Herbicide database README file

The following document is split into two parts:

1) Using the Microsoft® Excel® workbook; and

2) Information describing the commands and parameters used to generate the data reported in this study, as well as instructions on how to generate similar data for additional compounds.

Part 1: Using the Microsoft® Excel® workbook

The Microsoft® Excel® workbook contains three sheets, named 'Data', Histograms' and 'Scatter Plots', which are described below.

Data worksheet

Data contains 56 columns, which contain various types of information, including SMILES strings, CAS registry numbers, IUPAC names, common names and various physico-chemical properties. At the top of each column is a cell describing the contents of the column. Each of these title cells has a comment containing a brief explanation of the type of data listed. Beginning with row 6, each row contains data for a particular compound.

Filtering Data:

The dropdown boxes in row 5 of the Data worksheet can be used to filter data, either via criteria, or in an arbitrary way via checkboxes, to return only those rows that match a user-defined criterion, such as, for example, only showing compounds that have a molar mass greater than 400. The user can set filters for as many columns as they choose. If one or more filters are active, then the row numbers on the left of the worksheet are coloured blue instead of the usual black. By applying a filter, the user also creates a subset of compounds referred to here as "filtered data", as opposed to the "unfiltered data", which corresponds to all data points that are returned when no filter(s) are applied.

Histograms worksheet

This worksheet shows histograms for most of the physico-chemical properties shown in the data worksheet. In each histogram only filtered data is displayed.

On the left hand side of each histogram, there is a "histogram control panel" which controls the number of bins, width of bins and lower bound of the leftmost bin displayed on the histogram. Some cells in the panel are coloured white, which signifies that the cell value can be changed by the user to affect the bins and in turn affect the appearance of the histograms. Some of these cells require input values while others are optional. If all of the optional cells in a control panel are empty, then the corresponding histogram operates in "fully automatic mode".

In fully automatic mode, Excel® will examine the filtered dataset and decide how many bins the data should be divided into, how wide the bins should be and will centre these bins midway between the minimum and maximum values of the dataset. As a starting point, Excel® will aim to divide the data into the "ideal" number of bins as predicted by Sturges' Rule, which determines this number of bins

using $\log_2(n)$, where n is the number of data points. By entering one or more values into these optional cells, the user can force one or more parameters to values that they desire, which will affect the histogram displayed. If values are entered into all optional cells, then the bins will not change, irrespective of any changes you might then make to the filtered data. However, if your chosen bin parameters cause any of the filtered data to fall outside the displayed bin range, then an error will be displayed.

In the lower right corner of each histogram, there is a checkbox marked "Show Unfiltered Data". When one or more filters is currently applied in the Data worksheet, then checking this box will cause the histogram to additionally display unfiltered data behind the filtered data. However, it should be remembered that when operating in fully automatic mode, only the filtered data is taken into account when deciding the width and boundaries of bins used. If the unfiltered data has a greater natural range than the filtered data, then it is very probable that these values will not be displayed. To gain a meaningful comparison with between filtered and unfiltered data, **all** of the optional values should be set in the histogram control panel **before** any filter(s) is/are applied.

Log *S* and Log *P* are presented as stacked histograms. In a stacked histogram, the overall height of the bar in a particular bin represents the total number of values in that bin, while the different colours that make up the bar show the proportion of the total population attributable to each subtype. In our case, for Log *P*, there are two data subtypes, namely data that was experimentally derived, and data that was calculated. The experimentally derived values are plotted in dark blue at the bottom, with the values values in light blue on top. The populations of both subtypes in a particular bin are summed to give the value printed at the top of the column.

Scatter Plots worksheet

The Scatter Plots worksheet can be used to produce scatter plots between any two whole or filtered datasets. The datasets are chosen from a selection given by one of two dropdown boxes lying along the two axes. If a filter is applied to one or more columns in the Data worksheet, then the filtered data points are coloured blue while the remaining points are coloured grey.

Part 2: Generating physico-chemical data

Apart from Log *S* and Log *P* values, which were sourced via VCCLAB, all names and physico-chemical properties were generated using the `cxcalc` and `molvconvert` commands, which are provided by local installations of the Marvin Beans part of the Marvin Suite of programs, developed by ChemAxon. Both `cxcalc` and `molvconvert` can be operated in batch mode via the command line of various operating systems. In order to use these commands, first download and install Marvin Beans from: http://www.chemaxon.com/download/marvin-suite/. At the time of writing, a licence was required to use the software, however, free licences were available for academic use.

Note for Windows Users: Several batch files have been written to semi-automate the commands described below. Please skip down to the ***Generating data for an individual compound*** and ***Generating data for multiple compounds*** sections for more information.

The commands and associated parameters that were used to generate physico-chemical properties in the Data worksheet are listed below (Table 1). The commands can be entered directly into the

command line. “Input.txt” is replaced with the name of a text file containing one or many SMILES strings (one entry per line).

Table 1. Commands and parameters used to generate physico-chemical/IUPAC Name Data

Property Description	Command Parameters[a]
Count of **total number of oxygen and nitrogen atoms able to accept hydrogen bond(s)** for the major protonation form of the major fragment at pH 7.4	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES acceptorcount --pH 7.4`
Count of **total number of atoms** including hydrogen of the major fragment	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES atomcount`
Count of **total number of non-aromatic atoms excluding hydrogen** of the major fragment	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES aliphaticatomcount`
Count of **total number of aromatic atoms excluding hydrogen** of the major fragment	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES aromaticatomcount Input.txt`
Count of **total number O–H and N–H bonds (hydrogen bond donor sites)** for the major protonation form of the major fragment at pH 7.4	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES donorsitecount --pH 7.4`
Calculate **formal charge** of the major protonation form of the major fragment at pH 7.4	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES formalcharge --pH 7.4`
Generate **chemical formula** including all fragments	`molconvert -g smiles Input.txt | cxcalc -i SMILES formula Input.txt`
Calculate **Log *D*** (as the arithmetic mean of three Log *D* estimates each generated using the ALOGP method trained with one of three different training sets for the major tautomer of the major protonation form of the major fragment) at multiple pH values ranging from 1.4 to 13.4	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES logd --lower 1.4 --upper 13.4 --step 1 --weights 1:1:1:0 --considertautomerization`
Calculate **molar mass** of all fragments	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES mass`
Generate **IUPAC names** including all fragments	`molconvert -g name -e ..UTF-8 Input.txt`
Calculate **polar surface area** for the major protonation form of the major fragment at pH 7.4	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES polarsurfacearea --pH 7.4`
Count of **total number of rotatable bonds** of the major fragment	`molconvert -g smiles -F Input.txt | cxcalc -i SMILES rotatablebondcount`
Generate **sortable chemical formula** including all fragments. To consider only the major fragment, add “ `--single true`” to the end of the command.	`molconvert -g smiles Input.txt | cxcalc -i SMILES sortableformula -d 3`

[a] These commands will display the results on screen. In order to output results to a file named Results.txt, insert `-o Result.txt` immediately after `cxcalc` in each command (except for IUPAC names, where `-o Results.txt` should be inserted immediately after `molconvert`)

Lastly, the internal help menu for `cxcalc` and `molvconvert` can be found by entering `cxcalc -h` or `molconvert -h`, respectively, into the command line. For help on a particular calculation, insert the calculation name before `-h`. e.g. `cxcalc acceptorcount -h` provides help regarding the `acceptorcount` function.

Generating data for an individual compound

A batch script named *single_compound_calculator.bat* can be used to generate certain physico-chemical properties for a single compound. This script applies the same parameters as were used to generate the data found in the Data worksheet.

NOTE: These scripts require the use of WINDOWS operating system.

To use the script, simply drag and drop onto the *single_compound_calculator.bat* icon a file containing a single molecule in one of the following formats: mol, rgf, sdf, rdf, csmol, csrgf, cssdf, csrdf, cml, cxsmiles, abbrevgroup, sybyl, mol2, pdb, xyz, inchi, SMILES, CAS RN, common names, IUPAC name or a peptide string. Alternatively, double-click the script icon and enter (by typing in or copy/paste) a molecule identifier, such as CAS number or a common name, at the prompt.

When the calculations are complete the results are displayed on screen (Figure 1).

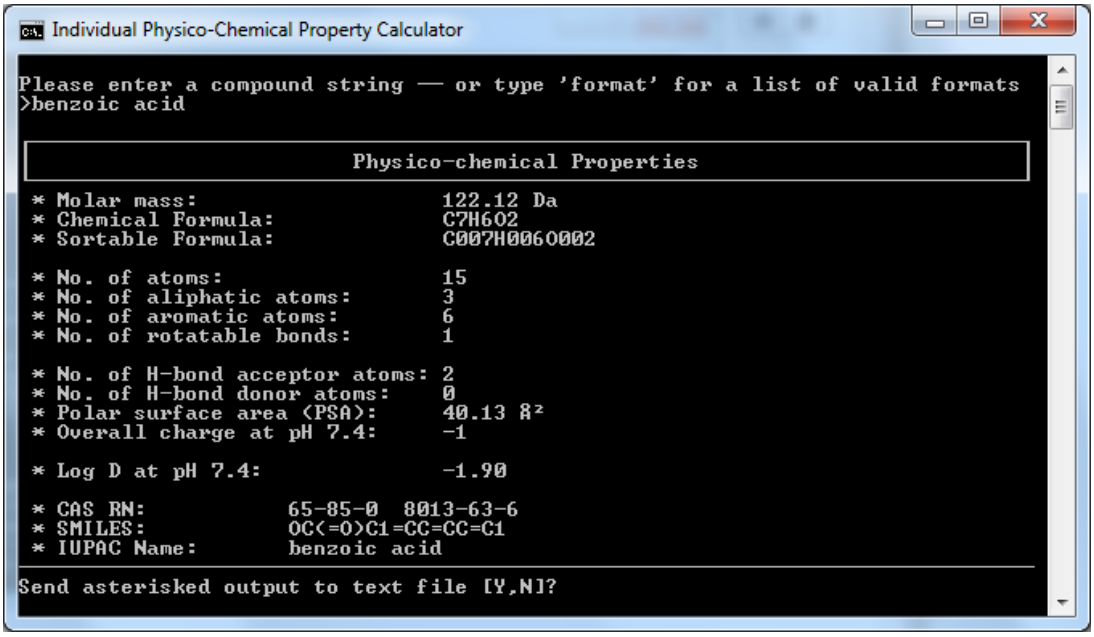

Figure 1. Example of on-screen display after a successful query for 'benzoic acid'

In cases where a multi-fragment compound is specified, the program will display appropriate data for both the entire molecule and the major fragment.

Finally, the user is given the option of saving output to a txt file, which may be imported into the Excel® workbook without further manipulation. An option is also given to generate png and svg imagery.

Generating data for multiple compounds

Three batch scripts are provided that can be used to generate images and physico-chemical properties either directly from SMILES strings or from CAS registry numbers (the latter are first converted to SMILES strings).

NOTE: These scripts require the use of WINDOWS operating system.

Initial Setup

Once Marvin Beans is successfully installed, decompress the three BATCH files (SMILES_generator.bat, Physicochem_calculator.bat and Image_generator.bat) and place them together in the same directory.

To generate images AND calculate physico-chemical data:

1. Prepare a text file containing CAS numbers (one entry per line).
2. Drag and drop your text file onto the SMILES_generator.bat icon.
3. You will be prompted for a unique foldername, herein denoted as {FOLDERNAME}. Enter this and follow the rest of the on-screen instructions.
4. The first task this script performs is to convert CAS to SMILES strings using an online lookup service. During this task, two output files are produced named {FOLDERNAME}_SMILES.txt, which contains a list of SMILES strings, and {FOLDERNAME}_CAS_SMILES.txt, which contains both CAS numbers and SMILES strings. If you get error messages, see the ERROR NOTE below.
5. Once the files containing the SMILES strings have been generated, you will be given the option to use this output to generate physico-chemical properties and images. If you choose "No", this can be done later by following the instructions below.

*****ERROR NOTE*****

The lookup service requires an active internet connection. Also, the lookup service used by this script can sometimes fail, even when a valid CAS number is provided. If this occurs, the offending CAS number(s) will be displayed on-screen. SMILES strings for these will need to be generated and added manually to the {FOLDERNAME}_SMILES.txt and {FOLDERNAME}_CAS_SMILES.txt output files. Once manually edited, these text files can be used directly to calculate properties and generate images via the instructions below.

To calculate ONLY Physico-chemical properties:

1. Prepare a text file containing SMILES strings (one entry per line). Alternatively, use the {FOLDERNAME}_SMILES.txt file generated above.
2. Drag and drop your text file onto the Physicochem_calculator.bat icon.
3. Follow the on-screen instructions.

To generate ONLY images:

1. In addition to SMILES strings, the input file also needs to provide a unique name for each image produced. Prepare a text file containing two columns, output filename in the first column and the SMILES strings in the second column. The first and second columns must be separated with a <TAB>. Alternatively, use the {FOLDERNAME}_CAS_SMILES.txt file generated above, which will use the CAS number as the filename.
2. Drag and drop your text file onto the Image_generator.bat icon.
3. Follow the on-screen instructions.