

# Multi-Variate Interactive Visualization of Data from Digital Laboratory Notebooks

F. Oellien and W. D. Ihlenfeldt

Computer Chemistry Center, Univ. of Erlangen-Nuremberg, Erlangen, Germany

K. Engel and T. Ertl

Institute of Computer Science, University of Stuttgart, Stuttgart, Germany

---

## Abstract

*The amount of available chemical reaction and activity data generated in chemical laboratories has dramatically increased in the last years due to the widespread introduction of robotic techniques. Classical methods to display and analyse the information stored in digital laboratory notebooks and reaction databases have become inadequate. We present novel methods for the interactive graphical visualization and mining of such datasets.*

## Keywords:

*Chemical reactivity, biological activities, data mining, information visualization, laboratory notebooks.*

---

## 1. Introduction

The number of reactions executed in chemical laboratories has notably increased in recent years, probably by one or two orders of magnitude compared to the usual output per chemists ten years ago. This can mainly be attributed to the introduction of new robotic technologies, such as combinatorial synthesis, parallel synthesis and high-throughput screening. Information on the outcome of attempts to run a specific reaction set-up or a generic class of chemical reactions, or the trends in the measured binding activity to a biological receptor are of prime importance to chemists. Reactions which do not work, or which can be made to work only under extreme or particularly precisely controlled conditions are of limited practical value, if they are to be used in the context of industrial processes. In many cases the ultimate goal is the development of a robust synthetic procedure. Similarly, the observation and analysis of trends in biological activity in successfully synthesized compounds is a most important guideline for the selection of further synthetic targets.

The results of reactions and biological target screenings traditionally have been meticulously noted in laboratory notebooks. Because of patent issues, the contents of these notebooks need to be precise and comprehen-

sive. Nowadays, because of the much larger data volume which long has exceeded the possibilities of documentation on paper, laboratory notebooks are large and often highly customized applications built upon an underlying database system. Nevertheless, their contents can still be considered a huge text corpus, and the export of the results of important experiments as text documents for manual signing to support later patent claims is still an important feature of these systems. Laboratory notebooks of large pharmaceutical companies with millions of annual entries are certainly very large collections of text.

The content is peculiar, because much of the important information is non-textual. Essential parts are reaction information, which contains both numeric data (conditions, yields, etc.) and graph data (nature of the chemical compounds involved, and the associated pattern of atom and bond rearrangements), plus the results of biological screening (generally numeric data).

The data contained in these sources should be used to learn about the factors influencing the outcome of the executing experiments and use the knowledge for future planning. Any reaction which needs not to be run to establish a stable and generally applicable synthetic route, or any compound which needs not to be synthe-

sized in the quest of finding selectively binding agents saves a significant amount of money. While it is possible to easily generate unprecedented amounts of data, these technologies are not cheap. Looking at printed tables with yields, conditions, and measured activities plus graphical plots of the structure of products has been a feasible approach for the analysis of data generated with previous generation technology, where reactions were run individually in flasks. They are no longer appropriate for today's lab equipment with robot-manipulated well plates, where often thousands of combinations of reagents, solvents, temperatures etc. are run in parallel or rapid succession.

## 2. Related and Prior Work

From the introduction above, it becomes obvious that the new methods for the rapid generation and deposition of reaction data need to be augmented by data analysis tools which allow the chemist to quickly gain insights into the trends and principles which are hidden in the dataset and could be leveraged for further experiment planning. Visual data exploration and analysis techniques are an obvious choice of tools for this kind of problem.

Roberts [1] et al. have described the LeadScope application, which was developed with the visualization and analysis of biological screening data in mind. Their approach centers around several variants of 2D bar-charts. They present the values of molecular properties which are correlated with a fixed set of structural filters. The user may select 'molecular filters' (implemented as slider-class GUI objects) to focus on interesting molecular property ranges and common structural features.

The general-purpose information visualization system Spotfire [2] has been extended with features such as structure and substructure display capabilities and chemical structure database linking to support the work with chemistry datasets. Spotfire has become a *de facto* standard for data-mining visualization in chemical applications, since it is marketed by MDL Inc, a major producer of chemoinformatics software.

Beyond those two described applications, publications about chemistry-specific data-mining visualization techniques are scarce. Generally, the available tools more mostly relying on standard 2D-display styles (scatter plots, bar-charts, etc.) or simple 3D point plots. Explorations into the realm of more sophisticated 3D-graphics, for example with property encoding as glyph shapes and attributes, or the use of volume rendering for the display of massive amounts of multi-dimensional data do not appear to have been reported for this application field. Nevertheless, these approaches to information visualization have in recent years gained increased attention in computer science circles [3].

Currently available tools are generally stand-alone applications, which require explicit installation and maintenance at the work place. We are not aware of Internet- and Browser-based implementations for tools in this problem domain.

The present paper focuses on the exploration of 3D glyph-based visualization techniques and volume-based rendering approaches for chemical information processing. These techniques go beyond what currently is being available in commercial software. We attempt to demonstrate, with real laboratory data, that the introduction of advanced 3D visualization techniques has a positive effect on the quality and speed of interpreting typical reactivity and activity datasets. We try to support chemists in their struggle to stay in control of the data acquisition process and avoid being overwhelmed by the volume of data pouring in.

## 3. A Glyph-based Chemistry Information Visualization Tool

We have implemented a prototype of an Internet-enabled glyph-based chemistry data visualization tool. The application is Java/Java3D-based and can be run both as a stand-alone program or as an applet embedded in a Web browser. It links to databases via JDBC and is thus able to access nearly arbitrary compound data from remote database tables.

Selected dimensions of the dataset are mapped to the three orthogonal axes, and additionally to shape, colour and size of scene objects. Furthermore, the display can be modified by the user by applying interactive filters from a filter panel to the dataset. Filtered subsets of the original data can be selected for rendering. Since the result of the visualization is a standard Java3D scene, it can be examined in detail from different viewpoints, or by standard operations such as rotation, zooming, etc.. Also, by changing the axis ranges, parts of the dataset can be temporarily hidden from the displayed scene. Suitable tools for scene content set-up, which operate by imposing constraints onto the selected and/or displayed data can be chosen at runtime by the user from a toolbox. Thus, depending on the data at hand, various control mechanisms, such as range sliders, item sliders or binary checkboxes can be freely chosen among.

Besides providing various tools to control the displayed content of the scene from the outside, we are also exploring the use of active components which are part of the scene graph. For example, we have implemented a 3D semitransparent selection box which can be moved around and resized within the scene for data subset selection. Data subsets may be transferred to a second scene graph in an additional window. The data glyphs are active objects as well and can perform various tasks.

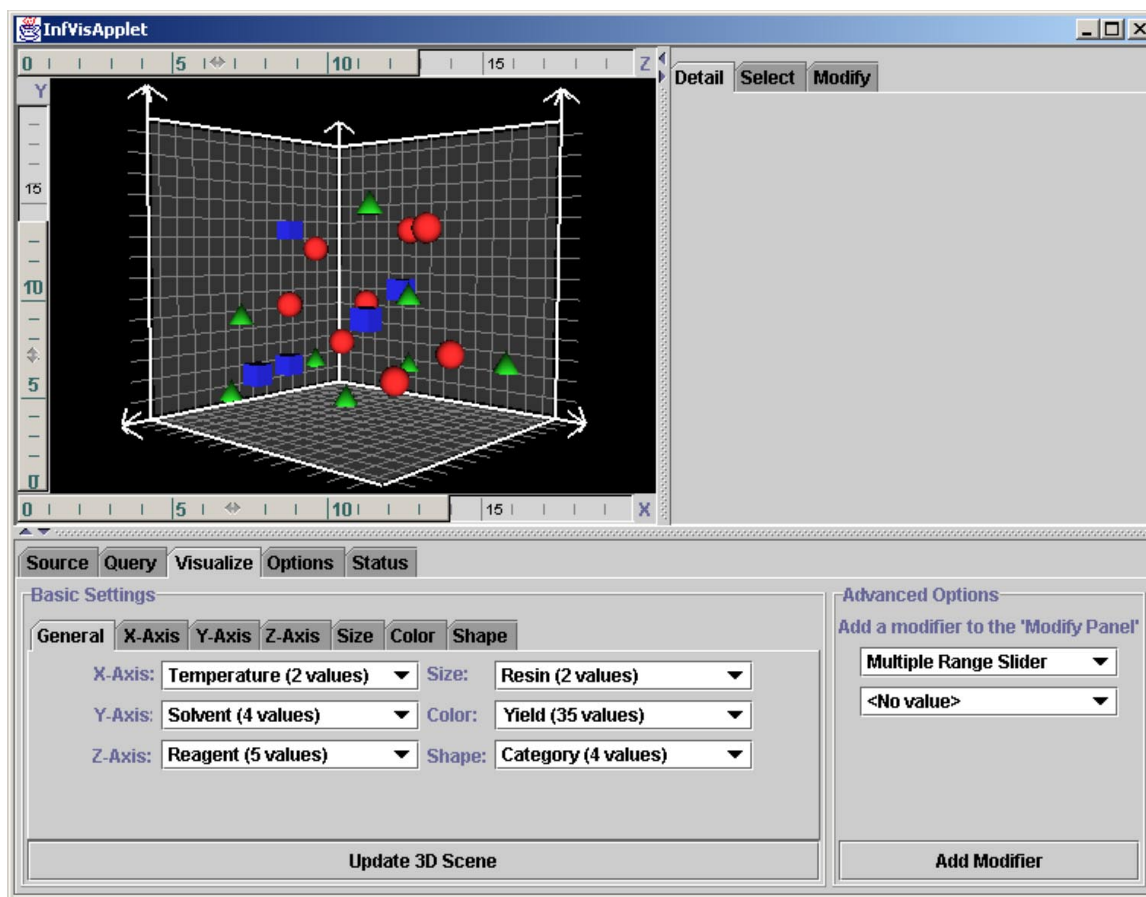


Figure 1: *Multidimensional Visualization of a Re-activity Experiment with Glyphs in 3D Space*

The identification of the associated chemical structure is the most obvious application.

We are also working on the implementation of a general client-server communication channel to allow more complex and domain-specific, yet location-independent manipulation of the data, such as clustering on property value sub-vectors or the *ad-hoc* computation of additional chemical properties, by server-side modules.

On client computers with reasonably up-to-date graphics hardware and processor speed, a few hundred or a few thousand data points can be handled with this tool. For the visualization of larger datasets, which are certainly not uncommon, alternative approaches are required. We are currently examining volumetric rendering approaches for this purpose.

#### 4. Volume-Based Chemistry Data Rendering

The rendering of large relational data sets by volumetric methods has become an area of active research.

Becker [3] describes a method for effective volume rendering of dense scatter plots of relational data by voxelization via binning and subsequent use of the density of the data points in the bins as parameters of an opacity function. The voxel colour is computed by averaging the binned point values of an additional dimension and mapping it to a colour map. Becker proposes to link other variables in the data set to external query sliders.

A prerequisite for the convenient exploration of volume data from sizable multi-variate data sets is a high interaction rate. The interactive manipulation of query sliders, transfer function parameters and other filtering and mapping modifiers is best achieved by hardware-accelerated visualization techniques. Unfortunately, since typical resolutions of the resulting volume data are comparably limited, strong artefacts would result. We have developed a novel texture-based volume rendering

approach which is especially suited for low- to midsize resolution volume data with non-linear transfer functions. This method, named *pre-integrated volume rendering* [4,] provides higher image quality with far less slice polygons than other approaches. Pre-integrated ray segments using dependent texture fetch operations are employed to obtain high accuracy. Because of the low number of slice polygons, rasterization requirements for the graphics hardware are moderate and thus high frame rates are possible, while still maintaining good display quality. Powerful filter operations may be added using hardware-accelerated per-fragment operations.

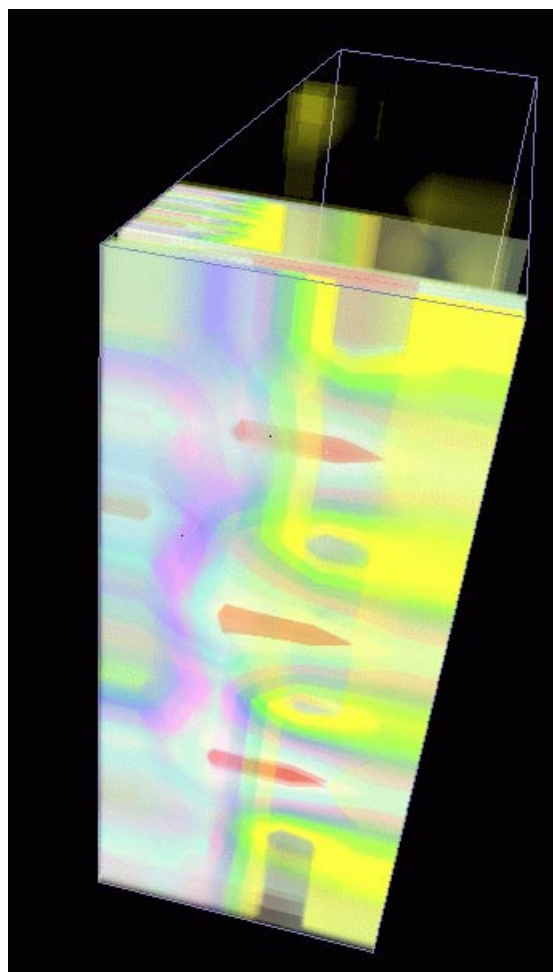


Figure 2: *Volumetric rendering of a large reactivity dataset*

The described volume rendering technique was implemented in a second prototype application. We have demonstrated that this approach is useful for rendering larger amounts of multi-variate data from our reactivity

test dataset which cannot be handled any longer by the glyph technique, due to performance and screen resolution limitations.

## 5. Conclusions

We have implemented two prototypes incorporating a pair of complementary techniques for the 3D rendering of extensive multi-variate chemical reactivity and activity data. This kind of data is usually stored as isolated experiment observation records in laboratory notebooks, which can be considered a highly specialized text form. Using the described set of tools, data can be aggregated and mined by visual inspection, yielding insights into the nature of the observed reactivity and activities which are otherwise difficult and certainly more time-consuming to obtain.

## References

1. G. Roberts, G. J. Myatt, W. P. Johnson, K.P. Cross, P. E. Blower, *Journal of Chemical Information and Computer Science*, **40**(6):1302, 2000
2. C. Ahlberg, Proc. SIGMOD **25**(4):25, 1996
3. B. G. Becker, Proc. IEEE Inf. Vis. Phoenix, Arizona, 1997
4. K. Engel, M. Kraus, T. Ertl, Eurographics/SIGGRAPH Workshop on Graphics Hardware '01 p. 9, Addison-Wesley Publishing Company, 2001