ELSEVIER

All scientific disciplines, including medicinal chemistry, are the subject of a revolution as data are generated at unprecedented rates and their analysis and exploitation become increasingly fundamental to innovation.

Data-driven medicinal chemistry in the era of big data

Scott J. Lusher[1,2], Ross McGuire[2,3], René C. van Schaik[1], C. David Nicholson[4] and Jacob de Vlieg[1,2]

[1] Netherlands eScience Center, Amsterdam, The Netherlands
[2] Computational Drug Discovery Group, Radboud University, Nijmegen, The Netherlands
[3] Bioaxis Research, Pivot Park, Oss, The Netherlands
[4] Bayer Crop Science, Monheim am Rhein, Germany

Science, and the way we undertake research, is changing. The increasing rate of data generation across all scientific disciplines is providing incredible opportunities for data-driven research, with the potential to transform our current practices. The exploitation of so-called 'big data' will enable us to undertake research projects never previously possible but should also stimulate a re-evaluation of all our data practices. Data-driven medicinal chemistry approaches have the potential to improve decision making in drug discovery projects, providing that all researchers embrace the role of 'data scientist' and uncover the meaningful relationships and patterns in available data.

Introduction

How we manage and explore data is not a new topic in pharmaceutical research but has become acutely more important as the perception grows that we are now producing data at a faster rate than we can analyse, interpret and base decisions upon it. Data-driven drug design is dependent on medicinal chemists (computational and synthetic) dealing with the growth in data volumes and finding ways to convert these resources into better decisions.

Data-driven research has two interconnected and equal branches:

i. Ensuring the most benefit can be extracted from the data you generate internally.
ii. Incorporating externally available data resources into your decision making.

The term 'big data' has recently entered the common lexicon with the mainstream media regularly discussing the implications and opportunities of data as the 'new oil'. The subject is most often presented in the context of business intelligence and the use of information resources and social media to understand and target consumers better. The apparent big-data revolution is, however, just as relevant in scientific research, with a growing need to manage increasing data resources and utilise the potential to enable a greater degree of data-driven decision making.

The universal importance of the big-data challenge across all scientific fields means that methods and approaches developed in one field could have potential applications in seemingly unrelated disciplines. Taking advantage of developments and protocols from other scientific

Corresponding author:. Lusher, S.J. (s.lusher@esciencecenter.nl)

Scott J. Lusher worked in pharmaceutical R&D (Organon, Schering-Plough and Merck) for ten years (2001–2011) providing the molecular basis and rationale for the selection and design of NCEs within cross-disciplinary projects. Additionally, he had a role ensuring the strategic management and project implementation of molecular informatics within research. Scott began his industrial career at Unilever Research (1998–2001) as project leader for an initiative utilising molecular informatics techniques for the discovery of bioactive compounds. Lusher joined NLeSC in 2011 as part of its management team and is currently Director Applied eScience developing new scientific applications of ICT. He has a part-time appointment with at the Radboud University Medical Center and has a PhD in computational drug design from the same institute.

Ross McGuire studied chemistry at the University of Edinburgh, gaining a PhD on novel heterocyclisation reactions. In 1987 he joined Rhône-Poulenc Agrochemistry in Essex, UK, as a synthetic chemist, also working in Lyon, France, training as a molecular modeller. He joined Organon in Scotland in 1991, working on CNS and neuromuscular blocking projects. In 2000, he moved to Organon's Oss research centre in The Netherlands, where he was appointed head of *In Silico* Drug Design in 2002. He led a multidisciplinary group in MSD with expertise in modelling, chem-/bio-informatics and SBDD. In 2011 he founded the European cheminformatics consultancy, BioAxis Research BV.

René C. van Schaik studied biophysical chemistry and informatics at the VU University of Amsterdam and the University of Groningen. His PhD research (RU Groningen and ETH Zürich) focused on 3D structure refinement of biomacromolecules using NMR or X-ray data. In 1993 he joined Unilever where he held several research positions including Skill-base Leader Bioinformatics. In 1998 he became Program Leader Bioinformatics and Head of IT at Keygene NV. In 2001 he joined Organon NV and, after merger with Schering-Plough, became Head of Molecular Informatics and Profiling for Europe. In July 2011 he joined The Netherlands eScience Center as CTO responsible for the development and implementation of innovative eScience solutions in multidisciplinary research projects.

C. David Nicholson is Head of R&D with Bayer CropScience, since March 2012. He graduated in pharmacology from the University of Manchester (1975) and obtained his PhD from the University of Wales in 1980. In 1978 he joined Beecham-Wülfing as Group Leader Cardiovascular Drugs. In 1988 he moved to Organon Bioscience and finally became Executive Vice President of Global R&D. When Schering-Plough acquired Organon in 2007 Dr Nicholson moved to the company's headquarters at Kenilworth (NJ, USA), where he was responsible for Global Project Management and Drug Safety. When Schering-Plough and Merck merged in 2009 he was nominated Senior Vice President Licensing and Knowledge Management based at Rahway (NJ, USA). He has been Member of the Board of several start-up companies including Actinium Pharmaceuticals.

Jacob de Vlieg studied biophysics at the University of Groningen and graduated *cum laude*. During his PhD research, he developed computational methods for 3D biostructure determination. Shortly thereafter, he joined EMBL, Heidelberg, to develop structural bioinformatics techniques. From 1990 until 2001, de Vlieg held a range of research and management positions at Unilever Research. Appointed in 2000, he is currently part-time professor, Computational Chemistry, at the Radboud University Nijmegen Medical Centre. De Vlieg joined Organon in 2001, as head of Molecular Design and Informatics, and in 2006 was appointed as VP R&D IT. In 2008, he was appointed Global Head Molecular Design & Informatics, Schering-Plough. In 2011 he began serving as CEO and scientific director of The Netherlands eScience Center.

 http://dx.doi.org/10.1016/j.drudis.2013.12.004

disciplines will benefit medicinal chemistry and data-driven drug discovery. Furthermore, the current interest in big data has also served as a reminder to review all our data practices and to evaluate many aspects of current research practices. Although many of the challenges posed by the rapid growth in data are technical, it is important that we pay equal attention to the behavioural changes required in project teams and at the individual researcher level. The role of the medicinal chemist (computational and synthetic) has never been static and continues to evolve [1]. The need to embrace further the role of data scientist or knowledge worker represents the latest stage of this evolution.

The new science

Data have always underpinned hypothesis-driven research, but the scale of its generation is now so great that science has to adapt to ensure it can fully exploit the opportunities it provides [2]. Additionally, the days of individual researchers working in isolated groups and focused only on their own, increasingly narrow expertise are numbered. Breakthroughs are increasingly made at the interface of disciplines by groups of scientists benefitting from the combination of their diverse skills. Modern research is characterised by four developments:

i. The access to, and requirement to, manage large amounts of data.

   Big data, although undoubtedly a hype term, is also a very real phenomenon driven by our increasingly digital world. Scientific data are generated at increasing speed owing to the miniaturisation and parallelisation of experiments, the deployment of remote sensors and the use of laboratory information management systems (LIMS) to integrate experimental tools with the internet and associated databases. In the internet age, data are shared as rapidly as they are generated, facilitating contemporary collaborative science and knowledge sharing. Good management of data is crucial to ensure reproducibility of earlier work, to develop larger populations of similar data to improve statistical analysis and for enabling data-driven research in the future. Medicinal chemistry is equally subject to the demands and opportunities of the big data era. This is driven by parallel synthesis, the generation of increasingly large and complex analytical and biological datasets associated with each new chemical entity (NCE) and the need to integrate publicly available information, including patent literature, into the design process.

ii. The requirement for all researchers to become data scientists. Although not every scientist is a mathematician, it is necessary for all scientists to have some mathematical capability. Modern data-driven approaches will require all researchers to become, in part, data scientists, which means developing and applying data analysis skills in addition to their existing experimental expertise. The role of the medicinal chemist (either from a synthetic or computational background) is to make decisions about which of the infinite possibility of new compounds should be made next (Fig. 1). As the amount and variety of data on which to base these decisions grows, so too must the data analysis skills of the medicinal chemist.

   The modern medicinal chemist should be able to recognise sources of relevant information, prepare raw data, use statistical tools, extract meaningful information, interpret results, recognise potential problems and make visualisations to communicate their findings. These are techniques rarely taught during organic chemistry degrees but must become fundamental medicinal chemistry tools. In fact, it can be questioned whether our organic chemistry graduates are significantly more data savvy, in this digital era, than previous generations, and if course curriculums have kept pace with the rapid changes in science.

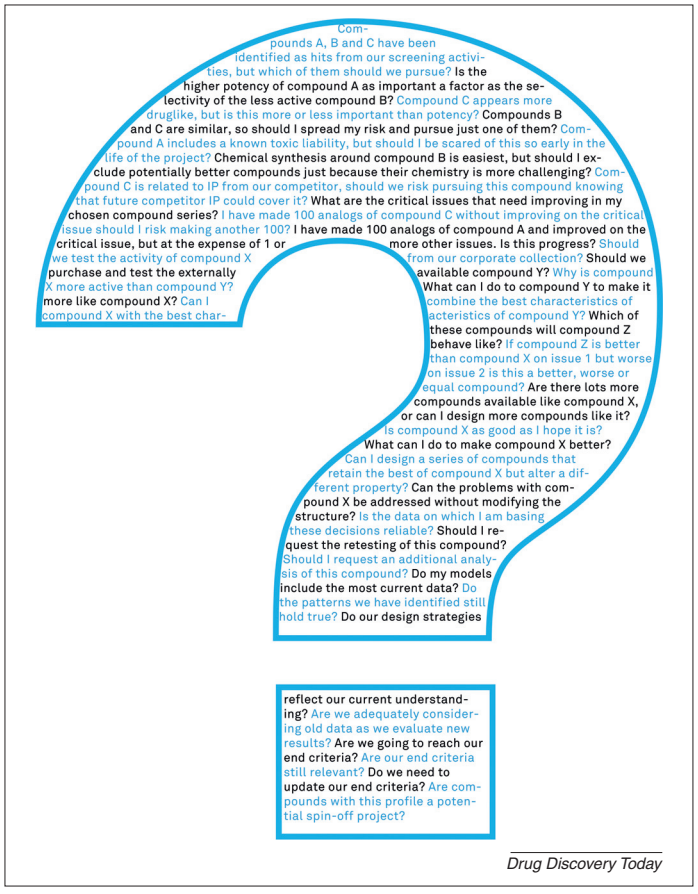

FIGURE 1

A representation of some of the common decisions that are faced by a medicinal chemist during a typical drug discovery project.

Scientists from all disciplines are struggling to manage and share data produced by their own laboratories as well as accessing and integrating data produced by others. A recent report: *supporting the changing research practices of chemists* [3], highlights the difficulties chemists are already facing in managing data and keeping abreast of current information resources with three key findings.

- Chemists need more and better support in data management, storage and sharing.
- Chemists often struggle to keep up to date with the relevant literature.
- Chemists are not very adept at using digital technologies to disseminate their research to a wide audience.

Correcting this situation will require improved education and increased access to data and information management tools. LIMS and electronic laboratory notebooks (ELNs) are helping to manage the data produced by synthetic chemists, but

integration with other data resources and external literature remains problematic.

iii. The increased complexity of research projects.
In relation to medicinal chemistry, complexity is currently illustrated by the need to undertake multiparameter drug discovery and balance numerous activities and characteristics within chemical series [4–7]. This is a direct result of pharma's attempts to reduce attrition rates in development by testing for a greater number of properties earlier in the process. This requires drug designers to manage a larger variety of property data and ensure incorporation into their design strategies. In the long-term this is likely to improve the quality of compounds being produced in research, but only if drug hunters have the skills, tools and mindset to ensure these data can be used to make new and improved decisions.
Understanding the complexity of biological processes to a sufficient degree that we are able to intervene chemically is the goal of rational drug design. A potential shift from the current 'one-target–one-drug' model to a multiple-target approach, predicted by many, will add further to this complexity [8] as networks of interactions grow and selectivity profiles become intricate. Quantifying the pharmacokinetic and pharmacodynamic (PKPD) requirements and outcomes of multitarget drugs will also require more sophisticated information management as well as medicinal chemical approaches capable of directing optimisation on multiple disparate targets and properties.

iv. The need to work effectively in larger teams with colleagues from multiple disciplines and at multiple locations.
Medicinal chemists undertake drug discovery projects as part of collaborative teams including pharmacologists, molecular biologists, informaticians and others. In other industries, in particular engineering, large multifactorial problems can be divided into smaller tasks and divided between disparate research and development teams who work independently [9]. This siloed approach is detrimental to drug discovery, which is dependent on collaboration, reflecting the general trend that discovery is increasingly made at the interface between multiple disciplines. Ensuring data generation from each discipline of the team is synchronised and shared efficiently is crucial, as is enabling the team to make decisions based on data. Furthermore, teams distributed across different sites, perhaps on different continents with differing time zones, are increasingly common, as is the need for close collaboration with partnering organisations in contract research organisations (CROs) and academia.
Sharing explicit data in these collaborations provides challenges (technical and social), but is certainly easier to achieve than sharing of the tacit knowledge that comes from the specific experience and insight of individual researchers and teams [10]. Tackling the need to share this knowledge is also crucial to successful collaborative research and is dependent on good knowledge management practices.

The big data challenge in medicinal chemistry

When discussing big data it is tempting to quantify the issue by including examples of relevant large data resources, for example the number of substances in the Chemical Abstract Services Registry (currently 74 million). However, scale alone is not sufficient to define the concept of big data. In addition to volume, the velocity at which data are generated and the variety, the so-called 3Vs of big data (defined by Doug Laney while an analyst at META group), are of equal importance. Since the first use of these terms to characterise big data, the concept has become routinely supplemented by two additional terms: value and veracity. It is the specific interplay and combination of these five factors (5Vs) that defines big data rather than the size of any specific database.

Volume

There is no single measure to define big data, or how much it comprises. Parallelisation and miniaturisation, combinatorial synthesis, high-content technologies and an increase in the number and diversity of measured parameters have contributed to a dramatic increase in the volume of data available to medicinal chemists. There has also been a rapid growth in the volume of publicly available data relevant to drug designers. However, compared with many other disciplines, for example astronomy, medicinal chemistry generates a relatively small amount of data, even for projects that might include an imaging component. The most common analogy used to quantify big data is: any amount that is too big to be managed by current conventional methods. This is however typically related to the capacity of hardware including processing power, data storage or network capacity. Even for the most intense medicinal chemistry projects, current computing architectures should be capable of managing these requirements. In this respect, volume is perhaps the least important of the 5Vs in the context of medicinal chemistry. A more relevant measure is whether the drug discovery team or organisation is able to extract the relevant information from their rapidly growing data resources. Are the methods we currently rely on scalable and robust? Does the team have sufficient data-science skills to exploit the resources at hand? Above all, are sufficient resources being allocated to data management and analysis?

Velocity

The goal of data-driven or intensive research is to improve decision making, perhaps in terms of speed or quality (preferably both). In the context of medicinal chemistry, decision making has to fit within the framework and timing of the drug design cycle (Fig. 2) and the optimisation of the evaluation phase [11]. As the speed of new data generation increases it becomes harder to ensure all factors are being duly considered in decision making. Ideally experimental data should be disseminated among the discovery team as it is generated or soon after. However, daily data updates put considerable pressure on researchers if we expect them to incorporate this new information into design decisions. New data are generated faster than decisions can be made, because every new data point is a potential reason to alter direction.

We know that drug discovery teams, like all project teams, are susceptible to a number of repeating bad practices that result in poor decision making [12]. These common errors in behaviour will undoubtedly increase as the amount and speed of data generation also increases. Earlier testimonies also tell us that growth of data and increased pressure to make rapid decisions can result in decreased evidence-based actions, with teams reporting that they consciously ignored new data [13]. Managing increasing data

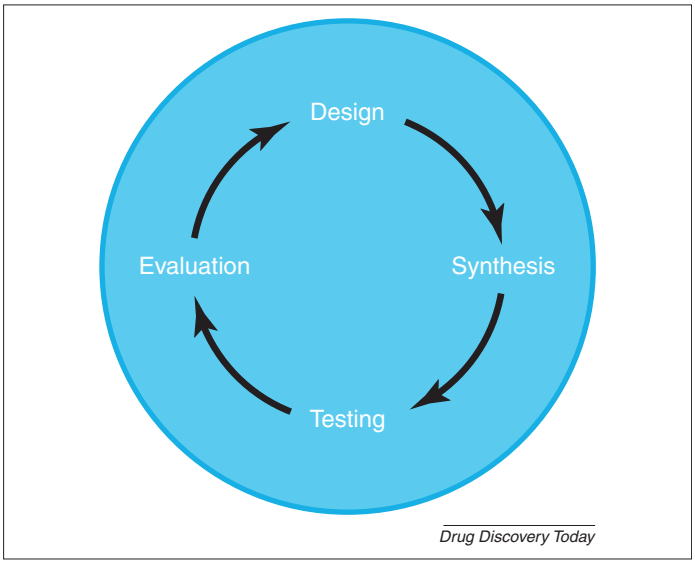

FIGURE 2

The classic depiction of the iterative cycle undertaken during the synthetic phase of drug discovery projects. In future, increased time and focus will be spent on the evaluation of new data and their integration into the design phase.

velocity, like most aspects of data-driven research, has a technical and a behavioural component. Intuitive data repositories and decision support tools are needed to disseminate and analyse data. Data generated in drug discovery projects, especially in early phases, is well understood in terms of structure and range which makes it well suited to pre-canned analysis. These methods, connected directly to the projects data repository, can update automatically as new data is generated and provide a rapid intuitive overview. Deeper analysis, with less-structured data, is also crucial but naturally requires deeper visualisation and analytics expertise.

Universal data repositories and decision support tools also ensure that everybody in the discovery team is looking at the same data and using the same analytics methods. This allows researchers to study new data before joining their colleagues in multidisciplinary team meetings to explore the data further and make comparisons with previous results and data from other sources. Even if the team is able to manage the daily stream of new data, there are likely to be times when even the best data-management strategies are put to the test. For example, when a CRO delivers a large volume of data on a single day (that they might have spent weeks or months generating) or when the key decision makers and data scientists in a project leave and new scientists have to get up to speed. Coping with these additional demands on data velocity requires skilled data scientists, enabled to make decisions and with a deep knowledge of the project.

Another potential pitfall is the generation of data at uneven velocity and the resulting bias on the use of the most readily available data at the expense of potentially more valuable resources. For example, co-crystals of NCEs bound to their target will usually take longer to generate and analyse than primary biochemical data but might provide valuable insight and direction for compound design. There must be capacity to modify chemical plans rapidly as new data become available.

Variety

Chemistry has always benefitted from the common nature of its representations. A chemical formula or structure is universally understood by researchers, regardless of their language or location. The digital age, and specifically the proliferation of data formats, has made this common communication more difficult. A great deal of a cheminformatician's time is spent managing this heterogeneous data and ensuring they can be used in conjunction with the equally heterogeneous array of software and databases employed by various users. The picture grows in complexity when we also consider the disparate nature of the biological data and experimental results that you might wish to associate to a compound. This will increasingly include image data, which have their own management and analysis challenges. Ensuring long-term availability of data in a world of changing formats and standards and unpredictable access to commercial software provides additional headaches to research informatics experts.

External data resources, such as peer-reviewed literature and patent records, of which drug-design organisations have limited control, are also poorly suited to supporting data-driven research. The open access movement is growing and making the final versions of articles available in increasing numbers, but this is often done in pdf formats, which poorly support mining and analysis. As Professor Peter Murray Rust explains in his online article *Data-driven science – a scientists view* [14], '. . .about 2 million chemical compounds are published each year (about half in patents) with insufficient semantics, metadata or hyperstructure. Vast effort is required to create useful data from these. . .'. Integrating this information with in-house tools is also not trivial, if possible at all. The result is an increase in the number of information resources the medicinal chemist must regularly access, each based on their own methodology and with a different user interface. The outcome of this is inevitably information overload. Furthermore, it is a key goal of data-driven research to find the hidden patterns and relationships between data points in disparate resources. As links and interactions are revealed, an ever more complex network develops that needs to be managed and understood.

Veracity

Naturally, data-driven approaches are dependent on the quality of the data that underpin them. This provides specific challenges within pharmaceutical research as a result of the use of surrogate models in place of human and animal testing. The translation of results between each level of reduction is fraught. How well do our biochemical assays reflect activities in cell-based approaches, and from cells to tissues to whole organisms (mice and rats > dogs > primates) and eventually humans? As recently observed, 60% of first-in-class drugs approved by the FDA between 1999 and 2008 resulted from phenotypic (cell based) screening rather than reductionist biochemical assays [15]. This is a discussion beyond the scope of this review, but the predictiveness of our experimental data determines the success of the decisions we base upon it.

Even if we believe in our assays at a conceptual level, there still remain operational barriers to reliable data creation. Generating meaningful data requires strictly adhered to protocols and authorisation steps as well as constant vigilance from users to identify

errors, systematic failures and the introduction of bias. A recent article [16] demonstrates some of the common reasons for the introduction of systematic measurement errors such as precipitation of reagents, variation in sample volume dispensing and plate-reading errors. Finding and correcting these problems requires dedicated effort, but is crucial for confidence in data and consequently the decisions we make based on it. It also appears that statisticians are underused during the earliest research phases where they could play an important part in the initial design of assays to ensure data are generated with significance [17–19]. Statisticians are ubiquitous in clinical research but often absent in preclinical phases where they could play an important part in calibration and standardisation of protocols and the overall design and analysis of data. If we are going to improve the output of preclinical research we must ensure data-driven decisions are based on statistically rigorous data that requires engagement with expert statisticians [17].

Given the challenge in ensuring the quality of internally generated data we might ask how we can have any confidence in externally generated data. It is certainly important to treat external data with a certain level of scepticism. How much reliance you place on external data will depend on the trustworthiness of the source and the amount of information available to describe how it was measured and validated. The relatively recent availability of public databases for chemistry and associated biological activities, such as PubChem [20–22] and ChEMBL [23,24], is of great benefit to the community, but we are now seeing the first evaluations of their quality [25]. Even when we have faith in external data (or even data generated at different points in the lifecycle of a project), combining them with other data sources generated in different laboratories, using slightly different protocols or measured with different apparatus, poses challenges for data normalisation [26].

Value

In multidisciplinary drug discovery projects it is not possible to determine in advance which piece of data will result in the valuable new design approach or, in fact, the cost effective early termination of a doomed project. All data should be given equal weight and consideration. Although it is true that automation, miniaturisation and parallelisation have greatly reduced the cost of generating data in drug discovery, especially in the earliest phases, the rarity and technical difficulty in collecting some biological samples should ensure collected data are treated with appropriate value. In many cases, especially in the case of animal models, failing to extract the greatest value from experiments is simply unethical.

Within drug discovery organisations (and all scientific pursuits) it should be unacceptable to begin complex and expensive drug discovery projects without adequate provision for ensuring that data are generated with statistical relevance, stored in a manner that will ensure they can benefit future studies, are communicated between relevant disciplines and, above all, analysed to ensure the most value can be extracted.

Data repositories and decision support tools

Undertaking new scientific endeavour, without sufficient focus on the IT systems needed to underpin these investments, results in technology islands and should be considered malpractice. Storing heterogeneous data, in specialist or inaccessible formats, with insufficient metadata (including protocol details) and in autonomous databases, is in contradiction to good data stewardship practices. Data and their location must be readily identifiable, searchable and accessible to drug designers, but with sufficient security to protect intellectual property.

From our own personal experience at Organon (later Schering Plough, later MSD), we remember biweekly discovery team meetings that began with pharmacologists handing out piles of printed Excel spreadsheets and curves. This would often be the first time the various chemists would have seen the data. Different pharmacologists in the team would have calculated similar, but crucially different, curves from the same data resulting in slightly different conclusions and inevitable confusion. Only the most recent data would be included and only the data from each pharmacologists own group. Relating new findings to previous results or from other groups [drug metabolism and pharmacokinetics (DMPK) or analytical chemistry for example] relied on the memory of medicinal chemists and their own spreadsheet summaries of earlier data, themselves printed and arranged in dossiers. It was a recipe for the single-parameter decision making and the blind chasing of potency repeated across the industry.

Compare this situation to the change in practice after the implementation of Organon's in-house decision support platform. The Integrated Project View (IPV) [27], comparable to ArQiologist [28] from ArQule, ADAAPT [29] from Amgen, OSIRIS [30] from Actelion and Johnson & Johnson's ABCD system [31], provided a single interface to all the biochemical, chemical, pharmacological, DMPK and analytical chemistry data generated within a project and was updated nightly. It was available to all researchers at their desktop and was the primary interface between researchers and their data. It allowed teams to see aggregated and raw data, undertake specific queries or browse new data or trends over time. Furthermore, it linked directly to standardised data analytics tools providing pre-designed data views common in many projects as well as enabling more in-depth data analysis.

Design chemists could now follow the latest developments in their project at their desktops, perform and repeat complex queries and export data for model building or other purposes. Access to raw data also helped establish the veracity of some data-driven decisions and resulted in an increase in the numbers of compounds retested. Researchers were now all working from identical data with equally rapid access and resulting from the same analysis tools (curve fitting, among others). The phenomenon of data only being available on individual computers or memory sticks stored in desk drawers soon ended with laboratory analysts keen to upload and certify their data as the future use became apparent. Anecdotally, it appeared the quality of some experimental work improved because it was recognised that colleagues were taking more time to analyse data in the knowledge that they would be stored and accessible indefinitely. Data had always been generated following carefully agreed and authorised protocols, but it became clearer to the data users when the protocol had changed (for example when moving to a new plate reader, among others) so that new data could be validated and models updated. Researchers arrived at team meetings having already gained familiarity with the new data and viewed them in the context of previous compounds and related resources. Project meetings

became increasingly data-driven and the printed spreadsheet a thing of the past.

Cyber commons

One consequence of teams becoming dependent on the decision support platform was the need to make it available in team meetings. It was also recognised quickly that access to data alone was not sufficient to support decision making. It was still necessary for experts, particularly those generating data, to give their opinions following the maxim that 'behind every data point is a story'. It was therefore decided to create a project and data visualisation suite to support cross-disciplinary research and data-driven decision making (Fig. 3). The facility, named the 'war room', after Winston Churchill's Cabinet War Rooms, was a crucial step in ensuring project meetings became a data-driven pursuit. The same tools available on the researchers desktop were now available to the team in a shared environment allowing people from each discipline to contribute their ideas and opinions during the design and analysis process.

The concept of the advanced project working and data visualisation suite, sometimes referred to as cyber commons or collaboratoriums, is growing in popularity with examples reported at organisations such as Proctor & Gamble and Monsanto. These facilities are distinct from the 3D visualisation rooms common in pharma organisations and are characterised by their focus on data analysis and project work. It is important that technologists appreciate that, despite the rapid growth in the power of computers, they still require humans to provide imagination and creativity. A cyber common, integrated into the regular research process, can play an important part in enabling data-driven and multidisciplinary research by bringing together research disciplines, but dos not replace the various informal networks that develop between collaborating colleagues over time [10].

Knowledge workers and data scientists

There is also a danger in the increasingly complex world of early drug discovery that scientists become increasingly specialised, focused only on their specific subdiscipline and the networks they maintain therein. As stated earlier, discovery is increasingly made at the interface of disciplines by collaborative researchers from multiple fields. We need to develop more broad-oriented scientists able to bridge disciplines and with a deep understanding of the drug discovery process and the generic scientific skills in data analysis and visualisation to exploit these insights.

People able to assimilate multiple conflicting data sources and identify trends and potential areas to exploit are rare and should be given suitable credit within their organisations. They are undoubtedly in demand in other industries. Digital scientists, able to work at the interface of their own scientific disciplines, data resources and advanced computing, are currently highly sought after. They should also be key members of drug discovery projects, with deep knowledge of medicinal chemistry, undertaking the most difficult analysis and building the key models for the project. This represents just the latest evolution of the informatician in drug discovery. Originally, synthetic chemists began to recognise the value of computational tools in managing and exploring chemical data and began the first computer-aided drug design groups. A second generation followed, with more-formal training in programming, able to develop the algorithms and software that still underpin many of the mostly widely used tools today. More recently, owing to the growing maturity and professionalisation of tools, the need to program has become less crucial and computational medicinal chemists are characterised by the ability to hop between different

Drug Discovery Today

FIGURE 3

The advanced visualisation and multidisciplinary project meeting room at Organon Biosciences, Oss, The Netherlands, in 2007.

tools and approaches as required by their projects. This continuing evolution will now be towards data specialists able to extract the most value from disparate data sources with perhaps more focus on understanding the implications of experimental data and less on *ab initio* prediction.

In addition to the need for data-specialists, there is a need for all researchers to become data scientists or knowledge workers [32], able to incorporate data analysis into their entire decision making. The researchers making the design decisions within projects should be the most knowledgeable of the data related to it. In the past this might have been manageable via spreadsheets and a good memory (although we doubt this was ever truly satisfactory), but the changing scientific landscape requires medicinal chemists to embrace data-driven methods to improve output in the future. Although there clearly remain technical and financial obstacles to expanding data-driven medicinal chemistry practices, perhaps the biggest challenge will be developing the necessary behaviour and actions within drug-design teams.

Data-driven decision making

With the exception of the talent, experience and imagination of the researchers themselves, a project's data resource is its greatest asset and should be treated with the respect that it deserves. That means building confidence in the quality of that data by constantly searching for discrepancies and errors in the experimental method. Unusual activity should be queried and retested if necessary, and models constantly updated and challenged by new findings with more emphasis placed on identifying trends and less on chasing outliers.

Above all we must require and stimulate project teams to make decisions based on data and to learn from previous experience [33]. This should be the case even if those data are at odds with conventional wisdom. For this to be possible the project team must be able to identify all the data and issues that are pertinent to the question at hand and not become overly focused on the most recent or easy to interpret data. At the same time, teams must avoid searching for data that support their predefined hypothesis. It is too easy to turn into a dogma an observation made early in a project, or from historical medicinal chemistry practice, that is not supported or is actually contradicted by the data. Perhaps the biggest challenge is for project teams to balance the need to change direction based on new trends in data, without every new data point becoming the basis for a new strategy. Projects have to learn to look for trends and patterns in data rather than reacting to each individual result as it becomes available. New data must be analysed in the context of earlier results and design decisions undertaken logically and without presumption.

Synthetic medicinal chemists are, for legitimate operational reasons, under a great deal of pressure to deliver quickly and cheaply on their projects. The result of this pressure is that they become trapped between two worlds. They wish to make their design decisions with a strong molecular basis and from compelling data, but too often feel the need to generate large numbers of compounds quickly at the expense of rationale [34]. Whereas in reality a combination of careful rational design and rapid chemical exploration is probably desirable, there is a danger that too often the balance is shifted too far towards ease of synthesis [35], which is unsurprising given the typical synthetic background of most medicinal chemists and their management. It will be interesting to see if the profile of senior medicinal chemists does change to reflect the growing biological chemistry and informatics components of the discipline.

When asked how they make their design decisions most medicinal chemists will describe chemical intuition as a key factor [36]. Although we recognise the importance of imagination and creativity in the chemical design process, that intuition should be guided by the available data and not conducted in isolation. Other contributing factors to the design process include the previous experience of the medicinal chemist and organisation (what worked before) [34], knowledge of the biological target (from a structure-based or ligand-based basis) and their personal toolboxes (their preferred reactions) [37,38]. Each of these three considerations can be enhanced within an organisation by good knowledge management practices. The more experience chemists have to draw from and the better their knowledge of their targets the less likely they will overly base decisions on synthetic ease. The use of wikis [35,39] and social media type approaches [40], as well as numerous other IT approaches to manage drug discovery and particularly synthesis planning (reagent selection, library design, among others) [41–47], is growing rapidly. We are even seeing the first descriptions of mobile apps to support medicinal chemistry and drug-discovery [48].

Another important development in recent years has been the use of a variety of metrics to quantify the progression of optimisation projects [33]. Ligand efficiency methods [49] and related approaches such as size-independent ligand efficiency [50], lipophilic ligand efficiency [51] and enthalpic efficiency [52] are all important drivers to help ensure drug-design decisions follow strategies that balance physicochemical parameters with efficacy.

Retaining knowledge

The process of drug discovery has undoubtedly become more complex in recent years as a result of an increase in the diversity and specialisation of technologies embedded within the process. An unfortunate consequence of the increased dependence on process-driven research has been an unspoken decrease in the value of individual researchers. This is perhaps most clearly demonstrated during the continued rounds of merger and reorganisation that persist in the pharmaceutical industry. The rationalisation that takes place when merging large research organisations tends to be dominated by the desire to optimise previous technology investments and maintain the newest (and typically largest) machinery. One molecular biologist or medicinal chemist is considered to be of similar value to any other without consideration for their knowledge or experience. Scientific breakthroughs tend to result from a combination of human knowledge and institutional knowledge contained within underlying processes and resources [10]. It cannot be a surprise that reducing the human knowledge component from the drug discovery process, in favour of ever larger technical capacity or islands of cheap outsourcing capacity, will have a negative effect [53]. The constant mergers throughout the industry might improve the pipelines and balance sheets of companies in the short term but can undermine capacity for drug discovery [54]. The recent years of corporate mergers has probably resulted in the loss of huge amounts of practical knowledge when data-stewardship practices have been insufficient.

Even during more stable times, experienced chemists retire or take on other roles. Ensuring their tacit knowledge is retained and shared throughout an organisation is challenging but can be achieved if their work has been well documented over time. Mining ELNs and corporate databases provides a huge opportunity to identify trends in the actions of medicinal chemists and increase our understanding of the decisions they make. We know that chemists are subjective in their decision making [33,36,41,55], which suggests that they could benefit from more-objective directions derived by knowledge-based exploration of earlier data.

Lipinski's seminal paper [56], profiling 2245 molecules to identify the shared properties of orally available drugs, represents an important step in the shift towards data-driven drug design [57]. Various datasets have since been analysed in various ways to define properties determining drug-likeness [51,58,59] and lead-likeness [58,60], or to identify compounds with a particular character such as capacity to cross the blood–brain barrier [61,62] or with an increased potential to target a particular protein class: kinase-likeness [63–66] or G-protein-coupled receptor (GPCR)-likeness [66–68]. Identifying bioisosters from previous projects [69–71], frequent hitters [72,73] that pollute HTS results or potential toxicophores [74–76] has been a key strategy of academic and commercial cheminformatics groups over the past 15 years. The more data available to base these analyses upon the better their predictive power. For pharmaceutical companies, the potential to combine public data with their in-house resources has increased the breadth of chemical series analysed which increases quality and future predictivity of models. Recent developments in the availability of open data are fuelling the big data revolution and ensuring that knowledge-based methods continue to grow in importance in medicinal chemistry.

Open data enabling data-driven research

In addition to the large amounts of data being generated internally within pharmaceutical companies there is also a rapid growth in the extraction, curation and dissemination of publicly available data of value to drug designers [77]. Although some public data are submitted to well-maintained sustainable databases, the majority is buried in unstructured publications, typically made available in pdf format and extremely difficult to incorporate into other analyses. For example, small molecule crystallographers submit their coordinates and metadata to the Cambridge Structural Database (CSD) [78] and structural biologists submit their X-ray and NMR structures to the Protein Data Bank (PDB) [79,80], but publication of new chemical series in medicinal chemistry journals or patents does not require similar submission. The Chemical Abstract Service (http://www.cas.org/) has the goal to 'find, collect and organise all publicly disclosed chemical substances' but does not record related biological data. As a consequence, large initiatives such as PubChem and ChEMBL are needed to curate the chemical literature, build databases and extract structures and related metadata from publications [81]. However, neither PubChem nor ChEMBL aims to curate all chemical and related information and therefore our access to data is incomplete. It is also the case that extracting data from publications is a less desirable approach than the submission of data, in a structured manner, by authors at the time of publication or before.

Given the importance of data and intellectual property to the competitive advantage of pharma companies, we do not foresee the total move from closed to open innovation [82] but do believe pharma will find benefits from tapping into the 'open' world and utilising tools, data and approaches within their own processes. Examples already exist for Novartis [83], Pfizer [84] and AstraZeneca [85] developing applications that combine public data, patents and proprietary data to provide a single view on the available chemical landscape.

Working with public data raises many technical challenges related to their integration with a company's own proprietary data. Public data will have invariably been measured on different machines with differences in reagents and protocols so how can we trust it is a complete and correct record? Integrating external and internal data therefore requires us to develop improved methods of data normalisation and comparison [86]. This big data challenge will have the side benefit of improving how we manage our internal data – often measured on different machines and with modified protocols over the length of a long drug discovery project.

Data reduction, visualisation and analytics

A consequence of the increased complexity, variety and volume of the datasets present in research has been a new concentration on the development of visualisation and analytics approaches [87]. There is a requirement to find ways to interact better with data by reducing their complexity and presenting them in a readily interpretable fashion. Data reduction and comparison approaches based on statistical approaches such as principal component analysis, linear regression, *K*-means clustering, Bayesian methods, hierarchical clustering and cross-validation underpin the most common data-analytic approaches such as predictive modelling (supervised learning), cluster analysis, data mining (unsupervised methods) and decision trees. Fortunately, these are also the methods that have underpinned model building and analysis in computational chemistry since its inception. Most pharmaceutical companies therefore already have expert teams of data scientists able to develop and apply these approaches. The challenge remains however to use novel and intuitive approaches to visualise results [86,88,89] and ensure visualisation and analytics is embedded in the process.

As the number of NCEs, and associated data, increases in each project it has become necessary for computational chemists to move away from studying compounds individually and instead study trends and patterns in properties. This is perhaps a return to an earlier time when QSAR dominated computer-assisted drug design. A commonly cited failing of computer-assisted drug design has been an inability of computational chemistry to keep up with the speed of modern synthesis and testing. Visual comparison of small numbers of compounds or co-crystals will still be important but we can expect the computational chemist to spend an increasing amount of time studying graphs, infographics and other forms of data representation as they try to make sense of large datasets at the same pace that synthesis and testing is undertaken.

Concluding remarks

Increasing the capacity of medicinal chemistry to undertake data-driven research has the potential to improve decision making in drug discovery and ensure the most benefit can be derived from the data produced internally and available externally. The rapid

increase in available data, the so-called big data era, makes harnessing these resources and optimising our research processes a prerequisite for future success. Data repositories are crucial to manage distinct data resources with a single user-friendly interface common to all project team members. Technicians should be rewarded for the rapid and consistent addition of their data to such a resource with senior management taking responsibility for the quality of these data. Analysis tools should be streamlined with the repository and adequate training provided to all users. In addition to providing data and analysis tools to individual researchers it is also crucial to support project work by making the same tools and data available during team meetings via project rooms tailored to data-intensive decision making at the group level.

It is however a mistake to consider harnessing big data to be a purely technical challenge. We need to improve the information literacy of medicinal chemists at a faster rate than we are currently doing so, and also ensure project teams are empowered and capable of data-driven decision making. All researchers, including medicinal chemists, will have to become comfortable working in data-rich environments. At the same time, there will be a growing need for data specialists, probably drawn from existing informatics departments, to build complex models and make connections between disparate data sources. The move towards data-driven drug design is an evolution of our existing approaches, rather than a revolution, but is required to ensure the most is gained from current investments in R&D.

References

1 Lombardino, J.G. and Lowe, J.A. (2004) The role of the medicinal chemist in drug discovery – then and now. *Nat. Rev. Drug Discov.* 3, 853–862
2 Mattmann, C.A. (2013) Computing: a vision for data science. *Nature* 493, 473–475
3 Long, M.P. and Schonfeld, R.C. (2013) *Supporting the Changing Research Practices of Chemists*. Available at: http://www.sr.ithaka.org/research-publications/supporting-changing-research-practices-chemists
4 Lusher, S.J. *et al.* (2011) A molecular informatics view on best practice in multi-parameter compound optimization. *Drug Discov. Today* 16, 555–568
5 Nicolaou, C.A. and Brown, N. (2013) Multi-objective optimization methods in drug design. *Drug Discov. Today Technol.* 10, 427–435
6 Nicolaou, C.A. *et al.* (2007) Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Dev.* 10, 316–324
7 Segall, M.D. (2012) Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* 18, 1292–1310
8 Medina-Franco, J.L. *et al.* (2013) Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov. Today* 18, 495–501
9 Slater, T. *et al.* (2008) Beyond data integration. *Drug Discov. Today* 13, 584–589
10 Zoller, F.A. and Boutellier, R. (2013) Design principles for innovative workspaces to increase efficiency in pharmaceutical R&D: lessons learned from the Novartis campus. *Drug Discov. Today* 18, 318–322
11 Plowright, A.T. *et al.* (2012) Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discov. Today* 17, 56–62
12 Chadwick, A.T. and Segall, M.D. (2010) Overcoming psychological barriers to good discovery decisions. *Drug Discov. Today* 15, 561–569
13 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
14 Murray-Rust, P. (2007) *Data-driven Science: A Scientist's View. NSF/JISC 2007 Digital Repositories Workshop*. Available at: http://www.sis.pitt.edu/~repwkshop/papers/murray.pdf
15 Swinney, D.C. and Anthony, J. (2011) How were new medicines discovered? *Nat. Rev. Drug Discov.* 10, 507–519
16 Hampton, D. *et al.* (2013) The hidden quality gap in discovery. *Drug Discov. Today* 18, 506–509
17 Peers, I.S. *et al.* (2012) In search of preclinical robustness. *Nat. Rev. Drug Discov.* 11, 733–734
18 Hothorn, L.A. (2005) Biostatistics in nonclinical and preclinical drug development. *Biometrical J.* 47, 282–285
19 Lendrem, D. (2002) Statistical support to non-clinical. *Pharm. Stat.* 1, 71–73
20 Li, Q. *et al.* (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today* 15, 1052–1057
21 Wang, Y. *et al.* (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633
22 Bolton, E.E. *et al.* (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* 4, 217–241
23 Gaulton, A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107
24 Wassermann, A.M. and Bajorath, J. (2011) BindingDB and ChEMBL: online compound databases for drug discovery. *Expert Opin. Drug Discov.* 6, 1–5
25 Williams, A.J. *et al.* (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discov. Today* 17, 685–701
26 Fay, N. (2006) The role of the informatics framework in early lead discovery. *Drug Discov. Today* 11, 1075–1084
27 Baede, E.J. *et al.* (2012) Integrated project views: decision support platform for drug discovery project teams. *J. Chem. Inf. Model.* 52, 1438–1449
28 Rojnuckarin, A. *et al.* (2005) ArQiologist: an integrated decision support tool for lead optimization. *J. Chem. Inf. Model.* 45, 2–9
29 Cho, S.J. *et al.* (2006) ADAAPT: Amgen's data access, analysis, and prediction tools. *J. Comput. Aided Mol. Des.* 20, 249–261
30 Sander, T. *et al.* (2009) OSIRIS, an entirely in-house developed drug discovery informatics system. *J. Chem. Inf. Model.* 49, 232–246
31 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
32 Loughney, D. *et al.* (2011) To measure is to know: an approach to CADD performance metrics. *Drug Discov. Today* 16, 548–554
33 Johnstone, C. (2012) Medicinal chemistry matters – a call for discipline in our discipline. *Drug Discov. Today* 17, 538–543
34 Hann, M.M. and Keserü, G.M. (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug Discov.* 11, 355–365
35 Robb, G.R. *et al.* (2013) A chemistry wiki to facilitate and enhance compound design in drug discovery. *Drug Discov. Today* 18, 141–147
36 Kutchukian, P.S. *et al.* (2012) Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. *PLoS ONE* 7, e48476
37 Roughley, S.D. and Jordan, A.M. (2011) The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* 54, 3451–3479
38 Jordan, A.M. and Roughley, S.D. (2009) Drug discovery chemistry: a primer for the non-specialist. *Drug Discov. Today* 14, 731–744
39 Mayweg, A. *et al.* (2011) ROCK: the Roche medicinal chemistry knowledge application – design, use and impact. *Drug Discov. Today* 16, 691–696
40 Hohman, M. *et al.* (2009) Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov. Today* 14, 261–270
41 Cheshire, D.R. (2011) How well do medicinal chemists learn from experience? *Drug Discov. Today* 16, 817–821
42 Williams, A.J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* 13, 502–506
43 Oprea, T.I. *et al.* (2000) Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graph. Model.* 18, 512–524 541
44 Lee, M.-L. *et al.* (2012) DEGAS: sharing and tracking target compound ideas with external collaborators. *J. Chem. Inf. Model.* 52, 278–284
45 Brodney, M.D. *et al.* (2009) Project-focused activity and knowledge tracker: a unified data analysis, collaboration, and workflow tool for medicinal chemistry project teams. *J. Chem. Inf. Model.* 49, 2639–2649
46 Yasri, A. *et al.* (2004) REALISIS: a medicinal chemistry-oriented reagent selection, library design, and profiling platform. *J. Chem. Inf. Comput. Sci.* 44, 2199–2206
47 Boström, J. *et al.* (2011) Exploiting personalized information for reagent selection in drug design. *Drug Discov. Today* 16, 181–187
48 Williams, A.J. *et al.* (2011) Mobile apps for chemistry in the world of drug discovery. *Drug Discov. Today* 16, 928–939
49 Hopkins, A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431

50 Nissink, J.W.M. (2009) Simple size-independent measure of ligand efficiency. *J. Chem. Inf. Model.* 49, 1617–1622
51 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
52 Ladbury, J.E. *et al.* (2010) Adding calorimetric data to decision making in lead discovery: a hot tip. *Nat. Rev. Drug Discov.* 9, 23–27
53 Branthwaite, P. (2010) The shrinking of the knowledge base – what is the impact of this on the speed and security of drug development? *Drug Discov. Today* 15, 1–2
54 Knutsen, L.J.S. (2011) Drug discovery management, small is still beautiful: why a number of companies get it wrong. *Drug Discov. Today* 16, 476–484
55 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* 47, 4891–4896
56 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26
57 Ghose, A.K. *et al.* (2006) Knowledge-based chemoinformatic approaches to drug discovery. *Drug Discov. Today* 11, 1107–1114
58 Lipinski, C. (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1, 337–341
59 Lipinski, C.A. (2001) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249
60 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
61 Pardridge, W.M. (2007) Blood–brain barrier delivery. *Drug Discov. Today* 12, 54–61
62 van de Waterbeemd, H. *et al.* (1998) Estimation of blood–brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. *J. Drug Target.* 6, 151–165
63 Aronov, A.M. *et al.* (2008) Kinase-likeness and kinase-privileged fragments: toward virtual polypharmacology. *J. Med. Chem.* 51, 1214–1222
64 Ingo, M. and Enyedy, I.J. (2004) Virtual screening for kinase targets. *Curr. Med. Chem.* 11, 693–707
65 Weinmann, H. and Metternich, R. (2005) Editorial: drug discovery process for kinase inhibitors. *ChemBioChem* 6, 455–459
66 Lowrie, J.F. *et al.* (2004) The different strategies for designing GPCR and kinase targeted libraries. *Comb. Chem. High Throughput Screen.* 7, 495–510
67 Balakin, K.V. *et al.* (2002) Property-based design of GPCR-targeted library. *J. Chem. Inf. Comput. Sci.* 42, 1332–1342
68 Sprous, D.G. *et al.* (2010) QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. *Curr. Top. Med. Chem.* 10, 619–637
69 Krier, M. and Hutter, M.C. (2009) Bioisosteric similarity of molecules based on structural alignment and observed chemical replacements in drugs. *J. Chem. Inf. Model.* 49, 1280–1297
70 Wagener, M. and Lommerse, J.P.M. (2006) The quest for bioisosteric replacements. *J. Chem. Inf. Model.* 46, 677–685
71 Devereux, M. and Popelier, P.L. (2010) In silico techniques for the identification of bioisosteric replacements for drug design. *Curr. Top. Med. Chem.* 10, 657–668
72 Roche, O. *et al.* (2002) Development of a virtual screening method for identification of frequent hitters in compound libraries. *J. Med. Chem.* 45, 137–142
73 Bleicher, K.H. *et al.* (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378
74 Kazius, J. *et al.* (2005) Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* 48, 312–320
75 Williams, D.P. and Naisbitt, D.J. (2002) Toxicophores: groups and metabolic routes associated with increased safety risk. *Curr. Opin. Drug Discov. Dev.* 5, 104
76 Price, D.A. *et al.* (2009) Physicochemical drug properties associated with in vivo toxicological outcomes: a review. *Expert Opin. Drug Metab. Toxicol.* 5, 921–931
77 Gaulton, A. and Overington, J.P. (2010) Role of open chemical data in aiding drug discovery and design. *Future Med. Chem.* 2, 903–907
78 Motherwell, S. (2004) Cheminformatics and crystallography. The Cambridge Structural Database. *Database* 1, 129–174
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 Rose, P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.* 41, D475–D482
81 Nicola, G. *et al.* (2012) Public domain databases for medicinal chemistry. *J. Med. Chem.* 55, 6987–7002
82 Judd, D.B. (2013) Open innovation in drug discovery research comes of age. *Drug Discov. Today* 18, 315–317
83 Zhou, Y. *et al.* (2007) Large-scale annotation of small-molecule libraries using public databases. *J. Chem. Inf. Model.* 47, 1386–1394
84 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815
85 Muresan, S. *et al.* (2011) Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data. *Drug Discov. Today* 16, 1019–1030
86 Howe, T.J. *et al.* (2007) Data reduction and representation in drug discovery. *Drug Discov. Today* 12, 45–53
87 Mente, S. and Kuhn, M. (2012) The use of the R language for medicinal chemistry applications. *Curr. Top. Med. Chem.* 12, 1957–1964
88 Ritchie, T.J. *et al.* (2011) The graphical representation of ADME-related molecule properties for medicinal chemists. *Drug Discov. Today* 16, 65–72
89 Ahlberg, C. (1999) Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug Discov. Today* 4, 370–376