Field-Based QSAR

Schrödinger Software Release 2015-1

Schrödinger Press

May 2015

Contents

Document Conventions

In addition to the use of italics for names of documents, the font conventions that are used in this document are summarized in the table below.

Font	Example	Use
Sans serif	Project Table	Names of GUI features, such as panels, menus, menu items, buttons, and labels
Monospace	`$SCHRODINGER/maestro`	File names, directory names, commands, environment variables, command input and output
Italic	*filename*	Text that the user must replace with a value
Sans serif uppercase	CTRL+H	Keyboard keys

Links to other locations in the current document or to other PDF documents are colored like this: Document Conventions.

In descriptions of command syntax, the following UNIX conventions are used: braces { } enclose a choice of required items, square brackets [ ] enclose optional items, and the bar symbol | separates items in a list from which one item must be chosen. Lines of command syntax that wrap should be interpreted as a single command.

File name, path, and environment variable syntax is generally given with the UNIX conventions. To obtain the Windows conventions, replace the forward slash / with the backslash \ in path or directory names, and replace the $ at the beginning of an environment variable with a % at each end. For example, `$SCHRODINGER/maestro` becomes `%SCHRODINGER%\maestro`.

Keyboard references are given in the Windows convention by default, with Mac equivalents in parentheses, for example CTRL+H (⌘H). Where Mac equivalents are not given, COMMAND should be read in place of CTRL. The convention CTRL-H is not used.

In this document, to *type* text means to type the required text in the specified location, and to *enter* text means to type the required text, then press the ENTER key.

References to literature sources are given in square brackets, like this: [10].

Introduction

Field-Based QSAR allows you to build 3D QSAR models based on fields, such as electrostatic, hydrophobic, or steric fields, for a set of aligned ligands. The models can be applied to other ligands, or stored with a pharmacophore hypothesis, or exported for later use.

A related method is atom-based QSAR, which allows you to build 3D QSAR models based on the spatial distribution of atom types or pharmacophore types. See Chapter 7 of the *Phase User Manual* for more information on these models.

1.1 Background

The field-based QSAR models are based on CoMFA [1] and CoMSIA [2, 3]. CoMFA field-based models are constructed by calculating the value of fields, such as the electrostatic field, on a rectangular grid that encompasses the molecules in the training set. The grid locations are the independent variables that are used in a partial-least-squares (PLS) fitting procedure to produce a relationship between the values of the fields and the activity of the training set molecules.

CoMSIA fields are also evaluated at points on a rectangular grid. The fields are calculated by summing the values of properties of a given atom, weighted by a Gaussian function of the distance between the grid point and the atom. The steric contribution is derived from the third power of the atomic radius; the electrostatic field from the partial atomic charges, and the hydrophobic field from estimated ALOGP values. Hydrogen-bond receptor and donor fields have a value of 1 at the projected point locations.

The field-based QSAR models are an implementation of the CoMFA and CoMSIA methods with a specific set of parameters. The Lennard-Jones steric potentials are taken from the OPLS_2005 force field, as are the atomic charges for the electrostatic fields (by default). Hydrophobic fields are based on the atom types and hydrophobic parameters from Ghose et al. [4]. Hydrogen-bond acceptor and donor fields are based on Phase pharmacophore feature definitions, with projected points, as are aromatic ring fields, with projected points 1.8 Å above and below the ring plane. As the models are not exactly the same as the standard CoMFA and CoMSIA models, different names have been used in Phase: Force Field for CoMFA-like models, and Gaussian for CoMSIA-like models.

1.2 Running Schrödinger Software

Schrödinger applications can be run from a graphical interface or from the command line. The software writes input and output files to a directory (folder) which is termed the *working directory*. If you run applications from the command line, the directory from which you run the application is the working directory for the job.

Linux:

To run any Schrödinger program on a Linux platform, or start a Schrödinger job on a remote host from a Linux platform, you must first set the `SCHRODINGER` environment variable to the installation directory for your Schrödinger software. To set this variable, enter the following command at a shell prompt:

csh/tcsh: `setenv SCHRODINGER` *installation-directory*

bash/ksh: `export SCHRODINGER=`*installation-directory*

Once you have set the `SCHRODINGER` environment variable, you can run programs and utilities with the following commands:

`$SCHRODINGER/`*program* `&`
`$SCHRODINGER/utilities/`*utility* `&`

You can start the Maestro interface with the following command:

```
$SCHRODINGER/maestro &
```

It is usually a good idea to change to the desired working directory before starting the Maestro interface. This directory then becomes the working directory.

Windows:

The primary way of running Schrödinger applications on a Windows platform is from a graphical interface. To start the Maestro interface, double-click on the Maestro icon, on a Maestro project, or on a structure file; or choose Start → All Programs → Schrodinger-2015-2 → Maestro. You do not need to make any settings before starting Maestro or running programs. The default working directory is the Schrodinger folder in your Documents folder.

If you want to run applications from the command line, you can do so in one of the shells that are provided with the installation and have the Schrödinger environment set up:

- Schrödinger Command Prompt—DOS shell.
- Schrödinger Power Shell—Windows Power Shell (if available).

You can open these shells from Start → All Programs → Schrodinger-2015-2. You do not need to include the path to a program or utility when you type the command to run it. If you want access to Unix-style utilities (such as `awk`, `grep`, and `sed`), preface the commands with `sh`, or type `sh` in either of these shells to start a Unix-style shell.

Mac:

The primary way of running Schrödinger software on a Mac is from a graphical interface. To start the Maestro interface, click its icon on the dock. If there is no Maestro icon on the dock, you can put one there by dragging it from the SchrodingerSuite2015-2 folder in your Applications folder. This folder contains icons for all the available interfaces. The default working directory is the Schrodinger folder in your Documents folder (`$HOME/Documents/Schrodinger`).

Running software from the command line is similar to Linux—open a terminal window and run the program. You can also start Maestro from the command line in the same way as on Linux. The default working directory is then the directory from which you start Maestro. You do not need to set the `SCHRODINGER` environment variable, as this is set in your default environment on installation. To set other variables, on OS X 10.7 use the command

`defaults write ~/.MacOSX/environment` *variable* `"`*value*`"`

and on OS X 10.8, 10.9, and 10.10 use the command

`launchctl setenv` *variable* `"`*value*`"`

1.3 Citing Field-Based QSAR in Publications

The use of this product should be acknowledged in publications as:

Field-Based QSAR, version 2015-2, Schrödinger, LLC, New York, NY, 2015.

Field-Based QSAR Tutorial

This chapter provides a tutorial exercise for developing a field-based QSAR model from Maestro.

2.1 Preparing for the Exercises

To run the exercises, you need a working directory in which to store the input and output, and you need to copy the input files from the installation into your working directory. This is done automatically in the Tutorials panel, as described below. To copy the input files manually, just unzip the `field_qsar` zip file from the `tutorials` directory of your installation into your working directory.

On Linux, you should first set the `SCHRODINGER` environment variable to the Schrödinger software installation directory, if it is not already set:

csh/tcsh: `setenv SCHRODINGER` *installation-path*

sh/bash/ksh: `export SCHRODINGER=`*installation-path*

If Maestro is not running, start it as follows:

- **Linux:** Enter the following command:

  ```
  $SCHRODINGER/maestro -profile Maestro &
  ```

- **Windows:** Double-click the Maestro icon on the desktop.

  You can also use Start → All Programs → Schrodinger-2015-2 → Maestro.

- **Mac:** Click the Maestro icon on the dock.

  If it is not on the dock, drag it there from the SchrodingerSuites2015-2 folder in your Applications folder, or start Maestro from that folder.

Now that Maestro is running, you can start the setup.

1. Choose Help → Tutorials.

   The Tutorials panel opens.

2. Ensure that the Show tutorials by option menu is set to Product, and the option menu below is labeled Product and set to All.

3. Select Field-Based QSAR Tutorial in the table.
4. Enter the directory that you want to use for the tutorial in the Copy to text box, or click Browse and navigate to the directory.

   If the directory does not exist, it will be created for you, on confirmation. The default is your current working directory.
5. Click Copy.

   The tutorial files are copied to the specified directory, and a progress dialog box is displayed briefly.

If you used the default directory, the files are now in your current working directory, and you can skip the next two steps. Otherwise, you should set the working directory to the place that your tutorial files were copied to.

6. Choose Project → Change Directory.
7. Navigate to the directory you specified for the tutorial files, and click OK.

You can close the Tutorials panel now, and proceed with the exercises.

2.2 Adding and Assigning the Ligands

To set up a QSAR model, you must add ligands and assign them to a training set and a test set. It is always important to have a test set so that you can assess the quality of the model.

1. Choose Tasks → QSAR → Field-Based or Applications → Field-Based QSAR.
2. For Add ligands, click From File.

   The Add From File file selector opens.
3. Select `cdk2_fqsar.maegz` and click Open.

   The Choose Activity Property dialog box opens.
4. Under Choose an activity property, select the pIC50 property.
5. Click OK.

   The Ligands table in the Field-Based QSAR panel is populated with the 71 ligands. The QSAR Set property is set to training, and the Activity property shows the pIC50 values.
6. Select rows 55 through 71 in the table (use shift-click).
7. Control-click the QSAR Set column for one of these rows.

   The value in the column changes to test for all of the selected rows.

2.3 Building the Models

When building a model, you must choose which fields to include in the model, and set parameters for building the model. In this exercise, the Gaussian fields will be used.

1. Click Build.

   The Build Field-Based Model dialog box opens.

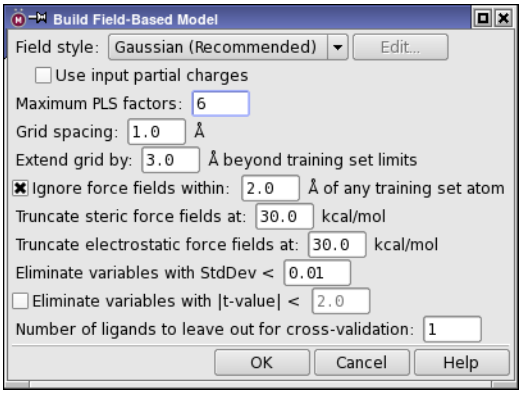

Figure 2.1. The Build Field-Based Model dialog box.

2. Set the Field style to Gaussian (Recommended), if it is not already set to this choice.
3. Enter 6 in the Maximum PLS factors text box.

   This value should be no more than the number of training set structures divided by 5, otherwise overfitting may occur. A model is built for each number of PLS factors up to this value.

   The remaining settings can be left at their default values.

4. Click OK.

   The dialog box closes. After a short while, the results columns in the Ligands table is filled in with the predictions for each of the 6 models, for both the training set and the test set, and the QSAR statistics and Field fractions tables are filled in with the statistics for the models.

2.4 Examining the Models

After building a set of models, the next task is to determine which of these models to use. For this purpose, both the training set and the test set results should be examined. To avoid

choosing an over-fit model, a good rule is to stop increasing the number of PLS factors when no more improvement is obtained in the results.

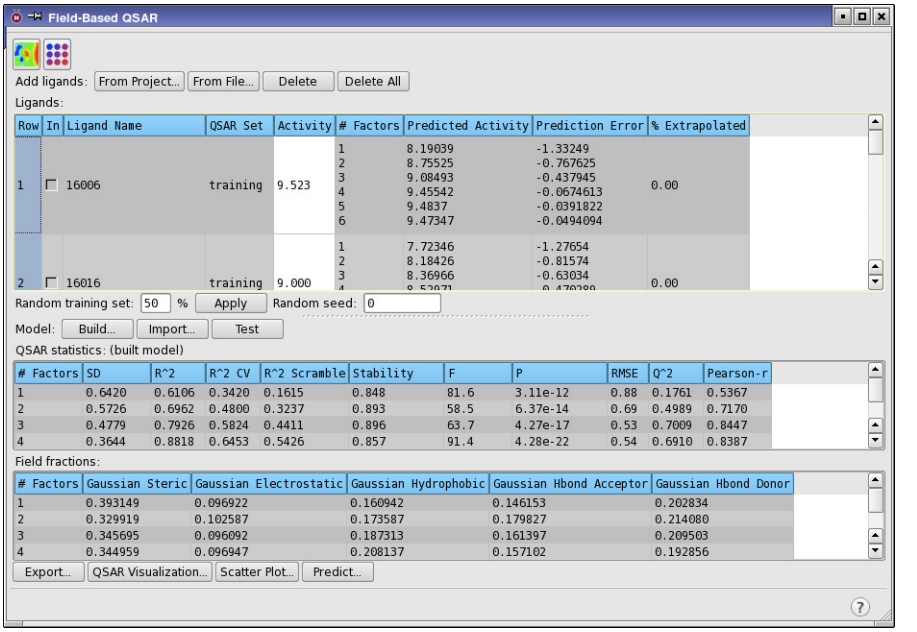

Figure 2.2. The Field-Based QSAR panel with results.

Examine the statistics in the QSAR statistics table.

- The standard deviation (SD) decreases and the correlation coefficient (R^2) increases as the number of PLS factors increases. R^2 Scramble also increases. This is normal, but does not really tell us much about which model to choose. It only tells us that increasing the number of variables in the fit reduces the error of the fit.
- Stability increases to 3 factors, then decreases. The 3-factor model predictions are the least sensitive to the composition of the training set, based on leave-1-out tests. For 4 or more factors, the R^2 value is larger than the stability value, which indicates that over-fitting starts at about 4 factors.
- The RMSE decreases from 1 to 3 factors, then doesn't change much with more factors. Likewise, Q^2 and Pearson-r increase up to 3 factors and don't change a lot subsequently. As these are test set results, they are a better indicator of the value of adding more factors.

On the basis of these observations, the 3-factor model is probably the best choice, as increasing the number of factors doesn't improve the test set predictions much, and the models with higher numbers of factors are probably over-fit.

2.5 Examining the Predictions

Having looked at the overall statistics, we now look at the predictions for both the training set and the test set, using the plotting facility.

1. Click Scatter Plot.

   The Phase QSAR - Scatter Plot dialog box opens.

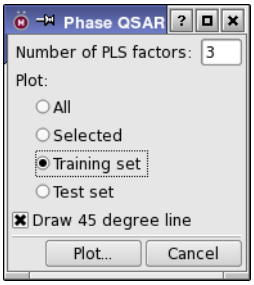

Figure 2.3. The Phase QSAR - Scatter Plot dialog box.

2. In the Number of PLS factors text box, enter 3.
3. Select Training set.
4. Ensure that Draw 45 degree line is selected.
5. Click Plot.

   After a short delay, a scatter plot of the training set activities is displayed, labeled plot-1. Nearly all of the points are close to the line, so the training set fit is generally good.

Now plot the test set data.

6. Click Scatter Plot (in the Field-Based QSAR panel).

   The Phase QSAR - Scatter Plot dialog box opens again.

7. Select Test set.
8. Click Plot.

   A scatter plot of the test set activities is displayed, labeled plot-2.

There are two main outliers, which we will examine.

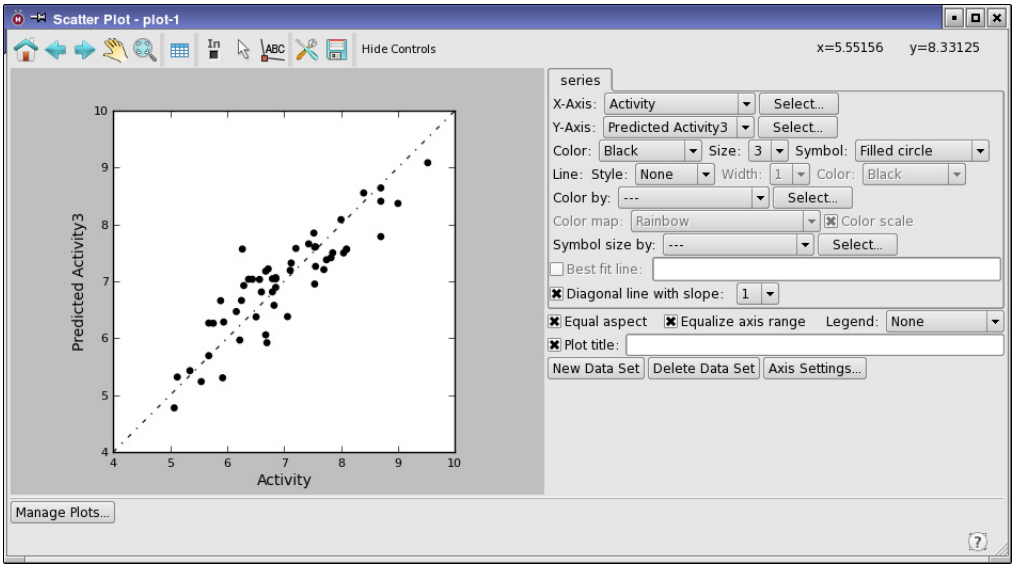

Figure 2.4. Plot of the training set results.

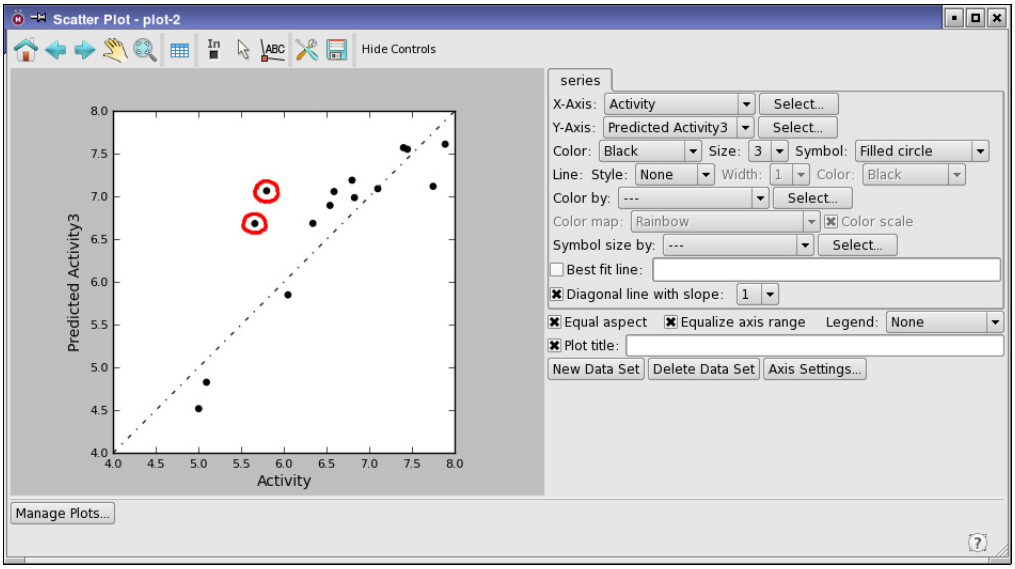

Figure 2.5. Plot of the test set results, with outliers marked.

9. Click the In button.

   

   This button allows you to pick points on the plot and have them displayed in the Workspace.

10. Pick the two outlying points on the plot, in turn.

    These are for ligands 16047 and 16076. Ligand 16047 has the largest error. There is a similar ligand in the training set, ligand 16058 (row 21) that has a small error.

11. In the Field-Based QSAR panel, click the In column in row 21.

    Ligand 16058 is placed in the Workspace.

12. Scroll down and control-click the In column in row 68 (ligand 16047)

    Ligand 16047 is also placed in the Workspace, superimposed on ligand 16058. The two ligands are very similar in shape and structure.

13. Click Tile.

    

    The two structures are displayed separately. The differences between the two can be seen more easily: the orientation of the pyridyl ring, and the bond order of a carbon-carbon bond in the middle ring of the fused ring system. The first difference would be eliminated if the pyridyl ring were rotated by 180º, which we will do in the next exercise.

14. Click Tile again, to exit tile mode.

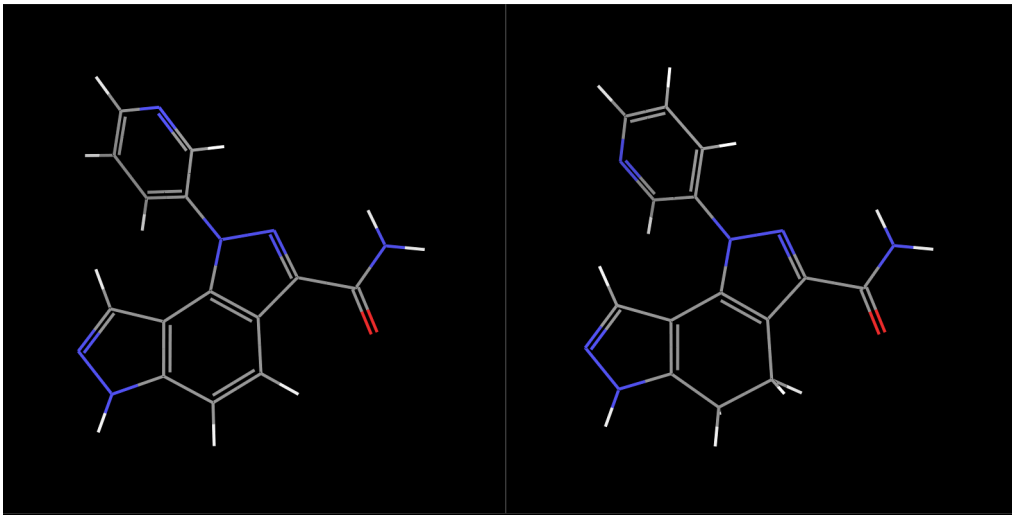

Figure 2.6. Ligands 16058 (left) and 16047 (right).

2.6 Making a Prediction

In this exercise, we will modify the structure of one of the outliers from the test set (ligand 16047) and predict its activity.

1. In the Entry List panel, enter 16047 in the filter (search) text box.

   The Entry List panel is docked in the main window by default. If it is not displayed, choose Project → Entry List.

   When you enter the text, only ligand 16047 is listed.

2. Click the row for the ligand to select just this ligand.

   The filter doesn't change the selection of entries, only the entries that are shown. The status report below the list of entries should show Entries: 71 total, 1 selected.

3. Right-click in the row and choose Duplicate → Ungrouped.

   The entry for this ligand is duplicated and becomes row 1, and the duplicate is included in the Workspace. If it is not, click the In column in row 1 to place it in the Workspace.

4. If the Edit toolbar is not displayed, click the Edit button on the Manager toolbar.

5. Click the arrow next to the Adjust button and choose Quick Torsion.

6. Click the bond between the pyridyl ring and the pyrazole nitrogen.

   A light blue arrow is displayed, pointing to the pyridyl ring.

7. Drag horizontally in the Workspace with the left mouse button to rotate the pyridyl ring by 180º.

   The initial angle (136.2º) and current angle are displayed in the status bar as you drag. The final angle should be close to –43.8º.

8. In the Field-Based QSAR panel, click Predict.

   The Choose Entries dialog box opens, so you can choose project entries to predict their activities. By default, only the selected entries are shown, so the ligand you just adjusted should be the only ligand listed.

9. Click Choose.

   The dialog box closes, and the prediction is made. The results for all 6 models are added to the Project Table as properties for this ligand.

10. Click the Table button on the Project toolbar.

    

    The Project Table panel opens.

11. Scroll the table horizontally so that the Activity column is showing.

    The measured activity value is about 5.8.

12. Scroll horizontally to view and compare the two sets of predicted activities.

    The first set (Predicted Activity*n*) is the set for the original structure, and was copied when the structure was duplicated. The second set (predicted activity*n*) contains the predictions for the modified structure.

    The two sets of predictions are almost identical. The model shows no difference resulting from the rotation of this group, and implies that the change in orientation of the pyridyl ring is not relevant to activity.

Running Field-Based QSAR from Maestro

Field-based QSAR models can be built and applied from Maestro in the Field-Based QSAR panel. To open this panel, choose Tasks → QSAR → Field-Based or Applications → Field-Based QSAR.

3.1 Preparing the Ligands

The first step is to prepare the ligands to use. The ligands you add must be fully prepared 3D structures that are properly aligned. The alignment should ideally include conformational variation as part of the alignment. No facility is provided in these panels for preparing the structures, or aligning the ligands.

Structure preparation can be done with LigPrep (see the *LigPrep User Manual*), and generation of conformers can be done with ConfGen (see the *ConfGen User Manual*) or MacroModel (see Chapter 8 of the *MacroModel User Manual*).

If you were using the Develop Common Pharmacophore Hypotheses panel to develop a pharmacophore hypothesis and the ligands were exported from this panel, they should already be prealigned to the pharmacophore model. No further alignment should be needed.

For best results, the alignment should include conformational variation. There are several ways in which you can align a set of ligands. Some of them involve a conformational search and alignment of the best conformer. Others perform a simple alignment without a search, and you would have to run the conformational search first.

- Use the Shape Screening panel (Tasks menu or Applications menu). This panel aligns the molecules to a query molecule on the basis of the shape or atom-type weighted shape. The method involves alignment of the best conformers from a conformational search. See Chapter 14 of the *Phase User Manual* for details.
- Use the Flexible Ligand Alignment panel (Tools menu). This panel does a quick conformational search and aligns the best conformer of the second and subsequent ligands to the first, replacing the structure with this conformer.
- Use the Superposition panel (Tools → Superposition) to align the ligands. The best choice is probably to align by a SMARTS pattern for the ligand core. You will also have to select the conformers that have the best alignment.

For example, you could run a conformational search for each ligand separately and store the results in separate entry groups, align each group separately to a chosen structure (for example, a known active), sort the ligands in each group by RMSD, then choose the entry with the lowest RMSD from each group.

- Align the ligands to a pharmacophore hypothesis, using one of the (Phase) Pharmacophore Screening panels to run a screening job. The hit file from this job contains the aligned ligands.

Whichever method you use, you should ensure that you get only one output structure for each ligand, which should be the best-aligned conformer.

3.2 Adding the Ligands for the Model

You can add ligands to the set to be used for the QSAR model from two sources, by clicking one of the Add ligands buttons:

- From Project—Opens the Add From Project dialog box, in which you can choose a set of entries; select an activity property, converting it into the appropriate units if need be; and select a property to define the training and test sets.

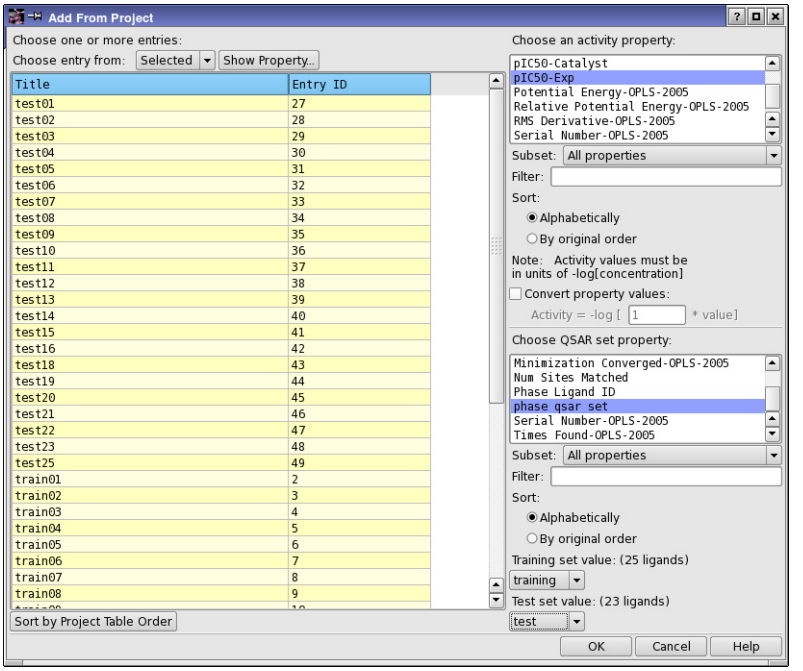

Figure 3.1. The Add from Project dialog box for QSAR models.

- From File—Opens a file selector, in which you can navigate to and select the file. When you click OK, the Choose Activity Property dialog box opens, in which you can select an activity property, converting it into the appropriate units if need be, and select a property to define the training and test sets.

You can use these buttons more than once to add multiple sets of ligands. The ligands you add are always appended to the Ligands table: there is no replacement of ligands, and no checking for duplicates is done. If you want to delete ligands, select them in the table and click Delete. This allows you to remove duplicates, or to remove ligands that you don't want in your model. To start again with a new set of ligands, click Delete All, then start adding the new ligands.

When you add ligands, you can assign them to the training and test sets on the basis of the values of a property. The choice is made in the same panel as the choice of the activity property. The assignment is made by choosing a single value of the property for the training set and a single value for the test set. If you want to use this feature, you will have to create an appropriate property beforehand. If you exported ligands from the Build QSAR Model step of the Develop Common Pharmacophore Hypotheses panel, you can use the phase qsar set property.

The ligands are displayed in the Ligands table when they are added. The # Factors, Predicted Activity, Prediction Error, and % Extrapolated columns are initially empty. The values are added after the QSAR model is built. The table columns are described in Table 3.1.

Table 3.1. Description of the Ligands table columns.

Column	Description
In	Inclusion status of the ligand. The button is colored black if the ligand is included in the Workspace, and is empty if the ligand is excluded. You can include and exclude ligands with click, shift-click, and control-click.
Ligand Name	The name of the ligand.
QSAR Set	Indicates whether a ligand is in the training set, the test set, or neither (the ligand is ignored). The column is blank if the ligand is ignored. Click the column repeatedly to cycle through the three possible states.
Activity	The ligand's activity. You can alter the activity values by editing the table cells.
# Factors	Number of PLS factors used for the QSAR model.
Predicted Activity	Activity predicted by the QSAR model. The number of rows in this column for each ligand is equal to the number of PLS factors specified in the Build QSAR Model - Options dialog box. Each row contains the prediction from a model containing the number of PLS factors indicated in the # Factors column.
Prediction Error	Error in the predicted activity.
% Extrapolated	Percentage of field values for the ligand that lie outside the range found in the training set.

3.3 Choosing a Training Set and a Test Set

The next task is to choose a training set and a test set, and exclude ligands that you do not want in either set. If you did not do this on the basis of a property when exporting the ligands, all of the ligands are initially included in the training set, and you must partition them.

To change the set membership of an individual ligand, click in the QSAR Set column for the ligand. The membership cycles between training, test, and blank, the last of which means that the ligand is excluded from both sets—that is, it is not used. To change the set membership for a group of ligands, select the ligands in the table using shift-click or control-click, then control-click in the QSAR Set column for any of the ligands.

You can select a random fraction of the ligands for the training set by entering a percentage in the Random training set text box and clicking Apply. The specified percentage of ligands is selected at random from the existing training and test sets and assigned to the training set. The rest are assigned to the test set. Ligands that are in neither set are not used in the selection.

If you select the training set randomly, you may want to do this in a reproducible way. By default, the random seed changes each time a random training set is selected, so you get a different training set each time you click Apply. If you change the value in the Random seed text box to any positive integer, the same random training set is created each time you click Apply. The default value of zero ensures that the assignment is always random.

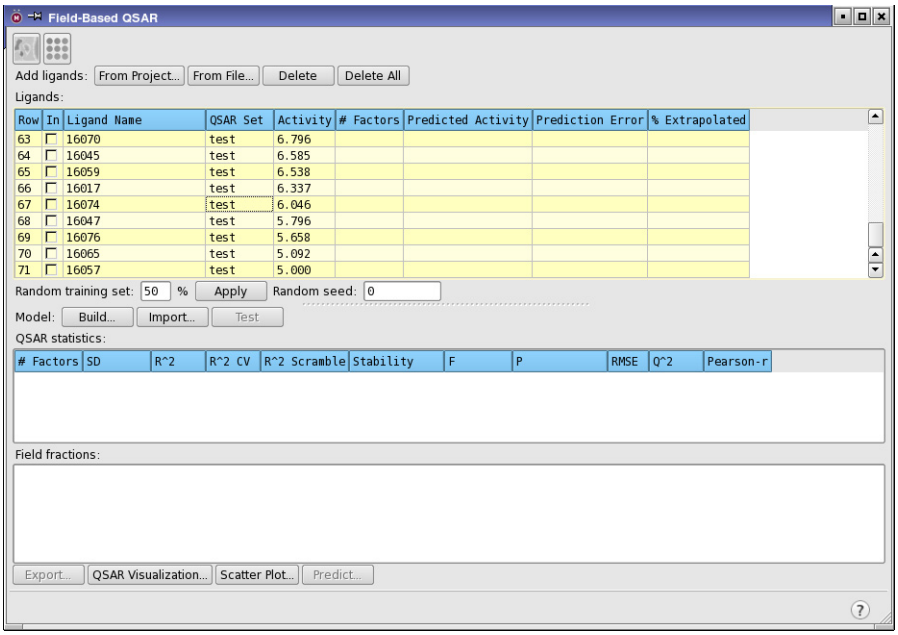

Figure 3.2. The Field-Based QSAR panel with ligands imported and assigned.

The following subsections describe how to build, test, and use the models.

3.4 Building and Testing the Model

Once you have chosen the training and test sets, click Build to build the QSAR models. The Build Field-Based Model dialog box opens. This dialog box has a range of settings for the fields and the data that are used to fit the fields.

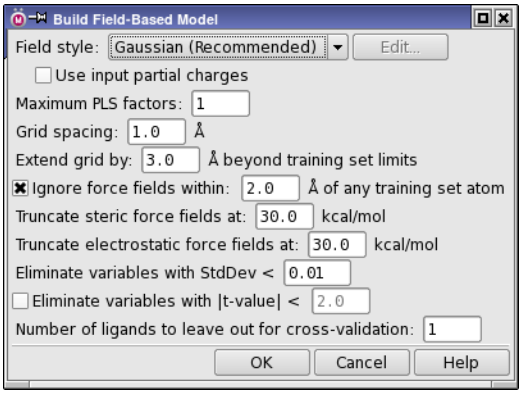

Figure 3.3. The Build Field-Based Model dialog box.

The first choice is to select the fields that you want to include in the model. The Field style options are:

- Force field—Use the force-field electrostatic and steric fields for the model (CoMFA).
- Gaussian—Use the five Gaussian fields for the model (CoMSIA).
- Extended Gaussian—Use the five Gaussian fields for the model (CoMSIA), plus the Gaussian Aromatic Ring field.
- Custom—Select a combination of force-field and Gaussian fields to use for the model. To make the selection, click Edit, and select the fields from the list in the Custom Field Style dialog box. In this dialog box you can also import feature definitions for the fields.

For the electrostatic fields, you can choose to use input partial charges rather than those from the OPLS_2005 force field, by selecting Use input partial charges.

You can specify the number of PLS factors to use in the Maximum PLS factors text box. A model is built for each number of PLS factors up to the specified maximum. There is no limit on the maximum number of PLS factors, but as a general rule, you should stop adding factors when the standard deviation of regression is approximately equal to the experimental error.

To set the spacing of the grid points on which the fields are evaluated, in angstroms, enter a value in the Grid spacing text box. You can extend the grid beyond the limits of the training set by entering a value in the Extend grid by N Å beyond training set limits text box. This allows you to make predictions for ligands that are larger than the training set ligands.

Once the grid is defined, you can make settings that determine how the fields are evaluated on the grid.

- Ignore force fields within N Å of any training set atom—Set the distance from the training set atoms inside which grid points are discarded when evaluating the fields. A grid point is discarded if it is within the specified distance of any of the atoms of any of the training set molecules. This option essentially prevents the model from being dominated by the strong fields close to the nuclei.
- Truncate steric force fields at *value*—Specify the cutoff value for truncating steric force fields, in kcal/mol. Any field that is greater than this value is set to this value.
- Truncate electrostatic force fields at *value*—Specify the cutoff value for truncating electrostatic force fields, in kcal/mol. Any field that is greater than this value is set to this value.

The values at each included grid point can be further processed to eliminate variations that are not significant.

- Eliminate variables with StdDev < *value*—Eliminate variables whose standard deviation is less than the value given in the text box. The smallest value you can set is 0.01. This setting allows you to remove variables that have no effect on the model.
- Eliminate variables with |t-value| < *value*—Select this option to use a t-value filter to eliminate independent variables whose regression coefficients are overly sensitive to small changes in the training set composition, and enter the threshold for eliminating variables in the text box. The resulting models have fewer uninformative variables and tend to give better predictions on test set compounds.

You can set the number of ligands to be used in the leave-N-out cross-validation statistics, in the Number of ligands to leave out for cross-validation text box. This value is also used for assessing the stability of the model (Stability property). The default is 1.

When you have finished making settings, click OK to build the model.

When the results are returned, the # Factors, Predicted Activity, and Prediction Error columns are filled in for both the training set and the test set, and the QSAR statistics and Field fractions tables are filled in.

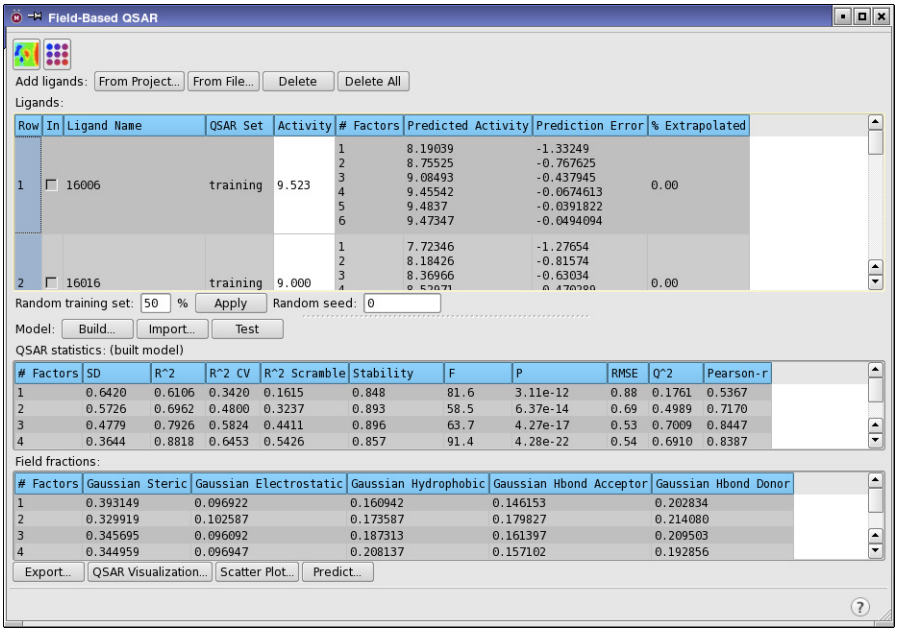

Figure 3.4. The Field-Based QSAR panel, showing results.

If you have ligands that you did not include in the test set, you can include them and click Test to calculate the predicted activity and update the QSAR statistics for the test set.

You can also import an existing model, instead of building it. To import the model, click Import and navigate to the desired `.qsar` file.

3.5 Examining the Model

There are several ways in which you can assess the accuracy or usefulness of the model, listed below. Some important tasks when examining the model are to decide which number of PLS factors should be used, and how good the models are in predicting activities.

- Examine the QSAR statistics, which are described in Table 3.2. Definitions of the statistics can be found in Appendix A.3 of the *Phase User Manual*.

  The most important statistics are the test set statistics: RMSE, Q^2, and Pearson-r, which indicate how good the predictions are. If the predictions are not improving (much) as the number of PLS factors increases, the extra factors are not adding to the model and the model is probably over-fit.

Table 3.2. Description of the QSAR statistics table columns.

Column	Description
# Factors	Number of factors in the partial least squares regression model.
SD	Standard deviation of the regression.This is the RMS error in the fitted activity values, distributed over *n*–*m*–1 degrees of freedom (*n* ligands, *m* PLS factors).
R^2	Value of R^2 for the regression (the coefficient of determination). A value of 0.80, for example, means that the model accounts for 80% of the variance in the observed activity data. R^2 is always between 0 and 1.
R^2 CV	Cross-validated R^2 value, computed from predictions obtained by a leave-N-out approach.
R^2 Scramble	Average value of R^2 from a series of models built using scrambled activities. Measures the degree to which the molecular fields can fit random data. A low value means that the model cannot fit random data, but a high value merely means that the variable set is fairly complete and can fit anything.
Stability	Stability of the model predictions to changes in the training set composition. Maximum value is 1. A high value indicates a model that is not sensitive to omissions from the training set. A stability value that is lower than the R^2 value is an indication of over-fitting.
F	The ratio of the model variance to the observed activity variance. The model variance is distributed over *m* degrees of freedom and the activity variance is distributed over *n*–*m*–1 degrees of freedom (*n* ligands, *m* PLS factors). Large values of F indicate a more statistically significant regression.
P	The significance level of F when treated as a ratio of Chi-squared distributions. Smaller values indicate a greater degree of confidence. A P value of 0.05 means F is significant at the 95% level.
RMSE	Root-mean-square error in the test set predictions.
Q^2	Value of Q^2 for the predicted activities. Directly analogous to R-squared, but based on the test set predictions. Q^2 can take on negative values if the variance in the errors is larger than the variance in the observed activity values.
Pearson-r	Pearson r value for the correlation between the predicted and observed activity for the test set.

Of the training set statistics, the Stability is an indicator of the sensitivity of the model to omissions from the training set. When the R^2 value is larger than the stability value, this is an indication that the data set is over-fit.

- Create a scatter plot of the experimental data against the predicted data. To do this, click Scatter Plot. The Phase QSAR - Scatter Plot dialog box opens, in which you can select the number of PLS factors and the ligands to include in the plot. When you click OK, the Manage Plots panel opens, and the Scatter Plot panel opens to display the plot.

- Visualize the QSAR model in the Workspace, as described in the rest of this section.

You can view a representation of the fields as contours (surfaces), or as color intensities of the fields on the grid. To do so, click the View Contours button or the View Intensities button at the top of the panel.


You can control what is displayed in the Workspace by using the Field-Based QSAR Visualization Settings panel. To open this panel, click QSAR Visualization.

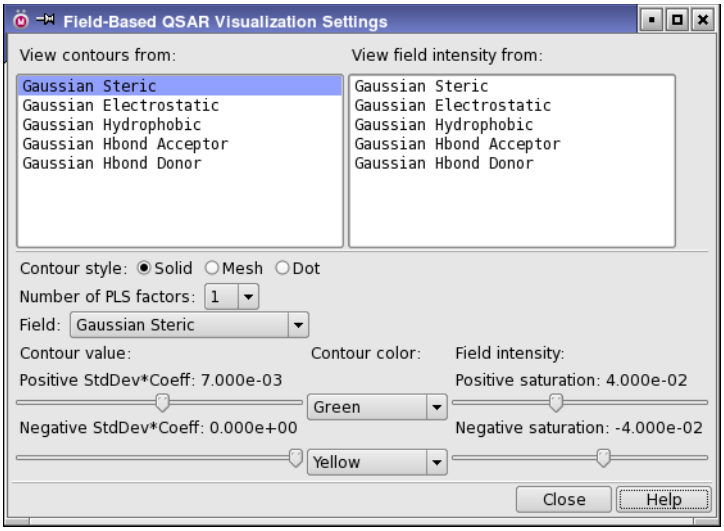

Figure 3.5. The Field-Based QSAR Visualization Settings panel.

Contours can be displayed in different styles. The default is Solid, but you can choose Mesh or Dot for the Contour style, and this will become the new default.

Multiple contours can be displayed at the same time. The contours that are displayed can be selected from the View contours from list. By default, the first contour is selected. To change the value of the field at which the contouring is done or the colors used, first select the field from the Field option menu, then use the Contour value sliders and Contour color option menus to make the changes. The default colors are given in Table 3.3.

Field intensities can only be displayed for one field at a time. The field that is displayed can be selected from the View field intensity from list. By default, the first in the list is displayed. You can change the field cutoffs for the color saturation with the sliders under Field intensity. Fields with greater absolute values than the cutoff are displayed at the maximum brightness.

Table 3.3. Colors for field contours.

Field Type	Positive	Negative
Force Field Steric	Green	Yellow
Force Field Electrostatic	Blue	Red
Gaussian Steric	Green	Yellow
Gaussian Electrostatic	Blue	Red
Gaussian Hydrophobic	Yellow	White
Gaussian Hbond Acceptor	Red	Magenta
Gaussian Hbond Donor	Purple	Cyan
Gaussian Aromatic Ring	Orange	Gray

3.6 Using the Model

Once you are satisfied with the model, you can make use of it in the following ways:

- Make predictions for other molecules, which must exist as entries in the Project Table. To do so, click Predict, and choose the entries in the entry chooser that is displayed. The predicted property for each number of PLS factors is then added to these entries in the Project Table.
- Use the Field fractions or Atom type fractions to assess the molecular features that are primarily responsible for the activity of the molecule. For example, if steric and hydrophobic Gaussian field fractions are much larger than the other types (as is often the case), that suggests that most of the binding energy is coming from hydrophobic interactions.
- Export it to an external file. The model can then be used in other projects or applications. To do so, click Export, and use the file selector that is displayed to name the file. The model is exported with a `.qsar` extension. Along with it, the ligands are exported to a file with the same base and a `_qsar_pred.mae` extension. The QSAR Set property is included in the ligand file, so you have a record of which ligands were used for training and test sets.

Running Field-Based QSAR from the Command Line

The field-based QSAR models can be built and tested from the command line with the utility `phase_fqsar`. The command syntax is

`phase_fqsar` *inFile outFile actProp* `-build|-test` [*options*]

On Linux and Mac hosts, you should open a terminal window, set the `SCHRODINGER` environment variable and prepend this command with `$SCHRODINGER/utilities/`. On Windows, in a Schrodinger Command Prompt shell, there is no need to add a prefix.

The input file, *inFile* is a structure file in Maestro or SD format, compressed or uncompressed, that contains the training set structures, test set structures, or both. The output file *outFile* contains the same structures with predicted activities added, and is in the same format as the input file.

actProp is the name of the experimental activity property in the input file. This is normally a quantity that is linearly related to free energy, such as pKi or pIC50.

The utility has two modes, `-build`, to build a new model, and `-test`, to test an existing model. For information on the command options, run the command `phase_fqsar -h`.

4.1 Using QM Electrostatic Fields

Quantum mechanical electrostatic fields can be generated with Jaguar, using the script `jag_qsar.py`. This script runs both `phase_fqsar` and Jaguar to generate the input, run the Jaguar jobs, and build the model. The command syntax is:

`run -FROM phase jag_qsar.py` [*options*] *inFile actProp* `'"`*buildOpt*`"'`

On Linux and Mac hosts, you should open a terminal window, set the `SCHRODINGER` environment variable and prefix this command with `$SCHRODINGER/`. On Windows, in a Schrodinger Command Prompt shell, there is no need to add a prefix.

The input file, *inFile*, can be a Maestro or an SD file that contains the training set structures, and may also contain test set structures.

The activity property, *actProp*, is the name of the activity property in the input structure file, e.g. `r_user_Activity` for a Maestro file. It should be one that is linearly related to the free energy, such as pK_i or pIC50.

The build options are options for the `phase_fqsar` utility that are needed to build a model with Jaguar electrostatic fields. You will need to include at least these options:

`-build -style qm_e -qmgrid` *gridFile* `-qmjob` *jaguarJob*

and any other options for building the model. These options can be displayed by using the option `-help_fqsar`, or with `phase_fqsar - h` (see above).

The other options that can be set include the Jaguar density functional and basis set, and a host option to distribute the Jaguar jobs over multiple processors. For a description of the options, run the script with the `-h` option. The output includes the Jaguar grid files that contain the electrostatic fields, and a file that lists the locations of these grid files.

To apply a model with QM electrostatic fields:

1. Make a backup copy of the QM grid file from the model building run (the file specified with `-qmgrid` *gridFile*).
1. Run `phase_fqsar` with the options `-build`, `-qmgrid`, and `-qmjob`, and with `qm_e` added to the `-style` settings. For example,

   ```
   phase_fqsar unknowns.mae results-out.mae.gz r_user_Activity -build
       -style qm_e -factors 5 -pt 0.75 -rand 123456789
       -qmgrid jag_grid.txt -qmjob jag_job
   ```

   This run simply creates the Jaguar input files needed to generate the electrostatic potentials. It does not build a model, even though `-build` is specified.
2. Replace the grid file from the new run with the backup copy of the model-building grid file.
3. If you used the options for the Jaguar basis set or density functional, edit the Jaguar input files to add the `basis` and `dftname` keywords.
4. Run the Jaguar jobs to generate the electrostatic potentials for your structures.

   `jaguar batch JOBS` *qmJobName*`_*.in`

   This job runs the Jaguar jobs sequentially and locally. If you want to run them on another host and on multiple processors, you can add a `-HOST` option to the command, but the host must have access to the directory that contains the input files. See Section 11.2.2 of the *Jaguar User Manual* for more information.
5. Run `phase_qsar` in test mode (with the `-test` option) to generate results for the new molecules.

References

1. Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

2. Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices Analysis in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.

3. Klebe, G.; Abraham, U. Comparative Molecular Similarity Index Analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 1–10.

4. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem A* **1998**, *102*, 3762–3772.

Getting Help

Information about Schrödinger software is available in two main places:

- The `docs` folder (directory) of your software installation, which contains HTML and PDF documentation. Index pages are available in this folder.
- The Schrödinger web site, http://www.schrodinger.com/, In particular, you can use the Knowledge Base, http://www.schrodinger.com/kb, to find current information on a range of topics, and the Known Issues page, http://www.schrodinger.com/knownissues, to find information on software issues.

Finding Information in Maestro

Maestro provides access to nearly all the information available on Schrödinger software.

To get information:

- Pause the pointer over a GUI feature (button, menu item, menu, ...). In the main window, information is displayed in the Auto-Help text box, which is located at the foot of the main window, or in a tooltip. In other panels, information is displayed in a tooltip.

  If the tooltip does not appear within a second, check that Show tooltips is selected under General → Appearance in the Preferences panel, which you can open with CTRL+, (⌘,). Not all features have tooltips.

- Click the Help button in the lower right corner of a panel or press F1, for information about a panel or the tab that is displayed in a panel. The help topic is displayed in the Help panel. The button may have text or an icon:

  

- Choose Help → Online Help or press CTRL+H (⌘H) to open the default help topic.
- When help is displayed in the Help panel, use the navigation links in the help topic or search the help.
- Choose Help → Documentation Index, to open a page that has links to all the documents. Click a link to open the document.

- Choose Help → Search Manuals to search the manuals. The search tab in Adobe Reader opens, and you can search across all the PDF documents. You must have Adobe Reader installed to use this feature.

For information on:

- Problems and solutions: choose Help → Knowledge Base or Help → Known Issues → *product*.
- New software features: choose Help → New Features.
- Python scripting: choose Help → Python Module Overview.
- Utility programs: choose Help → About Utilities.
- Keyboard shortcuts: choose Help → Keyboard Shortcuts.
- Installation and licensing: see the *Installation Guide*.
- Running and managing jobs: see the *Job Control Guide*.
- Using Maestro: see the *Maestro User Manual*.
- Maestro commands: see the *Maestro Command Reference Manual*.

Contacting Technical Support

If you have questions that are not answered from any of the above sources, contact Schrödinger using the information below.

Web:	http://www.schrodinger.com/supportcenter
E-mail:	help@schrodinger.com
Mail:	Schrödinger, 101 SW Main Street, Suite 1300, Portland, OR 97204
Phone:	+1 888 891-4701 (USA, 8am – 8pm Eastern Time)
	+49 621 438-55173 (Europe, 9am – 5pm Central European Time)
Fax:	+1 503 299-4532 (USA, Portland office)
FTP:	ftp://ftp.schrodinger.com

Generally, using the web form is best because you can add machine output and upload files, if necessary. You will need to include the following information:

- All relevant user input and machine output
- **Product** purchaser (company, research institution, or individual)
- Primary **Product** user
- Installation, licensing, and machine information as described below.

Gathering Information for Technical Support

The instructions below describe how to gather the required machine, licensing, and installation information, and any other job-related or failure-related information, to send to technical support. Where the instructions depend on the profile used for Maestro, the profile is indicated.

For general enquiries or problems:

1. Open the Diagnostics panel.
   - **Maestro:** Help → Diagnostics
   - **Windows:** Start → All Programs → Schrodinger-2015-2 → Diagnostics
   - **Mac:** Applications → Schrodinger2015-2 → Diagnostics
   - **Command line:** `$SCHRODINGER/diagnostics`
2. When the diagnostics have run, click Technical Support.

   A dialog box opens, with instructions. You can highlight and copy the name of the file.
3. Upload the file specified in the dialog box to the support web form.

   If you have already submitted a support request, use the upload link in the email response from Schrödinger to upload the file. If you need to submit a new request, you can upload the file when you fill in the form.

If your job failed:

1. Open the Monitor panel, using the instructions for your profile as given below:
   - **Maestro/Jaguar/Elements:** Tasks → Monitor Jobs
   - **BioLuminate/MaterialsScience:** Tasks → Job Monitor
2. Select the failed job in the table, and click Postmortem.

   The Postmortem panel opens.
3. If your data is not sensitive and you can send it, select Include structures and deselect Automatically obfuscate path names.
4. Click Create.

   An archive file is created, and an information dialog box with the name and location of the file opens. You can highlight and copy the name of the file.
5. Upload the file specified in the dialog box to the support web form.

   If you have already submitted a support request, use the upload link in the email response from Schrödinger to upload the file. If you need to submit a new request, you can upload the file when you fill in the form.

6. Copy and paste any log messages from the window used to start the interface or the job into the web form (or an e-mail message), or attach them as a file.
   - **Windows:** Right-click in the window and choose Select All, then press ENTER to copy the text.
   - **Mac:** Start the Console application (Applications → Utilities), filter on the application that you used to start the job (Maestro, BioLuminate, Elements), copy the text.

If Maestro failed:

1. Open the Diagnostics panel.
   - **Windows:** Start → All Programs → Schrodinger-2015-2 → Diagnostics
   - **Mac:** Applications → SchrodingerSuite2015-2 → Diagnostics
   - **Linux/command line:** `$SCHRODINGER/diagnostics`
2. When the diagnostics have run, click Technical Support.

   A dialog box opens, with instructions. You can highlight and copy the name of the file.
3. Upload the file specified in the dialog box to the support web form.

   If you have already submitted a support request, use the upload link in the email response from Schrödinger to upload the file. If you need to submit a new request, you can upload the file when you fill in the form.
4. Upload the error files to the support web form.

   The files should be in the following location:

- **Windows:** `%LOCALAPPDATA%\Schrodinger\appcrash`
  (Choose Start → Run and paste this location into the Open text box.)
  Attach `maestro_error_`*`pid`*`.txt` and `maestro.exe_`*`pid_timestamp`*`.dmp`.
- **Mac:** `$HOME/Library/Logs/CrashReporter`
  (Go → Home → Library → Logs → CrashReporter)
  Attach `maestro_error_`*`pid`*`.txt` and `maestro_`*`timestamp_machinename`*`.crash`.
- **Linux:** `$HOME/.schrodinger/appcrash`
  Attach `maestro_error_`*`pid`*`.txt` and `crash_report_`*`timestamp_pid`*`.txt`.

If a Maestro panel failed to open:

1. Copy the text in the dialog box that opens.
2. Paste the text into the support web form.

120 West 45th Street
17th Floor
New York, NY 10036

101 SW Main Street
Suite 1300
Portland, OR 97204

245 First Street
Riverview II, 18th Floor
Cambridge, MA 02142

8910 University Center Lane
Suite 270
San Diego, CA 92122

155 Gibbs St
Suite 430
Rockville, MD 20850-0353

Dynamostraße 13
D-68165 Mannheim
Germany

Zeppelinstraße 73
D-81669 München
Germany

Potsdamer Platz 11
D-10785 Berlin
Germany

Quatro House
Frimley Road
Camberley GU16 7ER
United Kingdom

8F Pacific Century Place
1-11-1 Marunouchi
Chiyoda-ku, Tokyo 100-6208
Japan

No. 102, 4th Block
3rd Main Road, 3rd Stage
Sharada Colony
Basaveshwaranagar
Bangalore 560079, India

SCHRÖDINGER®