Where are the Gaps in Your Data?

PEER REVIEWED

In drug discovery, data is king. Gathering the right information makes it easier to design better compounds, faster. How can you be sure that you are getting the most out of your SAR data? And how easy is it to see where the gaps are in your data?

Navigating a Complex Web of Information

A typical lead optimisation phase of a drug discovery project will gather many thousands of data points. Hundreds, or even thousands, of compounds will be synthesised over the course of the project and each of these compounds has an associated wealth of potency, selectivity and ADMET information.

Structure-activity relationships (SAR) make it possible to extract valuable information from this data, but thousands of data points translates to millions of comparison points. Navigating this complex web of information is a daily challenge for the project team. Their goal is to identify areas of critical activity in order to work out what synthetic decisions to make and to guide the project forward to the next lead candidate.

In addition to the data generated by the project, it's also important for the team to stay up to date by reading, extracting and summarising SAR information from published patent information and literature during the lifespan of a project. This is a time-consuming process, but new patent publications on a project of current interest will have a significant bearing on optimisation decisions on in-house series.

Getting up to Speed on New Projects

Whenever a new research project is initiated, or transferred across teams, familiarisation with the prior art for the project must be completed in the shortest possible time. The team must get an overview of the data that has been gathered to date in order to avoid wasting resources investing in directions already explored in the past. This leads to a steep familiarisation curve.

Dr Giovanna Tedesco recalls the merger between two companies that she was working with. "Projects were moved and new compounds were being made. You had to get familiar with the SAR really quickly to make sure you were not re-doing experiments or re-exploring hypotheses for which the solution was already known."

Even though such historical information, both in-house and published, is often available in electronic format, exploring the known SAR for a target can be a tedious and time-consuming exercise for the project team because of the volume of data involved.

Creating One Picture from Many SAR Data Points

Whether tracking an existing project, or getting to grips with an ongoing project, getting a useful handle on so much project data requires smart analysis solutions. Visualisation is a particularly powerful tool for interpreting large amounts of complex data.

The data journalist David McCandless works with big data and talks about the beauty of creating data maps when you are lost in information. His TED talk 'The beauty of data visualisation'[1] is an engaging introduction to the power of data visualisation and to the insights that can come about that might otherwise not be apparent. He describes the process as 'knowledge compression,' that is, 'a way of squeezing an enormous amount of information into a small space.'

SAR data for a discovery project does not enter the realm of big data, but it can still reach far beyond the level at which the human mind can draw useful conclusions from a spreadsheet. The ideal scenario is to be able to condense large data tables into a single picture that summarises structure-activity data into visual 3D maps. These maps can then be used to give an overall snapshot of the data and to inform the design and optimisation of new compounds.

For example, a visualisation tool that can show the electrostatic, hydrophobic and shape regions that have been fully explored by the project makes it easy to assess whether new designs are worth making, on the basis of whether or not they bring new knowledge to the project (Figure 1). This visualisation makes it easy to identify areas where there is little or no SAR data, helping scientists to target future investigations.

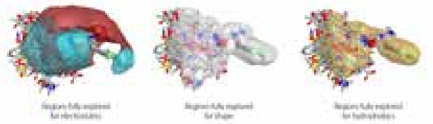

Regions fully explored for electrostatics
Regions fully explored for shape
Regions fully explored for hydrophobics

Figure 1: These visualisations of the electrostatic, hydrophobic and shape regions that have been fully explored by a project condense SAR data into a form that is easier to understand and interpret.

Converting Patent Data into 3D Maps of SAR

In this case study, the 3D SAR mapping software Activity Atlas[2] was used to explore the SAR of a large data set of orexin 2 receptor ligands taken from the US patent literature.

Activity Atlas is a probabilistic method of analysing the SAR of a set of aligned compounds as a function of their electrostatic, hydrophobic and shape properties. The method uses a Bayesian approach to take a global view of the data in a qualitative manner. This is useful for gaining a better understanding of the features which underlie the SAR of a compound series.

Crystal Structure of Suvorexant Bound to the Human Orexin 2 Receptor

The Orexin system is composed of two widely-expressed G-protein coupled receptors: the orexin 1 (OX$_1$R) and orexin 2 (OX$_2$R) receptors, which respond to the two peptide agonists orexin-A and orexin-B. These receptors work in the central nervous system to regulate sleep and other behavioural functions in humans[3].

Suvorexant is a first-in-class drug for the treatment of insomnia developed by Merck & Co with the trade name Belsomra. Suvorexant binds to both human OX$_1$R and OX$_2$R with sub-nanomolar affinity, potently inhibiting orexin receptor signalling in cell-based assays, and promoting the transition to rapid eye movement (REM) and slow wave sleep in animals and humans [4,5,6].

The X-ray structure of Suvorexant bound to human OX_2R was recently solved at 2.5Å resolution ([7] PDB code 4s0v), and as shown in Figure 2, was used to drive the alignment of the compounds in the data set chosen for SAR analysis.

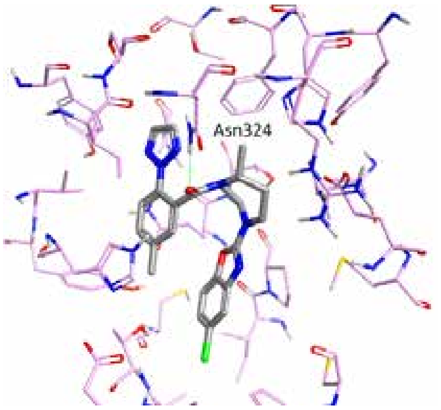


Figure 2: The crystal structure of Suvorexant bound to the human Orexin-2 receptor.

The Data Set

A large data set of approximately 380 compounds with OX_2R pKi activity data ranging from 5 to 8.5 was recently published by Janssen in the US patent literature[8]. The structures of the compounds and the related OX_2R activity data were downloaded from BindingDB[9]. The most potent compound in the series (Figure 3) was selected as the reference structure for the subsequent modelling work.

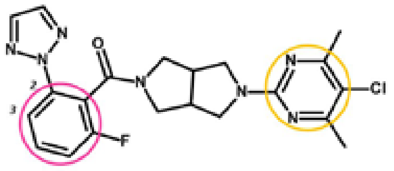


Figure 3 – the reference structure for the Janssen data set (OX_2R pKi 8.5)

When using any 3D analysis method, such as 3D-QSAR, it is necessary to generate accurate alignments for all of the compounds in the data set. Activity Atlas uses a 3D similarity metric to compare compounds, and the alignments were carried out using the program Forge [10]. Field-based alignment was used to superimpose the Janssen reference structure to the X-ray crystal structure of Suvorexant. All of the other compounds were aligned to the reference structure by maximum common substructure alignment. Activity Atlas models were then calculated for the aligned data set.

Results

The activity cliff summary 3D maps in Activity Atlas highlight, in a highly visual manner, the most critical regions in the SAR of the Janssen data set.

Looking at Figure 4, we can see that the SAR of the central phenyl ring (pink circle in Figure 3) is crucial for modulating OX_2R activity; the preferred substituents will be those which help create the correct pattern of positive and negative electrostatic fields around the molecule. Also, steric bulk in the 2-position is clearly favourable.

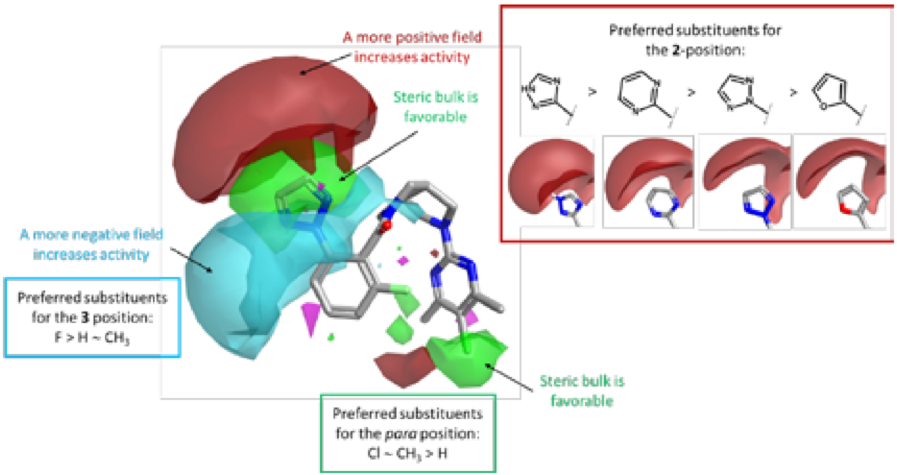

A more positive field increases activity
Steric bulk is favorable
Preferred substituents for the 2-position:
>
>
>
A more negative field increases activity
Preferred substituents for the 3 position: F > H ~ CH_3
Preferred substituents for the para position: Cl ~ CH_3 > H
Steric bulk is favorable

Figure 4: Activity cliff summary 3D maps for the Janssen data set with interpretation, observed SAR and positive electrostatic interaction potentials for substituents at the 2-position of the central phenyl ring.

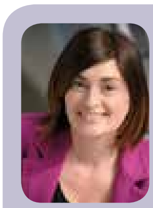
There is a critical SAR region also on the pyrimidine ring on the right side of the molecule (orange circle in Figure 3), where steric bulk in the *para* position is also beneficial for OX_2R activity.

Conclusions

This visual analysis method was very useful for quickly summarising, analysing and understanding the SAR of this large collection of compounds gathered from US patent information. The relevant SAR information was summarised into one interactive visual representation for this chemical series, demonstrating the applicability and utility of this method for the SAR analysis of large data sets.

A tool that can condense SAR data down into easily understandable visualisations makes it far easier and quicker to get to grips with the scope and quality of project data. Visualisation can help scientists to identify gaps in their project data, and to see where more investigation is required.

References
1. http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization
2. http://www.cresset-group.com/activity-atlas
3. Li, J., et al., Br. J. Pharmacol. 171, 332-350 (2014)
4. Michelson, D. et al., Lancet Neurol. 13, 461–471 (2014).
5. Winrow, C. J. & Renger, J. J., Br. J. Pharmacol. 171, 283–293 (2014).
6. Cox, C. D. et al., J. Med. Chem. 53, 5320–5332 (2010).
7. Yin, J., et al., Nature 519, 247-250 (2015)
8. US Patent 8,653,263 B2
9. https://www.bindingdb.org
10. http://www.cresset-group.com/products/forge



Dr Giovanna Tedesco is the Cresset Product Manager for computational chemists. Previously, she was a senior computational chemist at Glaxo where she supported a variety of drug discovery programs in the antibacterials and CNS areas, and led target-to-lead CNS programs.
Email: giovanna@cresset-group.com

Katriona Scoffin is a freelance science writer and marketing professional with extensive experience in the life science industry. She currently works for Cresset, an innovative company that uses software to help chemists discover, design and optimize the best small molecules for their project.
Email: katriona@cresset-group.com