

## Application Note

### Building ChemBank Compound Scaffolds using SARvision Plus™

*Analyzing high throughput screening (HTS) data continues to challenge most researchers today. Many scientists who rely on HTS experiments demand sophisticated but practical cheminformatics software tools to help them analyze, visualize and interpret these data. To illustrate a solution to these challenges, this report shows how an assay data set from ChemBank can be easily imported into SARvision, and how it can enable a chemist to quickly generate SAR tables, build compound scaffolds, and generate dynamic plots all in a single desktop application.*

#### Introduction

There are several public compound repositories available on the internet such as ChemBank, Zinc and PubChem. ChemBank (<http://chembank.broad.harvard.edu/>) is a web site that includes freely available data derived from small-molecule screens and has resources for studying these data. ChemBank maintains a varied set of measurements obtained from cells and other biological assay systems treated with small molecules. Specifically, the database offers information on hundreds of thousands of small molecules and hundreds of assays related to biomedicine. In order to help scientists explore, analyze, and interpret these data, an example assay data set from ChemBank was imported into SARvision Plus.

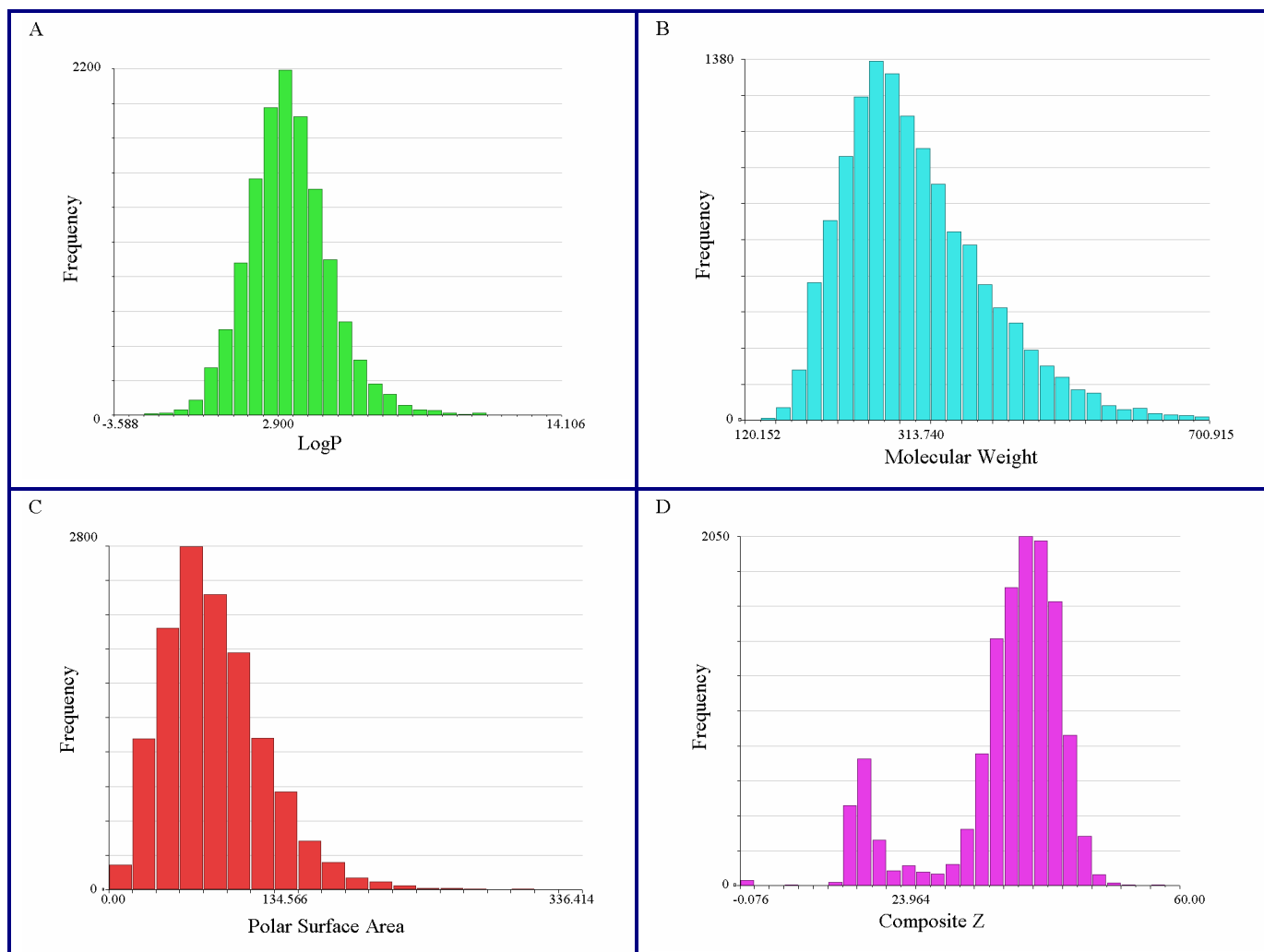
SARvision Plus is a desktop application to visualize, mine and organize chemical data. The software automatically identifies chemotypes in a chemical library and organizes data in a compound scaffold. SARvision Plus comes with capabilities to filter data by scaffold type and by any other associated data such as HTS results, docking scores or physicochemical data, and it identifies the chemotypes that satisfy user selected criteria. Multiple tools including powerful graphics facilitate the task of identifying relevant scaffolds and compounds.

#### Importing ChemBank into SARvision

ChemBank has hundreds of assay sets one can analyze, but a single data set from the BH3/Bcl-xL binding assay ([raw{Pol\(P\)}\(179.0035\)](#)) was selected to demonstrate the analysis functionality of SARvision Plus. ChemBank claims this particular assay is used to show compound inhibition of heterodimerization of Bcl-2 family member through the BH3 domain. A table of data in *csv* format was downloaded from the ChemBank website and imported into SARvision Plus. Before building a scaffold tree, the data table was modified and adjusted to improve the quality of the analysis process. First, any redundant chemical structures and rows were removed from the table. This step eliminated compounds from the original 13, 628. Next, the molecular weight (MW), logP, and polar surface area (PSA) was calculated for each compound, and the results were automatically updated in the molecule table. A summary of these properties were plotted with SARvision Plus along with the calculated and normalized Composite-Z value derived from ChemBank (Figure 1). The Composite-Z values ranged from -0.076 to

5565. Because 99.6% of the Composite-Z values were less than 60, the top 54 compounds with Composite-Z values greater than 60 were not included in the plot.

The distribution of LogP values for the compounds used in the assay appears normal, while the distribution of molecular weight and polar surface area values are slightly skewed to the left. The similar distribution trends between the molecular weight and polar surface area plots are not surprising because a lower molecular weight could correlate to a smaller surface area. Although the LogP, molecular weight, and polar surface areas of the test compounds show a reasonable Gaussian distribution, the Composite Z-scores calculated by ChemBank show a bimodal distribution. The first and second peaks are centered near the Composite Z-scores of 15 and 40, respectively.



**Figure 1: Distribution plots for A) LogP, B) Molecular Weight, C) Polar Surface Area, and D) Composite Z scores.**

## Building a Scaffold Tree

After examining the distribution of the various calculated properties, these values were filtered based on the following criteria: Composite Z value (8.5 to 60), logP (-0.4 to 5.6), and molecular weight (160 to 480). The remaining molecules (~13000) were used to generate a compound scaffold tree. The scaffold tree building process required three minutes to complete on a 64 bit Intel Core2 Quad CPU (Q6600) 2.40 GHZ processor with 3GB of RAM.

The resulting scaffold tree was pruned to show only data for compound structures with 2 – 5 rings, and it was sorted based on which compound had the highest ratio of filtered properties. The Molecular Table was sorted by lowest molecular weight, and the non-essential columns were hidden from the user view (Figure 2).

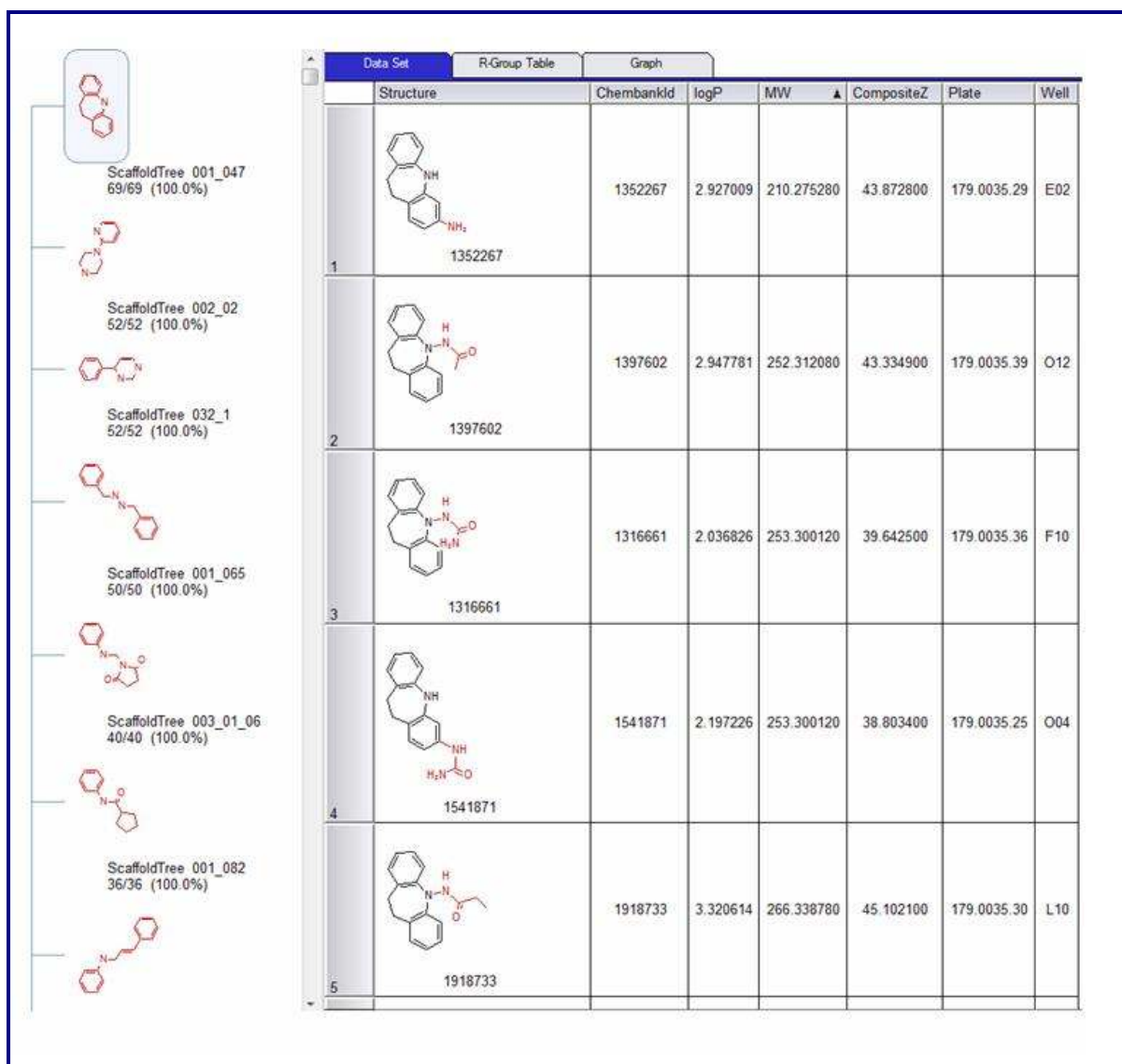


Figure 2: Pruned scaffold tree showing an individual scaffold with its associated molecules

The top scaffold was selected, and an R Group Table was generated for each compound associated with that scaffold (Figure 3). The top scaffold was comprised of 69 molecules, and it had two R groups labeled as R1 and R2. To facilitate the identification of key Composite Z scores, the Heatmap feature of the Molecular Table was used to quickly highlight any desirable value ranges.

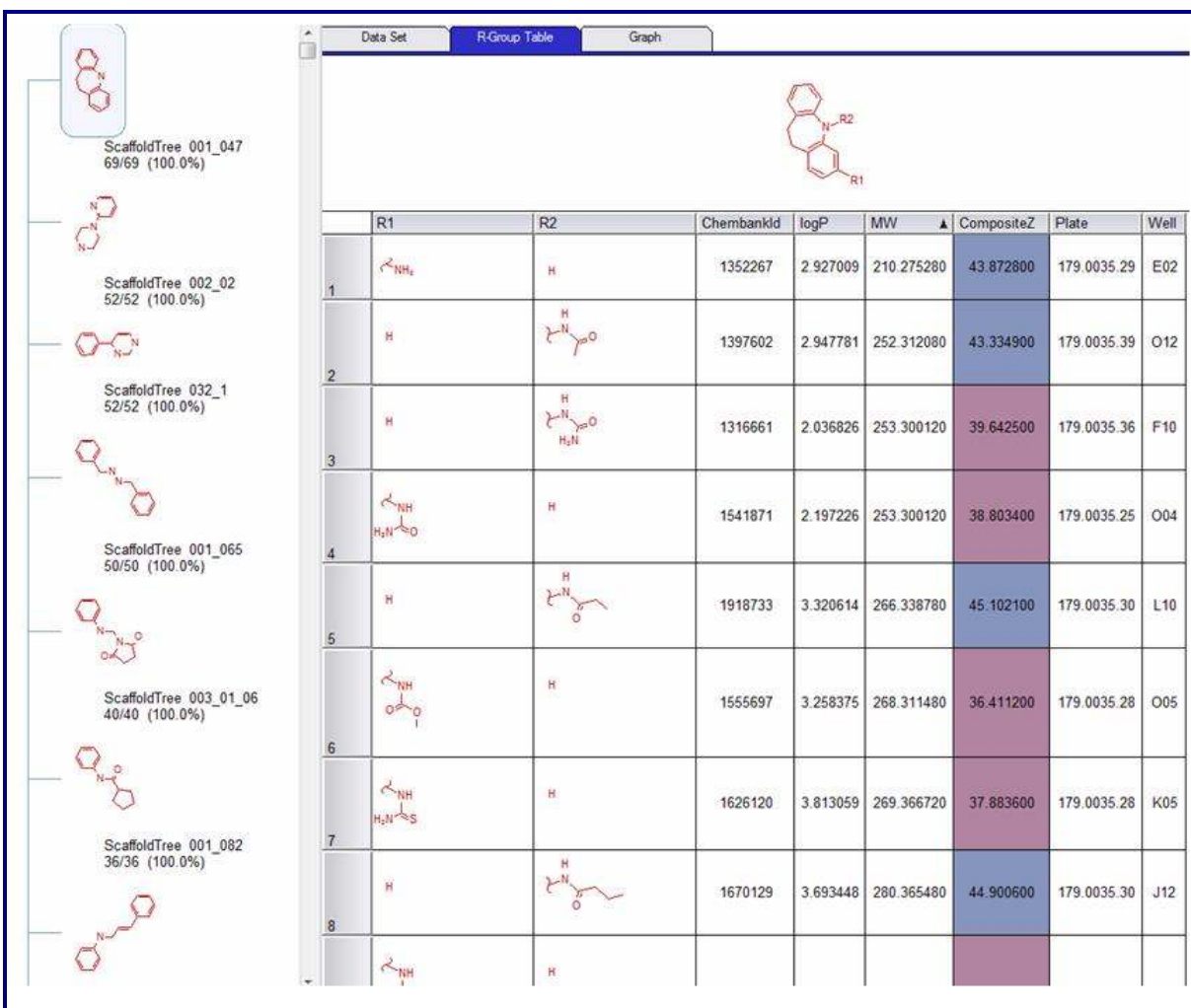


Figure 3: R-Group Table for a single scaffold with table heatmap highlighting of a column

## Summary

Analyzing data with SARvision Plus is an excellent way to identify important chemotypes from HTS assay data such as ChemBank. SARvision Plus is a practical, sophisticated desktop software tool for any scientist performing SAR analyses, high throughput screening experiments, or chemical analytics for patent exploration. The full capabilities of SARvision Plus and related software products can be experienced by visiting [www.chemapps.com](http://www.chemapps.com).