ELSEVIER

Pursuing the leadlikeness concept in pharmaceutical research

Mike M Hann[1] and Tudor I Oprea[2]

Lipinski and others, through concepts such as drug-likeness, re-focussed drug discovery back to the principles of medicinal chemistry in the high-throughput era as key to reducing attrition. More recently, the need to go further in defining what makes a good lead has been recognised with the concept of leadlikeness. Leadlikeness implies cut-off values in the physico-chemical profile of chemical libraries such that they have reduced complexity (e.g. MW below <400) and other more restricted properties. We examine these concepts in the context of Virtual (theoretically possible), Tangible (chemically feasible) and Real (physically available) worlds of molecules. In a thought experiment, we take the HTS concept to the extreme: screening an estimated 60 million 'Global Collection' on 5000 targets and realising that perhaps millions of drug candidates might be found that could not possibly be handled in reality. Sampling of the Virtual and Tangible worlds is therefore a necessity. We show that the world of Reals is significantly under-sampled as the MW of compounds increases. This supports the design and screening of 'reduced complexity' (leadlike) compound libraries, preferably with synthetic handles available for rapid chemical iteration and detected as interesting by careful screening or biophysical assays.

Addresses
[1]GlaxoSmithKline Research and Development, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK
e-mail: mike.m.hann@gsk.com
[2]Division of Biocomputing, University of New Mexico School of Medicine, MSC 08 4560, 1 University of New Mexico, Albuquerque New Mexico 87131-0001, USA
e-mail: toprea@salud.unm.edu

Current Opinion in Chemical Biology 2004, **8**:255–263

This review comes from a themed issue on
Combinatorial chemistry
Edited by A Ganesan and Anthony D Piscopio

Available online 7th May 2004

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2004.04.003

Abbreviations

ACD	Available Chemicals Directory
HAC	hydrogen bond acceptors
HAM	high-activity molecule
HDO	hydrogen bond donors
HTS	high-throughput screening
MDDR	MDL Drug Data Report
RNG	number of rings
RO5	rule of fives
RTB	rotatable bonds
VTR	Virtual/Tangible/Real
WDI	World Drug Index

Introduction

Preclinical drug research has placed an increased pressure on earlier stages of the discovery process [1–3], in particular on the choice of leads or drug prototypes [4] (i.e. the molecular structures that undergo the process of optimization before reaching candidate drug status [5]). In this review, we discuss the reasons for this pressure, briefly analyzing the evolution of concepts that aim at improving the quality of leads [6,7], and the understanding of lead-like space [8]. These concepts are currently used to assist the design and construction of virtual and physical compound collections for screening.

Two distinct scenarios occur. In the first, one does not have any specific target in mind at the time when the compound collection is assembled. In the physical world, this corresponds to most in-house collections for high-throughput screening (HTS) that have evolved in the pharmaceutical industry through the historical collection of samples synthesized and acquired over many years. In the virtual *in silico* world, this can be extended to the concept of a virtual library that is not target-specific. A subset of *possible (or tangible)* compounds in the library includes those that experience suggests can be physically assembled on demand through established chemistries. The second scenario occurs when constructing a more focused library for a specific target (or group of related targets), using target-based information to focus or bias the selection. In the physical world, this corresponds to target-specific (focused) libraries available from many chemical vendors or as designed and synthesized internally within a pharmaceutical company. In the virtual world, such compound sets can be derived from analyzing high-activity molecules (HAMs), or from known leads co-crystallized with the target of interest. The *possible* collection is in this case represented by a much more limited set of target molecules that meet, for example, specific pharmacophoric criteria, although they may still be part of a larger array, which is actually synthesized. This is because of the nature of combinatorial chemistry, which is usually done in n by m arrays.

One of the major efforts to revise the input/output (or the signal/noise) ratio with regards to the effectiveness of chemical aspects of drug discovery has been in the area of cheminformatics. In the strictest sense, chemical informatics integrates data via computer-assisted manipulation of chemical structures [9]. Chemical inventory and compound registration are vital to cheminformatics, but it is their combination with other theoretical tools from the wider realm of computational chemistry and their linkage to physical organic chemistry, pharmacodynamics and

pharmacokinetics (and eventually, to the amelioration/avoidance of undesirable pharmacology leading to toxicology) that brings unique capabilities in the area of lead and drug discovery. In recent years, cheminformatics has emerged as the informatics-driven technological push in preclinical research, since it attempts to link all the involved scientific partners, from virtual screening to animal toxicology via one central element: *chemical structure*. Cheminformatics has been given increased attention in the early stages of lead discovery, where the concept of leadlikeness has gained increased importance: the processes by which interesting starting points for medicinal chemistry can be found need to become cost effective.

Issues in early lead discovery

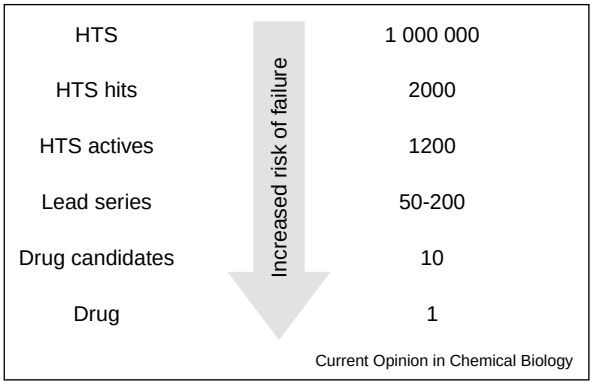
The innovation deficit [1] of pharmaceutical R&D, whereby there appears to be a lack of truly new therapies being developed, can in part be explained by the desire to have a 'best-in-class' strategy for products (thus securing a lasting product), in contrast to the 'first-in-class' strategy (i.e. being the first to market a new class of therapeutics). Because 'first-in-class' therapeutics rarely remain 'first-in-class' (e.g. Cimetidine was surpassed by Ranitidine, and Felodipine by Amlodipine), the incentive to be strong innovators is somewhat lacking unless a company also commits to improving on its own 'first-in-class' products. Pharmaceutical companies therefore often follow similar trends and molecular targets in a market-driven prioritization process [3], which can slow the pace of innovation with new targets. While companies will aspire to be truly innovative, the market experience often makes the prospects of developing truly new products daunting [3,10•].

Lipinski's seminal analysis of reasons why compounds fail to progress to become oral drugs and the resulting 'rule of fives' (RO5) [11] pointed out the dangers of ignoring pharmacokinetic properties in combinatorial library design. Given the time-lag between lead discovery and drug launch (usually 8–15 years [12]), we may still be witnessing the effects of progressing drug candidates from the pre-RO5 era. A decade after the initial shift in the lead discovery paradigm toward HTS and combinatorial chemistry, pharmaceutical R&D productivity remains low. In addition to ignoring or forgetting many of the principles of medicinal chemistry in the early years of the new technologies, the goalposts have also been continually moving. Thus, the criteria that candidate drugs must fulfill before approval are increasingly demanding.

HTS clearly works as a method for finding starting points for drug-discovery programs, but how can it be made more effective?

The preclinical drug-discovery cascade, starting from HTS and moving into the launched drug phase requires the screening of the order of one million compounds to find a suitable lead for one ultimately successful outcome [13] (Figure 1).

Figure 1



A typical drug discovery cascade [13]. Accurate figures are difficult to average across the pharmaceutical industry, so the number of compounds is for illustrative purposes. A 40% false positive rate is assumed in evaluating HTS hits, and one in 2–5 leads are assumed to progress from lead identification to drug candidate. The risk of failure increases as a molecule becomes a drug candidate because of high costs in clinical trials. Modified from [13].

If we knew *a priori* more about the relationship between chemotypes and target activity, this ratio would undoubtedly improve. Thus, HTS is usually more successful for so called 'tractable targets' (e.g. kinases or G-protein coupled receptors) [14]. Not withstanding this lack of chemotype-activity knowledge, there are also many process enhancements that should be considered as being helpful in improving the overall success in HTS. Often there is a high rate of false positives in single-dose single-experiment assays (see Figure 1); partly, this is the risk of doing $n = 1$ experiments. Post-HTS analyses [15] are often further clouded by the screening of reactive species or optically interfering components (which can be the result of sample degradation) in biochemical assays [16•], the tendency of some chemicals to aggregate [17••] or to turn up as frequent hitters [18•]. Further, the selection of HTS hits to follow up from the primary assay often remains subjective, as the definition of a 'HTS hit' may depend on the available information and experience of the chemist assigned to the project. Totally new targets and the desire to not rule out possible hits may force chemists to select 'hits' at 30% inhibition, whereas well-patented areas and decades of medicinal chemistry experience, coupled with an established assay, will allow chemists to select hits at 80% inhibition. Probability schemes have been devised to assist this process [19]. Cheminformatics tools are increasingly used to handle the vast amounts of data from HTS [15] and to bring rigor to the process of looking for genuine leads.

The *Virtual*, *Tangible*, *Global* and *Real* worlds of molecules

Virtual, *Tangible*, *Global* and *Real* collections

In the ideal scenario, it would seem appropriate to reliably screen the maximum number of molecules that we can afford against every appropriate target to find the highest number of leads and ultimately effective drugs. While this is the only way to ensure that we discover all possible leads (that already physically exist) for all targets, this is highly impractical. For instance, there is no effective limit for the number of compounds that can be made or acquired. It has been estimated that there are far in excess of 10^{60} druglike molecules that could be made [20]. This vast number of compounds is referred to as the *Virtual Collection* of compounds because they cannot be all made, but they are essentially a 'resource' that can be mined as needed. Having appropriate informatics systems to access these virtual compounds via 2D, 3D and other property spaces is a key part of lead discovery strategies. Those compounds that can be reliably made via an appropriate chemical route can be designated as *Tangibles* because they could be 'easily' synthesized or acquired in a timely manner from a supplier as needed. The total output of the pharmaceutical sector (including academic and commercial resources) represents the *Global Collection* (see below). Most pharmaceutical companies have yet smaller collections ready for screening. These include the compound samples that have been accumulated over many years, as well as novel compounds acquired from external sources or produced in-house by automated facilities together with compounds made in lead optimization projects. These are the *Reals* (i.e. those discrete entities that physically exist within a company and are actually available for screening). The Virtual/Tangible/Real (VTR) description of compounds provides a framework for considering how we design and build screening sets.

The magnitude of the ultimate screening experiment: a thought experiment estimate

The issues in lead discovery are better understood by gaining insights into the magnitude of the problem that might need to be faced if HTS was taken to extremes. We estimate that fewer than 120 million compounds have been synthesized worldwide and could be available for biological screening if everyone pooled their resources. This is based on the fact that one of the largest collections of commercially available structures, ChemNavigator [21], covers ca. 12 million structures (8 million unique molecules — the remaining 4 million are overlapping structures made by different chemical vandors), of which 90% are RO5 compliant. The combined output from the pharmaceutical sector world-wide, the *Global Collection*, is unlikely to exceed 10 times that number, in terms of unique chemical structures. This sets an upper limit for the *Global Collection* if we had access to the contents of everyone's *Reals*.

For reasons related to inadequate storage, compound purity and stability, and considering the compound quantities, we further estimate that only 50% of the *Global Collection* (i.e. 60 million structures) could become available for screening. Given the current capacity of HTS robots (we assume 100 000 compounds/day), screening 60 million compounds on 5,000 targets at a single dose, single experiment level would take over 8 years, using 1000 HTS robots operating at full capacity. Five thousand targets represent ten times more targets than currently addressed by therapeutic agents (N = 483 [2]). At a few cents per assay for reagents, the entire effort would cost many billions of dollars (not including man-years, equipment, assay/target preparation and chemical preparation costs). The budget of this 'global HTS' effort would be comparable to the entire research and development budgets ($32 billion) of the pharmaceutical industry in 2002 [22].

It is not just the cost of this experiment in reality that is daunting. If 8 bits are enough to store single results, and 4 bits are required to store assay conditions (i.e. 12 bits/result), the results of screening the *Global Collection* would further require more than 3352 gigabytes of storage space. While such space is feasible these days, it is unlikely that current software has the capacity to effectively navigate through the entire dataset — although each target *per se* would require less than 687 MB of storage. By conservatively assuming a 0.1% success rate and 40% false positives (as in Figure 1), this effort could yield 180 million HTS actives, up to 3 million drug candidates and up to 300 000 new drugs. This thought experiment shows without a doubt that the current lead discovery paradigm could reach an unprecedented scale, but would require steep changes both in terms of logistics and financial support. Even if mergers and acquisitions world-wide led to a single, meta-pharmaceutical entity, this would still be an extraordinarily daunting task that would require drastic changes in the decision-making process and clarity in the prioritization of molecules at the chemical level.

The druglike and leadlike concepts

Druglikeness

Because the *Global Collection* is likely to remain unavailable for lead discovery in the next decade, medicinal and combinatorial chemists are exploring the VTR concept in an effort to explore *in silico* which *Reals* are sensible to have available to 'represent' the larger *Global and Virtual* spaces. As discussed above, chemical space is effectively infinite. A further simple example of this is provided by considering the simple case of substituted *n*-hexanes with 150 substituents [23]: Weininger estimates that all the possibilities, from mono- to 14-substituted hexanes, regardless of synthetic feasibility, amount to 10^{29} n-hexanes [23]. The search for 'lost and emerging chemistry' [24] aims at identifying molecular scaffolds that go

beyond rings with 6, 10–13 or 17 atoms. More effective methods are needed to decide which of these vast numbers of compounds to select as potential starting points and ultimately which have any prospect of being developed into drugs.

Chemical fingerprints can serve as the basis [25,26] for a crude computer-based discrimination between 'drugs', represented by WDI (the World Drug Index) [27], or by the MDL Drug Data Report (MDDR) [28], and 'non-drugs', represented by ACD (the Available Chemicals Directory) [28]. Although this result was reproduced by other groups [29,30,31•], it has yet to become accepted by the chemistry community as a decision-enabling scheme. If it was truly effective, it could assist chemists to quickly evaluate, for example, what other chemists have considered worthy of synthesis (and patenting) before them. The problem is that good druglike scores do not make a molecule a drug. It is often assumed that Lipinski's RO5 criteria define druglike space. However we showed that this was not the case [32], as there are more compounds in ACD, or 'non-drugs', that are RO5 compliant, compared with compounds from MDDR, or 'drugs'. A recent study by Vieth *et al.* [33•] looked at the differences in the properties of drugs having a variety of routes of administration and confirmed that oral drugs have properties associated with lower MW, fewer hydrogen bond acceptors (HAC) and donors (HDO), and fewer rotatable bonds (RTB) compared with drugs that have other routes of administration (see also earlier work [11]). Despite this extension to RO5 criteria, there remains a gulf between these crude rules of thumb and true discriminating power for specific design purposes. It is therefore more appropriate to think of the RO5-type criteria as necessary, but not sufficient to create an oral druglike molecule.

Leadlikeness

Unlike the druglike scores, where large numbers of chemical structures have been submitted to statistical analyses, the leadlike concept [34] is based on significantly smaller datasets [6,7,35•]. Despite this, the concept of leadlikeness is already having a significant impact in the design of chemical libraries [36••]. This is, in part, because the concepts and methods related to leadlikeness are very intuitive and fit with the current experience of what typically happens [37•] in lead optimization. Based on current data, it appears that, on average, effective leads have lower molecular complexity [6] when compared with drugs, as well as a smaller number of rings (RNG) and rotatable bonds [7], have lower MW and are more polar [34].

Rishton extended the leadlike concept [16•] by including chemical properties. He suggests that leadlike structures should bind only in a non-covalent, reversible manner, should show chemical stability toward proteins, and should not be 'promiscuous inhibitors' [17••], 'frequent hitters' [18•] or 'warhead' compounds [16•]. Rishton's 'warheads' include electrophilic 'suicide inhibitors', phosphates, phosphonates, hydroxamates and thiol 'chelators' (i.e. groups known to react with proteins under HTS assay conditions).

Implications for library design

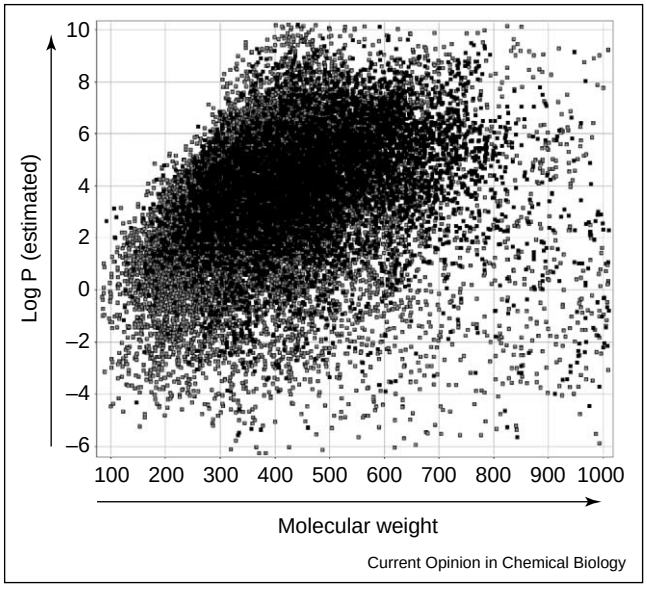
Having recognized that poor solubility and poor permeability are among the main causes of failure [38] in later stages of drug development (see also Figure 1), the medicinal chemistry community is now rethinking [39••] its drive to produce large, hydrophobic molecules by limiting these properties to values smaller than those suggested by Lipinski [11]. Our survey [8] of the chemical structures published between 1991 and 2000 in the *Journal of Medicinal Chemistry* [40] shows that 25.2% of the HAMs, or better than 10 nM, are large (MW $>$ 425 amu), hydrophobic (the logarithm of the octanol/water partition coefficient [41], LogP, is above 4.25) and poorly soluble (the logarithm of the intrinsic aqueous solubility, $\mathrm{LogS_w}$, is below -4.75). This should be compared with the 1.7% HAMs that are small (MW $<$ 300), significantly less hydrophobic (LogP $<$ 1.5) and soluble ($\mathrm{LogS_w} > -2$). Therefore, one can conclude that the benefits of the leadlike concept have yet to be translated into practice on a large scale.

As pointed out by Kuntz *et al.* [42] and confirmed in our earlier work [13], higher MW does not necessarily warrant higher activity. A close examination of the WOMBAT database [40] reveals that increased biological activity is not directly correlated [8] to an increase in size and hydrophobicity (Figure 2).

This result is relevant as one of the aims of combinatorial chemistry is ultimately to produce drugs, not leads [7]. The leadlike strategy, also proposed for virtual screening [43], has practical consequences for energy-based ranking of virtual hits [44], since an increase in the number of non-hydrogen atoms is likely to yield higher scores during virtual screening. Therefore, careful choice of virtual screening scoring schemes needs to be done if inappropriately large molecules are not to be selected by *in silico* screening for taking forward for real screening. We have also argued that such molecules actually have a lower chance of being hits because of the very high chance of getting interactions wrong in over-functionalized (i.e. large) molecules [6].

Placing our property-based analyses [8] in the context of preclinical drug discovery, we have formulated computational criteria for leadlike compounds [45]: MW $\leq$ 460, $-4 \leq$ LogP $\leq$ 4.2, $\mathrm{LogS_w} \geq -5$, RTB $\leq$ 10, RNG $\leq$ 4, HDO $\leq$ 5, HAC $\leq$ 9. Such criteria are expected to be applicable to chemical libraries during lead identification. However, the following experimental criteria, mostly

Figure 2



Size (MW) and hydrophobicity (LogP) in relationship to biological activity as captured in WOMBAT [40]: HAMs (6564 activities, black) are shown in contrast to low-activity molecules, LAMs (24124 activities below 1 μM, in gray); 61% of the LAMs and 41.6% of the HAMs can be labeled as 'leadlike' (MW $<$ 450 and LogP $<$ 4.5).

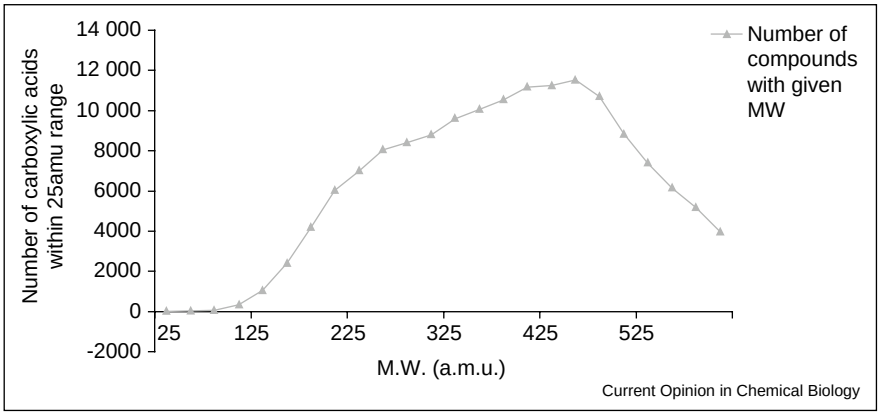
related to *in vivo* properties (e.g. in rat), become more relevant for individual compounds: bioavailability above 30%, low clearance (e.g. below 30 ml/min/kg in rat), $LogD_{7.4}$ (LogP at pH 7.4) between 0 and 3, poor (or no) binding to drug-metabolizing cytochrome P450 isozymes, plasma protein binding below 99.5%, lack of acute and chronic toxicity at the expected therapeutic window (e.g. assuming 500 mg/day P.O. regimen for 7 days), no genotoxicity, teratogenicity or carcinogenicity at doses 5–10 times higher than the therapeutic window. The experimental criteria should be applied to (most) compounds progressed from the lead identification to the lead optimization stage.

Developing leadlike screening sets

These and related concepts have led us and others to develop screening strategies that are complementary to more traditional HTS methods. Some companies (e.g. Astex [46], Plexxikon [47] and Vertex [48]) have gone so far as to have the concepts of screening fragments or very small leadlike entities (in connection with X-ray crystallography or NMR) as their principle lead generation paradigm [36••]. The general approach is to try to find start points for lead optimization that are more 'leadlike' and typically less complex than those derived solely on 'druglike' criteria.

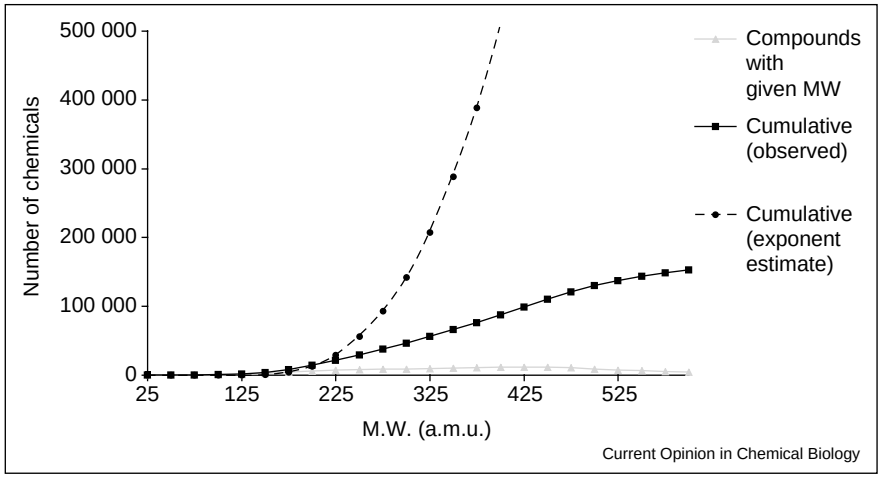
Another aspect of leadlikeness and reduced complexity that we have explored [49] concerns the sampling rates that can be achieved with *Reals* of a given complexity within the vast space of *Tangibles* or *Virtuals*. This can be explored with the aid of Figure 3, which shows the number of carboxylic acids (of all types) registered in the GSK registry system plotted as a molecular weight distribution. The gray curve shows the incremental number of acids in the collection for each 25 amu. increase and is effectively the rate of increase in the number of compounds in a particular MW range. The steep rise in the number of acids with MW follows an exponential curve initially, as expected since the number of *Tangibles* increases exponentially with the number of heavy (non-hydrogen) atoms in a molecule. However, at around 150 amu, the observed MW increase of these compounds ceases to be exponential. Why is the rise no longer exponential after 150 amu? Our explanation is that we significantly under-sample the potential carboxylic acids (i.e. the *Virtuals*), and that this under-sampling gets worse as MW and complexity increase.

Figure 3



Distribution of carboxylic acids in GSK collection.

Figure 4



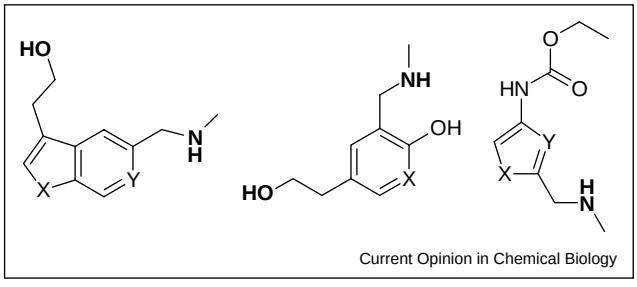
The potential number of chemicals with a carboxylic moiety, plotted against MW. Above MW = 350, there is an increased divergence between the number of observed compounds (*Reals*) and all theoretical combinations (*Virtuals*).

A different view is presented in Figure 4 of the same data. This combines the actual count of carboxylic acids within a given MW range (gray triangles, as per Figure 3), the cumulative observed count of these carboxylic acids (black squares) and a theoretical and truly exponential cumulative count (black circles). The *y*-axis is now changed to show the cumulative number of carboxylic acids (both real and virtual) that exist as MW increases. The precise numbers in the estimated exponent are not important, in fact there may be over 10^{10} carboxylic acids by MW of 450. The important point is that the Real (squares) and Virtual (circles) curves are dramatically diverging above about a MW of 300. Thus, at lower MWs the Reals represent a better sampling of the Virtual world compared with that at higher MW. Provided it is possible to find relevant biological activity associated with a chemotype in a lower MW representation, then using these molecules for screening purposes can provide an effective way to represent the more complex higher MW compounds that will ultimately result from lead optimization. This is an important aspect of the leadlike concept, and provides space for the process of lead optimization, in contrast to attempts to directly probe and find biological activity in the regions of exponentially larger numbers of higher MW compounds.

Starting points with lower MW are likely to have less potency and are not always clearly identifiable via HTS, if the screening concentration is typically of the order of 10 μM or less. The obvious solution is to screen compounds at higher concentrations (e.g. 500 μM), but this introduces problems related to solubility, purity and interference with readout (e.g. by fluorescence quenching). Nevertheless, with careful selection of compounds and robust screens we have been able at GSK to screen several targets (mainly enzymes) at up to 1 mM concentration and still extract useful information.

The so-called 'reduced complexity' screening set that we have used for this purpose was assembled using several computational criteria (e.g. average values for MW < 350, RTB ≤ 6, heavy atoms ≤ 22, HDO ≤ 3, HAC ≤ 8, ClogP ≤ 2.2), and matching certain 3D pharmacophoric patterns based on the GaP approach [50]. These criteria are similar to the 'Rule of Three' (MW < 300, ClogP < 3, HDO < 3, RTB < 3) proposed by Astex [51] for fragment libraries in lead discovery. The GSK selection criteria also require the presence of a 'synthetic handle' (i.e. chemical moieties that allow rapid synthesis of further analogues). Typical generic structures considered for the 'reduced complexity' screening set are shown in Figure 5. The choice of fragments for such a 'Reduced Complexity' library is best considered in the context of molecular recognition and the combinatorial chemists'

Figure 5



Examples of generic structures considered for the reduced complexity screening set. X and Y indicate possible heteroatoms. 'Synthetic handles' are shown in **bold**.

desire to work within a familiar chemistry space where robust reactions can be rapidly exploited. This needs to be balanced against the intrinsic and unique properties that such small fragments may have for exploring more space with less [52•]. Often, though, similarity searching for related compounds in the world of Reals (GSK compound collection and external suppliers) may provide an alternative and sometimes quicker follow-up procedure. Wherever possible we also aim to obtain experimental data on the binding mode of the compounds to the protein by X-ray or NMR methods.

Conclusions

In our opinion, the concept of leadlikeness will help refine the processes by which interesting starting points for medicinal chemistry can be found in a cost effective manner. We believe that leadlikeness is an integral part of the continual enhancement of the processes of HTS. It illustrates the use of conceptual and computational tools that are needed to avoid resorting to the heroics that would be needed if our 'all against all' thought experiment was literally followed to exhaustion. In looking for leadlikeness, one needs to exercise caution. Unlike the Planck constant, the cut-off values attributed to leadlikeness are context-specific: Should the absorption change from oral to inhaled these values would have to be adjusted to fit a different profile [53•]. Furthermore, the history of drug discovery abounds with counterexamples to the leadlike concept: tetrahydrofolate (MW = 574.5) served as lead for Methotrexate (MW = 454.4), and Tubocurarine (MW = 610.7) was the lead for Gallamine (MW = 510.8). As Rishton points out [16•], 'most drugs found in the compiled databases were classically discovered and developed using biological assays, selective cytotoxicity assays and animal models of disease, not using biochemical (e.g. HTS) assays.' In other words, these leads were optimized at a time where chemists could modify 1–10 molecules, have them screened and interpret the results before another design/make/test cycle would start. Today, there is a risk that high-throughput experiments reduce the opportunity for innovative and iterative thinking, as millions of molecules are screened simultaneously without the possibility of interpretation and analysis between the traditional rounds of experiments for this number of datapoints. We have to face the fact that the design/make/test cycle sometimes occurs only in the late stages of lead identification (secondary and follow-up screening), and mostly in lead optimization. The critical decision in preclinical discovery remains the choice of the lead compounds, which ultimately derive from what is in screening collections [13]. Therefore, the careful incorporation of the leadlike concept into screening collections becomes even more important.

Update

A new mathematical model for the design of a screening collection for HTS has recently been published by colleagues at GSK [54]. This model relates chemical structural similarity and cluster to biological activity, and hence to the probability of finding active lead series of compounds in high-throughput assays. The optimal screening collection content for a given fixed size of screening collection can then be derived and the competing demands of focussed and diverse sets explored. This is a good example of how mathematical modelling can be applied to bring further rigor and objectivity to understanding what should be in a screening collection and hence further improve our success rates in drug discovery.

Acknowledgements

MMH would like to thank Andrew Leach, Brian Evans, Jon Hutchinson, Peter Lowe, Chun-wa Chung, Dave Langley, Stephen Pickett, Darren Green, Gavin Harper and many other colleagues at GSK who have contributed to these concepts over several years. TIO would like to acknowledge the significant influence of Andy Davis, Simon Teague, Paul Leeson, Mike Cox and Mark Divers (AstraZeneca) in the evolution of the leadlike concept. The authors thank Drake Eggleston (GSK) for timely comments and pertinent observations regarding this manuscript. TIO was supported by New Mexico Tobacco Settlement funds.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Drews J: **Innovation deficit revisited: reflections on the productivity of pharmaceutical R&D**. *Drug Discov Today* 1998, **3**:491-494.

2. Drews J: **Drug discovery: a historical perspective**. *Science* 2000, **287**:1960-1964.

3. Horrobin DF: **Innovation in the pharmaceutical industry**. *J R Soc Med* 2000, **93**:341-345.

4. Sneader W: **Drug prototypes and their exploitation**. Chichester: Wiley, 1996.

5. DeStevens G: **Serendipity and structured research in drug discovery**. *Prog Drug Res* 1986, **30**:189-203.

6. 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.

7. 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.

8. Oprea TI: **Cheminformatics and the quest for leads in drug discovery**. In *Handbook of Cheminformatics vol. 4*. Edited by Gasteiger J, Engel T. New York: VCH-Wiley; 2003:1508-1531.

9. Brown F: **Cheminformatics: what is it and how does it impact drug discovery**. *Annu Rep Med Chem* 1998, **33**:375-384.

10. • Horrobin DF: **Modern biomedical research: an internally self-consistent universe with little contact with medical reality?** *Nat Rev Drug Discov* 2003, **2**:151-154.
The author questions the relevance of modern pharmaceutical research, by contrasting the lack of congruence between *in vitro* and *in vivo* models, and the diminishing interest in good clinical research.

11. 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.

12. Boyd DB: **Progress in the rational design of therapeutically interesting compounds**. In *Rational Molecular Design in Drug*

Research. Edited by Liljefors T, Jorgensen FS, Krogsgaard-Larsen P. Copenhagen: MUNKSGAARD; 1998:15-29.

13. Oprea TI: **Lead structure searching: are we looking at the appropriate property?** *J Comput Aided Mol Des* 2002, **16**:325-334.

14. Machin P: **Designing drugs - where next?** In *EuroQSAR 2002 - Designing Drugs and Crop Protectants: Processes, Problems and Solutions*. Edited by Ford M, Livingstone D, Dearden J, Van de Waterbeemd H. New York: Blackwell Publishing; 2003:1-4.

15. Oprea TI, Li J, Muresan S, Mattes KC: **High throughput and virtual screening: choosing the appropriate leads**. In *EuroQSAR 2002 - Designing Drugs and Crop Protectants: Processes, Problems and Solutions*. Edited by Ford M, Livingstone D, Dearden J, Van de Waterbeemd H. New York: Blackwell Publishing; 2003:40-47.

16. • Rishton G: **Nonleadlikeness and leadlikeness in biochemical screening**. *Drug Discov Today* 2003, **8**:86-96.
Excellent contrast between the druglike and the leadlike concepts.

17. •• McGovern SL, Caselli E, Grigorieff N, Shoichet BK: **A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening**. *J Med Chem* 2002, **45**:1712-1722.
This paper proves experimentally that some compounds active in HTS assays aggregate, forming particles 20–400 nm in diameter. These aggregates may, in turn, act as enzyme inhibitors.

18. • Roche O, Schneider P, Zuegge J, Guba W, Kansy M, Alanine A, Bleicher K, Danel F, Gutknecht EM, Rogers-Evans M *et al.*: **Development of a virtual screening method for identification of 'frequent hitters' in compound libraries**. *J Med Chem* 2002, **45**:137-142.
This paper evaluates frequent hitters using chemoinformatics tools.

19. Fogel P, Collette P, Dupront A, Garyantes T, Guedin D: **The confirmation rate of primary hits: A predictive model**. *J Biomol Screen* 2002, **7**:175-190.

20. Bohacek RS, Martin C, Guida WC: **The art and practice of structure-based drug design: a molecular modeling perspective**. *Med Res Rev* 1996, **16**:3-50.

21. The ChemNavigator homepage on http://www.chemnavigator.com

22. According to the PhRMA: (Pharmaceutical Research and Manufacturers of America) website on http://www.phrma.org/issues/researchdev/

23. Weininger D: **Combinatorics of small molecular structures**. In *Encyclopedia of Computational Chemistry*, Vol. 1. Edited by Von Ragué Schleyer P. New York: Wiley; 1998:425-430.

24. De Laet A, Hehenkamp JJJ, Wife RL: **Finding drug candidates in virtual and lost/emerging chemistry**. *J Heterocycl Chem* 2000, **37**:669-674.

25. Ajay, Walters WP, Murcko MA: **Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules**?. *J Med Chem* 1998, **41**:3314-3324.

26. Sadowski J, Kubinyi H: **A scoring scheme for discriminating between drugs and nondrugs**. *J Med Chem* 1998, **41**:3325-3329.

27. WDI: the Derwent World Drug Index, is available from Daylight Chemical Information Systems at http://www.daylight.com

28. MDDR and ACD are available from MDL Information Systems: http://www.mdli.com/products/finders/database_finder/index.jsp. MDDR is developed in cooperation with Prous Science Publishers at http://www.prous.com/index.html

29. Wagener M, Van Geerenstein VJ: **Potential drugs and nondrugs: prediction and identification of important structural features**. *J Chem Inf Comput Sci* 2000, **40**:280-292.

30. Frimurer TM, Bywater R, Naerum L, Lauritsen LN, Brunak S: **Improving the odds in discriminating ''druglike'' from ''nondruglike'' compounds**. *J Chem Inf Comput Sci* 2000, **40**:1315-1324.

31. • Brüstle M, Beck B, Schindler T, King W, Mitchell T, Clark T: **Descriptors, physical properties and drug-likeness**. *J Med Chem* 2002, **45**:3345-3355.
The discrimination scheme distinguishes compounds assigned to WDI (drugs) or Maybridge (nondrugs), using principal component analysis. The emerging Kohonen (neural net) map appears to discriminate between drugs and hormones as well.

32. Oprea TI: **Property distribution of drug-related chemical databases**. *J Comput Aided Mol Des* 2000, **14**:251-264.

33. • Vieth M, Siegel MG, Higgs RE, Watson IA, Robertson DH, Savin KA, Durst GL, Hipskind PA: **Characteristic physical properties and structural fragments of marketed oral drugs**. *J Med Chem* 2004, **47**:224-232.
The most exhaustive analysis on the properties of orally available drugs carried to date, this study also includes comparative analyses with prior publications.

34. Teague SJ, Davis AM, Leeson PD, Oprea TI: **The design of leadlike combinatorial libraries**. *Angew Chem Int Ed Engl* 1999, **38**:3743-3748.

35. • Proudfoot JR: **Drugs, leads, and drug-likeness: an analysis of some recently launched drugs**. *Bioorg Med Chem Lett* 2002, **12**:1647-1650.
This account is a clear illustration of the 'innovation deficit', showing that only 11 drugs launched in 2000 are actually innovative structures.

36. •• Davis AM, Teague SJ, Kleywegt GJ: **Application and limitations of X-ray crystallographic data in structure-based ligand and drug design**. *Angew Chem Int Ed Engl* 2003, **42**:2718-2736.
Pitfalls in atomic models derived from X-ray crystallography can influence structure-based design and virtual screening. Screening libraries based on small, polar templates (leadlike) and their use in virtual and HTS illustrate the complementarity between the two approaches.

37. • Kenakin T: **Predicting therapeutic value in the lead optimization phase of drug discovery**. *Nat Rev Drug Discov* 2003, **2**:429-438.
A pharmacologist's viewpoint on the possible dichotomy between *in vitro* research and intended therapeutic (human) use — which relates to the discrepancies noted by Horrobin [10•].

38. Lipinski CA: **Drug-like properties and the causes of poor solubility and poor permeability**. *J Pharmacol Toxicol Methods* 2000, **44**:235-249.

39. •• Kubinyi H: **Drug research: myths, hype and reality**. *Nat Rev Drug Discov* 2003, **2**:665-668.
One of the few papers that points out a common myth from drug discovery analysts: poor pharmacokinetic properties are *no longer* the main cause of attrition in drug discovery.

40. The WOMBAT: (World of Molecular BioAcTivity) database, is available from Sunset Molecular Discovery at http://www.sunsetmolecular.com

41. Leo A: **Estimating $LogP_{oct}$ from structures**. *Chem Rev* 1993, **5**:1281-1306.

42. Kuntz ID, Chen K, Sharp KA, Kollman PA: **The maximal affinity of ligands**. *Proc Natl Acad Sci USA* 1999, **96**:9997-10002.

43. Oprea TI: **Virtual screening in lead discovery: a viewpoint**. *Molecules* 2002, **7**:51-62.

44. Pan Y, Huang N, Cho S, MacKerell AD: **Consideration of molecular weight during compound selection in virtual target-based database screening**. *J Chem Inf Comput Sci* 2003, **43**:267-272.

45. Oprea TI, Bologa C, Olah M: **Compound selection for virtual screening**. In *Virtual Screening in Drug Discovery*. Edited by Alvarez JC, Shoichet B. New York: Marcel Dekker; 2004:in press.

46. Carr R, Jhoti H: **Structure-based screening of low-affinity compounds**. *Drug Discov Today* 2002, **7**:522-527.

47. Milburn MV: **Drug discovery on a proteomic scale**. Abstracts of Papers, 224th ACS National Meeting, Boston, MA, USA, August 18–22, 2002, COMP-042.

48. Moore J, Abdlu-Manan N, Fejzo J, Jacobs M, Lepre C, Peng J, Xie X: **Leveraging structural approaches: applications of NMR-based screening and X-ray crystallography for inhibitor design**. *J Synchrotron Radiat* 2004, **11**:97-100.

49. Hann MM, Leach AR, Green DVS: **Computational chemistry, molecular complexity and screening set design**.

In *Cheminformatics in Drug Discovery* Edited by Oprea TI. New York: Wiley-VCH: 2004:in press.

50. Leach AR, Green DVS, Hann MM, Judd DB, Good AC: **Where are the GaPs? A rational approach to monomer acquisition and selection**. *J Chem Inf Comput Sci* 2000, **40**:1262-1269.

51. Congreve M, Carr R, Murray C, Jhoti H: **A 'Rule of Three' for fragment-based lead discovery?** *Drug Discov Today* 2003, **8**:876-877.

52. • Fattori D: **Molecular recognition: the fragment approach in lead generation**. *Drug Discov Today* 2004, **9**:229-238.
The fragment approach, which has conceptually evolved from leadlikeness, is discussed in the context of molecular recognition and medicinal chemistry experience.

53. • Tronde A, Norden B, Marchner H, Wendel AK, Lennernaes H, Hultkvist Bengtsson U: **Pulmonary absorption rate and bioavailability of drugs in vivo in rats: structure-absorption relationships and physicochemical profiling of inhaled drugs**. *J Pharm Sci* 2003, **92**:1216-1233.
Theoretical and experimental studies show the property profile of drugs delivered via inhalation therapy — a significant market for the pharmaceutical sector.

54. Harper G, Pickett SD, Green DVS: **Design of a compound screening collection for use in high-throughput screening**. *Comb Chem High Throughput Screen* 2004, **7**:63-70.