COMP3007 Short research paper (Coursework C2 Report)

COMP3007 Student
University of Plymouth, Plymouth, UK, PL4 8AA.

1. Abstract

The study of visualisations in chemistry has been ongoing even before the advent of computers. However, previous research has only focused on visualising individual reactions, and not on visualising extensive collections of chemical reactions. This paper seeks to explore how knowledge from other fields can be utilised to enhance visualisation in chemistry.

2. Introduction

Chemistry has been using information visualisation for a long, even before computers were invented. However, despite years of research in this area, scientists have only focused on visualising a small amount of data, and no one has visualised large chemical datasets yet. Currently, the majority of chemical visualisation research focuses on visualising various spectrometry results. However, the current solutions require extensive training to begin understanding.

This report aims to gather techniques for visualising data from other fields and apply them to the visualisation of chemical reaction networks. Additionally, the report aims to assess how effective these visualisations are in assisting non-professional chemists in analysing a huge set of chemical reactions.

3. Data set

The data set chosen for this project is the Open Reaction Database (ORD), available at https://docs.open-reaction-database.org. This dataset offers detailed information for hundreds of thousands of chemical reactions. Mostly, the reactions in this database are organic, so most reactions contain Carbon.

The ORD does not contain the price of reagents, so a secondary dataset was available at https://web.archive.org/web/20111216144726/http://www.icis.com:80/chemicals/channel-info-chemicals-a-z/. The dataset is from 2011. However, more up-to-date data is not freely available. For this report, the outdated data is still relevant.

4. Current state-of-the-art

4.1 Reaction Network Visualisation

There is little research in the field of Visualising chemical reaction data, with the focus of Chemistry visualisation research going into spectrometry. The main area of research is Reaction Path Analysis (RPA) which aims to visualise all the reaction states. An “application of RPA is in model refinement” [1]. This type of visualisation is suitable for optimising reaction paths, but it fails to consider other use cases outside of academia. The visualisation does not consider that the most optimal route is not always the most cost-effective for businesses. The visualisation also assumes that the shortest reaction path is always the best when a longer path could use cheaper reagents. Finally, while the visualisation is very detailed, it could be confusing for other chemists who are not academics, as it shows unstable and free-radical states that would instantaneously collapse into a more stable version.

The RPA study contains valuable data encoding, including a visualisation that displays the expected yield percentage for each step. The thickness of the arrows represents the rate of reaction, which is helpful for chemists in understanding the value of catalysts.

4.2 Visualising Network Data in Hyperbolic Space

At a high level, the ORD data could be viewed as a network/graph of reactions, with each compound being linked to each other through a reaction.

In a study analysing the effectiveness of visualising social network data in hyperbolic space, the researchers found that it “conveys both relational and structural information for a given large temporal social network dataset” [2]. While this paper only looks at using hyperbolic browsers for social network data, the usability study focused only on using hyperbolic browsers and not the data. The study also found that “inexperienced users can quickly perform high-level analysis regarding the structural properties of actors in a large temporal social network dataset accurately” [2]. All this evidence shows that a hyperbolic browser is a usable and intuitive way of displaying highly relational and networked information.

4.3 Bubble Tile Maps

Bubble maps have been used to visualise geographic data to great success. While the ORD is not geographic data, the data can be represented by each reaction placed in the periodic table.

A study showed that bubble maps are “easily understandable” and provide “more data visibility than other applications” [3]. The study suggests that bubble maps offer high-level information briefly. Furthermore, the data is hard to “estimate differences” and cannot see multiple fields of that data [3].

Each bubble in the map can encode more than just the quantity of data in the points. For example, in Figure 1, the colour of the bubble encodes the party elected in that state and how *decisive the vote was.*

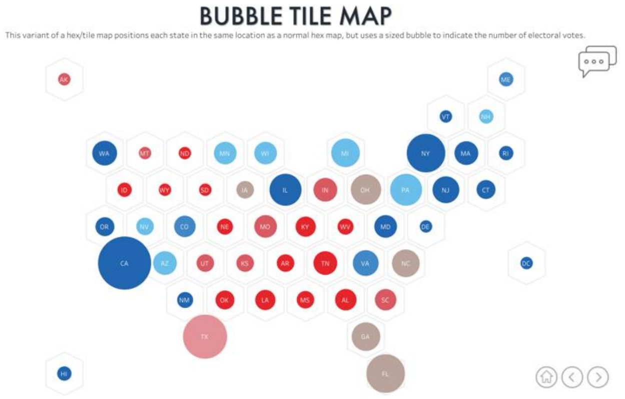

BUBBLE TILE MAP
This variant of a hex/tile map positions each state in the same location as a normal hex map, but uses a sized bubble to indicate the number of electoral votes.

Figure 1 Bubble map of the 2020 US Election.

4.4 Parallel Coordinates

As the users of the ORD can have many objectives when viewing the data, a solution specialising in multi-dimensional multi-objective visualisation is required.

Parallel coordinates are “able to visualise arbitrary dimensional data” and “aid in multi-objective search” [4]. Parallel coordinates can “prove challenging for users to understand at first”, but “users easily learn how to extract useful information from them.” [4]. Furthermore, the study investigates how parallel coordinates can be used to filter down the number of results. “To further decrease the complexity of the plot,

looking at a subset of [data points] is very helpful.” [4]. The study describes the features of *parallel coordinates well. However, it does not evaluate how effective the filtering is with users.*

4.5 Heat Maps

Heat maps have been applied in many fields. Heat maps usually show the density of a measured variable across an image or a map. *(*§Figure 2*).*

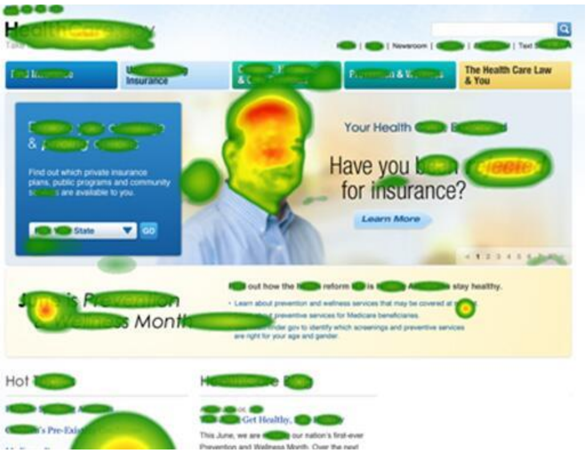

Insurance
The Health Care Law & You
Your Health
Have you been for insurance?
Learn More
Find out which private insurance plans, public programs and community services are available to you.
State
GO
Hot
Get Healthy.

Figure 2 Eye tracking heat map from

Source https://www.usability.gov/how-to-and-tools/methods/eye-tracking.html

“The visualisation of complex measurement data via a heat map approach is a valuable screening tool for quickly testing broad hypotheses regarding relationships” [5].

5. Innovation

5.1 Density data on the periodic table

A bubble tree map was used to visualise the entire dataset. Each chemical has a bubble in the place it would be on the periodic table, with the size of the bubble encoding the number of reactions containing this element.

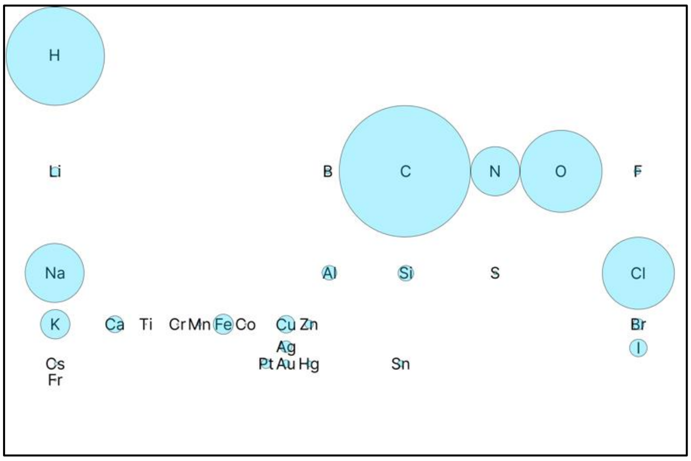


Figure 3 Bubble Map of the ORD.

As the periodic table encodes data on both the rows and columns, the row and column position must be maintained. Preserving the location of the element’s position also ensures the mastery of experienced chemists can be utilised.

This visualisation provides an overview of the entire dataset, allowing chemists to understand what the dataset contains. Any chemist can tell that the chosen dataset is mostly organic reactions from the large concentration of carbon-based reactions.

This visualisation only shows the number of reactions in the database. A mock-up to include the type of reaction, such as substitution or replacement, was created (§Figure 4). However, after getting feedback from five chemistry students, it was determined that the extra information added clutter and did not provide helpful information. Since the reaction type has little impact on the procedure, it was deemed unnecessary when selecting between reaction paths.

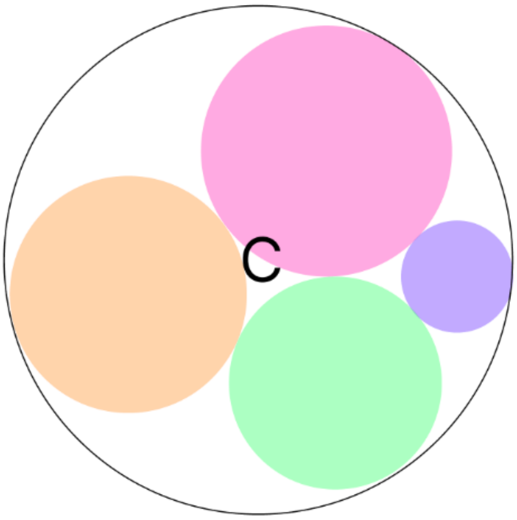


Figure 4 Encoding type of reaction for Carbon.

5.2 Networked reactions on a hyperbolic browser

Users need a zoomed-in visualisation When they click on an element to view it. Hyperbolic browsers are great for displaying networks of information like the link between reactants and products in the ORD. Most reactions have products that can be used in other reactions, creating a network. By linking one reaction's products to another's inputs, chemists can traverse this network to go from accessible chemicals to their desired product.

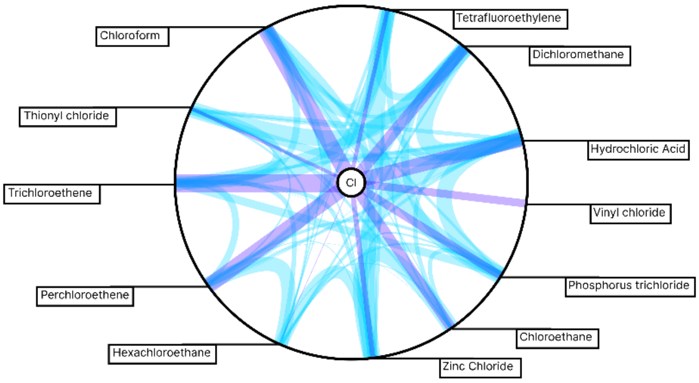


Figure 5 Hyperbolic Browser of compounds containing Chlorine.

In Figure 5, the points on the circle represent compounds in the dataset that contain Chlorine. They are all connected to each other through light blue lines. Additionally, each compound is linked to a central Chlorine node because they can be formed by reacting other compounds with Chlorine.

As this is still a high-level view of the dataset, filtered down to only one chemical, too much data encoding would make the visualisation too complex to understand. The percentage yield of the reaction is encoded into the line thickness.

Expected yields is the most essential information at this stage. Chemists would use Figure 5 to determine the most efficient path to get from the compounds they have on hand to the ones they need.

The Chlorine field in the middle combines reactants such as $H-Cl$, and Cl_2, that are commonly stocked in all laboratories and are cheap. This grouping had to be done as the ORD does not contain any reactions with Cl^-, or any other singular base element.

Displaying the reactions as a network from one compound to another is helpful in a commercial environment. Companies are always looking for ways to refine their procedure, which could result in huge savings if they find a higher percentage yield route between compounds.

5.3 Parallel coordinates

Although the hyperbolic browser can display a large number of data points, it cannot show many fields. As a result, if the user wants to compare all inputs to one output and filter down, they will need to use another visualisation. Parallel coordinates have proven to be effective in displaying both multidimensional and multiobjective data.

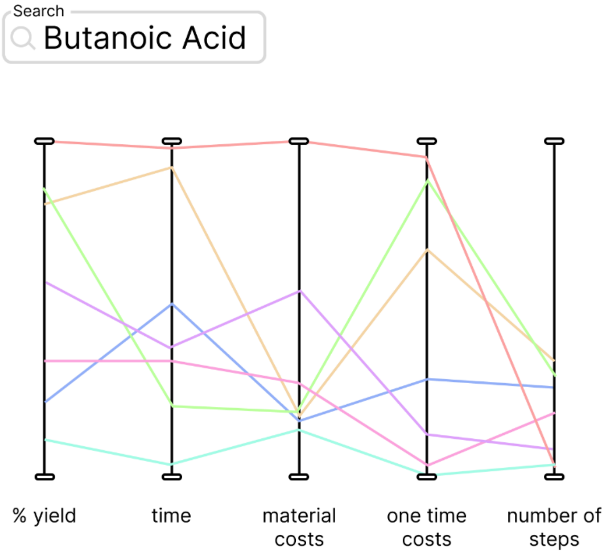


Figure 6 Parallel coordinates filtering results that have butanoic acid as the product.

In Figure 6, the user searches for reactions producing butanoic acid. All the necessary data for decision-making is available, allowing the user to view all possible reactions. This multi-objective view is helpful for chemists with different objectives. For instance, a big company aims to cut down on time and material costs, yield, at the expense of one-time expenses. On the other hand, chemists producing the product for one-time use will prioritise minimising the one-time cost.

The values of each column are encoded into the height at which the line is when it intersects the y-axis. The colour does not encode anything but aids in making following the lines easier.

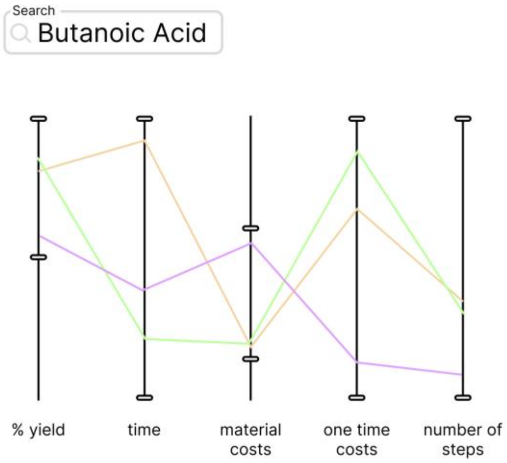
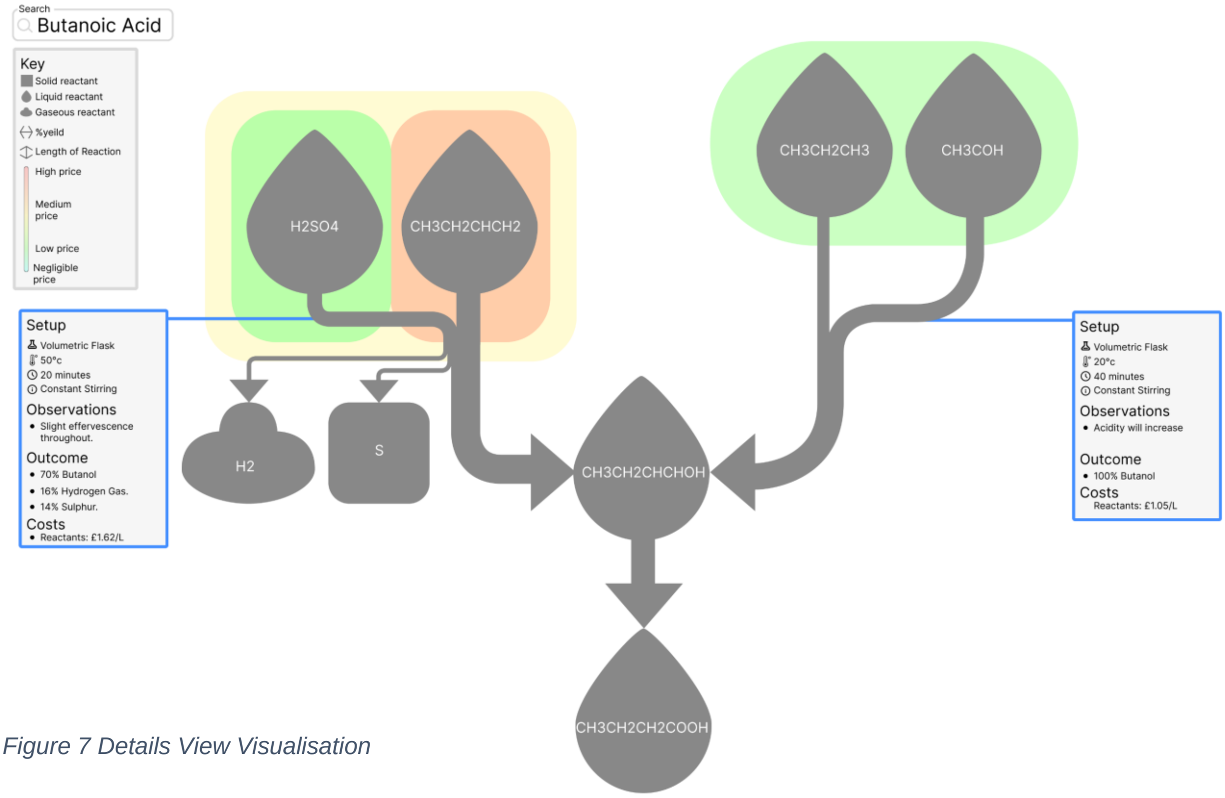


Figure 8 Filtering the paralell coordinates data

5.4 Comparing Reactions using Reaction Networks

Once the user has filtered the data down to a few points, a detailed view containing all the available data can be used. Drawing on the RPA visualisation, a simplified version removing any unstable or free radical states was created (§Figure 7).

The price data for the reagents is encoded into a heatmap in the background behind the inputs. The thickness of the arrows represents the percentage yield, while the vertical length of the arrows represents the time it takes for the reaction to occur. You can view all the details of



Setup
Volumetric Flask
50°c
20 minutes
Constant Stirring

Observations
- Slight effervescence throughout.

Outcome
- 70% Butanol
- 16% Hydrogen Gas.
- 14% Sulphur.

Costs
- Reactants: £1.62/L

Setup
Volumetric Flask
20°c
40 minutes
Constant Stirring

Observations
- Acidity will increase

Outcome
- 100% Butanol

Costs
Reactants: £1.05/L

Figure 7 Details View Visualisation

the reaction in a detailed view. Additionally, the state of the compound at each point in the process is shown.

Figure 7 demonstrates combining the RPA and heat maps approaches to show a detailed comparison between two reactions with the same product.

6. Evaluation

The visualisations were evaluated by users using the Figma prototyping feature. Each visualisation was linked together through click events. Five users all with an interest and at least an A-level in chemistry were interviewed. The number of errors and time to complete the task was measured. The participant's understanding of the results was also measured.

The first task was to look at Figure 3 and give an overview of the dataset. All participants quickly and accurately responded that the dataset mainly comprises organic compounds.

The second task was to navigate to the Chlorine network, which all participants completed quickly and without error.

While looking at Figure 5, the third task was to find the most efficient path from HCl to $\mathrm{CH3CCl3}$. The best path is $HCl \rightarrow CH3CHCl2 \rightarrow CH3CCL3$. This task took, on average, twenty seconds. Two participants found the optimal path, while three found the second most optimal path.

While looking at Figure 6, participants were asked to filter results down to two results. The objectives were the highest yield with the cheapest costs. The average time taken for this task was forty-seven seconds. Two participants did not realise that the bars could be filtered and spent more time looking at all the points,

Finally, the participants are shown Figure 7. This time they were tasked with finding the fastest option irrespective of yield or costs. All participants picked the correct path within ten seconds.

7. Conclusion

Despite little research in the field of visualising large amounts of chemical data, the data can be visualised using techniques present in other industries and interpreted in new ways.

The visualisations created throughout this paper offer good information and depth at various zoom and filter levels. Proven by a small usability study, the visualisations effectively convey trends in the data.

While most visualisations were quick and easy to understand, the hyperbolic browsers faced some issues. The hyperbolic browser proved to be a valuable tool to visualise how the reactions link together, and too much data made it hard to understand. Further research would be required into how the yields can be encoded without cluttering the view.

Using RPA, simplifying it, and combining it with a heat map proved to be a very good tool for comparing a small number of reactions. The evaluation of this visualisation proved that there was adequate information to make informed decisions for any objective.

8. References

[1] Gupta, U. and Vlachos, D.G. (2020) 'Reaction network viewer (renview): An open-source framework for reaction path visualisation of chemical reaction systems', *SoftwareX*, 11, p. 100442. doi:10.1016/j.softx.2020.100442.

[2] Cengiz Turker, U. and Balcisoy, S. (2014) 'A visualisation technique for large temporal social network datasets in hyperbolic space', *Journal of Visual Languages & Computing*, 25(3), pp. 227–242. doi:10.1016/j.jvlc.2013.10.008.

[3] Shaito, M., Elmasri, R. and Levine, D. (2022) 'Comparison of map visualisation techniques used for spatial and spatio-temporal data: An analytical survey applied to COVID-19 data', *Medical Research Archives*, 10(9). doi:10.18103/mra.v10i9.3072.

[4] Finsterwalder, R. and Grübel, G. (1991) 'A "parallel coordinate" editor as a visual decision aid in a multi-objective Concurrent Control Engineering Environment', *IFAC Proceedings Volumes*, 24(4), pp. 119–123. doi:10.1016/s1474-6670(17)54257-3.

[5] Pleil, J.D. *et al.* (2011) 'Heat map visualisation of complex environmental and biomarker measurements', *Chemosphere*, 84(5), pp. 716–723. doi:10.1016/j.chemosphere.2011.03.017.