For reprint orders, please contact reprints@future-science.com

Finding the right approach to big data-driven medicinal chemistry

“Data generation in pharmaceutical research has been industrialized without our capacity to manage, disseminate, analyze and base decisions upon these data keeping pace.”

Keywords: big data • data driven • data intensive • design cycle • fourth paradigm • medicinal chemistry

Data generation in pharmaceutical research has been industrialized without our capacity to manage, disseminate, analyze and base decisions upon these data keeping pace. Like most scientific disciplines, medicinal chemistry is becoming increasingly data intensive and dependent on our capacity to manage and exploit growing data resources. Appropriate data-intensive strategies are required to ensure most value can be gained from all new scientific endeavors by using information technology to improve experimental design, data management, data analysis and communication. Fundamental is the need for drug-discovery organizations to enable its drug hunters [1] to make decisions informed by the content of their internally generated data and their integration with external data [2]. Addressing these requirements is commonly referred to as the challenge of big data [3], referring to the analysis of datasets too large, unstructured, diverse or rapidly changing to be analyzed conventionally [2]. While synonymous with predictive (data) analytics, big data do not refer to any specific technology or solution, but rather a new scientific environment in which we all work. Disregarded by some as hype, there can be little doubt that our increasing data resources provide rich opportunity, but also numerous challenges such as:

- How to collect, interpret, manage and disseminate these data?
- How to combine biochemical, cellular, structural, drug metabolism and pharmacokinetics data with pharmacological

results and external patent and literature data?

- How to ensure data are of sufficient quality, especially if from external sources, to reliably drive decision making?
- How to extract the meaningful information, hidden patterns, unexpected relationships, developing trends and useful connections from the growing volume of data?
- How to support drug hunters and project teams addressing these challenges?

Industrializing drug discovery

Improvements in the productivity of synthetic organic chemistry, resulting from parallelization, combinatorial and click chemistry approaches have increased the numbers of compounds generated in drug-design projects. The capacity to measure greater numbers of biological and physicochemical characteristics of these compounds has also increased as a result of high-throughput assays and miniaturization of experiments. Furthermore, there has been a clear trend to reduce late-stage attrition by introducing assays, predictive of the eventual fate of potential drugs, earlier into the drug-discovery pipeline [4]. Even previously low-throughput approaches, such as x-ray crystallography of protein–ligand complexes, are demonstrating facility to be conducted at scale, again with the result of increasing the amount and complexity of data to be considered when making design choices.

Scott J Lusher

Author for correspondence:
Netherlands eScience Center,
Amsterdam, The Netherlands
and
Computational Discovery & Design
Group, Center for Molecular &
Biomolecular Informatics, Radboud
University Medical Center, The
Netherlands
Tel.: +31 020 460 4770
s.lusher@esciencecenter.nl

Tina Ritschel

Computational Discovery & Design
Group, Center for Molecular &
Biomolecular Informatics, Radboud
University Medical Center, The
Netherlands

**FUTURE
SCIENCE**

part of

fsg

Data confidence

In the pursuit of generating large, complex and varied datasets as efficiently as possible, via robotization and increasingly reductionist approaches, we must ensure we do not sacrifice quality for quantity. Experimental design and statistical robustness must not be overlooked. It is also easy to presume that data available in a well-curated database are intrinsically accurate. Ensuring confidence in data quality requires us to:

- Perform extensive assay validation;
- Monitor deviations in activity (especially from reference compounds over time);
- Identify systematic bias (e.g., plate reading errors);
- Perform regular retesting;
- Ensure the choice of experimental repetition (duplicate, triplicate, etc.) is statistically sound;
- Make primary data available for scrutiny (and not just averaged values).

Incorporating external data

A key aspect of the big data challenge is the capacity to incorporate external data resources into the decision-making process in conjunction with proprietary data. The growth of curated chemogenomic data has been rapid in recent years [5,6] providing opportunities to extract new design rules, identify drug likeness properties, explore chemical space (both generic and compound specific) and potentially enriching deficiencies of in-house data sources.

“The challenge is therefore to ensure that the wealth of available tools are appropriately applied, in a timely fashion, to high-quality data within teams willing to incorporate new insight into their design strategies.”

Assuring a sufficient level of confidence in the completeness, compatibility and quality of externally generated data is however problematic. Identifying relevant data from the huge number of disparate publicly available data sources is addressed by Open PHACTS [7], an initiative developing an open information environment to integrate multiple data sources. These types of open public-private initiatives are crucial to managing the diversity and complexity of public data and unlocking its huge potential value [8–10].

Data analytics for quantitative drug design

The term data analytics is inextricably linked to the concept of big data and refers to application of statistical methods, such as linear regression, principal component analysis, K-means clustering, Bayesian methods and cross-validation, together with machine learning approaches such as self-organizing maps, neural networks and support-vector machines and related technologies such as genetic algorithms. These approaches underpin supervised learning (predictive modeling), unsupervised learning (data mining), cluster analysis, decision trees and form the tool box of the data scientist. They have also been widely applied in chemometrics and cheminformatics for many years, demonstrating that there is no lack of tools available to the drug designer wishing to quantitatively analyze big data. The challenge is therefore to ensure that the wealth of available tools are appropriately applied, in a timely fashion, to high-quality data within teams willing to incorporate new insight into their design strategies [11].

One area in which there is potential for new developments is in the use of data visualization to reduce the complexity of multi-parameter drug design. The sheer volume and variety of data will require drug hunters to spend less time looking at individual molecules and to focus more on analyzing trends and patterns in a data-centric manner.

The design cycle

The ‘design, synthesis, testing and evaluation’ cycle has always underpinned chemical design with the goal to improve the overall properties of the compound series by balancing an array of often conflicting properties during successive rounds of design and synthesis. At each step, newly generated data should be evaluated with existing data and insight to inform the next round of synthetic choices. Achieving this is dependent on:

- Ensuring all data are generated in a timely synchronized fashion and treated with equal value;
- Data being readily available in user-friendly and comprehensive information systems.

The most important task of the drug hunter is to evaluate new biological testing in the context of known chemistry rules, general- and project-specific models and any other available information such as protein structures. As data resources increase:

- The relative amount of energy spent on evaluation of data (in comparison to design, synthesis and testing) will have to increase;

- The background of drug hunters (mostly drawn from a synthetic chemistry at present) will become broader;
- Drug hunters will become increasingly computer literate, comfortable with identifying, assimilating, analyzing and visualizing complex data;
- Compound evaluation will have to transition from the study of compounds as individual entities in favor of studying developing trends and patterns in available data;
- Drug hunters will be challenged to ensure all assays inform design (and are not just used for selection/prioritization);
- Additional effort will be dedicated to retrospective analysis of data and identifying new opportunities from old data including the repositioning and repurposing of existing drugs.

Regardless of the quality of data resources, drug discovery is still dependent on project teams and imaginative drug hunters creating links, identifying opportunities and making difficult decisions and prioritizations. In terms of using data, this requires project teams to:

- Allow data to shape ideas and decisions above any other consideration;
- Develop more compound-specific models;
- Allocate more resources on synthesizing 'informative compounds' to explore structure–activity relationships;
- Revisit data-driven decisions regularly as data resources increase to avoid developing new dogma;
- Re-evaluate their projects data portfolio at key points in projects;
- Seek independent evaluation of their data models and resources.

References

- 1 Bennani YL. Drug hunters incorporated! Is there a formula? *Drug Discov. Today* 20(1), 1–2 (2015).
- 2 Lusher SJ, McGuire R, Van Schaik RC, Nicholson CD, De Vlieg J. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today* 19(7), 859–868 (2014).
- 3 Lynch C. Big data: how do your data grow? *Nature* 455(7209), 28–29 (2008).
- 4 Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3(8), 711–715 (2004).
- 5 Guha R, Nguyen D-T, Southall N, Jadhav A. Dealing with the data deluge: handling the multitude of chemical biology

It is also important that data-centric approaches are recognized as a fundamental component of the team's responsibility and not the isolated activity of few or to be peripheralized by other design considerations. This requires all researchers to embrace aspects of knowledge working and become comfortable working in data-centric environments.

Conclusion

Industrialization of research and development in the pharmaceutical and biotech industries has resulted in huge investments being made in data generation. Unfortunately, investment and focus on methods to exploit these data resources to improve decision making, especially in lead finding and lead optimization, have not kept pace. Hype term or not, the big data era provides us an excellent opportunity to consider our management and utilization of the vast amounts of data generated by internal and external drug discovery efforts. Big data are a universal scientific challenge and medicinal chemistry may benefit from some of the heralded technical approaches being developed to exploit large complex data resources. However, it is unrealistic to imagine any single computational approach will address all data challenges. Rather, organizations able to make the procedural and personnel changes needed to exploit their valuable data resources will have a huge competitive advantage and the capacity to develop safer and more efficacious compounds.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

data sources. In: *Current Protocols in Chemical Biology*. John Wiley & Sons, Inc., NJ, USA (2009).

- 6 Gaulton A, Overington JP. Role of open chemical data in aiding drug discovery and design. *Future Med. Chem.* 2(6), 903–907 (2010).
- 7 Williams AJ, Harland L, Groth P *et al.* Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* 17(21–22), 1188–1198 (2012).
- 8 Hardy B, Douglas N, Helma C *et al.* Collaborative development of predictive toxicology applications. *J. Cheminform.* 2(1), 7 (2010).

- 9 Harland L, Larminie C, Sansone S-A *et al.* Empowering industrial research with shared biomedical vocabularies. *Drug Discov. Today* 16(21–22), 940–947 (2011).
- 10 Harrow I, Filsell W, Woollard P *et al.* Towards virtual knowledge broker services for semantic integration of life science literature and data sources. *Drug Discov. Today* 18(9–10), 428–434 (2013).
- 11 Lusher SJ, McGuire R, Azevedo R, Boiten J-W, Van Schaik RC, De Vlieg J. A molecular informatics view on best practice in multi-parameter compound optimization. *Drug Discov. Today* 16(13–14), 555–568 (2011).