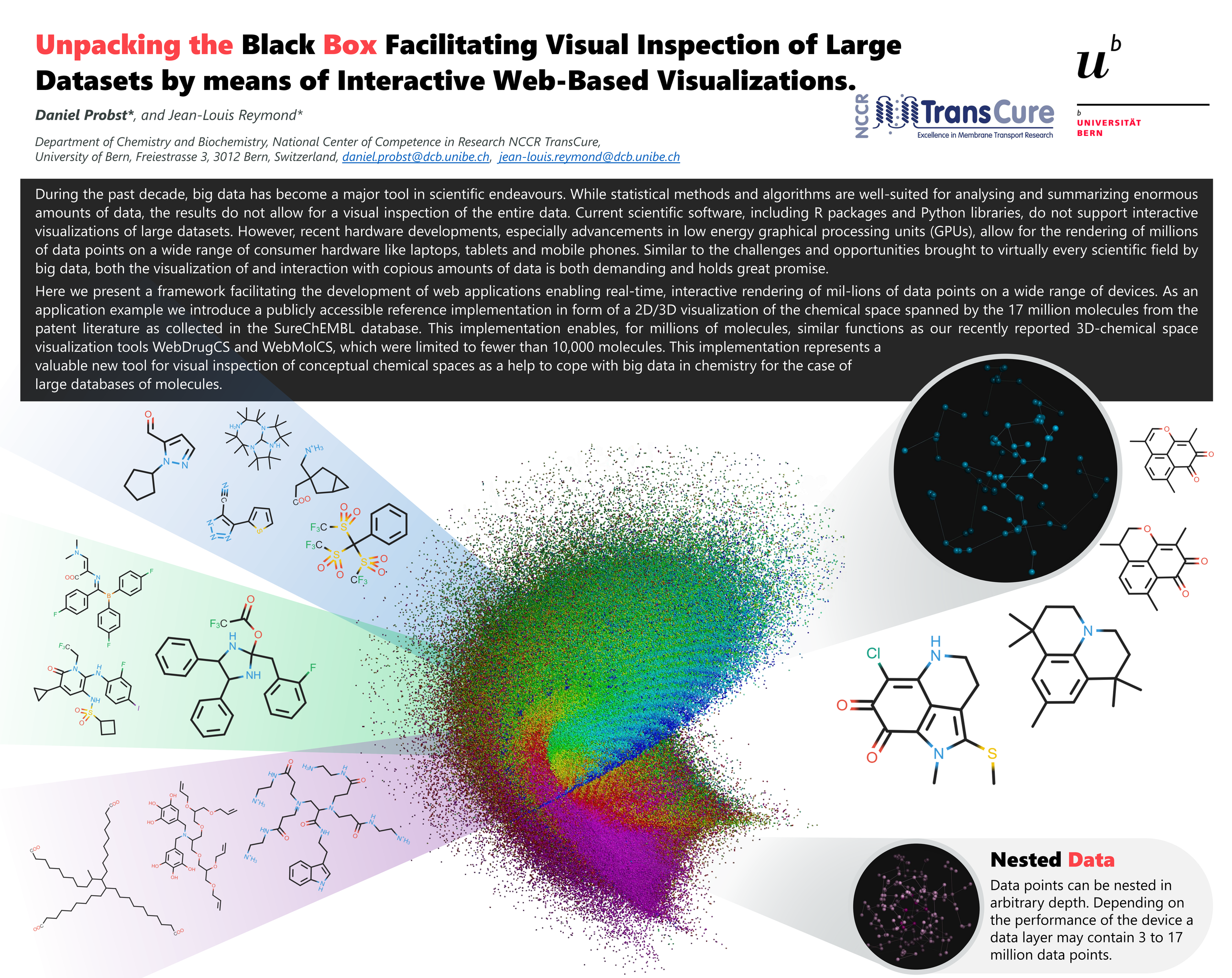
Unpacking the Black Box Facilitating Visual Inspection of Large Datasets by means of Interactive Web-Based Visualizations.

Daniel Probst*, *and Jean-Louis Reymond**

Department of Chemistry and Biochemistry, National Center of Competence in Research NCCR TransCure, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland, daniel.probst@dcb.unibe.ch, jean-louis.reymond@dcb.unibe.ch

During the past decade, big data has become a major tool in scientific endeavours. While statistical methods and algorithms are well-suited for analysing and summarizing enormous amounts of data, the results do not allow for a visual inspection of the entire data. Current scientific software, including R packages and Python libraries, do not support interactive visualizations of large datasets. However, recent hardware developments, especially advancements in low energy graphical processing units (GPUs), allow for the rendering of millions of data points on a wide range of consumer hardware like laptops, tablets and mobile phones. Similar to the challenges and opportunities brought to virtually every scientific field by big data, both the visualization of and interaction with copious amounts of data is both demanding and holds great promise.

Here we present a framework facilitating the development of web applications enabling real-time, interactive rendering of mil-lions of data points on a wide range of devices. As an application example we introduce a publicly accessible reference implementation in form of a 2D/3D visualization of the chemical space spanned by the 17 million molecules from the patent literature as collected in the SureChEMBL database. This implementation enables, for millions of molecules, similar functions as our recently reported 3D-chemical space visualization tools WebDrugCS and WebMolCS, which were limited to fewer than 10,000 molecules. This implementation represents a valuable new tool for visual inspection of conceptual chemical spaces as a help to cope with big data in chemistry for the case of large databases of molecules.



Nested Data

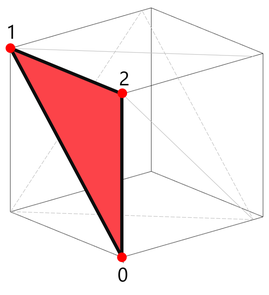
Data points can be nested in arbitrary depth. Depending on the performance of the device a data layer may contain 3 to 17 million data points.

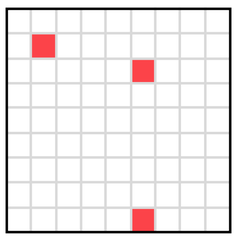
Data Visualization Engine

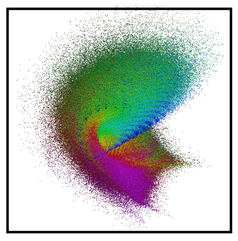
Core of the client application is the JavaScript 3D data visualization engine Lore.js, which wraps the WebGL API and is optimized for rendering large scatter plots and graphs. We chose to implement Lore.js to optimize the performance of all aspects of the WebGL rendering pipeline as well as to control main memory and CPU usage. The main performance enhancing features encompass (1) Data points are stored in an octree, a space partitioning data structure, reducing the time complexity of operations such as vertex picking or k-nearest-neighbor searches. (2) The sorting algorithm used throughout the engine is a radix sort implementation for 32-bit floats with a time complexity of $O(4n)$. (3) All data is stored in typed arrays, reducing the memory usage for 32-bit integers and 32-bit floats by 50 %. (4) Instead of rendering spheres as separate geometries, or instances of a geometry, a custom-built fragment shader simulates spherical vertices using the sphere equation and a linear shadow function to enhance the perception of depth in the orthogonal projection. (5) Adaptive framerate. While the user is not interacting with the rendered scene (except for vertex picking), the framerate is lowered by 50 to 75 %, reducing the number of context switches.

$$n \cdot (r,g,b) + n \cdot (x,y,z) = n \cdot 192\ bytes + n \cdot 192\ bytes = \boldsymbol{n \cdot 384\ bytes}$$

$$n \cdot (r,g,b) + n \cdot (x,y,z) = n \cdot 24\ bytes + n \cdot 96\ bytes = \boldsymbol{n \cdot 120\ bytes}$$







JAVASCRIPT (WEB BROWSER) → GPU MEMORY → VERTEX SHADER → FRAGMENT SHADER → FRAME BUFFER

CPU / Main Memory — GPU / Dedicated Memory

Top The user interface of the web application Faerun. Enabling the user to explore high-dimension chemical spaces on a wide range of devices.

Acknowledgements & References

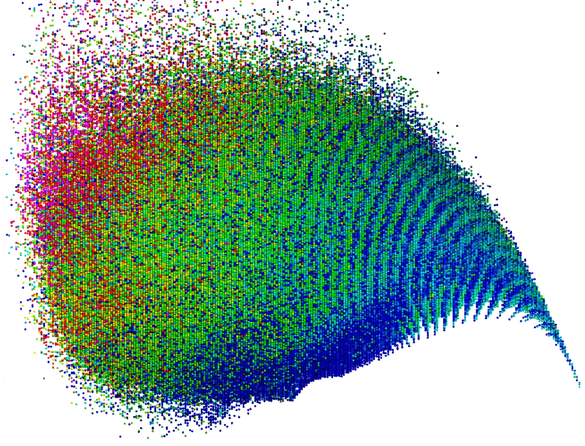
This work was supported by NCCR TransCure.

Papadatos, G. et al. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res. 44, D1220–D1228 (**2016**)

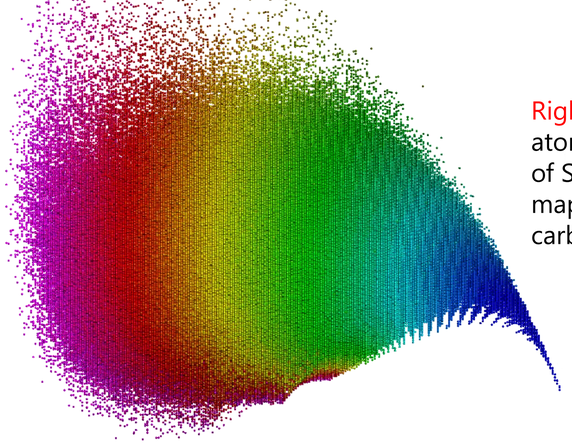
Awale, M., Probst, D. & Reymond, J.-L. WebMolCS: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. J. Chem. Inf. Model. 57, 643–649 (**2017**)

Tetko, I. V., Engkvist, O., Koch, U., Reymond, J.-L. & Chen, H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. Mol. Inform. 35, 615–621 (**2016**)

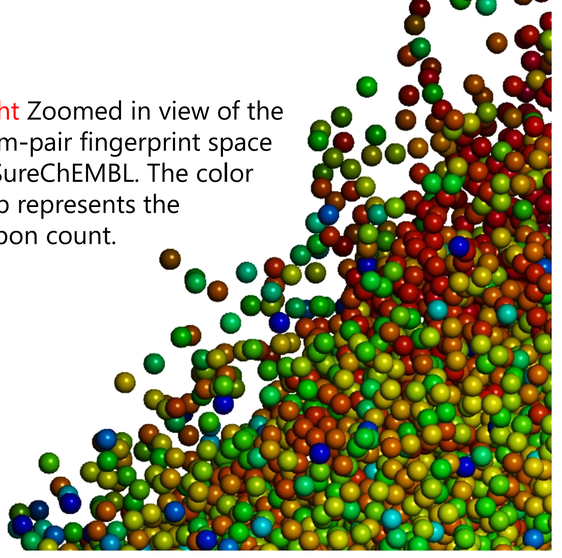
Probst, D. & Reymond, J.-L. FUn: A Framework for Interactive Visualizations of Large, High Dimensional Datasets on the Web, **Submitted**



Top A view of the atom-pair fingerprint space of SureChEMBL. The color map represents the ring count.



Top A view of the atom-pair fingerprint space of SureChEMBL. The color map represents the heavy atom count.



Right Zoomed in view of the atom-pair fingerprint space of SureChEMBL. The color map represents the carbon count.