Advanced Review

Understanding drug-likeness

Oleg Ursu,[1,2] Anwar Rayan,[3,4] Amiram Goldblum[3] and Tudor I. Oprea[1,2]*

Q1 'Drug-likeness', a qualitative property of chemicals assigned by experts committee vote, is widely integrated into the early stages of lead and drug discovery. Its conceptual evolution paralleled work related to Pfizer's 'rule of five' and lead-likeness, and is placed within this framework. The discrimination between 'drugs' (represented by a collection of pharmaceutically relevant small molecules, some of which are marketed drugs) and 'nondrugs' (typically, chemical reagents) is possible using a wide variety of statistical tools and chemical descriptor systems. Here we summarize 18 papers focused on drug-likeness, and provide a comprehensive overview of progress in the field. Tools that estimate drug-likeness are valuable in the early stages of lead discovery, and can be used to filter out compounds with undesirable properties from screening libraries and to prioritize hits from primary screens. As the goal is, most often, to develop orally available drugs, it is also useful to optimize drug-like pharmacokinetic properties. We examine tools that evaluate drug-likeness and some of their shortcomings, challenges facing these tools, and address the following issues: What is the definition of drug-likeness and how can it be utilized to reduce attrition rate in drug discovery? How difficult is it to distinguish drugs from nondrugs? Are nondrug datasets reliable? Can we estimate oral drug-likeness? We discuss a drug-like filter and recent advances in the prediction of oral drug-likeness. The heuristic aspect of drug-likeness is also addressed. *WIREs Comput Mol Sci* 2011 00 1–22 DOI: 10.1002/wcms.52

INTRODUCTION

The process of small molecule drug discovery is currently confronted with multiple difficulties at the societal, economic, and scientific levels. The industry faces a decade-long innovation deficit[1] which is compounded by an overall decline in the introduction of truly novel drugs. At the same time, more academic and nonprofit research groups worldwide are showing signs of increased pace in the arena of drug discovery, translational research, and drug repurposing. Against this backdrop, the issue of what truly constitutes a drug is no longer a philosophical one, but one of practical and immediate consequences. From a healthcare practitioner's standpoint, drugs (medicines) are well-defined entities that lead to clinical consequences: upon intake, these substances alter symptoms, kill microorganisms, balance metabolism, hormones or electrolytes, etc., with the purpose of restoring health or improving the subject's well-being. However, drugs are an ill-defined entity from a chemical standpoint. It is understood that, besides affinity to the intended target(s), therapeutic drugs must observe certain properties related to bioavailability, acute and chronic toxicity, mutagenicity, efficacy, etc. In contrast to other manufacturing industries that produce physical deliverables based on research and development, the biopharmaceutical industry does not have a well-defined understanding of what the end product looks like.

'Drug' (in the regulatory sense) is not an intrinsic property of chemicals.[2] Despite the wealth of literature on drug-likeness, discussed below, the quality of a drug is attributed by a regulatory body such as the US Food and Drug Administration (FDA), based on scientific evidence available at the time of the New Drug Application (NDA). In each country, committees weigh in all available evidence, compare the submission with existing therapies and drugs, and

*Correspondence to: toprea@salud.unm.edu

Q2 [1]Division of Biocomputing, Department of Biochemistry and Molecular Biology, School of Medicine, University of New Mexico, Albuquerque, NM, USA

[2]UNM Center for Molecular Discovery, School of Medicine, University of New Mexico, Albuquerque, NM, USA

[3]Molecular Modeling and Drug Design Lab and the Alex Grass Center for Drug Design and Synthesis, Institute of Drug Research, The Hebrew University of Jerusalem, Jerusalem, Israel

[4]Drug Discovery Informatics Lab, QRC-Qasemi Research Center, Al-Qasemi Academic College, Baqa-El-Gharbia, Israel

DOI: 10.1002/wcms.52

decide to approve, and sometimes withdraw, the 'drug' quality from the molecule on the basis of efficacy, safety, and cost-benefit. Therefore, despite the ubiquitous use of 'drug' as a quality to be studied with machine learning tools, 'drug' is not a natural property of chemicals. As the 'drug' label is time-dependent, our understanding of drug-likeness is a heuristic process which is likely to remain limited to the 'soft' arena of mathematical modeling.[3]

Furthermore, it is outside the scope of this paper to address the relationship between drugs and dosage. The father of toxicology, Theophrastus Bombastus Paracelsus, wrote: *Alle Ding' sind Gift, und nichts ohn' Gift; allein die Dosis macht, daß ein Ding kein Gift ist* (All things are poison and nothing is without poison, only the dose permits something not to be poisonous) [http://en.wikipedia.org/wiki/Paracelsus]. That is to say, all drugs can act as poisons when overdosed. In this review, we address our current level of understanding the 'drug-like' character of small molecules, and discuss ways in which machine learning tools have been used to model it.

MAPPING THE CHEMICAL SPACE OF BIOACTIVE SMALL MOLECULES

The need to identify drug-like molecules is rooted in our inability, so far, to populate a complete map of the chemical space of small molecules (CSSM). A complete CSSM will allow for identification and mapping of specific regions of CSSM with drug-likeness properties, such mapping will eliminate the need of filtering/modeling high-throughput screening (HTS) libraries for drug-like chemicals by simple determination of overlap between pockets of drug-likeness in CSSM and chemical libraries. According to Weininger,[4] all possible derivatives of n-hexane, starting from a list of 150 substituents, when completely enumerated, can lead to over 10^{29} structures (Box 1). CSSM is limited at one end by low molecular weight (MW) (see Box 1). In a thought experiment, the largest collection of physically existing molecules, collected from all chemical (private, public, and commercial) collections worldwide is estimated to be near 120 million.[5] The collections listed in Box 1 are likely to undersample[5] the space of molecules with MW > 300, which substantiates the limitations of systematic CSSM mapping.

To date, CSSM exploration has been systematically performed for molecules containing C, N, O, S, and Cl up to 13 atoms, based on estimated chemical stability and synthetic feasibility rules, which led to nearly one billion potential chemicals,[6] stored in the GDB-13 database. In a different approach, all possible CSSM scaffold topologies, which are mathematical representations of ring structures, were exhaustively enumerated for up to eight rings, resulting in 1,547,689 distinct scaffolds.[7] Of these, only 0.61% (9747 unique topologies) were mapped to the known CSSM, represented by 52 million compounds from eight different collections embodying drugs, natural products, chemogenomics compounds, environmental toxicants, and virtual molecules.[8] Although a significant area of biologically relevant chemical space is occupied by natural products, i.e., chemical entities produced by living organisms, the issue of mapping natural products in chemical space has been addressed elsewhere.[9] Chemical biologists, natural product scientists as well as drug hunters continue to seek CSSM pockets that contain biologically relevant and 'drug-like' compounds. Within this context, current methods to probe pharmaceutically relevant CSSM, in particular, for probe, lead, and drug discovery,[10] are placed under scrutiny with the goal of maximizing efficiency.

Further complexity to this process is added by factors external to the area of CSSM exploration such as intellectual property, increased costs in clinical research, and marketing-driven decisions, all of which are critical to the decision-making process. The intellectual property portfolio and proprietary information, a stronghold by which most companies attract investors, forces pharmaceutical companies to act as competitors, and to make important decisions

BOX 1: THE CHEMICAL SPACE OF SMALL MOLECULES IN NUMBERS

Monosubstituents and up to 14-substituted hexanes: over 10^{29} structures (4).
'Tangible' small molecules of pharmaceutical interest: over 120 million structures (5).
GDB-13 database (http://www.dcb-server.unibe.ch/groups/reymond/gdb/home.html): 977 468 314 chemicals (6).
Scaffold topologies for up to eight rings (http://topology.health.unm.edu/): 1,547,689 unique scaffolds (7).
PubChem Compounds: 26,384,357 molecules (*).
PubChem, Ro5 compliant: 18,504,698 molecules (*).
PubChem, Ro5 compliant, biologically tested: 11,937,428 molecules (*).
MDDR (http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp): over 150,000 compounds.
Launched drugs: over 4000 (estimated).
(*) Source: PubChem website (http://pubchem.ncbi.nlm.nih.gov/) as of March 24, 2010.

based on often incomplete data. It is quite rare that pharmaceutical companies share information critical to the discovery process, in particular, to CSSM mapping. One such issue relates to Pfizer's Rule of five (Ro5), and to the lead-like concept proposed by AstraZeneca. Both concepts are relevant to explain the evolution of understanding drug-likeness, and are described below.

Pfizer's Rule of Five

It has been observed that, given a choice among alternate drug administration methods, patients prefer orally formulated drugs (e.g., tablets, capsules). For practical reasons, oral formulations can rarely exceed 1 gm in quantity, which places additional burdens related to therapeutic efficacy, frequency of administration, and dosage, not to mention the large array of pharmacokinetic (PK) and toxicological properties. Indeed, for the most part, all pharmaceutical research seeks to develop orally bioavailable drugs (OBDs), to be administered in single dose, with no side effects (if possible) and low dosage (e.g., not exceeding 300 mg/day). Within —three to five years of introducing high-throughput combinatorial chemistry and screening technologies, the pharmaceutical industry was reminded about the importance of restricting small molecule synthesis in the property space defined by log *P*, the logarithm of the octanol–water partition coefficient,[11] MW, the number of hydrogen bond donors (HDO), and acceptors (HAC) in a seminal paper by Lipinski et al.[12]

This paper was based on the post-HTS analysis of the early (1994–1996) results of HTS and combinatorial chemistry at Pfizer, where most of the hits were large (high MW) and hydrophobic (high log *P*), which made their progression from hit to lead significantly more difficult. They analyzed 2245 compounds from the World Drug Index (WDI)[13] that had reached phase II clinical trials or higher, and looked at the 90th percentile for the distribution of MW ($\leq$500), log *P* ($\leq$5), HDO ($\leq$5) and the sum of nitrogens and oxygens, accounting for HACs (HAC $\leq$ 10). In this paper, log *P* was estimated using the ClogP software from Biobyte Corporation (http://www.biobyte.com/). Natural products and actively transported molecules were excluded from the Ro5 analysis. If any two of the above conditions are violated, the molecule was less likely to result in an orally active drug. This work enhanced the awareness of the medicinal chemistry community regarding the existence of a specific area of chemical space that potentially restricts the properties of orally available chemicals. More importantly, it significantly changed our perception regarding drug properties and chemical space exploration. Before long, most library design programs, based on combinatorial chemistry or compound acquisition,[14] included Ro5 filters. To date, the Ro5 is the only computationally derived filter that is unanimously recognized by pharmaceutical executives, academic and industrial medicinal chemists and combinatorial chemists, and by drug discovery practitioners alike.

In fact, Ro5 compliance was (and still is; this paper has been cited more than 2000 times to date) often considered to be important to label a molecule as 'drug-like', despite the restriction of Ro5 criteria to the issue of oral availability via passive transport only. Upon removal of the approximately 5000 drugs from available chemicals directory (ACD), a chemical catalog of reagents representing 'nondrugs', the remaining ~190,000 chemicals are almost as likely to pass the Ro5 criteria as a set of 400 orally available drugs from the Physician Desk Reference, and a set of nearly 80,000 pharmaceutically useful compounds indexed in the MDL Drug Data Report (MDDR).[15] In other words, the distribution of the four properties captured by the Ro5 step function does not differ significantly between MDDR, ~~PDR,~~ and ACD. Attempts to incorporate the four Ro5 properties in machine learning efforts distinguishing drugs from nondrugs[16] did not lead to improved discrimination and, indeed, were no better than random. However, optimizing the ranges of the four Ro5 properties improves the prediction of oral drug-like molecules, but not the best selection of discriminative properties.[17]

The Lead-Like Concept

Within the scope of drug discovery, the initial (still current) purpose of assembling chemical libraries for screening was to identify hits and perhaps leads, but rarely (if ever) directly *drugs*. Following the introduction of Pfizer Ro5 criteria, pharmaceutical drug discovery projects made frequent use of these filters *ad litteram*, i.e., MW $\leq$ 500 and Clog *P* $\leq$ 5, despite the fact that these values had been obtained from analyzing drugs, not leads. Many HTS campaigns yielded (and continue to yield) mostly *micromolar* hits. When filtered with the drug-derived Ro5 criteria, these HTS hits are not easily amenable to traditional medicinal chemistry lead optimization, since postoptimization they are likely to fall outside Ro5 space.[18] From an initial set of 18 lead–drug pairs, we further suggested[18] that lead-like libraries should be designed with lower MW ($\leq$300) and lower log *P* ($\leq$3.0) cutoffs, and questioned combinatorial technologies that concatenate several monomers using multicomponent reactions or split-and-mix protocols.

An analysis[19] based on 470 lead–drug pairs extracted from Walter Sneader's book[20] found that leads have property profiles which are 'left shifted' when compared to the resulting drugs. Their profiles included MW and Clog *P* (also discussed above), as well as the number of aromatic rings, Andrews' binding energy,[21] and the number of Bits set in the 1024 Daylight fingerprint; the latter was used as a measure of a molecule's internal complexity. On average, leads were found to have lower MW, lower Clog *P*, fewer aromatic rings, fewer HAC, lower Andrews' binding energy, and less bits set, compared to their drug counterparts. The authors further questioned combinatorial library design philosophies and the propensity of such libraries to yield overly complex molecules.

An extension of earlier work[18] was based on 96 lead–drug pairs,[22] and examined MW, Clog *P*, HDO, and HAC, log D_{74} (the calculated logarithm of the octanol/water distribution coefficient at pH 7.4), the number of rings, RNG, the number of nonterminal rotatable bonds, RTB, as well as two drug-like scores (DLS).[15] In agreement with Hann et al.,[19] we found that lead structures exhibit, on the average, lower complexity (lower MW, lower RNG, and RTB), less hydrophobicity (lower Clog *P* and log D_{74}) and lower DLS.[22] One year later, Proudfoot found even lower differences between leads and drugs,[23] based on set of 25 lead–drug pairs launched in 2000. This set contained four enantiopure forms (esomeprazole, perospirone, dexmedetomidine, and levobupivacaine) of previously launched racemic drugs (which now served as leads), five drugs where the change was either an addition or a rearrangement of a methyl or ethyl group, as well as five other compounds where minor chemical alterations in a single region of the molecule was introduced. Thus, the differences found by Proudfoot between leads and drugs launched in 2000 were even smaller than previously reported. The most extensive dataset, based on 385 documented leads and 1651 launched drugs, included thousands of Phases I, II, and III clinical trial drug candidates, as well as tens of thousands of literature-reported bioactive (or inactive) molecules.[10] This study reiterated earlier findings that, on average, leads are significantly smaller, more soluble, less hydrophobic, less flexible, and less complex than any other subset; and that MW and complexity,[24] as well as estimated aqueous solubility, dropped as one progressed from high-activity molecules ($N = 5784$ with at least one reported nanomolar affinity) to Phase I ($N = 801$), Phase II ($N = 1047$), Phase III ($N = 301$) and drugs (with leads at the lowest median value). Furthermore, the property profile of chemical probes ($N = 198$) was closer to that of leads, as opposed to other categories.

Perola[25] examined binding affinity and ligand binding efficiency (LBE) for 60 lead–drug pairs, and observed that, on average, p*K*i(drug) $\gg$ p*K*i(lead) yet Clog *P*(drug) = Clog *P*(lead). Thus, increased lipophilic ligand efficiency is one of the recurring trends of successful drug discovery programs. Maintaining a particular lipophilicity profile whilst increasing MW remains one of the keys to successful lead optimization. Most drug discovery programs often retain the lead scaffold, find ways to boost affinity via charge-charge interactions and identify the most efficient fragments of lead structures. Sometimes binding efficiency is reduced in order to improve other properties (e.g., solubility, plasma protein binding) in the late stages of a lead optimization program.

Although based on a small statistical sample (under 500 leads), the concept of lead-likeness appears to remain central in early drug discovery.[26] During its century-old history, the pharmaceutical industry has done a poor job in documenting the actual decision process, e.g., why certain chemical steps and moieties were embodied into particular compounds. As we seek to better understand what drugs are, the issue of what constitutes a good lead turns out to be a particularly relevant question, as the pressure is mounting to increase productivity, reduce costs and efforts, and to deliver high-quality candidate drugs. Perhaps best stated by DeStevens[27] in 1986, structured management does not work in the context of preclinical drug discovery, an observation that remains true to date.[28,29] A comprehensive analysis of advanced leads identified in the decade prior to 2009, traced back property differences to the nature of HTS hits and hit-to-lead optimization practices, further suggesting that organizational adjustments are required in order to reduce the attrition rate of clinical candidates.[30]

It appears more and more difficult to attribute the 'lead' label to a compound, unless one is completely confident that the molecule in question exhibits high activity on the target(s) in question, a good degree of selectivity against other targets and antitargets,[31] chemical features amenable for optimization, is part of a well-established structure–activity relationship (SAR) series, it shows favorable patent situation as well as good pharmacokinetic and toxicological profile. The progression HTS hits $\rightarrow$ HTS actives $\rightarrow$ lead series $\rightarrow$ drug candidate $\rightarrow$ launched drug has, in the last decade, shifted the focus from good quality candidate drugs to good quality leads.[32] Logically, improvements are also conducted

in the area of HTS actives selection,[33] as well as chemical library design.[14] Besides property profiling, the one common denominator in this arena has been 'drug-likeness' evaluation.

DRUG VERSUS NONDRUG DISCRIMINATION

Efforts to characterize drug-likeness of chemical substances were undertaking before publication of Pfizer's Rof5 by Gillet et al.,[34] where the authors tried to discriminate between drug-like substances (WDI) and non-drug-like substances (SPRESI) using property filters, substructural analysis, and genetic algorithms. The importance of restricting HTS actives selection to the property profile, defined by Pfizer's Ro5, was quickly implemented in many library design programs. Within less than a year after Lipinski's and Gillet's work, two groups attempted to define drug-like space based on large (tens of thousands) sets of small molecules: drugs, and nondrugs,[35,36] being the first to compute DLS. These scores offer the ability to discriminate 'drugs', in this case represented by WDI[13] and MDDR,[37] from 'nondrugs', here by ACD.[38] Rather than focusing on what is a drug (societal issue), the need to evaluate HTS hits/active and leads forced discovery scientists to address directly a molecule's probability for becoming a drug; this was regarded as an 'in silico', not a human problem. Our ability to distinguish between 'drugs' and 'nondrugs' by means of machine learning can be used to analyze large sets of molecules and to prioritize them for synthesis, for biological screening, and for in-depth evaluation. Thus, the development of DLS schemes can impact the resources and time required to succeed in preclinical drug discovery, whether by means of large-scale high-throughput synthesis and screening, or by evaluating (small) focused chemical libraries. DLS schemes could be used to construct molecular libraries on the basis of scaffolds and to increase the efficiency of virtual screening.[39,40] Such analyses should preferably specify the extent of fitness of a molecule to be a drug, and should also be associated with particular molecular features, in order to enable subsequent design of molecules.

Posing the Question

Given its complex nature, initial tools aimed at discriminating drugs from nondrugs were rooted in neural networks. These methods[35,36] utilized the training set of 'drugs', i.e., compounds of pharmaceutical relevance indexed in MDDR and WDI, from 'non-drugs', i.e., compounds generally regarded as reagents of little therapeutic relevance, in order to condition the network to identify similar compounds. In one study, a Bayesian neural network used the Comprehensive Medicinal Chemistry (CMC) dataset (5500 molecules) for learning, and MDDR (~80,000 entries) for testing, as surrogate for 'drugs', whereas ACD served as surrogate for 'nondrugs'. Over 90% of CMC, but only 10% of ACD compounds were classified as drug-like; the model correctly estimated 80% of MDDR (external set) as drug-like, as well.[35] In another study, a feed-forward neural network was trained on 169,331 'nondrugs' (ACD), and 38,416 'drugs' (WDI), with 83% (ACD) and 77% (WDI) classification accuracy, respectively.[36] This influential work stemmed a wide number of reports, summarized in Table 1, which describe successful attempts to discriminate drugs from nondrugs.

Such successful machine learning models and molecular descriptors (Table 1), as well as molecular property and functional groups filters have been applied on multiple datasets (MDDR, WDI, CMC, ACD, etc.), in order to optimize DLS and filters. Based on this literature survey, machine learning methods appear to outperform filter-based methods by a margin of 10–20%. However, when confronted with a large number of input variables and large training sets, machine learning tools are prone to overtraining and overfitting,[41,42,43] which in turn is likely to lead to a loss of external predictive ability (i.e., outside the validation sets). The drug-like scoring schemes from Table 1 that are based on machine learning have a prediction accuracy that ranges from 75% to 90%. These models often use more than half of the available data for training and the number of descriptors often exceeds 100. Our own results[44] suggest that almost 40% of ACD (after removing drug structures) contains chemicals with a relatively high content of drug-like fragments, i.e., as much as two-fifths of the chemicals in the ACD catalog bear a reasonable degree of similarity to drugs. Thus, machine learning models input a significant number of ACD structures that bear some similarity to drugs under the 'nondrug' label, when it would be best if these were ignored. This, in our opinion, compels machine learning models to utilize ill-defined (noisy) data, which in turn leads to apparently successful models at the cost of 'memorizing' data, overfitting, and reduced performance on external (blind prediction) sets. To assess the true predictive power of such models, *independent validation sets* for drugs/nondrugs [e.g., critical assessment of protein structure prediction (CASP) style (http://predictioncenter.org)] would be appropriate. However, such an independent evaluation has

TABLE 1 | Comparative Overview of Drug-Likeness Studies in Peer-Reviewed Literature. Since Citing each Descriptor System and Prediction Method would have Significantly Increased the Complexity of this Table, the Reader is Kindly Invited to Consult the Original Reference

Data Sources for Drugs	Data Sources for Nondrugs	Test Sets	Prediction Methods	Results	Comments	Reference
CMC (6522) + MDDR (5182) total and used for training	–	(a) Top selling drugs (1997)—79	Multilevel grouping analysis	(a)–60 drug compatible	Multilevel grouping analysis are filters based on atom environments (up to four atoms) statistics	45
		(b) Compounds under biological testing from MDDR (68,017)		(b)–27.4% drug compatible		
		(c) Anticancer drugs from CMC (461)		(c)–19.1% drug compatible		
		(d) Reactive and toxic compounds from ACD (57)		(d)–0 drug compatible		
WDI (38,416) total	ACD (169,331)	Test sets WDI (10,000) and ACD (10,000)	Decision trees	82.6% accuracy on validation set with no penalization (for misclassified drugs) and 91.9% with penalization with false positive rate for later of 34.3%	Descriptor set used is Ghose and Crippen atom types for log *P* calculation	46
5000 used for training	5000 used for training	Validation sets WDI (23416) and ACD (154,331)				
3000 MDDR (classified as real drugs) used for training set	70,000 ACD used for training set	Test sets MDDR (1400) and ACD (20000)	Neural networks	88% predicted correctly for ACD and MDDR	Descriptor set used is CONCORD atom types	16
68,523 filtered MDDR	150,310 filtered ACD	–	Properties filters	62.68% from ACD pass the filters	Descriptor set used is number of rings, rigid bonds, rotatable bond, Clog *P*, MW, HBA, HBD	15
				61.23% from MDDR pass the filters		
4836 from CMC		250,282 from ACD	Properties filter	28.3% from ACD are not drug-like	Descriptors used are based on frequencies of building blocks constituents of molecules	47

(*Continued*)

TABLE 1 | Continued

Data Sources for Drugs	Data Sources for Nondrugs	Test Sets	Prediction Methods	Results	Comments	Reference
6000 compounds from MDDR	6000 compounds from ACD	(a) 7170 compounds from CMC	PLS Discriminant Analysis	(a) 78/73% scored as drug-like using DFP/PPF	Daylight 4096 bit fingerprints (DFP) and 240 bit property and pharmacophore fingerprint (PPF)	48
		(b) 420 compounds from NCE		(b) 88/81% scored as drug-like using DFP/PPF		
		(c) 412 compounds from PDR		(c) 83/72% scored as drug-like using DFP/PPF		
		(d) 15,557 compounds from FALERT		(d) 83/81% scored as drug-like using DFP/PPF		
		(e) 83,405 compounds from MDDR		(e) 88/88% scored as drug-like using DFP/PPF		
		(f) 209,978 compounds from ACD		(f) 22/43% scored as non-drug-like using DFP/PPF		
5000 compounds from WDI	5000 compounds from ACD	(a) Launched, registered and under investigation (LRI) compounds from Cipsline database (6148)	PASS	(a) 73.4% predicted as drugs	Multilevel neighborhoods of atoms were used as descriptors	49
		(b) Launched, and registered (LR) compounds from Cipsline database (864)		(b) 78.5% predicted as drugs		
		(c) 9484 commercially available compounds as nondrugs test set		(c) 83.8% predicted as nondrugs		
		(d) Top 100 prescription pharmaceuticals (88 compounds)		(d) 87.5% predicted as drugs		
1322 from MDDR + 2617 from CMC	155,402 from ACD	78,028 from MDDR	Rules based on pharmacophore point filter	65.9% from MDDR pass the filters	Pharmacophore points defined on the following functional groups: amine, amide, alcohol, ketone, sulfone, sulfonamide, carboxylic acid, carbamate, guanidine, amidine, urea, and ester	50

(*Continued*)

TABLE 1 | Continued

Data Sources for Drugs	Data Sources for Nondrugs	Test Sets	Prediction Methods	Results	Comments	Reference
		4708 from CMC		60.6% from CMC pass the filter 36.5 from ACD pass the filter		
2105 drugs from WDI	52,712 from Maybridge library	42131 compounds from WDI with 150 atoms or less	Kohonen artificial neural net	73.5% from WDI predicted as drugs 20.7% from Maybridge predicted as drugs	AM1 derived descriptors encoding size, shape, and electrostatics	51
2417 from corporate database labeled with various drug-likeness scores	1563 from corporate database	1617 from corporate database for drugs	Artificial neural nets and Support Vector Machines	87.6% of compounds in test sets correctly classified	Descriptors used MACCS key, Crippen atom type, MOE 2D descriptors	52
Training set 800	Training set 800	763 from corporate database for nondrugs				
15,000 compounds from Ensemble database	15,000 compounds from Sigma-Aldrich catalog	Test set for drugs 3751 and for nondrugs 3749	Neural networks and Support Vector Machines	Neural networks 78.33% accuracy for drugs and 63.51% for nondrugs	Descriptors from ChemoSoft	53
Training set 7465/ Validation set 3755	Training set 7535/ Validation set 3744			Support Vector Machines 72.19% for drugs and 78.10% for nondrugs		
87,266 from MDDR + 6678 from CMC	293,487 from ACD	Chinese Natural Product Database	Properties filter	72.91% predicted as drug-like	Descriptors used are constitutional: number of sp3 carbons, nitrogen, oxygen, aromatic carbon, aromatics N, S, double bonds, triple bonds, aromatic bonds, rigid bonds, rings, etc.	54
34,549 compounds from WDI split into training/ validation/test	151,752 compounds from ACD split into training/ validation/test	10700 from WDI and ACD	Support Vector Machines	7.1% error rate for polynomial kernel and 6.9% error rate for RBF-kernel	Descriptor set used is Ghose and Crippen log *P* atom types	55

(*Continued*)

TABLE 1 | Continued

Data Sources for Drugs	Data Sources for Nondrugs	Test Sets	Prediction Methods	Results	Comments	Reference
43,185 compounds from WDI Training set 38581	307,624 compounds from ACD Training set 303,020	9208 compounds from WDI and ACD	Support Vector Machines	92.73% accuracy with RBF-kernel and 89.74% with linear kernel	Descriptors set used Pipeline Pilot ECFP_4	56
3117 compounds from Merck Index, G. Milne's compilation of drugs, etc.	2238 compounds from unknown sources	51 pharmaceutical compounds from journal 'Drugs of the feature' 01/2005–12/2006	Decision trees	76% of all nondrugs filtered out first step, applying more descriptors to succeeding step increases the performance to 92% of all nondrugs filtered out while less than 19% of drugs are lost	Descriptor set used MW, Xlog *P*, molar refractivity, SMARTS keys, AM1 quantum descriptors	57
KEGG Drug Database of approved pharmaceutical in USA and Japan (5294)	NatDiverse collection from Analyticon Discovery (17,402)	20% of compounds from screening libraries	Decision trees	91% of HitFinder compounds are classified correctly	Molecular Structure Generator Program molecular descriptions	58
92% of pharmaceutical were used for training and validation	HitFinder collection from MayBridge (14,400)	8% of compounds from approved pharmaceuticals				
	80% of compounds in the libraries were used for training and validation			99% of NatDiverse compounds are classified correctly		
Launched drugs database from GVK BioSciences (3767)		(a) Clinical candidates library from GVK (44,843) (b) Commercially available compounds from ZINC database (44,140) (c) AnalytiCon database (27,376) (d) IBS database (425,148) (e) Enamine database (1,316,159)	Statistical correlation coeffi-cients	DLS, lower values indicate closer to launched drugs dataset (a) 0.08 (b) 0.20 (c) 0.09 (d) 0.20	MW, log *P*, HBA, HBD, RTB, PSA	59

(*Continued*)

TABLE 1 | Continued

Data Sources for Drugs	Data Sources for Nondrugs	Test Sets	Prediction Methods	Results	Comments	Reference
		(f) ChemDiv database (662,630) (g) ASINEX database (399,583) (h) Vitas-M database (487,261) (i) Maybridge database (56,824) (j) ChemBridge database (422,087) (k) Bionet database (4233)		(e) 0.21 (f) 0.22 (g) 0.23 (h) 0.24 (i) 0.25 (j) 0.25 (k) 0.26		
CNS drugs (119)	CNS candidates (108) Pfizer diversity set (11,303)		Multiparameter optimization	77% of CNS drugs are aligned with ADME properties 54% CNS candidates aligned with ADME properties 49% of diversity set are aligned with ADME properties	Clog *P*, ClogD, MW, PSA, HBD, pKa	60

not been attempted for drug-likeness, quite likely because freely accessible and widely accepted validation datasets are simply nonexistent.

Property Characterization of Drugs—the Other Side of Drug-likeness

Another arena of 'drug-likeness' deals with the interplay between solubility,[61] permeability and other properties[62] has been recognized as one of the major areas of research in drug discovery. Leeson and Springthorpe examined the trends of common properties of drugs that have been used to model drug-likeness [MW, Clog *P*, polar surface area (PSA), HAC, HDO, RTB, RNG, etc.] for a set of 2118 launched drugs.[26] These were compared with 431 compounds in development (preregistration and Phases I–III), representing work from the top 25 pharmaceutical companies (extracted from the Prous Science Integrity database), as well as for 117,148 patented compounds from the 2001–2007 timeline, from four companies (AstraZeneca, GlaxoSmithKline, Merck, and Pfizer).[26] The trend for the oral drugs approved from 1983 to 2007 appears to be an increased value for the Ro5 parameters (MW, HDO, HAC) except Clog *P*, where less change was observed. There was a significant change in the median value of oral drugs compared to patented compounds: 350 versus 450 for MW, and 3.1 and 4.1 for Clog *P*, respectively. The authors explain the difference by the shift in today's drug targets and by increased requirements in potency and bioavailability. It is also suggested that lipophilicity is the most important property that should be kept as low as possible because it is directly connected to promiscuity and thus toxicity.[26] The degree of saturation, expressed as the ratio of sp3 carbons to the total number of carbons is altered during the progression of a compound from discovery phase (0.36) through clinical phases (0.38, 0.43, 0.45) to drugs (0.47) according to Lovering.[63] The authors further outline the number of stereo centers, which increases by 21% in drugs (64%) with one or more stereo centers compared to discovery compounds (53%). The impact of the number of aromatic rings contained in a molecule,

which inversely correlated with degree of saturation, on lipophilicity (Clog *P* and log *D* 7.4), aqueous solubility, serum albumin binding, cytochrome inhibition, and hERG inhibition which ultimately influences the oral drug candidate developability was outlined by Ritchie and Macdonald.[64] They examined 280 compounds in the GlaxoSmithKline pipeline, and found that compounds with more than three aromatic rings have poor developability and increased risk of attrition.

Based on the observation that hydrophobicity and hydrogen bonding can be conformation dependent, their variability was incorporated in the 'molecular sensitivity' concept.[65] Molecular sensitivity captures the ratio between the range of a given property (e.g., hydrophobicity) and conformational flexibility, which is computed as mean root mean square distance (RMSD) for all nonredundant conformers. It thus captures the variability of that property per shift in atomic positions.[65] This concept was explored for 125 biologically relevant molecules, for which conformational profiles were determined by Monte Carlo simulations; around 40% of the dataset was found to be sensitive to molecular flexibility. The authors evaluated molecular sensitivity in relationship to transdermal permeability, for which they found that log *P* alone does not correlate well with cutaneous penetration ($r^2 = 0.36$, where r^2 is the fraction of explained variance); however, by adding range and sensitivity log *P* to the correlation markedly improved it ($r^2 = 0.76$). This confirms results obtained by Hopfinger et al.,[66] who developed the membrane-interaction QSAR (MI-QSAR) paradigm for the same reasons. According to them, permeability and hydrophobicity depend upon the free energy of (aqueous) solvation, the extent of interaction of the drug with a model phospholipid monolayer, and on the conformational flexibility of the solute within the model membrane.[66]

Testa et al.[65] point out that commonly used hydrophobicity/H-bonding descriptors miss relevant information by lacking appropriate description to capture the dynamic nature of bioactive molecules, and conclude that designing such dynamical descriptors can increase the amount of chemical information to be used in drug discovery pipelines. The dynamic polar surface area (PSA_d) could be regarded as a free-energy-based (Boltzmann) equivalent of the 'molecular sensitivity' concept, since it is weighted based on all (low-energy) conformations identified via molecular mechanics calculations in vacuum and in simulated chloroform and water environments.[67] PSA_d is, in fact, related to hydrogen bonding since PSA correlates very well with the sum of HDO and HAC. Thus, molecular sensitivity analyses could offer additional information to support our understanding of drug-likeness in terms of property distributions.

Following the seminal work by Lipinski et al. on Ro5, several papers addressed the issue of oral absorption and central nervous system (CNS) penetration. For example, several cutoff values such as PSA $\leq$ 90 Å^2 (CNS permeable[68]), PSA $\leq$ 140 Å^2 (intestinal absorption[68]) and (RTB $\leq$ 10, PSA $\leq$ 140 Å^2, HDO + HAC $\leq$ 12) (rat oral bioavailability[69]) have been recommended. In another study, the following cutoff values have been proposed[70] in order to maximize the probability of a chemical library to contain CNS permeable compounds: (1) PSA < 90.0 Å^2; (2) HDO $\leq$ 2; (3) Clog *P* between 2 and 5; (4) MW < 450. Although these cutoff values are likely to improve the probability of identifying compounds that penetrate the blood–brain barrier via passive diffusion, they cannot take into account the influence of efflux pumps such as those from the ATP-binding cassette (ABC) transporter family. Their perturbation by small molecules can significantly influence a drug's CNS permeability, leading to the presence (or absence) of CNS-related side effects.[71] Therefore, although 'drug-like' properties can assist scientists in their effort to design better chemical libraries and perhaps better drugs, such properties cannot take into account all the variables that influence a drug's and pharmacodynamic (PD) properties. To address the influence of commonly used molecular descriptors (e.g., MW, Clog *P*, sum of oxygen and nitrogen, RNG, RTB, HAC, number of halogens) on the PK/PD properties of drugs, Vieth et al.[72] examined the differences between marketed oral drugs and other drug formulations (e.g., injectables, topical). They found that lower MW (median value 322.5 in drugs vs. 416.4 in injectables), balanced Clog *P* (median value 2.3 in drugs vs. 0.7 in injectables), and fewer flexible bonds (median RTB 5 in drugs vs. 7 in injectables) improves the likelihood of a compound to become a drug candidate with an oral administration route.

The discrimination between 'drugs' and 'nondrugs' has consistently been reproduced by many groups, which used a variety of descriptors, statistical methods and as well as different chemical databases. The concept of machine-based 'drug-likeness' has become widely accepted by the cheminformatics community as part of the decision tree, and is often used in conjunction with the Ro5 and 'lead-like' criteria. In its general interpretation, DLS can assist chemists to quickly evaluate, for example, what other chemists have considered worthy of synthesis (and patenting) before them, while still placing the chemicals in question closer to 'drugs' as opposed to 'nondrugs'. High DLS values do not make a molecule a drug, nor

do they ensure better toxicological or PK profiles. Rather, they indicate that more of the molecule's properties and features (depending on descriptors) are encountered in molecules from CMC, MDDR, and WDI, and fewer of its features are associated with ACD. To evaluate drug-likeness, one can incorporate besides DLS the estimation of certain physicochemical properties such as PSA, in order to maximize the likelihood of CNS penetration or intestinal permeability. Separate models (if available) need be considered in order to account for active transport, metabolism, and other PK/PD phenomena.

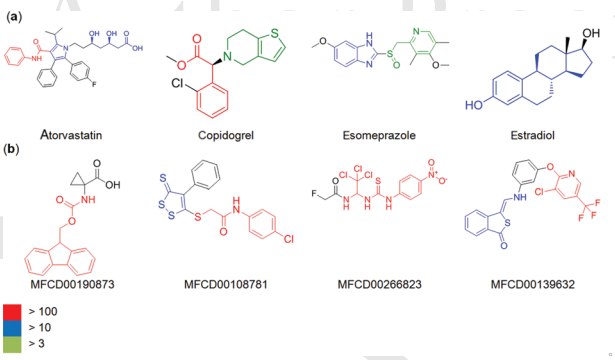

FIGURE 1 | ***Examples of representative nonoverlapping fragments with up to five bonds radius having high frequencies in DRUGS (a), and available chemicals directory (b). The depicted fragments are highlighted according to their occurrence in DRUGS/ACD data sets.

UNDERSTANDING DRUG-LIKENESS: A HEURISTIC PROCESS

Drug-Likeness Filters from Fragments

Filter-based methods, which are simplified decision trees, have comparatively good performance compared to other machine learning methods in drug-likeness evaluation, and are less likely to lead to overfitting. Filter methods are suitable for evaluation of very large data sets, where the decision (pass/fail) is rooted in the statistical distribution of input variables, as opposed to machine learning methods that employ (non) linear transformations on input variables. We refer to such methods as 'model-free' to emphasize their (relative) independence on noisy, mislabeled data, and the significantly less complex computations required. The rules underlying such filters are based on statistical evidence, e.g., the occurrence of fragments or the distribution of properties, and on chemical expertise, with the advantage of being linked to chemical moieties and direct interpretation. Perhaps this ease-of-use contributed to the wide adoption of filter-based techniques (e.g., Ro5, lead-likeness filters) in compound selection for high-throughput and virtual screening. The development of a drug-likeness filter based on molecular fragments occurrence in a drugs/nondrugs data set[44] is summarized in the followings.

Systematic Exploration of DRUGS/NONDRUGS Molecular Fragments Space

Among many categories of two-dimensional (2D) molecular descriptors, a set that performs well in virtual screening is referred to as circular fingerprints or atomic signatures.[73–75] These circular fingerprints, computed using the Morgan algorithm, produce canonical atomic environments and features that encode enough structural information to compute any 2D descriptors.[76–78] The information-rich content of these descriptors is attributed to systematic exploration of chemical space describing a set of chemical structures. Use of atomic signatures to characterize drug/nondrug chemical space provides a comprehensive mapping of all possible molecular fragments present in drug/nondrug chemical catalogs. Our in-house collection was derived by indexing all active ingredients (salts, formulations, or drug combinations were excluded) from over 10,000 approved drugs worldwide. This collection served as the drugs dataset ('DRUGS'; $N = 3823$). Version 2002.1 of ACD (from MDL/SYMYX) was used as the 'nondrugs' dataset; following duplicate and DRUGS molecules removal, this version of ACD contained 178,011 compounds. For external validation, we used the 2006.2 release of MDDR (MDL/SYMYX); after removal of duplicates and structures present in both DRUGS and ACD, MDDR contained 169,277 compounds. Circular fingerprints were computed using an in-house program written using Java and the JChem[79] application programming interface (API). Atomic environments of up to five bonds radius were collected and saved for further processing (see Figure 1).

The selection of those atomic environments or molecular fragments that are most relevant was performed using an occurrence-based scheme as follows: Each newly generated fragment was assigned two probability values: One associated with DRUGS, another one associated with ACD. Only molecular fragments with occurrence ≥ 3 ($\sim 0.1\%$) for DRUGS, and ≥ 100 ($\sim 0.1\%$) for ACD were processed further. The following types of fragments were discarded: (1) fragments for which the probability values were equal; (2) fragments from ACD which have the probability $P_{ACD} < 2^{*}P_{DRUGS}$ and fragments from DRUGS dataset which have probability $P_{DRUGS} < 2^{*}P_{ACD}$. A final list of 15,970 fragments was stored

TABLE 2 | Distribution of Molecular Properties in ACD Compounds that Pass DLF, Compared to MDDR and DRUGS

Molecular Property	Threshold	%ACD[a]	ACD[a] Median Value	%ACD[b]	ACD[b] Median Value	%DRUGS	DRUGS Median Value	%MDDR	MDDR Median Value
MW	≤500	98.50	278.24	98.60	310.46	86.42	318.52	71.61	423.46
Clog *P*	≤5	88.82	2.58	80.17	3.53	86.16	2.34	71.80	3.70
HDO	≤5	99.29	1	99.82	1	92.81	1	93.02	2
HAC	≤10	99.97	2	99.99	2	94.59	3	95.70	3
RNG	$1 \leq RNG \leq 4$	87.91	2	96.86	2	82.29	2	75.12	4
RTB	$2 \leq RTB \leq 8$	84.85	4	87.84	4	67.38	5	56.98	7
RGB	≥18	29.17	13	38.62	16	40.62	16	72.74	22
PSA	≤120	94.06	60.22	97.81	54.79	79.91	68.74	75.69	82.28

[a]Distribution of properties computed on subset of ACD library which pass the DLF.
[b]Distribution of properties computed on subset of ACD library which fails the DLF.

as SMARTS,[46] along with probability values for both the ACD and DRUGS datasets. Most frequent fragments found in DRUGS are: hydroxyl groups, amines, esters, amides, phenols, sulfonamides, and the conserved β-lactam, as opposed to ACD. In an associative manner, these fragments combined to other such fragments contribute to the drug-like character of a chemical. Similar functional groups were found to be predominant in drugs.[46,50,56] Naturally, these functional groups occur in 'nondrugs' as well, albeit with lower occurrence probability compared to drugs.

Fragment-Based Drug-Like Filter

Each of the three datasets (DRUGS, ACD, and MDDR) was submitted to the drug-like filter (DLF). For each input molecule the DLF fragments are matched, if a match is found than the probability values (DRUGS/ACD) associated with the matching fragments are summed up, and the final sums are compared. Molecules pass the DLF if the sum of probability values for drug fragments is higher than that of nondrugs; they fail the DLF otherwise. This filtering procedure is similar to the Naïve Bayes classifier, the difference between our procedure and the former being related to the way probabilities are used: In our procedure, probabilities are summed, whereas in Naïve Bayes probabilities are multiplied. We found by trial and error that the fragments-based filter developed here performs much better when rules fragment probabilities are used in an additive, not multiplicative manner. It should be noted here however, that similar approaches to fragment/substructure based analysis have been undertaking before by Cramer[80] and Hodes,[81] where the authors assigned statistical and probability based scores to predict chemical bioactivity.

Filter Performance

After DLF evaluation, the outcome for the DRUGS, ACD, and MDDR datasets was as follows: 87.05% of DRUGS, 39.65% of ACD, and 78.45% of MDDR structures were evaluated as more drug-like as opposed to ACD-like. The 90th percentile for DRUGS was MW = 562.04, compared to 417.28 for ACD, and 646.07 for MDDR, respectively. For Clog *P*, the 90th percentile for DRUGS was 5.48, compared to 5.55 (ACD), and 6.68 (MDDR), respectively. Based on these observations, most of the DRUGS have MW < 600, and because fragment occurrence is not size dependent we decided to include the MW ≤ 600 rule (optional) to the DLF procedure. Within the DRUGS dataset, 319 compounds (8.34%) have MW ≥ 600; of these, 315 (8.24%) passed the DLF. From MDDR, 23,037compounds (13.61%) have MW ≥ 600, of which 21,697 (12.82%) passed the filter. These high MW compounds (e.g., cyclosporine) have a high content of drug-like fragments to pass DLF; however, they represent only a small fraction of drug molecules. After adding MW ≤ 600 as an additional DLF rule, 78.81% of DRUGS, 40.17% of ACD, and 65.64% of MDDR passed through (DLF + MW). Almost 40% of ACD structures pass the DLF, which proves that ACD is far from perfect as a surrogate for 'nondrugs'. The statistical distribution of the molecular descriptors (Table 2) is similar to distributions observed and discussed elsewhere.[12,82,33,68,83]

The DLF-compliant ACD[a] subset has a property distribution profile closer to DRUGS, as opposed to MDDR. When comparing ACD[a] versus DRUGS, positive differences were observed for PSA, RTB, and MW, where at least 10% more ACD compounds are within these thresholds, while the largest negative difference is observed for ~~RGB~~, where there are at least 10% more DRUGS within the threshold. A likely Q12

reason for the RGB shift to lower values is that most ACD compounds have lower MW (and lower complexity) compared to MDDR, which was used to set the threshold value. More than 94% of the ACD[a] subset has PSA $\leq$ 120 Å^2, versus 80% of DRUGS with the same cutoff. On the other hand the noncompliant DLF ACD[b] subset has a somewhat similar distribution to DRUGS dataset with the exception Clog *P* and PSA where it appears to be more hydrophobic median Clog *P* value for ACD[b] is 3.53 versus 2.34 for DRUGS and 2.58 for ACD[a], and less polar, median PSA value for ACD[b] is 54.79 versus 68.74 for DRUGS and 60.22 for ACD[a]. The distribution profile for MW, HDO, HAC, RTB, RGB, and RNG in ACD[b] follows closely the DRUGS dataset, this leads to the conclusion that the ACD[b] subset is more hydrophobic with less O and N containing functional groups compared to DRUGS and ACD[a]. Drug-like molecular fragments have a higher occurrence rate in the ACD[a] subset compared to non-drug-like fragments (see Figure 1), which is consistent with the 'drug-like' property profile discussed earlier.

These observations advocate the need for such a filter, based on as few assumptions as possible with respect to 'drug' versus 'nondrug' categorization, where the pass/fail criteria are ultimately derived from simple fragment counts. Given the significant overlap between the ACD and DRUGS labels (almost 40%) with respect to chemical fragments and, implicitly, to 2D properties,[76–78] machine learning methods are likely to force the kernel function to use input data that is rather confounded, which ultimately results in classifier models with lower external prediction accuracy, thus modeling more noise than signal. Rule-based systems are more likely to highlight the overlap (noise) between sets, and highlight the differences that emerge from occurrence-based evidence. Instead of performing data compression or reduction on the entire dataset, we went through a simple statistical based elimination process (see above for fragment selection procedure) whereby most (1,402,652 fragment types) of the fragments (represented as SMARTS patterns) were removed. DLF is based on ~1.13% of the total number (1,418,622) of fragments. Filter rules based on ACD-occurring fragments increases the sensitivity of DLF for chemical structures where associative contributions from 'nondruglike' (more correctly, ACD-like) fragments outweighs the contribution of drug-like fragments. When it comes to label separation, it is quite likely that machine learning models (in particular, SVM) will outperform simple filters; however, this particular discrimination problem is based on rather noisy and ill-defined categories, in particular, with respect to 'non-drugs'. Furthermore, temporal validation indicates that predictive power on external datasets can deteriorate in as little as four months,[84] which highlights the need for constantly updating machine learning models. This hinders in particular the utility of drug-like classifiers built using third-party software because of the 'black box' approach, where the model-selected criteria for discrimination are hidden from the end-user and direct interpretation. By contrast, rule-based approaches provide clear output results that can be interpreted more directly.

Predicting Oral Drug-Likeness

For orally administered drugs, understanding and predicting oral bioavailability is of prime interest. Drug absorption through the intestinal tract and its subsequent distribution in the body have an important role in determining its therapeutic efficacy. A debate[26,30,85–87] has developed in the last decade regarding the applicability of Lipinski's Ro5 for oral drug likeness, and several papers proposed to refine and reassess the determination of potential oral drugs.[88–92,72,64] Veber et al.[69] proposed two additional rules for predicting rat oral bioavailability: RTB $\leq$ 10 and PSA $\leq$ 140Å^2; both properties should be observed if a molecule is orally bioavailable. It was later suggested that these rules are hardly applicable to human oral bioavailability and the ability to predict bioavailability by simple thumb rules has been criticized.[93] It is important to improve the methods for discriminating between OBDs and nondrugs (nOBDs), which was the subject of our recent work.[17] A method for indexing oral bioavailability of drugs was introduced, which may be used for *in silico* examination of a molecule's potential to become an oral drug.

From Classification to Indexing

The distinction between OBDs and others was performed by a novel optimization method from our lab.[94–96] Iterative stochastic elimination (ISE) is a general optimization method that finds best solutions to complex combinatorial problems that are functions of many variables. ISE has been used to optimize the ranges of variables' values in order to maximize the differences between two sets. A function (Matthews' Correlation Coefficient, MCC[16]) is used to score the optimizations based on the numbers of True and False Positives and Negatives. ISE leads to the formation of a set of filters, each consisting of a set of variable value ranges. Subsequently, individual molecules may be examined and scored for their ability to pass the

filter set. This score is the oral bioavailability drug-like index (OB-DLI) for individual compounds, which reflects a molecule's chance to belong to the database of OBDs. For testing purposes, the binary character of the decision naturally remains. But, the scalar presentation with OB-DLI allows one to make more elaborate decisions based on the position of a molecule along a scale, as well as on enrichment factors that can be calculated at different levels of the OB-DLI.

TABLE 3 | Average OB-DLI Values of a few Databases

Database (# Test Sets)	Average OB-DLI	# Molecules in each Set
ACD (2)	−0.91	9128
CMC (1)	0.86	1694
MDDR (3)	1.32	6070
ZINC (2)	0.46	9300

Bioavailability of Drugs and of Nondrugs

Indexing the oral bioavailability of compounds based on oral drugs alone does not discriminate well oral drugs from drugs with other modes of administration. We compiled a dataset of OBDs from CMC and MDDR, but included only those molecules that are both Ro5-compliant and lead-like.[15,61] The equivalent dataset of nondrugs from ACD was also reduced to those that obey the same two sets of rules.

The ISE algorithm was then applied to perform simultaneous selection and optimization of the ranges of k-descriptor sets in order to distinguish between OBDs and the others, presumably nOBD. From both databases we cleaned pesticides, UV-screens, etc.[50,97] as well as undesired atomic elements (different from C, S, O, N, P, H, Si, Cl, Br, I, F). The databases were further subjected to clean-up operations such as removal of counterion and solvents. Despite the potential source for confusion, we used single tautomers and the ionization state was determined based on the chemical functions.[98–100] A total of 184 2D descriptors were computed for both databases using MOE 2008.10[101]; of these, 146 noncorrelated descriptors were used. The orally bioavailable CMC (OB-CMC) database was employed as the basis for the training dataset of OBDs and the orally bioavailable ACD was employed as the training set of orally bioavailable nondrugs (nOBD). The 6776 compounds OB-CMC set was randomly partitioned into two portions of $\frac{3}{4}$ and $\frac{1}{4}$ of the dataset, thus providing 5082 training set compounds, with 1694 serving as the test set. Three random subsets from the OB-ACD were picked following the same clean-up and filtering described above: Each one of the three subsets included 9128 compounds; the first was the training set, whereas the two others served as test sets. Three random 6070 compound sets each were prepared from OB-MDDR by the same procedure; two more datasets composed, each, of 9300 molecules from OB-ZINC, were also included in this study. The MDDR sets were used as OBD test sets, in addition to the OB-CMC test set, whereas the ZINC sets and the ACD sets were used as nOBD test sets.

Oral Bioavailability Drug-Like Index

An OBD-like molecule should have a higher probability to belong to the OBD database, and ought to 'pass' DLFs, including those filters that are less successful than the best, i.e., the 'global optimum' with the highest MCC value. Such filters may be regarded as 'local minima' that can be quite close to the global minimum, but may differ due to different clustering. Designing new molecules by having their descriptor values pass a few or more such filters is complicated. However, identifying molecules that already have such characteristics is an easier task. Indeed, by using a set of such filters rather than the single 'best' one, the method benefits from the larger set of good filters, which increases its potential to prioritize molecules. A special equation was devised for calculating the OB-DLI.[17] Results for the average OB-DLI values are presented in Table 3.

We illustrate the effectiveness of this procedure, which discriminates compounds from the 'OBD' database from those in the 'nOBD' database in Figure 2. With this indexing, the qualitative classification is transformed into quantitative one via multiple filters that are the result of the ISE process. Yet, this remains a statistically based process: Those filters that constitute the basis for quantification emerge from examining a large set of molecules, and the MCC value reflects the percentage of predictive success of each filter; although OB-DLI scores single compounds, its statistical nature and the chance for erroneous prediction should be kept in mind.

One way to appreciate the success of OB-DLI in distinguishing OBD from nOBD is to randomly select molecules from each dataset, and position them on the same plot, (see Figure 3). Here, 300 molecules from OB-ACD and from OB-MDDR were randomly picked and positioned along the X-axis, with values of 1–300, as random compound numbers. The Y-axis is the OB-DLI values. It is obvious from Figure 3 that most compounds of OB-ACD have lower OB-DLI values while those from OB-MDDR have much higher values, with most of its compounds above an OB-DLI of 0.0. It may also be seen that OB-DLI may be used

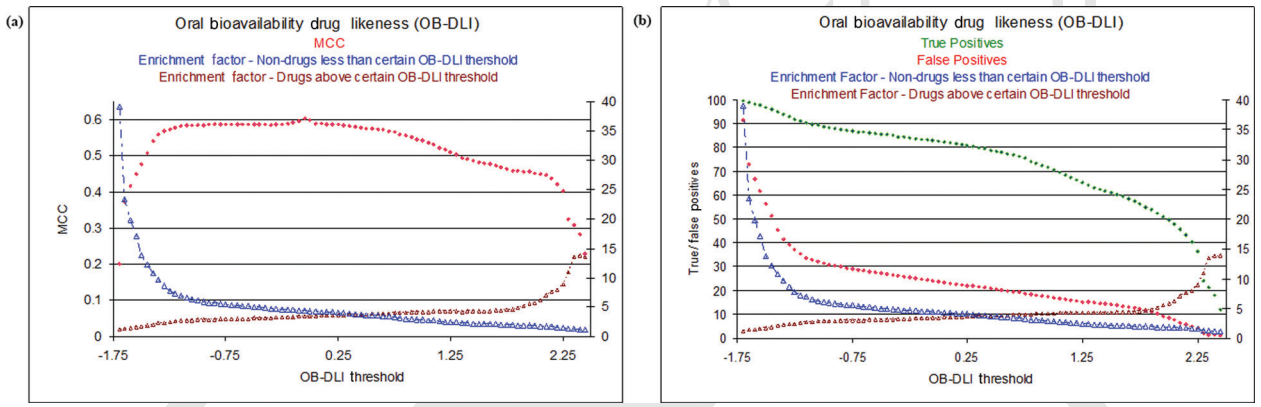

FIGURE 2 | Compounds that have an oral bioavailability drug-like index (OB-DLI) above a certain threshold are considered as potentially orally bioavailable drugs (OBDs) while others are nondrugs. Figure 2a demonstrates the change of MCC along the OB-DLI axis while Figure 2b focuses on the enrichment factors along that axis.

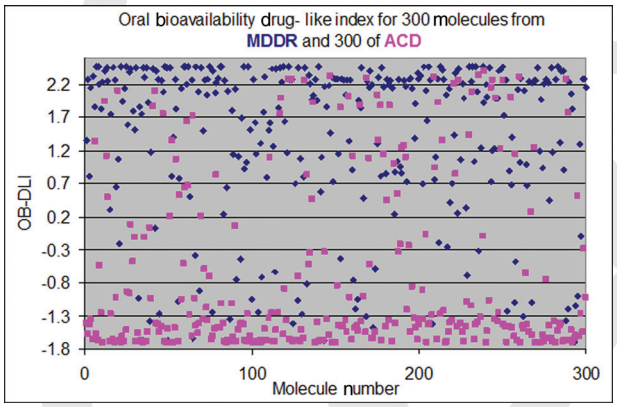

FIGURE 3 | Oral bioavailability drug-like index (OB-DLI) values of 600 randomly selected molecules from two databases—300 chemicals from available chemicals directory (red squares) and 300 drugs from MDDR-clinical (blue rectangles).

as a decision-making tool, in that a line drawn parallel to the X-axis at some higher OB-DLI value excludes most of the nOBDs and enriches the remaining set with many OBD candidates. Some of the grey squares (molecules from the non-OBD database, ACD) that have higher OB-DLI values, could be candidates for OBDs if they are active at some disease target. To conclude, Ro5 does not distinguish between the bioavailability of drugs and that of nondrugs. To address some of the debate associated with Ro5 filtering, we optimized the ranges of Ro5 variables (MW, HDO, HAC, and Clog *P*) in order to best distinguish between the two sets, OBD and nOBD. The optimal filter has MCC of 0.47 corresponding to discovering 93% of the true positives and only 49% of true negatives. This new filter is composed of only two descriptors: MW $\geq$ 240 and number of acceptors $\geq$1.

Launched drugs and MDDR compounds in clinical trials may have higher or lower OB-DLI. We compared the fraction of those drug candidates with high values (OB-DLI $\geq$1.0, 4321 drugs) to those with low values ($\leq$0.0, 1058 drugs). From histograms of the molecular properties we find that some properties differ significantly between those predicted to be more orally bioavailable clinical candidates and those predicted to be less orally bioavailable. Some of these differences are depicted in Figure 4. In addition to differences in molecular weight (see Figure 4a), number of rings (see Figure 4b) and number of rigid bonds (see Figure 4c) there are differences in other properties such as total hydrophobic/negative/positive VDW surface area, sum of atomic polarizabilities, first kappa shape index, Van der Waals surface area and volume. Molecular weight was however the most dominant in its effect on OB-DLI compared to the other descriptors of Lipinski and number of rings and number of rigid bonds from Oprea descriptors. From these results (see Figure 4) it seems that lower values of the properties contribute to lower drug-likeness, i.e., they could be associated with any of the PK factors that limit drug action and decrease efficiency of orally bioavailable molecules.

Oral Bioavailability and Toxicity

The oral bioavailability indexing does not use any information about the toxicity of compounds in the learning sets. Because it is inherently related to dosage (according to Paracelsus), the issue of toxicity remains somewhat ambiguous. On the one hand, drugs are molecules that clearly interact with macromolecular targets, proteins in particular. However, drugs and their metabolites frequently interact with other targets, which may result in off-target activity, i.e.,

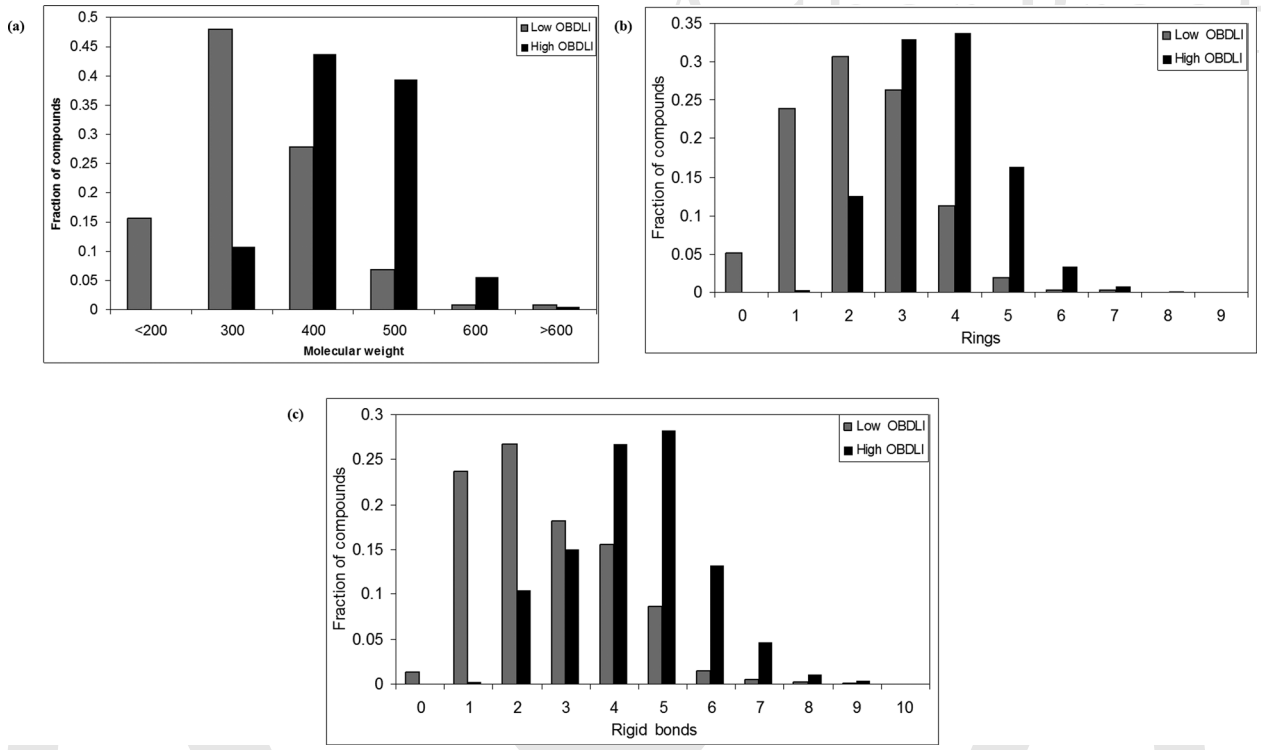

FIGURE 4 | Distribution of the molecular properties of the low/high oral bioavailability drug-like index (OB-DLI) drugs, (a) molecular weight, (b) rings, and (c) rigid bonds.

toxicity. Because toxicity is related to dose, there is a strong interest in identifying drugs that can be administered at a therapeutic window (efficacy) that is orders of magnitude below the toxicity threshold. CMC drugs were assumed in our study to be administered in the range of clinical doses, so that their toxicity is reduced. Therefore, distinguishing between OB-CMC and OB-ACD may capture the difference in toxicity between these two sets. The fact that OB-drugs of CMC are less toxic needs to be contrasted with the set of OB-nondrugs from ACD. However, their toxicity cannot be immediately assessed, as it depends in turn on their ability to interact with proteins and other macromolecules, and to permeate membranes.

CONCLUSION

Drug-likeness is the sum of properties characteristic to chemical substances known as drugs. These are assigned by regulatory agencies that have an implicit social component and cannot, as such, be reliably predicted for any individual compound. However, the pressure to evaluate large chemical libraries shifted the problem from human learning to machine learning. A number of papers deal with the evaluation of drug-likeness for small molecules, as summarized from eighteen such groups. Such work should be placed in the context of Pfizer's Ro5 evaluation as well as lead-likeness, and combined with additional drug-like properties. The discrimination between 'drugs' and 'nondrugs' has been observed by multiple groups using a variety of descriptors, statistical methods, and chemical databases. The concept of machine-based 'drug-likeness' has become widely accepted as part of the cheminformatics-based decision tree and is often used in conjunction with the Ro5 and 'lead-like' criteria. In its most general interpretation, DLS assists chemists to quickly evaluate what other chemists have considered worthy of evaluation, while classifying those chemicals closer to 'drugs' as opposed to 'nondrugs'. High DLS values do not increase the likelihood of FDA approval because they do not ensure better toxicological or PK profiles. Rather, they indicate that more of its properties and features are encountered in molecules from CMC, MDDR, and WDI; and fewer of these molecules are likely to have near neighbors in ACD.

The inherent difficulty in estimating drug-likeness comes from its very definition, i.e., namely that only a relatively small dataset of molecules comprises marketed drugs, which is often replaced by larger datasets. Other heuristic issues include: (1) the 'drug' character of individual compounds may change

over time, as drugs are sometimes withdrawn from the market; (2) the drugs dataset has a rather high heterogeneity, as drugs range in MW from 3 (lithium, the active principle in the antidepressant lithium carbonate) to over 1200 (e.g., Cyclosporine). Despite these issues, the computer-based evaluation of DLS has been incorporated as one of the tools in early drug discovery, and is used to filter out compounds with undesirable properties, or to enrich libraries with compounds having a higher drug-like character.

Our survey of literature data shows that drug-likeness has been predominantly modeled using machine-learning methods, which we believe are compelled to discriminate 'drugs' from 'nondrugs' using a rather confounded set of labels (in particular, ACD). Drug-likeness might be better evaluated using filter-based tools based on molecular fragments because these enable compound selection for chemical structures (and implicitly 2D properties) closer to known drugs, in contrast to those chemical structures and properties that are closer to 'nondrugs'. Both DLS and DLF tools are ultimately used to assist with compound (e.g., HTS hit) prioritization. Furthermore, there is a constant interest to identify orally available drugs. The Iterative Stochastic Elimination optimizer was used to develop of a set of property filters, which form the basis of OB-DLI. Although the decision is binary in character, the scalar representation of OB-DLI allows the end-user to make more elaborate decisions based on the position of an individual compound along a scale. However, one of the limitations of drug-like evaluation is the inability to incorporate dosage, which could potentially assist with toxicity evaluation.

SUPPLEMENTARY INFORMATION

Note to WIRE: Some of the material related to drug-like filters (DLF) and to oral drug-likeness could be moved to Supplem Info. However, we believe it should be an integral part of this text.

ACKNOWLEDGMENTS

We thank David Marcus for contributing to the oral bioavailability indexing. A.G. and A.R. thank the EU CancerGrid Consortium http://cancergrideu.w3h.hu/ for financial support. This study was supported, in part, by NIH grant 5U54MH084690–03 (OU and TIO) and by the University of New Mexico sabbatical leave program (TIO).

REFERENCES

1. Drews J. Innovation deficit revisited: reflections on the productivity of pharmaceutical R&D. *Drug Discov Today* 1998, 3:491–494.
2. Oprea TI. Sense and nonsense in drug discovery: a chemical perspective. In: Kruse CG, Timmerman H, eds. *Towards Drugs of the Future*. Amsterdam: IOS Press; 2008; 29–36.
3. Stone M, Jonathan P. Statistical thinking and technique for QSAR and related studies. Part I. General theory. *J Chemom* 1993, 7:455–475.
4. Weininger D. Combinatorics of small molecular structures. In: Von Ragué Schleyer P, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, eds. *Encyclopedia of Computational Chemistry*. New York: Wiley; 1998, 425–430.
5. Hann MM, Oprea TI. Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* 2004, 8:255–263.
6. Blum LC, Reymond J-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc* 2009, 131:8732–8733.
7. Pollock SN, Coutsias EA, Wester MJ, Oprea TI. Scaffold topologies. 1. Exhaustive enumeration up to eight rings. *J Chem Inf Model* 2008, 48:1311–1324.
8. Wester MJ, Pollock SN, Coutsias EA, Allu TK, Muresan S, Oprea TI. Scaffold topologies. 2. Analysis of chemical databases. *J Chem Inf Model* 2008, 48:1311–1324.
9. Rosén J, Gottfries J, Muresan S, Backlund A, Oprea TI. Novel chemical space exploration via natural products. *J Med Chem* 2009, 52:1953–1962.
10. Oprea TI, Allu TK, Fara DC, Rad RF, Ostopovici L, Bologa CG. Lead-like, drug-like or "pub-like": how different are they? *J Comput Aided Mol Des* 2007, 21:113–119.
11. Leo A. Estimating LogPoct from structures. *Chem Rev* 1993, 5:1281–1306.
12. 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.

13. http://thomsonreuters.com/products_services/science/science_products/a-z/world_drug_index. (Accessed March 31, 2010). Q14
14. Olah MM, Bologa CG, Oprea TI. Strategies for compound selection. *Curr Drug Discov Technol* 2004, 1:211–220.
15. Oprea TI. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* 2000, 14:251–264.
16. 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, 40:1315–1324.
17. Rayan A, Marcus D, Goldblum A. Predicting oral druglikeness by iterative stochastic elimination. *J Chem Inf Model* 2010, 50:437–445.
18. Teague SJ, Davis AM, Leeson PD, Oprea TI. The design of leadlike combinatorial libraries. *Angew Chem Int Ed Engl* 1999, 38:3743–3748.
19. Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J Chem Inf Comput Sci* 2001, 41:856–864.
20. Sneader W. *Drug Prototypes and their Exploitation*. Chichester: Wiley; 1996.
21. Andrews PR, Craik DJ, Martin JL. Functional group contributions to drug-receptor interactions. *J Med Chem* 1984, 27:1648–1657.
22. 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, 41:1308–1315.
23. Proudfoot JR. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg Med Chem Lett* 2002, 12:1647–1650.
24. Allu TK, Oprea TI. Rapid evaluation of synthetic and molecular complexity for *in silico* chemistry. *J Chem Inf Model* 2005, 45:1237–1243.
25. Perola E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J Med Chem* 2010, 53:2986–2997.
26. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* 2007, 6:881–890.
27. DeStevens G. Serendipity and structured research in drug discovery. In: Jucker E, ed. *Progress in Drug Research*. Birkhauser: Basel; 1985, 189–203.
28. Horrobin DF. Innovation in the pharmaceutical industry. *J R Soc Med* 2000, 93:341–345.
29. Cuatrecasas P. Drug discovery in jeopardy. *J Clin Invest* 2006, 116:2837–2842.
30. Keseru GM, Makara GM. The influence of lead discovery strategies on the properties of drug candidates. *Nat Rev Drug Discov* 2009, 8:203–212.
31. Vaz RJ, Klabunde T. *Antitargets: Prediction and Prevention of Drug Side Effects*. Weinheim: Wiley-VCH; 2008.
32. Oprea TI. Current trends in lead discovery: are we looking for the appropriate properties? *J Comput Aided Mol Des* 2002, 16:325–334.
33. Muegge I. Selection criteria for drug-like compounds. *Med Res Rev* 2003, 23:302–321.
34. Gillet VJ, Willett P, Bradshaw J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J Chem Inf Comput Sci* 1998, 38:165–179.
35. Ajay A, Walters WP, Murcko MA. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J Med Chem* 1998, 41:3314–3324.
36. Sadowski J, Kubinyi H. A scoring scheme for discriminating between drugs and nondrugs. *J Med Chem* 1998, 41:3325–3329.
37. http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp. (Accessed March 31, 2010). Q15
38. http://www.symyx.com/products/databases/sourcing/acd/index.jsp. (Accessed March 31, 2010). Q16
39. Walters WP, Stahl MT, Murcko MA. Virtual screening—an overview. *Drug Discov Today* 1998, 3:160–178.
40. Oprea TI, Matter H. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* 2004, 8:349–358.
41. Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. *J Chem Inf Comput Sci* 1995 35:826–833.
42. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci* 2003, 44:1–12.
43. Burges CJC. A Tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 1998, 2:121–167.
44. Ursu O, Oprea TI. Model-Free Drug-likeness from fragments. *J Chem Inf Model* 2010, 50:1387–1394.
45. Wang J, Ramnarayan K. Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J Comb Chem* 1999, 1:524–533.
46. Wagener M, van Geerestein VJ. Potential drugs and nondrugs: prediction and identification of important structural features. *J Chem Inf Comput Sci* 2000, 40:280–292.
47. Xu J, Stevenson J. Drug-like index: a new approach to measure drug-like compounds and their diversity. *J Chem Inf Comput Sci* 2000, 40:1177–1187.
48. Oprea TI, Gottfries J, Sherbukhin V, Svensson P, Kühler TC. Chemical information management in

drug discovery: optimizing the computational and combinatorial chemistry interfaces. *J Mol Graph Model* 2000, 18:512–524.

49. Anzali S, Barnickel G, Cezanne B, Krug M, Filimonov D, Poroikov V. Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *J Med Chem* 2001, 44:2432–2437.
50. Muegge I, Heald SL, Brittelli D. Simple selection criteria for drug-like chemical matter. *J Med Chem* 2001, 44:1841–1846.
51. Brustle M, Beck B, Schindler T, King W, Mitchell T, Clark T. Descriptors, physical properties, and drug-likeness. *J Med Chem* 2002, 45:3345–3355.
52. Takaoka Y, Endo Y, Yamanobe S, Kakinuma H, Okubo T, et al. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J Chem Inf Comput Sci* 2003, 43:1269–1275.

Q17

53. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci* 2003, 43:2048–2056.
54. Zheng SX, Luo XM, Chen G, Zhu WL, Shen JH, et al. A new rapid and effective chemistry space filter in recognizing a druglike database. *J Chem Inf Model* 2005, 45:856–862.
55. Muller KR, Ratsch G, Sonnenburg S, Mika S, Grimm M, Heinrich N. Classifying 'drug-likeness' with kernel-based learning methods. *J Chem Inf Model* 2005, 45:249–253.
56. Li QL, Bender A, Pei JF, Lai LH. A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification. *J Chem Inf Model* 2007, 47:1776–1786.
57. Schneider N, Jackels C, Andres C, Hutter MC. Gradual in silico filtering for druglike substances. *J Chem Inf Model* 2008, 48:613–628.
58. Schierz A, King R. Drugs and drug-like compounds: discriminating approved pharmaceuticals from screening-library compounds. In: Istrail S, Pevzner P, Waterman M, eds. *Pattern Recognition in Bioinformatics*. Berlin/Heidelberg: Springer; 2009, 331–343.
59. Ohno K, Nagahara Y, Tsunoyama K, Orita M. Are there differences between launched drugs, clinical candidates, and commercially available compounds? *J Chem Inf Model* 2010, 50:815–821.
60. Wager TT, Hou X, Verhoest PR, Villalobos A. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chem Neurosci* 2010, 1:435–449.
61. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 2000, 44:235–249.
62. Oprea TI. Chemical space navigation in lead discovery. *Curr Opin Chem Biol* 2002, 6:384–389.
63. Lovering F, Bikker J, Humblet C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J Med Chem* 2009, 52:6752–6756.
64. Ritchie TJ, Macdonald SJF. The impact of aromatic ring count on compound developability—are too many aromatic rings a liability in drug design? *Drug Discovery Today*. 2009, 14:1011–1020.
65. Vistoli G, Pedretti A, Testa B. Assessing drug-likeness—what are we missing? *Drug Discov Today* 2008, 13:285–294.
66. Kulkarni A, Han Y, Hopfinger AJ. Predicting caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J Chem Inf Comput Sci* 2002, 42:331–342.
67. Palm K, Luthman K, Ungell A-L, Strandlund G, Beigi F, et al. Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J Med Chem* 1998, 41:5382–5392.
68. Clark DE. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration. *J Pharm Sci* 1999, 88:815–821.
69. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 2002, 45:2615–2623.
70. Hitchcock SA. Blood–brain barrier permeability considerations for CNS-targeted compound library design. *Curr Opin Chem Biol* 2008, 12:318–323.
71. Broccatelli F, Carosati E, Cruciani G, Oprea TI. Transporter-mediated efflux influences CNS side effects: ABCB1, from antitarget to target. *Mol Inf* 2010, 29:16–26.
72. Vieth M, Siegel MG, Higgs RE, Watson IA, Robertson DH, et al. Characteristic physical properties and structural fragments of marketed oral drugs. *J Med Chem* 2004, 47:224–232.
73. Bender A, Mussa HY, Glen RC, Reiling S. Molecular similarity searching using atom environments, information-based feature selection, and a Naive Bayesian classifier. *J Chem Inf Comput Sci* 2003, 44:170–178.
74. Bender A, Mussa HY, Glen RC, Reiling S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J Chem Inf Comput Sci* 2004, 44:1708–1718.

75. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, et al. Comparison of topological descriptors for similarity based virtual screening using multiple bioactive reference structures. *Org Biomol Chem* 2004, 2:3256–3266.
76. Faulon J-L, Visco DP, Pophale RS. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J Chem Inf Comput Sci* 2003, 43:707–720.
77. Faulon J-L, Churchwell CJ, Visco DP. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J Chem Inf Comput Sci* 2003, 43:721–734.
78. Faulon J-L, Collins MJ, Carr RD. The Signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J Chem Inf Comput Sci* 2004, 44:427–436.
79. ChemAxon: JChem Base version 5.3.1. Budapest, 2010.
80. Cramer RD, Redl G, Berkoff CE. Substructural analysis. Novel approach to the problem of drug design. *J Med Chem* 1974, 17:533–535.
81. Hodes L, Hazard GF, Geran RI, Richman S. A statistical-heuristic method for automated selection of drugs for screening. *J Med Chem*. 1977, 20:469–475.
82. Oprea TI. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* 2000, 14:251–264.
83. Clark DE. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J Pharm Sci* 1999, 88:807–814.
84. Gavaghan C, Arnby C, Blomberg N, Strandlund G, Boyer S. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J Comput Aided Mol Des* 2007, 21:189–206.
85. Abad-Zapatero C. A Sorcerer's apprentice and the rule of five: from rule-of-thumb to commandment and beyond. *Drug Discov Today* 2007, 12:995–997.
86. Kubinyi H. Drug research: myths, hype and reality. *Nat Rev Drug Discov* 2003, 2:665–668.
87. van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov* 2003, 2:192–204.
88. Bai JPF, Utis A, Crippen G, He HD, Fischer V, et al. Use of classification regression tree in predicting oral absorption in humans. *J Chem Inf Comput Sci* 2004, 44:2061–2069.
89. Bergstrom CAS, Strafford M, Lazorova L, Avdeef A, Luthman K, Artursson P. Absorption classification of oral drugs based on molecular surface properties. *J Med Chem* 2003, 46:558–570.
90. Martin YC. A bioavailability score. *J Med Chem* 2005, 48:3164–3170.
91. Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD. A comparison of physiochemical property profiles of development and marketed oral drugs. *J Med Chem* 2003, 46:1250–1256.
92. Yoshida F, Topliss JG. QSAR model for drug human oral bioavailability. *J Med Chem* 2000, 43:2575–2585.
93. Hou TJ, Wang JM, Zhang W, Xu XJ. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *J Chem Inf Model* 2007, 47:460–463.
94. Glick M, Rayan A, Goldblum A. A stochastic algorithm for global optimization and for best populations: a test case of side chains in proteins. *Proc Natl Acad Sci USA* 2002, 99:703–708.
95. Rayan A, Barasch D, Brinker G, Cycowitz A, Geva-Dotan I, et al. New stochastic algorithm to determine drug-likeness. *Abstr Pap Am Chem Soc* 2003, 226:U297-U297.
96. Rayan A, Senderowitz H, Goldblum A. Exploring the conformational space of cyclic peptides by a stochastic search method. *J Mol Graph Model* 2004, 22:319–33.
97. Zuccotto F. Pharmacophore features distributions in different classes of compounds. *J Chem Inf Comput Sci* 2003, 43:1542–1552.
98. Kubinyi H. From narcosis to hyperspace: the history of QSAR. *Quant Struct-Act Relat* 2002, 21:348–356.
99. Pospisil P, Ballmer P, Scapozza L, Folkers G. Tautomerism in computer-aided drug design. *J Recept Signal Transduct Res* 2003, 23:361–371.
100. Tetko IV, Bruneau P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J Pharm Sci* 2004, 93:3103–3110.
101. Vilar S, Cozza G, Moro S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem* 2008, 8:1555–1572.

FURTHER READING

Balakin KV. *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery*. Hoboken, NJ: John Wiley & Sons; 2010.

Bultinck P. *Computational Medicinal Chemistry for Drug Discovery*. New York: Marcel Dekker; 2004.

Chorghade MS. *Drug Discovery and Development*. Hoboken, NJ: Wiley-Interscience; 2006.

Fischer J, Ganellin CR. *Analogue-Based Drug Discovery*. Weinheim: Wiley-VCH; 2006.

Gad SC. *Drug Discovery Handbook*. Hoboken, NJ: Wiley-Interscience; 2005.

Samuelsson G. *Drugs of Natural Origin: A Textbook of Pharmacognosy*. Stockholm: Swedish Pharmaceutical Press; 2004.

Wermuth CG. *The Practice of Medicinal Chemistry*. San Diego, CA: Academic Press; 2003.

Queries

Q1: AU: Please check that language changes made throughout this article are OK.

Q2: AU: Please check affiliations for correctness.

Q3: AU: Please define 'GBD'.

Q4: AU: Please check the term 'ClogP' for correctness.

Q5: AU: Please define 'MDDR' and 'PDR'.

Q6: AU: Please define 'RNG' and 'RTB'.

Q7: AU: Please check '–' sign in this entry.

Q8: AU: Please check '–' sign in this entry.

Q9: AU: Please check '–' sign in this entry.

Q10: AU: Please check '–' sign in this entry.

Q11: AU: Please define 'hERG'.

Q12: AU: Please define 'RGB'.

Q13: AU: Please check heading 'Oral Bioavailability Drug-Like Index' for correctness.

Q14: AU: Please provide the source and title of web page for Ref. 13.

Q15: AU: Please provide the source and title of web page for Ref. 37.

Q16: AU: Please provide the source and title of web page for Ref. 38.

Q17: AU: As per the journal style guide, at least ten authors can go in the reference list. Please provide at least ten authors for Refs. 52, 54, 67, 72, 75, 88, and 95 if these references contain more than ten authors.