A molecular informatics view on best practice in multi-parameter compound optimization

Scott J. Lusher[1], Ross McGuire[1], Rita Azevedo[1], Jan-Willem Boiten[1], Rene C. van Schaik[1] and Jacob de Vlieg[1,2]

[1] Department of Molecular Design & Informatics, Merck Research Laboratories, Oss, The Netherlands
[2] Computational Drug Discovery Group, Radboud University, Nijmegen, The Netherlands

The difference between biologically active molecules and drugs is that the latter balance an array of related and unrelated properties required for administration to patients. Inevitability, during optimization, some of these multiple factors will conflict. Although informatics has a crucial role in addressing the challenges of modern compound optimization, it is arguably still undervalued and underutilized. We present here some of the basic requirements of multi-parameter drug design, the crucial role of informatics and examples of favorable practice. The most crucial of these best practices are the need for informaticians to align their technologies and insights directly to discovery projects and for all scientists in drug discovery to become more proficient in the use of *in silico* methods.

SCOTT LUSHER

Scott Lusher moved to the department of Molecular Design & Informatics at Organon (later to become Schering-Plough and then Merck) in 2001. During the past 10 years, he has worked on multiple discovery projects, reaching the position of Research Fellow. In addition to his project responsibilities, Scott has a strategic role to improve the project application of all aspects of molecular informatics within discovery. Scott was born in the UK and educated at the University of Leeds and DeMontfort University before joining the Computational Biomolecular Discovery group at Unilever Research (NL) in 1997.

Introduction

In all industries with a high technical component, from architecture to car design, the use of computers to improve the quality of work and to increase the efficiency has been embraced. Computers are now crucial tools in all stages and aspects of drug design and development, from the identification of disease-related genes to the interpretation of clinical trial data, illustrated by the growing number of marketed drugs for which computational methods have had a significant contributing factor in their design. These include norfloxacin, losartan, zolmitriptan, dorzolamide, zanamivir and amprenavir [1]. Despite this, the pharmaceutical industry, with science and innovation at its core, has been slow to use information technology fully within many areas of the traditional drug design sciences. More specifically, the industry needs to challenge the preconception that the role of informatics is merely to support existing practices and, in future, must identify new ways of approaching problems that can only be addressed computationally.

As the drug discovery industry looks to decrease attrition rates in discovery, there is the recognition that a recurring historical failure of compound optimization has been the emphasis on individual properties [2,3], with potency the most likely to be chased early [4,5]. The life of a compound optimization project often begins with a rush to improve potency, with limited regard for the crucially important drug-like properties that separate biologically active compounds from commercial drugs. Only after sufficient potency is achieved, will attention switch to optimizing the numerous other criteria required [6]. If the discovery team is fortunate, this might require a limited

Corresponding author:. Lusher, S.J. (scott.lusher@merck.com)

1359-6446/06/$ - see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2011.05.005

amount of further modification, but frequently the project researchers will have painted themselves into an area of undesirable chemical space, requiring dramatic and time-consuming new modifications or the acceptance of an ultimately unsatisfactory compromise. The evidence is that drug designers appreciate these issues, but have been either unable, or at least slow, to heed these lessons and change the way in which they approach compound optimization, with the physical properties of compounds currently being synthesized in projects still differing significantly from current oral drugs [4]. During the 1990s, increasing access to drug-likeness assays earlier in the discovery process has reduced attrition rates for pharmacokinetics (PK) in recent years [7]. Although this suggests a move in the right direction, it might just be that the same compounds are being excluded earlier in the pipeline, which, although desirable, is not as valuable as systematically focusing design on more drug-like compounds to begin with. It should also be recognized that generating terabytes of data alone is not sufficient to ensure that projects move away from chasing potency and other restricting single parameter approaches. In fact, generating ever larger data sets without strategies, processes and tools for their analysis can so overwhelm teams that the number of data-driven decisions made decreases [8].

Here, we begin by introducing some of the key issues in multi-parameter drug design. We go on to present some of the molecular informatics practices required to achieve these goals.

Solving the Rubik's cube

The process of multi-parameter optimization is comparable to solving a Rubik's cube (http://www.Rubiks.com). Each face represents a required character (i.e. potency, stability or selectivity) and changing one face will affect another face, perhaps detrimentally. As with a Rubik's cube, addressing one parameter (solving one face of the cube) is relatively straightforward if one ignores each of the other parameters, but will not result in a completed puzzle. Sometimes, solving the puzzle requires the sacrifice of a completed face; that is, taking a local step back, to make a global step forward.

For the Rubik's cube, all faces are of equal value and weight, but is this the case for drug-design parameters? It can be argued that, for a specific compound, stability is less important than potency, but how does one separate parameters for which their total sum equals biological activity? Why decide at the beginning of a project to improve target engagement by increasing potency when it might be possible to improve PK and gain the same biological outcome?

All drugs are a result of compromise, but a more balanced approach to compound optimization would give all drug activity and property parameters equal weight and search for satisfactory compromise among them from the beginning rather than as a desperate late-stage concession. This approach also has the potential to deliver several compounds, with differing, but equivalent compromises among the relevant properties rather than a set of compounds dominated by one property. Delivering a set of non-dominated solutions with varying properties is also more attractive to project pipelines compared with a stream of 'me-too' compounds.

Discovery informatics providing a foundation

The ability to generate meaningful activity and property data in a timely fashion that is rapid enough to keep pace with design cycles, is the foundation required to support all multi-parameter drug design. Project teams need access to accurate and precise heterogeneous data, generated by multiple and potentially geographically dispersed disciplines. This requires strictly adhered-to protocols and authorization steps [9], and the need to disseminate data by the minimum number of user interfaces using globally consistent analysis processes.

Historically, teams would be responsible for gathering project-relevant data and storing them in whatever generic application was available, often Excel, which would then be shared among the team on an irregular basis. Mistakes in these composites accumulated over time, with every manual data intervention a potential source of introducing error. Collecting these data sets depended on the team being able to navigate a network of data sources, identifying all relevant information and ensuring that it was collected in a timely fashion. Analysis depended on whichever statistical method a particular researcher favored (which might be different from other members of the team).

As a basis for multi-parameter drug discovery, this situation is unsatisfactory, although it remains the status in many organizations [10]. However, it can be addressed by the design of tailored information management systems [11], the first descriptions of which include, 'ArQiologist' from ArQule [12], 'ADAAPT' from Amgen [13], 'OSIRIS' from Actelion [14] and the 'ABCD' system of Johnson & Johnson [10].

Organon Biosciences (now part of Merck & Co) has developed its own 'Integrated Project View' (IPV), providing drug designers with access to all pharmacological activity data generated for each project alongside data generated by its Drug Metabolism and Pharmacokinetics (DMPK) groups, analytical chemistry data and calculated data generated *in silico*. The IPV system is available at each team member's desktop and also in team meeting rooms, enabling data to be discussed and challenged by individual researchers or teams.

Analysis of raw data, without proper consideration of how they were generated, is dangerous, and drug-designers should be aware of general factors underlying data generation and be able to discuss issues underpinning interesting compounds and outliers. It is therefore crucial that those who generate data are in regular discussion with those who use the data to design new compounds. The use of IPVs at Organon 'ended' the sharing of data via Excel spreadsheets, as well as the previous hazard of data being generated and stored in hard copies in researchers' desk drawers. Information management tools enable scientists to browse recent results or specifically query the activity of thousands of compounds against multiple assays simultaneously. The IPV system links directly to industry standard data-mining tools, further facilitating multi-dimensional data analysis for even the novice.

There is evidence to suggest that the most productive source of lead compounds is from previous optimization projects [15]. Compound optimization is so challenging that if a series progresses along the pipeline, and is shown to be drug-like and safe, one should be obliged to identify every other application for which that series might be used. This can include follow-up compounds for the same target [16] or transfer to related targets. Systems to record all previously generated data, and tools to mine these resources, will facilitate this form of lead identification. The most important contributing factor to the success of these information systems is bench scientists, who see their value and actively generate and deposit data in routine fashion following agreed protocols. This can be an issue, especially as bench scientists are often not the end-user of the data that they generate and, therefore, do not always benefit directly from access to these systems.

Dealing with unexpected results and outliers in data

Data should drive drug discovery, even if those data are at odds with conventional wisdom, with teams allowing data to shape ideas rather than searching for data that support predefined hypotheses. One hindrance to multi-parameter drug design has been the focus on outliers during the design process. After making 100 non-selective compounds, it is tempting to promote the first selective compound to the top of the interest list and redesign all chemical plans around it. This strategy can work, but when the resulting compounds lack selectivity, researchers blame the multiple conflicting parameters, rather than consider that the base compound was simply an outlier or one-off (i.e. a unique set of properties combining to produce a unique outcome). Uniquely active substitutions restrict flexibility that the research team might need later and also risk stranding projects when the substituent does not survive the increasingly complex and diverse assays carried out later in lead optimization.

Despite this, outliers in data are often the most interesting results as they potentially indicate novel biological activity. However, caution should be exercised as they might just as easily indicate issues with the generation or management of the experimental data. Retesting and reanalysis of chemical composition might be time consuming, but is more efficient in the long term compared with basing design decisions on erroneous data. Teams should routinely evaluate their series on all relevant properties (not just potency) to identify outliers, as well as identifying compounds that have unusual property combinations, including New Chemical Entities (NCEs) that behave inconsistently across assays or for which the combination of two properties in one compound is unique.

It might not always be obvious which compounds are outliers, especially early during projects, when data for comparison are limited. Therefore, project teams should perform a periodic reevaluation of their total data pool, looking for compounds that buck trends or show unusual combinations of behaviors. These 'interesting' compounds are potentially valuable to projects and can be identified with very limited resources, provided adequate data management systems are in place.

Provided the activity of the outlier can be confirmed with some confidence, resources should be allocated to understanding its molecular basis. As with all aspects of chemical design, the deeper the understanding the easier an outcome is to replicate or improve. Determining the molecular basis for activity can be investigated in parallel to rapid exploration of the chemical space around outliers, but not replaced by it.

Unexplained activity is often a result of a lack of knowledge or understanding, demonstrating a need to improve the appreciation that design chemists have of protein–ligand interactions. This includes the balance between enthalpy and entropy, and the contribution of weak interactions, such as halogen bonds, including distances and geometries [17,18]. Owing to their marginal net value, weak interactions are difficult to design rationally into complexes, but might help to explain nuances in structure–activity relationship (SAR) data. These can then direct new syntheses, although they can also provide justification for overanalysis [18].

Related to this issue is the question of how many compounds must be generated to be confident that the specific area has been fully explored. The less that is understood about the factors contributing to the activity of a compound, the more NCEs need to be synthesized to feel confident that the project has been comprehensive. In the reality of time and resource-restricted projects, there is the frequent danger of cutting corners here and basing decisions on insufficient examples of compounds and data.

Is there a 'God's number' for compound optimization?

To return to the Rubik's cube analogy, a standard cube consists of 54 outer squares in six different colors, allowing for 43 252 003 274 489 856 000 different configurations (http://www.Rubiks.com). Even this number is small compared with the estimated 10^{60} drug-like molecules it might be possible to synthesize [19]. Estimating one second per turn, sampling all possible configurations of the Rubik's cube would require 1400 trillion years, but despite this, no possible configuration of the cube is ever more than 20 moves from being solved (the so-called 'God's number'; http://www.cube20.org). Although it might always be theoretically possible to solve any Rubik's cube formation in so few moves, in practice, it usually takes more.

For most drug-design projects, hundreds or even thousands of compounds must be generated to progress from an initial hit to development candidate, but in retrospect, how many individual chemical steps are the initial hits away from the final development candidate or drug? Anecdotally, the progress of internal projects from screening hit to development candidate requires hundreds of compounds to be synthesized, but less than 20 retrospective steps and typically less than 10 steps are required to achieve the goal. Earlier studies have identified how remarkably similar launched drugs are to their leads [15], which are themselves most often derived from earlier optimization efforts or from endogenous ligands (Box 1).

Others have investigated the ideal path for generating optimized drug leads, specifically investigating how ligand-efficiency changes can help determine whether projects should continue exploration at one position or whether to abandon the approach in favor of other directions [20]. In terms of guiding project teams through the challenges of drug design, a better understanding of successful and unsuccessful decision making in compound optimization might provide some guidance on how long to pursue one chemical subseries. Additionally, research to further quantify the collective actions of medicinal chemists, and specifically, preferred directions for compound optimization, will help improve *de novo* design tools and other *in silico* methods. For example, a recent paper demonstrated that the 3745 licensed single-entity drugs in the KEGG DRUG database are all based on just 236 conserved core structures and 506 peripheral fragments, whose 125 chemical attachment patterns depend on the core [21].

Although there remains a great deal more knowledge to extract from these analyses of earlier projects, one of the most important lessons seems to be to ensure that teams do not focus excessive resources exploring one area of chemical space around their leads at the expense of other potentially valuable directions. Project

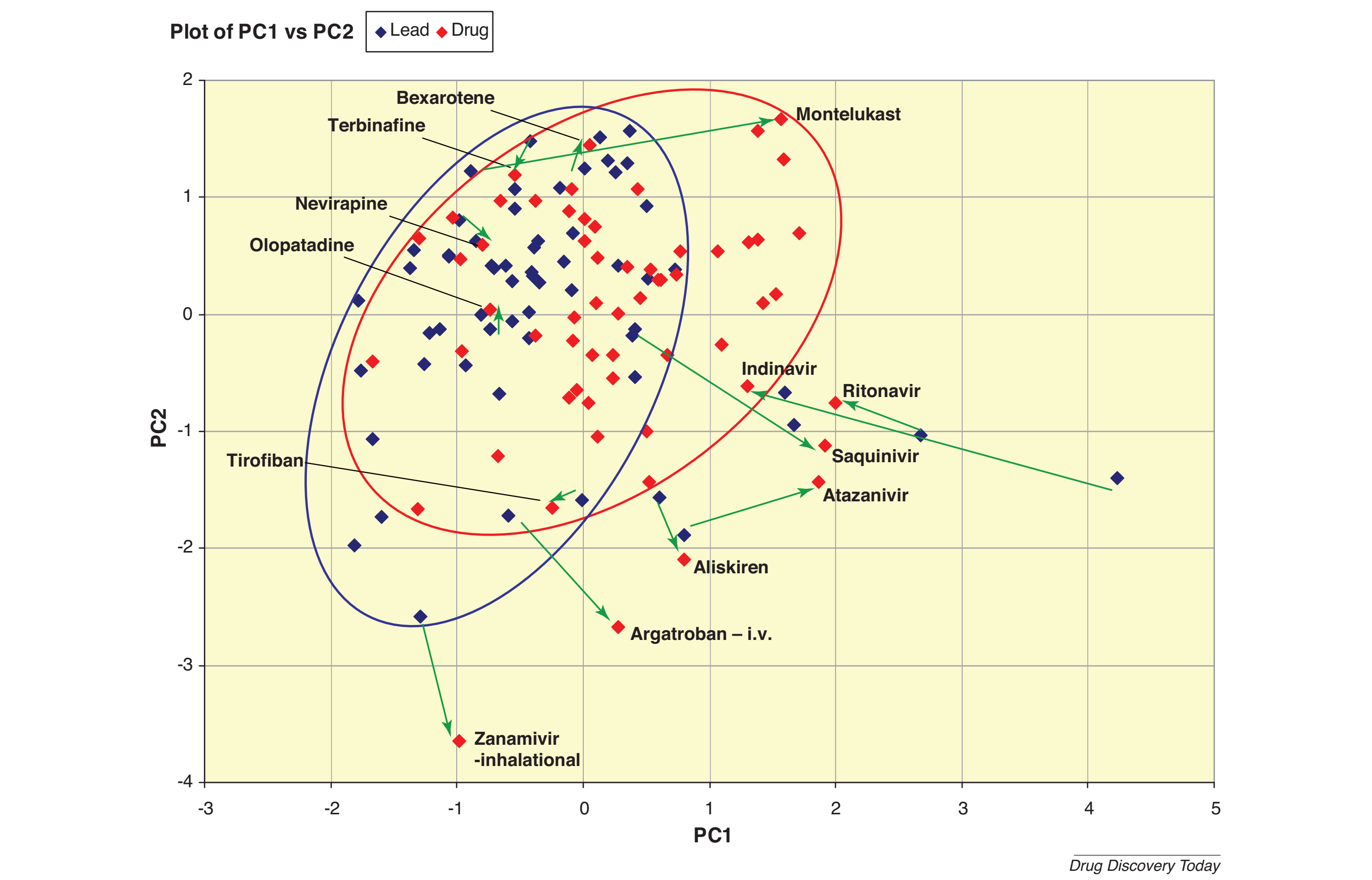


FIGURE 1

Principle component plot comparing pairs of leads (blue diamonds) and eventual drugs (red diamonds).

teams should not treat drug design as a journey along a straight path, with each new compound following the last, but rather imagine the process to be similar to navigating a maze, where barriers will be encountered requiring teams to retrace their steps and follow alternative directions.

Navigating the maze of compound optimization

The key to drug design and discovery is being able to judge when to stick by a choice previously made and to explore fully its opportunities and when to step back and follow a different direction. A key requirement of achieving this is for drug designers to resist the natural tendency to hope that '...the next compound I make might be the final compound', toward a mindset that says '...the the next compound I make should be the next step toward my final goal'. Every compound, active or inactive, is a step on a journey, with both helping to build the understanding necessary to make eventually the compound that will become the drug. This requires projects to begin with a mindset of exploration and to retain it until strong evidence indicates that a focus on a specific area is required.

The progression of a discovery project should be typified by a growing knowledge and understanding of the problems at hand and the potential to address them, with the latter stages characterized by bringing together previously successful approaches (molecular groups) to deliver interesting candidates. This can only be possible if the team has determined a sufficiently diverse SAR by adequately exploring the total chemical space around their series. It should then be possible to look back to earlier work to identify an R-group providing selectivity when the current selectivity inducing R-group has to be sacrificed in search of improved stability.

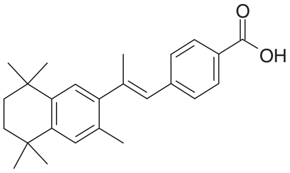
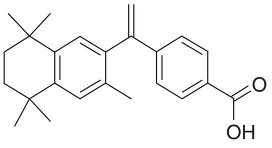
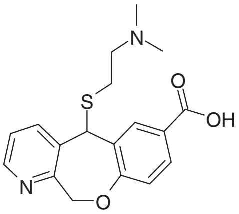
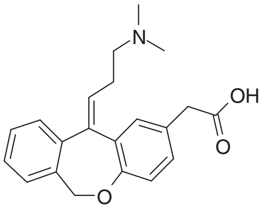
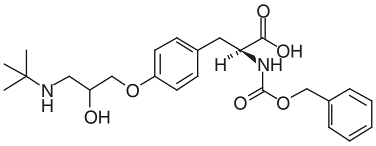
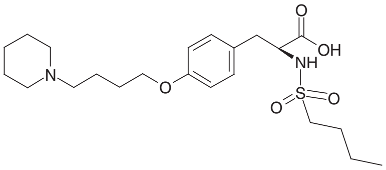
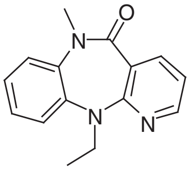
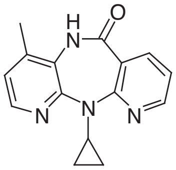
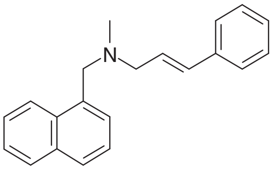
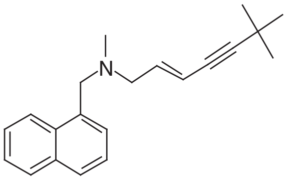
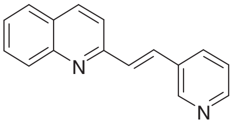
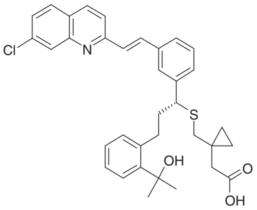
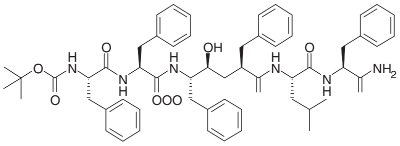
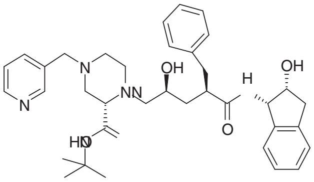
Similar to trying to escape a maze, at a certain point, progression will be hindered, leaving one literally or figuratively lost. When crossing the maze, the natural action when this state is reached is to track back to a previous crossing, the point at which a decision was made, and to choose a different path. Nothing is gained from repeatedly covering the same ground. Although it sounds trivial, the courage required to step back from a potent subseries toward a less potent more drug-like series is often the most important differentiator of success. Rewarding chemists based on the numbers of compounds they produce, or on the numbers of compounds above a certain activity threshold, has previously been a common practice within the industry, with the danger of reducing the openness of chemists to step back from compound subseries that are delivering bonuses even if they are unlikely to deliver drugs.

Building on the strongest possible foundation

It has been suggested that the presence of compounds with poor drug-like properties within discovery pipelines can be traced back to the nature of high-throughput screening hits and the associated hit-to lead practices [22]. Improving the overall quality of corporate screening collections has been an industry-wide goal for several years, with better collections generating higher quality

TABLE 1

Examples of leads and the resulting drugs for which (a) limited or (b) larger chemical modifications have been necessary

Lead structure	Drug structure
(a)	
(i) Bexarotene	
	
(ii) Olopatadine	
	
(iii) Tirofiban	
	
(iv) Nevirapine	
	
(v) Terbinafine	
	
(b)	
(vi) Montelukast	
	
(vii) Indinavir	
	

BOX 1

Compound Optimization Steps – 'God's Number' would require a list of marketed drugs and the initial lead from which they are derived, as well as defined rules on what constitutes a 'step'. There is no comprehensive resource of drugs and their leads, and a reported attempt to generate such a list describes the numerous hurdles involved, including lack of reporting and multiple leads contributing to individual drugs [101], but does include a useful partial list. Walter Sneader authored a book containing 'drug prototypes' of drugs [102], which has already been used to compare the physical and chemical differences between leads and drugs [103]; more recently, a collection of 60 lead–drug pairs, including binding data to their target of interest, has been published [3]. What is striking when browsing these 60 pairs is how similar many of the final drugs are to their lead, as shown in Table 1a, with examples of more significant modifications from lead to drug shown in Table 1b for comparison.

The limited number of steps between the leads and final drugs in so many cases illustrates the requirement to initiate compound optimization from the best possible starting points, and perhaps also the need to explore fully the chemistry around the lead before making larger synthetic steps. This also raises the concern that the limited steps seen for these successful pairs might demonstrate an inability to optimize compound series when bigger and more involved design intervention is required. If this is the case, it further reflects a failure in the ability to optimize 'difficult' compounds routinely and indicates a need to improve the approach.

Examination of the change in simple physicochemical properties on-going from lead to drug reveals that the separation between lead-like and drug-like physiochemical space is relatively small in most cases. Fig. 1 shows a principal component plot, where PC1 and PC2 represent 80% of the variance in a range of calculated descriptors (*c* log *P* [3], Molecular Weight, H-bond acceptors, H-bond donors, calculated logS, number of rings and number of rotatable bonds, data generated using MOE [104] descriptors and procedures) for the 60 lead–drug pairs. Ellipses are drawn, each covering 90% of leads (blue) and 90% of drugs (red).

Of course, a drug might end up having similar physicochemical properties to its lead, despite having undergone radical structural rearrangements; Fig. 1 nonetheless shows that the lead and drug principal component areas overlap to a remarkable degree and, in many cases, drugs will not differ dramatically from their leads. Illustrating the extremes in difference between lead and drug, Table 1 shows the five compounds closest in physicochemical space to their leads and the two drugs farthest from their given starting point [3]. For example, compared with its lead structure, bexarotene has only a slightly different linker between two aromatic rings, whereas olopatadine swaps an aromatic nitrogen for a carbon, a sulfur atom for a methine and sees one additional CH_2 group. Generally, pairs i–v in Table 1 show only slight alterations (e.g. an aromatic nitrogen for an aromatic carbon, a very small change in H-bonding groups and alterations in aliphatic/aromatic groups). Most changes on-going from lead to drug are greater than these five, but most only move ca one unit in Fig. 1. At the other end of the scale, the greatest changes are seen for montelukast and indinavir, pairs vi and vii in Table 1. The montelukast lead is at an extreme point on the blue lead oval in Fig. 1 and montelukast itself ends up close to the extremes of the red oval. Interestingly, the lead structure for indinavir is at an extreme point on plot X in Fig. 1, far from the lead or drug-like regions, yet the transformations made mean that indinavir itself falls just inside the red oval.

Overall, six drugs lie outside the red 'drug-like' oval of principal component space. Interestingly, three of these drugs are not delivered orally and need not be bound by the same physicochemical limits as oral drugs. The remaining four (Table 2) are aliskiren, a renin inhibitor and three HIV-1 protease inhibitors, saquinivir, atazanavir and ritonavir. These compounds show that it remains possible to deliver drugs with properties failing to satisfy the Lipinski rule of 5, provided that strong design rationale is in place. This is especially the case if supported by SBDD and intelligent bioisosteric substitutions [105–107], leading to clinically important drug molecules despite violating Lipinksi guidelines.

and more progressable hits. It is among the most important of all cheminformatics tasks to identify weaknesses in screening collections, such as poor physical properties or lack of diversity, and then facilitate the selection of compounds to purchase or synthesize to address these limitations [23].

If the quality of leads is the key determinant for success in lead optimization, and absorption, distribution, metabolism, excretion - toxicity (ADME-Tox) issues are more difficult to optimize than is potency [24], then prioritization during HTS triage must be directed to selecting more progressable compounds even at the expense of potency. Despite it being generally recognized that hit prioritization is the most important medicinal chemistry decision to be taken during a project [24], this evaluation will be based too often on limited biological data, unreliable 'chemical-eye' [25] and 'gut-feeling'. Given this limited information, it is not surprising that potency and ease of chemical synthesis tend to dominate selection rationale [26]. To be successful, a validated hit should have a surmountable number of liabilities and a maximum potential for progression, including a balanced profile of potency, efficiency, drug likeness and selectivity; in addition, medicinal chemists must consider each of these issues to improve the selections that they make [27].

The use of ligand-efficiency measures [28] can be a useful aid, provided sufficient reliable data are available, but must be accompanied by other knowledge approaches to HTS triage. This will include approaches to manage systematic errors in screening, such as plate effects, cell toxicity or enzyme degradation, as well as knowledge-based removal of known toxicophores and frequent hitters. *In silico* models to predict other undesirable activities, such as human Ether-á-go go Related Gene (hERG) liabilities and gene toxicity, have value for prioritization of hits even if a safety alert this early is no guarantee that the liability will still be present at the end of compound optimization. Cheminformatics methods are also important for the clustering of hits, analog identification and the rapid design of libraries around hits.

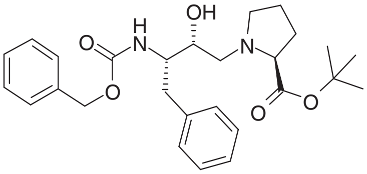
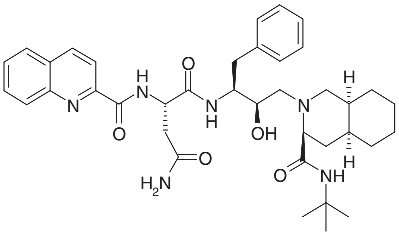
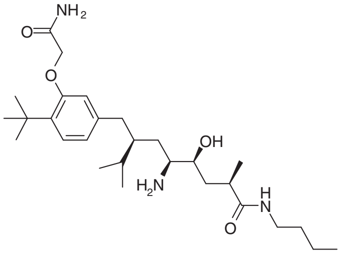
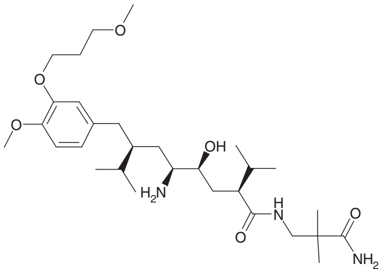
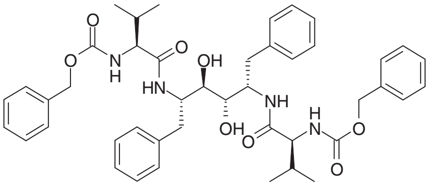
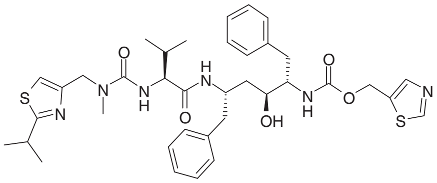
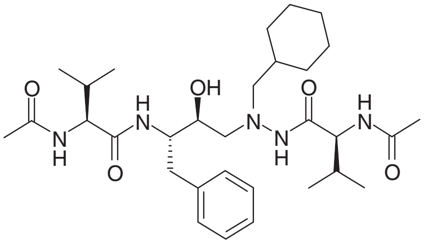
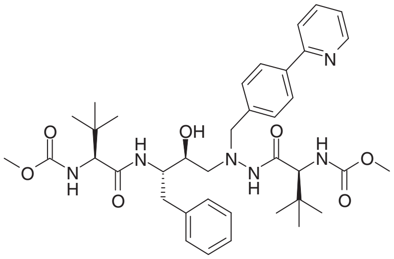
Although not a long-term strategy for successful drug design, there are cases, especially for new targets, where a rapid increase in potency might be required to allow validation of the biological hypothesis underpinning the project [26]. However, even in these situations, desirable drug likeness will facilitate comparison of *in vitro* and early *in vivo* results [5]. This will help avoid the common stumbling block of potent compounds *in vitro* showing no *in vivo* activity without teams being able to conclude if the biological rationale is wrong, if the compound is still not sufficiently potent, or if PK is the crucial limiting issue.

In silico methods for lead finding

The use of *in silico* methods for the identification of drug leads from databases of small molecules, commonly referred to as virtual

TABLE 2

Structures of oral leads and drugs outside the drug-like red oval (Fig. 1)

Lead structure	**Drug structure**
(i) Saquinivir	
	
(ii) Aliskiren	
	
(iii) Ritonavir	
	
(iv) Atazanavir	
	

screening (VS), has been extensively reviewed and critically evaluated in recent years (e.g. [29–31]). The first examples of compounds derived from VS approaches appear to be entering the clinic and even the market [30], but the overall impact of the approach remains open for discussion.

The efficacy of VS in 'positive design' for the selection of biologically active compounds directed to a specific target [31], similar to all aspects of Computer-Assisted Drug Discovery (CADD), can be improved by increasing the molecular understanding of the target structures or SAR that VS queries are built upon, and by ensuring the chemical integrity (stereo-isomers, tautomeric states, etc.) of the library compounds being searched [32]. There remain deficiencies in the methodologies used, although this is less evident in ligand-based VS compared with structure-based VS, where scoring, protein flexibility and the inclusion of water remain 'holy grails' [30].

'Negative design', for the elimination of most inactive or inappropriate compounds [31], to remove reactive, toxic or non-drug-like compounds, was originally a popular VS approach to pre-filter databases. However, this has since found a role in improving the quality of corporate screening collections [33].

The crossover between VS methods and HTS workflows is neither a new concept [33,34] nor limited only to 'negative design', with informatics methods also having an important role in selecting subsets of compounds to test (focused screening). Focused screening is a standard practice when dealing with some target classes, such as protein kinases [35], demonstrating huge enrichment improvements compared with full HTS [36]. The procedure enables teams to focus on higher quality chemical space and use lower throughput assays with a higher precision that is not compromised by the requirements of speed and capacity [37]. Again, the kinase family provides a clear demonstration, with the now common application of cross-screening (kinase-profiling or selectivity-screening) strategies where a small subset of kinase inhibitor-like compounds is tested against a panel of kinase targets rather than the single target more common in HTS [38]. The value of this chemogenomic approach is that both compound potency and selectivity can be evaluated simultaneously, which facilitates a more compound-centric approach to early discovery, with the quality of lead compounds contributing to the prioritization of targets and not pharmacology alone [38]. As ADME-Tox assays become increasingly high throughput, including them in these types of strategy will further improve the ability to identify progressable leads from primary screening [5], facilitating multi-parameter optimization in the future.

Chemogenomics: informatics bridging chemistry and biology

The broadest definition of chemogenomics is the understanding of the interaction between all possible ligands and all possible targets, but, at a more practical level, is characterized by the focused screening of libraries or compounds against multiple targets [39,40], and the related informatics methods for ligand and target classification and selection. Knowledge management is so fundamental to these activities that chemogenomics is often considered to be an informatics discipline [41], where targets are classified and studied as protein families, to facilitate the transfer of insight from one member of a family to another [42]. The first systematic application of such methods is often credited to the SARAH (SAR homology) project at GSK, relating protein sequences to small molecule SARs [43]. Additionally, less formal approaches allowing multidisciplinary researchers opportunity to leverage technology and knowledge gained from work on one target to other related targets have also been common.

Molecular informatics resources for chemogenomics or family-based research in general, are growing in breadth and depth from both a target and a ligand perspective. As an example, nuclear receptor target information is gathered by both the NUREBASE project [44] and the NucleaRDB project [45]. The NucleaRDB is accompanied by the Nuclear Receptor Structure Analysis Server (NRSAS) for structure prediction and related services [46], and the Nuclear Receptor Mutation Database [47] as well as further domain-specific analysis [48]. These tools allow for informed selection of selectivity panels and homology modeling, among other protein bioinformatics activities.

Chemogenomics tools from the ligand perspective center around databases of molecules linked to bioactivity data (quantitative and qualitative), including raw screening data in PubChem [49] and Chembank [50] and commercial SAR databases, such as WOMBAT [51], among others [52–54]. These tools facilitate the determination of specific properties differentiating between activity on specific target families, the elucidation of privileged scaffolds or the identification of interesting new bioisosteres. As with VS, chemogenomics therefore has the potential to provide better start points and directions for compound optimization.

The values of these ligand- and target-based chemogenomics resources are without question, but can be improved by their direct integration with in-house data resources. Big-pharma, in particular, have decades of screening and related data that might lack the chemical diversity of the publicly available data, but are likely to be of higher quality and, more importantly, multi-parameter in nature. Integrating this legacy data into chemogenomics projects is a way to leverage that historical knowledge and provide a competitive advantage [54]. Examples of this include Novartis linking 2.5 million in-house compounds to several public databases [55] and Pfizer, who built a data warehouse containing nearly 5 million chemical structures from their own internal repositories combined with data from public sources [56].

Information-rich compounds

Each step on the journey toward the delivery of a compound into development should increase knowledge and molecular understanding in the form of improved SAR, which, despite difficulties in extrapolation to the design and evaluation of new compounds, is the central pillar of all medicinal chemistry projects [57]. SAR has traditionally been elucidated on paper and in the heads of medicinal chemists. This suffers from subjective analysis, becomes increasingly more difficult as the number of compounds and properties increase, and is not suitable for considering multiple parameters simultaneously. SAR elucidation is therefore increasingly a 'data-mining' task, with both bench chemists and computational chemists making use of tools to visualize and interpret multivariate data for their compound series [58,59].

All SAR and quantitative SAR (QSAR) are dependent on the data from which they are derived, relying on the presence of informative compounds with a broad range of distinguishable physical–chemical properties. Previously, Craig plots and Topliss schemes have been used to help select synthetic modifications that will ensure either systematic chemical exploration or the generation of a property-rich data set for QSAR studies. As researchers wrestle with the complexities of multi-parameter drug design, these types of information-generating method might sound attractive, but have passed out of fashion since the implementation of parallel synthesis. It is now easier to make all available substituents [60] rather than carefully selecting an information-rich subset. This is not necessarily a disadvantage and should still allow the drawing of the same conclusions as from the consecutive application of either Topliss or Craig approaches; however, in practice, energy is focused on generating analogs of the most active compounds without investing time in understanding the patterns and trends generated. Convincing project teams to commit resources to generate information-rich compounds, that researchers suspect are unlikely to be active, but will be useful in improving understanding, is required if the project is to have access to good models early in the process.

Defining constraints for predictive models

A key responsibility of any computational drug discovery group is the design of general and bespoke predictive models. Provided data are available, models can be generated to address any issue in which activity or another property is related to compound structure. The quality of any model will be determined by the method used, the ability to select the most appropriate molecular descriptors, its applicability to the compounds to which it will be applied, and the validity of the data the model is based on [9]. Regardless of which models are used, teams must ensure that they re-evaluate their applicability routinely and, in the case of project-specific models, regularly update them. Last year's design insight can quickly become this year's misleading dogma.

The quality of models alone does not determine their value within projects, with technical [61] and psychological factors [62] playing roles. Informatics often relies on complex methods that might seem inaccessible to the non-expert and, therefore, practitioners must make the effort to find ways to explain concepts if they want colleagues to support them fully and to ensure that project teams recognize and appreciate the relative accuracy of any theoretical prediction of the activity of a compound.

Crucially, practitioners must be able to determine when the application of a model is useful, that is, optimally retaining or excluding the relevant set of desirable or undesirable compounds. This will depend on the predictive power of the model and likelihood that the properties it selects on will be present in the application set (low presence requires more accurate models to impact decision making) [61].

Drug designers should also appreciate that predicted tolerance across a range of a property (i.e. a clogP of between 2 and 5) is likely to have a bell-shaped curve distribution, with either extremity being less desirable chemical space than the mode (peak of curve value). This is particularly pertinent when dealing with multiple properties and criteria. A compound might satisfy every defined constraint in lead optimization, but if all these criteria are only just met, what is the likelihood that the compound will survive to the market? A compound achieving only minimum levels of solubility, potency and stability will probably require higher dosing, putting more pressure on its selectivity and toxicity profile. It is therefore important to avoid the 'tick-box' approach to multi-parameter design, checking off each parameter in isolation of others, and to evaluate holistically the value of the most progressed compounds.

Applying parallel models, predicting multiple end points (solubility, stability and potency) and excluding those that do not satisfy specific constraints, can be a powerful way to address some of the challenges of multi-parameter optimization [63]. Tools such as Pipeline Pilot, KNIME and InforSense provide infrastructure, via single workflow systems [64]. These can apply consecutive models, excluding failing compounds at each step [2], or be used to calculate all values for all compounds allowing one to identify balanced profiles.

Computational methods for multi-objective optimization

Recently, the first examples of techniques for objectively defining compromise between multi-parameters, referred to as multi-objective optimization (MOOP), have been described [65–68]. MOOP techniques try to optimize numerous dependent properties simultaneously to deliver a series of satisfactory compromises, and to avoid single objective dominant solutions. The two approaches currently described to achieve this are weighted scoring functions and Pareto-based methods [69]. Weighted scoring functions sum each individual design parameter (novelty, solubility, potency, etc.) with a weighting function associated with each parameter. Pareto-based methods are used to identify multiple solutions representing various compromises between the drug design parameters by selecting solutions that make at least one individual parameter better off without making any other parameter worse [70]. Examples of MOOP methods in cheminformatics can be found in the following recent reviews on the topic [69,71], and include its use in combinatorial library design [72] and *de novo* ligand design [73].

The medicinal chemistry knowledge worker

Without a doubt, the ability of pharmaceutical companies to manage and exploit their corporate knowledge is a crucial differentiator for success [74]. In the same way that the -omics technologies, dealing with their own 'data explosion', are now inseparable from bioinformatics, medicinal chemistry must become inseparable from cheminformatics [75]. Molecular biology has rapidly increased its ratio of *in silico* scientists compared with bench scientists and improved overall knowledge of informatics methods across its entire community. Medicinal chemistry has, in part, resisted this necessary change, but will have to address this deficiency in the short term, learning to treat computer literacy as a core skill of the bench chemist [76]. As projects increase the percentage of work being outsourced and the volume of the data to analyze, medicinal chemists must increasingly become knowledge workers, able to use informatics methods to manage their projects, track literature and competitor intelligence resources as well as to extract the most detail from their own SAR. Increasingly this is becoming the norm, with medicinal chemists using various knowledge-based tools to achieve this.

Less common is for bench chemists to use more complex informatics tools, which is unfortunate as experienced design chemists with additional understanding and insight of their SAR derived from the use of informatics tools will inevitably design better compounds. Achieving this requires the provision of training, time to apply these approaches and intuitive user interfaces, with most organizations using web-enabled systems to deliver these services [77]. Engaging bench chemists to undertake their own *in silico* experiments, supported by informatics experts, who are themselves freed to focus on more complex and difficult computational challenges, should be a crucial goal for modern drug discovery.

A commonly cited argument for encouraging chemists to undertake their own modeling is to increase their feeling of ownership, as they are more likely to consider a model they have developed themselves and then devote time to synthesizing molecules based on it [78]. Such anti-collaborative attitudes, if they really do exist, should not be tolerated, but it is undeniable that a feeling of shared ownership does encourage the use of models.

In addition to model building, the modern chemist must embrace a move from 2D to 3D drug design. Determining how differences between compounds, including conformation and electrostatics, alter their activity requires tools to compare them in 3D, and a feeling for how different groups influence shape and conformation. Scaffold hopping, consideration of pharmacophores, compound overlaying as well as any side-group modification, without sufficient 3D consideration, runs the risk of false comparison. Bench chemists must be able to consider the 3D implications of the design decisions that they make daily, requiring access to software, an understanding of the basic principles of force fields and the differences between local and global minima. These considerations are also important when studying ligand–receptor complexes, as are the skills to map different properties onto binding pockets, overlay multiple complexes, compare pockets, visualize non-bonded interactions for common fragments and the use of probe tools to identify regions of the binding pocket that are favorable for the addition of specific groups.

Numerous tools exist to facilitate these activities [79], but perhaps the sheer volume and diversity available is a hindrance to their broader use.

Modelers as core discovery team members

Most drug discovery organizations will already have teams of scientists within their molecular modeling, cheminformatics and CADD groups adept at handling and using multiple data sets [80], but too often these skills and methods sit at the periphery of the project teams [37]. Any successful informatics strategy must be focused on meeting the requirements of the modern drug design, which is only possible when fully integrated within the discovery process [37]. Most pharmaceutical companies use a project team structure with a small team of pharmacologists, molecular biologists and chemists leading the drug discovery project. The inclusion of computational chemists as core decision makers within these teams is important, as this is the only way to ensure that the modelers are fully informed about the problems at hand, able to define the questions they will tackle, and have influence to ensure that their methods are taken up. CADD must be considered a core technology within discovery projects rather than a service, with modelers sharing equal ownership and responsibility for projects. This challenges modelers to become broader and more knowledgeable of all aspects of the drug discovery process and behave as drug hunters and not just informatics or IT support for projects. It is also an absolute requirement that the molecular modeling experts within an organization are in close contact with their bench colleagues. Cross-fertilization and stimulation of ideas between the two disciplines will increase the quality and efficiency of both groups.

A recurring hindrance to the timely application of molecular informatics has been the availability of resources. Too often modelers are assigned to multiple projects, dividing their time over numerous diverse problems. This situation prevents the computational chemist from focusing sufficiently on each problem, resulting in the use of suboptimal approaches and delaying the delivery of bespoke models and insight. Modelers should be equally engaged as lead chemists, who are rarely asked to work on multiple projects in lead optimization, and share with them an equal appreciation of SAR, planned synthesis and other project issues.

Modeling groups also tend to be composed largely, if not exclusively, of PhD-level scientists without access to technicians, whose presence would allow research scientists to focus more time on the interpretation and exploitation of data and less time on its generation and management. Another hurdle to the timely application of CADD can be the desire of computational chemists to improve their models before applying them in the hope that its next iteration will be an improvement on the previous version. As scientists, researchers are naturally eager to produce the most accurate and relevant output, always striving to improve the quality of their work, but this can result in repeated delays. Researchers have to become better at saying, ‘good is good enough’ and to accept that the search for perfection will actually diminish the positive impact on drug design. This requires modelers to accept the risk of being wrong, itself requiring management to foster an environment where presenting a challenging, but potentially valuable new idea is encouraged.

It is important for the success of CADD groups that they have an expert in each of the most important fields (methods or target families) who can define best practice and monitor and critically evaluate developments in their area. However, it is at least as important that all CADD scientists, representing this skill base within projects, have a broad knowledge of the field and are able to apply a diverse range of computational medicinal chemistry approaches. A failing of molecular modelers in the past has been a reluctance to apply a more appropriate *in silico* technique outside the realm of their particular expertise. Productive computational medicinal chemists will identify the question most important to the team and then identify which of a broad array of tools can be used to devise strategies to formulate answers.

Experimental design, without consideration for data analysis, is flawed, and the generation of increasingly large data resources within projects, without provision of the resources required for its analysis and exploitation ensures that research fails to benefit fully from the investments it makes at the bench or within its automated laboratories.

Integrating computational and synthetic chemistry

Bench-based medicinal chemists and their computational medicinal chemistry colleagues share the same primary goal, which is the design of new and commercially relevant chemical entities. This shared objective can create tension between these two overlapping disciplines, with a series of often repeated criticisms directed in both directions. The most common objection from computational chemists toward their bench colleagues is a lack of willingness to synthesize the compounds that they suggest. From the opposite direction one often hears the comment, ‘Your idea for a new NCE looks great on the computer screen, but how will I be able to synthesize it?’.

Experienced and knowledgeable modelers are able to ensure that compounds suggested by CADD can be synthesized within the constraints of a normal design strategy. Designing compounds using available building blocks, intermediates and the reactions already being used within a project might also help encourage the synthesis of NCEs based on modeling and informatics. The drawback is that this narrow approach does not allow CADD methods to challenge the existing presumptions and dogmas of the project and might result in simply pre-empting choices the bench chemist would have made anyway.

Modelers should not hesitate to challenge the synthetic chemist to make difficult compounds if supported by reliable models or insight. A compound being more difficult to make is insufficient argument not to make it, but it is the responsibility of the modeler to demonstrate to the synthetic chemist why they strongly believe that it should be considered. The synthetic chemist, in turn, has the responsibility to make the compound if the rationale is strong enough, even if this means not making compounds of their own design. There is always a reason not to make a compound, as no compound, even those on the market, is ever perfect. It is therefore possible to place unrealistic and artificial barriers in front of any challenging idea, inevitably at the cost of the project.

There also remains a common misconception that the promise of computational approaches in rational drug design is to allow the reduction of bench chemists employed by pharma. There is in fact a growing argument that, for CADD to have impact, more chemists are required, not fewer. CADD will suggest new directions and new chemical space to explore, and will require chemists to undertake newer and more difficult synthesis than they might choose otherwise. All of this will be in addition to the chemistry that the bench chemists want to investigate themselves, and only partially compensated for by the ability to identify and discontinue unproductive chemical space. Therefore, projects with a high CADD component are likely to require more, rather than fewer, chemists. The benefit to the drug-discovery company will be more diverse compounds coming from projects with a wider range of properties and characteristics eventually resulting in more and better drugs in the pipeline, where efficiency and value are really generated. Given an estimate of 95% attrition in clinical development [7], just a 5% reduction in this failure would achieve a doubling of the number of compounds reaching the market [4]. This should be more efficient and better value than the alternative of doubling the number of compounds going into development.

In the current climate of pharma mergers, consolidation and streamlining, arguments for greater manpower might fall on deaf ears. However, with the increasing availability of outsourced chemistry [81], it should still be possible to support sufficiently more information-driven drug design.

Getting more from structure-based drug design

Protein–ligand co-crystals are one of the most valuable resources for molecular design, providing an excellent opportunity to optimize compounds rationally. It can, however, be argued that their application within multi-parameter design projects is not yet being fulfilled. The first barrier has been the mindset that structure-based drug design (SBDD) is only a tool for improving potency or selectivity by altering the number and strength of intermolecular interactions. SBDD can be just as useful in identifying regions that will tolerate the addition of solubilizing groups or selecting which functions can be safely removed or protected without losing activity on the target.

Historically, there has been a separation between the structural biologists generating new protein–ligand complexes and the modelers and medicinal chemists responsible for new compound design, but that gap can be bridged. Modelers should become more structural biology aware, making use of electron density and R-factors as measures of flexibility and not just final coordinates [37,82]. We have found that giving modelers the training to be able to calculate density and models from X-ray diffraction data independently, along with direct support from structural biology experts, ensures a fuller appreciation of the subtleties and nuances in the structures that they are basing decisions upon, resulting in more informed and reliable compound design.

X-ray structures, as with all forms of data, are dependent on timely availability to ensure their true value. Historically, SBDD was undertaken later in projects and used to improve already well-optimized compound series, but there is growing acceptance that access to co-crystals will have greater impact the earlier in the process they can be obtained. The later projects have access to this resource, the more entrenched the poor characteristics of compound series are likely to be. Instead of hoping that SBDD can help dig one out of a hole, projects will be better served using it to avoid traps in the first place. The growing influence of fragment-based drug design (FBDD) contributes to this view.

FBDD and related informatics

FBDD hits, owing to their reduced size and complexity, tend to have lower potency than do HTS hits, but higher or at least comparable ligand efficiencies and are more progressable. Owing to the complexity of most compounds in corporate screening collections, there is a good possibility that HTS hits will include both good and bad interactions with the target. FBDD hits, by contrast, are feature poor, reducing the potential to include bad interactions. FBDD is also able to sample chemical space more efficiently than is HTS owing to the lower number of potential fragments compared with drug-like molecules [83]. Since Abbott's first description of the practice [84], its adoption industry wide has been rapid, with 17 clinical candidates derived from FBDD reported in a 2009 review [85]. This success is despite the fact that FBDD is often applied only after the failure of traditional screening methods.

As with all screening campaigns, the quality of the fragment library will determine the eventual outcome and, therefore, building fragment libraries is the first crucial step in FBDD. The most frequently applied design rules for fragment libraries are based on Astex's 'rule of three' [86] (MW $\leq$300 Da, $c\log P \leq 3$, H-bond donors $\leq$3 and rotatable bonds $\leq$3) as well as additional measures, such as polar surface area $\leq$60 Å^2 [85] and consideration of synthetic tractability [87]. Additional target-specific restrictions can be included based on available knowledge, for example, our own fragment library includes kinase hinge-binding fragments created from an analysis of the kinase crystal structure entries of the Protein Data Bank.

After the design of fragment libraries, either generic or target focused, and the triage of potential hits, the next informatics goals revolve around the expansion of fragments to better occupy the binding pocket or the linking together of multiple fragments in one pocket. As such, fragment hits typically require access to experimentally determined structural data describing binding to targets for direct optimization [88], which is especially the case owing to the lack of *in vitro* measured data for SAR analysis [87]. Methods to predict theoretically binding modes are hampered by the difficulty of using docking methods to predict binding orientations of small molecules, and the ubiquitous SBDD issues addressed earlier.

Informaticians are used to approaching drug design from a fragment perspective as they often reduce complex molecules to smaller substructures for the calculation of various properties, including logP and potential toxicology [89]. It is therefore not a surprise that *de novo* design methods already provide an array of methods for the optimization and joining of fragments in the context of their targets [90,91].

De novo design is also the subject of recent publications related to its application in MOOP studies [73,92,93] as its role in FBDD makes it essential to ensure that optimization, even from these attractive start-points, follows tractable paths. *De novo* design methods also pose the challenge of translating *in silico* ideas into synthetically amenable plans, which continues to be difficult, not least because synthetic feasibility is a subjective measure dependent on a chemist's experience and knowledge [94].

The requirement for access to X-ray co-crystals and availability of protein in suitable form for screening, limits the number of amenable targets to which FBDD can be applied. Abbott, a founder and current leader in FBDD, estimates that it is a method applicable to just 30% of their targets [88]. Medicinal chemists, reluctant to work on low potency hits, are a further hindrance to the application of FBDD [83].

Technology development

Molecular informatics remains subject to rapid new technology development within academia, commercial research organizations (CROs) and directly within industry. Ensuring this effort is directed to solving the actual problems that drug designers face within their discovery projects is important to maximize this effort and avoid the generation of 'white elephant' applications. Strong partnerships between industry and academia are crucial here, and should involve broad collaboration, long-term commitment and bidirectional sharing of knowledge.

There is an argument that real scientific breakthroughs in computational drug design methodologies have been scarce in recent years [37], but there is scope to ensure that the most value can be extracted from existing methods. This means optimizing how existing methods can be embedded in the drug discovery process and tailoring current applications to research needs. Examples within Merck include expert user tools, such as Fleksy [95], a flexible protein docking method developed in collaboration with the Computational Drug Discovery group at the Radboud University in Nijmegen, and based on the existing FlexX [96] and FlexE [97] docking methods. Non-expert tools include SyGMA [98], a method for the prediction of metabolic transformations based on rules derived from the MDL Metabolite Database (http://www.mdl.com/products/predictive/metabolite) and IBIS [99], a bioisoster generator relying on a combination of theoretically proposed and experimentally derived bioisosters from the BIOSTER database [100] and the mining of our in-house corporate databases. Both SyGMA and IBIS are web tools available to, and heavily used by, bench scientists via a common web interface.

IBIS is a result of mining 188 assays covering 61 targets from six target families, from Organon and Schering-Plough projects, to identify and record every tolerated chemical change made during compound optimization. This enables anyone to query which groups have previously shown similar or better activity compared with their compound and, as such, is an important tool for retaining corporate knowledge. This will return the bioisosters that all experienced medicinal chemists can name, but also many less well-known examples, putting the knowledge of generations of drug designers at everyone's fingertips. Queries against this database can be more specific; for example, it can be used to suggest only bioisosters that are more polar than the input and already appear regularly in drug databases. As an additional chemogenomics consideration, all bioisosters are additionally classified by target family, allowing for more specific decision making. If knowledge is the most important resource in the pharmaceutical and biotech industries [60], then these methods are crucial to ensure that it is retained and exploited.

Concluding remarks

We often compare drug discovery to finding a needle in a haystack. If this was the case, then one might presume that the increasing scale and industrialization of drug discovery should be bearing fruit as researchers have spent the past few decades developing tools to discard hay more rapidly. Unfortunately, this does not appear the case. The haystack analogy also fails to consider that each piece of hay can easily be discarded as it is clearly not a needle and that, when you eventually find the needle, you can be sure that your task is complete.

Drug discovery can therefore be better compared to searching for a very specific needle in a very large stack of other needles. Similar to molecules in chemical space, many of the needles will appear to be identical or at least very similar. Finding the small differences that separate most of the needles from the small group of desired needles is a significant problem, but cannot be solved by focusing on just one or two characteristics. Informatics will provide the most important tools within the medicinal chemist's arsenal for the exploration and understanding of these issues.

Acknowledgements

The authors would like to thank and acknowledge their colleagues within the department of Molecular Design and Informatics for countless shared ideas and experiences over the past 10 years.

References

1 Clark, D.E. (2006) What has computer-aided molecular design ever done for drug discovery? *Exp. Opin. Drug Discov.* 1, 103–110

2 Wunberg, T. *et al.* (2006) Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* 11, 175–180

3 Perola, E. (2010) An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem.* 53, 2986–2997

4 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* 6, 881–890

5 Kerns, E.H. and Di, L. (2003) Pharmaceutical profiling in drug discovery. *Drug Discov. Today* 8, 316–323

6 Jorgensen, W.L. (2004) The many roles of computation in drug discovery. *Science* 303, 1813–1818

7 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715

8 Macdonald, S.J.F. and Smith, P.W. (2001) Lead optimization in 12 months? True confessions of a chemistry team. *Drug Discov. Today* 6, 947–953

9 Brown, S.P. *et al.* (2009) Healthy skepticism: assessing realistic model performance. *Drug Discov. Today* 14, 420–427

10 Agrafiotis, D.K. *et al.* (2007) Advanced biological and chemical discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model.* 47, 1999–2014

11 Waller, C.L. *et al.* (2007) Strategies to support drug discovery through integration of systems and data. *Drug Discov. Today* 12, 634–639

12 Rojnuckarin, A. *et al.* (2005) ArQiologist: an integrated decision support tool for lead optimization. *J. Chem. Inf. Model.* 45, 2–9

13 Cho, S.J. *et al.* (2006) ADAAPT: Amgen's data access analysis, and prediction tools. *J. Comp. Mol. Design.* 20, 249–261

14 Sander, T. *et al.* (2009) OSIRIS, an entirely in-house developed drug discovery informatics system. *J. Chem. Inf. Model.* 49, 232–246

15 Proudfoot, J.R. (2002) Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg. Med. Chem. Lett.* 12, 1647–1650

16 Zhao, H. and Guo, Z. (2009) Medicinal chemistry strategies in follow-on drug discovery. *Drug Discov. Today* 14, 516–522

17 Kuntz, I.D. *et al.* (1999) The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9997–10002

18 Bissantz, C. *et al.* (2010) A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* 53, 5061–5084

19 Bohacek, R.S. *et al.* (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16, 3–50

20 Hajduk, P.J. (2006) Fragment-based drug design: how big is too big? *J Med. Chem.* 49, 6972–6976

21 Shigemizu, D. *et al.* (2009) Extraction and analysis of chemical modification patterns in drug development. *J. Chem. Inf. Model.* 49, 1122–1129

22 Kesurü, G.M. and Makara, G.M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* 8, 203–212

23 Agrafiotis, D.K. *et al.* (2007) Recent advances in chemoinformatics. *J. Chem. Inf. Model.* 47, 1279–1293

24 MacCoss, M. and Baillie, T.A. (2004) Organic chemistry in drug discovery. *Science* 303, 1810–1813

25 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* 47, 4891–4896

26 Schnecke, V. and Boström, J. (2006) Computational chemistry-driven decision making in lead generation. *Drug Discov. Today* 11, 43–50

27 Leeson, P.D. *et al.* (2004) Drug-like properties: guiding principles for design – or chemical prejudice? *Drug Discov. Today: Technol.* 1, 189–195

28 Hopkins, A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431

29 Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature* 432, 862–865

30 Clark, D.E. (2008) What has virtual screening ever done for drug discovery? *Exp. Opin. Drug Discov.* 3, 841–851

31 Schneider, G. (2010) Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* 9, 273–276

32 Klebe, G. (2004) Lead identification in post-genomics: computers as a complementary alternative. *Drug Discov. Today: Technol.* 1, 225–230

33 Stahura, F.L. and Bajorath, J. (2004) Virtual screening methods that complement HTS. *Combinat. Chem. High Throughput Screen* 7, 259–269

34 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894

35 Akella, L.B. and DeCaprio, D. (2010) Cheminformatics approaches to analyse diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* 14, 325–330

36 Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today* 11, 277–279

37 Stahl, M. *et al.* (2006) Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discov. Today* 11, 326–333

38 Goldstein, D.M. *et al.* (2008) High-throughput kinase profiling as a platform for drug discovery. *Nat. Rev. Drug Discov.* 7, 391–397

39 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262–275

40 Maréchal, E. (2008) Chemogenomics: a discipline at the crossroad of high throughput technologies, biomarker research, combinatorial chemistry, genomics, cheminformatics, bioinformatics and artificial intelligence. *Combinat. Chem. High Throughput Screen* 11, 583–586

41 Harris, C.J. and Stevens, A.P. (2006) Chemogenomics: structuring the drug discovery process to gene families. *Drug Discov. Today* 11, 880–888

42 Jacoby, E. (2006) Chemogenomics: drug discovery's panacea? *Mol. BioSyst.* 2, 218–220

43 Frye, S.V. (1999) Structure–activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. *Chem. Biol.* 6, R3–R7

44 Duarte, J. *et al.* (2002) NUREBASE: database of nuclear hormone receptors. *Nucleic Acids Res.* 30, 364–368

45 Horn, F. *et al.* (2001) Calculating and harvesting biological data: the GPCRDB and NucleaRDB. *Nucleic Acids Res.* 29, 346–349

46 Bettler, E. *et al.* (2003) NRSAS: nuclear receptor structure analysis servers. *Nucleic Acids Res.* 31, 3400–3403

47 van Durme, J.J.J. *et al.* (2003) NRMD: nuclear receptor mutation database. *Nucleic Acids Res.* 31, 331–333

48 Folkertsma, S. *et al.* (2005) The nuclear receptor ligand-binding domain: a family-based structure analysis. *Curr. Med. Chem.* 12, 1001–1016

49 Wheeler, D.L. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35, D5–D12

50 Seiler, K.P. *et al.* (2007) A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36, D351–D359

51 Olah, M. *et al.* (2005) WOMBAT: world of molecular bioactivity. In *Cheminformatics in Drug Discovery*, (Vol. 23) (Oprea, T.I., ed.), pp. 249–272, Wiley-VCH

52 Oprea, T.I. and Tropsha, A. (2006) Target, chemical and bioactivity databases – integration is key. *Drug Discov. Today* 3, 357–365

53 Doddareddy, M.R. *et al.* (2009) Chemogenomics: looking at biology through the lens of chemistry. *Stat. Anal. Data Mining* 2, 149–160

54 Senger, S. and Leach, A.R. (2008) Chapter 11 SAR knowledge bases in drug discovery. *Annual Reports in Computational Chemistry*, (Vol. 4), pp. 203–216, Elsevier

55 Zhou, Y. *et al.* (2007) Large-scale annotation of small-molecule libraries using public databases. *J. Chem. Inf. Model.* 47, 1386–1394

56 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815

57 Delaney, J. (2008) Modelling iterative compound optimization using a self-avoiding walk. *Drug Discov. Today* 14, 198–207

58 Wawer, M. *et al.* (2010) Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov. Today* 15, 630–639

59 Howe, T.J. *et al.* (2007) Data reduction and representation in drug discovery. *Drug Discov. Today* 12, 45–53

60 Hopkins, A.L. and Polinsky, A. (2006) Knowledge and intelligence in drug design. *Annu. Rep. Med. Chem.* 41, 425–437

61 Segall, M.D. and Chadwick, A.T. (2010) Making priors a priority. *J. Comp. Mol. Discov.* 24, 957–960

62 Chadwick, A.T. and Segall, M.D. (2010) Overcoming psychological barriers to good discovery decisions. *Drug Discov. Today* 15, 561–569

63 Shimada, J. *et al.* (2002) Integrating computer-based *de novo* drug design and multidimensional filtering for desirable drugs. *Targets* 1, 196–205

64 Tiwari, A. and Sekhar, A.K.T. (2007) Workflow based framework for life science informatics. *Computat. Biol. Chem.* 31, 305–319

65 Agrafiotis, D.K. (2002) Multiobjective optimization of combinatorial libraries. *J. Comput. Mol. Design* 16, 335–356

66 Gillet, V.J. *et al.* (2002) Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Computat. Sci.* 42, 375–385

67 Nicolotti, O. *et al.* (2002) Multiobjective optimization in quantitative structure–activity relationships: deriving accurate and interpretable QSARs. *J. Med. Chem.* 45, 5069–5080

68 Ekins, S. *et al.* (2002) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *Mol. Divers.* 5, 255–275

69 Nicolaou, C.A. *et al.* (2007) Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Dev.* 10, 316–324

70 Schneider, G. and Baringhaus, K.-H. (2008) *Molecular Design: Concepts and Applications*. Wiley

71 Ekins, S. *et al.* (2010) Evolving molecules using multi-objective optimization: applying to ADME/Tox. *Drug Discov. Today* 15, 451–460

72 Agrafiotis, D.K. (2002) Multiobjective optimization of combinatorial libraries. *J. Comp. Mol. Design* 16, 335–356

73 Dey, F. and Caflisch, A. (2008) Fragment-based *de novo* ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* 48, 679–690

74 Hoffmann, T. (2010) The future of discovery chemistry: *quo vadis?* Academic to industrial – the maturation of medicinal chemistry to chemical biology. *Drug Discov. Today* 15, 260–264

75 Bajorath, J. *et al.* (2001) Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* 6, 989–995

76 Davenport, T.H. *et al.* (2005) Human aspects of the management of drug discovery knowledge. *Drug Discov. Today: Technol.* 2, 205–209

77 Muchmore, S.W. *et al.* (2010) Cheminformatic tools for medicinal chemists. *J. Med. Chem.* 53, 4830–4841

78 Gund, P. *et al.* (2009) Informatics and computation for drug discovery – specialist or mainstream tools? *Curr. Opin. Drug Discov. Dev.* 12, 337–338

79 Kirchmair, J. *et al.* (2008) The protein data bank (PDB), its related services and software tools as key components for *in silico* guided drug discovery. *J. Med. Chem.* 51, 7021–7040

80 Andersson, S. *et al.* (2009) Making medicinal chemistry more effective – application of Lean Sigma to improve processes, speed and quality. *Drug Discov. Today* 14, 598–604

81 Clark, D.E. and Newton, C.G. (2004) Outsourcing lead optimisation – the quiet revolution. *Drug Discov. Today* 9, 492–500

82 Davis, A.M. *et al.* (2008) Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today* 13, 831–841

83 Congreve, M. *et al.* (2008) Recent developments in fragment-based drug discovery. *J. Med. Chem.* 51, 3661–3680

84 Shuker, S.B. *et al.* (1996) Discovering high affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534

85 Law, R. *et al.* (2009) The multiple roles of computational chemistry in fragment-based drug design. *J. Comp. Mol. Design.* 23, 459–473

86 Rees, D.C. *et al.* (2004) Fragment-based lead discovery. *Nat. Rev. Drug Discov.* 3, 660–672

87 Hubbard, R.E. *et al.* (2007) Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discov. Dev.* 10, 289–297

88 Hajduk, P.J. and Greer, J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.* 6, 211–219

89 Villar, H.O. and Hansen, M.R. (2007) Computational techniques in fragment based drug discovery. *Curr. Topics Med. Chem.* 7, 1509–1513

90 Schneider, G. and Fechner, U. (2005) Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* 4, 649–663

91 Mauser, H. and Guba, W. (2008) Recent developments in *de novo* design and scaffold hopping. *Curr. Opin. Drug Discov. Dev.* 11, 365–374

92 Nicolaou, C.A. *et al.* (2009) *De novo* drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.* 49, 295–307

93 Kutchukian, P.S. *et al.* (2010) *De novo* design: balancing novelty and confined chemical space. *Exp. Opin. Drug Discov.* 5, 789–812

94 Boda, K. *et al.* (2007) Structure and reaction based evaluation of synthetic accessibility. *J. Comp. Mol. Design.* 21, 311–325

95 Nabuurs, S.B. *et al.* (2007) A flexible approach to induced fit docking. *J. Med. Chem.* 50, 6507–6518

96 Rarey, M. *et al.* (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489

97 Claussen, H. *et al.* (2001) Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* 308, 377–395

98 Ridder, L. and Wagener, M. (2008) SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* 3, 821–832

99 Wagener, M. and Lommerse, J.P.M. (2006) The quest for bioisosteric replacements. *J. Chem. Inf. Model.* 46, 677–685

100 Ujvary, I. (1997) BIOSTER – a database of structurally analogous compounds. *Pestic. Sci.* 51, 92–95

101 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315

102 Sneader, W. (1996) *Drug Prototypes and their Exploration.* John Wiley & Sons

103 Hann, M.M. (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41, 856–864

104 Vilar, S. *et al.* (2008) Medicinal Chemistry and the Molecular Operating Environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr. Topics Med. Chem.* 8, 1555–1572

105 Rahuel, J. *et al.* (2000) Structure-based drug design: the discovery of novel nonpeptide orally active inhibitors of human rennin. *Chem. Biol.* 7, 493–504

106 Wlodawer, A. and Vondrasek, J. (1998) Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* 27, 249–284

107 Bold, G. *et al.* (1998) New aza-dipeptide analogues as potent and orally absorbed HIV-1 protease inhibitors: candidates for clinical development. *J. Med. Chem.* 41, 3387–3401