

Application Note

Extracting and analyzing compounds from U.S. Patent & Trademark Office website using ChemDox™

Extracting compound structures from patent applications continues to challenge many researchers today. Many researchers who rely on these data sources demand practical software tools to help them extract and analyze these data. To demonstrate a solution to these challenges, this report shows how a collection of patent applications can be easily analyzed with ChemDox, and how it can enable a researcher to quickly generate compound structures and perform further analyses.

Introduction

There are a variety of informational sources that contain compound names with no downloadable chemical structures. These sources include such items as published manuscripts, internal documents, and web pages. For example, patent applications can be searched on the U.S. Patent & Trademark Office (USPTO) website (<http://appft.uspto.gov/netahtml/PTO/search-adv.html>) where claimed compound names can be reviewed within a web browser. Although many patents contain hundreds of compound names, it is difficult to extract these compounds and view them as editable chemical structures. In order to help researchers identify, extract, and convert these compounds, an example set of patent applications from the USPTO website were processed with ChemDox and analyzed in SARvision Plus.

ChemDox is a desktop application to convert compound names from text to editable chemical structures. The software automatically identifies compound names in a document source and transforms them into a useful collection of chemical structures for further analysis.

Extracting compound names from U.S. Patent Applications

The USPTO online database has thousands of patents one can download, hence a specific query was created to search for compound claims for the protein called p38. Using the advanced USPTO query syntax, the following query was used: PD/1/1/2009->8/14/2009 and ABST/p38 and ACLM/compound to find similar patent applications. A total of 36 patent applications were returned, however only fifteen of them contained compound names. ChemDox was launched and compound names from each online patent application were copied (Figure A). The ChemDox interface automatically identified compound names and displayed them as black text. If a string of terms was not recognized, the text was colored red (Figure 1B). The compound names displayed in black were then extracted and viewed in a molecule structure table (Figure 2).

Analyzing extracted compounds

The results obtained from ChemDox are summarized in Table 1. The number of compounds in each selected patent application ranged from 3 to 340. The percentage of successfully extracted compounds ranged from 67 to 100 with an average percentage value of 89.

Patent Application Number	Number of Compound Names	Number of Extracted Compounds	% Extracted
90074676	6	4	67
90203702	19	15	79
90149443	13	13	100
90023725	24	23	96
90005401	7	7	100
90012299	37	35	95
90048307	3	3	100
90042877	12	11	92
90143422	35	31	89
90149459	27	22	81
90156597	6	6	100
90017036	182	127	70
90005377	301	290	96
90054382	340	279	82

Table 1: Compound name extraction results

The table created in ChemDox was saved and exported in SD format. The table was also exported to MS Excel, MS Word, and SARvision Plus with a single click of a button. Additional analysis of the extracted compounds was carried out with SARvision Plus. For example, SARvision Plus was used to calculate the distribution of LogP and molecular weight values for all of the extracted compounds for U.S. Patent Number 90005377 (Figure 3).

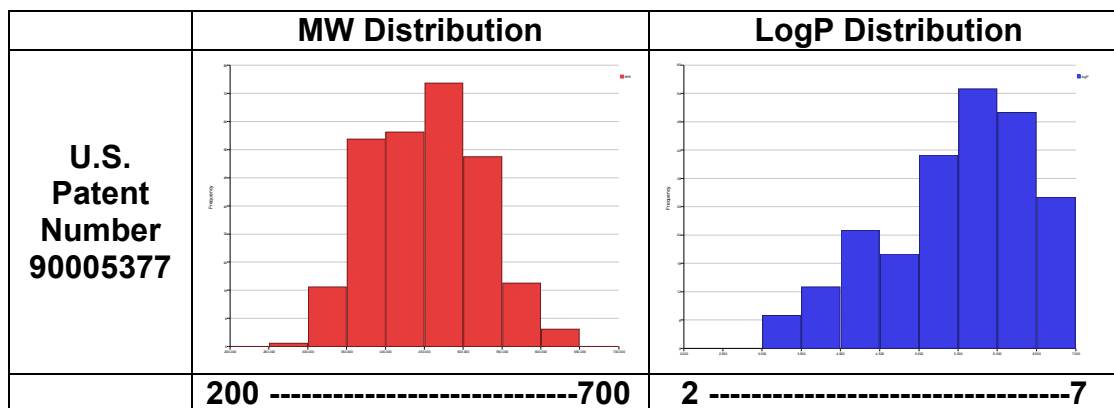


Figure 3: Molecular weight and LogP distribution of extracted compounds for U.S. Pat No. 90005377

SARvision Plus was also used to create a scaffold tree for the 290 compounds extracted from U.S. Patent Number 90005377, and seven major scaffolds were identified (Figure 4).

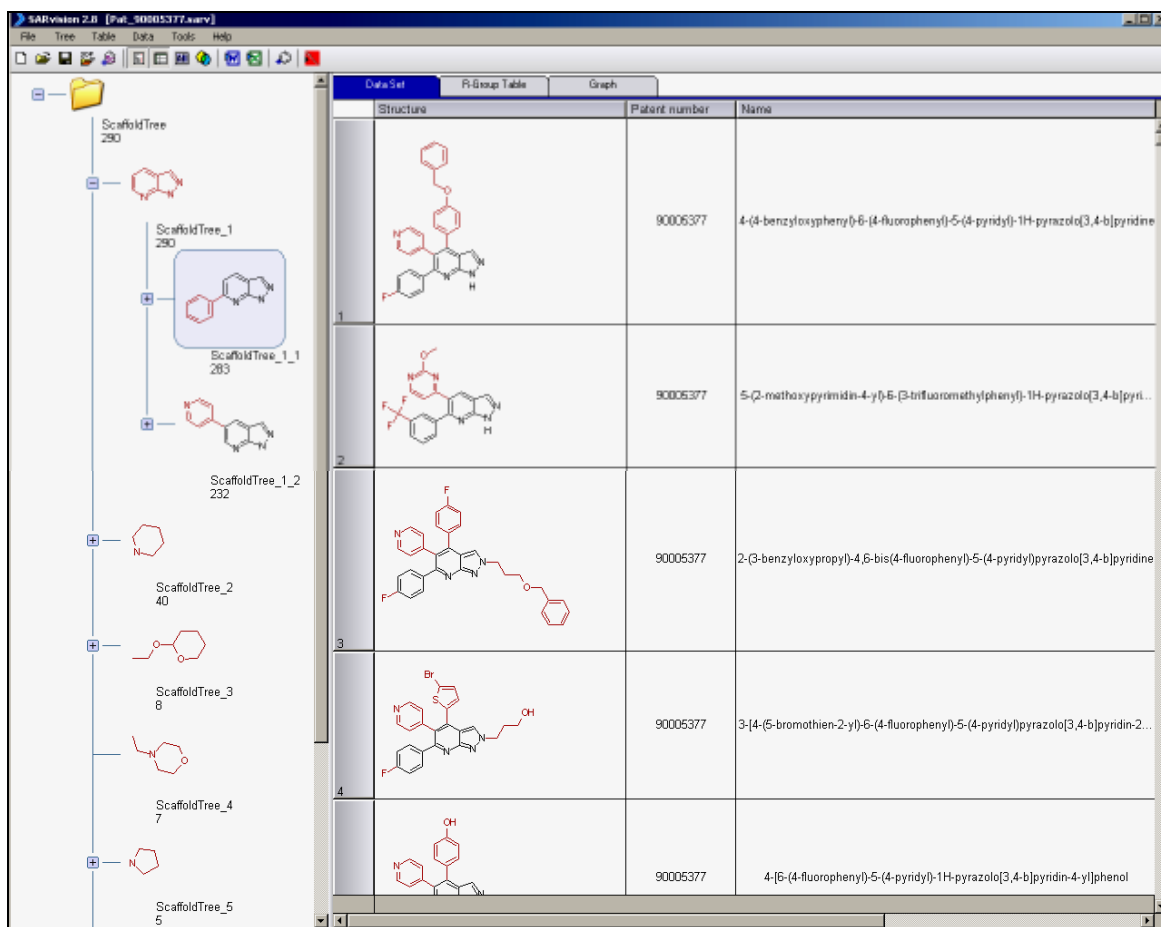


Figure 4: SARvision Plus scaffold tree based on compounds extracted from U.S. Pat No. 90005377

Summary

Extracting compounds with ChemDox is an easy way to quickly collect important compounds from document sources such as U.S. Patent and Trademark Office. ChemDox is an easy-to-use desktop software tool for any researcher performing chemical analytics for patent exploration. The full capabilities of ChemDox and other related software products can be experienced by visiting www.chemapps.com.