Bioorganic & Medicinal Chemistry 20 (2012) 5324–5342

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Early phase drug discovery: Cheminformatics and computational techniques in identifying lead series

Bryan C. Duffy, Lei Zhu, Hélène Decornez, Douglas B. Kitchen *

AMRI, 26 Corporate Circle, PO Box 15098, Albany, NY 12212-5098, USA

ARTICLE INFO

Article history:
Available online 5 May 2012

Keywords:
Hit-to-lead
Cheminformatics
Chemical databases
Lead-like properties
Drug-like properties
High throughput screening analysis

ABSTRACT

Early drug discovery processes rely on hit finding procedures followed by extensive experimental confirmation in order to select high priority hit series which then undergo further scrutiny in hit-to-lead studies. The experimental cost and the risk associated with poor selection of lead series can be greatly reduced by the use of many different computational and cheminformatic techniques to sort and prioritize compounds. We describe the steps in typical hit identification and hit-to-lead programs and then describe how cheminformatic analysis assists this process. In particular, scaffold analysis, clustering and property calculations assist in the design of high-throughput screening libraries, the early analysis of hits and then organizing compounds into series for their progression from hits to leads. Additionally, these computational tools can be used in virtual screening to design hit-finding libraries and as procedures to help with early SAR exploration.

1. Introduction

1.1. Early stages of drug discovery

The drug discovery and development path from novel drug target to FDA approved drug to marketing of a novel molecular entity (NME) is long and expensive. The recent approval of Factor Xa anticoagulation drugs illustrates the potential length of time from biochemical target discovery (the 1970s) to the entry of investigational new drugs (IND) into the clinic (late 1990s) and finally approval of Factor Xa based-NMEs (2010).[1] In order to increase drug discovery efficiency, new technologies such as high throughput screening (HTS) and fragment-based drug discovery are used to identify molecules (hits) with a given biological activity. A greater understanding of new biochemical targets through genomics and chemical biology has also increased the number of novel drug targets for which biological screens are being developed. The combination of new biological screening technology and the information-rich environment created from modern computational methods combined with early in vitro ADMET tools provides an opportunity to speed the process of identifying new drug candidates. The first challenge is selecting the drug target and identifying hit compounds that have the best chance of forming the basis for chemical optimization and then ultimately translating into INDs and NMEs. High-throughput screening of diverse libraries is an often used approach to identify new lead series. However, the fairly low bar of identifying new hits series often proves elusive as illustrated by Macarron et al.[2] who reported that 50–84% of HTS make it to the chemical optimization stage. In this review, we will describe computer-based techniques that organize the data in a chemically intuitive fashion and which we believe will help to improve the success rate in hit-finding programs (such as virtual screening and HTS).

The fields of cheminformatics and computational chemistry are intertwined. In this issue, the editor has provided a thorough review of the definitions of the many tools that are loosely categorized under the term ‘cheminformatics’.[3] In spite of the possibility of broad applications throughout early drug discovery processes, often cheminformatics has been limited to the large-scale calculation of physical properties and then their application as filters for library design and experimental results. Clearly, filtering by biological and calculated properties is now a useful part of the screening campaigns, hit selection and other early drug discovery steps. However, these filtering steps are often performed without regard to the chemical relationships between data points. Collecting and sorting biological data by core-common substructures is a natural approach in lead optimization as it is at the core of traditional structure–activity relationships (SAR) in the medicinal chemistry toolbox. Grouping compounds according to their chemical similarity, clustering and chemical scaffolding calculations can also be valuable to help design screening libraries and to analyze very crude data. When experimental data is grouped with consideration for chemical relationships, we have found that some screening campaigns identify higher quality or more attractive chemical series for lead optimization.

* Corresponding author.

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.04.062

We begin this review with the description of a generic process for identifying hits and turning them into leads. We continue with an overview of useful cheminformatic techniques to design compound collections which result in an improved probability of identifying screening hits. Next, we emphasize the role of computational and cheminformatic tools that assist in the selection, validation and prioritization of screening hits into series and the SAR tools that will guide the transformation of these prioritized hit series into leads. We especially consider the integration of computed values with various experimental methods, a crucial area to de-risk the process leading all the way to lead generation. Today, the successful execution of hit identification and hit-to-lead steps results in optimizable lead series and eliminates non-druggable biology targets. Cheminformatics has become crucial to ensure that drug discovery resources are applied efficiently.

1.2. Definition of hits and leads

Most drug discovery programs now involve the process known as 'hit-to-lead' (or lead generation) to develop 'hits' that come from an initial screen where biological activity could be validated. The criteria for the selection of compounds and when to call them a hit or a lead vary from one organization to another. Features which impact hit and lead selection also vary based on organizational preferences. Examples of other factors which impact target product profiles are competitive environment and knowledge of the standard of care in the relevant therapeutic area.

1.3. From hit identification to hit series

In Figure 1 we illustrate some common steps in the hit-to-lead process and we particularly emphasize very early steps in the process because these steps. Often, a hit-finding library is assembled for use in many different biological screens. These libraries are assembled from various physical samples which may be purchased or the results of drug discovery programs. These libraries are then used as the 'input' to various primary screens. The screens are typically single-concentration experiments and those samples which meet some threshold signal (preliminary hits) are then subjected to secondary screening to validate the original measurement. These validated hits will undergo various chemical analyses to assure that each positive biological response is due to a single chemical structure. The hit selection process prioritizes and organizes the chemical structures so that the project team can perform a rapid SAR explorations and refine the list available lead series. This iterative process can be termed hit-to-lead. We use the term '*hit*' in the phrase 'hit-to-lead' to describe a molecule that binds to the intended biological target and promotes the biological response above a certain threshold of desired activity in a biological screen. Hit compounds are intended to be valid starting points upon which additional structural modifications can be made to improve effects on biological activity and drug properties such as off-target selectivity, pharmacodynamics and pharmacokinetic parameters are grouped into hit series. Lead molecules are representative examples of a large series of chemical analogs developed from a common scaffold identified in the hit series. Organizations will often commit additional resources to these hit series. The terms hit and lead have many possible modifiers. Table 1 provides a summary of compound descriptions encountered in hit and hit series identification. A preliminary hit from a primary screen will not justify significant additional resources until after validation steps. After the experimental measurement on a sample has been validated, other chemical analysis is required to confirm its identity. Even at this stage, a validated hit may require many additional studies to promote it to the status of a true 'hit'.

1.4. From hits to leads

The overall goal for the hit-to-lead process itself is the simultaneous selection of the many properties and characteristics that transform compounds from a hit series into a small number of compounds suitable for development and eventual use as drugs. This is a very broad description of a goal that has very specific requirements to be achieved. This generalized process involves the confirmation, expansion, selection, and transformation of the initially identified active compound series, hits from a high-throughput screen or other sources, into lead compounds possessing the chemical, biological, and physical properties suitable for lead development.[4]

The steps required to go from a hit to a lead compound are individualized depending on the project goals, but generally follow the sequence of: hit identification, activity validation, structure validation, clustering, hit expansion, early SAR exploration, Structure–property relationship (SPR) determination and lead selection. During the process of transforming a 'hit', which is a known entity, to a 'lead', which will translate to a new drug invention, there must be a concurrent development of intellectual property. The interplay of techniques in early hit-to-lead is exemplified in Figure 1. During the process, decisions are made as data becomes available and compounds progress through a funnel-like process. Table 2 provides a description of the decision filters that are commonly applied during the hit-to-lead process.

2. Designing hit-finding libraries to facilitate hit identification

Identifying a large number of hits is the critical first step in a drug discovery program. Different organizations may have different screening technologies and chemical libraries available. A common method for identifying hits is through a HTS campaign and that is the focus of this review. However, hits are often identified from other sources, such as virtual screening, and can be combined with HTS hits. Fragment-based screening[5] and natural products[6] have also been excellent sources of hits as starting points for lead optimization. The impact of a well-designed hit-finding libraries are obvious: if good compounds (e.g., in a novel chemistry space) do not exist in the screening library then they will never be explored. However, if too many inappropriate compounds are screened, then it is more likely that some good compounds will not be identified as hits and that inordinate amounts of resources will be expended in eliminating compounds that could have been avoided in the library design stage.

Often a common shortfall in hit-to-lead efforts has been to view the process as a linear one with various tools and techniques applied only after certain steps were taken. It is often assumed that the hit-to-lead campaigns begin at the end of the primary screen, but the planning of the screening library and goals of the lead program are important contributors to the overall success of the drug discovery process and must be considered sooner.[7] Until recently, the common approach was to apply chemistry knowledge, off-target receptor activities and toxicity evaluations or computational filtering techniques only after completion of a HTS of a very large database of compounds. More recently, clustering and scaffolding techniques have been introduced to optimize the process of compound selection for a hit-finding library and later for validating preliminary actives. In the case of experimental techniques, advances permit the evaluation of a larger number of compounds than previously possible resulting in the design of new hit-finding libraries. Experimental methods to screen large numbers of compounds typically are more error-prone and produce more hits. It is often too expensive to follow-up on all hits and therefore

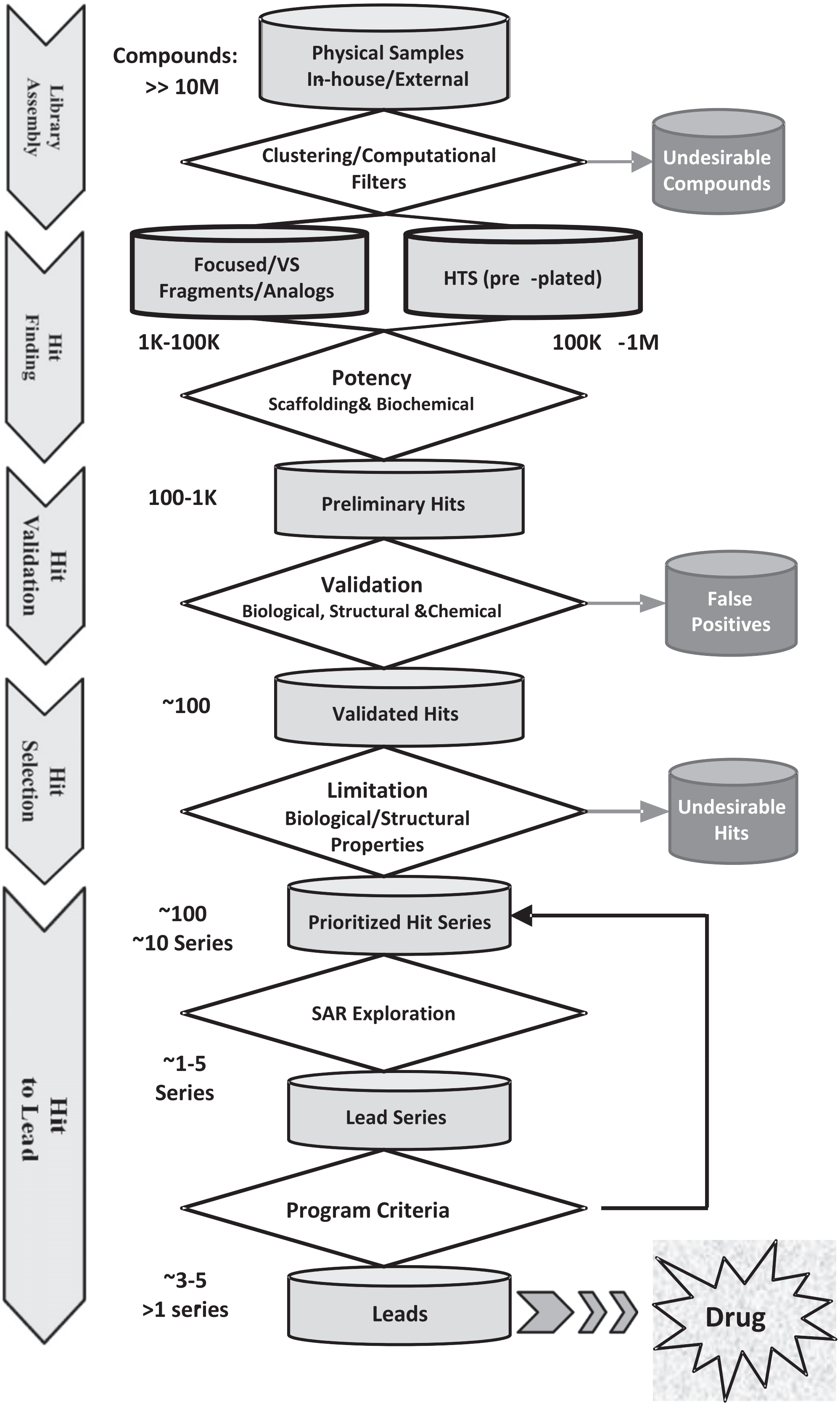


Figure 1. A schematic view of a typical small molecule early drug discovery process from library assembly to hit-to-lead exploration. A general description of compound sets along the process is given in Table 1. The filters applied at each step are detailed in Table 2 with common values recommended in the literature given in Table 3. In a typical program, assembled physical samples are filtered and clustered by structure. Compounds for which undesirable properties can be identified early are removed at this early stage. The remaining clustered samples are run through a virtual screen to identify promising candidates or through an experimental assay to determine potency. The active samples are collected as preliminary hits which become hit series after validation, prioritization, and SAR exploration. Further SAR exploration and successful optimization of properties according to program criteria result in lead compounds.

Table 1
Classification of compounds in the hit-to-lead process in Figure 1

Compounds set	Description and general considerations
Physical samples	Annotated with structure and properties for rapid searches. Potential sources of chemical samples are: In-house chemistry collections Historical medicinal chemistry projects External commercial collections Natural products
Undesirable compounds	Computational filters used to remove compounds with undesirable properties Computational filters frequently used: drug-like; lead-like; reactive groups; toxic groups
Focused databases (DB)	For use in conjunction with lower throughput assays. NMR; X-ray crystallography; full dose–response biological assays. Types: Analogs (similarity or common substructure) Fragments Virtual screening: docking, pharmacophore, similarity searches
High-throughput screening (HTS DB)	Large in-house collections of pre-plated compounds Subsets can be selected but all compounds on a given plate usually are screened
Preliminary hits	A set of compounds with positive readouts from the primary assay
False positives	A set of compounds with positive assay readouts resulting from artifacts in sample Examples: measurement interference (aggregation), degraded samples, irreproducible positive assay results from a repository sample or failure in an orthogonal assay
Validated hits	A set of compounds for which the assay activity can be attributed to sample interacting as intended in assay
Undesirable hits	A set of valid compounds possessing properties which would make them challenging to use as a starting point for a hit-to-lead campaign Reasons for de-prioritization include: off-target toxicity, off-target selectivity, ADMET liability, structure too simple to follow-up or too complex with a lack of handles for SAR expansion, IP positioning, etc.
Prioritized hit pool	A set of valid compounds prioritized for hit-to-lead campaign
Lead series/leads	Set of compounds resulting from hit-to-lead efforts. Compounds can results from analog sourcing or chemistry efforts. Application of program criteria results in a few lead series being followed with the goal of obtaining a handful of leads

Table 2
General description for the decision filters applied during early small molecule drug-discovery as depicted in Figure 1. For more detail on the filters see Table 3

Filter	Description
Computational	Structural: clustering, diversity, similarity Properties: lead-like, drug-like Liabilities: reactive groups, toxic groups, known off-target chemotypes, ADMET Structure-based and target-based prediction: virtual screening (VS) for activity or selectivity against one or a given set of targets. Molecular docking, pharmacophore searching, QSAR, etc.
Potency	Biochemical: Activity evaluation with a biochemical assay Computational (if not previously applied): Structure-based and target-based prediction: virtual screening (VS) for activity or selectivity against one or a given set of targets. Molecular docking, pharmacophore searching, QSAR, etc.
Validation	Biological assay readout duplicated. Higher level of validation includes replicating positive result with sample from alternate source and positive readout from an orthogonal assay Structural: sample has been checked to be reported structure Chemical: stability, solubility, lack of aggregation, etc.
Limitation	Biological: off-target activity or lack of selectivity Structural: Structure challenging to follow-up (too simple, too complicated, known in the literature or crowded IP) Property: poor solubility, poor ADMET Other: Lack of SAR among similar compounds, singletons with high MW Suspect reactive or toxic groups not previously identified
SAR exploration	Rapid analog identification in in-house and external libraries Synthesis of analogs Continuous filtering with cheminformatics tools associated with activity, potency and the lack of liabilities described above
Program criteria	Specific values for compound progression through various stages of testing in a given program to optimize activity, potency while maintaining a lack of toxicity or other liabilities

cheminformatics has become essential to ensure that expensive experimental resources are not used to validate undesirable hits.

2.1. Experimental considerations for Hit-finding libraries

HTS uses large libraries (often >1,000,000), in a miniaturized format (e.g., 396 well plates) and often fairly crude single-concentration screening methods to identify compounds with some weak activity against a target. The preliminary hit rates (defined by biological response at a single concentration) obtained from HTS libraries are quite low (less than 1%). The definition of a preliminary hit immediately after the primary screen is often based on objective criteria of the signal-to-noise ratio and then adjusted for the available resources for follow-up. Often the same cutoff is used during the preliminary HTS and the initial confirmation screen of the preliminary hit list. The hit-list (both preliminary and confirmation) can be greatly affected by the choice of the cutoff and in particular the treatment of compounds close to the cutoff. Compounds below the cutoff may never make the hit-list and preliminary hits and compounds just above the cutoff in the HTS may appear inactive upon confirmation event though there are only small differences in measured response.

2.2. Effects of library composition and experimental testing on hit-finding

Retrospective analysis of prior drug discovery processes from a diverse range of programs and organizations exemplifies the benefits of library planning. Several groups have analyzed large numbers of drug discovery programs that started with HTS campaigns and reported the factors that most often led to successful lead generation programs and their clinical success.[8–10,2] Macarron et al.[2] reported that 50–84% of HTS campaigns make it to the chemical optimization stage and the success of these campaigns has greatly improved over the last decade. Macarron et al. attribute some of the possible causes of failed screening campaigns project related issues that are independent of the screening process, for example, unvalidated biological targets, unexpected toxicities and inadequate animal models. Some failures may be due to experimental HTS systems (e.g., artificial or non-physiological screening methods) and insufficient confirmation resources. However, many of the causes of failed campaigns are related to the nature of compound screening collections. Among the remaining possible causes, it is also suggested that screening libraries may fail to yield good results because they are too small to cover the necessary diversity required of general purpose screening libraries, are non-drug-like and some programs lack sufficient 'informatics capacity'. The lack of broad chemical diversity in a screening library may mean that there are no active compounds in the library for a given biological target. We will present cheminformatic approaches that begin to address some of the causes of failure during the formulation of a screening deck. In particular, screening libraries can be improved by selecting more drug-like compounds and organizing compounds into structural classes and clusters. The latter technique can be used to ensure chemical diversity. As will be reviewed, clustering compounds also helps to balance the limited follow-up resources and provides a natural grouping of similar compounds with their biological responses. Given that the assembly of these large libraries is very expensive and requires highly specialized equipment for pre-plating it is important to improve the quality of the hits during the compound selection stage of library assembly.

Besides library assembly, assay methods themselves should be sensitive, selective, reproducible, and resistant to interference from biological, chemical, and mechanical sources. The design of a screening workflow giving considerations to reduce the errors from the specific screening methods used for the project is also a crucial factor for HTS campaigns. The screens are commonly complex cell pathway measurements, binding or enzyme inhibition assays. The miniaturization of an assay can increase its speed but decrease its reliability. Therefore, while an HTS run may generate large amounts of data suggesting that various compounds are potential hits (statistically significant changes in signal versus positive and negative controls), the values need to be verified in confirmation assays and orthogonal assays when possible to validate the observed activity against a target.[11,10] Various statistical parameters for the assay robustness (e.g., Z′ and Z-factor[11]) serve as accepted evaluation tools for the statistical quality of an assay but do not indicate assay sensitivity, that is the potential of a biological test system to identify weakly active compounds. Duplicate measurements and orthogonal assays, while costly, can increase sensitivity to weak hits.[12] The cost per well may affect the number of compounds that can be screened, which may prompt preselection of compounds by virtual screening in order to focus the chemical space on likely areas of interest. The use of virtual screening to preselect HTS libraries may result in lower screening resource cost, use of more robust screening validation assays, selective coverage of chemical space, and a theoretically improved signal-to-noise ratio of the potential hits. However, the efficiency and success of pre-selection depends on the quality of the virtual screen model and composition of the compound library whether a virtual database or a database of available compounds.

2.2.1. Exclusion of problem compounds

An early step in library selection is to remove compounds which are known to be reactive or likely to interfere in the assay conditions. The PAINS compounds[13] as well as other frequent hitters[14–16] are filtered by substructure matching and other similarity methods.[17] Pre-filtering of these known problematic compounds which are likely to be discarded at a later time is a time and cost saving measure for the HTS and hit-to-lead process.

All HTS methods rely on some type of physical detection method to record the output. Commonly detection is performed by either UV/visible absorbance or fluorescence techniques. Some compounds would trigger false positive or false negative results based on their spectroscopic profile.[18] Some less obvious screening compound characteristics are related to HTS mechanics in that compounds must be considered for their stability under the assay and storage conditions,[19] solubility in the assay solvent, and biological incompatibility in the HTS assay. If possible, compounds possessing these characteristics should be identified in the planning stage to prevent their interference in the HTS screening process. Pearce et al. found that their in-house collection was reduced by 12% when promiscuous and interfering compounds were eliminated and they provide a list of functional group filters that are easily applied.[20] Large HTS runs can be very sensitive to screening concentrations due to poor solubility of most compounds above 10 μM and the likelihood that most hits will be approximately >1 μM binders. Di and Kerns discussed implications of DMSO insolubility for bioassay measurements including lower hit-rates (less compound in solution at the screening concentration), impurities that produce more false positives and variable data.[21]

2.2.2. Effects of structural and physical compound property profiles

The contents of an optimal screening library are generally desired to be smaller, lower molecular weight compounds, because many of the physical properties which predict druggability[22] have rather strict upper limits. But the availability of compounds can also play a large role in library design. Often, million compound libraries are acquired at a low cost per sample which generate a preference for synthetic cost rather than drug-like value. Alternatively, the compounds may come from medicinal chemical programs within an organization. While filtering to eliminate all unnecessary library compounds has an idealistic appeal, it may not be practical within available and affordable chemistry space.

HTS libraries are routinely evaluated using calculated properties. The choice of cutoff for properties is usually at the lead-like rather than drug-like level. This preliminary filtering is performed, because the goal of the HTS is to select compounds that are ideal for optimization to drug-like compounds, not to identify compounds that are drug-like themselves.[23] The inclusion of drug-like compounds should be kept to a minimum with the primary exceptions being when these compounds are the sole representative of a unique series or highly similar to known binders of drug targets. Using lead-like criteria for selection reduces the size of the compounds in the potential library and provides compounds which have better development potential. Some organizations use side-samples from previous programs to build their HTS library. These samples often meet drug-like criteria but have already surpassed the lead-like properties criteria. Fragment based drug discovery pushes the minimalistic size argument to an extreme. One of the arguments for screening with a library of very small molecules, known as fragments, is that these compounds have more room for modification. However, if a compound is small its activity is likely comparatively low and may be overlooked or missed entirely by typical biochemical screening methods. In the case of structure-based fragment screening, binding events are detected at fairly high concentrations that would be impractical for use in HTS runs. In contrast, the challenge in the development of larger, drug-like molecules (e.g., larger than determined by lead-like physical properties) requires that considerable experimentation with substituents in a possible series will be needed to drive the activity to high-potency, including the replacement or even complete removal of functional groups. This usually results in a more costly and labor intensive approach to early SAR development.

2.2.3. Investigation of structural compound properties and clustering

Compounds in the HTS libraries that may have met the early lead-like filter criteria may still contain overly complex scaffolds. A scaffold can be viewed in terms of the composition of the smallest core ring system substructure that is common among a set of the compounds. Different metrics are used to judge both the synthetic and structural complexity of compounds.[24] The complexity of the scaffolds contained in an HTS library may impact the choice of leads for future exploration. Overly complex hit structures can result in challenging chemistry, thereby potentially slowing SAR development.[25] The number of chiral centers can make synthesis difficult and interpretation of biological activities somewhat more complex since enantiomers and diasteriomers will usually have differential in vivo and in vitro biological effects. Scaffolds included from an internal medicinal chemistry program library can be small and yet contain a high level of complexity, while purchased scaffold libraries are commonly less complex. Complex library compounds developed for previous in-house programs likely have established chemistry, which facilitates the task of synthesizing future analogs and negates the synthetic difficulty concern. In the case of HTS libraries assembled from external sources, the scaffolds for related hits can represent large common core structures which arose from external high-throughput combinatorial syntheses. Although these hits are synthetically accessible, these library compounds may also be undesirable as starting points for a lead series, because they may possess more properties and chemical moieties for synthetic ease rather than lead-like properties. The authors recommend against the inclusion of overly complex scaffolds in screening libraries, unless the synthesis is previously known or the scaffold syntheses is based on an obtainable natural product.

There is a very low likelihood of directly finding a final drug molecule during a HTS program. Hit-to-lead optimization usually involves the addition of functional groups, ring systems, hydrogen bond donor and acceptor groups, and generally increases the molecular weight. Similar structural modifications of already drug-like compounds would negatively impact the drug-like properties of the molecule making them unsuitable as a lead compound. Different optimization approaches are required for drug-like compounds and would involve instead the removal of functional groups, ring systems, etc. Unfortunately, this reverse approach often leads to more onerous synthetic approaches yet result in compounds which would have been unsuitable for consideration as lead-like compounds. The screening of drug-like library compounds thus increases the cost of the HTS, but adds limited value to the hit-to-lead process or overall program. In addition, complex or drug-like library compounds of interest can be re-incorporated at later stages into the hit set by follow-up clustering and SAR expansion during hit development.

2.2.4. Balancing requirements for overall diversity and redundancy of scaffolds

From an experimental perspective, after the initial HTS is run, the next step is to verify the primary screen measurements by duplicate screens, orthogonal assays and/or a concentration response experiment. Other follow-up experiments may be cell-based toxicity measurements and/or interference measurements to eliminate false-positives. This process is often performed robotically however, the steps are usually resource-limited. It is important to note that this early validation step is usually only a confirmation of the original single-concentration measurement by performing a dose response of a biochemical screen, and measurements on HTS samples that do not reproduce their original values, also known as false positives, are often eliminated from further consideration. The result of the confirmation process should not be judged by the number of potent compounds but the number of surviving structurally similar series or starting points for the next steps in hit-to-lead. McFadyen et al. discusses the importance of hit confirmation within clusters to eliminate false-positives and more importantly to increase the recovery of false-negatives.[26] The 'Top-X' problem occurs when a single cutoff is chosen for a screen based on statistical analysis of the data or based on the number of samples that can be processed by confirmation experiments. Because error bars are larger in high-throughput single-concentration measurements than confirmation assays there is a significant risk that resources will be expended on uninteresting and undesirable samples while false negatives that are just below the cutoff will be missed. The use of cheminformatic clustering helps to direct confirmation assays towards chemically interesting series and lifts some hits that are measured just below the Top-X criteria, or are false-negatives, onto a confirmation list. Therefore, a well designed library includes clusters of similar compounds in order to make this early screening task more productive.

A major consideration of library planning is to ensure scaffold redundancy, an often overlooked tool in library design. Most HTS libraries are assembled for repeated use against a wide variety of biological targets. Testing large HTS libraries with good coverage of chemical space is important to identify numerous hits across widely varying biological targets and this favors very large and chemically diverse libraries. Chemical similarity calculations using descriptors and fingerprints provide numerical estimates of similarity across different databases of compounds.[27] However, when comparing two large databases the calculations to determine similarity between all pairs of compounds can be too time-consuming. Additionally, the results from a numerical evaluation of similarity, though systematic, may lack the intuitive interpretation that human analysis provides. Recently, many systematic approaches to cluster compounds by common substructure components have been developed. These methods rely on systematic identification of ring systems within chemical structures through algorithms. ScaffoldTree,[28] HierS[29] and similar analysis allow the fairly rapid analysis of databases with linear computational time dependence and a chemically intuitive assignment of structurally-related clusters. The combination of similarity and scaffold analysis can expedite the selection of diverse compounds for the enhancement of an HTS set.

There is always apprehension in HTS concerning the possible omission of a key scaffold due to false negative or very weak positive results. Early work by Nilakantan et al.[30] indicated that by proper design of a library diminishes the probability of missing a possible hit series and can be made negligibly small by including 50–100 representatives for each compound class in a screening library. The total size of a library and chemical diversity of a screening deck may become unsuitable or too expensive since there are likely in the millions of possible scaffolds. More recent work suggests ways to balance the number of scaffolds and the number of representatives of each library in order to maintain core scaffold redundancy and have a statistical advantage in the hit-to-lead process.[31] These methods of library design take advantage of the principle that any active cluster (i.e., active series of similar compounds) is expected to be composed of approximately 20–30% active compounds based on similarity property principle.[32,33] Therefore, in order to identify a series as possibly active, only one of compound from a series needs to be measured and confirmed as a true hit. These analyses help to pose the problem of hit-series selection in general enough terms so that libraries can be designed with multiple analogs in mind while keeping the overall size of the library workable for screening. Examples of various series and measurement outcomes are schematically represented in Figure 2 and will be discussed in more detail in Section 3.2.1, below.

Singular representatives of a scaffold or cluster, otherwise known as 'singletons', are potentially the only representative of a starting point for a series and therefore their inclusion in a library can greatly increase the diversity of a library. For a singleton to be measured as a hit requires that the screen work correctly for that well, that is the compound is soluble and every other aspect of the screening system works well. If a singleton is a false negative, there may be no way to detect this compound as a valid starting point. Therefore, enhancing singleton clusters can improve the performance of a library though this enhancement needs to be balanced with the need for controlling the size of the collection.

We find that using cluster and scaffold analysis on all hit-finding libraries is important to the success of hit-identification. Likewise, a docking calculation has inherently large errors and grouping by chemical similarity and scaffold provides some additional confidence that if a scaffold is active one member will be detected during the experimental screen. Likewise, the availability of samples is often low and requesting multiple examples of a scaffold is more likely to produce sufficient samples for testing to validate the activity of the entire series.

2.2.5. HTS Libraries as dynamic compound collections

Ideally, new compound collection would be assembled and plated after careful consideration of the requirements of the program. Realistically, organizations have already assembled large libraries over many years and continually acquire new samples as either replacements or additions. At times it may be desirable to eliminate unwanted samples, but this is often impractical within the restrictions of robotics and plating schemes. HTS libraries are

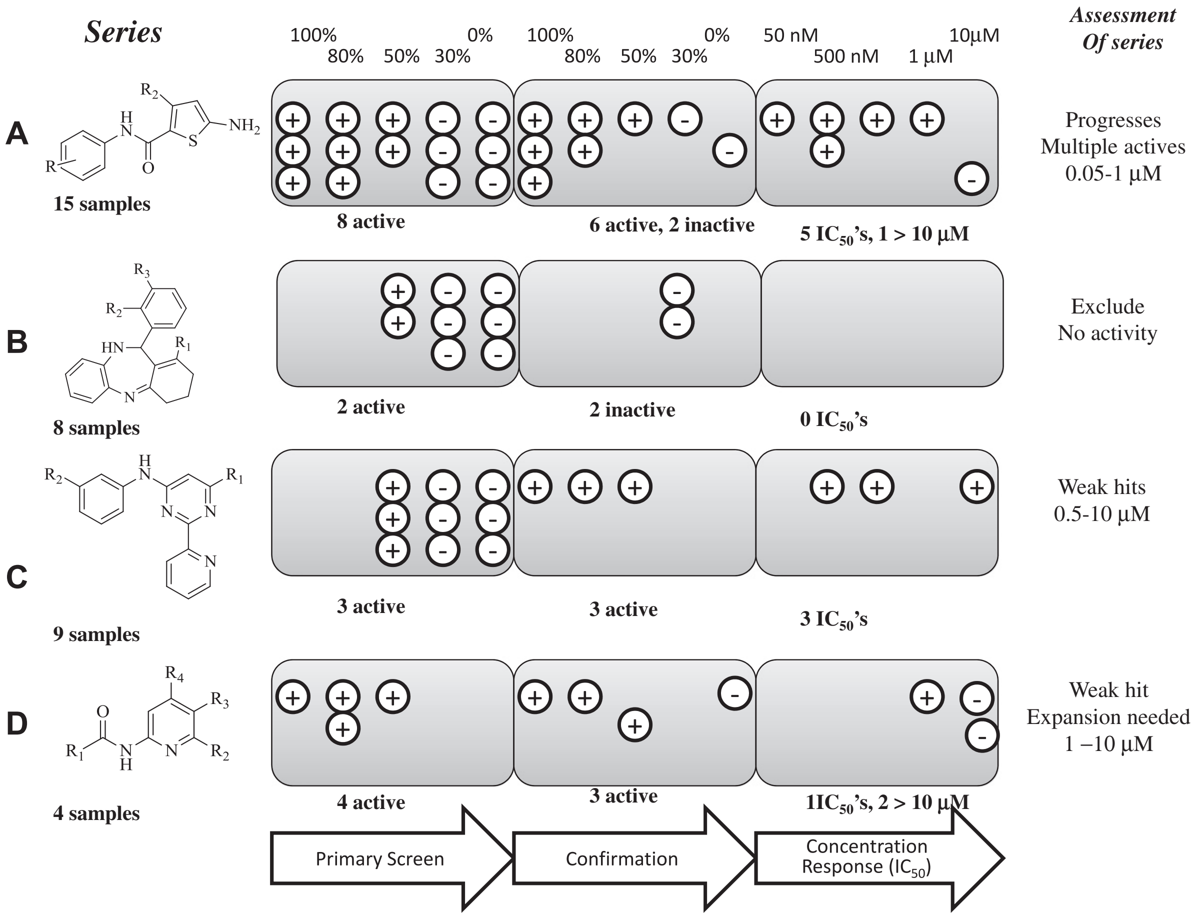


Figure 2. A hypothetical screening set progressing through typical triaging processes: primary screen, confirmation screen and determination of concentration–response curves (e.g., an IC_{50}). Hypothetical measurements on compounds are represented as small circles within each of four clusters of compounds. The plus (active) or minus (inactive) signs represent the assessment of each measurement based on a cutoff at each stage and represent positive and negative signal in a biology screen. Only compounds labeled as possible actives proceed to the next level of hit confirmation. For each box the most active compounds are to the left and the activity scale is indicated at the top. For the primary and confirmation columns, the scale represents a hypothetical activity measurement at a single concentration. The right column provides an assessment of the series based on the IC_{50} values. Any active series (i.e., active series of structurally similar compounds such as series A) is expected to be composed of many active compounds. In order to identify a series as possibly productive, only one of the compounds from a series needs to be measured and confirmed as a true hit, with multiple active compounds increasing the likelihood that activity is accurately labeled. When going from one screen to the next, the distribution of activity measurements for the compounds can shift due to normal experimental uncertainty. Series A is unlikely to be missed as active because several compounds are true positives and it is highly unlikely that all will be mis-measured. In series B, if this negative result was the result of experimental uncertainty, it would be difficult to recover the series. In all likelihood, Series B is not an active series worth pursuing. Series C is a latent hit series and three weak IC_{50}'s are obtained. Among the series, series D has the greatest risk of being mis-assigned as the only active compound in the primary screen tested as inactive in the confirmation screen. However, the distribution of results suggests that it is a latent hit series since there are several compounds that may be weakly active at the various measurement stages.

dynamic collections of compounds which can be easily enhanced by augmenting with compounds from under-represented areas of chemistry but which cannot be as easily curated due to practical considerations (e.g., removing a single compound from a plate is not cost efficient but removing an entire plate of compounds is). There are many reasons for the removal of compounds and include most frequently: exhausted follow up supplies or newly discovered undesirable properties (solubility, toxicity, etc.). Compounds to augment a library may be acquired from a variety of internal and external sources and usually are filtered to remove those with problematic structural features or with poor physical properties. HTS libraries are usually not static and require the addition of interesting compounds that are complementary to the existing screening collection. This complementarity is usually expressed by a combination of enhanced chemical diversity and augmented scaffold series both of which can be assessed by using chemical similarity measures and identification of chemical scaffolds. This approach was used to successfully enhance our own screening collection of ~100,000 compounds.[40]

2.2.6. Other types of compound collections

Other starting points for drug discovery programs can bypass the need for screening of large compound databases. Fragment based drug discovery overcomes some of the limitations of screening large libraries by covering diversity space thoroughly in a much smaller number of simpler compounds (tens of thousands of compounds versus millions in HTS).[34] Fragment hits can also be excellent starting points because of their very small size. Furthermore, screening techniques based in structural biology such as X-ray crystallography and NMR provide additional information about the nature of the bound complex.[35,36] Virtual screening of databases of compounds relies on docking or ligand-based methods to identify compounds for acquisition and offers higher hit rates versus HTS. Lastly, active analogs or scaffold-hopping may be used as a program starting point where the hit compounds or known drug molecules have been disclosed.[37]

Mayr and Bojanic provide analysis of the advances in HTS describing the pros and cons of combining HTS and other starting points such as iterative, fragment and virtual screening. They consider the potential experimental protocols and the role that each plays in early hit identification and how advances in robotics help to make these combinations more practical than in the past.[38] An excellent case study can be found in a recent compound library design for neglected diseases.[39]

The careful analysis of HTS generated hits usually delineates the official start of the hit-to-lead process. If purity, compound structure, and biological activity with a suitable signal-to-noise value have been rigorously determined, the hits are carried on. Otherwise, the HTS library screening hits need each of these properties rigorously evaluated before progressing them further.

3. HTS results and hit treatment

3.1. Screening metrics

The overall quality of HTS and virtual screening results should always be considered when analyzing the hit compounds. Statistical methods are commonly applied in virtual screening to evaluate the predictive power of models. These methods include the calculation of Receiver Operator Characteristic curves (ROC curves),[41] various signal-to-noise calculations, and selectivity statistics to evaluate the predictive power of the virtual screen.[42,43] Most of these methods rely on the use of internal positive and negative controls and the use of decoy compound sets to quantitate their accuracy and signal-to-noise ratio.[44]

3.2. Biochemical and biological validation of potential hits

3.2.1. Initial confirmation

The initial confirmation of a biochemical result may be expensive in terms of supply of compounds or biological reagents. The number of hits that can be confirmed is a practical limitation and the quality of the signal from HTS is an additional consideration in selecting which samples to confirm. Reviewing the primary screening data and the reconfirmation in biological screens is best done by collecting the hits derived from HTS, virtual screens, and any other screening technique; sorting by calculated properties and grouping by chemical structure. The merged hit-set derived from virtual screening, HTS, and fragment screening is useful to understand the chemical space available for a given biological target. Top-X approach (e.g., sorting only by primary screening measurements) often overlooks the early SARs in the data. Therefore, a preferable method to optimize the hit confirmation set is to cluster compounds by similarity calculations and/or scaffolds so that structural overlaps can be carefully examined.

Confirmation of the initial biochemical measurement is often done first from the original, pre-plated wells. This step of hit-picking can be somewhat laborious and the follow-up assays may be expensive in terms of time and cost of chemical and biological materials. Therefore, the clustering and filtering step helps to optimize the confirmation list. The concept of latent hits was defined as compounds which can be easily derived from a weakly active compound that is observed in a screen.[45] Varin et al. have illustrated the more general principle of 'latent hit series' where this series is based on compounds that individually would not be considered 'hits' based on activity thresholds but for which the results indicate that there may be an underlying active scaffold when considered collectively.[46,47] The value of sorting compounds into chemically related groups has been found to be useful in identifying hits and helps to overcome some of the inherent experimental difficulties in screening measurements. When samples are measured during a confirmation experiment, the results are expected to show normal experimental variation where a compound could be measured at higher or lower values. Near the hit-cutoff threshold, the confirmation runs can be particularly susceptible to switching compounds from 'active' to 'inactive' labels in spite of rather modest changes in measured responses. Formally, the primary assay values may be replicated within experimental uncertainty but the list of compounds that are 'confirmed' as hits may change drastically with large numbers of compounds falling off the hit list, that is false-positives will be eliminated more readily and false-negatives will not be recovered.

We use Figure 2 to illustrate the importance of analyzing hit-finding libraries and their 'preliminary hits' by chemical similarity and scaffold. The hypothetical test results are intended to show the effects of experimental uncertainties and the value of grouping experimental data together with chemical series. This association of data and structure permits a more nuanced decision making process. Series A is clearly labeled as a hit series regardless of whether clustering or Top-x is used to process compounds through hit validation. Series B indicates a generally weak series and probably best labeled as inactive. However, upon confirmation two compounds may have very weak activity. Series C contains one fairly potent hit and two weak hits when viewed by IC_{50} results. However, the SAR in the primary and confirmation screens indicate the presence of additional weakly active compounds. Series D has one surviving member with a measurable IC_{50}. The series will require more early exploration but supporting data from the primary screen and confirmation results suggest that there is some activity in the series and may well be a latent hit series. We do not illustrate a true singleton case. In a screening setting, a compound which contains a unique scaffold may fall slightly below the arbitrary cutoff for a screen. In order to identify this potential false negative, the data set can be sorted by the number of compounds in a scaffold or cluster and then by measured activity. Singletons need to be identified so that if their measured value is near the statistical noise level, then they can be incorporated into the confirmation stages. Singletons with modest activity may be of much greater interest for confirmation studies than additional compounds from more highly populated series. Statistical enrichments can be calculated (such as described by Varin et al.[46]) in order to highlight groups of compounds which show a higher than expected level of activity.

3.2.2. Structural confirmation

Once a group of hits are obtained from a HTS, or other hit generation method, confirmation of the structure of the hits is necessary. The structure confirmation process involves two parts. The structure that is believed to have been tested must be compared to physical data to be sure the initial structure assignment was correct. More importantly, the compound eliciting the observed assay response must be confirmed to be identical to what is believed to have been tested. The former challenge is one of synthetic and analytical chemistry while the latter is a control for possible reactions during the complex HTS assay process. The re-synthesis step can be particularly expensive but if a cluster of hits is obtained with a common core structure, only a few of the hits in the cluster need to be re-synthesized to validate the group of compounds as a possible series. By properly organizing hit-lists, scaffold and similarity analysis can help to make this particular step in hit validation much more efficient.

3.2.3. Identification of false positives

False positive identification is performed at this stage. It is usually advantageous, and highly advised, to begin re-synthesis, or re-isolation in the case of natural products, of the hit compounds. Retesting of the freshly prepared, or isolated, compounds often eliminates many commonly observed idiosyncratic false positives that are costly to follow-up. Whether using material from compound library stocks or re-synthesis, additional biochemical analysis may be required. For example, the Hill-slope and shape of an IC_{50} curve may be evaluated to look for non-stoichiometric issues.[48–50] Additionally, a lower throughput orthogonal detection biological assay may be needed to ensure that compounds bind to the expected protein or interfere in the correct portion of a pathway.[31,51] Additional experiments that use detection of physical binding such as NMR,[52] Surface Plasmon Resonance (SPR)[53] and X-ray co-crystallization with a target protein can unambiguously provide validation of the original hit.

3.2.4. Project specific property evaluation

Although most of the compounds with obviously reactive functional groups would have already been filtered from the library, libraries in many disease areas retain some reactive groups. In addition, some groups may become reactive by nature of the biological target or through metabolism. Covalent binders, soft drugs and/or pro-drugs[54] are sometimes appropriate choices and therefore strict filters are not always applied. In general, most drug discovery programs are seeking non-covalent mechanism; therefore, a secondary, more rigorous reactive group evaluation can be performed by calculation methods. Another example is the presence or absence of an acid group which is a desirable feature when intended against protease targets but becomes a liability for central nervous system (CNS) targets since acid group are correlated with a poor penetration of the blood brain barrier. Properties such as electrophilicity, nucleophilicity, and Fukui functions[55] can be used to estimate local reactivity. Manual structure inspection should also be performed on the hit compounds to look for less obvious modes of unintended reactivity.

The high-throughput measurement of ADMET properties is one area that has seen tremendous improvements over the past decade. As a result, high-throughput ADMET assays are now available for many property evaluations. The results from these assays and experiments are now routinely considered in early hit-to-lead development and the hit selection process. The incorporation of these measurements has the advantage of rapidly identifying areas of strength or of weakness in a hit compound. This additional data can impact both the prioritization of the preferred hits as well as the selection of additional compounds for secondary screening.

3.3. Incorporation of virtual screening and high throughput screening data

When information about the biological target such as X-ray crystal structure or NMR structure is available, virtual screening is usually performed in parallel with other HTS methods.[56] Virtual screening is often viewed as a competitor to HTS methods; however, as discussed in other analysis, the experimental screening and virtual screening methods are complementary.[57] Common practice is to assemble the largest, most diverse 'virtual' library practical before virtual screening is performed. The starting virtual compound library is filtered by a similar set of rigorous criteria as was used in filtering the HTS library before actual screening is started. Large virtual libraries can be used when the computational cost of the model is low. Ligand-based search methods such as substructure, pharmacophore and similarity searching are all used for extremely large libraries. There are now many comparisons of docking and ligand-based methods which allows reasonable guidance on the choice of computational search. However, with some computational approaches such as complex docking methods[58] extensive filtering may be necessary to select smaller libraries to compensate for the higher computational cost. More complex virtual models often provide the benefit of more accurate and reliable results despite the initially smaller library size. Commonly, the hit structures from the HTS are also evaluated in a virtual screen model as a confirmation of the virtual model and to provide calibration and possible structure predictions.

The merged hits from HTS and VS are clustered using cheminformatic algorithms[59] such as molecular fingerprint analysis, scaffold analysis, substructure searching, or other similarity calculations. The application of cheminformatic scaffold similarity searching, pharmacophore searching, and scaffold cluster matching with expanded global libraries are also used to re-examine a hit set for possible false negatives that are structurally similar to hit compounds. As previously discussed, it is important to maximize the statistical benefit of screening a series of related hit compounds, rather than unique, singular hits.[60] A possible hit series can be statistically inferred from rather crude primary data and therefore justifies increased experimental effort on the series.

Any newly introduced or recaptured compounds from the HTS should undergo all of the previous hit validation tests if the compound is physically available through purchase or synthesis. Any virtual screen hits should be tested in the biological screening assay if they are readily available through purchase or synthesis. Additional computational analysis such as docking and similarity calculations can be used to prioritize compounds for acquisition. Often, a small set of compounds that are similar to hits are visually inspected and selected to enable a rapid SAR exploration. The hit set is generally considered complete, and at its maximum size at this stage.

4. Selection and prioritization of hits and hit series

4.1. Calculated properties

Cheminformatic physical property are usually pre-calculated and are useful to provide a profile of descriptors for each compound in a hit set. These calculations fall under three categories: (1) constitutive properties descriptors, such as molecular weight, molecular formula, and heavy atom count; (2) calculated prediction of physical properties, such as solubility log*P*, and (3) other properties which are not physical observables (such as polar surface area) and/or are dependent on choice of geometry (dipole moment). In these cases, calculation speed may be a factor and in the case of PSA, the calculation is often replaced by topological polar surface area (tPSA) which relies on functional group look up.[61] The estimation of physical properties can be problematic[62] and therefore calculated values of many properties such as log*P* (octanol/water) depend on the software used. Computational methods to predict solubility are available[63–66] but solubility can be particularly challenging to calculate because this property depends on lipophilicity, hydrophilicity (usually log*P* is a surrogate for these two values) and the crystallinity of a sample, all properties which are difficult to predict accurately.[67] Because of these subtleties, filters and scores based on computed physical properties must vary. Most chemical software packages also include many additional advanced descriptors such as liver microsomal metabolism liability,[68] blood–brain barrier penetration,[69,70] cell permeability[71,72] and efflux,[73] oral bioavailability,[74,75] and ease of synthesis.[25] These calculated properties are usually derived from various knowledge bases which are often of limited scope. Therefore, quantitative predictions outside typical drug-molecules can be quite inaccurate. Nevertheless, the underlying parameters are often informative. Calculated permeability estimates (blood–brain barrier, intestinal absorption and parallel artificial membrane permeability assay (or PAMPA) are often derived from complex equations that factor in number of rotatable bonds, polarity, log*P* and other parameters that reflect the difficulty of the passive permeation of compounds through micellar, charged membrane which requires desolvation, attraction to a anionic surface, passage through a hydrophobic, rigid membrane channel and then re-solvation after exiting the intracellular membrane surface. While some descriptor calculations are relatively accurate, other calculations can be highly variable. A good practice is to employ descriptors from several software packages as either single selected values or in an averaged form and then for the end users to be cognizant of the potential type and scale of errors for each value.

The nearly universally calculated physical descriptors used in the hit-to-lead process include: molecular weight, log*P*, log*D*, log*S*, topological polar surface area, and hydrogen bonding group counts. It has also been observed for certain classes of compounds that rules may need to be tightened or a certain subset of the rules applied. For example, in recent work, the effect of carboxylic acids in drugs has been cataloged and the various limits adjusted to account for the consistently poorer properties of carboxylic acids in drug. There are many biological targets which require a carboxylic acid or its isostere but the increased polarity and lower solubility introduced by the group need to be counteracted elsewhere in the molecule.[76] We have attempted to collect as many experimental and calculated properties and their accepted range in Table 3. This table also includes criteria used in a recent publication to define a lead series.[77] It should be noted that filters can be much looser for preliminary hits than for leads and therefore we have included some frequently used cutoff values for each stage. It is often preferable to use descriptors to sort compounds rather than using them to strictly filter out compounds. There are several reasons for sorting rather than filtering. First, Table 3 reveals that there are many descriptors which may be intended to monitor similar potential liabilities, eg hydrogen bond acceptors and count of N and O atoms. Second, some descriptors are very sensitive to calculation method, for example, *c*log*P*. These ambiguities can be problematic in assigning good cutoffs for filters. Third, sorting a series or scaffold by various descriptors and properties ensures that a potential series with systematic problems with important calculated properties is identified and those series with only a few individual compound issues rise to the top. Also, the application of filters may eliminate an interesting singleton where later the particular violation may be overcome with a few new analogs.

4.2. Scoring functions

The scoring functions applied to cheminformatic descriptors can be simple, unweighted functions, such as Lipinski's rule-of-5.[78] Multiple physical descriptors are often combined and weighted using formulas to generate a scoring scheme.[79] These schemes can be slightly more complex such as the Central Nervous System Multiparameter Optimization (CNS MPO) score[80] which is a combination of unweighted linear functions, or significantly more complex formulas utilizing weighted spline functions such as those found in the OSIRIS Property Explorer used at Actelion Pharmaceuticals Ltd.[81] Individual companies, as well as individual programs, may develop their own proprietary scoring functions to aid in decision making. One such example is the bioavailability multiparameter score (see Fig. 3). This score utilizes multiple calculated physical descriptors to predict the bioavailability of compounds based on the aggregated physical descriptors of known oral drugs.[75]

While Lipinski's landmark work[78] is most frequently used to estimate drug likeness, continued work has indicated many other calculated quantities that can distinguish drug-like, lead-like and fragment-like compounds. Muchmore et al. have collected an analysis of many of the calculated properties and provide some guidelines to aid in the correct application of calculations.[62] Physical property estimates such as log*P* are subject to variation among software package and therefore their application in filtering and use in scoring functions requires adjustment to cutoffs. Calculations of very simple constitutive descriptors can be unambiguously performed and can have very useful applications. For example, the number of aromatic atoms is an unambiguous calculation and has been found to correlate with lower developability.[82] Additionally, the fraction of sp^3 carbons (Fsp^3) is also easily calculated[83] and is found to be a good descriptor of 3D structural complexity. Increased sp^3 composition was found to correlate entry of compounds into clinical testing with increased solubility of more saturated compounds being a possible explanation for the trend. An additional descriptor is the presence, and possibly even the count, of chiral carbon centers within a compound. The structure of drug targets are complex 3-dimensional shapes, which commonly have chiral centers. Computational scoring based on the physical descriptors describing the 3D structural properties of hit compounds can result in hit series with increased propensity to bind the intended target; however, these metrics usually parallel increases in synthetic difficulty.

4.3. Data-storage, retrieval, display and annotation

Once the descriptor data is generated, it must be stored in an organized and rapidly accessible fashion and in a manner where it can be easily merged with experimental data and decisions. The data storage component is now most often referred to as a data warehouse, wherein the underlying storage mechanisms may be

Table 3
Hit confirmation check-list populated with data for a typical hits and lead compounds[a]

Experimental methods	Hit profile	Lead profile
Chemical		
Confirm structural identity, purity; chemical stability, not fluorescent or otherwise interfering[111,112,10]	Y	
Solubility (mg/mL)	>50[77]	>10 mg/mL[77]
log *P* (O/W)		1–3[78,113]
log *D*	1.6[77]	<3.0[77]
Tractable synthetic group	Good	Y[111,112,10]
Potential chemical handles for structure modification	(literature synthesis, <5 steps[77])	
Favorable intellectual property position	Y[111,112,10,77]	Y
Number of active analogs	5 analogs (pIC$_{50}$ >6)[77]	Y[77]
Analogs tested with exploitable SAR	Y	Y[77]
Biochemical		
Affinity/potency (pK_i/pIC$_{50}$)	6.2[77]	~100 nM[114,77]
Ligand efficiency (kcal mol^{-1} per non-H atom)	>0.44[77,99,97]	
Selectivity over closely related targets	>5-fold[77]	y[77]
Good dose response curve, functional assay (agonist/antagonist)	Y[111,112,10]	
Mechanism of action: for example, competitive inhibitor		
ADMET		
Ca flux IC$_{50}$		<0.1 μM[77]
Preliminary PK		Y[111,112,10]
Not cytotoxic	Y[111,112,10]	Y[111,112,10]
Human liver microsomes (ml/min/mh protein)	30[77]	<23 μL/min/mg[77]
Rat hepatocytes (ml/min/10^6 cells)	70[77]	<14[77]
Rat microsomes (ml/min/mg protein)	60[77]	
Plasma stability, $t_{½}$ (min)	120[77]	
Rat iv pharmacokinetics		<35 mL/min/kg, V_{ss} >0.5 L/kg, $T_{1/2}$ >0.5 h[77]
Rat po bioavailability		>10%[77]
Oral bioavailability		F >10%, PPB <99.5%[77]
CYP inhibition (pK_i/pIC$_{50}$)	6.1 (3A4)[77]	
hERG pK_i (dofetilide binding)	5.1[77]	
Plasma protein binding		>99.5%[77]
Computational and cheminformatic calculations		
Calculated physical properties		
*c*log*P*, *S*log*P*		−4 to 4.2[115]<3.0[77]
Molar refractivity; CNS-MPO; apK_a, bpK_a		
Polar surface area (Å^2)		<80 CNS permeable[116,117]
Drug-like indexes, Andrews intrinsic binding energies		[118,119]
Structural properties/descriptors		
Molecular weight		<460[115]; <450[77]; <350[114]
Number of rotatable bonds; number H-bond donors; number H-bond acceptors		⩽10[115]; ⩽5[115]; ⩽9[115]
Number of N's and O's		⩽10 (druglike)[78]
nC, nN, nO, nS, nN/nHeavy; nO/nHeavy, nHdonors(solvated) nH acceptors (solvated)[120]		
Number of halogens[120]		<5[121]
n(O + N) n(OH + NH)		[121]
Number of each bond type:nC–C, nC–N, nC–O, nC–S, nC–X[120]		
Number of rigid bonds[122]	Y[111,112,10]	
Degree of unsaturation		
Methylene groups in unbranched chain[123]		<9
Size of macrocycle		<23 atoms[121]
Fused aromatic systems		<4[121]
ring fusion degree[120]		
Formal charge		−1 to 1[113]
Number of heavy atoms		(Fragment <20)
Number of rings		⩽4[121]
Number of chiral centers		[121,124]
Filters and undesirable groups		
Not promiscuous inhibitor 'frequent hitter'	Y[111,112,10]	Y[16]
No chemical reactive groups	Y[111,112,10]	Y[16,125]
No undesirable chemical groups		
No suspected toxic or mutagenic groups	Y	Y1[13,121,126]

[a] Y indicates that the profile must be met for this stage in the process (hit or lead-stage).

of varying database philosophies due to query requirements. These differences are due to the necessity of storing different data types such as raw image data and chemical structures. The chemical information aspect of databases has been reviewed by Miller.[84] The organization of 2D and 3D structural information and its retrieval requires a specialized set of database functions and storage schemes. Some earlier reviews[85] illustrate some considerations for organizing information so that properties can be stored and queried in very complex ways. Additionally, in this work, the Neurogen group added a social-network approach so that complex discussions can take place and more complex information recorded. Several other systems have also been published which link disparate data types for display, analysis and prioritization.[86] A well designed data storage scheme will facilitate concurrent viewing of experimental and cheminformatics data and allow for capture of decisions on individual compounds or entire compound series as the project progresses.

The challenge of maintaining the internal database is already daunting, but research is often simultaneously integrated with collaborator or competitor data outside of the organization. Maintaining data integration between a project at the hit identifications or hit-to-lead stage and outside literature information can be very helpful, even when looking at a fairly novel target with sparse data as information frequently becomes available over the relatively long lifespan of drug discovery projects. Often identical hits, or structurally very similar compounds, occur in the literature along with published SARs of related proteins and other series. The current PubChem[87] and Chembl[88] databases are two public systems organizing more than 20 million structures and associated biological data. Smaller and more heavily curated chemical and biological databases are also freely downloadable such as DrugBank[89] and Binding DB.[90] The data mining of hit compounds in these external databases is valuable in assessing the likelihood that a hit is valid since similar compounds tend to have similar activities and similar proteins bind similar compounds.[91] Importantly, a singleton hit can be viewed as more valid if a similar compound binds a similar protein in the literature. A complete and robust analysis can help to prioritize compounds up or down by making make them less desirable due other considerations such as their 'frequent-hitter' status, intellectual property (IP) coverage etc.

Systems that allow user defined, complex queries and display formats are becoming more common. Scaffold clustering,[29,28] compound clustering,[92] scoring, and ranking help to mine the important values and trends contained in these large stores of data. Once the important information has been identified and extracted, several display schemes are used to deliver the information to the medicinal chemist.[93,94] However, a universal user interface is quite difficult to develop in an entirely satisfying approach, since different projects often have inconsistent nomenclature and varying breadth of data. For example, the comparison of selectivity data is often made from several families of proteins, in a particular subgroup of G-protein coupled receptors or kinases; however, the various experimental data values may be obtained from many different combinations of internal collaborators and vendors. Project teams make decisions based on both these combined experimental data and many different computational predictions. These decisions need to be captured as data in the storage scheme and are an integral part of the project so as to avoid loss of time when looking for latent series. Additionally, many decision points may be fuzzy and a de-prioritized series may need to be revisited or provide useful information for other series of compounds. Recording failures can also help future project teams avoid repeated mistakes around similar compounds that may be identified in new biology programs.

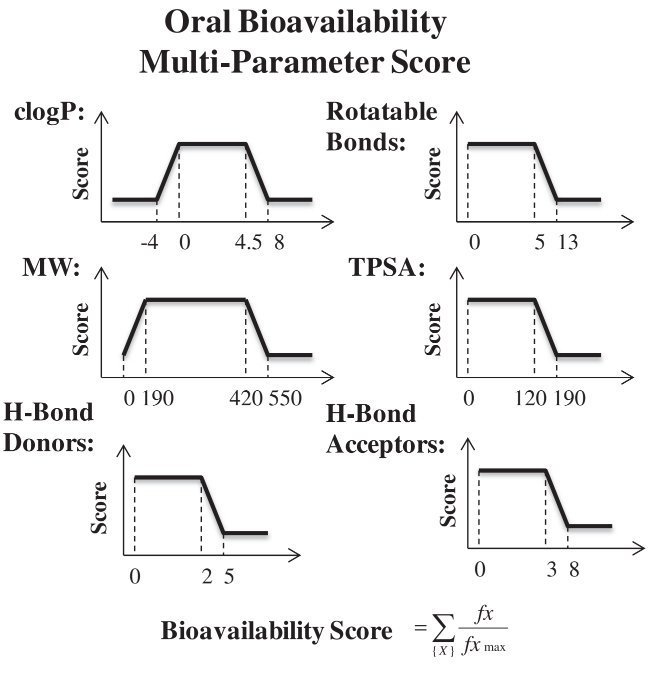


Figure 3. Oral bioavailability multiparameter score. Plots showing the scoring of individual parameters used to calculate the total oral bioavailability multiparameter score.

4.4. Practical considerations of scoring and filtering

There is a continual debate with regard to a filtering paradigm, versus a paradigm of hit scoring and prioritization based on the available cheminformatic information. Hit prioritization is likely a better method to choose lead compounds than simple filtering. Many good starting points can be derived from hits that have only a few liabilities. Therefore, decision making in drug discovery should not be strictly binary gated, but rather a continuum of best to worst probabilities. The reasoning is that all generated data required some amount of resource investment and therefore has value. Removing compounds from consideration for being on the incorrect side of an arbitrarily determined cutoff value is counter-productive, unless error tolerances are known and the cutoff value was chosen for a good reason. Within the context of a single project, a good strategy is to assign a less-desirable score to a given compound, which decreases its priority in the rank order. Using this method, compounds are not removed from valuable trend analysis. All information is considered to be useful and can be retained, even in the event the hit becomes de-prioritized during the process. Limited resources need to be applied to the compounds of highest priority[95] and the preferred method is to use a preplanned scoring mechanism to apply an unbiased score to every hit. These collections of data associated with hit compounds can be stored with traffic-light analysis, radar plots and other visual strategies used to organize and evaluate data.[94,93] It is important to have a method of prioritizing hits very early, as early as during a high throughput screen where many good hits may be lost and extensive experimental effort wasted on undesirable compounds.

The most important properties of scoring functions are consistency and applicability to the biological target. In a study published in 2004, Lajiness et al.[96] evaluated compound selections made by a group of medicinal chemists. Their study showed that while some sets of structures were commonly selected for follow up by different chemists, a large variability was found in the selection of other sets. These results clearly showed a level of bias in the selection of compounds for follow up which likely depends on the medicinal chemists' particular areas of expertise and experience. While this bias may be viewed as knowledge gained through experience, the scoring mechanism should reflect this collective experience and not rely on individual choices. Sorting the hits by overall score based on unbiased parameters dictates the priority of resources which should be committed to each hit or cluster of hits. It is not uncommon for preliminary SAR results to emerge from the clustered data during sorting and prioritization. These early SAR trends are a favorable factor in determining which lead series to carry forward.

4.5. Ligand efficiency

Mathematical functions combining experimental and cheminformatic data are some of the most useful calculations for scoring and prioritization. Ligand efficiency (LE)[97] and lipophilic efficiency (LLE)[98] are valuable tools in evaluating the compounds and the progress of the hit-to-lead program. For practical reasons, the negative logarithm of the K_i (pK_i) or IC_{50} (pIC_{50}) is used in place of the free energy of binding (ΔG). Equation 1 expresses the efficiency in terms the number of heavy atoms and Equation 2 expresses the efficiency in terms of calculated $\log P$ ($c\log P$) and is termed lipophilic efficiency.

$$LE = \frac{\Delta G}{N_a} \tag{1}$$

$$LLE = \Delta G - c\log P \tag{2}$$

Ligand efficiency is then a convenient way to weight the effect of additions to a molecule to simultaneously optimize the size and activity of compounds. While ligand efficiency has been exploited in compound design, it can also be useful in prioritizing hits and possible lead series. In principle, various formulations of ligand efficiency indices (LEI) can be calculated in terms of any easily calculated property.[99,100] Abad-Zapatero and Metz illustrate formulas to use MW and polar surface area as efficiency measures. Additionally, they suggest ways to use single concentration experiments in HTS to calculate LEI to sort compounds at the hit stage and thereby moving the smallest, most ligand efficient hits to the top of the list.

The concept of LE has roots in several empirical relationships between binding energies and the number and types of atoms and functional groups. In the case of the Andrews binding energy,[101] optimal contributions for a collection of functional groups were calibrated. In principle, this calculated quantity would be the maximum binding energy for a compound. It is expected that compounds will not reach these maximum values which implies that there is an inefficient use of functional groups. Kuntz et al.[102] proposed a maximal contribution per atom and that these contributions exhibit an approximately linear relationship with the number of heavy atoms for molecules with less than approximately 15 atoms. However, as the size of molecules increases beyond 15 atoms, additional atoms do not provide equal contributions on a per-atom basis. Reynolds et al.[103] investigated ligand binding efficiency using more recent collections of data and some prototypical calculations in order to identify some reasons for the loss of ligand efficiency as molecules grow in size. Among the important reasons for larger molecules to be less ligand efficient are the inability of larger molecules to find a complementary fit for every ligand atom in a binding site, a decrease in accessible surface area/atom and to a lesser extent increased entropic penalties in larger molecules (e.g., increased numbers of rotatable bonds). During the lead optimization process, new compounds tend towards higher MW, complexity, rings, number of heavy atoms and rotatable bonds in order to increase potency. Statistically, these additional atoms and molecular features will be less optimal and less ligand efficient. Therefore compounds and series that survive the hit selection and hit-to-lead process should be the smallest structures possible so that there is room to add various functionalities to the series in order to solve other non-target issues.

Historically, the most polar compounds of similar ligand efficiency values were given preference, because they were better development candidates likely due to their better water solubility and overall bioavailability.[104] In early hit triaging techniques, smaller and lower $\log P$ compounds should be sorted to the top of the evaluation because very highly ligand-efficient compounds

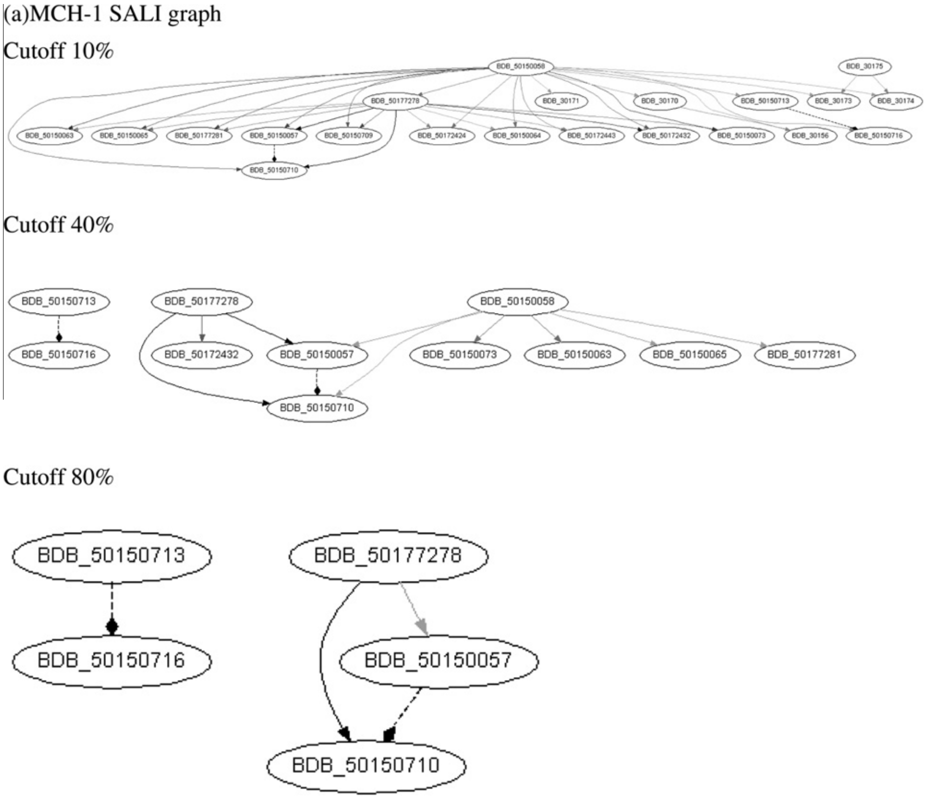


Figure 4. Examples of SAR visualization tools. Compound IC_{50}'s were obtained from Binding DB for MCH-1 antagonists. (a) SALI graph[108] Numbers are Binding DB identifiers. An edge occurs between two nodes having a SALI value for that pair greater than the cutoff. The direction of an edge indicates bioactivity from low to high for the corresponding pair of nodes. (b) SAR analysis from MOE software.[127] Darker shading indicates better activity. Three correlation tables: Scaffold vs R_1, Scaffold vs R_2 and R_1 vs R_2. (c) A traditional SAR table with traffic light analysis: Activity: <100 nM (green), 100–1000 nM (yellow), >1000 nM (red); MW: <450 (green), 450–600 (yellow), >600 (red); $c\log P$: −4 to 4.2 (green), 4.2–6.0 (yellow), >6.0 (red).

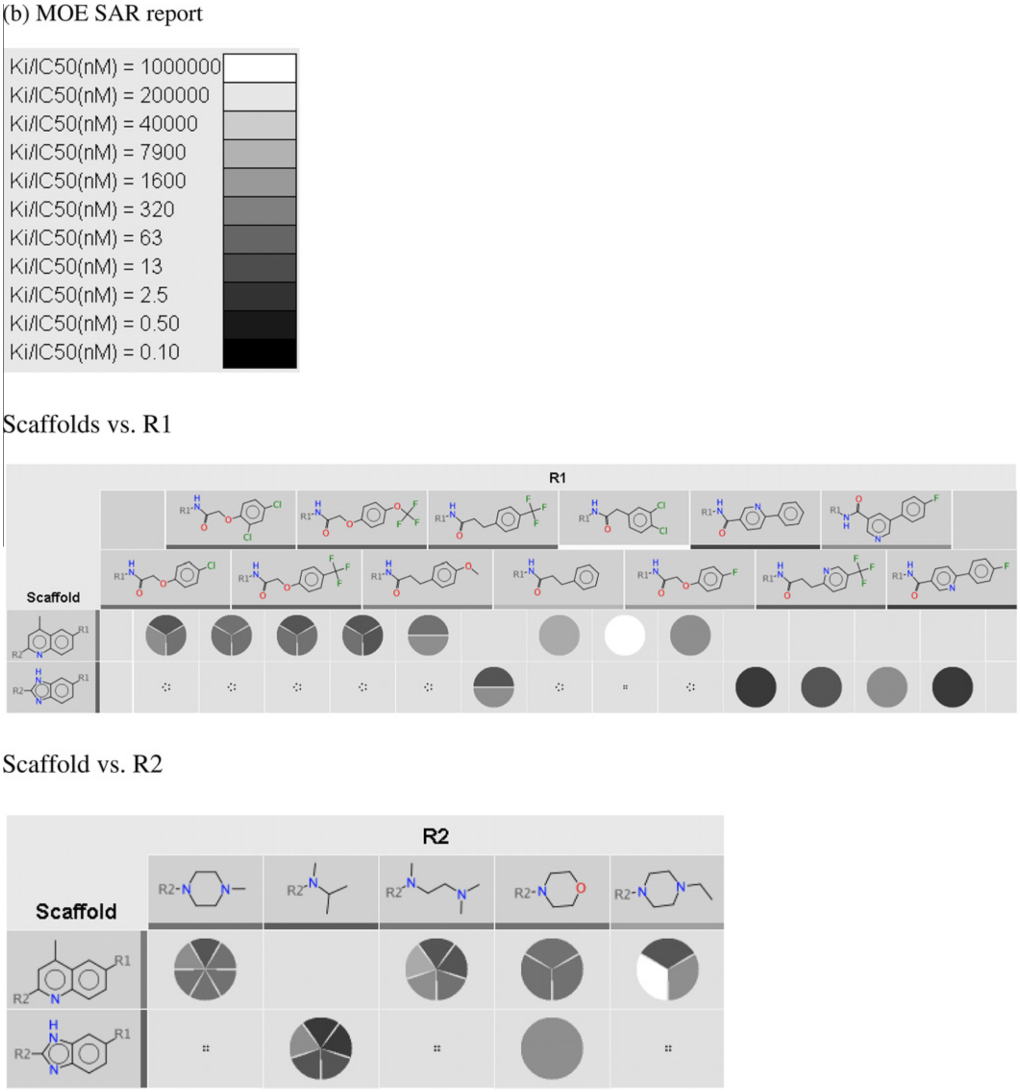


Fig. 4 (*continued*)

may appear as weak hits in primary screening results. The value of these compounds as starting points for lead series may far exceed larger compounds that may be only marginally more potent. There may be interesting cases where a weak biological response is obtained for fairly large molecules (low ligand efficiency). Before eliminating these compounds from further consideration, it may be appropriate to identify the source of suboptimal biological responses. For example, unnecessary rotatable bonds, halogen atoms or protecting groups may be present. A few synthetic or purchased analogs that remove functionalities while maintaining potency (i.e., increasing ligand efficiency) may provide support for a series or the usefulness of a hit.

5. Developing hits into leads: Synthetic considerations and SAR analysis tools

As a result of HTS, virtual screening, high-throughput ADMET, descriptor calculations, and clustering of experimental, cheminformatic and structural data, there is a tremendous amount of biological, physical, and cheminformatic information available to the drug discovery teams. The availability of physical compounds is an important and often overlooked aspect to drug discovery. The commercial availability of starting materials and the methods for preparing the core scaffold structure are important factors in the attractiveness of the hit series. Other important considerations are given to having synthetic targets that can be built by a divergent synthesis at late stages. Long linear sequences to install new functional groups for SAR testing are very costly in terms of time, synthetic efficiency, and material efficiency. Not all synthetic problems are readily predictable. Some unexpected, costly hurdles to efficient synthetic SAR throughput are non-divergent syntheses, failure or low yield of key synthetic steps in planned routes, reactive/unstable intermediates, and difficult purifications, starting material shortages, safety considerations that limit scale or throughput, and structural analysis. Conducting actual wet-chemistry synthetic trials with small groups of chemists is a very good strategy to predict synthetic accessibility.

Information from newly synthesized compounds is merged with the information from that provided by the original hit series to provide a working landscape of the chemical space. The avoidance of scaffolds showing extremely flat SARs[105] and identification of structure activity cliffs[106] is important during early SAR evaluation. There are many ways to organize and display SARs and many advances have been made to facilitate the complex interrelationships between compounds and series. We will discuss several different methods of representing an SAR where the methods range

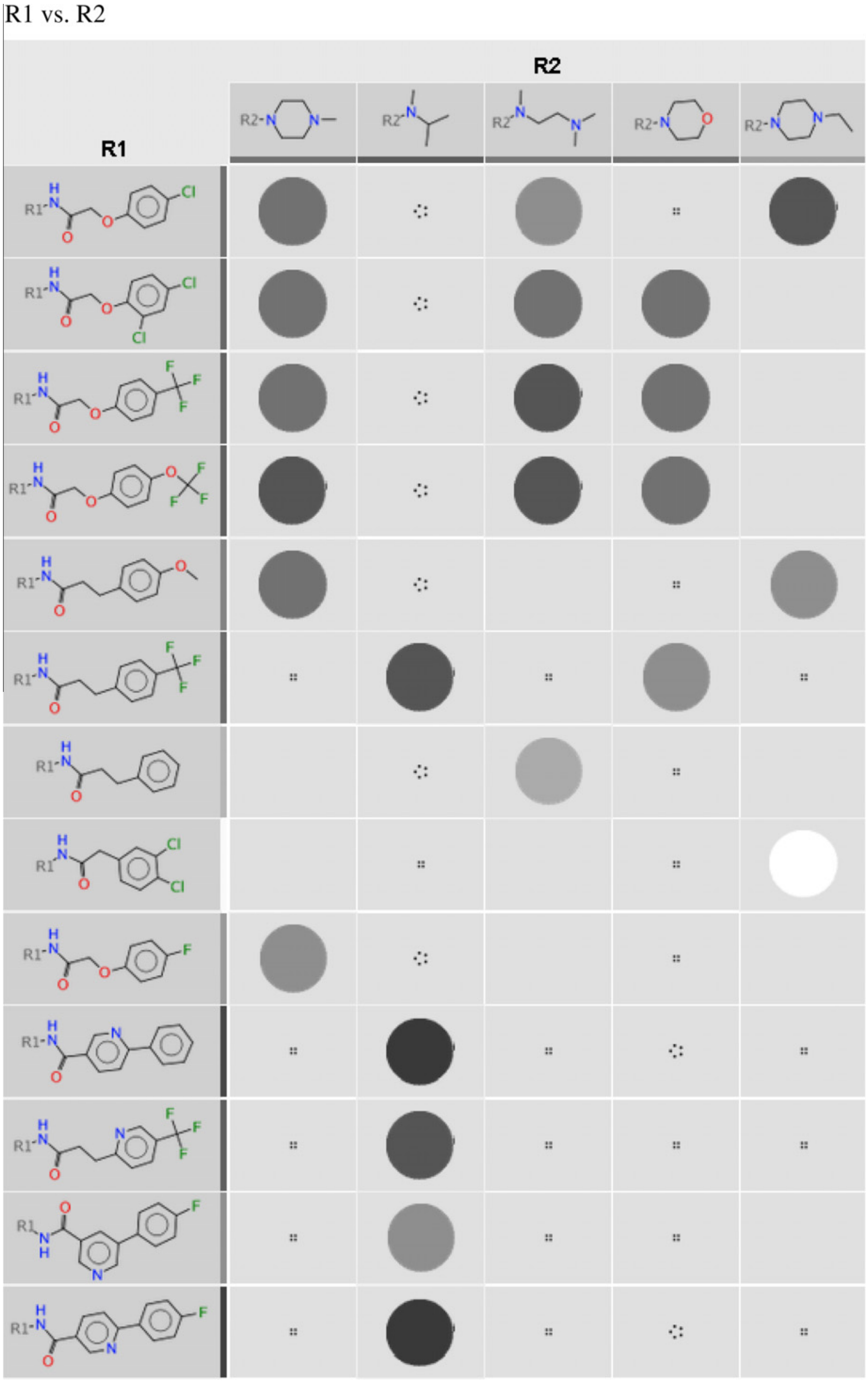


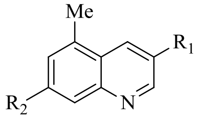
Fig. 4 (*continued*)

from fairly complex maps with connections among many different compounds to very simple and traditional SAR tables.

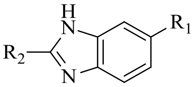
Recently, a new area of investigation is to quantitate and then map SAR landscapes of compounds.[107] By weighting chemical similarity by biological responses connections between different clusters of compounds can be calculated with SAR indices. These numeric values allow the characterization of data sets containing multiple scaffolds and clusters as continuous, discontinuous, heterogeneous-relaxed, and heterogeneous-constrained. In particular, the continuous SAR reflects multiple clusters of compounds which have diverse structures but are similar in activity. Using this method of analysis during hit-series selection, the continuous classification of screening results would indicate that many distinct classes of compounds are available for lead optimization. However, if a dataset reveals a discontinuous SAR (large biological differences due to small structural changes) then the landscape is likely more favorable for lead optimization of a single scaffold.

Figure 4 illustrates several other approaches that allow users to analyze SAR in graphical format. The structure–activity landscape index (SALI) is another tool used to which seeks to display the landscape of a medicinal chemistry data set.[108] We have collected several sample outputs from the SALI software in Figure 4a. The SALI plot of an SAR obtained from Binding DB indicates where relatively small changes in structure result in large changes in activity

(c) SAR Tables



1



2

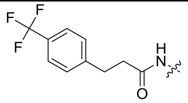
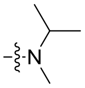
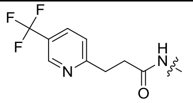
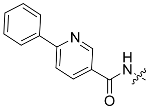
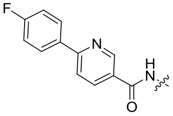
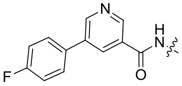
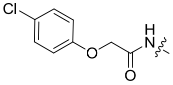
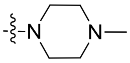
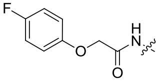
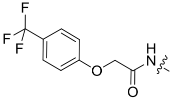
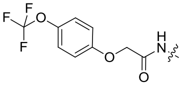
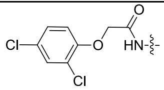
Scaffold	R1	R2	BDB ID	IC_{50} (nM)	MW	cLogP
2			BDB_30156	🟩 21	🟩 404	🟨 4.8
2			BDB_30171	🟩 41	🟩 405	🟩 3.8
2			BDB_30173	🟩 8.8	🟩 385	🟨 4.3
2			BDB_30174	🟩 5.5	🟩 403	🟨 4.5
2			BDB_30175	🟩 900	🟩 403	🟩 4.0
1			BDB_50150057	🟩 71	🟩 424	🟨 4.3
1			BDB_50150058	🟥 1682	🟩 408	🟩 3.9
1			BDB_50150063	🟩 63	🟩 458	🟨 4.7
1			BDB_50150073	🟩 36	🟩 474	🟨 5.3
1			BDB_50150709	🟩 89	🟩 458	🟨 4.7

Fig. 4 (*continued*)

(a so called activity cliff). The cutoff reflects all connections of SALI scores below a percentage of the range of SALI values. At the 10% cutoff level, the broadest view of the SAR landscape is provided. As the cutoff is increased only the steepest cliffs are revealed. Early in hit-to-lead, identifying a steep improvement in activity is desirable; conversely, as the SAR exploration proceeds on already potent compounds it is optimal to identify positions for chemical modifications in a series where activity remains high despite a variety of structural changes. The next approach to displaying SARs is the SAR table shown in Figure 4b and it shows activity with respect to multiple scaffolds and R-groups.[109] Another related technique is SAR Maps.[110] These types of diagrams and organizing data in tables can help to identify key pharmacophores and common features between series. Early in the evaluation of a hit series, these maps can point out missing analogs as well as simple trends and patterns. Lastly, Figure 4c illustrates an example of traffic light

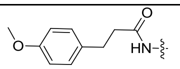
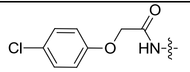
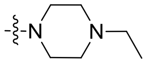
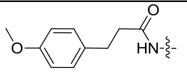
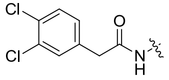
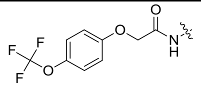
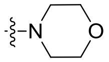
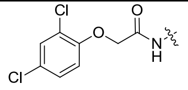
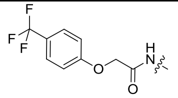
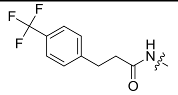
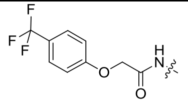
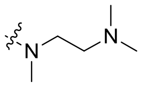
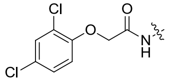
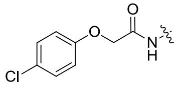
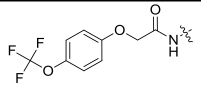
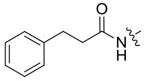
1			BDB_50150716	250	418	4.4
1			BDB_50150710	30	438	4.6
1			BDB_50150713	730	432	4.7
1			BDB_50172417	727000	456	5.6
1			BDB_50172424	82	461	5.1
1			BDB_50172443	59	445	4.7
1			BDB_50150064	124	445	4.5
2			BDB_30170	428	418	3.8
1			BDB_50150065	22	460	4.9
1			BDB_50172432	66	460	5.1
1			BDB_50177278	807	426	4.6
1			BDB_50177281	31	476	5.5
1			BDB_50177303	10000	390	4.8

Fig. 4 (*continued*)

analysis on a set of compounds which is often used to ensure that all important criteria are met for a compound or series at key decision points. A series may be prioritized fairly quickly if the exploratory SAR shows a tendency towards 'green' and conversely, such a simple display may also indicate that changes in the molecule do not help in potency and/or other properties.

6. Other factors in hit-to-lead and series prioritization

Another large factor during the hit development process is intellectual property. While individual compounds in hit containing libraries may be clear of patent encumbrances, an estimate must be made of how broad the available IP space needs to be in order to conduct meaningful SAR. The availability of patent space can vary greatly among biological targets. While the defining reason for drug discovery is to relieve human pain, suffering and death the research and development endeavor must also be commercially viable in order to justify the large financial investment. Therefore one of the primary goals in lead generation and lead optimization is to develop novel intellectual property that can be adequately protected by new patents.

During the hit-to-lead process, decisions are made from a wealth of data to select the best compounds for ongoing investment. Some problems with compounds or series are nearly impossible to overcome during the later stages of drug development. These ‘red flags’ may be problems with poor solubility, high lipophilicity, low ligand efficiency, toxicity, reactivity, large lipophilic fragments that are required for good SAR activity, overly flat SAR trends, and promiscuous binding in single or multiple biological classes are the most common development roadblocks. These problems are historically difficult to correct and compound clusters bearing these traits should have already been de-prioritized by prudent use of scoring. However, when these problems arise late in a hit-to-lead program, selection of alternate scaffolds for lead series development is the best recourse.

7. Summary

Proceeding from a large number of hits to a few good starting points for hit-to-lead is an inherently risky process with many opportunities for poor choices. Published examples readily demonstrate the importance of careful library, assay, and cheminformatic planning before the HTS even begins. At each stage of drug discovery, the cost of experimentation increases; therefore, the use of cheminformatics, computational, and database systems are a significantly aid in organizing possible lead series. Clusters of similar compounds help to validate a series without expending extensive wet chemistry resources on weak hits. While virtual screening and cheminformatics do not replace traditional biological and chemical experimentation, these in silico techniques aid in the allocation of resources and decision making process. The rapid evaluation and organization of early screening data into many visualization formats is an important aid in the identification of subtle, and sometimes hidden, trends. The combined use of cheminformatic techniques such as physical descriptor calculation, filtering, scaffold clustering, similarity searching, virtual screening, SAR landscape evaluation, and hit prioritization is effective at reducing the time, cost, and resources required for lead series generation.

Acknowledgments

The authors would like to thank Bruce Molino and Robb Lewis for their input and suggestions and the referees for their review and valuable comments.

References and notes

1. Weitz, J. *Nat. Rev. Drug Disc.* http://www.nature.com/nrd/posters/warfarin/warfarin_poster.pdf (accessed January 3, 2012).
2. Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. *Nat. Rev. Drug Disc.* **2011**, *10*, 188.
3. Vogt, M.; Bajorathh, J. *Bioorg. Med. Chem.*, in press. http://dx.doi.org/10.1016/j.bmc.2012.03.030.
4. Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. *Nat. Rev. Drug Disc.* **2003**, *2*, 369.
5. Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. *J. Med. Chem.* **2008**, *51*, 3661.
6. Newman, D. J.; Cragg, G. M.; Snader, K. M. *J. Nat. Prod.* **2003**, *66*, 1022.
7. Keserü, G.; Makara, G. M. *Drug Discovery Today* **2006**, *11*, 741.
8. Smith, A. *Nature* **2002**, *418*, 453.
9. Macarron, R. *Drug Discovery Today* **2006**, *11*, 277.
10. Gribbon, P.; Sewing, A. *Drug Discovery Today* **2005**, *10*, 17.
11. Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. *J. Biomol. Screen.* **1999**, *4*, 67.
12. Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. *Nat. Biotechnol.* **2006**, *24*, 167.
13. Baell, J. B.; Holloway, G. A. *J. Med. Chem.* **2010**, *53*, 2719.
14. Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. *J. Med. Chem.* **2002**, *137–142*, 45.
15. Goode, D. R.; Totten, R. K.; Heeres, J. T.; Hergenrother, P. J. *J. Med. Chem.* **2008**, *51*, 2436.
16. McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. *J. Med. Chem.* **2002**, *45*, 1712.
17. Grant, J. A.; Haigh, J. A.; Pickup, B. T.; Nicholls, A.; Sayle, R. A. *J. Chem. Inf. Model.* **2006**, *46*, 1912.
18. Simeonov, A.; Jadhav, A.; Thomas, C. J.; Wang, Y.; Huang, R.; Southall, N. T.; Shinn, P.; Smith, J.; Austin, C. P.; Auld, D. S.; Inglese, J. *J. Med. Chem.* **2008**, *51*, 2863.
19. Hochlowski, J.; Cheng, X.; Sauer, D.; Djuric, S. J. *Comb. Chem.* **2003**, *5*, 345.
20. Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. *J. Chem. Inf. Model.* **2006**, *46*, 1060.
21. Di, L.; Kerns, E. H. *Drug Discovery Today* **2006**, *11*, 446.
22. Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. *J. Med. Chem.* **2003**, *46*, 1250.
23. Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. *J. Chem. Inf. Model. Sci.* **2001**, *41*, 1308.
24. Clark, A. M. *J. Chem. Inf. Model.* **2010**, *50*, 37.
25. Ertl, P.; Schuffenhauer, A. J. *Cheminformatics* **2009**, 1.
26. McFayden, I.; Walker, G.; Alvarez, J. Enhancing Hit Quality and Diversity within Assay Throughput Constraints. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, 2004; pp 143–173.
27. Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
28. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. *J. Chem. Inf. Model.* **2007**, 47.
29. Wilkens, S. J.; Janes, J.; Su, A. I. *J. Med. Chem.* **2005**, *48*, 3182.
30. Nilakantan, R.; Immermann, F.; Haraki, K. *Comb. Chem. High Throughput Screening* **2002**, *5*, 105.
31. Posner, B. A.; Xi, H.; Mills, J. E. J. *J. Chem. Inf. Model.* **2009**, *49*, 2202.
32. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *J. Med. Chem.* **2002**, *45*, 4350.
33. Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*, 1st ed.; John Willey & Sons: New York, 1990.
34. Makara, G. M. *J. Med. Chem.* **2007**, *50*, 3214.
35. Hesterkamp, T.; Whittaker, M. *Curr. Opin. Chem. Biol.* **2008**, *12*, 260.
36. Hajduk, P. J. *J. Med. Chem.* **2006**, *49*, 6972.
37. Wermuth, C. G. *Drug Discovery Today* **2006**, *11*, 348.
38. Mayr, L. M.; Bojanic, D. *Curr. Opin. Pharmacol.* **2009**, *9*, 580.
39. Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. *ChemMedChem* **2008**, *3*, 435.
40. Decornez, H. Y.; Duffy, B. C.; Kitchen, D. B. SP2 2011, Sep. 23–24.
41. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H. O. *J. Med. Chem.* **2005**, *48*, 2534.
42. Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. *J. Comput. Aided Mol. Des.* **2008**, *22*, 179.
43. Knox, A. J. S.; Meegan, M. J.; Carta, G.; Lloyd, D. G. *J. Chem. Inf. Model.* **2005**, *45*, 1908.
44. Wallach, I.; Lilien, R. J. *Chem. Inf. Model* **2011**, *51*, 196.
45. Mestres, J.; Veeneman, G. H. *J. Med. Chem.* **2003**, *46*, 3441.
46. Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J.-H.; Raman, P.; Ertl, P.; Schuffenhauer, A. *J. Chem. Inf. Model.* **2010**, *50*, 2067.
47. Varin, T.; Didiot, M.-C.; Parker, C. N.; Schuffenhauer, A. *J. Med. Chem.* **2012**, *55*, 1161.
48. Habig, M.; Blechschmidt, A.; Dressler, S.; Hess, B.; Patel, V.; Billich, A.; Ostermeier, C.; Beer, D.; Klumpp, M. *J. Biomol. Screen.* **2009**, *14*, 679.
49. Shoichet, B. K. *Drug Discovery Today* **2006**, *11*, 607.
50. Shoichet, B. K. *J. Med. Chem.* **2006**, *49*, 7274.
51. Von Ahsen, O.; Schmidt, A.; Klotz, M.; Parczyk, K. *J. Biomol. Screen.* **2006**, *11*, 606.
52. Kieffer, B.; Homans, S.; Jahnke, W. Nuclear Magnetic Resonance of Ligand Binding to Proteins. In *Biophysical Approaches Determining Ligand Binding to Biomolecular Targets: Detection, Measurement and Modelling*; Podjarny, A., Dejaegere, A., Kieffer, B., Eds., 1st ed.; Royal Society of Chemistry, 2011; p 40.
53. Cooper, M. A. *J. Mol. Recognit.* **2004**, *17*, 286.
54. Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. *Nat. Rev. Drug Disc.* **2011**, *10*, 307.
55. Olah, J.; Alsenoy, C. V.; Sannigrahi, A. B. *J. Phys. Chem.* **2002**, *106*, 3885.
56. McInnes, C. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494.
57. Bajorath, J. *Nat. Rev. Drug Disc.* **2002**, *1*, 882.
58. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Disc.* **2004**, *3*, 935.
59. Brown, N. *WIREs Comput. Mol. Sci.* **2011**, *1*, 716.
60. Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. J. *Chem. Inf. Comput. Sci.* **2005**, *45*, 1784.
61. Ertl, P.; Rohde, B.; Seizer, P. *J. Med. Chem.* **2000**, *43*, 3714.
62. Muchmore, S. W.; Edmunds, J. J.; Stewart, K. D.; Hajduk, P. J. *J. Med. Chem.* **2010**, *53*, 4830.
63. Sanghvi, T.; Jain, N.; Yang, G.; Yalkowsky, S. H. *QSAR Comb. Sci.* **2003**, *22*, 258.
64. Jorgensen, W. J.; Duffy, E. M. *Adv. Drug Discov. Rev.* **2002**, *54*, 355.
65. Ran, Y.; Jain, N.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208.
66. Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773.
67. Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. *J. Chem. Inf. Model.* **2008**, *48*, 220.
68. Trunzer, M.; Faller, B.; Zimmerlin, A. *J. Med. Chem.* **2009**, *52*, 329.
69. Goodwin, J. T.; Clark, D. E. *JPET* **2005**, *315*, 477.
70. Fan, Y.; Unwalla, R.; Denny, R. A.; Di, L.; Kerns, E. H.; Diller, D. J.; Humblet, C. *J. Chem. Inf. Model.* **2010**, *50*, 1123.
71. Norinder, U.; Österberg, T.; Artursson, P. *Pharm. Res.* **1997**, *14*, 1786.
72. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R. *J. Pharm. Sci.* **2007**, *97*, 1946.

73. Dolghih, E.; Bryant, C.; Renslo, A. R.; Jacobson, M. P. *PLoS Comput. Biol.* **2011**, *7*, 1.
74. Rayan, A.; Marcus, D.; Goldblum, A. *J. Chem. Inf. Model.* **2010**, *50*, 437.
75. Biswas, D.; Roy, S.; Sen, S. J. *Chem. Inf. Sci. Model.* **2006**, *46*, 1394.
76. Böcker, A.; Bonneau, P. R.; Hucke, O.; Jakalian, A.; Edwards, P. J. *ChemMedChem* **2010**, *5*, 2102.
77. Zoran Rankovic, R. M. *Lead Generation Approaches in Drug Discovery*, 1st ed.; Wiley, 2010.
78. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.
79. Leeson, P. D.; St-Gallay, S. A. *Nat. Rev. Drug Disc.* **2011**, *10*, 749.
80. Wagner, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. *ACS Chem. Neurosci.* **2010**, *1*, 435.
81. Sander, T.; Freyss, J.; von Korff, M.; Reich, J. R.; Rufener, C. *J. Chem. Inf. Model.* **2009**, *49*, 232.
82. Ritchie, T. J.; Macdonald, S. J. F. *Drug Discovery Today* **2009**, *14*, 1011.
83. Lovering, F.; Bikker, J.; Humblet, C. *J. Med. Chem.* **2009**, *52*, 6752.
84. Miller, M. A. *Nat. Rev. Drug Disc.* **2002**, *1*, 220.
85. Manly, C.; Chandrasekhar, J.; Ochterski, J.; Hammer, J.; Warfield, B. *Drug Discovery Today* **2008**, *13*, 99.
86. Agrafiotis, D. K.; Alex, S.; Dai, H.; Derkinderen, A.; Farnum, M.; Gates, P.; Izrailev, S.; Jaeger, E. P.; Konstant, P.; Leung, A.; Lobanov, V. S.; Marichal, P.; Martin, D.; Rassokhin, D. N.; Shemanarev, M.; Skalkin, A.; Stong, J.; Tabruyn, T.; Vermeiren, M.; Wan, J.; Xu, X. Y.; Yao, X. *J. Chem. Inf. Model.* **2007**, *47*, 1999.
87. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. *Nucleic Acids Res.* **2012**, *40*, D400.
88. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012**, *40*, D11.
89. Wishart, D. *DrugBank—Open Data & Drug Target Database*. http://www.drugbank.ca/.
90. Gilson, M. K.; Liu, T.; Nicola, G.; Hwang, L. *Binding DB*. http://www.bindingdb.org.
91. Banville, D. L. *Curr. Opin. Drug Discov. Devel.* **2009**, *12*, 376.
92. Haranczyk, M.; Holliday, J. *J. Chem. Inf. Model.* **2008**, *48*, 498.
93. Ritchie, T. J.; Ertl, P.; Lewis, R. *Drug Discovery Today* **2011**, *16*, 65.
94. Lobell, M.; Hendrix, M.; Hinzen, B.; Keldenich, J.; Meier, H.; Schmeck, C.; Schohe-Loop, R.; Wunberg, T.; Hillisch, A. *ChemMedChem* **2006**, *1*, 1229.
95. Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Schieber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. *J. Chem. Inf. Model.* **2007**, *47*, 1319.
96. Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. *J. Med. Chem.* **2004**, *47*, 4891.
97. Hopkins, A. L.; Groom, C. R.; Alex, A. *Drug Discovery Today* **2004**, *9*, 430.
98. Ryckmans, T.; Edwards, M. P.; Horne, V. A.; Correia, A. M.; Owen, D. R.; Thompson, L. R.; Tran, I.; Tutt, M. F.; Young, T. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4406.
99. Abad-Zapatero, C.; Metz, J. T. *Drug Discovery Today* **2005**, *10*, 464.
100. Abad-Zapatero, C. *Expert Opin. Drug Discov.* **2007**, *2*, 469.
101. Andrews, P. R.; Craik, D. J.; Martin, J. L. *J. Med. Chem.* **1984**, *27*, 1648.
102. Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997.
103. Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. *J. Med. Chem.* **2008**, *51*, 2432.
104. Peltason, L.; Hu, Y.; Bajorath, J. *ChemMedChem* **2009**, *4*, 1864.
105. Peltason, L.; Bajorath, J. *Methods Mol. Biol.* **2011**, *672*, 119.
106. Peltason, L.; Bajorath, J. *J. Med. Chem.* **2007**, *50*, 5571.
107. Guha, R.; Van Drie, J. H. *J. Chem. Inf. Model.* **2008**, *48*, 646.
108. Clark, A. M.; Labute, P. *J. Med. Chem.* **2009**, *52*, 469.
109. Kolpak, J.; Connolly, P. J.; Lobanov, V. S.; Agrafiotis, D. K. *J. Chem. Inf. Model.* **2009**, *49*, 2221.
110. Morwick, T.; Berry, A.; Brickwood, J.; Cardozo, M.; Catron, K.; DeTuri, M.; Emeigh, J.; Homon, C.; Hrapchak, M.; Jacober, S.; Jakes, S.; Kaplita, P.; Kelly, T. A.; Ksiazek, J.; Liuzzi, M.; Magolda, R.; Mao, C.; Marshall, D.; McNeil, D.; A.P., III; Sarko, C.; Scouten, E.; Sledziona, C.; Sun, S.; Watrous, J.; Wu, J. P.; Cywin, C. L. *J. Med. Chem.* **2006**, *49*, 2898.
111. Steinmeyer, A. *ChemMedChem* **2006**, *1*, 31.
112. Rishton, G. M. *Drug Discovery Today* **2003**, *8*, 86.
113. Wunberg, T.; Hendrix, M.; Hillisch, A.; Lobell, M.; Meier, H.; Schmeck, C.; Wild, H.; Hinzen, B. *Drug Discovery Today* **2006**, *11*, 175.
114. Barker, J.; Hesterkamp, T.; Whittaker, M. *Drug Discov. World* **2008**, 69.
115. Clark, D. E. *J. Pharm. Sci.* **1999**, *88*, 807.
116. Clark, D. E. *J. Pharm. Sci.* **1999**, *88*, 815.
117. Xu, J.; Stevenson, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177.
118. Leach, A. R.; Bradshaw, J.; Green, D. V.; Hann, M. M.; Delany, J. J., III *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161.
119. Vistoli, G.; Pedretti, A.; Testa, B. *Drug Discovery Today* **2008**, *13*, 285.
120. Olah, M. M.; Bologa, C. G.; Oprea, T. I. *Curr. Drug Discov. Technol.* **2004**, *1*, 211.
121. Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. *J. Comput. Aided Mol. Des.* **2007**, *21*, 113.
122. Oprea, T. I. *Molecules* **2002**, *7*, 51.
123. Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem.* **2005**, *48*, 312.
124. Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjögren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; Saal, W. v. d.; Zimmermann, G.; Schneider, G. *J. Med. Chem.* **2002**, *45*, 137.
125. Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897.
126. Chemical Computing Group, Inc. Molecular Operating Environment (MOE), 2011. v2011.10.
127. Mark, Alan. E.; W.F.v.G. *J. Mol. Biol.* **1994**, *240*, 167.