HIT AND LEAD GENERATION: BEYOND HIGH-THROUGHPUT SCREENING

Konrad H. Bleicher, Hans-Joachim Böhm, Klaus Müller and Alexander I. Alanine

The identification of small-molecule modulators of protein function, and the process of transforming these into high-content lead series, are key activities in modern drug discovery. The decisions taken during this process have far-reaching consequences for success later in lead optimization and even more crucially in clinical development. Recently, there has been an increased focus on these activities due to escalating downstream costs resulting from high clinical failure rates. In addition, the vast emerging opportunities from efforts in functional genomics and proteomics demands a departure from the linear process of identification, evaluation and refinement activities towards a more integrated parallel process. This calls for flexible, fast and cost-effective strategies to meet the demands of producing high-content lead series with improved prospects for clinical success.

A GUIDE TO DRUG DISCOVERY

DRUGABILITY
The feasibility of a target to be effectively modulated by a small molecule ligand that has appropriate bio-physico-chemical and absorption, distribution, metabolism and excretion properties to be developed into a drug candidate with appropriate properties for the desired therapeutic use.

F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, CH-4070, Basel, Switzerland. Correspondence to A. A. e-mail: alexander.alanine @roche.com
doi:10.1038/nrd1086

All drugs that are presently on the market are estimated to target less than 500 biomolecules, ranging from nucleic acids to enzymes, G-protein-coupled receptors (GPCRs) and ion channels[1] (FIG. 1). Although the target portfolio of large pharmaceutical companies is continuously changing, all of the main target classes are likely to be represented. The relative distribution of classes varies from company to company on the basis of the disease area they focus on, and also because some target families are more numerous than others. Presently, GPCRs are the predominant target family addressed, and more than 600 genes encoding GPCRs have been identified from human genome sequencing efforts[2].

The balance of such targets and their relative novelty are the domain of the companies' early project-portfolio management strategy[3]. Of course, a balance always has to be struck between the requirements of the disease area for efficacious new therapies, business considerations and, most crucially, the chemical tractability or DRUGABILITY of targets for small-molecule intervention[4]. It is well accepted within the medicinal chemistry community that, independently of the technology applied, certain protein families are more readily modulated by small-molecule intervention than others. In this context, target selection plays a pivotal role in the final outcome of HIT and LEAD identification activities. A retrospective analysis of past discovery programmes reveals that much higher success rates have been demonstrated for aminergic GPCRs compared with large peptide receptors, for example. This is not surprising, as modulating protein–protein interactions — often involving large surface areas — by a small chemical entity is far more demanding than competing against an endogenous small-molecule ligand.

Apart from the intrinsic biochemical and kinetic challenges in identifying an appropriate modulator for a target, the range of meaningful assays and ligand-identification technologies can also significantly influence the chances of success. Considering a representative target portfolio, HIGH-THROUGHPUT SCREENING (HTS) is presently the most widely applicable technology delivering chemistry entry points for drug discovery programmes. However, it is well recognized that even when compounds are identified from HTS they are not always suitable for the initiation of further medicinal chemistry exploration (FIG. 2). The potential for success is nevertheless demonstrated by a variety of development candidates and marketed drugs that have resulted

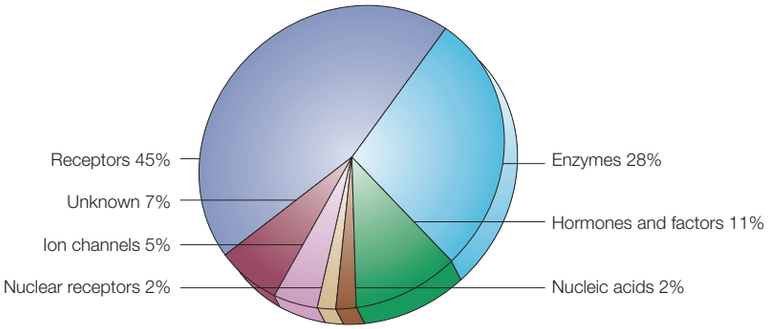


Figure 1 | **Therapeutic target classes.** All current therapeutic targets can be subdivided into seven main classes, wherein enzymes and receptors represent the largest part. Adapted with permission from REF. 1 © American Association for the Advancement of Science (2000).

from hits generated by HTS campaigns. It is evident that in the future the overwhelming number of emerging targets will dramatically increase the demands put on HTS and that this will call for new hit and lead generation strategies to curb costs and enhance efficiency[5].

HIT
A primary active compound(s), with non-promiscuous binding behaviour, exceeding a certain threshold value in a given assay(s). The 'active' is followed up with an identity and purity evaluation, an authentic sample is then obtained or re-synthesized and activity confirmed in a multi-point activity determination to establish the validity of the hit (validated hit).

Reducing attrition

The late-stage attrition of chemical entities in development and beyond is highly costly, and therefore such failures must be kept to a minimum by setting in place a rigorous, objective quality assessment at key points in the discovery process (FIG. 3). This assessment needs to begin as early as possible and must be of high stringency to prevent precious resources being squandered on less promising lead series and projects. The earliest point at which such knowledge-driven decisions can be made is in the lead-generation phase. Here, the initial actives, or 'hits', are progressed into lead series by a comprehensive assessment of chemical integrity, synthetic accessibility, functional behaviour, STRUCTURE–ACTIVITY-RELATIONSHIPS (SAR), as well as bio-PHYSICOCHEMICAL and absorption, distribution, metabolism and excretion (ADME) properties. This early awareness of the required profile (a given selectivity, solubility, permeation, metabolic stability and so on) is important for the selection and prioritization of series with the best development potential. In this regard, it is important that at least two lead series of significantly different pharmacological and/or structural profile are advanced as reserve, or 'back-up', lead series. This insures against unexpected failures due to unpredictable factors, such as toxicological findings in later animal studies. The effect of such a rigorous process at an early stage is to achieve greater awareness of key liabilities, which can be addressed in adequate time and with sufficient resources. The net effect is to reduce attrition in the costly clinical phases by intercepting many crucial ADME-related issues before they are discovered too late to be resolved.

Traditionally, hit identification is assumed to be the crucial bottleneck for lead generation success, but this is not the case. Rather, it is the overall characteristics of a compound class that make it an attractive starting point for medicinal chemists. Depending on the threshold set, an HTS campaign will always deliver active compounds, but it is the potential to optimize them into drug-like and information-rich lead series that is evidently far more important for the downstream success of the entities. This is clearly illustrated by the observation that despite the massive growth in screening compound numbers over the past 15–20 years, no corresponding increase in successfully launched new chemical entities has resulted.

Figure 2 | **Don't panic…** Turning an organic compound into a HIGH-CONTENT CHEMICAL LEAD SERIES is a challenging and sometimes extremely complex endeavour, as numerous hurdles beyond activity and selectivity have to be overcome. It is vital to identify high-quality actives, or 'hits', as the molecular starting point is crucial in determining the later potential for success. Hit discovery and lead generation is therefore far more than just the identification of active compounds; it is the multi-disciplinary process of selecting the most promising lead candidates from rigorously assessed molecular series.

Multi-property optimization

During the past few years, there has been an increasing awareness of the need for developing drug-like properties of a molecule. These are the balance of biophysicochemical requirements for the molecule to reach its site of action in man at the given concentration, for the necessary duration and with an adequate safety window in order to answer the therapeutic principle hypothesis[6].

In the past, lead-finding activities were mainly directed towards affinity and selectivity rather than molecular properties, metabolic liabilities and so on. It was not uncommon for a confirmed single primary active compound to be considered a 'lead' structure, or, in the case of a cluster of actives with SAR, a 'lead series'. Frequently, attention was not paid to characteristics of the molecules other than perhaps their chemical stability and synthetic accessibility. A consequence of these insufficient lead criteria — varying significantly not just between companies but also often within them — was that full project teams were assembled with only a single superficially evaluated 'lead'. A thorough consideration of other important drug features was often postponed until late in the optimization phase, when the *in vitro* affinity and selectivity had been fully optimized at the expense of other facets, such as solubility, permeability or metabolic stability.

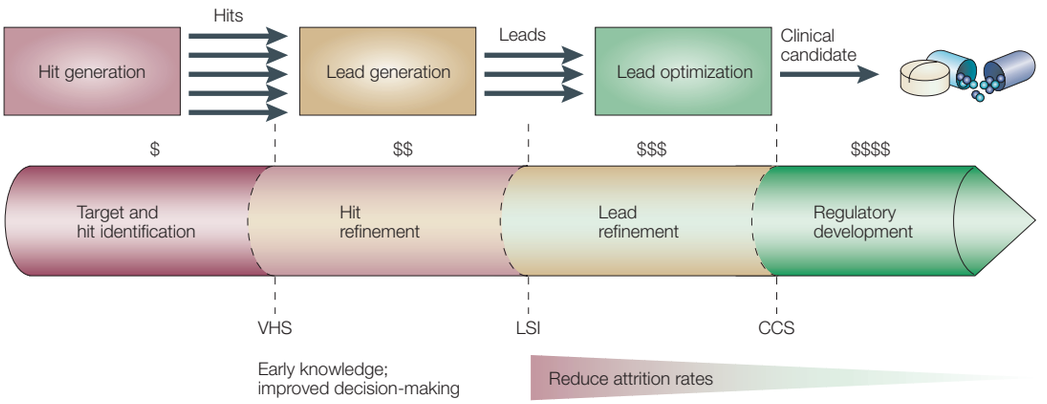


Figure 3 | **Stage-by-stage quality assessment to reduce costly late-stage attrition.** Typical important milestones are VALIDATED HIT SERIES (VHS), LEAD SERIES IDENTIFIED (LSI) and clinical candidate selection (CCS), which ensure that only drug candidates with an appropriately high-potential profile are advanced to the next phase.

LEAD
A prototypical chemical structure or series of structures that demonstrate activity and selectivity in a pharmacological or biochemically relevant screen. This forms the basis for a focused medicinal chemistry effort for lead optimization and development with the goal of identifying a clinical candidate. A distinct lead series has a unique core structure and the ability to be patented separately.

HIGH-THROUGHPUT SCREENING
Screening (of a compound collection) to identify hits in an *in vitro* assay, usually performed robotically in 384-well microtitre plates.

HIGH-CONTENT LEAD SERIES
A lead series in which representatives have been extensively refined in not only their structure–activity relationship and selectivity, but also in their physicochemical and early absorption, distribution, metabolism and excretion properties, and safety measures, such as metabolic stability, permeation and hERG liabilities. Correlations have been elucidated and all crucial parameters have shown themselves to be modulated in the series.

STRUCTURE–ACTIVITY RELATIONSHIP
The consistent correlation of structural features or groups with the biological activity of compounds in a given biological assay.

PHYSICOCHEMICAL PROPERTIES
Physical molecular properties of a compound. Typical properties are solubility, acidity, lipophilicity, polar surface area, shape, flexibility and so on.

VALIDATED HIT SERIES
A set of hits clustered into substructurally related families, representatives of which have been evaluated for their specificity, selectivity, physicochemical and *in vitro* ADME properties to characterize the series.

Unfortunately, as the lead molecule becomes increasingly more potent, selective and tailored for the target, there is generally less tolerance for introducing significant changes to affect biophysical properties without a large intrinsic affinity penalty. Such unbalanced, sub-optimal candidates entering clinical studies have attractive *in vitro* profiles but poor ADME attributes that often preclude them from progressing and being fully evaluated in the clinic due to, for example, dose-limiting solubility, poor absorption, CYTOCHROME P_{450} interactions or metabolic instability. Clearly, poor initial leads with weak entry criteria into lead optimization often can not be refined to generate compounds with an appropriate profile, resulting in high attrition rates at the clinical candidate selection stage. This point has been highlighted in a recent analysis of launched drugs, which indicates that, generally, relatively minor changes in structural and physical molecular properties take place between the lead and the launched drug candidate[7]. This emphasizes once more that the quality of the lead is crucial in most cases to the success of the refinement and development process. If the clinical entry criteria are lax, the attrition is moved further into pilot safety testing or early clinical-phase studies. The optimization process has historically been largely sequential in nature, addressing one issue at a time, with the hope that all necessary modifications could be accommodated within the PHARMACOPHORE optimized for affinity only. This approach led to a very high and expensive failure rate in the clinic for all major pharmaceutical companies. During the mid-90s, this view changed to embrace a more holistic attitude towards lead optimization and subsequently to hit-to-lead generation. The required trade-off for balancing these properties, in conjunction with pure affinity to achieve an equilibrated potential therapeutic drug molecule, resulted in a change of approach from sequential to MULTI-DIMENSIONAL OPTIMIZATION.

Hit and lead generation strategies

The entry point for any chemistry programme within drug discovery research is generally the identification of specifically acting low-molecular-weight modulators with an adequate activity in a suitable target assay. Such initial hits can be generated in a number of ways, depending on the level of information available[8]. It is therefore important to employ alternative hit-identification strategies that are able to tackle a variety of biological macromolecular targets effectively, and to identify proprietary, synthetically tractable and pharmacologically relevant compounds rapidly (FIG. 4).

These methods can be subdivided into those that require very detailed ligand and/or target information, and those that do not. The former include techniques such as mutagenesis, NUCLEAR MAGNETIC RESONANCE (NMR) and X-ray crystallography, as well as the recognition information that can be derived from endogenous ligands or non-natural small-molecule surrogates retrieved from literature and patents. At the other extreme are the technologies that do not require any prior information on target or ligand, and which use serendipity-based search strategies in either a given physical or virtual compound subset. Examples of so-called 'random' or pseudo-biased hit-identification strategies include biophysical and biochemical testing that employ one or other method of detecting a molecular-binding event, usually in a high-throughput format[9].

Between these extremes are more integrated approaches, including targeted libraries and chemogenomics[10]. The marriage of HTS with computational chemistry methods[11] has allowed a move away from purely random-based testing, towards more meaningful and directed iterative rapid-feedback searches of subsets and focused libraries. The prerequisite for success of both approaches is the availability of the highest-quality compounds possible for screening, either real or virtual.

Quality versus quantity

Besides the debate about how large a corporate compound collection should be, the questions of how to judge the quality of the inventory, and how to ultimately improve it, are important issues[12]. The collections of

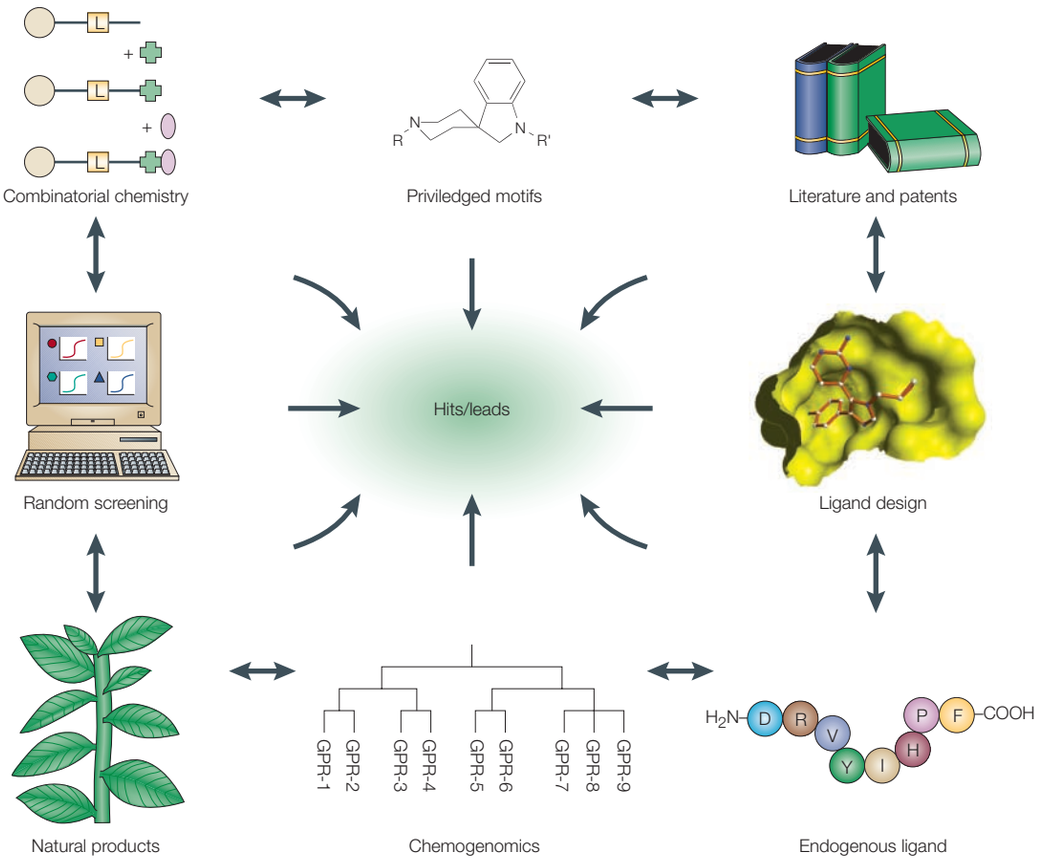


Figure 4 | **Hit-identification strategies.** The most commonly applied hit-identification strategies today range from knowledge-based approaches, which use literature- and patent-derived molecular entities, endogenous ligands or biostructural information, to the purely serendipity-based 'brute-force' methods such as combinatorial chemistry and high-throughput screening. The amalgamation of both extremes is anticipated to deliver more high-content chemical leads in a shorter period of time.

large pharmaceutical companies are approaching approximately one million entities, which represents historical collections (intermediates and precursors from earlier medicinal or agrochemical research programmes), natural products and COMBINATORIAL CHEMISTRY libraries. This is about an order of magnitude higher than ten years ago when HTS and combinatorial chemistry first emerged. Although this number is somewhat arbitrary, logistical hurdles and cost issues make this inventory size an upper limit for most companies. Many research organizations subsequently scaled back their large compound-production units after the realization that the quality component needed to get reliable and information-rich biological readouts cannot be obtained using such ultra high-throughput synthesis technologies favoured in the early 1990s. Today, instead of huge internal combinatorial chemistry programmes, purchasing efforts in every pharmaceutical company are directed towards constantly improving and diversifying the compound collections, and making them globally available for random HTS campaigns.

Although the chemical integrity of compounds can be checked by various analytical techniques, determining whether the chemical entities are useful in general as starting points for a hit-to-lead programme is far more complex. Besides the variable perceptions of medicinal chemists of what makes a valuable hit (lead-like versus drug-like), the issue concerning structural similarity, and in particular the overlap of chemical space (inventory versus vendor library), is frequently debated. Here, various computational algorithms are applied for the validation of compound collections to be purchased in terms of their DRUG-LIKENESS[13], chemical DIVERSITY[14] and similarity to the existing corporate compound inventory[15]. Although prediction tools for physicochemical properties, or 'FREQUENT-HITTER' LIABILITIES[16], and so on are successfully applied in a routine fashion, the issue concerning diversity is largely unresolved. The value of structure- or TOPOLOGY-oriented diversity DESCRIPTORS is not in question, although determining pharmacologically relevant similarity is far more complex and cannot be described accurately by any single metric, such as binding affinity. Conversely, it is difficult to describe the similarity (or dissimilarity) of two compounds that display the same activity, but possess, for example, different functionality, selectivity, toxicological liabilities and so on. Similarity is a context-dependent parameter and therefore the context must define the appropriate

LEAD SERIES IDENTIFIED
A peer-reviewed milestone, the requirements to be fulfilled are closely linked to the clinical candidate profile. Initial criteria are defined when hits are first identified; they include activity, selectivity and pertinent physicochemical properties, plus an evaluation of ADME and certain safety attributes. *In vivo* activity is not a mandatory requirement, provided the obstacles are appreciated and considered to be surmountable based on evidence.

CYTOCHROME P_{450}
A family of promiscuous iron-haem-containing enzymes involved in oxidative metabolism of a broad variety of xenobiotics and drug compounds.

PHARMACOPHORE
The spatial orientation of various functional groups or features necessary for activity at a biomolecular target.

MULTI-DIMENSIONAL OPTIMIZATION
The process of parallel optimization of several relevant drug-property parameters in concert with activity, to produce a drug candidate with balanced property profiles suitable for clinical development.

NUCLEAR MAGNETIC RESONANCE
A spectroscopy tool used for the assignment and confirmation of chemical structure of a compound or biological macromolecule. Sophisticated multi-dimensional methods are used to characterize larger and more complex biomolecules.

COMBINATORIAL CHEMISTRY
Synthesis technologies to generate compound libraries rather than single products. Robotic instruments for solid- and solution-phase chemistry, as well as high-throughput purification equipment, are applied.

DRUG-LIKENESS
A scoring metric (computational) for the similarity of a given structure to a representative reference set of marketed drugs.

metric, otherwise it is meaningless[17]. In any case, to increase the quality of a compound inventory, certain filtering techniques have to be applied for weeding out compounds that contain unfavourable chemical motifs. Database-searching tools have been developed that allow the differentiation of desired and undesired compounds. These computational algorithms are often based on sub-structural analysis methods, similarity-searching techniques or artificial neural networks[18]. Besides the application of those for filtering physically available compound collections or vendor databases, such algorithms can of course also be used to screen and validate virtual combinatorial libraries. It is in this setting that computational screening can have the greatest impact, owing to the overwhelmingly large number of compounds that are synthetically amenable using combinatorial chemistry technologies (BOX 1).

Focusing for libraries

The 'combinatorial explosion' — meaning the virtually infinite number of compounds that are synthetically tractable — has fascinated and challenged chemists ever since the inception of the concept. Independent of the library designs, the question of which compounds should be made from the huge pool of possibilities always emerges immediately, once the chemistry is established and the relevant building blocks are identified.

The original concept of 'synthesize and test', without considering the targets being screened, was frequently questioned by the medicinal chemistry community and is nowadays considered to be of much lower interest due to the unsatisfactory hit rates obtained so far. The days in which compounds were generated just for filling up the companies inventories, without taking any design or filtering criteria into account, have passed. In fact, most of the early combinatorial chemistry libraries have now been largely eliminated from the standard screening sets due to the disappointing results obtained after biological testing. The first generation of such combinatorial libraries were unattractive for most screening groups due to overloaded molecular complexity[19], poor drug-like features and low product purity. As a consequence, there is now a clear trend to move away from huge and diverse 'random' combinatorial libraries towards smaller and focused drug-like subsets. Although the discussion of how focused or biased a library should be is still an ongoing debate, the low hit rate of large, random combinatorial libraries, as well as the steady increase in demand for screening capacity, has set the stage for efforts towards small and focused compound collections instead.

Guided by the target. Biostructural information derived from mutagenesis data, as well as NMR or X-ray crystallographic analysis, has long been used for drug discovery purposes. Although the emphasis was initially focused more on single compound synthesis, a shift towards designing specific compound libraries is more commonplace today[20–22]. Recognizing that PARALLEL SYNTHESIS procedures cannot be applied to every structural motif makes the integration of biostructure-based design and combinatorial chemistry far more challenging, as synthesis protocols for compound arrays are often the limiting factor in the choice of useful designs. The close collaboration of computational scientists and chemists is therefore essential for formulating library proposals that fit with the target structure requirements and that are simultaneously amenable to parallel synthetic assembly. Finally, an understanding of the mechanism of action of a biological target, which is often available for many families of enzymes, is an important aid in biasing compound collections. These mechanism-based libraries have been applied successfully to a variety of proteins to generate transition-state mimics using either parallel solution- or solid-phase synthesis techniques[23].

Privileged structures or motifs. Another widely used approach concerning the generation of targeted compound collections is ligand motif-based library design. This is particularly relevant for targets for which very limited or no biostructural information is available. It is here that elements of known biologically active molecules are used as the core for generating libraries encompassing these 'PRIVILEGED STRUCTURES'[24]. Especially in the area of GPCRs, such design tactics have been applied successfully[25]. An inherent issue linked to this approach is the fact that these motifs can show promiscuous activity for whole target families, so selectivity considerations have to be addressed very early on. The restricted availability of privileged structures, and resulting issues concerning intellectual property, clearly limits the scope of this ligand-based approach to some extent. As a result, there is a continued need to identify novel proprietary chemotypes, and computational tools such as Skelgen[26] and TOPAS[27] have already shown their potential in this area.

'Cherry picking' from virtual space. A highly sophisticated way to avoid the synthesis of trivial analogues is the application of virtual screening tools in order to search through chemical space for topologically similar entities using known actives (seed structures) as references. In addition, biostructural information can also be applied if available[28]. Principally, one can subdivide such a virtual screening exercise into three main categories, namely virtual filtering, virtual profiling and virtual screening (BOX 2). The first focuses on criteria that are based on very fundamental issues concerning pharmacological targets in general. In this filtering step, all candidates are eliminated that do not fulfill certain generally defined requirements. These elimination criteria can either be based on statistically validated exclusion rules, substructural features or on training sets of known compounds. A retrospective analysis of drug molecules that demonstrated appropriate bioavailability formed the basis for the 'rule of five' guidelines, which make use of simple descriptors, such as molecular mass, calculated lipophilicity and hydrogen-bond donors/acceptors, in order to assess the probability of compounds being absorbed intestinally[29]. DEREK and TOPKAT are prediction tools for toxicological liabilities based on substructural analysis[30]. Artificial neural networks

DIVERSITY
A property–distance metric reflecting the dissimilarity of objects (molecules). Various molecular descriptors (indices) are used to define compounds in a numerical fashion so that they can be readily compared. Such measures must be considered within an appropriate context to be meaningful.

'FREQUENT-HITTER' LIABILITIES
An empirically derived metric by which compounds are assigned a probability to produce (false) positive results (hits) frequently in diverse screening assays.

MOLECULAR TOPOLOGY
A graph-based method of describing molecular structure using atom connectivity through the molecular framework and assigning atoms or substructural domains with various property types: lipophilic, H-bond acceptor/donor, positively/negatively charged and so on.

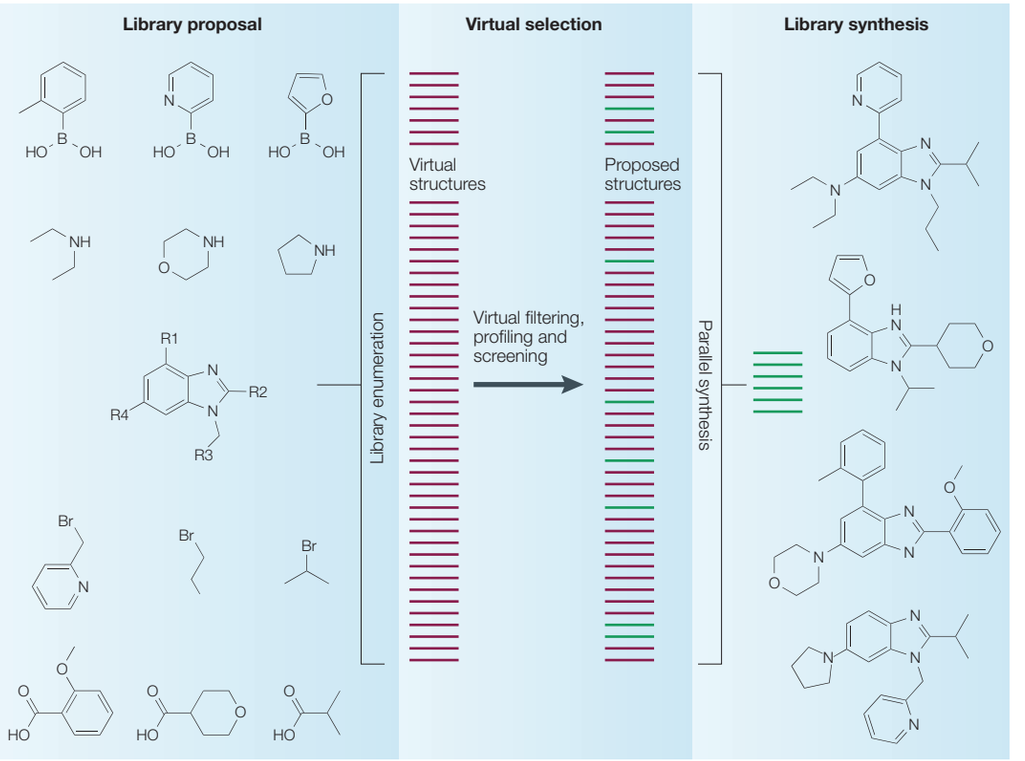
DESCRIPTORS
Metrics used to numerically describe a structure or certain molecular attributes of a compound (for example, Tanimoto, Ghose and Crippen, BCUT and so on).

PARALLEL SYNTHESIS
The process by which a set of individual compounds is made simultaneously using common chemical building blocks and homologous reagents.

PRIVILEGED STRUCTURE
A specific core or scaffolding structure that imparts a generic activity towards a protein family or limited set of its members independently of the specific substituents attached to it.

Box 1 | **The issue of chemical space**

The number of synthetically tractable compounds can be taken to be practically unlimited. The resulting chemical space is hard to comprehend, but the issues encountered are easily exemplified. Benzimidazoles, for example, are one of many interesting classes of molecules for which chemists can immediately devise various synthesis access routes. The shown example starts from the corresponding modified phenylenediamines by elaboration with carboxylic acids, alkyl- or (hetero)arylmethyl-halides, primary or secondary amines and boronic acids. Having only 100 entities of each building block available, a library of $100 \times 100 \times 100 \times 100 = 10^8$ benzimidazoles is conceivable. Even though only a fraction of those molecules are probably pharmacologically relevant, the huge number of possibilities indicates that the compound collection could span a large portion of chemical property space containing members with biological activities against many different pharmacological targets. Obviously the synthesis and testing of all combinations is neither feasible nor meaningful. Virtual screening technologies help to filter out the unfavourable combinations and predict actives out of such a library proposal if particular target and/or ligand information is available.



have been described that can discriminate between drug-like and non-drug-like compounds[31], molecules with high likeliness for cytochrome P450 interactions[32] or compounds that might show hERG liabilities[33] to further profile compounds in more depth.

Although the filtering and profiling steps can be applied to qualify particular compounds as being more or less drug-like, the virtual screening part instead encompasses specific project information to predict a certain binding propensity. Computational tools based on two-dimensional topological descriptors have proved to be very valuable for rapidly screening huge databases using known molecules as seed structures to generate activity-enriched libraries[34]. A big advantage in this context is the fact that besides the speed, a single molecule can be sufficient to identify compounds that are very different structurally, but which show similar biological activity, out of a particular virtual (or physically available) compound library. Searching at a higher level of sophistication can be achieved by using three-dimensional pharmacophores; however, this requires much more knowledge in terms of ligand information and conformation, as well as far greater computational power and time. Generating three-dimensional CONFORMERS of all library members clearly limits the size of the virtual library to a great extent, but having a two-dimensional searching step integrated for pre-selection helps to reduce the number of possible candidates. Virtual pharmacophore searches have been widely used during recent decades, and the impact within structure- or property-based drug discovery and lead design is becoming more prevalent[35].

The ultimate step of a virtual screening campaign is the introduction of the 'fourth dimension', namely, the target structure itself. The closest virtual approach to real bio-screening is virtual DOCKING AND SCORING, in which compounds are selected by defining interaction patterns of virtual compounds with the binding site of

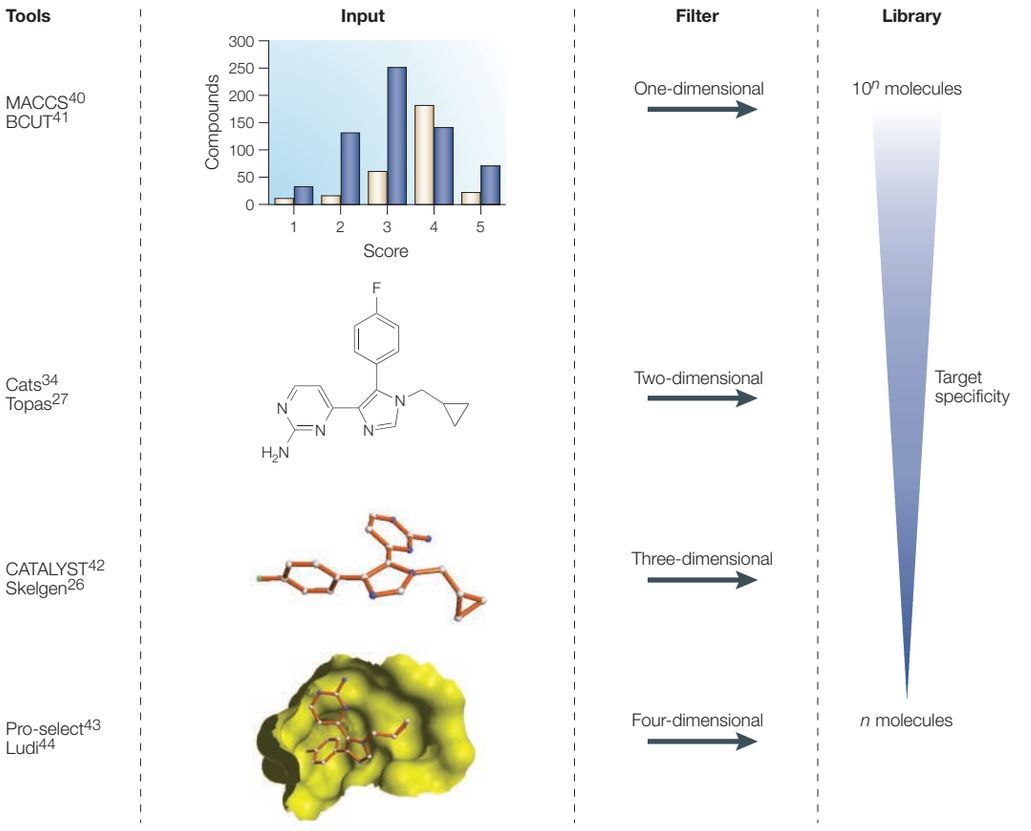
CONFORMERS
Distinct three-dimensional forms of a molecular structure of a given atomic connectivity, which results from internal rotations about single bonds between atoms.

the target protein. The requirement of crystallographic data, detailed knowledge of the binding mode and inherent issues concerning affinity-scoring functions still limits this approach to a great extent when considering the screening of large virtual libraries[36]. The iterative sequence of selection and refinement using one-dimensional descriptor, two-dimensional ligand and three-dimensional pharmacophore screening reduces the selected candidates to a manageable number. This leaves the highest-ranking molecules for further filtering by biostructure-based docking and scoring, and so provides both high activity enrichment and structural novelty.

As promising and valid as many of these computational algorithms are, they clearly can only be regarded as prediction tools. A continuous validation of proposed actives by rapid synthesis and testing is essential. So, the tight integration of database generation (library proposals), virtual screening, synthesis and multidimensional testing (affinity, selectivity, physicochemical properties and so on) is mandatory to ensure a successful process. Rather than synthesizing large compound libraries, only to find out later that hits do or do not materialize, rapid feedback loops are far more useful, as they allow the flexible adaptation of either the computational tools, the information used as input or the library proposals themselves. Adaptive cycles need to be established to guide the 'journey through chemical space' by computational scientists and the resulting data generated by chemists and biologists. Needless to say, it is imperative that logistical hurdles are overcome and that artificial boundaries between many disciplines are eliminated for maximizing the potential success of this approach.

Box 2 | **Virtual screening**

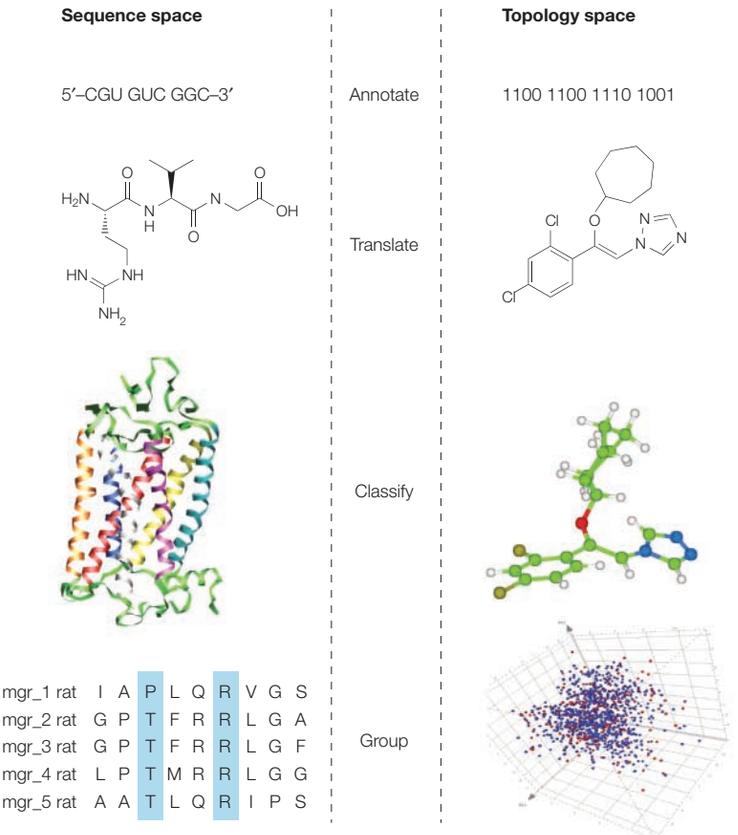
Individual steps in a virtual screening cascade can be subdivided into four principal components that distinguish the level of complexity delivered as input. First, general criteria are applied to eliminate all chemical structures that possess reliably predicted detrimental features, which would make them intrinsically less attractive as potential drugs. Molecular size, lipophilicity or potential metabolic liabilities, for example, could be used to reduce the number of possible candidates significantly. These filters are regarded as one-dimensional, as they are typically scalar and no detailed information on project-specific criteria is used. Topological searches from known ligands are often successfully applied in virtual screening when seeking compounds showing similar biological activities but with different structural characteristics (template hopping). Three-dimensional pharmacophore models are also applied for virtual screening when more detailed information concerning ligands and three-dimensional pharmacophore orientations is available. Even more structural knowledge is required when applying docking to a target structure and scoring the fittest chemical entities to be prioritized.



DOCKING AND SCORING
The process of computationally placing a virtual molecular structure into a binding site of a biological macromolecule (docking) and flexibly or rigidly relaxing the respective structures then ranking (scoring) the complementarity of fit.

Box 3 | **Similarity searching**

Similarity-searching algorithms applied in chemo- and bioinformatics serve to identify and annotate DNA or protein targets, as well as potential small-molecule modulators, at different levels of sophistication. On the basis of genomic information, proteins can be translated from their DNA sequence. Similarly, in chemistry, SMILES or BIT STRINGS are applied to annotate compounds and, more importantly, large compound databases (the chemical genome of a library). Those can be further classified, for example, by their three-dimensional pharmacophore representation, just as proteins can be classified by their function. Homology alignments of DNA or proteins by sequence similarity make the grouping of targets into target families possible. This is analogous to similarity-based virtual screening in which compounds are grouped on the basis of their annotation. Matching both topology and target space allows the identification of novel targets and ligands simultaneously.



SMILES
A character-based line notation for chemical structures.

BIT STRINGS
A contiguous set of characters that consists entirely of 1s and 0s, which can be used to encode, for example, the presence or not of structural elements in a compound.

Chemogenomics

There is no doubt that computational tools (both ligand- and biostructure-based tools) can be very useful to prioritize compounds that are more likely to be active at a particular target compared with others. The synthesis of these predicted actives ranked by any similarity metric results in a set of focused compounds that cover a certain portion of chemistry space. Owing to the target-related input that was used for biasing this set of compounds, a link to the corresponding proteins is established. The fuzziness inherently incorporated due to the imprecise nature of any prediction method is in fact of great benefit for a chemogenomics approach in which focused libraries of small drug-like molecules are used for the identification and validation of novel targets in a highly parallel fashion. So, computational technologies play a central role in chemogenomics not only for the generation of biased libraries, but also for the identification and clustering of biological targets.

The main question still to be answered is whether it is possible to overlay a certain chemistry or topology space (defined by the compound libraries) with a particular target space (defined by the sequence of the proteins). This depends on reliable biological and chemical annotation systems, as well as the possibility of linking them to each other. Experimental evidence for the chemogenomics concept has been delivered by medicinal chemistry for a long time. It is well known that more or less conservative changes in a molecular structure not only affects the activity, but also the selectivity, of a compound. The same phenomena are observed on the target side where protein mutations might lead to a complete loss of ligand activity or show no effect at all. In other words, the probability and extent of ligand binding can be tuned by the similarity of the chemical entities on the one hand, as well as by phylogenetic distance of the targets on the other. Understanding both, and being able to systematically annotate target- and ligand-space on a pharmacologically relevant basis, makes possible the identification of novel ligands and targets simultaneously (BOX 3). Receptor de-orphanization is no longer restricted to the identification of endogenous ligands, but can be achieved in a prospective manner by applying the similarity principle on both the ligand and the target side. Once a target protein is identified and correlated to a particular family cluster, the testing of focused compound libraries biased towards that subfamily should deliver hits for this particular target as well. So, the first step within a chemogenomics endeavour is the hunting for novel genes and proteins.

The constant growth in the amount of data emerging from genomics and proteomics studies clearly requires alternative methods to support the classical strategies for target assessment. Computational methods are making an increasing contribution, and bioinformatics plays a crucial role[37]. Successful applications of *in silico* target identification through bioinformatic approaches have been described recently in which sequence-similarity searching was performed using known DNA or protein sequences as seed information. To take a specific example, a recent publication describes the identification of four GPCRs from genome databases that were searched using various GPCR sequences as queries. Transcripts for all four genes were experimentally detected in the brain, which indicated that these receptors might be novel targets for central nervous system research[38]. Eventually, focused compound libraries using either the privileged structure approach, or more advanced virtual screening-based compound arrays, should allow us to identify small-molecule agonists, further validate the target and simultaneously move forward with drug-like compounds into the lead-generation and lead-optimization phases. The ongoing debate as to whether two-dimensional versus three-dimensional or ligand- versus target-based input is required for

Figure 5 | **Where there's a will, there's a way…** The discovery and development of new medicines is regarded as one of the most complex areas of research in both industry and academia. The expertise of many disciplines is essential to resolve the multifaceted challenges facing this discovery endeavour, ranging from pathway analysis to late-stage clinical development. In an increasingly complex and fast-paced research environment, the tight integration of complementary disciplines and technologies is becoming more essential than ever. This process will expand our current understanding of drug discovery, necessitating a move away from serial programmes towards increasingly multi-parametric parallel processing.

library focusing clearly shows that the computational community is still in the process of identifying the most valuable tools and strategies at each stage.

Conclusion

Hit and lead generation are key processes involved in the creation of successful new medicinal entities, and it is the quality of information content imparted through their exploration and refinement that largely determines their fate in the later stages of clinical development. It is in the early phases of drug discovery that changes in process, such as the early interception of key ADME parameters, can have the maximum impact on later-stage success and timelines. The present high attrition rates, especially after lead-optimization phases, indicate that drug discovery as a sequential alignment of independent disciplines is ineffective for delivering high-quality medicines of the future, and that issues beyond activity and selectivity must be addressed as early as possible in a flexible, parallel fashion. In our view, the combination of virtual screening and parallel medicinal chemistry, in conjunction with multi-dimensional compound-property optimization, will generate a much-improved basis for proper and timely decisions about which lead series to pursue further.

Applying this strategy shifts the bottleneck from hit identification to lead optimization. Therefore novel processes will have to be developed downstream for the full exploitation of high-quality and high-throughput technologies in chemistry, biology, molecular property analytics and ADME assessment in an integrated fashion. Continuous efforts are necessary to develop even more reliable, predictive virtual screening tools and knowledge-based algorithms for a better estimation of the key biophysical and even *in vivo* toxicological liabilities before synthesis is initiated. The combination of these tools into an integrated process (FIG. 5) will allow modern drug discovery to progress to a new level of sophistication. A quantitative improvement of the success rates is essential to cope with the ever-increasing expectations on pharmaceutical research, as well as the many new therapeutic targets expected to derive from intensified genomics and proteomics programmes.

Outlook ahead

In a manner similar to the tremendous development and maturation of oligonucleotide chemistry since the 1980s to the situation today in which DNA primers can be ordered by e-mail and are delivered the next day, we expect that parallel organic chemistry will progress analogously. This will allow the synthesis of focused compound libraries very rapidly on demand. The enormous chemical space that can already be covered by well-established chemical procedures (being much larger than any compound inventory will ever be), linked with ever better virtual screening and prediction tools, will give the chemist the opportunity to propose certain chemotypes to be 'squeezed' into the relevant pharmacology space by appropriate decoration. Increasingly, examples will follow where chemistry is dictated by the chemist and not by chemicals identified after random screening. Downstream optimization work will become increasingly effective not only because the chemistry is established, but also because a broader choice of templates and building blocks will be readily available. The application of HTS technologies will certainly move from pure random testing of huge compound pools to iterative RAPID FEEDBACK SCREENING of smaller, but more focused, compound ensembles. Therefore, the time spent on obtaining the relevant information, rather than the sheer capacity of synthesis and testing, will determine the success of a research programme. Stepping back from the rather 'safe' HTS paradigm to discovery in virtual space, which is still not yet fully developed, certainly needs courageous management decisions, not only in terms of financial investments, but also in organizational evolution. Breaking down artificial boundaries between different disciplines is a prerequisite for making full use of their potential. The large number of emerging targets, which are expected from functional genomics, demands novel and effective approaches for the hit and lead generation process as well as the lead optimization phase. Therefore chemistry will increasingly be applied upstream for both target identification as well as assessment[39] where the refined tools for investigating receptor pharmacologies will already encompass the properties of future potential drugs.

RAPID FEEDBACK SCREENING
Rapid feedback provided by assaying small compound sets (< 1,000) through a medium-throughput assay to guide the SAR for rapid iterative design and synthesis cycles.

1. Drews, J. Drug discovery: A historical perspective. *Science* **287**, 1960–1964 (2000).
2. Lander, E. S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Knowles, J. & Gromo, G. Target selection in drug discovery. *Nature Rev. Drug Discov.* **2**, 63–69 (2003).
4. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
5. Lenz, G. R., Nash, H. M. & Jindal, A. Chemical ligands, genomics and drug discovery. *Drug Discov. Today* **5**, 145–156 (2000).
6. Hodgson, J. ADMET — turning chemicals into drugs. *Nature Biotechnol.* **19**, 722–726 (2001).
7. Proudfoot, J. R. Drugs, leads, and drug-likeness: An analysis of some recently launched drugs. *Bioorg. Med. Chem. Lett*. **12**, 1647–1650 (2002).
8. Alanine, A., Nettekoven, M., Roberts, E. & Thomas, A. Lead generation — enhancing the success of drug discovery by investing into the hit to lead process. *Combin. Chem. High Throughput Screen.* **6**, 51–66 (2003).
9. Boguslavsky, J. Minimizing risk in 'Hits to Leads'. *Drug Discov. & Develop*. **4,** 26–30 (2001).
10. Bleicher, K. H. Chemogenomics: bridging a drug discovery gap. *Curr. Med. Chem.* **9**, 2077–2084 (2002).
11. Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **1**, 882–894 (2002).
This review article covers the current concepts of integrating both virtual and high-throughput screening.
12. Teague, J. S., Davis, A. M., Leeson, P. D. & Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* **38**, 3743–3748 (1999).
13. Walters, P. & Murcko, M. A. Prediction of 'drug-likeness' *Adv. Drug Deliv. Rev.* **54**, 255–271 (2002).
14. Martin, E. J. & Critchlow, R. E. Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem*. **1**, 32–45 (1999).
15. Menard, P. R., Mason, J. S., Morize I. & Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design and compound selection. *J. Chem. Inf. Comput. Sci.* **38**, 1204–1213 (1998).
16. Roche, O. *et al*. Development of a virtual screening method for identification of 'Frequent Hitters' in compound libraries. *J. Med. Chem*. **45**, 137–142 (2002).
17. Balkenhohl, F., von dem Busche-Hünnefeld, C., Lansky, A. & Zechel, C. Combinatorial synthesis of small organic molecules. *Angew. Chem. Int. Ed. Engl.* **35**, 2288–2337 (1996).
18. Böhm, H.-J. & Schneider, G. (eds). *Virtual Screening for Bioactive Molecules* (Wiley–VCH, Weinheim, 2000).
An excellent compendium of current virtual screening methods.
19. Hann, M. M., Leach, A. R. & Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864 (2001).
20. Crossley, R. From hits to leads, focusing the eyes of medicinal chemistry. *Modern Drug Discov.* **5**, 18–22 (2002).
21. Van Dogen, M., Weigelt, J., Uppenberg, J., Schultz, J. & Wikström, M. Structure-based screening and design in drug discovery. *Drug Discov. Today* **7**, 471–477 (2002).
22. Carr, R. & Jhoti, H. Structure-based screening of low affinity compounds. *Drug Discov. Today* **7**, 522–527 (2002).
23. Huang, L., Lee, A. & Ellman, J. A. Identification of potent and selective mechanism-based inhibitors of the cysteine protease cruzain using solid-phase parallel synthesis. *J. Med. Chem.* **45**, 676–684 (2002).
24. Patchett, A. A. & Nargund, R. P. Privileged structures — an update. *Annu. Rep. Med. Chem.* **35**, 289–298 (2000).
25. Bleicher, K. H., Wütherich, Y., Adam, G., Hoffmann, T. & Sleight, A. J. Parallel solution- and solid-phase synthesis of spiropyrrolo-pyrroles as novel NK-1 receptor ligands. *Bioorg. Med. Chem. Lett.* **12**, 3073–3076 (2002).
26. Stahl, M. *et al.* A validation study on the practical use of automated *de novo* design. *J. Comput.-Aided Mol. Des*. **16**, 459–478 (2002).
27. Schneider, G. *et al*. Virtual screening for bioactive molecules by *de novo* design. *Angew. Chem Int. Ed. Engl*. **39**, 4130–4133 (2000).
28. Schneider, G. & Böhm, H.-J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **7**, 64–70 (2002).
29. Lipinski, C., Lombardo, F., Dominy, B. & Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev*. **23**, 3–25 (1997).
A landmark publication based on retrospective data analysis for bioavailability resulting in the 'rule-of-five'.
30. Cariello, N. F. *et al*. Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. *Mutagenesis* **17**, 321–329 (2002).
31. Sadowski, J. & Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem*. **41**, 3325–3329 (1998).
32. Zuegge, J. *et al*. A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant. Struct.-Act. Relat.* **21**, 249–256 (2002).
33. Roche, O. *et al*. A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *Chembiochem* **3**, 455–459 (2002).
34. Schneider, G., Neidhart, W., Giller, T. & Schmid, S. 'Scaffold hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem Int. Ed. Engl*. **38**, 2894–2896 (1999).
35. Mason, J. S., Good, A. C. & Martin, E. J. 3-D Pharmacophores in drug discovery. *Curr. Pharm. Des.* **7**, 567–597 (2001).
36. Bissantz, C., Folkers, G. & Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **43**, 4759–4767 (2000).
37. Duckworth, D. M. & Sanseau, P. *In silico* identification of novel therapeutic targets. *Drug Discov. Today* **7**, 64–69 (2002).
38. Lee, D. K. *et al*. Identification of four human G-protein-coupled receptors expressed in the brain. *Mol. Brain Res*. **86**, 13–22 (2001).
This paper describes the successful identification of orphan G-protein-coupled receptors initiated by bioinformatic approaches.
39. Alaimo, P. J., Shogren-Knaak, M. A. & Shokat, K. M. Chemical genetic approaches for the elucidation of signaling pathways. *Curr. Opin. Chem. Biol.* **5**, 360–367 (2001).
40. McGregor, M. J. & Pallai, P. V. Clustering of large databases of compounds: using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comp. Sci*. **37**, 443–448 (1997).
41. Stanton, D. T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Com. Sci.* **39**, 11–20 (1999).
42. Sprague, P. W. Automated chemical hypothesis generation and database searching with CATALYST. *Perspect. Drug Discov. Design* **3**, 1–20 (1995).
43. Liebeschuetz, J. W. *et al.* PRO_SELECT: combining structure-based drug design and array-based chemistry for rapid lead discovery. 2. The development of a series of highly potent and selective Factor Xa inhibitors. *J. Med. Chem*. **45**, 1221–1232 (2002).
44. Boehm, H.-J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from *de novo* design or 3D database search programs. *J. Comput.-Aided Mol. Des*. **12**, 309–323 (1998).

Acknowledgement

Dr. Simona Ceccarelli is cordially thanked for providing the cartoons 'Don't panic….' and 'Where there's a will, there's a way…'.

Online links

FURTHER INFORMATION

Society for Biomolecular Screening:
http://www.sbsonline.com

Access to this interactive links box is free online.