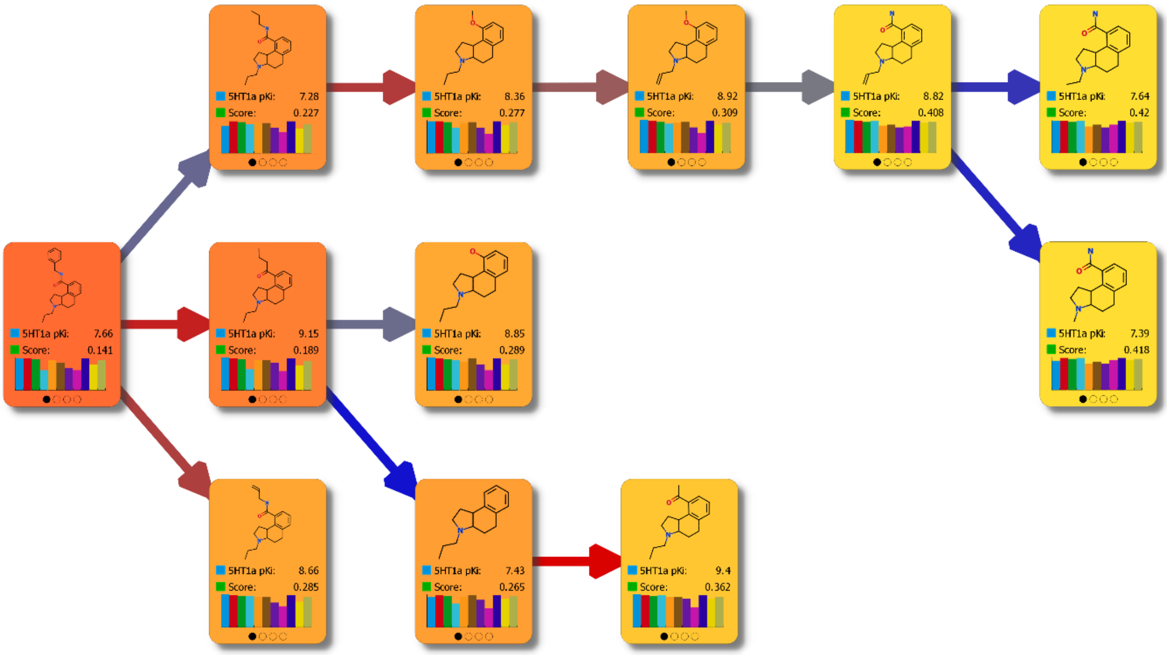
Breaking Free from Chemical Spreadsheets

Matthew Segall, Ed Champness, Chris Leeding, James Chisholm, Peter Hunt, Alex Elliott, Hector Garcia-Martinez, Nick Foster, Samuel Dowling

Optibrium Ltd., 7221 Cambridge Research Park, Beach Drive, Cambridge, CB25 9TL

Abstract

Drug discovery scientists often consider compounds and data in terms of groups, such as chemical series or clusters, and relationships, representing similarity or structural transformations, which help to navigate the complex process of compound selection and optimisation. This is often supported by chemoinformatics algorithms that analyse complex compound data and extract relevant patterns, for example clustering and matched molecular pair analysis. However, the software that supports drug discovery chemistry almost always presents these data as spreadsheets or form views; essentially long lists that make it hard to find relevant patterns or even conveniently compare related compounds. In this paper we review methods that are commonly used to extract information from chemistry data and the ways in which these data are typically viewed. We then introduce a new framework that breaks free from the restrictions of chemical spreadsheets to work with drug discovery data in the way that scientists think about them. We also illustrate how this approach can be used to view and interact with the output of algorithms to quickly and intuitively identify key structure-activity relationships with which to guide further optimisation.



Introduction

In drug discovery, project scientists think about their compounds in many different ways. At the level of an individual compound, we want to know its biological and physicochemical properties: Is it active against the intended target(s)? Does it have appropriate absorption, distribution, metabolism and excretion (ADME) properties? Is it likely to have off-target effects or cause toxicity, etcetera?

However, although the ultimate goal of any discovery project is the nomination of a high quality development candidate, this outcome is typically the result of investigations of many compounds. Therefore, to help us to navigate this selection and optimisation process, we often organise compounds using a variety of conceptual frameworks. We often consider compounds in groups, such as chemical series, clusters or 'bins' of compound (e.g. progress, reject and study further). We also consider relationships between compounds: optimisation steps or transformations to find modifications that improve activity or other properties; synthetic steps; structure-activity relationships (SAR) that will guide the further compound optimisation; and retrospective analysis of project progression in the hope of learning lessons for future projects.

Given the many and varied ways that project scientists consider compounds, their data and relationships, it is perhaps surprising that software to support drug discovery chemistry almost always presents those data as spreadsheets or form views (see Figure 1). These are essentially long lists that make it hard to find relevant patterns, focus on subsets of data or even conveniently compare a small number of related compounds. To overcome this constraint, we have even seen project teams print their compounds on sheets of paper and spread them out on a table! Some technological approaches have been explored to address this, for example by Roche (http://youtu.be/3qrQTLs1hPs); but in the age of modern, touch interfaces (Figure 2), from the perspective of user interaction, chemistry software has largely been stuck in the 1990s…

Often, advanced chemoinformatics algorithms are used to analyse complex compound data and extract important patterns and SAR. However, while these can identify inconspicuous relationships between compounds and their properties, the output also tend to be presented as yet more tables or spreadsheets, making it difficult to interpret and act upon the results and often needing an expert to pore over the output to reach a conclusion.

Data visualisations, such as scatter plots, box plots, pie charts, histograms and SAR tables can help and illustrative examples are shown in Figure 3. However, beyond a link to a spreadsheet of data, so that points in a plot can be selected to highlight the corresponding rows, these are static displays of the raw data. They don't allow a scientist to impose their own order on the data to represent the way in which they are thinking about the project's compounds. Perhaps paradoxically, we would like to visualise structured data in an unstructured way.

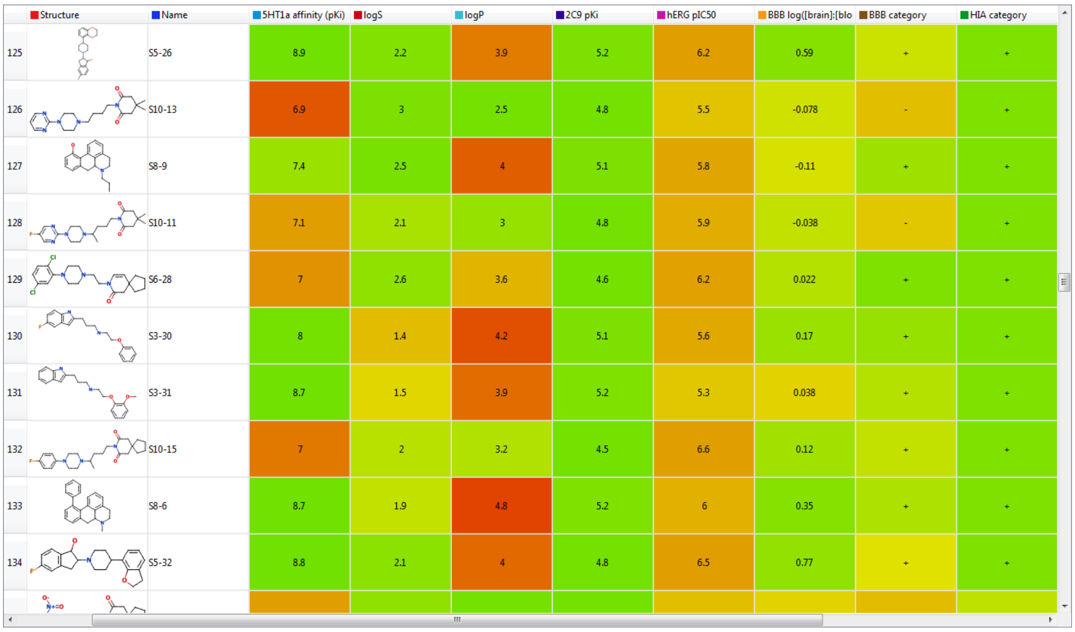
In this paper, we will review the methods that are commonly used to impose order on, and extract information from, chemistry data and the ways in which this information is typically viewed. We will then introduce a new framework, in which compound structures and data are arranged as cards, which can be positioned, stacked and linked, under the complete control of a scientist. We'll illustrate how this can be used to work with drug discovery data in the way that we think about them and how it can be used to view and interact with the output of algorithms to quickly identify key SAR with which to guide further compound optimisation.

Chemoinformatics Algorithms

Numerous chemoinformatics algorithms are routinely applied to compound data sets. These find patterns and highlight relationships that help to select compounds or series and guide further optimisation [1]. Here we briefly describe some common methods and the ways in which they are applied.

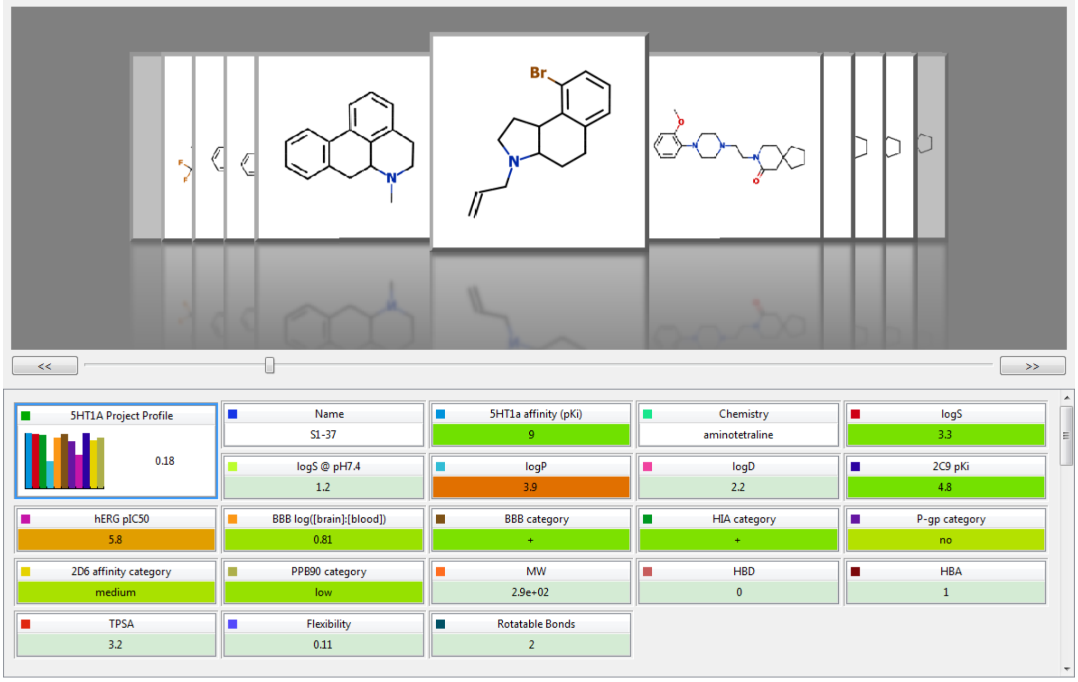
Clustering

Clustering algorithms group together compounds which are similar in terms of structure or properties. A common application of clustering is to identify series of similar compounds within a diverse data set, for example to identify hit series from a high throughput screening campaign. There are numerous clustering methods that can be applied to this challenge, including K-means [2], Jarvis-Patrick [3] and dbclus [4], based on distance between compounds in a descriptor space or measures such as Tanimoto similarity [5] between structural fingerprints. Common substructure [6] or scaffold detection [7] methods can also be used to cluster compounds that share a significant structural motif.



	Structure	Name	5HT1a affinity (pKi)	logS	logP	2C9 pKi	hERG pIC50	BBB log([brain]:[blo	BBB category	HIA category
125		S5-26	8.9	2.2	3.9	5.2	6.2	0.59	+	+
126		S10-13	6.9	3	2.5	4.8	5.5	-0.078	-	+
127		S8-9	7.4	2.5	4	5.1	5.8	-0.11	+	+
128		S10-11	7.1	2.1	3	4.8	5.9	-0.038	-	+
129		S6-28	7	2.6	3.6	4.6	6.2	0.022	+	+
130		S3-30	8	1.4	4.2	5.1	5.6	0.17	+	+
131		S3-31	8.7	1.5	3.9	5.2	5.3	0.038	+	+
132		S10-15	7	2	3.2	4.5	6.6	0.12	+	+
133		S8-6	8.7	1.9	4.8	5.2	6	0.35	+	+
134		S5-32	8.8	2.1	4	4.8	6.5	0.77	+	+

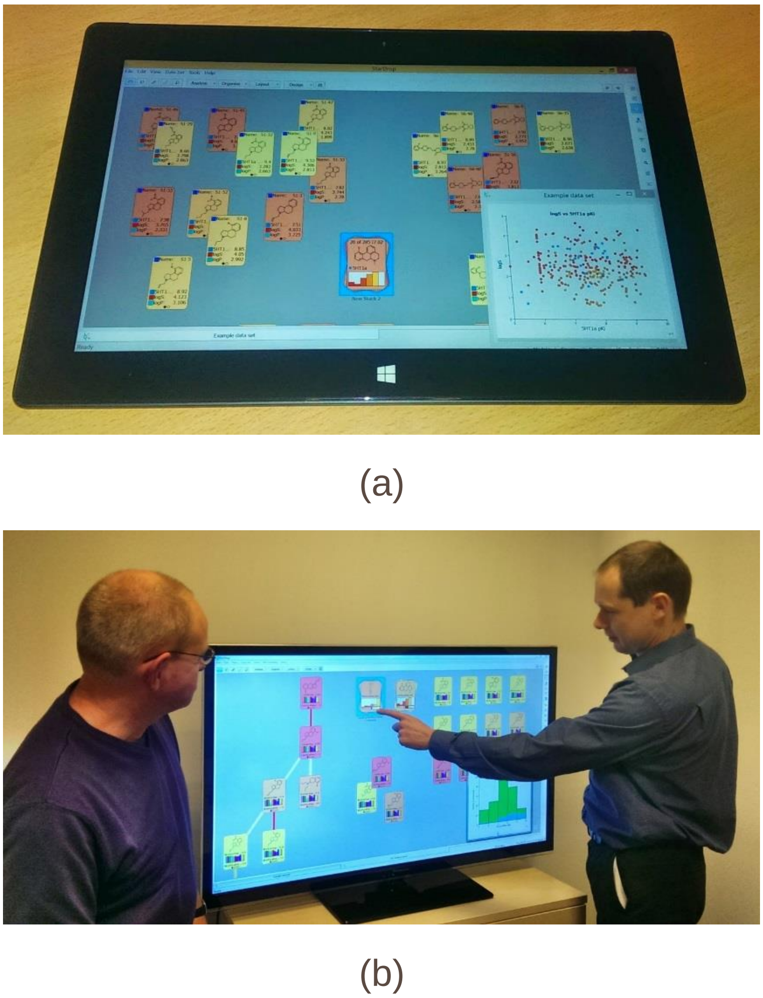
(a)



5HT1A Project Profile 0.18	Name S1-37	5HT1a affinity (pKi) 9	Chemistry aminotetraline	logS 3.3
	logS @ pH7.4 1.2	logP 3.9	logD 2.2	2C9 pKi 4.8
hERG pIC50 5.8	BBB log([brain]:[blood]) 0.81	BBB category +	HIA category +	P-gp category no
2D6 affinity category medium	PPB90 category low	MW 2.9e+02	HBD 0	HBA 1
TPSA 3.2	Flexibility 0.11	Rotatable Bonds 2		

(b)

Figure 1. Examples of spreadsheet and form views of compound data sets. (a) shows a spreadsheet in which each row represents a compound and the columns contain data including the compound structure and identifier, experimentally measured and calculated properties. The data cells in the spreadsheet have been coloured to produce a heat map on a colour scale from ideal values in green to critically poor values in red. (b) is an example of a form view in which the properties of a single compound (shown centred) are summarised. The properties for which criteria have been defined are also highlighted in a heat map on the same colour scale as in (a).



(a)

(b)

Figure 2. Modern touch interfaces provide a natural and intuitive way to interact with data: (a) tablets ideal for individuals to explore their data and (b) larger touch screens can facilitate interactive group discussions.

Other methods have been applied to visually represent chemical diversity and identify trends across a data set of compounds, such as a library, chemical series or the 'chemical space' explored by a project [8]. These include simple, linear methods such as principal component analysis [9], multi-dimensional scaling (MDS) [10] and visual clustering techniques such as the t-distributed stochastic neighbour embedding method (t-SNE) [11].

The main limitation of clustering methods is that the results often do not correspond to a 'chemist's eye view' of what constitutes a chemical series and hence the results often need to be manually refined before they can be used effectively; a time consuming and tedious process.

Activity Landscapes and Cliffs

Popularised by Guha and Van Drie [12] and Bajorath [13,14], activity landscapes compare all of the compounds in a data set to identify the most structurally similar and highlight where there are significant differences in property values between similar compounds. This is sometimes represented by a structure-activity landscape index (SALI) [12] defined as:

$$SALI_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)},$$

where A_i is the activity of compound *i*, A_j is the activity of compound *j* and *sim(i,j)* is the structural similarity between compounds *i* and *j*. Regions of the landscape with high SALI values indicate that small structural changes yield large changes in an activity, representing interesting SAR. In regions where SALI values are small, this indicates a 'flat spot' where there is little strong SAR and may indicate that the opportunities for optimisation of that activity are limited.

A related index, the structure-activity relationship index (SARI) has been proposed by Peltason and Bajorath [15] that provides a quantitative score for a set of compounds reflecting whether the SAR within the set is continuous, discontinuous or heterogeneous.

A simpler application of this analysis can be used to compare a single reference compound with other, related compounds. This is often applied to identify ‘activity cliffs’, which are small changes in structure that give rise to large changes in a property [16,17]. These discontinuities can highlight important interactions or shape constraints that represent useful SAR; alternatively, they may flag outliers that require further investigation.

Matched Molecular Pairs Analysis

Matched molecular pair analysis (MMPA) identifies pairs of compounds that differ by a single, small contiguous fragment, i.e. where there is a single point of variation such as a change in R-group, linker or a ring change. By analysing existing data sets containing similar compounds for which the same properties have been measured, MMPA can identify transformations that have a consistent, significant impact on a property of interest, such as target activity, physicochemical or ADME properties [18-20] These transformations may provide useful strategies for optimisation of novel compounds.

The principle of MMPA appears to provide a useful strategy to guide optimisation, based on analysis of large, diverse data sets of compounds and measured activities. However, when applied across data sets covering diverse chemistries and targets, the changes in activities associated with a transformation are most often distributed roughly symmetrically with an average of zero [21]. This means that statistically significant conclusions cannot often be drawn about the likely impact of a given transformation.

We can understand this because the impact of a transformation will be highly dependent on the context in which it is applied; it will be related to the binding environment in the vicinity of the substitution point, which, in turn, is influenced both by the intended target and the overall structure and properties of the compound to which the transformation is applied. Approaches to improving the reliability of predictions generated by analysis of simple transformations have been proposed, through analysis of matched series (i.e. series of more than two transformations applied to the same scaffold), through which more information on the binding environment can be inferred [22,23]. However, MMPA relies on easily identifying the context in which a transformation will apply and, conversely, where it is unlikely to be effective.

Limitations of Application

Chemoinformatics algorithms can provide useful analyses of complex data sets to identify SAR that guide compound optimisation, as evidenced by examples in the references above and the increasing adoption of these methods. However, they are not applied as effectively as they might be, due to the fact that their outputs are often challenging to understand, requiring an expert to interpret the results and provide recommendations to a project team. This interpretation can be very time consuming, particularly in light of the limitations discussed above, and the delay in providing feedback means that decisions regarding a further design iteration may be made before the results are available. More intuitive representations of the results, to make them more accessible to non-experts, would encourage more interactive application and timely feedback. This would, in turn, increase the impact of such analyses on the decisions made in the course of an optimisation project.

Data Visualisation

Data visualisation is widely used to help with the interpretation of complex data, to identify trends, find SAR and select compounds for progression [24]. A host of visual representations for chemistry data have been implemented in a wide variety of software packages, some of which are summarised in Table 1.

Table 1. Examples of software applications used for chemistry data visualisation

Software	Description	Developer	Link
Cytoscape and Chemviz	Chemviz is a chemoinformatics layer on top of the open source Cytoscape platform for network visualisation	UCSF (ChemViz)	www.cytoscape.org/ and www.cgl.ucsf.edu/cytoscape/chemViz/
Data Warrior	Free chemistry data visualisation tool	Actelion Pharmaceuticals	www.openmolecules.org/datawarrior/
Sentira	A chemically aware desktop tool for data visualisation	Optibrium	www.sentira-software.com
Seurat	A data sharing and visualization tool for all members of a discovery team	Schrödinger	http://www.schrodinger.com/Seurat/
SpotFire	A chemically aware layer built on an extensive data analytics platform.	Tibco and Perkin Elmer	http://www.cambridgesoft.com/ensemble/spotfire/
StarDrop	Comprehensive compound optimisation platform including data visualisation and Card View™	Optibrium	www.optibrium.com/stardrop
Vortex	Data visualisation and analysis solution with full chemical structure intelligence	Dotmatics	http://www.dotmatics.com/products/vortex/

Figure 3 shows some common examples of visual representations used to interpret data and represent the output of chemoinformatics analyses. A detailed overview of data visualisation techniques is beyond the scope of this review, but some illustrative examples include:

- Heat maps or ‘traffic light’ displays, are commonly used to provide context within a spreadsheet of data by colouring cells green for property values that ‘pass’ a criterion, red for those that ‘fail’ and yellow/orange for those that are ‘close’, as illustrated in Figure 1.
- Plots, charts and graphs are often used to present and explore multi-dimensional data (see Figures 3(a) and 3(b) for examples). Results from analyses such as a ‘chemical space’ can also be displayed as a scatter plot, to quickly identify ‘hot spots’ of active compounds or those that have desirable properties, as shown in Figure 3(c).
- SAR plots, such as the example in Figure 3(d), display the results of an R-group analysis within a chemical series. In these, the distribution of a property can be shown for different combinations of substituents at two positions on a common scaffold.
- Network diagrams are often used to represent activity landscapes, in which compounds are shown as points that are linked to structurally similar compounds [12,17,25], as illustrated in Figure 3(e). The colour of link can reflect the change in a property value,SALI or SARI, while the size and colour of the points can represent additional properties.
- The results of clustering can be represented as a dendrogram or tree structure that shows relationships within and between clusters of similar compounds. As shown in the example in Figure 3(f), each ‘leaf’ represents a compound and the ‘tree’ can be traversed to identify clusters of a desired size and diversity.

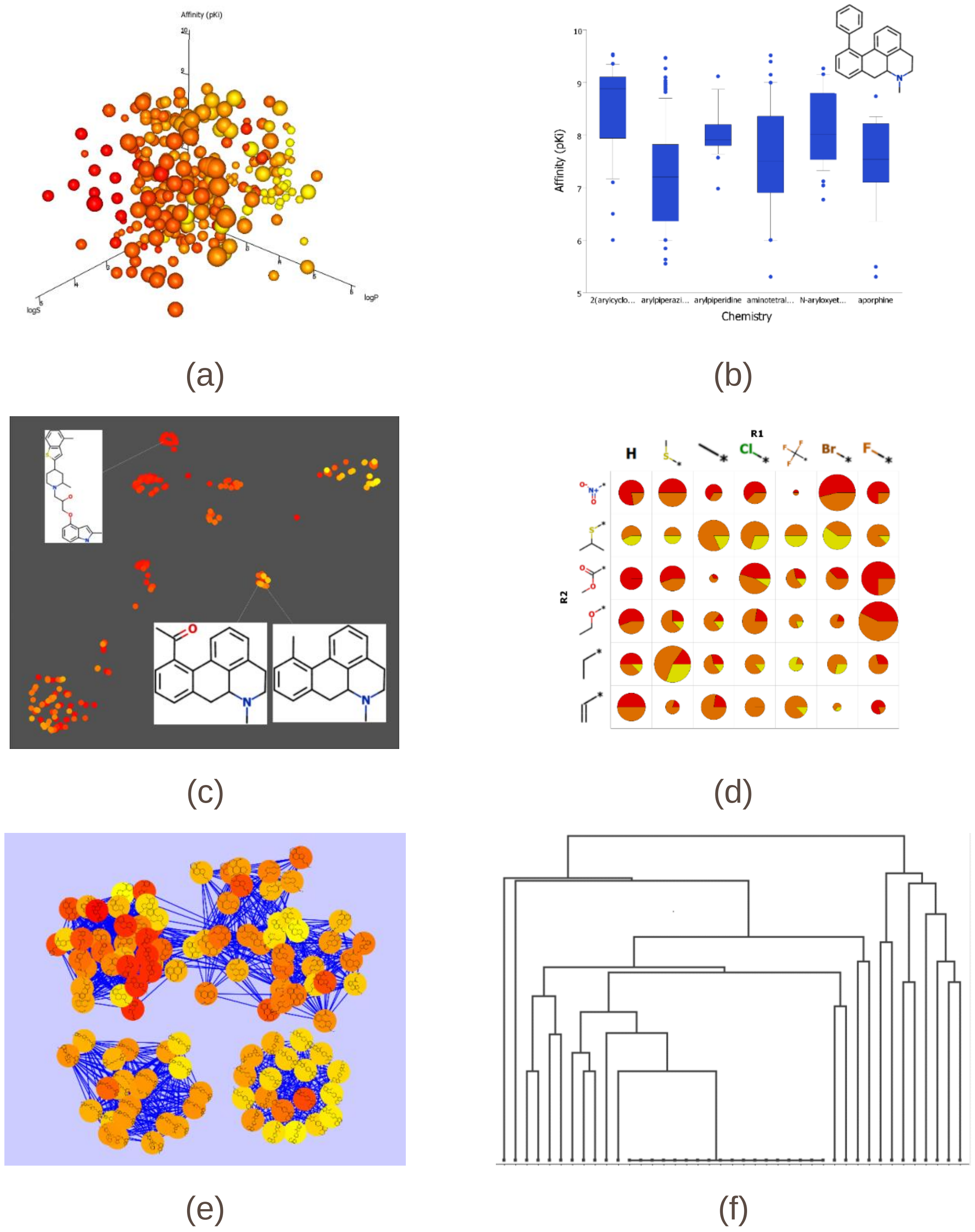


Figure 3. Examples of common approaches to visualise data.

a) A three-dimensional scatter plot for a data set of compound, in which the affinity for a therapeutic target (pK_i), calculated logarithm of the octanol:water partition coefficient (logP) and logarithm of the aqueous solubility in micromoles (logS) have been plotted. The points are coloured by predicted inhibition of the human Ether-a-go-go (hERG) ion channel from low (red) to high (yellow) and the size of each point is proportional to the logarithm of the concentration ratio between brain and blood. Created with StarDrop™ [22]

b) A box plot is a good way to compare the distributions of a property between groups of compounds. In this case chemical series are plotted on the x-axis and the affinity for a therapeutic target (pK_i) on the y-axis. The horizontal line shows the median value for each series, the top and bottom of the box indicate the 75th and 25th quartiles respectively, the whiskers show the 90th and 10th centiles and outliers are shown as individual points. One example point is annotated with the corresponding compound structure. Created with StarDrop™ [22]

Continued overleaf…

Figure 3 caption continued:

c) **A ‘chemical space’ plot of a compound library in which each point represents a single compound and the proximity of points indicates their structural similarity (2D path-based similarity calculated by a Tanimoto index [5]). Points that are close together are structurally similar, while those that are far apart are diverse on the scale of the library, as illustrated by the three points for which structures are shown. The points are coloured by the score against a multi-parameter profile from the highest scoring in yellow to the lowest scoring in red. Created with StarDrop™ [22]**

d) **An SAR plot showing the distribution of logS for different combinations of substituents at positions R1 and R2 in a series of compounds with a common scaffold. The distribution of logS for each combination is represented as a pie chart with three categories: high (yellow), medium (orange) and low (red). The size of the pie chart indicates the number of compounds with each combination of R1 and R2 from the smallest representing a single compound to the largest representing 14. Created with StarDrop™ [22]**

e) **A similarity network in which each compound is represented by a circle and links are shown between compounds with a Tanimoto similarity [5] greater than 0.7. Created with Cytoscape [23] and Chemviz (http://www.cgl.ucsf.edu/cytoscape/chemViz/).**

A dendrogram is often used to illustrate hierarchical clustering. In this, each ‘leaf’ at the bottom represents a single compound and clusters are indicated by bifurcation points. Moving up the tree corresponds to the formation of larger clusters of less similar compounds. The proximity of the leaves corresponds approximately to the structural similarity between the corresponding compounds. Created using KNIME (www.knime.org)There are, of course, many other useful forms of visualisation not shown in Figure 3, in the interests of space, including categorical plots showing probabilistic relationships between binned ranges of property values, ‘radar’ or ‘spider web’ plots that enable several compound properties to be viewed simultaneously and pie charts that show property distributions for groups or clusters of compounds. Some additional examples are shown in the Supporting Information for this paper.

Data visualisations are aesthetically appealing and provide a way to present conclusions in an impactful way. However the complexity of data often makes it difficult to draw conclusions in a rigorous way. This is exacerbated by the limited ways in which a scientist can interact with the views of their data: compound structures can be associated with elements of the visualisation, such as a point on a scatter plot, and selecting an element can highlight the corresponding row(s) in a spreadsheet. Multiple views of data can also be linked dynamically to aid exploration across multiple parameters or properties. However, having created a visualisation, the display is essentially static and it is not possible to further refine the view based on the experience or opinion of a user or share this with colleagues.

Playing with Cards

Much greater flexibility can be introduced in the way we represent and interact with compound structures and data by ‘breaking free’ from the traditional chemical spreadsheet metaphor. The objectives of this are to:

- Provide an intuitive way that a scientist can organise compounds to reflect the way they are thinking about their project
- Represent the results of chemoinformatics methods to more easily interpret and act on their output, combining a computer’s ability to analyse complex data, with an expert’s understanding of the underlying chemistry and data
- Share and present results to clearly communicate conclusions and provide the flexibility for scientists from different disciplines to view the data in which they are most interested
- Enable interactive discussions between members of a project team, possibly using a large screen

To illustrate this we will describe one such approach, called Card View™, which is implemented in the StarDrop™ software platform. Other examples of environments that have explored some of these concepts are Scaffold Explorer [26] and Scaffold Hunter [27].

Cards

In Card View, each compound is represented as a ‘card’, which can be positioned with complete freedom to create a layout according to a user’s interpretation and understanding. Cards can also be coloured by a chosen property to highlight interesting compounds for further investigation.

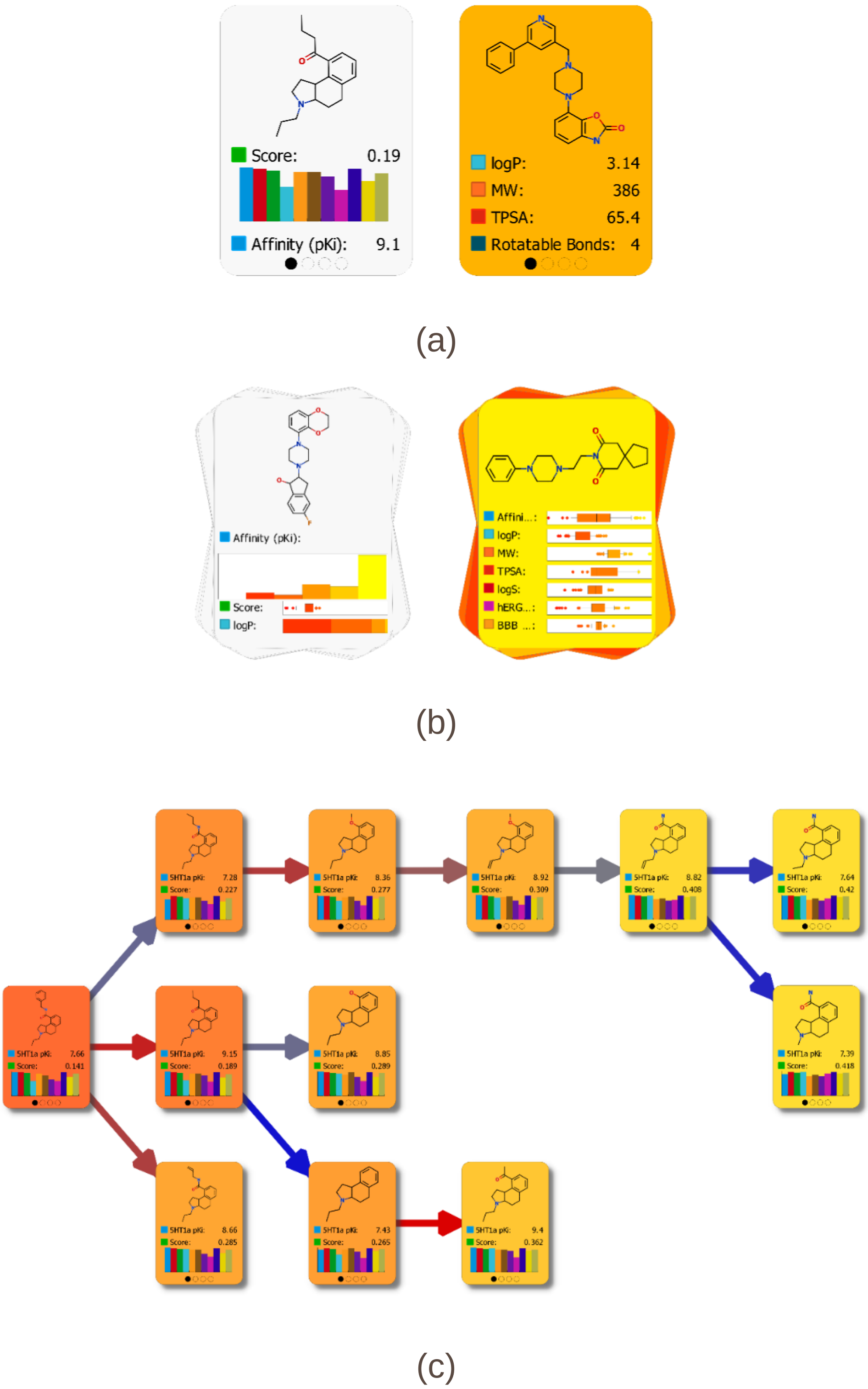


Figure 4. Illustrations of the key elements of Card View.

a) A card represents a single compound. In these examples, each card shows the structure of the corresponding compound. The card on the left also displays the score against a multi-parameter profile of property criteria, in which the impact of each individual property is indicated by the accompanying histogram and the measured affinity against the compound's therapeutic target. On the right, the card shows a summary of a compound's 'drug-like' properties: logP, molecular weight (MW), topological polar surface area (TPSA) and count of rotatable bonds. The card on the right is coloured to indicate the affinity against the therapeutic target.

Continued overleaf…

Figure 4 caption continued:

b) Stacks represent groups of compounds. Each stack shows the structure of a representative example of the compounds in the stack. The stack on the left shows the distribution of the affinities of the compounds in the stack as a histogram, the distribution of scores against a multi-parameter profile as a box plot and a 'compact histogram' showing the distribution of logP values, where the colour indicates the logP value from low (red) to high (yellow) and the area corresponds to the proportion of compounds. On the right, the cards in the stack are coloured by pK_i against the therapeutic target of the compounds and the box plots give a summary of the distributions of key properties: target affinity (pKi), logP, MW, TPSA, logS, hERG inhibition and blood-brain barrier penetration (BBB).

A link indicates a relationship between two compounds. In this example, the links represent optimisation steps and the colour of each link indicates the change in affinity for the therapeutic target in the direction of the arrow, from a large decrease (blue) to a large increase (red), with zero change indicated by a grey link. The colours of the cards represent the score against a multi-parameter profile, also shown on the cards, from low (red) to high (yellow). A card displays a structure and the most relevant data for a compound, making it easy to compare compounds across multiple criteria; some example cards are shown in Figure 4(a). The data displayed on a card can be chosen to reflect the most relevant information for a project, different sizes of cards can be used to display more or less data and a card can have multiple 'pages', enabling quick access to additional information. Card 'templates' can also save pre-defined designs that can be applied to any data set, to instantly change the view of compounds to reflect the most relevant properties for each scientist.

Simply placing cards next to one another makes it easy to compare the corresponding compounds to choose between small groups and understand property differences in terms of structure.

Stacks

Cards can be grouped to create 'stacks' of compounds, by 'dropping' one card on top of another and further cards can be added to the stack in a similar way. A stack summarises the distribution of properties of the compounds within the stack, as illustrated in Figure 4(b). As with cards, the most relevant data can be summarised, by displaying numerical averages, histograms or box plots and templates make it easy to change between views.

Stacks can be used to conveniently compare groups of compounds, whether they represent series, clusters or arbitrary groupings.

Links

A link represents a relationship between two compounds, for example a synthetic step, a more general transformation or structural similarity. Links can be directional (with an arrow) or non-directional and can be coloured to highlight large changes that correspond to interesting SAR.

Links can be used to automatically create layouts such as trees or networks to clearly highlight relationships within the context of a data set or project, such as an optimisation flow, as illustrated in Figure 4(c). This view could be used to perform a retrospective analysis of a project to understand how progress was made, where barriers were encountered and learn lessons that could be applied to the conduct of further projects.

Annotations

In addition to laying out cards, stack and links, it is important to capture 'freeform' information that reflects the thoughts and ideas of a scientist or project team. This has often been described as a project 'whiteboard'. Notes can be made on individual cards to record comments on individual compounds and card layouts can be annotated by drawing or adding text labels.

These annotations can be saved with a data set and also copied as images to communicate conclusions in presentations or reports.

A Flexible Environment for Understanding

Environments such as Card View enable scientists to easily organise their data to make and present decisions based on their understanding of the compounds and data they are exploring. However, this also provides a novel and intuitive way in which to view and interact with the results of data analysis algorithms, such as those described above.

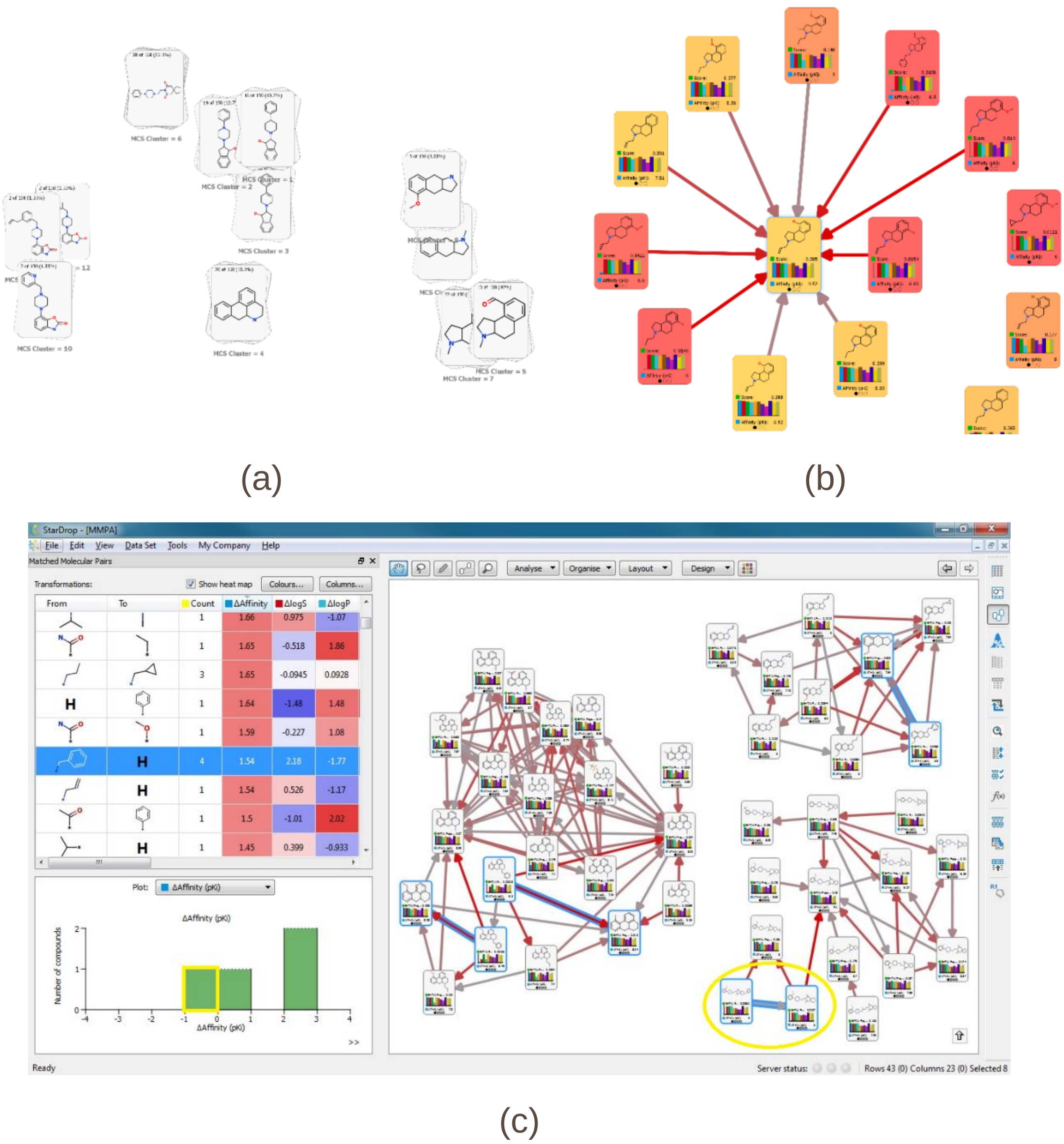
For example, the output of a clustering algorithm can be shown as stacks, facilitating the comparison of the resulting clusters to find high quality chemical series. Furthermore, the stacks can be arranged to show their relationships, for example by positioning similar clusters close to one another, as shown in Figure 5(a). This can help to spot where the clusters don't reflect a chemist's view, as discussed above, and the results can be easily 'fine-tuned', to merge clusters or remove compounds that should not be associated with a series or common scaffold. This interaction is illustrated in a video in the Supporting Information to this paper (http://youtu.be/YiIwSvwJQJc). A common, alternative approach in which clustering results are presented as a large table of structures and cluster labels makes this fine tuning process time consuming and laborious.

Activity cliffs around a reference compound can be conveniently identified by displaying the 'neighbourhood' of similar compounds using links, as illustrated in Figure 5(b). In this view, the compounds are arranged as a spiral with the reference compound in the centre and the remaining compounds in order of decreasing structural similarity from the centre outwards. The nearest neighbours are indicated by links, where the arrow shows the direction of increase of a chosen property and the colour of the link corresponds to the magnitude of the increase. Thus, activity cliffs representing important SAR around the reference compound are clearly highlighted, as short, brightly coloured links, for further investigation. For example, from the analysis in Figure 5(b) it can be clearly seen that a polar group at position C-8 is strongly preferred over the same group at C-5. Furthermore, potency in this series appears to be enhanced by a small, lipophilic group substituted on the Nitrogen, while larger substituents at this position appear to decrease the potency.

Activity landscape analysis and MMPA, which explore relationships across an entire data set, can be represented as a network, as suggested by Guha and Van Drie [12] and Stumpfe and Bajorath [17] among others. However, instead of points, the nodes of these networks are cards, enabling easy comparison between related compounds across multiple properties. Furthermore, a similar representation can help to address the context issue for MMPA, discussed above, by linking a conventional view of matched pairs as a table of transformations with a network view, as illustrated in Figure 5(c). This overview identifies the contexts in which a matched pair produces a consistent change in a property or where this relationship does not apply and, furthermore, the availability of additional data on the cards allows the impact of a specific transformation on other, important properties to be easily identified. In the example shown in Figure 5(c) we can see that the selected transformation, corresponding to the removal of a benzylic group, has been explored in three series and the results show a consistent increase in affinity for two, similar series. However, in a third series, where the context in which the transformation occurs is different, this trend does not apply. A further video in this paper's Supporting Information demonstrates the interactivity of this approach (http://youtu.be/WSmgHotXUa0).

Layouts, such as those in Card View, can also be linked with 'conventional' data visualisations, so that making a selection in one view will highlight the corresponding compounds in all others (see Figure 5(c) for an example). This can leverage the strength of each approach to investigate trends and patterns within a project's data.

There are numerous analyses that can be represented in similar ways and, above all, it is essential to the value of this approach that the results are not represented as a static view, but enable a scientist to interact with, investigate and refine the results to quickly reach a decision on how to proceed.



From	To	Count	ΔAffinity	ΔlogS	ΔlogP
		1	1.66	0.975	-1.07
		1	1.65	-0.518	1.86
		3	1.65	-0.0945	0.0928
H		1	1.64	-1.48	1.48
		1	1.59	-0.227	1.08
	H	4	1.54	2.18	-1.77
	H	1	1.54	0.526	-1.17
		1	1.5	-1.01	2.02
	H	1	1.45	0.399	-0.933

Figure 5. Example layouts in Card View generated by chemoinformatics algorithms.

(a) An example of clustering based on common substructure, in which each cluster is represented by a stack of cards. Each stack displays the substructure that all compounds in the stack have in common, as well as the number of cards in the stack and the proportion of all cards that this represents. Stacks with similar common substructures are positioned close to one another, making it easy to identify when the same series is represented by multiple clusters and improve the classification of compound by interactively merging stacks, as illustrated in the video in the Supporting Information (http://youtu.be/YilwSvwJQJc).

(b) This output conveniently identifies activity cliffs around a reference compound (centre). The other compounds are arranged in order of decreasing structural similarity with the reference in a spiral from the centre outwards. Links are shown between the reference and the ten most similar compounds; the arrow shows the direction of increase of the target affinity (pK_i) and the colour indicates the magnitude of the increase from 0 (grey) to the maximum difference of 3.5 log units (red). The cards are coloured by the overall score against a multi-parameter profile from the lowest in red to highest in yellow.

Continued overleaf…

Figure 5 caption continued:

(c) MMPA can be represented as a network of cards in which each pair of cards representing a matched molecular pair are linked. The arrow on each link indicates the direction in which the shows the direction of increase of the target potency (pK_i) and the colour indicates the magnitude of the increase from 0 (grey) to the maximum difference of 2.9 log units (red). The network view indicates that there are three separate series in which transformations have been explored. Also shown is a table view of the matched pair transformations showing the number of times each transformation occurs in the data set and the average resulting change in affinity (pK_i) and predicted logS and logP. This table is coloured as a heat map from the maximum decrease of each property in blue, through white representing no change, to the maximum increase in red. One row of the table is selected (highlighted in blue) and the corresponding links are also highlighted in the network. The transformation in the selected row represents removal of a benzylic group and results in an average increase in affinity of 1.5 log units. Below the table, a histogram also shows the distribution of the changes in affinity for the four examples of this transformation One example of the transformation, highlighted by a yellow boundary, is inconsistent with the trend for the other three, highlighting a context in which this trend does not apply.

Conclusions

In this paper we have discussed the limitations of the ubiquitous chemical spreadsheet that is used to present drug discovery data in chemistry software. We have considered these limitations in the context of commonly applied methods for visualisation and analysis of chemistry data and described a novel approach to working with data that provides much greater flexibility and interactivity.

However, flexibility must not come at the cost of interpretability and ease of use. The Card View metaphor is intuitive and enables fluid and dynamic interactions with data, whether using a mouse, touchpad or, increasingly available, touch-sensitive screens.

We have illustrated some of the benefits of this approach to manually organise data and represent individual compounds and their relationships. Furthermore, the results of algorithms that analyse larger and more complex data sets fit naturally in this environment. These methods go beyond the practical limitations of manual analysis to identify important SAR, but their effective use has previously been limited to experts by the challenge of interpreting their outputs.

The ultimate objective, of course, is not only to produce aesthetically appealing images, but to enable better decisions to be made more quickly and move drug discovery projects forward to their goals of delivering high quality candidate drugs.

Acknowledgements

The authors would like to thank Jon Tyzack for his comments on the draft manuscript.

Supporting Information

Two videos are provided as supporting information that illustrate the user interaction with Card View and how this can be used to explore sets of compound data and the output of chemoinformatics analyses. The first, http://youtu.be/YiIwSvwJQJc, illustrates the use of clustering and how the results can be 'fine tuned' by the user to reflect their understanding of the chemistries. The second, http://youtu.be/WSmgHotXUa0, shows user interactions with the results of a MMPA on a project data set.

A further document provides some additional examples of data visualisations commonly used in drug discovery.

References

1 Bajorath, J. *Chemoinformatics for Drug Disovery*. John Wiley & Sons, Hoboken, 2014.

2 MacQueen, J.B. (1967) Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.* 1, 281-297.

3 Jarvis, R.A. and Patrick, E.A. (1973) Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Comput.* C-22, 1025-1034.

4 Butina, D. (1999) Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* 39, 747-750.

5 Rogers, D.J. and Tanimoto, T.T. (1960) A computer program for classifying plants. *Science* 132, 1115-1118.

6 Raymond, J.W. and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* 16, 521-533.

7 Bemis, G.W. and Murcko, M.A. (1996) The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* 39, 2887-2893.

8 Segall, MD et al. (2006) Focus on Success: Using in silico optimisation to achieve an optimal balance of properties. *Expert Opin. Drug Metab. Toxicol.* 2, 325-337.

9 Jolliffe, I.T. *Principal Component Analysis, Second Edition*. Springer, New York, 2002.

10 Agrafiotis, D.K. (2001) Multidimensional scaling and visualization of large molecular similarity tables. *J. Comp. Chem.* 22, 488-500.

11 van der Maaten, L.P.J and Hinton, G.E. (2008) Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.

12 Guha, R. and Van Drie, J.H. (2008) Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* 48, 646-658.

13 Bajorath, J. (2012) Modeling of activity landscapes for drug discovery. *Expert Opin. Drug Discov.* 7, 463-473.

14 Bajorath, J. et al. (2009) Navigating structure–activity landscapes. *Drug Discov. Today* 14, 698-705.

15 Peltason, L. and Bajorath, J. (2007) SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* 50, 5571-5578.

16 Maggiora, G.M. (2006) On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535.

17 Stumpfe, D. and Bajorath, J. (2012) Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* 55, 2932-42.

18 Leach, A.G. et al. (2006) Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* 49, 6672-6682.

19 Dossetter, A.G. et al. (2013) Matched Molecular Pair Analysis in drug discovery. *Drug Discov. Today* 18, 724-731.

20 Warner, D.J. et al. (2010) WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* 50, 1350-1357.

21 Hajduk, P.J. and Sauer, D.R. (2006) Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *J. Med. Chem.* 51, 553-564.

22 Wawer, M. and Bajorath, J. (2011) Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. *J. Med. Chem.* 54, 4944-2951.

23 O'Boyle, N.M. et al. (2014) Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem.* 57, 2704-2713.

24 Ritchie, T.J. et al. (2011) The graphical representation of ADME-related molecule properties for medicinal chemists. *Drug Discov. Today* 16, 65-72.

25 Wawer, M. et al. (2008) Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Acitivty Relationship Indices. *J. Med. Chem.* 51, 6075-6084.

26 Agrafiotis, D.K. and Wiener, J.M. (2010) Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure−Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* 53, 5002-5011.

27 Klein, K. et al. P. Scaffold Hunter: Facilitating Drug Discovery by Visual Analysis of Chemical Space. In *Computer Vision, Imaging and Computer Graphics. Theory and Application* (Heidelberg 2012), Springer Berlin, 176-192.

28 Smoot, M.E. et al. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431-432.

29 Optibrium. [Internet]. [cited 2015 January 8]. Available from: http://www.optibrium.com/stardrop.