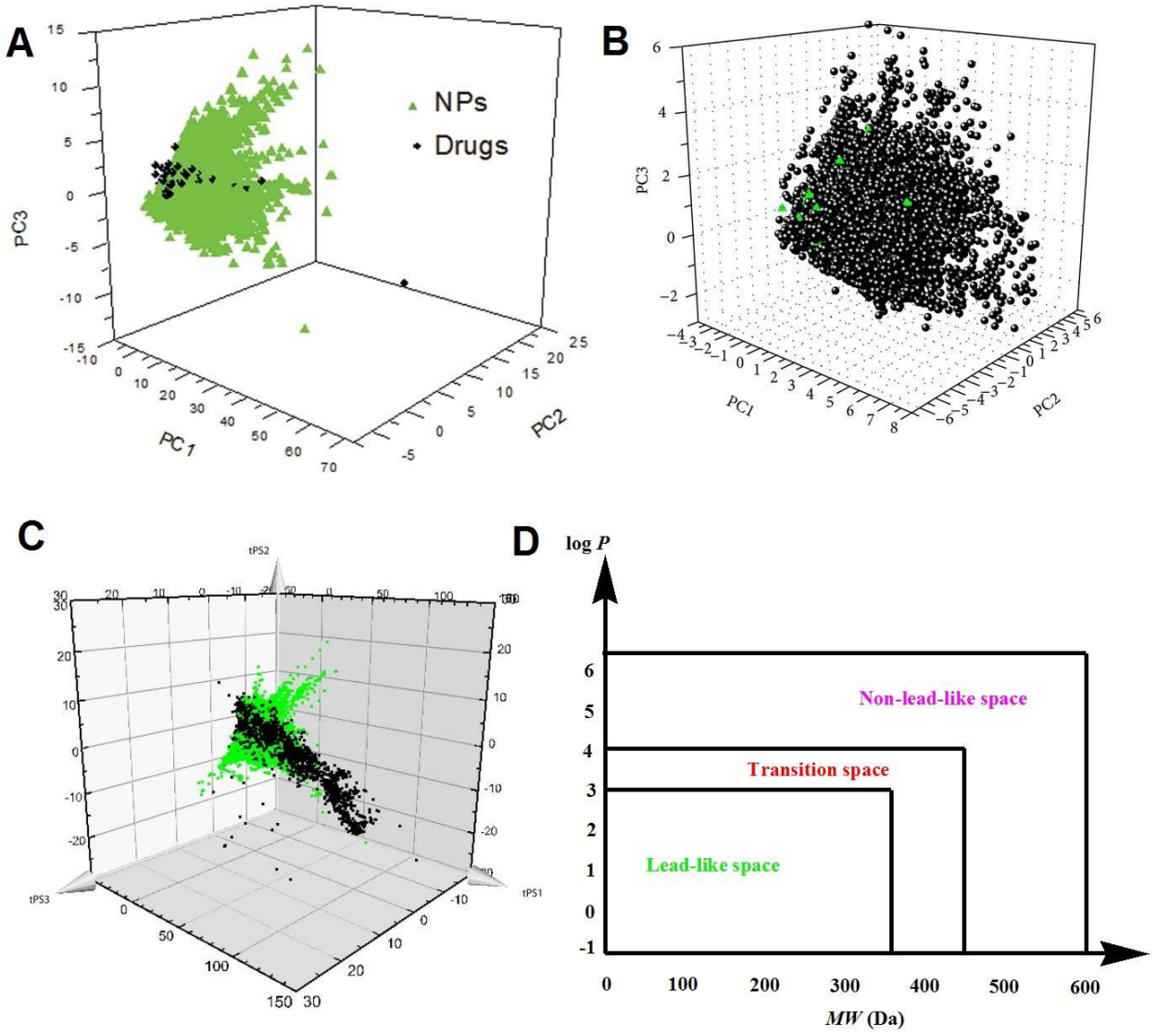

Figure 5: The distribution of biologically relevant chemical space of NPs, when compared with SDs: (A) PCA analysis of NPs in the Universal Natural Products Database (UNPD) and FDA-approved drugs. The green triangles and black dots represent natural products and FDA-approved drugs, respectively [30]; (B) PCA analysis of NPs contained in medicinal plants and 25 FDA-approved drugs for the treatment of type II diabetes mellitus (T2DM). The black dots and green triangles represent natural products and FDA-approved drugs, respectively [32]; (C) Predicted score (tPS) plots of NPs (in green) and bioactive medicinal chemistry compounds from the World of Molecular Bioactivity (WOMBAT) database (in black) [31]; (D) Property space representation for lead-like molecules of some selected chemical libraries [24, 35]. Figures reproduced by permission.

4.2 Antidiabetic medicinal plant-based bioactive natural products *versus* known antidiabetic drugs

In this study, the authors developed a docking score-weighted prediction model based on drug-target network in order to evaluate the efficacy of medicinal plants for the treatment of type II diabetes mellitus (T2DM). The docking dataset was composed of >208,000 medicinal plant-based NPs from retrieved from the UNPD *versus* drugs from DrugBank [40], which were FDA-approved for T2DM treatment. The both datasets were docked

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

against X-ray or NMR for each protein from RCSB protein databank (PDB) [41] which was related to T2DM pathogenesis, based on information of these proteins from KEGG Pathway database [42] and DrugBank. The binding free energy-based docking score (pK_i) was used to evaluate the affinity between each compound and each protein and compared with the experimental binding affinities of each of the FDA-approved T2DM drugs against their respective target proteins. It could be inferred most of the NPs would be drug-like. Besides, the wide distribution of the investigated NPs in chemical space (Figure 5B) showed that there would be vast structural and functional diversity. Moreover, the large overlap between NPs and the 25 FDA-approved small-molecule drugs for T2DM demonstrated that the NPs contained in the medicinal plants had a hopeful prospect for drug discovery for T2DM.

4.3 Analysis of the World of Molecular Bioactivity (WOMBAT) dataset against the Dictionary of Natural Products (DNP)

In an attempt to prove NPs to be a rich source of novel compound classes and new drugs, researchers from the working group of Tudor Oprea used the chemical space navigation tool ChemGPS-NP to evaluate the chemical space occupancy by NPs (from the DNP) and bioactive medicinal chemistry compounds from the World of Molecular Bioactivity (WOMBAT) database. Euclidean distances D_{pq} between points $P = (p_1, p_2,..., p_n)$ and $Q = (q_1, q_2,..., q_n)$ in Euclidean n-space, computed using ChemGPS-NP scores of the compounds, based on computed molecular descriptor were determined using the formula (Equation 1):

$$D_{pq} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots . (p_n - q_n)^2} \qquad (1)$$

It was observed that two sets differed in coverage of chemical space (Figure 5C). Besides, several "lead-like" NPs were found to cover regions of chemical space not present in WOMBAT. The authors also used property based similarity calculations to identify NP neighbors of approved drugs and showed from this method that several of the NPs exhibited the same activities as their drug neighbors in WOMBAT. It could be concluded that NPs could be identified via this method as useful lead compounds for drug discovery in searching for novel leads with unique properties. From Figure 5C, it could be clearly seen that NPs cover parts of chemical space not represented in medicinal chemistry compound space, showing that these areas of chemical space are yet to be investigated and which could be of interest in drug discovery.

5 The design of "drug-like" natural product libraries and implications for drug discovery

5.1 Strategy for designing a library with focused properties

Classical natural product drug discovery is only able to undertake drug-likeness analysis after the compounds are isolated and their structures elucidated. However, there are success stories using approaches that address front-loading of both extracts and subsequent fractions with desired physico-chemical properties prior to screening for drug discovery [24, 35]. If NPs are often referred to as 'sources of inspiration', it simply implies that 'lead-like' libraries could be designed, starting from NP scaffolds, with many examples available in the literature [24, 35, 43-46]. However, when an NP is used as the guiding structure for the creation of 'NP-like' libraries, controlling certain molecular descriptors (e.g. MW, clogP, etc.) during the synthetic process is of major importance for the generation of 'lead-like' libraries [24, 35]. This simply means preparing a RO5-compliant library can ensure the timely development of natural product lead compounds at a reasonable rate. The reader is invited to carefully read reference [24] for a summary of what to take most seriously when preparing a NP drug-like library.

5.2 Case study

NPs are known for containing fused medium-sized rings (Figure 6). In an attempt to mimic such NPs, Ventosa-Andrés et al. synthesized several molecular scaffolds containing medium-sized fused heterocycles using amino acids [46]. This is because amino acids are known to be useful building blocks used in natural reservoirs as well

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

as chemistry laboratories to create structural diversity. The authors employed a traditional Merrifield solid-phase peptide synthesis, and cyclization was carried out through acid-mediated tandem endocyclic *N*-acyliminium ion formation. The last steps were nucleophilic addition with internal nucleophiles. These led to seven-, eight-, and nine-membered ringed molecular scaffolds with newly generated stereogenic centers in most cases, using variety of heteroatoms contained in the bicycles, e.g. N, O, and S. The details of the synthetic strategy are beyond this discussion. The reader is invited to consulted the original paper for further details [46].

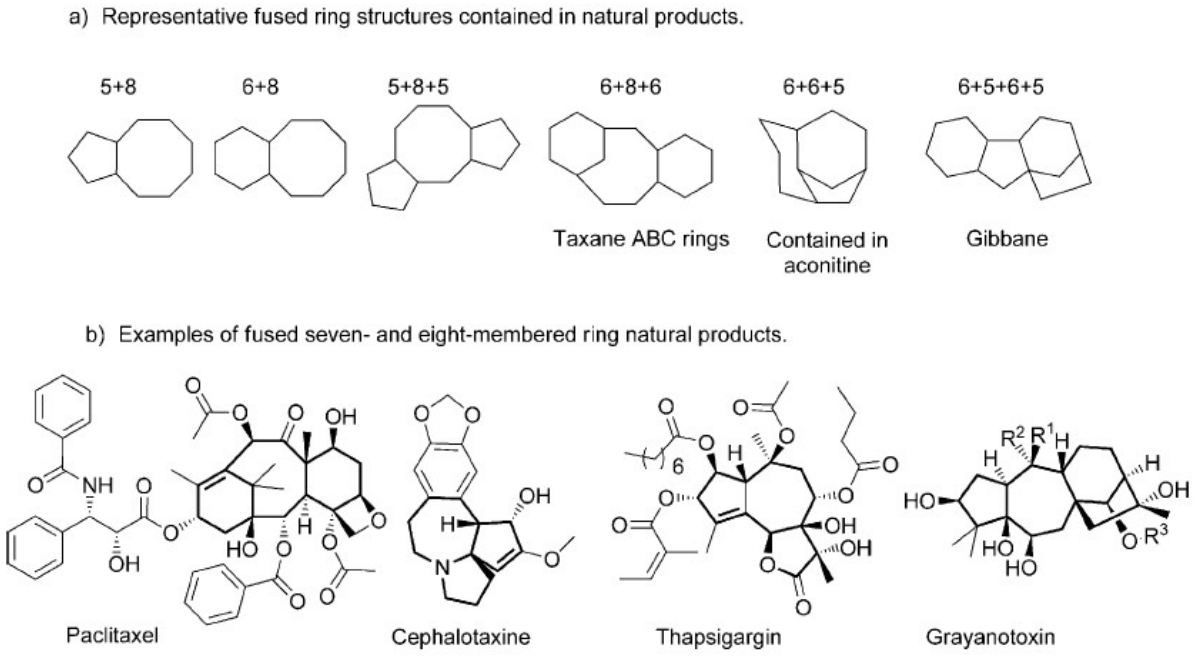

Figure 6: Fused structures and examples of natural products containing fused medium-sized rings. Figure reproduced by permission.

5.3 "Drug-likeness" prediction on available electronic natural products libraries for chemoinformatics analysis

An entire chapter on NP databases and datasets for virtual lead discovery is available in this collection [47]. Researchers within the Kirchmair group have also provided a recent analysis of available NP virtual and physical (vendor and academic) compound libraries which are highly useful for lead compound discovery [48, 49]. These include 25 virtual and 31 physical NP datasets employable for chemoinformatics projects, e.g. chemical space exploration, fragment based design, NP mimicking, and virtual screening. For each library, the authors provide detailed information on the extent of available structural information, and the overlap between the different datasets. From the analysis, it was observed that at least 10% of known NPs belong to the readily purchasable space (including small sized NPs for fragment-based design and macrocycles) and that with the renewed interest in NPs as lead structure, many more NPs and NP derivatives are being made available through on-demand sourcing, extraction and synthesis services.

5.3.1 Virtual libraries

Chen et al. recently described a large number of NP libraries, most of which can be freely accessed for chemoinformatics purposes towards lead compound discovery [48], further characterizing the chemical space thereof [49]. Most of these libraries were curated by academic groups based on literature information. Our recent analysis of these libraries showed ~1,500 unique compound entries, the major limitation being the absence of biological activities and compound sample accessibility information.

5.3.2 Physical collections, vendor libraries and their drug target space

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

This collection was done by Chen et al, by keying in the data vendor catalogues from compound suppliers and collections from academic groups [48]. With the goal of characterizing the chemical space of extent of coverage of chemical space by known and readily obtainable natural products and by individual natural product databases, the authors compiled comprehensive data sets of known and readily obtainable natural products from 18 virtual databases (including the Dictionary of Natural Products), 9 physical libraries, and the PDB [49]. After removing all sugars and sugar-like moieties, which are not of interest in drug discovery projects, the authors were able to show that the readily obtainable NPs are highly diverse and populate regions of chemical space that are of high relevance to drug discovery. In some cases, substantial differences in the coverage of natural product classes and chemical space by the individual databases are observed, while >2,000 NPs were found to be co-crystallized with at least one biomacromolecule in an X-ray crystal structure within the PDB.

6 Computational methods for estimating drug-likeness and ADME/T

It has been regrettably observed that many drugs often fail to enter the market due to poor pharmacokinetic (ADME/T) profiles [50]. This has necessitated the inclusion of pharmacokinetic considerations at earlier stages of drug discovery programs [51, 52]. However, due to the high cost of such experiments, the use of computer-based methods is often sufficient at early stages of lead discovery to save time and cost [53-55]. It requires, for example, less than 1 minute to screen 20,000 molecules in an in silico (computer-based) model, when compared with 20 weeks in the "wet" laboratory to do the same exercise [51]. In silico modeling of drug-likeness often employs standard filters that have been established using the accumulated ADME/T data in the late 1990s. Thus, many pharmaceutical companies now prefer computational models that, in some cases, are replacing the "wet" screens [51]. This has spurred up the development of several theoretical methods and software programs for ADMET prediction [56-59], even though some of the predictions could be disappointing [60]. Most software tools currently used for ADMET prediction make use of statistical models like quantitative structure-activity relationships (QSAR) modeling [60, 61] or knowledge-based methods [62-64]. A promising lead compound may, therefore, be defined as one which combines an interesting biological activity against a drug target (potency) with an attractive ADMET profile. This saves time and cost by discarding compounds with uninteresting predicted ADMET profiles from the list of potential drug candidates early enough, even if these prove to be highly potent. Otherwise, the DMPK properties are "fine-tuned" in order to improve their chances of making it to clinical trials [65]. Machine learning has now become very useful in the ADME/T profiling and drug-likeness prediction of compounds aimed at drug discovery [66].

7 Computational methods for estimating natural product-likeness

7.1 The natural product-likeness score

The concept of 'NP-likeness' has been around for about a decade [25]. It simply connotes the similarity of a molecule to the structure space covered by natural products, is a useful criterion in screening compound libraries and in designing new lead compounds.

Ertl et al. used a Bayesian measure which allows for the determination of how molecules are similar to the structural space covered by NPs. The NP-likeness score is an efficient approach to separate NPs from synthetic molecules (SMs). This score is very useful in virtual screening, prioritization of compound libraries toward NP-likeness, and the design of building blocks for the synthesis of 'NP-like' libraries [25]. The NP-likeness score (NP_{score}) in Equation 2 ranges from -5 to 5 and is computed for a whole molecule, as a sum of contributions of M fragments, f_i, (considered to be independent of each other, Equation 3) in the molecule, normalized relative to the molecule size:

$$NP_{score} = \sum_{i=0}^{M} f_i \qquad (2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

$$f_i = \log\left(\frac{A_i}{B_i} \cdot \frac{B_{tot}}{A_{tot}}\right) \quad (3)$$

where A_i is the number of NPs which contain fragment i, B_i is the number of SMs which contain fragment i, A_{tot} is the total number of NPs, and B_{tot} is the total number of SMs in the training set.

7.2 Implementations of the natural product-likeness score

The NP-likeness measure has now been implemented in several open-source, open-data tools, e.g. in a Taverna 2.2 workflow [67], which is available under Creative Commons Attribution-Share Alike 3.0 Unported License [68]. It is also available for download as an executable stand-alone java package under Academic Free License [69]. This scoring system can be used as a filter for metabolites in computer assisted structure elucidation or to select natural-product-like molecules from molecular libraries for the use as leads in drug discovery. A distribution of the scores for the training (synthetic molecules and natural products) and the test datasets have been shown in Figure 7.

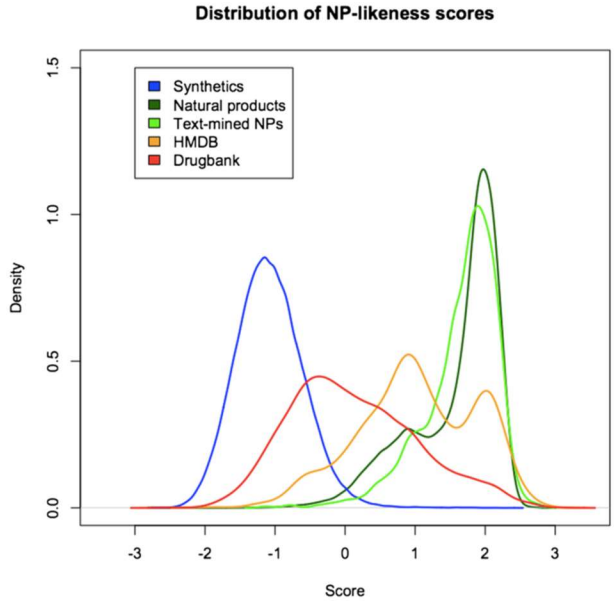

Figure 7: Distribution of NP-likeness score for the training (synthetic molecules and natural products) and the test datasets [66]. Figure reproduced by permission.

8 Machine learning methods to classify drugs from non-dugs

Attempts to define molecules likely to be drugs have been limited to simple numerical rules related to computed physico-chemical parameters, based on the RO5, which was derived following a statistical analysis of known drugs, e.g. 70% of the "drug-like" compounds are known to have 0 to 2 HBDs, 2 to 9 HBAs, 2 to 8 rotatable bonds, and 1 to 4 rings. Although such models are quite simple to implement and very fast to compute; by simply putting off molecules that fail two or more of the criteria, more sophisticated computational models of "drug-likeness" have been developed using machine learning techniques (e.g. neural networks or decision trees). Machine learning models begin with a training set of compounds with divergent properties, e.g. drugs and non-drugs. A number molecular descriptors are computed for each dataset. The model is then developed using the training set and its computed descriptors. Using a dataset of drugs from the WDI and a set of compounds from Available Chemicals Directory (ACD) with no known activities (regarded as non-drugs), and using a set of Ghose-Crippen atom type count descriptors, Sadowski and Kubinyi developed neural network model with 92 input nodes, 5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

hidden nodes and 1 output node to predict "drug-likeness" [70]. This model could correctly predict 77% of the WDI drugs and 83% of the ACD molecules as drugs and non-drugs, respectively. Similar results have been obtained using neural network [71, 72] and decision trees [73], using the same databases of drugs and non-drugs and the same set of descriptors. The performance of the decision tree model was comparable to that of the neural networks, correctly predicting ~83% of a validation set not included in the initial model.

9 A binary QSAR model to classify natural products from synthetic molecules

Researchers within the working group of Jurgen Bajorath were able to build a model in order to distinguish between NPs from the DNP and SDs from ACD based on Shannon entropy (SE) analysis [74]. The authors computed values of 98 descriptors from 2D representations of 199,420 ACD molecules and 116,364 NPs from the DNP, respectively. SE values were then defined as in Equation 4:

$$SE = -\sum p_i \ \log_2 p_i \tag{4}$$

where p is the probability of observing a particular descriptor value, computed from the number of compounds with a descriptor value that falls within a specific histogram bin, or "count" (c), for a specific data interval i. Thus, p is calculated as in Equation 5:

$$p_i = c_i / \sum c_i \tag{5}$$

The SE concept, initially employed in digital communication theory, is now popularly used in molecular descriptor analysis, since it is often combined with binary QSAR methodology to correlate structural features and properties of compounds with a binary formulation of biological activity (i.e., active or inactive). The authors adapted this approach to correlate molecular features with chemical source (i.e., natural or synthetic) by applying different combinations of such descriptors and variably distributed structural keys to the training sets of natural and synthetic molecules and used it to derive predictive binary QSAR models. The derived models were then applied to predict the source of compounds >80% prediction accuracy for the best models.

10 Conclusions

NPs have often been said not to abide by the RO5, as noted by Chris Lipinski himself [75], although about 60% of compounds from the DNP showed no violation of any of these "rules" [21]. In this chapter, we have navigated from simple rule-based approaches for determining what could likely be orally bioavailable of 'drug-like", 'lead-like' of 'natural product-like' for more advanced approaches like neural networks, decision trees and a combination of Shannon entropy and binary QSAR. We have shown that naturally-occurring compounds represent a significant proportion of known drugs and that the chemical space occupied by natural compounds is much wider than those of synthetic compounds and known drugs, implying that a large proportion of possible 'drug-like' space is yet to be investigated.

Table 2: List of abbreviations and definitions used in the text.

ACD	Available Chemicals Directory
ADME/T	Absorption, distribution, metabolism, excretion and toxicity
ChemGPS	Chemical global positioning system

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2018 doi:10.20944/preprints201811.0561.v1

DNP	Dictionary of Natural Products
Drug-likeness	The ability to resemble drugs
FDA	Food and Drug Administration
HTS	High-throughput screening
MW	Molecular weight
NOC	Naturally-occurring compound
NPs	Natural products
PCA	Principal component analysis
PDB	Protein databank
RO5	"Rule of 5"
SDs	Synthetic drugs
SE	Shannon entropy
SMs	Secondary metabolites
T2DM	Type II diabetes mellitus
UNPD	Universal Natural Products Database
WDI	World Drug Index
WOMBAT	World of Molecular Bioactivity

Acknowledgments

FNK acknowledges a return fellowship and an equipment subsidy from the Alexander von Humboldt Foundation, Germany. Financial support for this work is acknowledged from a ChemJets fellowship from the Ministry of Education, Youth and Sports of the Czech Republic awarded to FNK.

References

[1] Abegaz BM, Kinfe H. Secondary metabolites, their structural diversity, bioactivity, and ecological functions: an overview. Phys Sci Rev. 2018. DOI: 10.1515/psr-2018-0100.

[2] Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod. 2016;79:629–61.

[3] Dobson CM. Chemical space and biology. Nature. 2004;432:824–8.
[4] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997;23:3–25.
[5] Oprea TI, Davis AM, Teague SJ, Leeson PD. Is there a difference between leads and drugs? a historical perspective. J Chem Inf Comput Sci. 2001;41:1308–15.
[6] Hann MM, Oprea T. Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol. 2004;8:255–63.
[7] Feher M, Schmidt JM. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. J Chem Inf Comput Sci. 2003;43:218–27.
[8] Saldívar-González FI, B. Pilón-Jiménez BA, Medina-Franco JL. Chemical space of naturally occurring compounds. Phys Sci Rev. 2018. DOI: 10.1515/psr-2018-0103.
[9] Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015;14:111–29.
[10] Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. Nat Chem. 2016;8:531–41.
[11] Wani MC, Taylor HL, Wall ME, Coggon P, McPhail AT. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. J Am Chem Soc. 1971;93:2325–7.
[12] Walsh CT, Fischbach MA. Natural products version 2.0: Connecting genes to molecules. J Am Chem Soc. 2010;132:2469–93.
[13] Stratton CF, Newman DJ, Tan DS. Cheminformatic comparison of approved drugs from natural product versus synthetic origins. Bioorg Med Chem Lett. 2015;25:4802–7.
[14] Li JWH, Vederas JC. Drug discovery and natural products: end of an era or an endless frontier? Science. 2009;325:161–5.
[15] Pan L, Chai HB, Kinghorn AD. Discovery of new anticancer agents from higher plants. Front Biosci. (Schol. Ed.) 2013;4:142–56.
[16] Harvey AL. Natural products in drug discovery. Drug Discov Today. 2008;13:894–901.
[17] Koehn FE, Carter GT. The evolving role of natural products in drug discovery. Nat Rev Drug Discov. 2005;4:206–20.
[18] Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci USA. 2017;114:5601–6.
[19] Skinnider MA, Magarvey NA. Statistical reanalysis of natural products reveals increasing chemical diversity. Proc Natl Acad Sci USA. 2017;114:E6271–2.
[20] Grabowski K, Schneider G. Properties and architecture of drugs and natural products revisited. Curr Chem Biol. 2007;1:115–27.
[21] Quinn RJ, Carroll AR, Pham NB, Baron P, Palframan ME, Suraweera L, Pierens GK, Muresan S. Developing a drug-like natural product library. Nat Prod. 2008;71:464–8.
[22] Lee M-L, Schneider G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. J Comb Chem. 2001;3:284-9.
[23] Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK. Quantifying biogenic bias in screening libraries. Nat Chem Biol. 2009;5:479–483.
[24] Camp D, Davis RA, Campitelli M, Ebdon J, Quinn RJ. Drug-like properties: guiding principles for the design of natural product libraries. J Nat Prod. 2012;75:72–81.
[25] Ertl P, Roggo S, Schuffenhauer A. Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model. 2008;48:68–74.
[26] Ertl P, Schuffenhauer A. Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. Prog Drug Res. 2008;66:217, 219–35.
[27] Larsson J, Gottfries J, Muresan S, Backlund A. ChemGPS-NP: tuned for navigation in biologically relevant chemical space. J Nat Prod. 2007;70:789–94.
[28] Oprea TI, Gottfries J. Chemography: the art of navigating in chemical space. J Comb Chem. 2001;3:157–66.
[29] Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. PCA. In Multi- and MegaVariate Data Analysis; Umetrics Academy: Umeå, 2001; pp 43-69.
[30] Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and

network pharmacology. PLoS One. 2013;8:e62839.
[31] Rosén J, Gottfries J, Muresan S, Backlund A, Oprea TI. Novel chemical space exploration via natural products. J Med Chem. 2009;52:1953–62.
[32] Gu J, Chen L, Yuan G, Xu X. A drug-target network-based approach to evaluate the efficacy of medicinal plants for type II diabetes mellitus. Evid Based Complement Alternat Med. 2013;2013:203614.
[33] Lachance H, Wetzel S, Kumar K, Waldmann H. Charting, navigating, and populating natural product chemical space for drug discovery. J Med Chem. 2012;55:5989–6001.
[34] López-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL. Expanding the medicinally relevant chemical space with compound libraries. Drug Discov Today. 2012;17:718–26.
[35] Pascolutti M, Quinn RJ. Natural products as lead structures: chemical transformations to create lead-like libraries. Drug Discov Today. 2014;19:215–21.
[36] Reaxys, version 1.7.8; Elsevier; 2012; RRN 969209, Last revised: 9. January 2014
[37] Shen JH, Xu XY, Cheng F, Liu H, Luo XM, Shen J, et al. Virtual screening on natural products for discovering active compounds and target information. Curr Med Chem. 2003;10:2327–42.
[38] He M, Yan XJ, Zhou JJ, Xie GR. Traditional Chinese medicine database and application on the Web. J Chem Inf Comput Sci. 2001;41:273–7.
[39] Qiao XB, Hou TJ, Zhang W, Guo SL, Xu XJ. A 3D structure database of components from Chinese traditional medicinal herbs. J Chem Inf Comput Sci. 2002;42:481–9.
[40] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for "Omics" research on drugs. Nucleic Acids Res. 2011;39:D1035–41.
[41] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein databank. Nucleic Acids Res. 2000;28:235–42.
[42] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. Kegg for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14.
[43] Zuegg J, Cooper MA. Drug-likeness and increased hydrophobicity of commercially available compound libraries for drug screening. Curr Top Med Chem. 2012;12:1500–13.
[44] Shi BX, Chen FR, Sun X. Structure-based modelling, scoring, screening, and in vitro kinase assay of anesthetic pkc inhibitors against a natural medicine library. SAR QSAR Environ Res. 2017;28:151–63.
[45] Sánchez-Rodríguez A, Pérez-Castillo Y, Schürer SC, Nicolotti O, Mangiatordi GF, Borges F, et al. From flamingo dance to (desirable) drug discovery: a nature-inspired approach. Drug Discov Today. 2017;22:1489–1502.
[46] Ventosa-Andrés P, La-Venia A, Ripoll CA, Hradilová L, Krchňák V. Synthesis of nature-inspired medium-sized fused heterocycles from amino acids. Chemistry. 2015;21:13112–9.
[47] Koulouridi E, Valli M, Ntie-Kang F, Bolzani VS. A primer on natural product-based virtual screening. Phys Sci Rev. 2018. DOI: 10.1515/psr-2018-0107.
[48] Chen Y, de Bruyn Kops C, Kirchmair J. Data resources for the computer-guided discovery of bioactive natural products. J Chem Inf Model. 2017;57:2099–111.
[49] Chen Y, Garcia de Lomana M, Friedrich N-O, Kirchmair J. Characterization of the chemical space of known and readily obtainable natural products. J Chem Inf Model. 2018;58:1518–32.
[50] Darvas F, Keseru G, Papp A, Dormán G, Urge L, Krajcsi P. In silico and ex silico ADME approaches for drug discovery. Top Med Chem. 2002;2:1287–304.
[51] Hodgson J. ADMET – turning chemicals into drugs. Nat Biotechnol. 2001;19:722–6.
[52] Navia MA, Chaturvedi PR. Design principles for orally bioavailable drugs. Drug Discov Today. 1996;1:179–89.
[53] Lombardo F, Gifford E, Shalaeva MY. In silico ADME prediction: data, models, facts and myths. Mini Rev Med Chem. 2003;3:861–75.
[54] Gleeson MP, Hersey A, Hannongbua S. In-silico ADME models: a general assessment of their utility in drug discovery applications. Curr Top Med Chem. 2011;11:358–381.
[55] DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003;22:151–85.
[56] OCHEM - A platform for the creation of in silico ADME / Tox prediction models (http://www.eadmet.com/en/ochem.php)
[57] Meteor software, version 13.0.0, Lhasa Ltd, Leeds, UK, 2010.

 doi:10.20944/preprints201811.0561.v1

[58] QikProp software, version 3.4, Schrödinger, LLC, New York, NY, 2011.
[59] Cruciani G, Crivori P, Carrupt PA, Testa B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. J Mol Struc-Theochem. 2000;503:17–30.
[60] Tetko IV, Bruneau P, Mewes H-W, Rohrer DC, Poda GI. Can we estimate the accuracy of ADMET predictions? Drug Discov Today. 2006;11:700–7.
[61] Hansch C, Leo A, Mekapatia SB, Kurup A. QSAR and ADME. Bioorg Med Chem. 2004;12:3391–400.
[62] Greene N, Judson PN, Langowski JJ. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ Res. 1999;10:299–314.
[63] Button WG, Judson PN, Long A, Vessey JD. Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. J Chem Inf Comput Sci. 2003;43:1371–7.
[64] Cronin MTD. Computer-assisted prediction of drug toxicity and metabolism. In Modern Methods of Drug Discovery. Hillisch A, Hilgenfeld R. (Editors); Basel: Birkhäuser; 2003, pp. 259–278.
[65] Hou T, Wang J. Structure-ADME relationship: still a long way to go? Expert Opin Drug Metab Toxicol. 2008;4:759–70.
[66] Pires DEV, Blundell TL, Ascher DB. *pk*CSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem. 2015;58:4066–72.
[67] Jayaseelan KV, Steinbeck C, Moreno P, Truszkowski A, Ertl P. Natural product-likeness score revisited: an open-source, open-data implementation, BMC Bioinform. 2012;13:106.
[68] Available at http://www.myexperiment.org/packs/183.html Accessed on 20 November 2018.
[69] Available at http://sourceforge.net/projects/nplikeness/ Accessed on 20 November 2018.
[70] Sadowski J, Kubinyi H. A scoring scheme for discriminating between drugs and nondrugs. J Med Chem. 1998;41:3325–9.
[71] Ajay A, Walters PW, Murcko MA. Can we learn to distinguish between "drug-like" and "non drug-like" molecules? J Med Chem. 1998;41:3314-24.
[72] Frimurer TM, Bywater R, Naerum L, Lauritsen LN, Brunak S. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. J Chem Inf Comput Sci. 2000;40:1315–24.
[73] Wagener M, van Geerestein VJ. Potential drugs and nondrugs: prediction and identification of important structural features. J Chem Inf Comput Sci. 2000;40:280–92.
[74] Stahura FL, Godden JW, Xue L, Bajorath J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. J Chem Inf Comput Sci. 2000;40:1245-52
[75] Lipinski CA. Chris Lipinski discusses life and chemistry after the rule of five. Drug Discov Today. 2003;8:12–6.