Version 3.0 for UNIX, Linux, and Windows systems

Technical Manual

SCHRÖDINGER, INC.

1500 SW First Ave. Suite 1180
Portland, OR 97201
Phone: (503) 299-1150
Fax: (503) 299-4532
E-mail: help@schrodinger.com

Table of Contents

Introduction

QikProp™ provides rapid predictions for physically-significant descriptors and pharmaceutically-relevant properties of organic molecules. Roughly 50,000 compounds can be processed per hour on an R10000 SGI or 1 GHz Pentium. Output includes the descriptors and predictions for

octanol/water partition coefficient (log Po/w)
aqueous solubility (log S)
brain/blood partition coefficient (log BB)
CNS activity
Caco-2 cell permeability (PCaco)
MDCK cell permeability (PMDCK)
% Human absorption
Lipinski Rule of Five
serum protein binding (log K'hsa)
HERG K^+ channel blockage
skin permeability (log K_p and J_m)
free energy of solvation in hexadecane
free energy of solvation in octanol
free energy of solvation in water
polarizability
ionization potential
electron affinity

The output includes a comma-separated file (*QP.CSV*) containing the descriptors and predictions; this can be imported into a spreadsheet program such as Excel or a statistical package such as JMP, and provides a basis for further QSPR or QSAR analyses and for design of compound libraries.

Incorporation of In-House Data

Users can develop regression equations for their custom data and have QikProp output the results. This is useful for properties or activities not covered by QikProp, or if the user has confidential data sets, or if the conditions for in-house measurements are different from those for the experimental data used in developing the QikProp regression equations.

To have QikProp output results from your custom QSPR/QSAR fits: (1) run QikProp for your compound library, (2) import the resultant *QP.CSV* file into your statistical package such as JMP and add a column with your property or activity data, (3) perform the regression analysis using the QikProp descriptors and predicted properties, and (4) enter the resultant coefficients for the descriptors in a copy of the *QPmyfits* file. Place it in the directory in which you are running

Qikprop. Subsequent execution of QikProp will output the results from your regression equation(s) in the printed output file *QP.out* as well as in *QP.CSV*.

Key New Features in QikProp 3.0

1. The molecules in the *Similar.CSV* file that are most similar to the input one are optionally retrieved. Similarity is determined by a comparison of 15 QikProp descriptors and computed properties; the most heavily weighted are molecular weight, numbers of hydrogen-bond donors and acceptors, rotatable bonds, QPlogPow, and ring desriptors. The number to be retrieved is set in *QPlimits*. *Similar.CSV* is a standard *QP.CSV* file; the provided one is for ca. 1700 known oral drugs, as described in J. R. Proudfoot, *BMCL* **15**, 1087-1090 (2005). The user can replace the provided file with an alternative that he/she creates by running *QikProp* on an sd file. The file is kept in the *QikProp/CSVfiles* directory; it is copied to the execution directory as long as a *Similar.CSV* does not already exist there. *Similar.CSV* is not over-written, so different ones could be used in different directories. Note: if you are processing a large number of molecules in an sd file, you may want to set the number of similar molecules retrieved to zero as this feature does slow processing since the *Similar.CSV* file is read for each input structure.

If you are running in batch mode, you need to use the new xQPROP and xQPCLN scripts for version 3.0 as well as the new QPlimits and QPmyfits files.

2. Several new structural descriptors were added to facilitate the similarity matching. They are output in *QP.CSV*: the total number of atoms in rings, the number of atoms in 3 and 4-membered rings, the number of atoms in 5 and 6-membered rings, the number of atoms in rings that can not conjugate (e.g., 3 for cyclopentene), and the number of non-hydrogen atoms.

3. Similarity information is also now used to improve the predictions for octanol/water log P (QPlogPow) and aqueous solubility (QPlogS). The new models are highly accurate. If the similarity between the input molecule and the closest one in the database is greater than 90%, the QikProp prediction is adjusted by the database information until it equals the database value at 100% similarity. The unadjusted and adjusted values are output in *QP.out*, while only the adjusted values are output to *QP.CSV* for QPlogPow and QPlogS.

4. The models for % Human Absorption have been improved. The influence of active transport, e.g., by peptide transporters, is not considered. The qualitative model (yielding low, medium and high rankings) considers the results for log S, PCaco, log Pow, rotatable bonds, and number of primary metabolites in a decision tree.

5. The HERG database has been expanded to 102 molecules and the model has been enhanced.

6. Processing of sd files (mol files with multiple molecules) will now optionally append all information from data fields in the output *QP.CSV* file. The data can be in any order and the maximum number of unique data fields is 50. This option is requested in the *QPlimits* file. For example, one might have a database of solubilities stored as an sd file with a solubility data entry for each molecule. Running QikProp on this sd file will then yield a QP.CSV in the usual way, but with the solubilities from the sd file appended as the last column. This facilitates comparisons between the QikProp predictions and database.

Key New Features in QikProp 2.3

1. Two predictions are now made for the human oral absorption of a compound into the GI tract. A qualitative analysis is provided based primarily on consideration of the computed Caco-2 cell permeability, solubility, number of primary metabolites, and rotatable bonds yielding rankings of low, medium and high. A quantitative estimate is also provided through a regression equation that uses these terms. The two predictions are designated HumanAbs and QP%HumanAbs in the output *QP.CSV* file.

2. Prediction of Caco-2 cell permeability has been enhanced by development of a combined dataset from Affymax, Boehringer-Ingelheim, Astra, and Pharmacia. The total number of unique compounds is 150. The diversity is very good, and the model has been significantly improved. The databases and regression equations for many other predicted properties have also been updated.

3. Polar surface area (PSA) is now output. This is the van der Waals surface area for nitrogens, oxygens, and attached hydrogens. It is computed using a solvent probe radius of zero, while solvent-accessible surface areas are computed with a water-sized probe radius of 1.4 Å. The corresponding SASA term is the QikProp FISA. In developing the regression equations for QikProp 2.3, both PSA and FISA were considered. In all cases in which such a term was relevant, FISA yielded the regression with lower error. The PSA results from QikProp 2.3 agree well with literature values (K. Palm et al., *J. Med. Chem.* **41**, 5382 (1998)), e.g., for atenolol 92.7 *vs* 90.0 $Å^2$, and for alprenolol 37.8 vs 37.1 $Å^2$.

4. The Lipinski rule of five is evaluated and the number of violations is output. In the *QP.CSV* file, this is designated RuleOf5. Though the analyses by QikProp are far more extensive, some users expressed interest in this addition as a descriptor and potential filter. As a component of RuleOf5, the sum of nitrogens and oxygens is also output. References: C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev*. **23**, 3 (1997); C. A. Lipinski, *J. Pharmacol. Toxicol. Methods* **44**, 235 (2000).

5. Two predictions for aqueous solubility are made by QikProp; QPlogS is the original scale, however, it shows some conformation dependence. QikProp regressions have been developed using extended structures, which are typically provided by 2D to 3D conversion programs like Concord and Corina. An alternative prediction is also made, CIlogS, that is conformation independent. It is now output in *QP.CSV* directly after QPlogS. The two values are generally within a few tenths of a log unit and provide a range for the expected solubility.

6. Overall, the number of items output to *QP.CSV* has increased from 37 in QikProp 2.1/2.2 to 45 in QikProp 2.3. The new items are CIlogS, HumanAbs, QP%HumanAbs, SAfluorine (SASA for fluorines), SAamideO (SASA for amide oxygens), PSA, #NandO, and Ruleof5.

7. The *QPlimits* and *QPmyfits* files have been changed to reflect the changes in output descriptors and properties. *QPlimits* now includes PSA, and *QPmyfits* now covers 40 descriptors. You need to use the *QPlimits* and *QPmyfits* files that are provided with QikProp 2.3; older versions will not work. *QPlimits* allows user definition of file names and property limits. *QPmyfits* allows the user to define QSPRs based on QikProp descriptors.

Key Features from Earlier Releases

There are four output files generated by QikProp:
QP.out is the summary output.
QP.CSV is the output comma-separated file.
QPwarning is the output file for any error messages.
QPSA.out is the additional output including SASA breakdown. Output of the *QPSA.out* file can be suppressed by designation in the *QPlimits* file. This might be desirable if one is processing a very large number of compounds.

Limits:
The limit on number of input atoms in a molecule in normal mode is 250 (150 H, 100 non-H); the limit on total number of atoms in FAST mode is 1000. The maximum length of a molecule name is 40 characters.

QikProp 2.0 – 2.2

1. FAST processing mode - 150,000 compounds per hour on a 1 GHz Pentium
A FAST processing mode has been made optional. This is selected in the *QPlimits* file. If chosen, the PM3 single-point calculation is not performed and one does not obtain the dipole moment (μ), μ^2/volume, ionization potential, and electron affinity as output descriptors (they are 0.0 in the *QP.CSV* file). The dipole terms were used in many of the QSPR equations. So, the equations were refit without these terms, fortunately with negligible loss of accuracy. However, in FAST mode QikProp executes ca. 50 times faster than in NORMAL mode for typical drug-like molecules. See *QikProp/notes/timing* for more information.

2. FAST processing mode - allows processing of ions and most atom types
In FAST mode, most ions will be processed. A warning appears in the *QPwarning* file, the output descriptors are correct, and the output properties are invalid - user beware. In FAST mode, atoms with any atomic number are allowed, while in normal mode the allowed PM3 atoms remain H, C, N, O, F, Al, Si, P, S, Cl, Br, I. Note that the descriptors will generally be valid, but the regression equations for the computed properties have not been specifically parameterized for other than the PM3 atom types. A warning is issued in *QPwarning* for non-PM3 atom types.

An input structure should be a single molecular entity, e.g., not an ion pair or two molecules. However, if a non-compliant structure is input, it will be processed, a warning will be issued in *QPwarning*, and the output properties are not valid.

3. Cell Permeability Predictions
For Caco-2 cell permeabilities (PCaco), a single predicted value, QPPCaco, replaced the two former quantities from the separate Boehringer and Affymax data sets. The predicted values correspond to the experimental conditions used by Affymax; Boehringer values are typically five times smaller. Predictions are also made for MDCK cell permeabilities; the results are designated QPPMDCK in the *QP.CSV* file.

4. Prediction of HERG K^+ Channel Blockage

A model has been added for prediction of IC_{50} values for blockage of HERG K^+ channels. Such blockage may be associated with cardiotoxicity and *Torsades de Pointes*; it is reflected in prolongation of the QT interval in electrocardiograms. The predicted log IC_{50} is output as QPlogHERG.

5. Linux

The default Linux/IRIX distribution from W. Jorgensen is Linux. IRIX executables are in the subdirectory IRIX.

6. mol2 File Input

QikProp accepts the input of molecular structures as Tripos mol2 files as well as the previous choices, MDL mol files, PDB files, and BOSS Z-matrix files. The only items read from a mol2 file are the name of the molecule, number of atoms, atomic coordinates, and atom types. All other information is ignored. Multiple molecules can be processed (1) by input of a file that contains a list of the file names for the individual molecules, (2) by separating the molecules with $$$$ entries in a single .mol file, or (3) by starting each molecule entry in a mol2 file with an @<TRIPOS>MOLECULE record.

7. Miscellaneous

a. The subdirectory logw is provided. This directory contains the database of free energies of hydration and free energies of solvation in hexadecane that were used in developing the corresponding QSPRs in QikProp 2.1.001. The prior databases have been much expanded to include data for 421 compounds from M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. Leo, and R. S.Taft, *J. Chem. Soc. Perkin Trans. 2*, 1777-1791 (1994).

b. A utility program "extract" has been provided to extract data from a .mol file and output it to a CSV file that can be read by JMP, Excel, etc. This is useful for developing QSAR or QSPR equations where the property or activity value is stored in the .mol file entry. E.g., if log Lw data were present for each entry in a file mols.mol with the following form

> <log.Lw>

Then one can execute extract to extract each log Lw entry. E.g., for Windows

```
C:\QikProp> extract
 Enter the name of the input file: mols.mol
 Enter the name of the output file:
  Output file is called out.
 Enter field to extract, e.g. MOLNAME: log.Lw
```

c. For augmented .mol files, if there is a > <MOLNAME> entry, QikProp reads the following line for the molecule name that is output in all files. If there is a > <MDLNUMBER> or > <MOLCODE> entry, they are output after the molecule name in the *QP.out* file. See the file *QikProp/molfiles/flucon.mol* for an example.

d. Ca. 30 types of reactive functional groups, which would cause false positives in HTS assays, are noted, if present, in the *QP.out* file. A count of the reactive functional groups is output in *QP.CSV* for use as a filter for library design.

e. In the "QP Breakdown" at the bottom of the *QP.out* file, a < marker is attached to any descriptor that exceeds the range of the training sets for log Po/w, log S, log PMDCK, and log BB.

f. Globularity descriptor - for a reference, see A. Y. Meyer, *Chem. Soc. Rev.* **15**, 449-475 (1986), page 457. This is basically the ratio of the surface area for a sphere with the same volume as the molecule over the true surface area. Glob = 1 for a sphere and less for real molecules.

g. Commas in the molecule name are replaced by semicolons in the output comma-separated file. Even though the name is placed in quotes, some programs will still take any comma inside the quotes as a field delimiter. This avoids that problem.

System Requirements

SGI

- R4*00, R5000, R8000, R10000, or R12000 processor
- IRIX 6.5.2 or later

Linux

- Pentium, Celeron, or AMD processors
- RedHat 5.2 or later, S.u.S.E. 6.2 or later
- Linux-supported network card with a configured network interface

Windows

- Pentium, Celeron, or AMD processors
- Windows 98, 2000, NT, or XP

Sun

- SunOS 5.7 (Solaris 7)

Installation

The QikProp files are normally provided by ftp or on a CD. Instructions are given for installation of the files in each case.

For Windows, if you already have a directory *c:\QikProp*, rename or delete it. A self-installer is provided; just double-click on the icon. The files are installed in the directory *c:\QikProp*

For UNIX/Linux the path to the top QikProp directory needs to be specified in the *.cshrc* file with a command like
setenv QPdir ~/QikProp
The key files are placed in the QPdir disk directory, normally called QikProp.

There is a *00readme* file that contains useful information that should be read.

Launching and Executing QikProp

UNIX/Linux

- Copy the *xQPROP* script from QPdir to the directory that contains the structure files you wish to evaluate, or alternatively, copy it to usr/local/bin so that you may use the script to launch QikProp from any directory.

- Once you have set the QPdir environment variable and copied the *xQPROP* script, launch QikProp using the command:

  `xQPROP <filename>`

  where <filename> is the name and path of the file containing the structure(s) whose properties you want to predict. QikProp generates two files - *QPSA.out* and *QP.out* - for each structure, and it appends each structure's property data to a third file, *QP.CSV*. Additionally, if QikProp has difficulty processing the structure, it will generate a fourth file, named *QPwarning*, for that structure. See the *Output* section below for more information about these file types.

- If you would like to preserve the data in an output file, use the `mv` command to save the file under a different name. If you do not explicitly save the output files, QikProp will overwrite them when it evaluates the next compound. The exception to this is the *QP.CSV* file. QikProp generates only one *QP.CSV* file, and it appends the property predictions for all subsequently evaluated compounds to this same file. If you would like to empty the *QP.CSV* file, use the `rm` command to delete the file. The next time you run QikProp, it will initiate a new *QP.CSV* file. You can also delete all files generated by QikProp with the script xQPCLN in QPdir.

Windows

Command Prompt Window

- In a Microsoft **Command Prompt** (CMD) window, copy the script *xQPROP.bat* to your current directory or to a directory declared in the execution path, and enter the command:

  `xQPROP <file>`

  where <file> is the name of the 3D structure(s) file that you wish to evaluate. This is the normal procedure when processing large numbers of molecules.

QikProp Graphical User Interface

- Alternatively, launch the **QikProp Viewer**, the graphical interface for QikProp, by double-clicking on the *QView.exe* icon in the QikProp folder (note that your Windows setup may

prevent the .exe suffix from being displayed). You may wish to make a copy of its shortcut for your desktop.

- Click on the **Run Molecule(s)** button on the **QikProp Viewer** panel. The **Open & Run** file selector will appear. If necessary, change the file selector's filter to correspond to the file type you wish to display. Navigate to the file that contains the structure whose properties you wish to predict, and select that file in the list.

- To perform property predictions for the structure in the file you have designated, click on the **Run** button, located on the **Open & Run** file selector. QikProp will process the results and return you to the **QikProp Viewer** panel.

- To view output files for the last structure evaluated, click on the **View Output Files** button on the **QikProp Viewer** panel. QikProp makes entries in two files - *QPSA.out* and *QP.out* - for every structure, and it appends the results for each structure to a third file, *QP.CSV*. Additionally, if QikProp has difficulty processing a structure, it generates a fourth file, named *QPwarning*, for that structure. See the *Output* section below for more information about these file types. When you click on the **View Output Files** button, a window displaying the most recently generated *QP.out* file will appear. You can print any of the output files with the **Print** button; e,g, you can easily get a hardcopy of the *QP.out* file for your lab notebook.

- If you wish to view or print other output files for the same structure, select the desired file type from the **Select Output File:** option menu.

- If you would like to save any output data, click on the **Save...** button, and use the **Save File** file selector to designate the directory and file to which you would like to save the data. If you do not explicitly save the *QP.out, QPSA.out,* or *QPwarning* file for a structure, QikProp will overwrite them when you evaluate the next compound. The exception to this is the *QP.CSV* file. QikProp generates only one *QP.CSV* file, and it appends the property predictions for all subsequently evaluated compounds to this same file.

- To empty the *QP.CSV* file, either (a) use the **Clear QP.CSV** button or (b) exit QikProp and delete the file. The next time you run QikProp, the program will initiate a new *QP.CSV* file. If you would like to initiate a new *QP.CSV* file, but save the existing one, **Save** the old *QP.CSV* file to a different file.

- To view a structure in a mol file, open the RasMol viewing program. Do this by either: 1) clicking on the **Run Molecule(s)** button on the **QikProp Viewer** panel, selecting a file using the **Open & Run** file selector, and then clicking on the **Open in RasMol** button, or 2) double-clicking on the *raswin.exe* icon in the **QikProp** folder window and using the RasMol **Select Molecular Coordinate File** file selector to open a file and display the enclosed structure.

- Note that you can navigate to any directory with **QikProp Viewer** and process structures in that the directory. The output *QP.out, QPSA.out, QP.CSV,* and *QPwarning* files will be in that directory.

Input

Molecular Structures

The key input is a file with the 3D (*note not 2D*!) structure of a molecule, i.e., x,y,z coordinates and atomic numbers. Within these files, *hydrogens on input structures must be explicit*, and, with the exception of mol2 and $$$$-delimited MDL SD mol files, each file must contain only one structure. Molecules need to be input with their functional groups in their neutral state, e.g., amines should not be protonated and acids should not be deprotonated. The program will make adjustments for the proper protonation state in its predictions.

Three file formats are accepted:

- MDL mol files, also known as 3D SD files
  The directory *QPdir/molfiles* contains sample MDL SD mol files for 30 common drugs.
- Tripos mol2 files.
  An example, *zoloft.mol2*, is provided in *QPdir*.
- PDB files
  An example is in *QPdir*.
- BOSS/MCPRO Z-matrix files
  An example is in *QPdir*.

For mol, mol2 and PDB files, the allowed atom types in normal mode are H, C, N, O, F, Al, Si, P, S, Cl, Br, and I. Fast processing mode permits almost all atom types.

Please check that the format of your mol2 or mol files is the same as in the provided examples. QikProp uses mol format as specified by MDL in *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244-255. I.e., from page 252:
Line 1 – Molecule name. (First 40 characters are retained by QikProp for printing.)
Line 2 - Information on the origin of the file. (Ignored by QikProp.)
Line 3 – Blank
Line 4 – Atom and bond counts.
Atom entries (x, y, z in Å, atomic symbol)
Bond entries. (Ignored by QikProp.)
Any additional entries. (Ignored by QikProp except as noted below.)
$$$$

Some additional entries to standard .mol files are processed as follows. If there is a
> <MOLNAME> entry, QikProp reads the following line for the molecule name that is output in all files. If there is a > <MDLNUMBER> or > <MOLCODE> entry, they are output after the molecule name in the *QP.out* file. See the file *QikProp/molfiles/flucon.mol* for an example.

Processing Multiple Molecules – As a Concatenated mol File (an sd file):
The mol files for multiple molecules can be concatenated into a file, e.g., see **QPdir/molfiles***/many.mol*. The entry for each molecule needs to end with the usual $$$$ delimeter. Then, to process the molecules sequentially,

`del QP.CSV` (Windows; to restart the *QP.CSV*, if you wish to clear its contents)

```
rm QP.CSV     (Unix/Linux)
xQPROP many.mol
```

Processing Multiple Molecules – As a Concatenated mol2 File:
The mol2 files for multiple molecules can also be concatenated in a file. Entries for each molecule should begin with an @<TRIPOS> MOLECULE record.

Processing Multiple Molecules in Separate Files:
Alternatively, to process multiple molecules, one can just place the names of their individual files in the input file. This works for mol, mol2, pdb, and BOSS Z-matrix files. This is illustrated with the file *many.names* in the *QPdir/molfiles* directory. Then, the commands

del QP.CSV (Windows)
rm QP.CSV (Unix/Linux)
xQPROP many.names

generate the output for each entry in *many.names*. The maximum length of a file name is 80 characters; characters after the 80th are ignored.

- **UNIX/Linux or Windows:** The program reads the input file and automatically detects the input file type, i.e., mol, mol2, pdb, BOSS Z-matrix, or a file list. File extensions such as .mol are not needed.

Custom Fits

Users can have QikProp output results from their custom QSPR/QSAR fits by providing the coefficients for the QikProp descriptors and predicted properties in the file *QPmyfits*. Please see *QPdir/QPmyfits* or *c:\QikProp\QPmyfits* (Windows) Only linear equations are implemented. The results are output to the *QP.out*, *QPSA.out*, and *QP.CSV* files. If you use *QPmyfits*, a copy of your *QPmyfits* file needs to be in the directory that contains the input structure file. The maximum number of custom fits is four.

QPlimits File

If you wish to change output file names, the ranges for observed descriptors and properties, the number of similar molecules to output, and to designate FAST or normal processing mode, you can (a) edit the *QPlimits* file in *QPdir* or in *c:\QikProp*, if you want the changes to apply for all executions of QikProp. Or, (b) copy the default *QPlimits* file from *QPdir* to your directory and edit it. The changes will then only apply for executions from that directory. Executing QikProp in a given directory will not overwrite a *QPlimits* file, if one already exists in the directory. The default *QPlimits* file from *QPdir* is only used when an alternative is not specified.

Similar.CSV

The molecules in the *Similar.CSV* file that are most similar to the input one are optionally retrieved. The number to be retrieved is set in *QPlimits*. *Similar.CSV* is a standard *QP.CSV* file; the provided one is for ca. 1700 known oral drugs, as described in J. Proudfoot, BMCL (2004). The user can replace the provided file with an alternative that he/she creates by running *QikProp* on an sd file. The file is kept in the *QikProp/CSVfiles* directory; it is copied to the execution directory as long as

a *Similar.CSV* does not already exist there. *Similar.CSV* is not over-written, so different ones could be used in different directories. Note: if you are processing a large number of molecules in an sd file, you may want to set the number of similar molecules retrieved to zero as this feature does slow processing since the *Similar.CSV* file is read for each input structure.

Output

There are 4 output files:

(1) The *QP.out* file lists the output descriptors and predicted properties. Comparisons are made of the compound's properties and those of typical drugs.

(2) The *QPSA.out* file contains additional information, specifically the covalent-neighbors list, atom-type assignments, and surface area analysis. Output of this file can be suppressed by making the appropriate designation at the beginning of the *QPlimits* file.

(3) The *QP.CSV* file contains the key output (descriptors and predicted properties) on one comma-separated line per molecule. Each time the program is executed, the results are added to the end of *QP.CSV*. If you want to start a new *QP.CSV* file, you need to delete the old one.

The *QP.CSV* file is suitable for input to a spreadsheet program such as Excel or an analysis program such as JMP. The QikProp descriptors and predictions, once imported into such programs, can then be used as a basis for other QSPR or QSAR analyses by adding a column of property or activity data.

In order to avoid potential input problems with programs that read *QP.CSV*, any commas in the molecule name are replaced by semicolons and blanks are removed.

(4) The *QPwarning* file - warnings about potential problems with the input structures are in *QPwarning*. This flags, for example, input ions, structures with multiple fragments, empty input files, and missing hydrogens. If multiple structures are being processed, any structures with fatal errors, e.g., a disallowed atom type, are normally skipped over and an entry is made in *QP.CSV* stating that the processing of the molecule failed. If QikProp encounters no problems with the processed structure(s), the *QPwarning* file is not generated. After executing QikProp for multiple structures, one should (a) display *QPwarning*, and (b) search for any occurrences of "empty" or "failed" in *QP.CSV*. In Unix/Linux:

```
cat QPwarning
grep empty QP.CSV
grep failed QP.CSV
```

Quality of Results

Experimental results for ca. 1000 compounds including ca. 700 drugs and related heterocycles were used in developing QikProp. The following table summarizes the fits for QikProp Version 3.0.

Statistics on QikProp 3.0 Fits to Experimental Data

Property	N	r^2	rms error	MW range
polarizability – $Å^3$	78	0.97	1.04	20-200
log P (hexadecane/gas)	392	0.93	0.37	20-200
log P (octanol/gas)	118	0.91	0.59	20-250
log P (water/gas)	421	0.93	0.58	20-250
log P (octanol/water)*	485	0.92	0.54	20-916
log S (water/solid)*	426	0.90	0.63	20-765
log S - conformation independent	470	0.88	0.70	20-765
log K'hsa (serum protein binding)	90	0.82	0.25	130-765
log IC_{50} for HERG blockage	102	0.72	0.74	160-840
log BB (brain/blood)	127	0.70	0.42	20-614
log PCaco-2	150	0.71	0.56	60-631
log PMDCK	52	0.73	0.57	130-430
% Human Absorption	102	0.79	14.9	60-750
log Kp (skin permeability)	61	0.80	0.67	20-600

* without the accuracy boosting from QP 3.0 similarity analyses.

In the *QP.out* file for the log P (octanol/water), log S, log P MDCK, and log BB output, if the value for a utilized descriptor exceeds the range for the experimental training set, it is flagged. Otherwise, the molecular weight ranges above indicate the domain of validated applicability of the regression equations.

The following tables illustrate results from QikProp 2.1 for test sets of molecules that were not in the QikProp 2.1 training set. The 3D structures were obtained directly from SciFinder, then to gauge the effect of optimization, they were optimized with the BOSS program using the OPLS-AA force field and CM1P charges. Both sets of structures were processed by QikProp 2.1. In developing QikProp 3.0, log S predictions were made for a test set of 46 drugs; the average error was 0.5 log unit. A test was also made for log Pow for ten drugs; the average error was 0.4 log unit.

Please also see the **QPdir/notes** directory for notes on various topics such as the conformational dependence of the results and details of the regressions for the PCaco fits.

Table 1. Log S Results for a Test Set of Molecules

Molecule	MW	exptl log S	QPlogS[a]	QPlogS[b]
N-methylmorpholine	101.1	1.00	1.85	1.66
2,5-dimethylpiperazine	114.2	0.49	1.47	1.36
Isoniazid	137.1	0.01	-0.79	-0.85
3,3-dimethyl-1-butanol	102.2	-0.50	-0.81	-0.88
3-methyl-3-hexanol	116.2	-1.00	-1.99	-1.53
Bis-(2-chloroethyl) sulfone	191.1	-1.50	-1.48	-1.42
Minoxidil	209.3	-1.98	-2.07	-1.92
2,4-D	221.0	-2.51	-2.21	-2.78
Heptabarbital	250.3	-3.00	-2.29	-2.30
Sulfadiazine	250.3	-3.51	-2.08	-2.06
Terbutyrne	241.4	-4.00	-2.89	-3.57
1,2,4-tribromobenzene	314.8	-4.50	-4.00	-4.05
Quinonamid	318.5	-5.03	-3.34	-3.89
Benfluralin	335.3	-5.53	-4.08	-4.49
Fluoranthene	202.3	-6.00	-6.44	-6.61
o,p'-DDD	320.0	-6.51	-6.53	-6.90
7,12-dimethylbenz(a)anthracene	256.3	-7.02	-6.40	-6.90
2,2',3,4,6-PCB	326.4	-7.43	-8.04	-7.71
benzo(j)fluoranthene	252.3	-8.00	-8.24	-8.63
2,2',4,4',5,5'-PCB	360.9	-8.56	-10.75	-8.50
q^2			0.89	0.95
Rms			1.01	0.70
ave. error			0.77	0.55

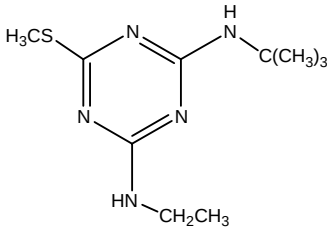
[a] Using structures from SciFinder. [b] Using optimized structures.

Table 2. Log $P_{o/w}$ Results for a Test Set of Molecules

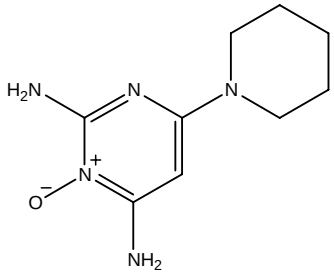
Molecule	MW	exptl log $P_{o/w}$	QPlogPo/w[a]	QPlogPo/w[b]
Isoniazid	137.1	-0.70	-0.24	-0.04
Nicotinamide	122.1	-0.37	-0.13	-0.20
N-methylmorpholine	101.1	-0.33	-0.65	-0.52
Sulfadiazine	250.3	-0.09	-0.06	0.02
Pyrazole	68.1	0.26	0.21	0.20
Sulfamethiazole	270.3	0.54	-0.40	-0.39
Minoxidil	209.3	1.24	1.00	1.13
Heptabarbital	250.3	2.03	0.82	0.86
2,4-D	221.0	2.81	2.03	2.27
Ethisterone	312.5	3.11	3.78	4.35
Pericyazine	365.5	3.52	3.19	3.34
Terbutyrne	241.4	3.74	2.52	2.82
Captafol	349.1	3.80	2.83	3.13
Doxepin	279.4	4.29	3.41	4.00
Biquinoline	256.3	4.31	4.25	4.35
Prochlorperazine	373.9	4.88	3.82	3.79
Fluoranthene	202.3	5.16	5.03	5.13
Benfluralin	335.3	5.29	3.99	4.76
Amitraz	293.4	5.50	5.60	5.86
7,12-dimethylbenz(a)anthracene	256.3	5.80	5.28	5.63
q^2			0.86	0.92
Rms			0.84	0.65
Ave. error			0.63	0.47

[a] Using structures from SciFinder. [b] Using optimized structures.

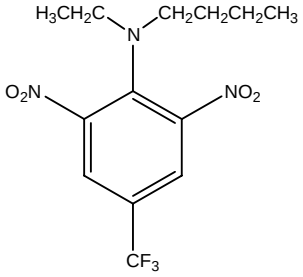
Some of the structures from the test sets are illustrated below:



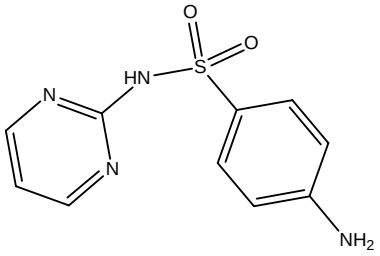
terbutyrne



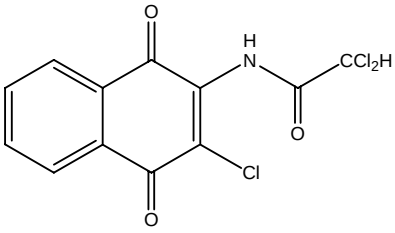
minoxidil



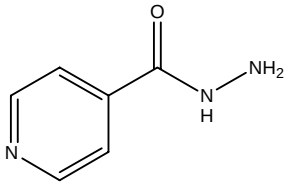
benfluralin



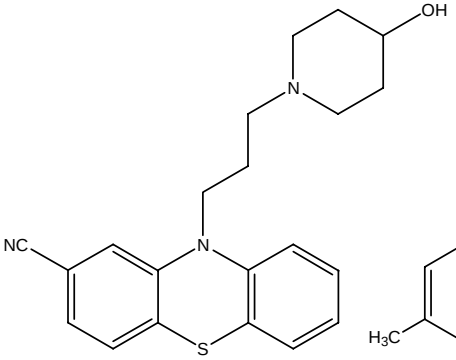
sulfadiazine



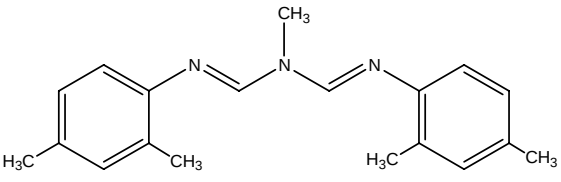
quinonamid



isoniazid



pericyazine



amitraz

QikProp References

"Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water", E. M. Duffy & W. L. Jorgensen, *J. Am. Chem. Soc*., **122**, 2878-2888 (2000).

"Prediction of Drug Solubility from Monte Carlo Simulations", W. L. Jorgensen and E. M. Duffy, *Bioorg. Med. Chem. Lett*., **10**, 1155-1158 (2000).

"Prediction of Drug Solubility from Structure", W. L. Jorgensen and E. M. Duffy, *Advanced Drug Delivery Reviews*, **54**, 355-366 (2002).

"The Many Roles of Computation in Drug Discovery", W. L. Jorgensen, *Science*, **303**, 1813-1818 (2004).

Data Sources and QikProp vs Experimental Results

Plots of QikProp predictions vs experiment are provided below and in the *plots* directory. The sources of the experimental **log P** and **log S** values are described in the above references. Additional data were kindly provided by Pharmacia Inc.

The databases of **free energies of hydration** and **free energies of solvation in hexadecane** are from M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. Leo, and R. S.Taft, *J. Chem. Soc. Perkin Trans. 2*, 1777-1791 (1994).

The **log BB** values are mostly listed in J. M. Luco, *J. Chem. Inf. Comput. Sci.*, **39**, 396-404 (1999), and D. Pan et al., *J. Chem. Inf. Comput. Sci.*, **44**, 2083-2098 (2004). The predicted **CNS activities** are based largely on log BB with some adjustments and correspond well with the CNS activities reported in Ajay, G.W. Bemis & M.A. Murcko, *J. Med. Chem.*, **42**, 4942-4951 (1999), Table 1 – Supporting Information. CNS activity is predicted on a -2 (--) to +2 (++) scale. This provides a filter for library design with or without CNS activity.

The **Caco-2 cell permeabilities** are from Boehringer-Ingelheim (Yazdanian et al., *Pharm. Res.,* **15**, 1490-1494 (1998) with minor additions; *Bioorg. Med. Chem. Lett*. **13**, 719-722 (2003)), Affymax (Irvine et al., *J. Pharm. Sci.*, **88**, 28-33 (1999)), Astra-Zeneca (Stenberg et al., *J. Med. Chem.*, 44, 1927 (2001)), and Pharmacia (A. R. Hilgers et al., *Pharm. Res.*, **20**, 1149-1155 (2003)). The experimental data have been put on a single scale consistent with the Affymax data. Thus, the predicted values correspond to the experimental conditions used by Affymax. The Astra-Zeneca and Pharmacia values are consistent with the Affymax data; the B-I results are generally five times smaller. The number of unique molecules used for the Caco-2 fit is 150 (gabapentin, lisinopril, and BI35 have been excluded). The **MDCK** (Madin-Darby Canine Kidney) cell permeabilities are also from the Affymax paper; 52 data points were used in the fit. Please see the file *QPdir/notes/Caco* and *QPdir/notes/CacoMDCK* for more information.

The data for **% human absorption** (%HA) into the GI tract are from the above cell-permeability papers and from Y. H. Zhao et al., *Pharm. Res*. **19**, 1446-1457 (2002). %HA is not the same as bioavailability. The latter refers to the % of compound that gets into the blood stream and includes first-pass metabolism. %HA is normally the upper limit for % bioavailability. All data of this type is noisy since it can depend on many details of formulation, administration, active transport, and the subjects. The 102 compounds that have been selected for development of the *QikProp* model have well-distributed %HA values. Lovastatin is an extreme outlier, with high predicted and low observed %HA. It and several compounds that are known to be actively transported were excluded, namely, amoxicillin, cefalexin, and loracarbef. Also, if all available data are used there are too many high %HA values and resultant models become biased towards high %HA predictions.

The experimental IC_{50} values for blockage of mammalian **HERG** K^+ channels are from:
(1) Cavalli et al., *J. Med. Chem.*, **45**, 3844-3853 (2002)
(2) De Ponti et al., *Eur. J. Clin. Pharmacol*., **57**, 185-209 (2001)
(3) http://www.fenichel.net
(4) R. A. Pearlstein et al., Bioorg. Med. Chem. Lett., 13, 1829-1835 (2003).
(5) G. M. Keseru, Bioorg. Med. Chem. Lett., 13, 2773-2775 (2003).
There is significant scatter in the experimental data amounting to uncertainties of factors of 2 – 50. The database was expanded from 71 molecules for *QikProp* 2.3 to 102 molecules for *QikProp* 3.0. Drugs that have been withdrawn owing to QT-prolongation problems also exhibit a large range of IC_{50} values, as shown in the following Table. Thus, the allowable limit for IC_{50}s depends on the class of compounds, dosage, and bioavailability. In general, it is found that most drugs associated with torsade de pointes are also associated with HERG K^+ channel blockage at concentrations close to the free plasma concentration found in clinical use (W. S. Redfern et al., *Cardio. Res.*, **58**, 32-45 (2003)).

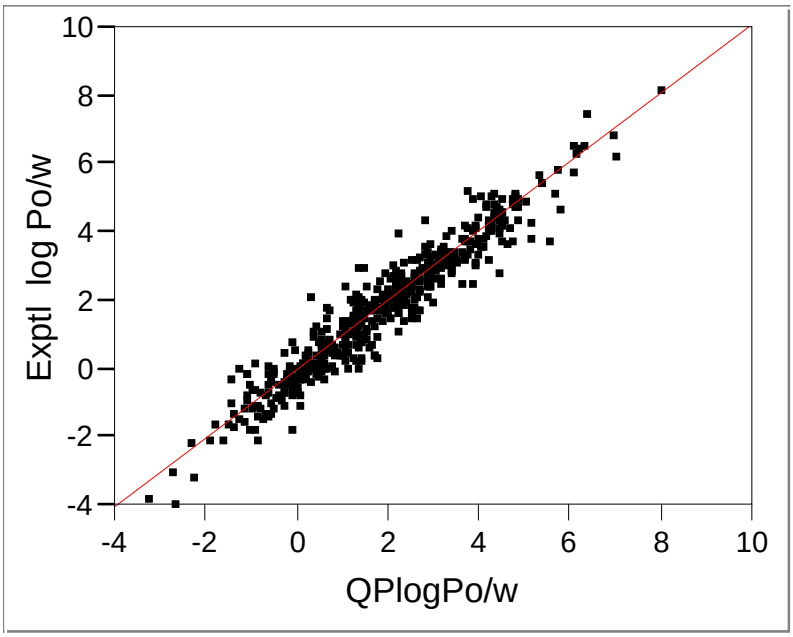
HERG Channel Blockage

Drug	log IC_{50} Exptl.	log IC_{50} QikProp 3.0	Status
astemizole	-8.0	-8.0	*withdrawn*
Cisapride	-7.4	-7.3	*withdrawn*
clarithromycin	-4.2	-5.4	warning
Dofetilide	-8.0	-7.2	warning
domperidone	-6.8	-7.0	warning
droperidol	-7.5	-7.3	*withdrawn*
grepafloxacin	-4.3	-3.3	*withdrawn*
halofantrine	-6.7	-6.9	warning
Ibutilide	-8.0	-6.9	warning
lidoflazine		-8.4	*withdrawn*
Pimozide	-7.3	-6.6	warning
Sotalol		-5.6	*withdrawn*
Sertindole	-8.0	-7.1	*withdrawn*
sparfloxacin		-3.5	warning
terfenadine	-7.3	-8.5	*withdrawn*
Terodiline		-6.3	*withdrawn*
thioridazine	-6.4	-6.5	*withdrawn*

Prediction of the **skin permeability** coefficient, Kp (cm/hr), as log Kp is provided along with the maximum rate of transport across skin, Jm (mg/cm^2-hr). These quantities refer to an aqueous solution of the solute in contact with human skin. The QP regression equation has been fit to data noted in R. O. Potts & R. H. Guy, Pharm. Res. 9, 663 (1992); 12, 1628 (1995). Please see *Qpdir/notes/skinKp* for more information.

The log K'hsa data for **binding to human serum albumin** are from G. Colmenarejo et al., *J . Med. Chem*., **44**, 4370-4378 (2001).

log P(octanol/water)



Linear Fit

Linear Fit

log Po/w = 0.0009652 + 1.0038052 QPlogPo/w

Summary of Fit

RSquare	0.92062
RSquare Adj	0.920455
Root Mean Square Error	0.544778
Mean of Response	1.811608
Observations (or Sum Wgts)	485

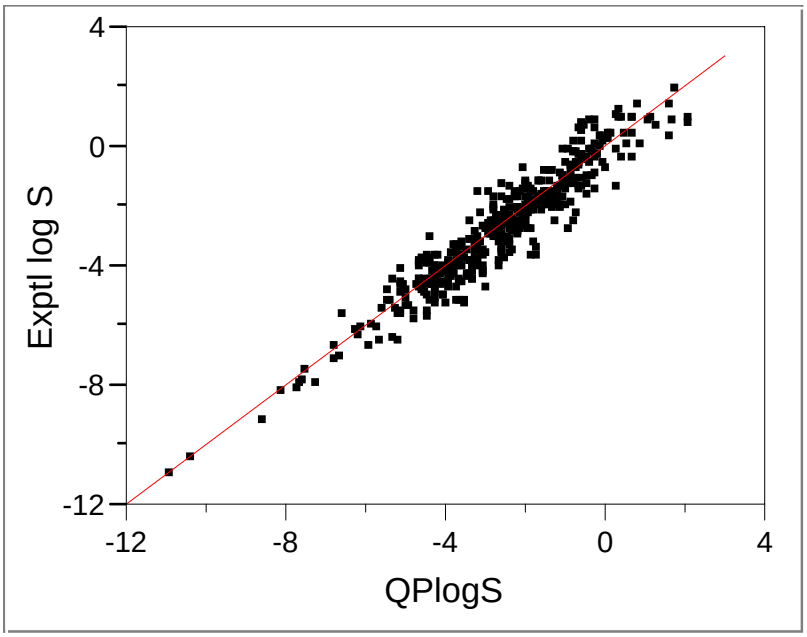
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1662.4677	1662.47	5601.631
Error	483	143.3461	0.30	Prob > F
C. Total	484	1805.8137		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0009652	0.0346	0.03	0.9778
QPlogPo/w	1.0038052	0.013412	74.84	<.0001

log S *vs* QPlogS



Linear Fit

Linear Fit

log S = 0.0300332 + 1.0008512 QPlogS

Summary of Fit

RSquare	0.902338
RSquare Adj	0.902107
Root Mean Square Error	0.62715
Mean of Response	-2.61978
Observations (or Sum Wgts)	426

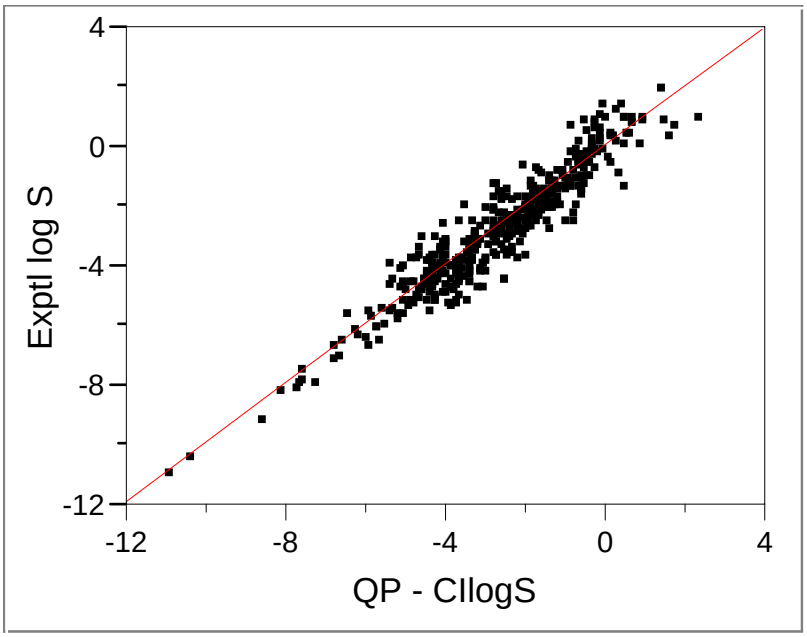
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1540.8176	1540.82	3917.492
Error	424	166.7665	0.39	Prob > F
C. Total	425	1707.5841		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0300332	0.052112	0.58	0.5647
QPlogS	1.0008512	0.015991	62.59	<.0001

log S *vs* Conformation Independent QPlogS



Linear Fit

Linear Fit

log S = 0.0378927 + 0.9927276 CIlogS

Summary of Fit

RSquare	0.893778
RSquare Adj	0.893527
Root Mean Square Error	0.654057
Mean of Response	-2.61978
Observations (or Sum Wgts)	426

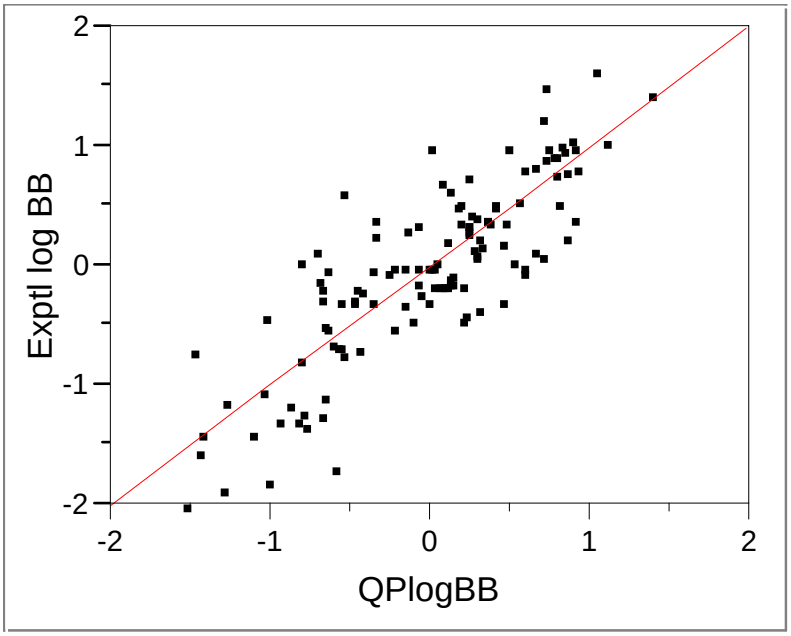
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1526.2011	1526.20	3567.639
Error	424	181.3831	0.43	Prob > F
C. Total	425	1707.5841		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0378927	0.054626	0.69	0.4883
CIlogS	0.9927276	0.01662	59.73	<.0001

log BB [brain]/[blood]



Linear Fit

Linear Fit

log BB = -0.009813 + 0.9967314 QPlogBB

Summary of Fit

RSquare	0.702763
RSquare Adj	0.700385
Root Mean Square Error	0.415251
Mean of Response	-0.03678
Observations (or Sum Wgts)	127

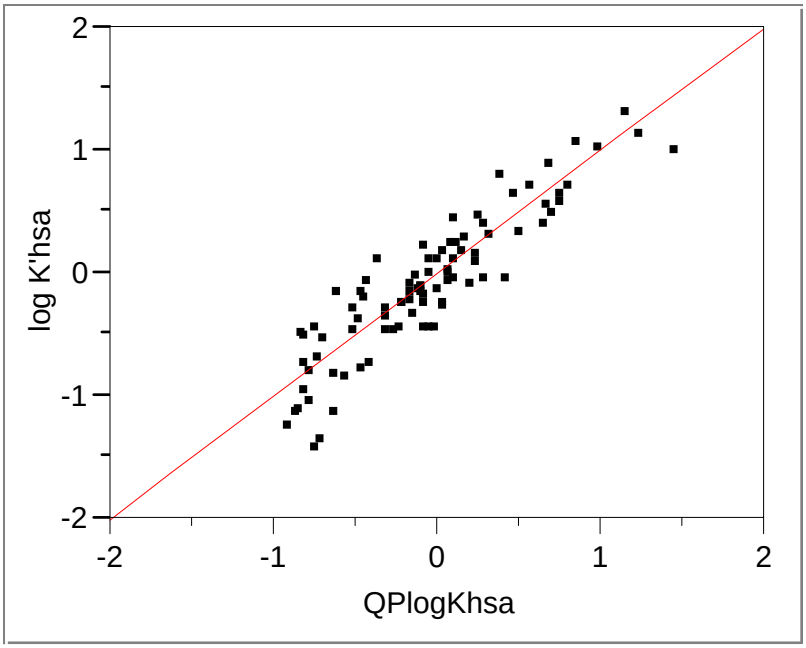
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	50.960963	50.9610	295.5400
Error	125	21.554173	0.1724	Prob > F
C. Total	126	72.515136		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.009813	0.036881	-0.27	0.7906
QPlogBB	0.9967314	0.057979	17.19	<.0001

Binding Affinity for Human Serum Albumin - log K'hsa



Exptl Data: G. Colmenarejo et al., J . Med. Chem. 2001, 44, 4370-4378.

Summary of Fit

Rsquare	0.82499
RSquare Adj	0.81005
Root Mean Square Error	0.252192
Mean of Response	-0.068
Observations (or Sum Wgts)	90

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	24.584558	3.51208	55.2205
Error	82	5.215282	0.06360	Prob > F
C. Total	89	29.799840		<.0001

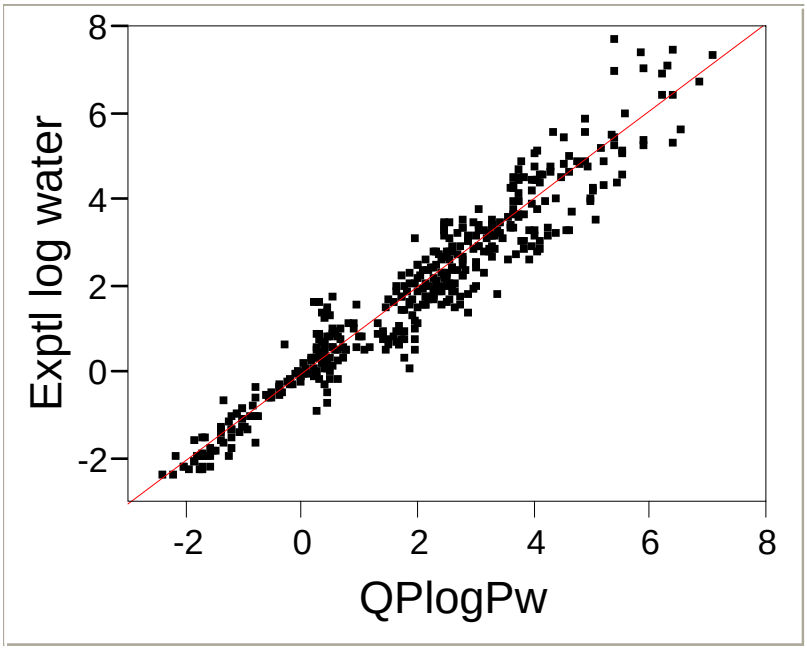
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.538625	0.132628	-11.60	<.0001
volume	0.0031255	0.000225	13.89	<.0001
donorHB	-0.177019	0.05179	-3.42	0.0010
accptHB	-0.261692	0.026188	-9.99	<.0001
ACxDN^.5/SA	41.247956	9.920293	4.16	<.0001
#acid	-0.163497	0.058579	-2.79	0.0065
#amide	-0.379871	0.06478	-5.86	<.0001
#rotor	-0.048244	0.013334	-3.62	0.0005

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
volume	1	1	12.275242	193.0039	<.0001
donorHB	1	1	0.743046	11.6829	0.0010
accptHB	1	1	6.351098	99.8585	<.0001
ACxDN^.5/SA	1	1	1.099562	17.2884	<.0001
#acid	1	1	0.495440	7.7898	0.0065
#amide	1	1	2.187037	34.3868	<.0001
#rotor	1	1	0.832605	13.0911	0.0005

Free Energies of Hydration = -2.3RT log Lw



Linear Fit

Linear Fit

log.Lw = -0.001043 + 1.0009413 QPlogPw

Summary of Fit

RSquare	0.926435
RSquare Adj	0.926259
Root Mean Square Error	0.581428
Mean of Response	1.956532
Observations (or Sum Wgts)	421

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1783.8108	1783.81	5276.626
Error	419	141.6467	0.34	Prob > F
C. Total	420	1925.4575		<.0001

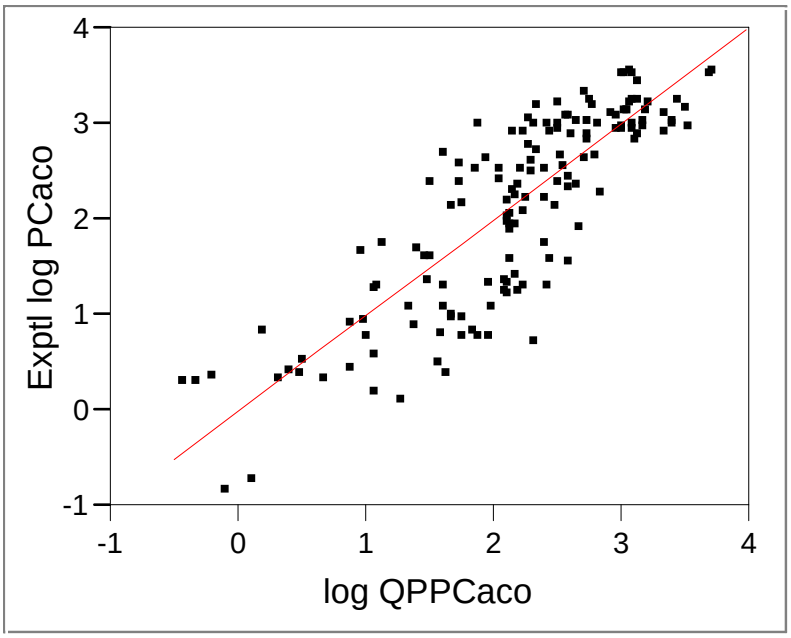
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.001043	0.039105	-0.03	0.9787
QPlogPw	1.0009413	0.013779	72.64	<.0001

Exptl data: M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. Leo, and R. S.Taft, *J. Chem. Soc. Perkin Trans. 2*, 1777-1791 (1994).

log PCaco

(units for PCaco are nm/s)



Linear Fit

Linear Fit

Exptl log PCaco = -0.002746 + 1.0009421 log QPPCaco

Summary of Fit

RSquare	0.70818
RSquare Adj	0.706208
Root Mean Square Error	0.556311
Mean of Response	2.147558
Observations (or Sum Wgts)	150

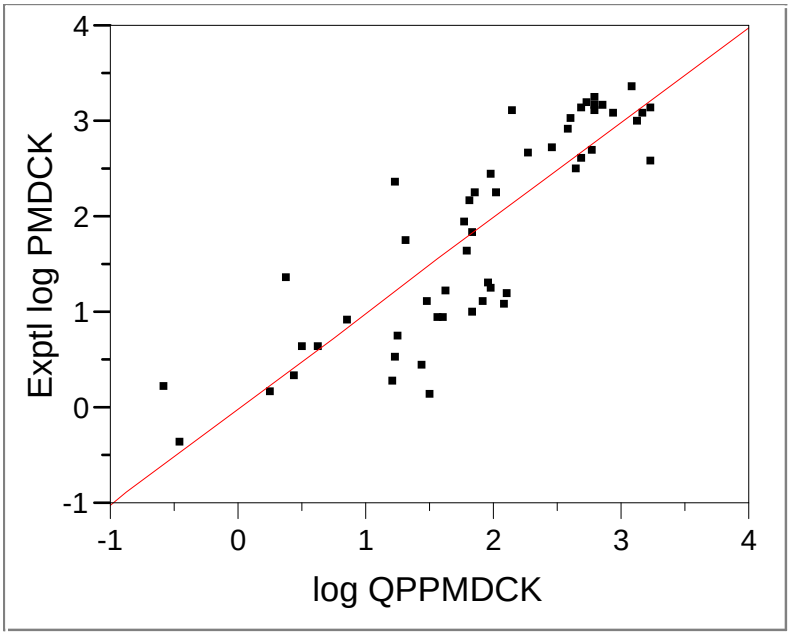
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	111.15415	111.154	359.1623
Error	148	45.80329	0.309	Prob > F
C. Total	149	156.95744		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.002746	0.122217	-0.02	0.9821
log QPPCaco	1.0009421	0.052816	18.95	<.0001

log PMDCK (Affymax data)



Linear Fit

Linear Fit

Exptl log PMDCK = -0.000859 + 1.0002067 log QPPMDCK

Summary of Fit

RSquare	0.729359
RSquare Adj	0.723947
Root Mean Square Error	0.570747
Mean of Response	1.87335
Observations (or Sum Wgts)	52

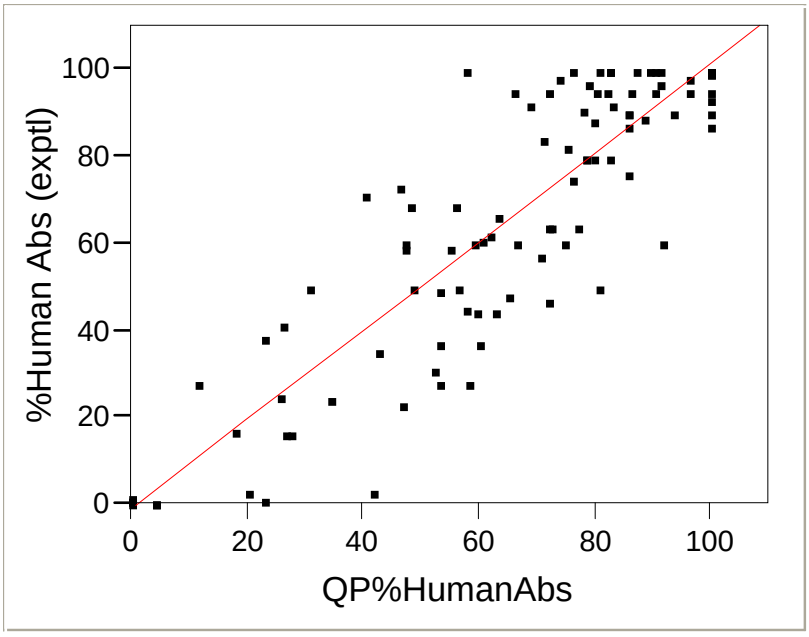
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	43.894082	43.8941	134.7469
Error	50	16.287607	0.3258	Prob > F
C. Total	51	60.181690		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.000859	0.179814	-0.00	0.9962
log QPPMDCK	1.0002067	0.086165	11.61	<.0001

% Human Absorption



Linear Fit

Linear Fit

%Human Abs = -1.130595 + 1.0206866 QP%HumanAbs

Summary of Fit

RSquare	0.790942
RSquare Adj	0.788851
Root Mean Square Error	14.92566
Mean of Response	64.85196
Observations (or Sum Wgts)	102

Analysis of Variance

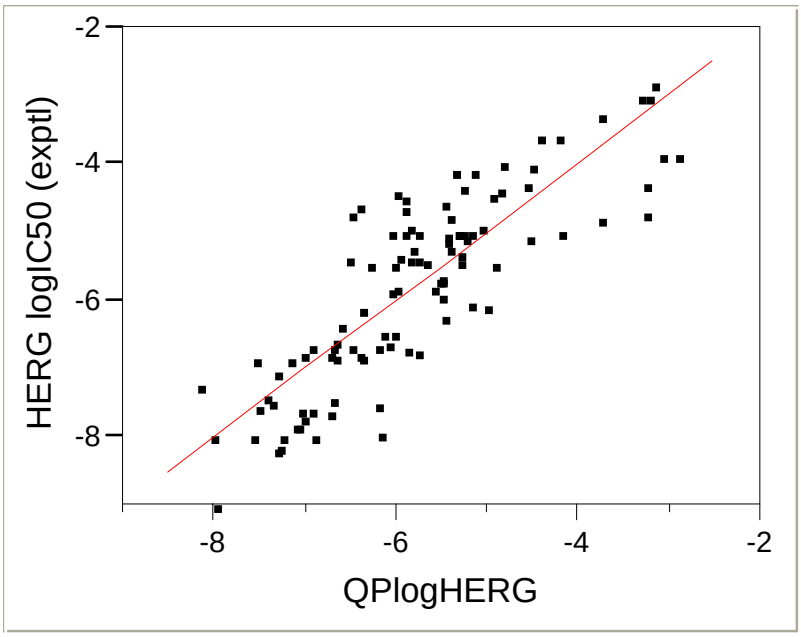
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	84283.99	84284.0	378.3362
Error	100	22277.54	222.8	Prob > F
C. Total	101	106561.53		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.130595	3.70021	-0.31	0.7606
QP%HumanAbs	1.0206866	0.052475	19.45	<.0001

This is a relatively difficult quantity to predict owing to the 0-100 range and noise in the experimental data.

Log IC_{50} for Blockage of Mammalian HERG K^+ Channels



Linear Fit

Linear Fit

HERGlogIC50 = -0.005422 + 0.999584 QPlogHERG

Summary of Fit

RSquare	0.715016
RSquare Adj	0.712166
Root Mean Square Error	0.742779
Mean of Response	-5.80069
Observations (or Sum Wgts)	102

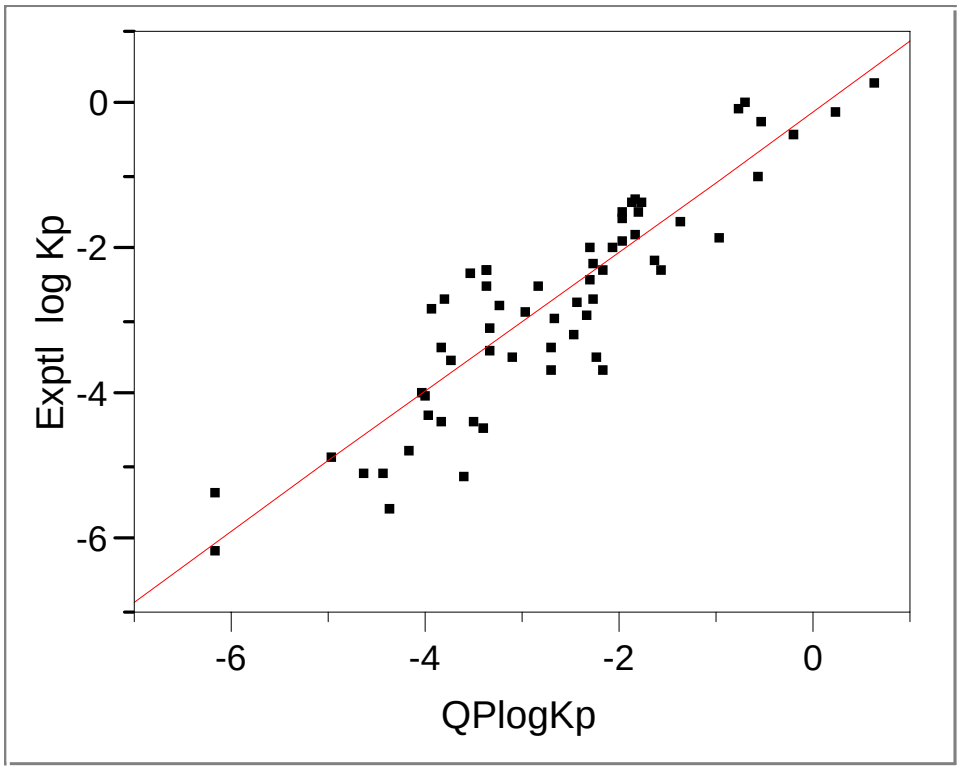
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	138.42532	138.425	250.8972
Error	100	55.17213	0.552	Prob > F
C. Total	101	193.59745		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.005422	0.373188	-0.01	0.9884
QPlogHERG	0.999584	0.063106	15.84	<.0001

log Kp (skin permeability)



Linear Fit

Linear Fit

log Kp = -0.083896 + 0.9642652 QPlogKp

Summary of Fit

RSquare	0.803801
RSquare Adj	0.800476
Root Mean Square Error	0.666147
Mean of Response	-2.69852
Observations (or Sum Wgts)	61

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	107.26143	107.261	241.7151
Error	59	26.18133	0.444	Prob > F
C. Total	60	133.44277		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.083896	0.188566	-0.44	0.6580
QPlogKp	0.9642652	0.062022	15.55	<.0001