

Advances in Data Reporting, Visualisation and Decision-making in the Drug Discovery Process

a report by
Chad L Stoner

Principal Scientist, Pfizer Global Research & Development

Drug discovery scientists today are very different from their predecessors. Today's scientists assess an increasing amount of information from multiple disciplines for a larger number of targets and a broader chemical space. Multiparameter optimisation (MPO) is becoming a standard approach across the pharmaceutical industry as the pressure of rising costs forces dramatic improvements in fundamental drug discovery processes.¹⁻³ Some of these important scientific advances are being driven by the way in which data are reported and visualised and decisions are made. Informatics colleagues and drug discovery scientists will play significant roles in helping to integrate safety, absorption, distribution, metabolism and elimination (ADME) and potency data into the design, testing and subsequent optimisation loops that will lead to tomorrow's new medicines. This review will describe some of the advances in data collection using electronic lab notebooks (ELNs), integration via laboratory information management systems (LIMS) and visualisations and how these new tools can help to advance decision-making.

In less than 20 years there has been a rapid and significant evolution of what it means to be a drug discovery scientist and the disciplines that are engaged. Drug discovery was historically dominated by two disciplines: drug discovery scientists were either medicinal chemists or biologists, working in isolation, and commonly tracking all their data on paper or using spreadsheets. There were relatively few compounds, so manual tracking was feasible, though cumbersome. Throughout the late 1980s and early 1990s, most pharmaceutical companies began to expand drug discovery, incorporating drug metabolism scientists and pharmaceutical scientists into drug discovery teams. In the early 1990s, significant advances in automation and miniaturisation drove gains in experimental throughput and productivity. In addition, the number of properties assessed in the characterisation of new chemical entities (NCEs) in drug discovery has significantly expanded.⁴ The most recent

discipline to move from the development realm into drug discovery is toxicology. High-profile withdrawals such as Vioxx®, Baycol® and Rezulin® have highlighted the need to screen for, and build safety into, NCEs early in their development. Consequently, most research organisations have responded by routinely using *in silico* and *in vitro* filters for safety.^{5,6}

Technological advances in analytical chemistry have also enabled substantial improvements in metabolic and toxicology (safety) screening. Of particular importance are the advances in mass spectrometry (MS). The developments in automated sample preparation, data acquisition, automated data processing and review

Since the 2003 completion of the Human Genome Project, it has been estimated that there are some 5,000 human drug-able targets.

procedures have made ADME and safety profiling for large numbers of compounds feasible. A relatively new development is the introduction of safety screens in drug discovery; the potential impact and cost savings it brings merits serious consideration.^{7,8} The addition of these data types, along with advances in chemoinformatics, has also allowed for the addition of computational chemists/biologists and informatics scientists to significantly enhance drug discovery teams.

While the generation of large amounts of data has been ongoing, the corporate databases used to house this information have not been easy to access, and mining out meaningful knowledge had been virtually impossible. Recent advances in data warehousing are allowing for rapid, simple exploration of drug discovery data. These warehouses allow for integration of data from various LIMS, networks and data stores.⁹

Another recent advance that has taken place across the industry is the widespread adoption of ELNs, which now exist for chemistry, biology and various types of analytical data and allow effective data collection and storage. ELNs ensure a level of consistency in terms of reporting critical information that can then be passed on to the corporate data warehouses while the experiments are conducted.¹⁰ The ability to have instant access to ongoing experiments allows discovery teams to rapidly plan future experiments and incorporate new findings immediately.



Chad Stoner is a Principal Scientist at Pfizer's La Jolla California Laboratories, working in the Department of Pharmacokinetics, Dynamics and Drug Metabolism. He is also Co-Chair of Pfizer's Global Computational Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) working group, which serves as a bridge across disciplines and sites to increase the speed and efficiency of ADMET decision-making by improving the design, selection and screening of compounds. Mr Stoner has been in the pharmaceutical industry for 14 years, working for Parke-Davis, Warner-Lambert and Pfizer Global Research and Development. He is an advocate of informatics in the laboratory, working to better integrate data generation, visualisation and design, and his current research interests revolve around better ways of visualising and exploring data, including automated decision-making. He has been a champion of the use of *in silico* ADMET properties to optimise drug discovery, and linking these learnings with design principles to improve ADMET properties.

Corporate data warehouses containing vast quantities of information are now explorable by drug discovery teams commonly looking to optimise multiple properties simultaneously. Since the 2004 report on ADME optimisation at Pfizer,¹¹ we have added more data types, developed global *in silico* models for the various assays, further increased automation to decrease scientist processing time and introduced a better Oracle infrastructure to support data exploration. While Spotfire® continues to be heavily used across the industry to explore drug discovery data, a number of new approaches have been introduced. For example, Pfizer and Johnson & Johnson have both recently reported (Chemoinformatics Conference 2006) on corporate-wide research tools that enable rapid data exploration and decision-making in realtime as the data are being collected in their global data warehouses. A remarkable attribute of these tools is the flexibility they accord researchers who use the data: they enable an in-depth analysis of a few compounds, or broad analysis of chemical series and libraries.

The importance of allowing scientists flexibility in terms of how to visualise and analyse data to allow for new designs and further testing is of utmost importance. The following example will attempt to highlight different ways in which the same type of information can be visualised. In this example, four different tools/approaches are shown that can be applied to a compound or series with a metabolic liability. In approach A, a 3D docking approach is used to determine how the compound is entering the P450 moiety. Designs could be explored that prevent the NCE from entering the P450 active site through interaction with the protein side chains. Approach B demonstrates how an NCE lines up with the P450 heme as a means of identifying the specific site of metabolism. Approach C takes a generic look at the protons that are most easily extracted during metabolism, where more 'red' equals more reactive. This visualisation technique could be used to identify sites for chemical modification to improve metabolic stability. Approach D, Metasite, uses a 3D map of interaction energies between the P450 protein and various chemical probes, and then compares those with the 3D structure of the discovery compound, again allowing for visualisation of labile sites on the NCE. All of these approaches allow for answering the same question; however, depending on the chemotypes and the team's preference, one approach may be more successful than another.

Another approach to visualisation that is gaining popularity is the radar plot, which is also called 'Spider' plot and is a visualisation technique that allows drug discovery scientists to examine many properties of a single compound or multiple compounds in a single visualisation. It allows normalisation of scales and definition of an optimal zone for each property (shaded turquoise in this example). This allows the viewer to see at a glance the properties not in the optimal turquoise zone. In this example (see Figure 1), each line represents a compound from a different chemotype. The discovery team can quickly decide the series that will require the least amount of modification to make progress. The properties in this example are:

- potency on the top is reported in nanomolar;
- human liver microsomal half-life (HLM) is represented in minutes;
- rat liver microsomal half-life (RLM) is also represented in minutes;
- dofetilide (Dof) interaction for human ether-à-go-go-related gene (hERG) cardiovascular liability is represented in nanomolar;
- P-glycoprotein interaction (cMDR) is predicted by an *in silico* model of efflux ranging from 1 (no efflux) to 10 (significant efflux); and

Figure 1: Radar Plot for a Typical Drug Discovery Programme

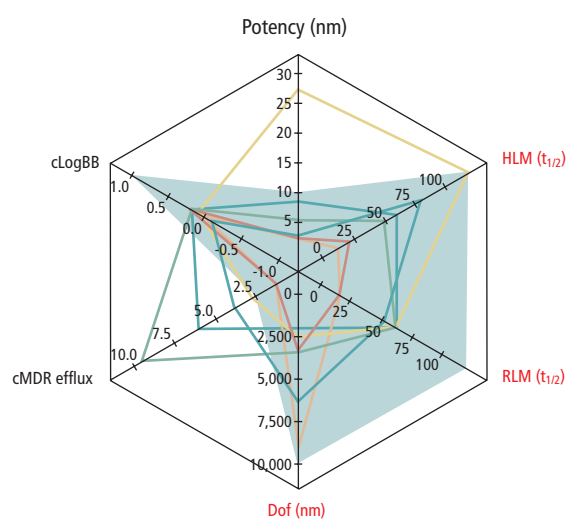
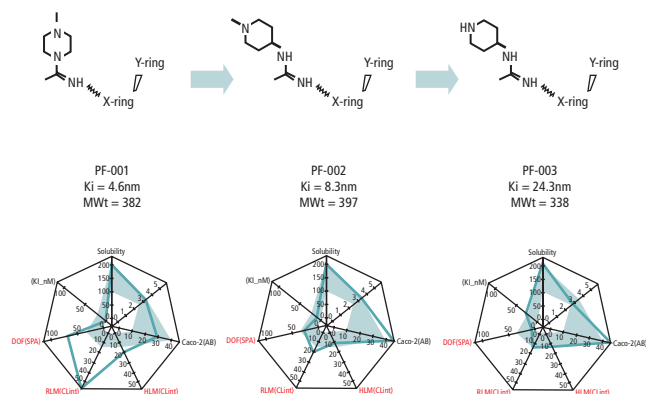


Figure 2: Application of Structure–Activity Relationship Using Radar Plots



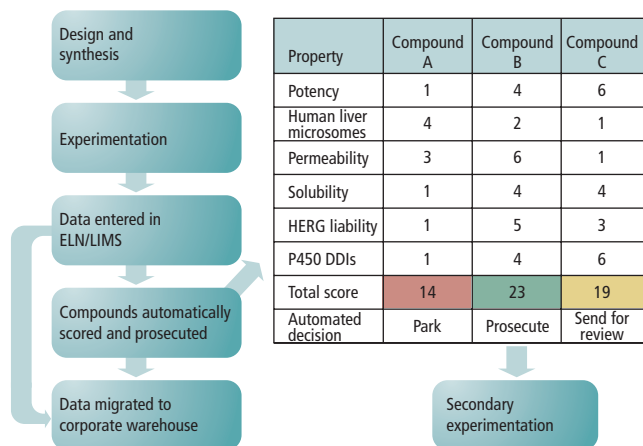
- brain penetration is predicted by an *in silico* model (cLogBB) ranging from 1 (no central nervous system (CNS) distribution) to +1 (for CNS distribution).

A discovery team can quickly and easily identify which liability is facing each of the compounds/series and prioritise which to advance, which to work on further and which to drop. This type of view can also be linked to the underlying structures and experimental data represented by each line to allow for further analysis. These radar plots are flexible and properties can be added or dropped as needed, and the optimal turquoise target zones can also be changed as a programme progresses from lead development into candidate seeking.

An application of radar plots can be seen in Figure 2. In this example, each compound is represented by a single blue line. A switch from piperazine to 4-aminopiperidine shows improved Dof interaction and RLM stability, and further switching from n-methyl to n-H further decreases the Dof interaction.

In addition to automatic data collection, decision-making in terms of compound progression is also being automated to accelerate screening funnels. Pre-defining what makes a 'good' compound for each target allows a discovery team to have an ideal profile and

Figure 3: Generic Screening Funnel



Incorporating automatic compound prosecution based on six early *in vitro* screening assays. Scores range from 1 = worst assessment to 6 = best assessment. ELN = electronic lab notebook; LIMS = laboratory information management systems; DDIs = drug–drug interactions.

prevent information overload. Decreasing decision-cycle times (and overall discovery-cycle times) has been identified as a critical need to increase productivity of most discovery organisations.⁵

Figure 3 shows a generic screening funnel and how a number of disparate properties and assays can be used to rank compounds in terms of a drug-like desirability function. In this example, three compounds are designed, synthesised and sent to the various drug discovery labs for evaluation. As the experiments are completed, data are automatically migrated from the LIMS or ELN to the corporate warehouse, where the data will be automatically scored. For the sake of simplicity, only six experimental variables were considered in this example. The discovery teams can pre-define the importance of each property towards the development of a successful candidate, and even have the most promising compounds automatically sent on to the next round of experimentation.

The concept of automating compound progression has become more and more prevalent in big pharma and has, in fact, been commercialised by two companies. The first tool, ADMENSA™ by Inpharmatica, not only allows discovery teams to track chemistry space visually, but also allows for defining optimal ADME space. Presumably, one could also explore selectivity space, safety space, etc. to help in the selection of the best chemical matter.¹² A significant advantage of this approach is that it allows the colleagues at the bench to define what is important to the project, allowing for flexibility within projects and across therapeutic areas. The automation also removes the need for human intervention and prevents personal bias towards one compound over another.

A second tool of note, Pipeline-Pilot (PLP) from Accelerlys, allows for automatic data queries and decision-making. PLP queries are gaining significant exposure across the industry as teams are becoming more advanced in the utilisation of this application.

Future data visualisation and automated decision-making tools will not only identify which structural features contribute to potency, selectivity and stability, but will also suggest isosteric replacements to rapidly change the chemistry to match the desired product profile.

Conclusion

While there have been significant advances in computing power, automation and overall discovery screening capacity over the last 20 years, substantial improvements still need to be made. Drug discovery scientists must adopt new and innovative ways of doing their jobs in order to increase productivity, decrease costs and, ultimately, improve human health. Since the 2003 completion of the Human Genome Project it has been estimated that there are some 5,000 human drugable targets.¹³ To enable the prosecution of these targets, drug discovery scientists must be able to evaluate their data in new, more flexible ways. Scientists will need to integrate 2D and 3D data along with new sources of information (e.g. biomarker data). The practice of tracking the structure–activity relationship (SAR) on paper or in flat Excel files is history. The drug discovery scientist of the future will spend a substantially larger amount of time analysing data and designing the appropriate experiments and less time in ‘wet’ experimentation and manual programme tracking. Continued advances in computational biology, ADME and toxicology will increase the amount of data that teams must consider to find optimal chemical matter, and to refine that matter into tomorrow’s medicines.¹⁴ To take advantage of these gains will require further advances in data analysis, reporting and databasing. Reports will take the form of living documentation that is dynamically updated. These interactive reports will allow scientists to review raw data and to rapidly advance compounds for further testing. Discovery teams will need to actively define and refine screening cascades to take full advantage of automation and further reduce the cycle times and associated costs. ■

Acknowledgements

The author would like to thank all the Pfizer colleagues who contributed experimental and computational results to this work. In addition, gratitude is expressed to all the informatics colleagues who provided the support and maintenance for all these applications.

A version of this article containing additional figures can be found in the Reference Section on the website supporting this briefing (www.touchbriefings.com).

- Huwe CM, Synthetic Library Design, *Drug Discov Today*, 2006;11(15–16):763–7.
- Chapman T, Drug Discovery: The leading edge, *Nature*, 2004;430:109–15.
- Riester D, Wirsching F, Salinas G, et al., Thrombin inhibitors identified by computer-assisted multiparameter design, *Proc Natl Acad Sci U S A*, 2005;102(24):8597–8602.
- Kerns EH, Di L, Pharmaceutical profiling in drug discovery, *Drug Discov Today*, 2003;8(7):316–23.
- Peakman T, Franks S, White C, Beggs M, Delivering the power of discovery in large pharmaceutical organizations, *Drug Discov Today*, 2003;8(5):203–11.
- van de Waterbeemd H, Gifford E, ADMET *in silico* modelling: towards prediction paradise?, *Nature*, 2003;2:192–204.
- Faller B, Wang J, Zimmerlin A, et al., High-throughput *in vitro* profiling assays: lessons learnt from experiences at Novartis, *Expert Opin Drug Metab Toxicol*, 2006;2(6):823–33.
- Kassel, Daniel B, Application of high-throughput ADME in drug discovery, *Curr Opin Chem Biol*, 2004;8:339–45.
- Barrett JS, Koprowski SP Jr, The epiphany of data warehousing technologies in the pharmaceutical industry, *Int J Clin Pharmacol Ther*, 2002;40(3):S3–S13.
- Scoffin R, The new wave in Electronic Laboratory Notebook Systems, *Chem Biol Drug Des*, 2006;67:184–5.
- Stoner C, Gifford E, Stankovic C, et al., Implementation of an ADME enabling selection and visualization tool for drug discovery, *J Pharm Sciences*, 2004;93(5):1131–41.
- Gola J, ADMET property prediction: the state of the art and current challenges, *QSAR Comb Sci*, 2006;12:1172–80.
- Pang YP, *In silico* drug discovery: solving the ‘target-rich and lead poor’ imbalance using the genome to drug lead paradigm, *Clin Pharmacol Ther*, 2007;81:30–34.
- Howe TJ, Mahieu G, Marichal P, et al., Data Reduction and representation in drug discovery, *Drug Discovery Today*, 2007;12:45–53.