Phase 4.3

User Manual

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 to Phase

Phase is a versatile product for pharmacophore perception, structure alignment, activity prediction, and 3D database searching. Given a set of molecules with high affinity for a particular protein target, Phase uses fine-grained conformational sampling and a range of scoring techniques to identify common pharmacophore hypotheses, which convey characteristics of 3D chemical structures that are purported to be critical for binding. Each hypothesis is accompanied by a set of aligned conformations that suggest the relative manner in which the molecules are likely to bind.

A given hypothesis may be combined with known activity data to create a 3D QSAR model that identifies overall aspects of molecular structure that govern activity. This model may be used in conjunction with the hypothesis to mine a 3D database for molecules that are most likely to exhibit strong activity toward the target.

Phase provides support for lead discovery, SAR development, lead optimization and lead expansion. Phase may also be used as a source of molecular alignments for third-party 3D QSAR programs.

Phase is integrated into Maestro, the graphical user interface (GUI) for all Schrödinger products. For more detailed information on Maestro, see the Maestro online help or the *Maestro User Manual*. An overview of the Phase interface is given in Chapter 2.

Phase is fully supported on Linux, Windows, and Mac platforms.

For a tutorial introduction to Phase, see the *Phase Quick Start Guide*. For installation instructions, see the *Installation Guide*.

1.1 Phase Workflows

Phase consists of the following four workflows:

- Building a pharmacophore model (and an optional QSAR models) from a set of ligands
- Building a pharmacophore hypothesis from a single ligand (and editing it)
- Preparing a 3D database that includes pharmacophore information
- Searching the database for matches to a pharmacophore hypothesis

Each of these workflows is supported by a Maestro panel. The first workflow, building a pharmacophore model, involves the following steps:

- Preparing the ligands, including 2D-3D structure conversion and the generation of ligand conformations. This step is described in detail in Chapter 3.
- Defining and identifying the pharmacophore sites in the ligands. This step is described in detail in Chapter 4.
- Creating hypotheses by finding common pharmacophores. This step is described in detail in Chapter 5.
- Scoring the hypotheses, and adding any excluded volumes to the hypotheses. This step is described in detail in Chapter 6.
- Building and examining 3D QSAR models. This step is described in detail in Chapter 7.

Building a pharmacophore hypothesis from one or more ligands manually is an alternative to building a pharmacophore model from a set of ligands by the automated process described above. Details of this task can be found in Chapter 8.

Preparing the 3D database involves the following tasks, which are described in Chapter 10:

- Preparing the molecules, including 2D-3D conversion
- Adding molecules to the database

Depending on how you want to use the database, you can perform the following tasks, which are also described in Chapter 10:

- Generating conformations for each molecule
- Defining and identifying the pharmacophore sites
- Creating subsets

Searching the 3D database for matches to a hypothesis includes various filtering and scoring mechanisms. This workflow is described in Chapter 11.

You can also run these three workflows from the command line, as described in Chapter 12 and Chapter 13. Chapter 14 describes searching a file for matches, rather than a 3D database.

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 Starting Jobs from the Maestro Interface

To run a job from the Maestro interface, you open a panel from one of the menus (e.g. Tasks), make settings, and then submit the job to a host or a queueing system for execution. The panel settings are described in the help topics and in the user manuals. When you have finished making settings, you can use the Job toolbar to start the job.

You can start a job immediately by clicking Run. The job is run on the currently selected host with the current job settings and the job name in the Job name text box. If you want to change the job name, you can edit it in the text box before starting the job. Details of the job settings are reported in the status bar, which is below the Job toolbar.

If you want to change the job settings, such as the host on which to run the job and the number of processors to use, click the Settings button. (You can also click the arrow next to the button and choose Job Settings from the menu that is displayed.)

You can then make the settings in the Job Settings dialog box, and choose to just save the settings by clicking OK, or save the settings and start the job by clicking Run. These settings apply only to jobs that are started from the current panel.

If you want to save the input files for the job but not run it, click the Settings button and choose Write. A dialog box opens in which you can provide the job name, which is used to name the files. The files are written to the current working directory.

The Settings button also allows you to change the panel settings. You can choose Read, to read settings from an input file for the job and apply them to the panel, or you can choose Reset Panel to reset all the panel settings to their default values.

You can also set preferences for all jobs and how the interface interacts with the job at various stages. This is done in the Preferences panel, which you can open at the Jobs section by choosing Preferences from the Settings button menu.

Note: The items present on the Settings menu can vary with the application. The descriptions above cover all of the items.

The icon on the Job Status button shows the status of jobs for the application that belong to the current project. It starts spinning when the first job is successfully launched, and stops spinning when the last job finishes. It changes to an exclamation point if a job is not launched successfully.

Clicking the button shows a small job status window that lists the job name and status for all active jobs submitted for the application from the current project, and a summary message at the bottom. The rows are colored according to the status: yellow for submitted, green for launched, running, or finished, red for incorporated, died, or killed. You can double-click on a row to open the Monitor panel and monitor the job, or click the Monitor button to open the Monitor panel and close the job status window. The job status is updated while the window is open. If a job finishes while the window is open, the job remains displayed but with the new status. Click anywhere outside the window to close it.

Jobs are run under the Job Control facility, which manages the details of starting the job, transferring files, checking on status, and so on. For more information about this facility and how it operates, as well as details of the Job Settings dialog box, see the *Job Control Guide*.

The Develop Common Pharmacophore Hypotheses workflow does not have a Job toolbar, as it has its own buttons for running the jobs that are part of the workflow.

1.4 Citing Phase in Publications

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

Phase, version 4.3, Schrödinger, LLC, New York, NY, 2015.

Please also cite the following reference:

Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A., “PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and

3D Database Screening. 1. Methodology and Preliminary Results," *J. Comput. Aided Mol. Des.*, **2006**, *20*, 647-671.

Running Phase from Maestro

Phase consists of the following workflows, each of which is supported by a Maestro panel:

- Building a pharmacophore model and an optional QSAR model
- Preparing a 3D database that includes pharmacophore information
- Building or editing pharmacophore hypotheses
- Searching the database for matches to a pharmacophore hypothesis

An overview of each of these workflows is given in the sections below, along with an overview of the supporting Maestro panel. The stages are described in detail in the following chapters.

The Maestro interface for the first workflow is wizard-like, and takes you through each step of the process. The interface is more flexible than a wizard, however, because it allows you to exit at any step and resume the process later at the same step-or a different step, provided that you have the data to support that step. Default options are provided that should give good results, but you can also choose from a range of other options to suit your purposes. The interface for the remaining three workflows are single panels.

To open any of the panels, choose the appropriate item from the menu in the main window:

- Applications → Phase
- Tasks → Pharmacophore Modeling

All Phase jobs can be started from Maestro. Many of these jobs generate a large amount of data, and most of them can be distributed across multiple processors. When you click a Start button or an action button that starts a job, a panel is displayed that allows you to set job control information.

Phase can also be run from the command line. For information on command-line use, see Chapter 12, Chapter 13, and Chapter 14.

2.1 Developing a Pharmacophore Model

Developing a pharmacophore model from a set of active molecules is the main means in Phase of generating pharmacophore hypotheses, which are subsequently used in database searching.

In Phase, there are five steps in the process of developing a pharmacophore model: preparing the ligands, creating pharmacophore sites from a set of features, finding common pharmacophores, scoring the hypotheses, and building a QSAR model.

The process of developing a pharmacophore model is called a *run*. The data for each run is stored as a separate entity, which you can open from the File menu. When a new run is created, links are made to common data, to avoid duplication. The run is stored as part of a Maestro project.

2.1.1 General Panel Layout

The Develop Common Pharmacophore Hypotheses panel is wizard-like in design, with five steps. Each step occupies the center of the panel, and consists of a title with a brief description of the step at the top, a set of controls and tables for results in the center, and a Back and a Next button at the bottom. The features of each step are described in detail in the online help.

In addition to the step features, the panel contains a menu bar, a toolbar, and a job status button at the top; and a step guide (the Guide) at the bottom. These features are described below.

The File Menu

The File menu allows you to work with the runs that are available in the project.

New	Create a new run
Open	Open an existing run from the submenu. If there are more than 4 runs, choose More to open a dialog box and select a run.
Save As	Save the current run with a new name
Rename	Rename the current run
Delete Run	Delete the current run

The Display Menu and the Toolbar

The Display menu provides options for viewing hypotheses and related attributes in the Workspace. These options are also available as toolbar buttons, and are described below. The items are only available in Step 4 (Score Hypotheses) and Step 5 (Build QSAR Model).

Hypothesis — Displays the selected hypothesis as a spatial arrangement of feature symbols. For a description of these symbols, see Table 4.1 on page 34.

Hypothesis Labels — Displays feature labels for the selected hypothesis.

Excluded Volumes — Displays excluded volumes for the selected hypothesis.

QSAR Model — Displays the selected QSAR model for the hypothesis. Only available in the Build QSAR Model step.

Site Measurements — Opens the View Site Measurements panel, in which you can select the intersite distances and angles of the hypothesis for display in the Workspace.

This menu also has an item for displaying ligand properties in the Workspace, Workspace Feedback. This item opens the Phase Workspace Feedback dialog box, in which you can turn feedback on and off, choose the properties to display from those available in the current Ligands or Alignments table, and choose whether to display property names. The settings apply to the current project, but can be stored as preferences by clicking Save Preferences in the dialog box. Separate settings can be made for each step.

The Step Menu

The Step menu contains an item to display or hide the Guide, and items for each of the steps. The current step is marked with a red diamond. If the Guide is displayed, it is marked with a red square. The items for the steps that are not available are dimmed. You can go to any available step by choosing the corresponding menu item.

The Job Status Button

When a job has been launched and is running, the job status button at the upper right of the panel turns green and the icon spins. When the job stops, the button turns pink and the icon stops spinning. It is replaced by an exclamation point if the job is incorporating, and returns to the original when incorporation has finished. To monitor the job using the Monitor panel, click the button. For more information about monitoring jobs, see the *Job Control Guide*.

The Guide

The Guide displays the steps in the model as a set of buttons linked by lines. The buttons for the steps that are not available are dimmed. The current step is highlighted with a white background. You can go to any available step by clicking its button in the Guide. The Guide can be displayed or hidden from the Step menu. If you go back to an earlier step and make changes, you are prompted to save the existing data and create a new run with the changed data.

To navigate the steps, you can click the Back and Next buttons, click the desired step in the Guide, or choose the desired step from the Step menu.

2.1.2 The Prepare Ligands Step

In this step, you add to the run the molecules that you want to use as the basis for the pharmacophore model and any other molecules that you want to use to build or test the QSAR model, and you select both active and inactive molecules for the set that is used for the pharmacophore model. If you include activity data with the ligands when you add them, you can select the active and the inactive sets of ligands either using cutoffs or manually.

To develop a pharmacophore model, you should ensure that you have all-atom 3D structures, and generate different conformations for each structure. If the structures need to be converted from 2D to 3D or otherwise need cleaning up, you can do so in this step. You can also convert the structures to the most probable ionization (protonation) state at a given pH, and generate different chiralities for the structures in this step. The molecules that you add are automatically grouped into conformer sets. You can choose to group stereoisomers in the same set or different sets. If you add only one conformer for a given molecule, you can generate the rest in this step.

Once you have added the molecules, cleaned up the structures, generated conformers, and selected the set of ligands, you can proceed to the next step.

2.1.3 The Create Sites Step

In this step, you use a set of chemical structure patterns to identify pharmacophore features in each ligand. Once a feature has been mapped to a specific location in a conformation, it is referred to as a *pharmacophore site*. The number of occurrences of each feature in each ligand is tabulated, and you can display the locations of the pharmacophore sites in the Workspace.

While the built-in set of features is adequate for many purposes, you might want to add new patterns to the built-in features, ignore patterns in the built-in features, or add custom features. You can add functional groups, defined as SMARTS patterns, to the definition of a feature. You can also designate functional groups that should be excluded from consideration as part of a feature, and you can choose to ignore functional groups.

2.1.4 The Find Common Pharmacophores Step

In this step, you perform a search for common pharmacophores among the set of high-affinity (active) ligands that you chose in the first step. The search spans one or more families of pharmacophores, known as *variants*. You can choose the number of site points in the pharmacophore, filter out variants that have too many or too few of a particular kind of feature, and select a set of variants from the filtered list. You can also set a lower limit on the number of ligands that must match a pharmacophore before it can be considered to be a hypothesis.

The search proceeds by enumerating all pharmacophores of a given variant and partitioning them into successively smaller high-dimensional boxes according to their intersite distances. Each n-point pharmacophore contains $n(n–1)/2$ unique intersite distances, so each box contains $n(n–1)/2$ dimensions. Pharmacophores that are clustered into the same box are considered to be equivalent and therefore common to the ligands from which they arise. The size of the box defines the tolerance on each intersite distance, and therefore how similar common pharmacophores must be. You can set parameters to control the minimum box size and you can exclude pharmacophores for which any intersite distance is below a certain threshold.

Boxes that contain pharmacophores from the minimum required number of ligands are said to *survive* the partitioning process. Each surviving box contains a set of common pharmacophores, one of which is ultimately singled out as a hypothesis.

Once all desired variants have been processed, you can continue to the scoring step.

2.1.5 The Score Hypotheses Step

In this step you apply a scoring function that identifies the best candidate hypothesis from each surviving box, and provides an overall ranking of all the hypotheses. You can finish at this point, or select hypotheses for the generation of QSAR models and continue to the next step, or select hypotheses and proceed to find matches to the hypotheses. You can also add to the hypothesis volumes that should not be occupied by atoms in any active molecule, known as *excluded volumes*.

The scoring algorithm includes contributions from the alignment of site points and vectors, volume overlap, selectivity, number of ligands matched, relative conformational energy, and activity. You can adjust these in the survival score, and you can create a custom score. You can also penalize hypotheses by scoring inactives and subtracting a multiple of this score from the survival score. The scores for each hypothesis are displayed in a table. You can select a hypothesis and view scores for the ligands that match the hypothesis, and the energy of the ligand relative to the lowest conformation.

If you have both active and inactive molecules that match a hypothesis, you can use their structures to define excluded volumes. Any region of space that is occupied by part of an inactive molecule and is not occupied by the active molecules is a good candidate for an excluded volume. The excluded volumes are used to filter out molecules in the database search that are likely to be inactive.

When you have selected one or more hypotheses, you can proceed to the next step.

2.1.6 The Build QSAR Model Step

In this step, you build QSAR models for the selected hypotheses using the activity data for molecules that match at least three points in the hypothesis. You can use molecules with any level of activity, including those that may be inactive due to steric clashes with the target receptor. The QSAR model partitions space into a grid of uniformly sized cubes, and characterizes each molecule by a set of binary-valued independent variables that encode the occupancy of these cubes by six atom classes or a set of pharmacophore feature types. Partial least-squares regression is applied to these variables to build a series of models with successively greater numbers of factors.

You can visualize the QSAR model in the Workspace, and analyze it by atom or feature class and ligand. This can be used to identify ligand features that contribute positively or negatively to the predicted activity.

When you have developed QSAR models, you can continue to the database search and use the QSAR models to predict activities for matches, or return to the previous step to select another set of hypotheses for QSAR model development.

2.2 Building or Editing Hypotheses

As an alternative to building a pharmacophore model, you can build pharmacophore hypotheses directly from known active molecules, in the Edit Hypotheses panel. In this task, Phase identifies the possible pharmacophore sites in the molecule you select, based on a set of pharmacophore feature definitions. You then select the features that you want to include in the hypothesis. No jobs are run in this workflow.

If you create a hypothesis from a known receptor or receptor-ligand complex, you can use the receptor to automatically generate excluded volumes. This task is performed in the Excluded Volume Receptor panel.

2.3 Preparing a 3D Database for Searching

The Manage 3D Database panel provides tools for preparing a structure database that can be searched for matches to a pharmacophore hypothesis. The database must contain all-atom 3D structures that are reasonable representations of the experimental structures. Preparing a database involves adding structures, cleaning the structures if necessary, generating conformers if necessary or desired, creating pharmacophore sites from selected features, and creating subsets of molecules for database searching as desired.

Databases are not connected to Maestro projects.

The Manage 3D Database panel is a single panel, with a menu bar and an octagon button at the top. When a job has been launched and is running, the gray octagon at the upper right of the panel turns green and spins. To monitor the job using the Monitor panel, click the octagon. For more information about monitoring jobs, see the *Job Control Guide*.

2.4 Finding Matches to a Hypothesis

The Find Matches to Hypothesis panel is a single panel, with four sections. In the top two sections, you specify the database to search and the hypothesis to use in the search. In the bottom two sections, you set parameters for the search and for the subsequent display of hits.

The search is performed in two steps: finding matches to the hypothesis, and fetching hits. The second step can be repeated with different processing options without repeating the first step. The processing options include adjusting the fitness score, by which hits are sorted, applying numerical cutoffs on the number of hits, applying excluded volumes to filter hits, and calculating activities using the QSAR model, if one was generated for the hypothesis.

2.5 Running Jobs

When you click an action button that is associated with a job or a Start button, the Start dialog box opens. In this dialog box, you can select the host on which you want to run the job, set the user name on that host, if it differs from the user name on the host on which you are running Maestro, and enter the number of processors to use for the job. The maximum number of processors available on the selected host is displayed in parentheses after the host name. For more information on this dialog box, see Section 2.2 of the *Job Control Guide*.

Phase jobs run under the Schrödinger job control facility. This facility allows you to monitor the progress of jobs within Maestro, both local and remote. It also provides the list of hosts in the Start panel from which you make a selection when you start a job.

The list of hosts is read from the `schrodinger.hosts` file, which is installed in the `$SCHRODINGER` directory. At installation time, this file should be set up to define the hosts on which Schrödinger software will be run. Instructions for setting up this file are given in the *Installation Guide*. You can copy this file to your home directory to customize it.

The time-consuming parts of Phase can be distributed across multiple processors. You can set up multiprocessor hosts in the `schrodinger.hosts` file, either as hosts on which you run jobs directly or as batch queues, with a specified number of processors. If you run a database search on a multiprocessor host, such as a cluster, the following requirements must be met:

- The database must be located in a directory that is uniformly accessible to all nodes of the cluster on which jobs will be run.

- If the file system where the database is stored is only accessible to the cluster, you must run Maestro on the manager node of the cluster to launch jobs.
- In the `$SCHRODINGER/schrodinger.hosts` file, each parallel queue that is used for database jobs should have a `tmpdir` entry with a path that is accessible to all nodes. For details of setting up these entries, see the *Installation Guide* or the *Job Control Guide*.

2.6 Preferences

Some options that affect the use of Phase can be set as preferences or with Maestro commands. You can add these options to the `maestro.cmd` file in your user profile directory. See Chapter 13 of the *Maestro User Manual* for more information about this file and the user profile directory.

- **Default custom feature definition file**—To define a default feature definition file other than that in the software distribution, you can set a Maestro preference. The command for setting this preference is as follows:

  `prefer phasedefaultfeaturedefinitions=`*filename*

  The file name can be either an absolute path or a relative path. If you specify a relative path, Maestro will look for the file at this location relative to its current working directory. You can enter this command into the Commands text box in the main window, and it will be added to your preferences when Maestro exits. Alternatively, you can add this command to `maestro.cmd` in your user profile directory.

- **Color of Phase labels in the Workspace**—You can set the color of Phase labels with a Maestro command, `phasemarkersettings`. This command controls all aspects of Phase markers, including feature colors and sizes, label colors, and so on. For Phase labels, you can use the following command:

  `phasemarkersettings labelred=`*r* `labelgreen=`*g* `labelblue=`*b*

  where *r*, *g*, and *b* are the fractions of the red, green, and blue components of the label color, expressed as a real number between 0 and 1. For more information on this command, see the *Maestro Command Reference Manual*. This command is not written as a preference, and must be added to `maestro.cmd` in your user profile directory.

Preparing Ligands for Pharmacophore Hypothesis Development

The first step in developing a pharmacophore model is to select the molecules that you want to use and to prepare them for use. This step is performed in the Prepare Ligands step of the Develop Common Pharmacophore Hypotheses panel.

The molecules you select should at a minimum include highly active molecules that you want to use as the basis of the pharmacophore model. You can also include inactive or moderately active molecules, which can be used to test pharmacophore hypotheses for specificity, for building and testing a QSAR model, and for the purpose of defining excluded volumes. When you add molecules, you can select an associated activity property to use for activity scoring and for the dependent variable in the QSAR model. This property can also be used to define the active and inactive molecules to use as the basis for the pharmacophore model.

Developing a pharmacophore model requires all-atom 3D structures that are realistic representations of the experimental molecular structure. Most ligands are flexible, so it is important to consider a range of conformations in order to increase the chances of finding something close to the bound structure.

Under some circumstances it can be important to generate variations on the input structures, such as varying the chirality or choosing the most probable protonation state. Varying the chirality of atoms in the molecules can be important if the chiralities are not known. The process of identifying common pharmacophores can then sample the different stereoisomers and locate the one that matches best. Also, if the input structure is not in the most common form at physiological pH values, the identification of common pharmacophores might give incorrect results, because the active form is protonated or deprotonated.

In the Prepare Ligands step, you add the molecules with their activity values to the Phase run. If you have 2D structures or united-atom 3D structures, you can convert them to all-atom 3D structures and locate the minimum energy structure using molecular mechanics. In the process, you can vary the chirality of atoms in the structures and assign the protonation state. Once you have the structures and any variations, you can generate the low-energy conformers for each structure. The conformers are automatically grouped into sets for each molecule. If you already have all-atom, 3D structures with their conformations, you must still pass them through the ligand preparation steps, so that Phase can verify that there are no unusable structures and that the geometry of the structures is sufficiently accurate.

The tasks involved in this step and how to accomplish them are described in detail below.

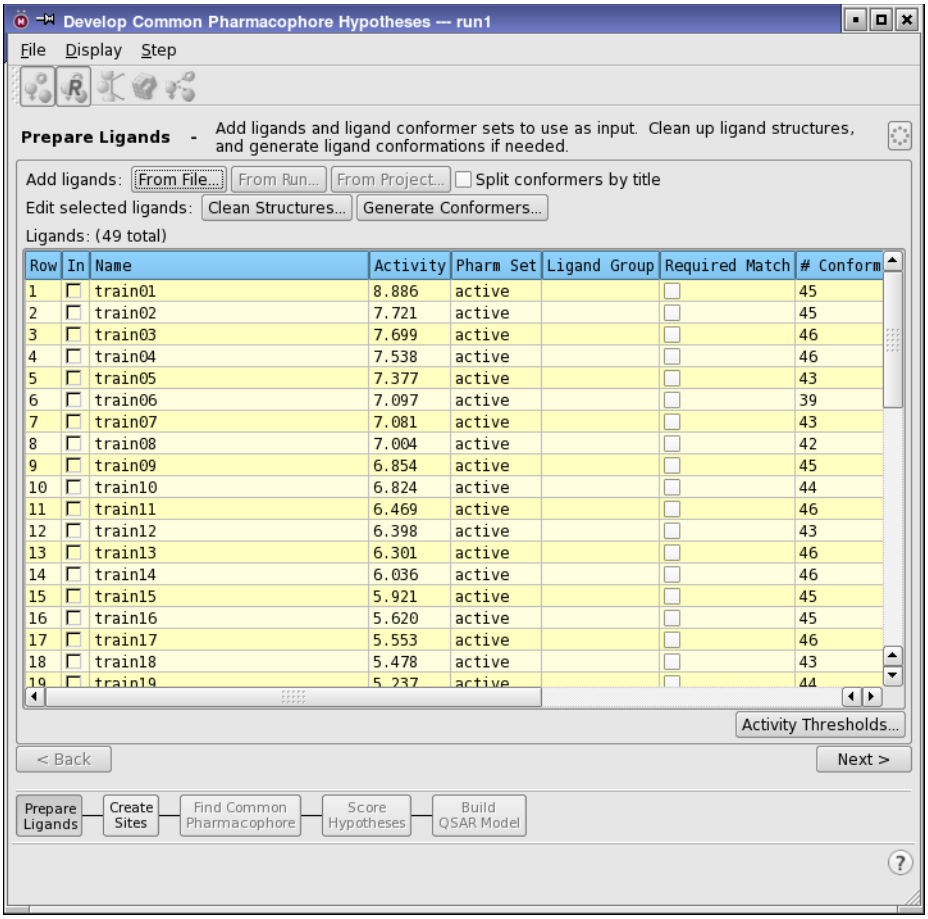

Figure 3.1. The Prepare Ligands step, with ligands.

3.1 Adding Ligands to a Run

Each “ligand” in Phase is actually a set of conformations of a ligand structure. When you add ligands, they are automatically grouped into conformer sets. Stereoisomers are not considered to be separate ligands by default, although the activities of stereoisomers can be different. You can merge or separate ligand conformer sets by stereoisomer using the Ligands table shortcut menu (see Table 3.2). You can also separate ligand conformer sets on import using the title, by selecting Split conformers by title. This feature is useful if, for example, the stereochemical data is not recognized and you want to separate stereoisomers. If you add ligands from a previous run, the sets are preserved with all their associated data. If you do not have conformations for the ligands you add, you can generate them in this step.

You can add ligands to a Phase run from a file, from a previous run, or from a Maestro project. To add ligands, click one of the Add Ligands buttons. Each button opens a dialog box in which you can read or copy the ligands. Activity data can be added with the ligands. Once you have added the ligands, they are displayed in the Ligands table. If the ligands do not have activity data, you can add the data by editing the table cells.

If you want to delete ligands, select them in the table, then right-click in the table and choose Delete from the contextual menu.

3.1.1 Adding Ligands From a File

You can read ligands directly from a file into the Phase run, without importing them into Maestro. To do so, click From File. A file selector opens, so that you can navigate to and select one or more files. You can filter the list of files displayed by choosing Custom File Extension from the Files of type option menu. Only Maestro format is supported, and properties are read with the structures. When you click Open, the Choose Activity Property dialog box opens. This dialog box contains a list of properties, with tools for filtering and sorting the list. You can choose a single property for the activity of the ligands, and convert the activity to a logarithmic scale if necessary. The activity must be a positive quantity that increases with increasing activity.

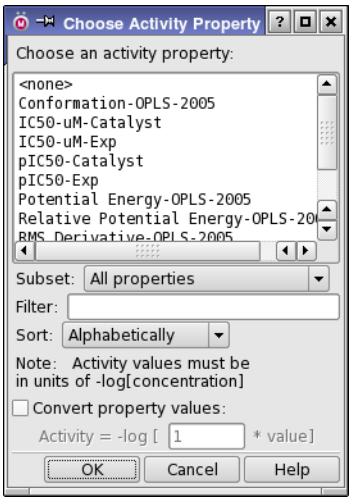

Figure 3.2. The Choose Activity Property dialog box.

3.1.2 Adding Ligands from Another Phase Run

To copy ligands from another Phase run in the current project, click From Run. The Add From Run dialog box is displayed. The dialog box contains a list of all ligands available from all other runs in the current project, with the run name, the ligand name, and the number of conformers. You can choose multiple ligands to add to the current run. The activity values and the membership of the active set are extracted and added with the ligand. If a ligand was used in more than one run, the list of ligands will contain duplicates. If you select duplicates, only one is added, with the activity data from the first run chosen.

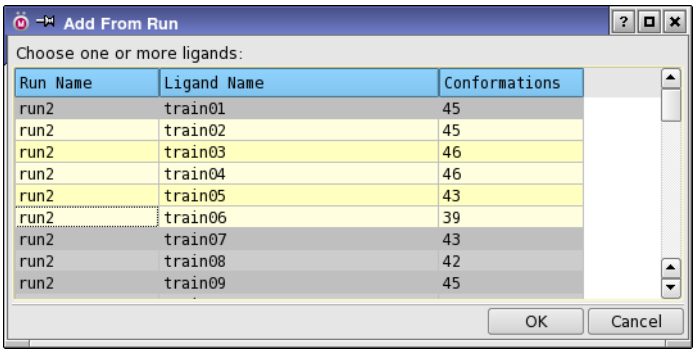

Figure 3.3. The Add From Run dialog box.

3.1.3 Adding Ligands from the Project

If you already have ligands in the Maestro project that you want to use, you can copy them from the project into the Phase run. To do so, click From Project. The Add From Project dialog box is displayed. This dialog box contains two lists: a list of entries, and a list of properties.

You can choose multiple entries to be added to the Ligands table (using shift-click and control-click). The set of entries that is displayed in the entry list in the dialog box is determined by the choice made from the Choose entry from option menu: all entries, selected entries, or included entries. You can sort the list by clicking one of the column headings, or by clicking Sort by Project Table Order. If you want to display some other property in the list, such as an activity property, click Show Property and choose the property from the list in the dialog box that is displayed. You can then sort the entries by the values of this property to aid in your ligand selection. The ligands are copied into the run, so that any changes made by Phase have no effect on the original ligands in the Project Table.

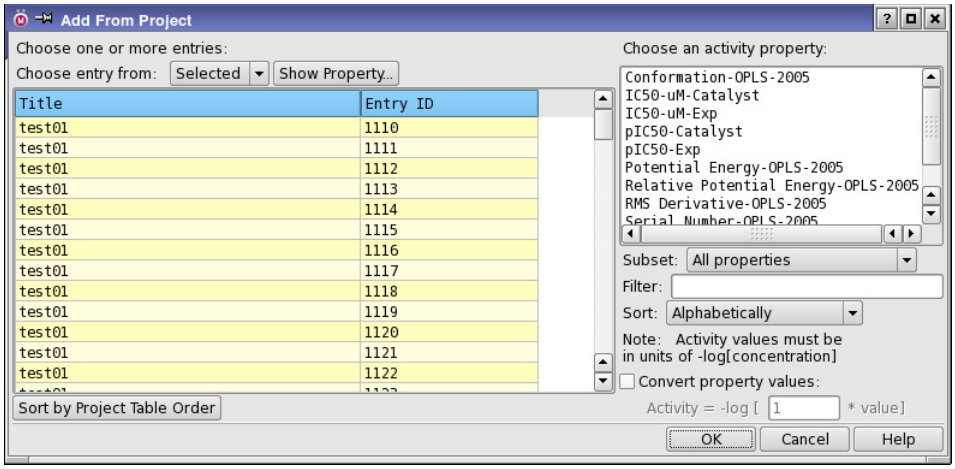

Figure 3.4. The Add From Project dialog box.

You can choose a single property for the activity of the ligands from the Choose an activity property list. The list can be restricted to a given property family by choosing from the Subset option menu, and it can be filtered by entering a text string (including wildcard characters * and ?) in the Filter text box. The ligand activity must be a property that has units of $-\log_{10}[x]$. If the property is in units of concentration [x], you can scale the property values and convert them to a logarithmic scale in this dialog box. The converted values are copied to the Ligands table.

3.2 Cleaning Up Ligand Structures

If the ligand structures are two-dimensional, lack hydrogen atoms, or include counter ions or solvent molecules, you must clean them up before proceeding. If the structures do not have the desired chirality or ionization (protonation) state, or if you want structures with different chirality, you can use the Clean Structures facility to generate them. Clean Structures is an interface to LigPrep with a range of options that is most appropriate for Phase. For more detailed information about the process, see the *LigPrep User Manual*.

In the cleanup process, the following actions are performed as necessary or as requested:

- Convert structures from 2D to 3D
- Add hydrogen atoms to ensure that the structure is an all-atom structure
- Remove counter ions and water molecules
- Add or remove protons to produce the most probable ionization state at the target pH
- Generate stereoisomers
- Remove noncompliant structures
- Perform an energy minimization

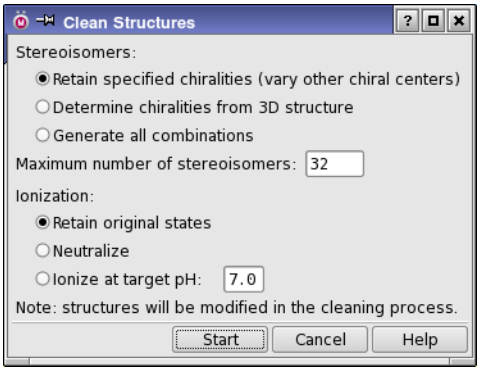

Figure 3.5. The Clean Structures dialog box.

The cleanup process is applied to the ligands that are selected in the Ligands table. You can therefore perform the cleanup with different options for different sets of molecules by making different selections. To clean up the selected structures, click Clean Structures. The Clean Structures dialog box is displayed. In this dialog box, you can set options for generation of stereoisomers and ionization states, then click Start to run the job to perform the cleanup. When you click Start, a dialog box is displayed, in which you can choose the host to run the job. You can distribute this job over multiple hosts.

3.2.1 Generating Stereoisomers

There are three options for generating stereoisomers, described below. For each option, any unspecified chiralities are varied, up to the number given in the Maximum number of stereoisomers text box. When you vary the stereochemistry, the process starts at the configuration with all chiral atoms to be varied set to R, and systematically varies the configuration. If you select fewer stereoisomers than the maximum, there is a chance that you might not generate the most important stereoisomers.

Retain specified chiralities (vary other chiral centers)

If the ligand has chirality information, this information is retained and used to ensure that the chiral atoms all have the correct chiralities. Chirality information includes parities and bond directions from SD files and the chirality property from Maestro files. If the configuration or chirality of one or more chiral centers is not specified, the chiralities for these centers is varied.

Determine chiralities from 3D structures

This option discards any information from the input file and determines the chirality from the 3D geometry. These chiralities are held fixed. For centers whose chirality is indeterminate, the two possible chiralities are generated.

Generate all combinations

This option discards chirality information and generates all possible configurations that result from the combination of chiralities on each chiral center.

3.2.2 Generating Ionization States

In the Ionization section you can choose from three options for generating the appropriate ionization state:

Retain original states

This option bypasses the generation of ionization states. If the ligands all have the correct ionization state for acidic and basic groups, choose this option.

Neutralize

This option converts all acidic and basic groups into their neutral form. For example, zwitterion groups are converted from a carboxylate and an ammonium to a carboxylic acid and an amine.

Ionize at target pH

This option generates the most probable ionization state at the given target pH, for which the default value is 7.

The ionization is performed with the `ionizer`. If you want to use Epik to generate ionization states (or tautomers), you must do so before you add the structures to the Phase run.

3.3 Generating Conformers

Once you have a set of cleaned-up ligands, you can run a conformational search to generate a set of conformers for each ligand. If you already have the conformations you need, you can skip this step. If you have used the option to separate conformers by title, you should not generate conformers, because you might produce conformers that invalidate the separation, and therefore produce erroneous results. You should generate the conformers outside Phase and import them.

To set up parameters for the conformational search, click Generate Conformers. The Generate Conformers dialog box is displayed. The dialog box has options for the search mode and solvation treatment, and allows you to limit the number of conformations generated, either to a specific number or by energy, which is evaluated in aqueous solution with a continuum solvation model. After setting options, click Start to run the conformational search job. A Start dialog box is displayed, in which you can choose the host to run the job, specify the number of

subjobs to run, and distribute this job over multiple processors. When the job finishes, the Ligands table displays the number of conformers generated for each ligand.

Some options have a greater impact than others on the outcome of pharmacophore model development. Options with the greatest impact include the maximum number of conformations, maximum relative energy difference, minimum atomic deviation and number of post-minimization iterations. The default settings—rapid search, distance-dependent dielectric solvation model, and no post-minimization iterations—are likely to be adequate for many purposes. However, for consistency, you should use the same options in the pharmacophore model development as you use in the database search.

The options for controlling the conformational search are described below.

3.3.1 Output Options

Current conformers

When you generate conformers, you can discard the existing conformer set or you can keep it. If you keep the existing set, the new conformations are appended to the set. The set might therefore contain redundant conformers.

Number of conformers per rotatable bond

This value specifies the maximum number of conformers to generate in the conformational search for each rotatable bond in the structure. The number given in the text box is multiplied by the number of rotatable bonds in the molecule to arrive at the maximum number of conformers.

Maximum number of conformers

This value limits the number of conformers returned from the generation process. If the number of conformers generated is higher than this value, a sample of all the conformers generated is returned. If the maximum specified here is lower than that derived from the number of conformers per rotatable bond, the lower value is used.

3.3.2 Search Method, Sampling, and Minimization Options

Conformer generation can be performed with one of two search methods, ConfGen or Mixed MCMM/LMOD. For each method, sampling of conformational space can be done in Rapid or Thorough mode. Experience to date suggests that the final pharmacophore model is not usually significantly improved by a thorough search.

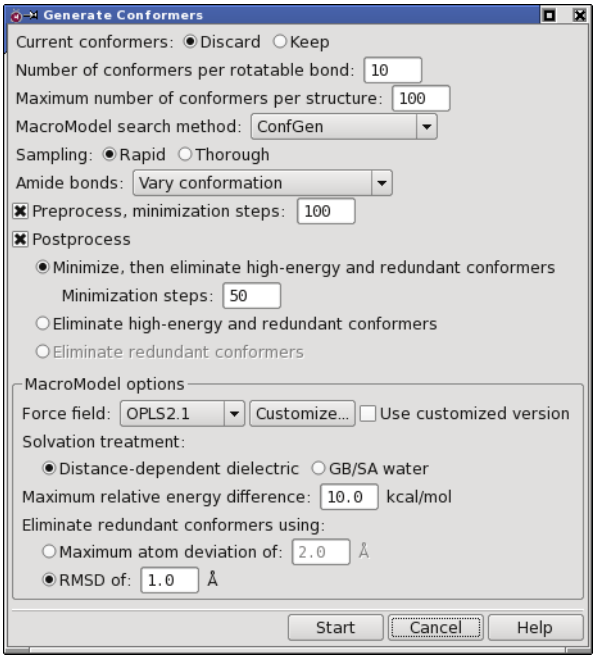

Figure 3.6. The Generate Conformers dialog box.

During the search, hydrogen-bonding interactions are suppressed, because conformations in which the ligand bonds to the receptor are needed in the model, not just conformations with internal hydrogen bonding.

3.3.2.1 ConfGen Search Method

In a ConfGen search, the molecule is divided into a core and a periphery. The peripheral groups have only one rotatable bond between the terminal groups and the rest of the molecule. All the nonperipheral rotatable bonds are assigned to the core. The conformational search generates all core configurations and then varies the peripheral configurations, either one-by-one or in a complete search. The Sampling options have the following meanings:

- Rapid—All the core conformations are generated, and the conformations of the peripheral (rotamer) groups are sampled one by one.
- Thorough—A complete set of conformations is generated for both the core and the peripheral groups.

When this option is selected, the Amide bond option menu is displayed below the MacroModel search method option menu. There are three options for treatment of amide bonds in the search. You can retain the original amide bond conformation in the input structure, you can set

the conformation to trans, or you can vary the conformation. Varying the conformation allows the amide dihedral angle to take any value, not just cis or trans.

The ConfGen search produces structures rapidly, and the conformations generated might not be optimal. For this search method, options for preprocessing and post processing are provided. Both the Preprocess and the Postprocess options perform MacroModel tasks, including minimization, using the options in the MacroModel options section. Preprocessing is done on the input structure, and is confined to minimization with a specified number of iterations. Postprocessing is done on the set of conformers generated by the conformational search, and has three options:

- Minimize, then eliminate high-energy and redundant conformers—Minimize the conformers after eliminating high-energy and redundant conformers. This option has a Steps text box, in which you can enter the number of minimization steps. This is the most expensive option.
- Eliminate high-energy and redundant conformers—Eliminate redundant conformers, evaluate the MacroModel energy and eliminate any conformers that exceed the energy threshold given in the Macromodel options section.
- Eliminate redundant conformers—Perform only the redundant conformer elimination step. This is the least expensive option, as it does not require any energy evaluation, and is only available when preprocessing is turned off.

The minimization and the energy calculation are done with MacroModel using the selected force field.

If you do not minimize the energy, the generation of conformers runs much faster. Many of the conformers are rejected because their energy is too high, so the number of conformers is usually smaller than if you do the energy minimization.

3.3.2.2 Mixed MCMM/LMOD Search Method

The alternative search method is a combined Monte-Carlo Multiple Minimum/Low Mode (MCMM/LMOD) search, and is more accurate than the ligand torsional sampling method, but as a consequence takes longer. The difference between Rapid and Thorough sampling is in the number of steps taken per rotatable bond, which is much larger for thorough sampling. There is no need for the amide bond sampling options with this method. Minimization of the conformers generated by a mixed MCMM/LMOD search is recommended. You can specify the number of steps taken in the minimization of each conformer. The minimization is applied to the input structure, which for MCMM/LMOD is treated as the first conformer in the set. For more information on this method, see Chapter 8 of the *MacroModel User Manual*.

3.3.3 MacroModel Options

In this section, you can select the force field and solvation treatment, and set thresholds to limit the number of conformations generated and determine when two conformers are considered to be identical.

Force Field

The default force field is OPLS_2005, but you can also select MMFFs. For details on these force fields, see Section 2.1 of the *MacroModel User Manual*.

Solvation treatment

Two continuum solvation treatments for water are provided.

- Distance-dependent dielectric
- GB/SA water

The distance-dependent dielectric model is somewhat faster than the GB/SA model, and usually produces similar results.

Maximum relative energy difference

This value sets an energy threshold relative to the lowest-energy conformer. Conformers that are higher in energy than this threshold are discarded. The energy is evaluated with MacroModel using the selected force field.

Eliminate redundant conformers using

Two cutoff criteria are available for eliminating redundant conformers:

- Maximum atom deviation of *N* Å—All distances between pairs of corresponding heavy atoms must be below this cutoff for two conformers to be considered identical.
- RMSD of *N* Å—The root-mean-square deviation of all pairs of corresponding heavy atoms must be below this cutoff for two conformers to be considered identical.

The cutoff is only applied after the energy difference threshold, and only if the two conformers are within 1 kcal/mol of each other. In addition to the cutoff above, a threshold of 60° is used for torsion angle differences for polar hydrogens. This threshold cannot be changed.

3.4 The Ligands Table

The Ligands table lists the ligands that you added, grouped into conformer sets. You can select table rows in the usual way with click, shift-click and control-click. You can sort the columns by clicking the column header, and you can resize the columns by dragging the column boundary. The table columns are described in Table 3.1.

Table 3.1. Description of the Ligand table columns.

Column	Description
In	Check boxes for inclusion of the ligand in the Workspace. This column functions in the same way as the In column of the Project Table. The molecule that is displayed is the first conformer of the set. To view other conformers, you must export them to the Project Table (right-click menu). The included ligands are added as a scratch entry to the Workspace. Inclusion and exclusion of ligands has no effect on the entries in the Project Table.
Name	The name of the ligand. The default name is taken from the Title property of the ligand, if you added it from a project or from a file in which the title is defined. Otherwise a name is created for the ligand. You can edit the name by clicking in the cell, changing the text, then pressing ENTER. The name does not have to be unique.
Activity	Contains the value of the activity you selected when you added the ligands. If you did not select an activity, the table cells are empty. You can edit the activity by clicking in the cell, changing the text, then pressing ENTER.
Pharm Set	Indicates whether a ligand is in the set of active molecules or the set of inactive molecules that will be used to develop the pharmacophore model (the "pharm set"), or is ignored. For these three states the column contains the text active or inactive, or is blank. You can cycle through these states by clicking the table cell. To cycle through the states for all selected rows, control-click any of the selected cells.
Ligand Group	Indicates which group the ligand belongs to. The cell is noneditable. You can change the group by selecting rows and choosing Group Ligands or Ungroup Ligands from the shortcut menu.
Required Match	Check boxes that indicate whether a match must be found for the active when searching for common pharmacophores. Requiring particular actives to match is only relevant if not all actives are required to match.
# Conformations	The number of conformations stored for the ligand. You will normally want to generate multiple conformers for each ligand, unless, for example, you are developing a pharmacophore model from x-ray structures.

If you right-click in the table, a shortcut menu is displayed, from which you can select an action to be performed on the selected ligands. The actions are described in Table 3.2.

Table 3.2. Ligands table shortcut menu items

Item	Description
Export Table Data	Export the data in the table to a CSV file or an HTML file as a table.
Group Ligands	Assign the selected ligands to the same ligand group.
Group Ligands by Title	Assign the selected ligands to ligand groups, using the title (Name) to define the groups.
Ungroup Ligands	Remove the selected ligands from their ligand groups. The selected ligands must be in the same group before this action is applied.
Merge Stereoisomers	Merge the conformers from stereoisomers into a single conformer set for each structure, for the selected ligands. Conformers that were separated by title but don't have recognized stereochemistry are not merged.
Separate Stereoisomers	Separate the conformers of the selected ligands into sets, one set for each stereoisomer.
Merge Conformers, Ignoring Title	Merge the conformers of the selected ligands into a single conformer set for each ligand, ignoring any separation by title or stereochemistry.
Separate Conformers by Title	Separate the conformers of the selected ligands into sets based on their titles.
Add Conformers to Project Table	Add the selected ligands to the Project Table. The structures for each ligand are placed in a separate entry group for the ligand if the ligand has multiple conformers.
Export Conformers to File	Export the selected ligands to a file. Opens a file selector, in which you can navigate to the location and name the file.
Select All	Select all ligands
Invert Selection	Invert the selection of ligands: selected ligands are deselected, and unselected ligands are selected.
Delete	Delete the selected ligands from the table.

When you display a ligand in the Workspace, you can also display information about the ligand in the Workspace, chosen from the properties in the current Ligands table. You can select the information in the Phase Workspace Feedback panel, which you open with Display → Workspace Feedback. If you want to save the selected properties as a preference for other projects or Maestro sessions, click Save Preferences, otherwise the choices apply only to the current project. The feedback can be selected for each step of the workflow independently.

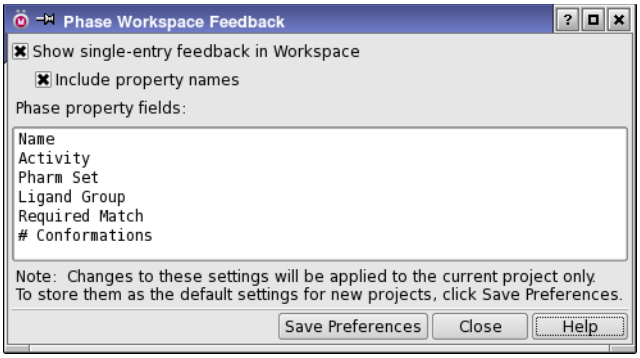

Figure 3.7. The Phase Workspace Feedback panel.

3.5 Defining the Ligand Set for Model Development

There are two ways in which you can define the set of ligands (the "pharm set") that will be used for model development: by setting a threshold, and by manual selection. The ligand set must include some active ligands, and can also include inactive ligands. The ligands marked as active in the Pharm Set column of the Ligands table will be used to develop the model.

To set thresholds for active and inactive ligands, click Activity Thresholds. In the Activity Thresholds dialog box, you can set a threshold for the active ligands and a threshold for the inactive ligands. Ligands with activity greater than or equal to the active threshold are marked as active and included in the pharm set. Ligands with activity less than the inactive threshold are marked as inactive and included in the pharm set. Ligands whose activity lies between the thresholds are excluded from the pharm set.

To add ligands manually to the pharm set, select the ligands (using click, shift-click, or control-click), then control-click the Pharm Set column of the Ligands table. This action changes the status of all selected ligands; a click or a shift-click changes the status of a single ligand.

Note that it is not always necessary to assign every active molecule to the pharm set. If you have groups of highly similar ligands with nearly the same level of activity, you may want to select only one or two ligands from each group. You might also want to reserve some active ligands to test QSAR models.

If you have active ligands that do not have the same connectivity (such as tautomers or ionization states), you can assign them to the same ligand group, so that they will be treated as "the same ligand" for the purpose of finding common pharmacophores. When matching is done to find common pharmacophores, only one of the ligands in the group needs to match. It can also be useful to group inactives or non-actives, for later use in developing QSAR models.

To group ligands, select the ligands that you want to group and choose Group Ligands from the table shortcut menu. The ligands in each group must have the same Pharm Set value, so if you select ligands with multiple Pharm Set values, you will get multiple groups. Each ligand group is assigned a unique number. The group number is listed in the Ligand Group column of the Ligands table, and the rows of the group members are colored with a unique color.

If the ligands that you want in each group have common titles, you can select all the ligands you want to group and choose Group Ligands by Title from the table shortcut menu. The ligands are divided into groups according to their titles and Pharm Set values. This situation is often the case when the structural variations have been generated from a source ligand, for example by LigPrep.

To ungroup ligands, select all the ligands in the group and choose Ungroup Ligands from the table shortcut menu. By default, ligands are not assigned to a ligand group.

If you plan to match less than the full number of actives when searching for common pharmacophores, you can choose actives to which a match is required, by selecting the check box in the Required Match column for that active. Actives within a group must all have the same matching requirement: if one is checked, all of them are checked. This means that a match must be found in the group, not that all members of the group must match. If you want all members of a group to match, you must ungroup them.

3.6 Step Summary

To prepare the ligands for pharmacophore model development, follow the steps below.

1. Import the ligands into the Phase run, by clicking From File, From Run, or From Project.
2. Separate stereoisomers if necessary by selecting the relevant ligands in the table and choosing Separate stereoisomers from the shortcut (right-click) menu.
3. If you want to build a QSAR model or perform activity scoring, enter activity data for the ligands if it is not already present.
4. Clean up the ligand structures and generate variations on stereochemistry or ionization state by clicking Clean Ligands.
5. Generate sets of conformers for each ligand by clicking Generate Conformers.
6. Define the pharm set, either by setting the activity thresholds (click Activity Threshold), or by selecting ligands in the Ligands table.
7. If desired, assign ligands to groups, and specify required matches for partial matching.
8. Click Next to proceed to the next step.

Creating Pharmacophore Sites

The second step in developing a pharmacophore model is to use a set of pharmacophore features to create pharmacophore sites (site points) for all the ligands. This step is performed in the Create Sites step of the Develop Common Pharmacophore Hypotheses panel.

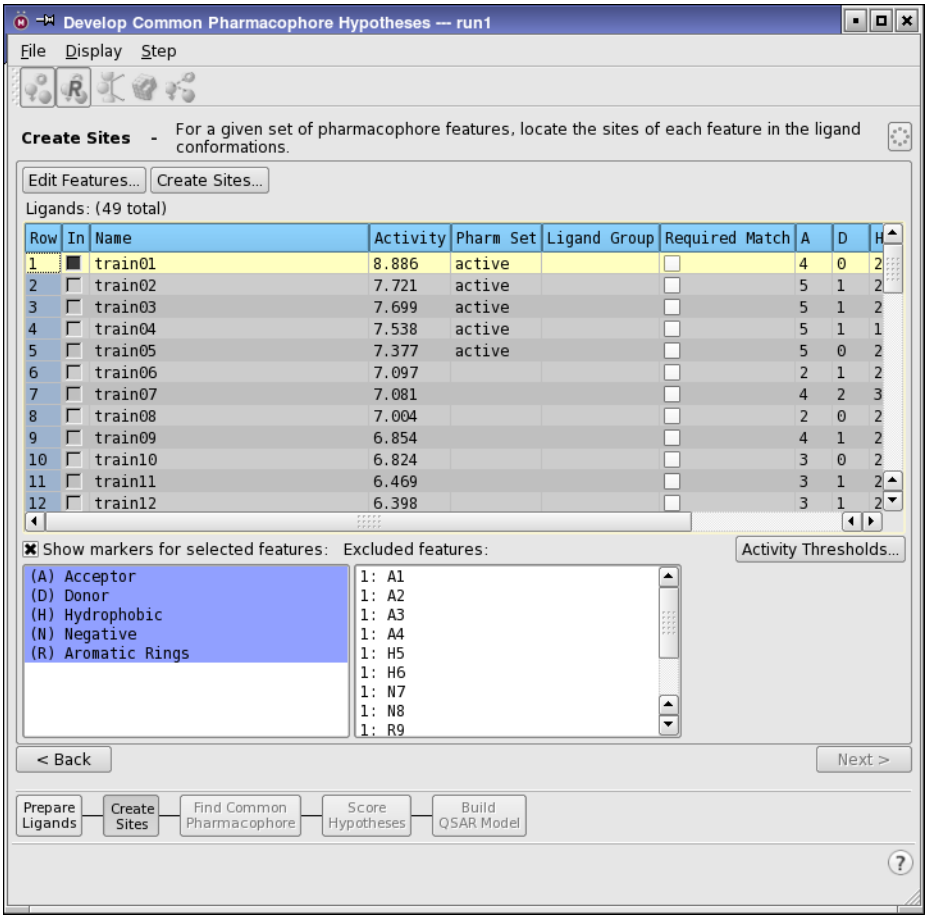

Figure 4.1. The Create Sites step.

Phase supplies a built-in set of six pharmacophore features:

- Hydrogen bond acceptor (A)
- Hydrogen bond donor (D)
- Hydrophobic group (H)
- Negatively charged group (N)
- Positively charged group (P)
- Aromatic ring (R)

Each pharmacophore feature is defined by a set of SMARTS patterns, with the exception of the default hydrophobic features (H) and aromatic rings (R). For these, special algorithms are applied to detect these features automatically and more efficiently than would be possible using SMARTS patterns. All user-defined patterns are specified as SMARTS patterns and assigned one of three possible geometries, which define physical characteristics of the site:

- Point—the site is located on a single atom in the SMARTS pattern.
- Vector—the site is located on a single atom in the SMARTS pattern, and it will be assigned directionality according to one or more vectors originating from the atom.
- Group—the site is located at the centroid of a group of atoms in the SMARTS pattern. For aromatic rings, the site is assigned a direction defined by a vector that is normal to the plane of the ring.

There is an important difference between a *vector feature* and *vector geometry*. "Vector feature" is a more general term that refers to any pharmacophore feature that contains directionality. This includes hydrogen-bond acceptors, hydrogen-bond donors and aromatic rings. "Vector geometry" is more specific, and refers to the particular types of directionality associated with hydrogen-bond acceptors and donors. Thus vector geometry implies vector feature, but vector feature does not necessarily imply vector geometry.

While the built-in feature definitions are adequate for many purposes, you may find it necessary to expand them to include new patterns. For example, the presence of electron-withdrawing groups may cause an otherwise non-acidic hydrogen to be significantly dissociated at pH 7. If the built-in negative ionic definitions do not cover this case, then you may want to supplement the definitions with the appropriate SMARTS pattern.

In some cases, you may feel that a particular built-in definition should not be used, so you can choose to ignore it. Or there may be instances where a built-in definition matches a functionality that you feel does not qualify. In that case you can add a pattern to exclude the functionality in question.

You may also wish to add your own custom features types (X, Y, Z) to account for chemical functionalities that are not covered by the built-in feature types (A, D, H, N, P, R), or to lend

special significance to a particular type of pharmacophoric element. If, for example, you know that all actives must contain a piperidine ring, then you could define a custom feature X with a corresponding SMARTS pattern to match piperidine. Or, perhaps you want to force the pharmacophore model to map C=O acceptors only to other C=O acceptors. This could be achieved by creating a custom acceptor feature Y that contains only the SMARTS pattern for C=O.

The pharmacophore features can be previewed in the Workspace for any ligand. This allows you to verify that the definitions of the features are what you expect, before proceeding to generate site points for the entire set of ligand conformations. It also allows you to exclude particular features from inclusion in the pharmacophore model.

4.1 Viewing Pharmacophore Features

Before you change the definitions of pharmacophore features, or submit the job to create site points using the pharmacophore features, you might want to view the features for each of the ligands. In this way you can check that the features are correctly identified.

Displaying features requires the creation of site points for one conformation of each ligand. This is done automatically when you enter the Create Sites step. If you change the feature definitions, you can create these site points and view the features by clicking Preview. In either case, a job is run locally to create the sites for the first conformer of each ligand. When the job is done (it should be quick), the feature counts are entered in the columns of the Ligands table, and you can display the features in the Workspace.

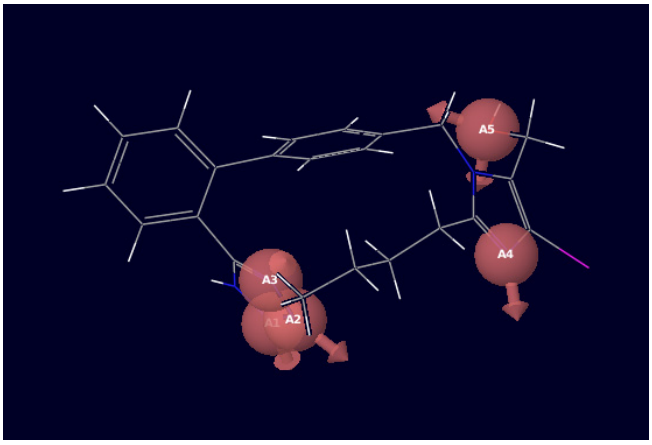

Figure 4.2. Pharmacophore features.

The first seven columns of the Ligands table are the same as in the Prepare Ligands step, and have the same behavior; the selection behavior of the rows is the same, and the right-click

menu is the same—see Section 3.4 on page 26 for a description. In place of the Conformations column is a series of columns, one for each pharmacophore feature. These columns are populated with feature counts (the number of times a feature is present in a ligand).

To display a ligand and its pharmacophore features in the Workspace, click the In column of the Ligands table for the ligand, select Show markers for selected features, and choose the feature types from the list below this option. You can select multiple features from the list. The appearance of the features is described in Table 4.1. To view features for a different ligand, include it in the Workspace using the In column of the Ligands table.

Table 4.1. Visual appearance of pharmacophore features in the Workspace.

Feature	Appearance
Acceptor (A)	Light red sphere centered on the atom with the lone pair, with arrows pointing in the direction of the lone pairs
Donor (D)	Light blue sphere centered on the H atom, with an arrow pointing in the direction of the potential H-bond
Hydrophobic (H)	Green sphere
Negative (N)	Red sphere
Positive (P)	Blue sphere
Aromatic Ring (R)	Orange torus in the plane of the ring
Custom	Colored sphere, with a unique color. Sphere includes arrows if the feature is a vector feature.

4.2 Editing Pharmacophore Features

If you want to supplement the built-in features, create custom features, or load features from another location, you can do so in the Edit Features dialog box. Features are defined in terms of SMARTS patterns. You can add patterns to both standard features and up to three custom features. You can edit and delete custom patterns, and you can exclude or ignore both standard and custom patterns in a feature.

To open the Edit Features dialog box, click Edit Features.

The Edit Features dialog box, displayed in Figure 4.3, contains a section for loading and storing feature sets, described in the next section, and a section for defining features. The feature definition section has controls for selecting, adding and deleting features, a table listing the SMARTS patterns that define the feature, and controls for adding, deleting, and moving patterns.

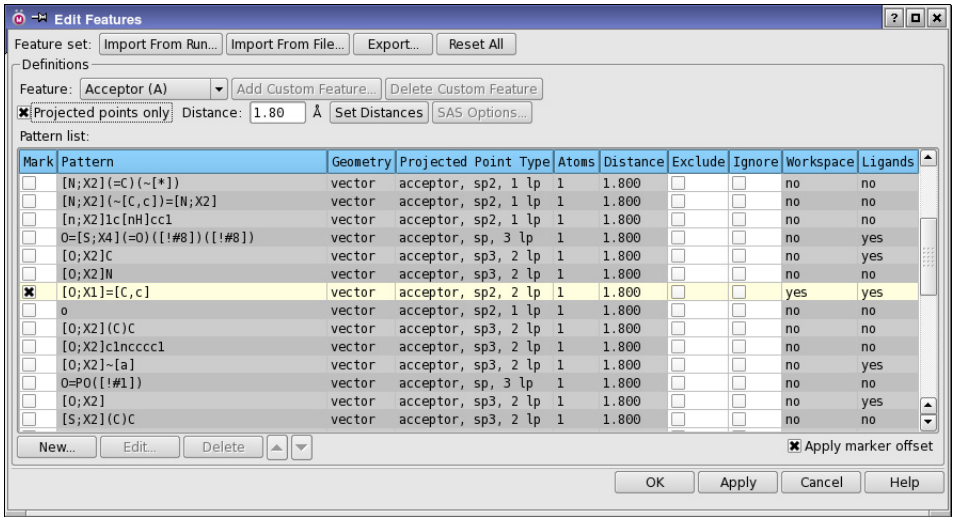

Figure 4.3. The Edit Features dialog box.

The Pattern list table lists all the patterns that are used to define the pharmacophore feature. You can only select one row at a time in the table, and the text fields are not editable, with the exception of the Distance column, which you can edit to set individual distances to projected points. The table columns are described in Table 4.2.

4.2.1 Loading and Storing Feature Sets

The built-in feature sets are stored in the Phase product distribution. To reload them, click Reset All. This button also clears the custom sets.

Feature sets are stored with each run. To import a feature set from another run, click Import from Run, and select the desired run in the dialog box that is displayed.

Feature sets can also be stored in a file. As you do not have access to runs from other projects, you must store feature sets that you want to use in other projects in a file. To save a feature set to a file, click Export, and specify the file location in the file chooser that is displayed. To import a feature set from a file, click Import from File, and navigate to the feature file. You can also use a saved feature file as the default—see Section 2.6 on page 14 for instructions.

Table 4.2. Pattern list table columns

Column	Description
Mark	Column of check boxes. Selecting a check box marks the pattern on any ligands that are displayed in the Workspace.
Pattern	Pattern definition. With the exception of default hydrophobic features and aromatic rings, the definitions are all SMARTS strings.
Geometry	Designates physical characteristics of site. Can be point, vector, or group, as described previously.
Projected Point Type	Defines the directionality of vector features. Can be an aromatic ring, a donor, an acceptor with one or more lone pairs, or none (i.e., a nonvector feature).
Atom Numbers	The list of atoms that determine the location of the pharmacophore site, numbered according to the SMARTS string. Point and vector geometries use a single atom, whereas group geometry uses multiple atoms.
Distance	Distance of the projected point from the ligand atom. This column only applies when Projected points only is selected. To change the distance for a pattern, you can edit the value in this column.
Exclude	Column of check boxes. Selecting a check box excludes the atoms in this definition from being mapped by other definitions. This is essentially a NOT operator. Excluded patterns are processed first when searching for features.
Ignore	Column of check boxes. Selecting a check box ignores the pattern when searching for features. Equivalent to deleting the pattern, but keeps the pattern in the table.
Workspace	Indicates whether the pattern was found in the Workspace structure
Ligands	Indicates whether the pattern was found anywhere in the ligand set (all ligands, not just actives).

4.2.2 Adding and Editing Custom Patterns

If the patterns in a given feature do not cover all the functional groups that you want to include in the feature, you can add extra patterns. To add a new SMARTS pattern to a feature, first choose the feature from the Feature option menu. You then have two options for adding the SMARTS pattern:

- Click the New button below the Pattern list table. The New Pattern dialog box is displayed. In this dialog box, you can enter a SMARTS pattern, define the feature geometry and projected point type and the atoms that represent the feature.
- Right-click on an existing SMARTS pattern that is similar to the one that you want and choose Duplicate Pattern from the shortcut menu. Click Edit to open the Edit Pattern dialog box, in which you can change the SMARTS pattern, the feature geometry and projected point type, and the atoms that represent the feature.

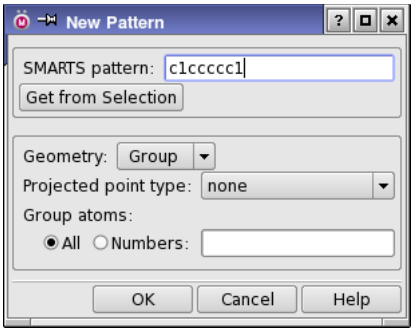

Figure 4.4. The New Pattern dialog box.

The New Pattern and Edit Pattern dialog boxes have the same controls. When you have made your choices, click OK to add or update the pattern. The choices are described in detail below.

To add a pattern to a feature, you must provide the SMARTS string for the desired arrangement of atoms, and define the corresponding pharmacophore site. The pharmacophore site can be a group, such as an aromatic ring; a single point, such as an atom; or a vector, such as a hydrogen bond acceptor or donor.

1. If you are creating a new pattern, type the SMARTS string into the SMARTS pattern text box, or click Get From Selection to use the selected atoms in the Workspace to define a SMARTS string.

   If you use the Workspace selection to define the SMARTS pattern, or you duplicated an existing pattern, you might want to edit it before proceeding.

2. Choose Group, Point, or Vector from the Geometry option menu.

   The remaining controls in the dialog box depend on the choice you make from this menu.

   - Group—The pattern contributes a group of atoms to the pharmacophore feature definition, with the pharmacophore site placed at the centroid. The Projected point type menu has only none and aromatic ring options available, and the Group atoms controls are displayed.
   - Point—The pattern contributes a single atom to the pharmacophore feature definition, with the pharmacophore site placed at that atom. The only available item on the Projected point type menu is none, and the Point atom text box is displayed.
   - Vector—The pattern contributes an atom with one or more directions to the pharmacophore feature definition, with the pharmacophore site placed at the atom. The Projected point type menu has items for donor and acceptor groups, and the Vector atom text box is displayed.

3. Choose the point type from the Projected point type option menu:
   - Group: Choose aromatic ring if the SMARTS pattern defines an aromatic ring, otherwise choose none.
   - Point: none is the only available choice.
   - Vector: Choose donor if the pattern represents a hydrogen bond donor, or choose the acceptor, sp*n*, *m* lp item that defines the type of acceptor (hybridization and number of lone pairs at the acceptor) if the pattern represents a hydrogen bond acceptor. If you want to use projected points, enter the desired distance between the projected point and the ligand atom in the Distance text box.
4. Choose the atoms that define the pharmacophore site:
   - Group: Select All if all atoms in the SMARTS pattern are to be used to define the group centroid, or select Numbers and type the atom numbers for the group centroid in the text box, separated by commas.
   - Point: Type the atom number for the pharmacophore site in the Point atom text box.
   - Vector: Type the atom number for the pharmacophore site in the Vector atom text box. This should be the donor or acceptor atom.

   The atom numbers refer to the order of the atoms in the SMARTS string.

Once you have added a pattern, you can edit it by clicking Edit. The Edit Pattern dialog box is displayed. This dialog box has the same controls as the New Pattern dialog box. If you no longer need the pattern, you can click Delete to delete it. However, you can also ignore it, if you want to keep it in the definition for other applications, but not use it—see the next section. Both of these buttons are only available when you select a custom pattern. Custom patterns are highlighted in blue in the Pattern list table.

4.2.3 Choosing How Patterns Are Used

Matching of patterns to ligand structures is done in the order specified in the Pattern list table. For example, if the first pattern maps a particular nitrogen in the ligand as an acceptor, that same nitrogen will not be mapped as an acceptor by any subsequent pattern. If you have added custom patterns, you can move them up and down the list with the arrow buttons below the table to set their priority. You cannot change the order of the built-in patterns.

If you want to exclude functional groups represented by a pattern from the feature, you can select the check box in the Exclude column for the pattern. For example, you might want to exclude a carboxylic acid group from being considered as a hydrogen bond donor, because it will be ionized under physiological conditions. Excluded functional groups are processed before included groups, so their position in the table does not matter.

If you want a pattern to be ignored, you can select the check box in the Ignore column. Ignored patterns are equivalent to deleted patterns. If you want to save a custom pattern for later use, but not use it in the current feature, select the check box in the Ignore column.

4.2.4 Viewing Patterns

The patterns that define a feature can be viewed individually in the Workspace for each ligand. To display a pattern for a particular ligand, select the ligand in the Ligands table (in the Define Pharmacophore Model panel), then select the check box in the Mark column of the Pattern list table (in the Edit Features dialog box) for the desired pattern. Any occurrences of the pattern are marked in the ligand structure.

You can display markers for more than one pattern, but the markers do not distinguish between patterns. You can display markers for more than one ligand by including the ligands in the Workspace. To see the atoms and bonds as well as the markers, select Apply marker offset.

4.2.5 Adding Custom Features

Phase allows you to define up to three custom features. By default, these features are listed in the Feature option menu as Custom (X), Custom (Y), and Custom(Z), and have the default aromatic vector and surface and aliphatic surface feature definitions included, but ignored. You can add patterns to these features and set their status as described in the sections above.

If you want to delete a custom feature, click Delete Custom Feature. To add a custom feature back, click Add Custom Feature. The Add Custom Feature dialog box is displayed, in which you can specify a name and choose a code letter for the new custom feature. The custom feature is added to the Feature option menu and selected, and the Pattern list table is populated with the default features as described above. Deleting then adding a custom feature is the only way to rename the feature.

4.2.6 Using Projected Points

By default, donors and acceptors are represented by vectors originating at the donor (hydrogen) or acceptor atom. The alignment of these vectors is used to determine whether ligands share the associated feature. Sometimes, two active ligands can form a hydrogen bond to the same receptor site, but from different directions. The projected point is in the same location but the ligand features are not. With the default representation, these two ligands would not contribute to the same pharmacophore hypothesis.

You can replace the vectors with points at a specified distance from the ligand donor or acceptor atom. These points simulate the corresponding acceptor or donor in the receptor, and are called *projected points*. In the default feature set, the projected points are implicit.

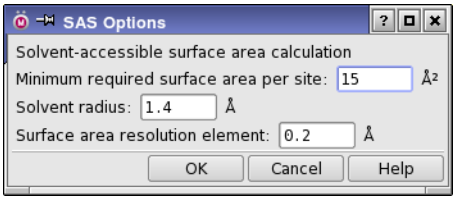

Figure 4.5. The SAS Options dialog box.

To use projected points, select Projected points only. To use the same distance for all patterns, enter a distance in angstroms in the Distance text box, and click Set Distances. To set a distance for an individual pattern, edit the value in the Distance column of the Pattern list table for the pattern.

With this option set, only the patterns that have a vector geometry and a defined projected point type contribute to the feature. All other patterns that are not excluded are ignored. Vector alignments are not used because the vectors have been replaced by points.

4.2.7 Surface Area Calculations for Hydrophobic Features

For a site to be accepted as hydrophobic, the solvent-accessible surface area of the hydrophobic group must exceed a certain threshold. If the area is too small, the site will be a very weak hydrophobe, if it is a hydrophobe at all. You can adjust several parameters that are used to calculate the surface area in the SAS Options dialog box, which you open by clicking the SAS Options button in the Edit Features panel. This button is only available if you have chosen Hydrophobic from the Feature option menu.

To set the minimum area threshold, enter a value in the Minimum required surface area per site text box. The solvent radius used in the surface area calculations is 1.4 Å, which corresponds to water. You can change this value, if for example you want to use a different solvent. The Surface area resolution element text box provides a way of controlling the accuracy of the calculated surface area. This value is used in the partitioning of the atomic spheres for the surface area calculation.

4.3 Excluding Pharmacophore Features

If you know that certain features on one or more of the active ligands cannot contribute to a pharmacophore model, you can mark these features as excluded, and they will not be used in the search for common pharmacophores. Doing this can reduce the number of possible hypotheses that you have to examine.

To exclude features, first place the ligands in the Workspace. You can include more than one ligand in the Workspace, to pick features for multiple ligands at the same time. This might be useful if the ligands are aligned and have the same feature that should be excluded.

When the ligands are placed in the Workspace, the Excluded features list is populated with a complete list of features for each ligand. The feature is represented by the row number in the Ligands table and the feature identifier (type and index).

You can now use one of the following methods to select the features to exclude:

- Select features in the Excluded features list (use shift-click and control-click).
- Pick the features to exclude in the Workspace. To do this you should select Show markers for selected features, and choose one or more feature types to display. Repeated picks select and deselect the feature.

The color of the feature labels in the Workspace is changed to red for the excluded features.

4.4 Defining the Ligand Set for Model Development

If you have not already done so, you can define the active and inactive ligands that are used to develop the pharmacophore model in this step. The controls for doing so are the same as in the Prepare Ligands step. You can click Activity Thresholds to set thresholds for the activity, or you can select ligands for the active set in the Pharm Set column of the Ligands table. See Section 3.5 on page 28 for more information.

4.5 Creating the Sites

Once you are satisfied with the feature set, click Create Sites to start the job that creates and stores the site points for each conformer of each ligand. A Start dialog box opens, in which you can make settings and start the job. The job status icon turns green and rotates until the job finishes, when it turns red and stops rotating.

If the sites already exist for a conformer set (because you copied them from another run, for example), a link is made to this set instead of running the job.

4.6 Step Summary

To create site points for each ligand:

1. Click Create Sites.
2. Click Start in the Start dialog box to run the job.
3. Click Next to proceed to the next step.

Optional tasks:

- Add to the existing features, create custom features, and exclude or ignore patterns by clicking Edit Features.
- Select the use of projected points for acceptors and donors rather than treating them as vector features.
- Define the active and inactive ligands by clicking Activity Thresholds or clicking in the Pharm Set column of the Ligands table.
- Select features to exclude from the search for common pharmacophores.

Finding Common Pharmacophores

In the Find Common Pharmacophores step, pharmacophores from all conformations of the ligands in the active set are examined, and those pharmacophores that contain identical sets of features with very similar spatial arrangements are grouped together. If a given group is found to contain at least one pharmacophore from each ligand, then this group gives rise to a *common pharmacophore*. Any single pharmacophore in the group could ultimately become a common pharmacophore *hypothesis*—an explanation of how ligands bind to the receptor.

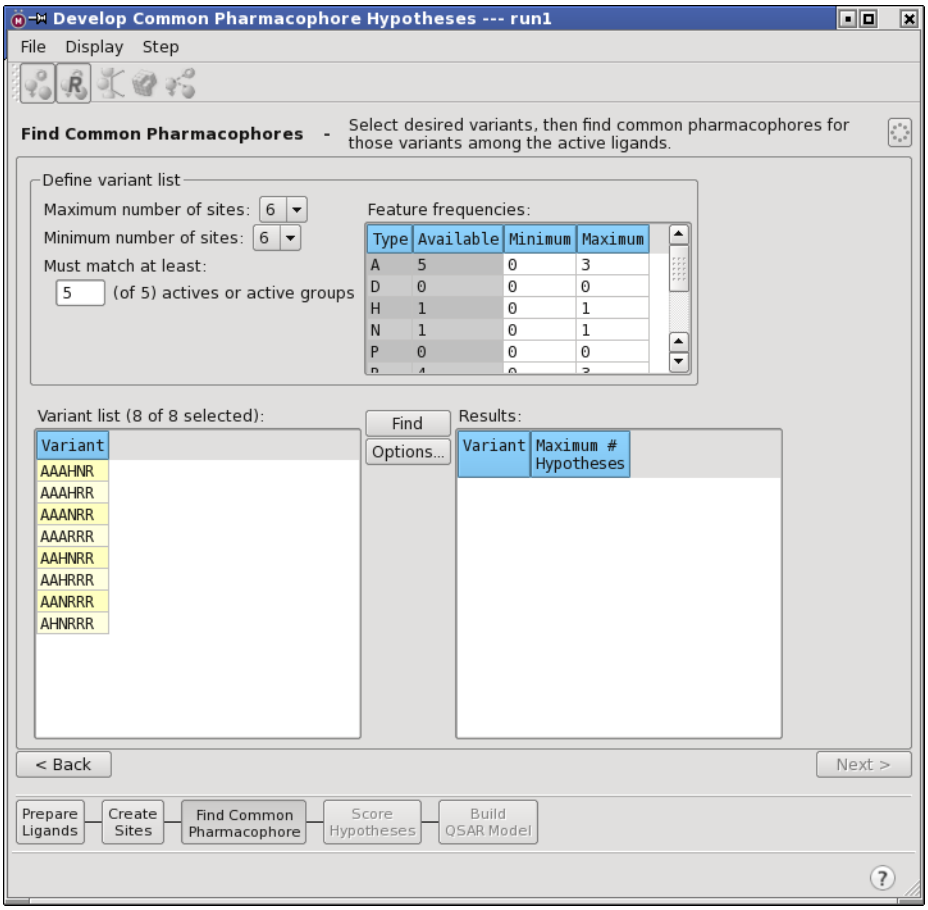

Figure 5.1. The Find Common Pharmacophores step.

Common pharmacophores are identified from a set of *variants*. A variant is a set of feature types that define a possible pharmacophore—for example, the variant ADHH contains a hydrogen-bond acceptor, a hydrogen-bond donor, and two hydrophobic groups.

Phase searches for common pharmacophores with a given number of pharmacophore sites. You can specify from 3 to 7 sites: hypotheses with more sites are not likely because each site represents a 2-3 kcal/mol interaction with the receptor. In addition, you can control how many ligands must match to form a valid hypothesis, and how many of each kind of feature must be included in the match. After the search is complete, the variants for which common pharmacophores were found are passed to the next step.

5.1 The Search Method

Common pharmacophores are identified using a tree-based partitioning technique that groups together similar pharmacophores according to their *intersite distances*, i.e., the distances between pairs of sites in the pharmacophore. Each *k*-point pharmacophore is represented by a vector of *n* distances, where $n = k \cdot (k-1)/2$. Each intersite distance *d* is filtered through a binary decision tree, such as in Figure 5.2.

The tree in Figure 5.2 has a depth of four and partitions distances (in angstroms) on the interval (0, 16] into bins that are 2- Å wide. If each of the *n* distances in a pharmacophore is filtered in this manner, an *n*-dimensional partitioning of the pharmacophore is created. This representation is referred to as an *n*-dimensional box, where the sides of the box are equal to the bin width. Thus a pharmacophore is mapped, according to its intersite distances, into a box of finite size. All pharmacophores that are mapped into the same box are considered to be similar enough to facilitate identification of a common pharmacophore. So if each of the minimum required number of active-set ligands contributes at least one pharmacophore to a particular box, then that box represents a common pharmacophore. Such boxes are said to *survive* the partitioning procedure, while all others are eliminated.

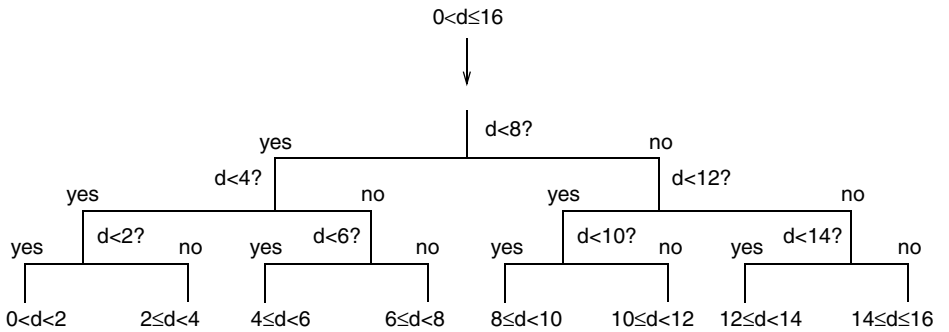

Figure 5.2. Binary decision tree.

5.2 Defining the Scope of the Search

Searching for all possible common pharmacophores could take a long time. From your knowledge of the system of interest, you might not want to search for pharmacophores that have too many or too few site points, or that have too many or too few features of a particular type. Phase provides the means to narrow the search to the variants of interest.

When you enter this step for the first time, a list of all available variants in the set of ligands designated as active is computed, from the number of available sites of each type for each ligand. This list is usually shorter than the theoretical maximum length, because the ligands don't necessarily include all possible variants. The list is filtered with the default settings for the number of sites before it is displayed in the Variant list table. The frequencies of occurrence of the features are used to determine how many occurrences of each feature could be found in a valid hypothesis, given the number of ligands that must be matched. These values are listed in the Available column of the Feature frequencies table, which is described in Table 5.1.

The first task is to decide how many site points to include in the hypothesis. You can choose a maximum number and a minimum number. The default is 5 for both maximum and minimum, but you can choose any number between 3 and 7, inclusive, from the Maximum number of sites and Minimum number of sites option menus. If the number of sites is too large, you might not find any common pharmacophores, but if the number of sites is too small, the common pharmacophores might not contain all required features, and therefore might not discriminate between actives and inactives very well. The search starts with the maximum number of sites, and if it does not find any common pharmacophores, it decreases the number of sites and runs the search again, until it either finds common pharmacophores or passes the minimum number of sites. Thus, the results returned are always for a particular number of site points.

Table 5.1. Description of Feature frequencies table columns.

Column	Description
Type	Lists the features by code letter. Noneditable.
Available	Number of sites of this type that are available, defined as the largest number of occurrences of this feature for which a match can be found for the number of ligands to be matched. For example, with ten ligands, of which three ligands have 4 Acceptors, six have 5 Acceptors, and one has 6 Acceptors, the number of Acceptors available is 4 if all ten ligands are matched, but is 5 if seven ligands are matched. Noneditable; updated if the number of ligands to match changes.
Minimum	Minimum number of features of the given type allowed in any variant. Editable: you can set the value to restrict the possible variants. The default is zero.
Maximum	Maximum number of features of the given type allowed in any variant. Editable: you can set the value to restrict the possible variants. The default is the maximum possible number, given the number available and the number of sites.

If you want to generate and examine hypotheses with different numbers of site points, you can create a new run for each number of points. To create a run that stores the information to date, choose Save As from the File menu, and name the new run. The original run is preserved, and you are now working in the new run. To revert to the original run, choose the run from the Open submenu of the File menu.

By default, all of the active-set ligands or groups ("actives") must contain a given variant for that variant to be listed. However, Phase allows you to relax this criterion so that a common pharmacophore need only match a subset of the chosen actives. This is often a necessity when more than one binding mode is observed among the actives. If you want to widen the search, you can set the number of actives that must contain the variant to a number less than the total, in the Must match at least text box. The number must be between 1 and the maximum, inclusive. The maximum number (the number of chosen actives or active groups) is displayed to the right of the text box. The fewer actives you require a match to, the more variants will be listed. However, if you required a match to any of the actives, the number of variants is restricted by the variants that are present in the required active ligand or group.

Not all the variants are likely to be useful: for example, a variant with five acceptors might be physically unreasonable, and should be excluded from the search. You can limit the number of occurrences of any of the features by entering a minimum and maximum number in the Minimum and Maximum columns of the Feature frequencies table. For example, you might want variants that have between 1 and 3 acceptors. In this case you would enter 1 in the Minimum column of the A row, and 3 in the Maximum column.

After each change, the variant list is automatically updated.

5.3 Modifying the Search Parameters

In the Find Common Pharmacophores - Options dialog box, you can specify the parameters that govern the search for common pharmacophores. Specifying the parameters is a balance between the size of the features in the hypotheses, the size of the ligands, and the time taken and storage requirements for the search. To open this dialog box, click Options in the Find Common Pharmacophores step. The text boxes are described below.

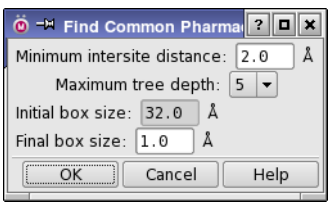

Figure 5.3. The Find Common Pharmacophores - Options dialog box.

Minimum intersite distance

Specifies the minimum distance allowed between two features. If the features in the ligand are closer than this distance, the hypothesis is rejected.

Maximum tree depth

Specifies the number of binary partitioning steps used to sort the pharmacophores into similar groups. This is the maximum recursion level in the partitioning, and the depth of the resulting binary partitioning tree.

Initial box size

Noneditable. Specifies the size of the initial box in the partitioning algorithm, computed from the final box size and the maximum tree depth:

$$\text{Initial box size} = (\text{Final box size}) * 2^{(\text{Maximum tree depth})}.$$

This size should be roughly the size of the binding pocket, or of the smallest ligand. You should therefore choose the final box size and the maximum tree depth to ensure that the initial box size is big enough. The default is 32 Å.

Final box size

Specifies the size of the boxes that contain intersite distances that are considered to be equivalent. This option governs the tolerance on matching: the smaller the box size, the more closely pharmacophores must match. However, the smaller the box size, the longer the search takes. If you choose a smaller final box size, you might have to increase the maximum tree depth so that the initial box size is large enough. If the final box size is too small, the tolerance on matching might be too strict to produce any common pharmacophores.

5.4 Starting the Search

When you have defined the list of variants, you can proceed to the search for common pharmacophores. The search is performed on the variants that are selected in the Variant list table. By default, all variants are selected. You can select variants from the Variant list using the usual combinations of click, control-click, and shift-click. You must have at least one variant selected to run the search.

The common pharmacophores are identified using a binary partitioning algorithm, in which the pharmacophores are split into progressively smaller and more similar groups based on intersite distances. If you want to change the parameters of the search, you can do so by clicking Options and setting the values in the dialog box that is displayed—see Section 5.3 on page 46.

To start the search, click Find. A dialog box is displayed, in which you can select the host, the number of CPUs to use, and the user name on the host. You can distribute this job over multiple processors.

The results of the search can take a large amount of disk space, depending on the number of ligands and their size and flexibility. The search results are kept inside the run, which is stored in the Maestro project. You should make sure that you have adequate disk space: in the temporary storage on the host or hosts on which you run the job, in the Maestro I/O directory, and in the project. Because of the disk space requirements, it is not advisable to run from a scratch project, which is kept by default in `/home/`*username*`/.schrodinger`. Instead, you should save the project to a disk that has plenty of free space.

The results of the run are displayed in the Results table. This table shows the maximum number of hypotheses that could be produced by each variant. You can sort the table by clicking the column headers. Some variants may have no common pharmacophores. These variants are not passed to the next step.

5.5 Step Summary

To find common pharmacophores:

1. Choose the number of sites from the Number of sites option menu.
2. Specify the number of active groups to match in the Must match section.
3. Set limits on the minimum and maximum number of features of each type in the Feature frequencies table.
4. Select variants from the Variant list.
5. (Optional) Set search parameters by clicking Options and entering values in the Find Common Pharmacophores - Options dialog box.
6. Start the search by clicking Find.
7. Click Next to proceed to the next step.

Scoring Hypotheses

In the Score Hypotheses step, common pharmacophores are examined, and a scoring procedure is applied to identify the pharmacophore from each surviving *n*-dimensional box that yields the best alignment of the chosen actives. This pharmacophore provides a hypothesis to explain how the active molecules bind to the receptor. There will of course be many hypotheses, because there are many boxes. The scoring procedure provides a ranking of the different hypotheses, allowing you to make rational choices about which hypotheses are most appropriate for further investigation.

Following the scoring of the hypotheses, the remaining molecules can be used to provide extra information in the hypothesis, based on their structure. To make comparisons, Phase uses *partial matching* to obtain alignments for these ligands. If at least three sites in the hypothesis can be matched, an unambiguous alignment is obtained. For each ligand not designated active, Phase searches for matches involving the largest possible number of sites, and identifies the match that yields the highest fitness score.

If the pharmacophore is an adequate hypothesis, it should discriminate between active and inactive molecules. By aligning and scoring known inactives, you can check the validity of the hypotheses that you generated. If inactives score well, the hypothesis could be invalid because it does not discriminate between actives and inactives, and therefore does not explain how active molecules bind but inactives do not. The hypothesis could also be incomplete because it lacks either a critical site that explains the binding or information on what prevents inactives from binding.

The pharmacophore features that were identified are not the only features that may be useful in defining a good hypothesis. Inactive molecules that have the same pharmacophore features could have functional groups in regions of space not occupied by the active molecules. It is reasonable to suppose that these regions are occupied by the receptor. These regions can then be added to the hypothesis as *excluded volumes*, and used in the database search to screen matches to the hypothesis.

Inactive molecules could also have different functional groups in the same location as functional groups in the active molecules, or be missing functional groups that are in the active molecules. Visual inspection of the aligned ligands can help you understand the structural differences. These differences can also be quantified by building a QSAR model (in the next step), which can be used for screening matches in the database search as well as for identifying functional groups that contribute, positively or negatively, to activity.

6.1 The Scoring Process

A surviving box contains a set of very similar pharmacophores culled from conformations of a minimum number of active-set ligands, and certain of these ligands may contribute more than one pharmacophore to a box. Each pharmacophore and its associated ligand are treated temporarily as a *reference* in order to assign a score. This means the other *non-reference* pharmacophores in the box are aligned, one-by-one, to the reference pharmacophore, using a standard least-squares procedure applied to the corresponding pairs of site points.

The quality of each alignment is measured in three ways: (1) the *alignment score*, which is the root-mean-squared deviation (RMSD) in the site-point positions; (2) the *vector score*, which is the average cosine of the angles formed by corresponding pairs of vector features (acceptors, donors, and aromatic rings) in the aligned structures; and (3) a *volume score* based on the overlap of van der Waals models of the non-hydrogen atoms in each pair of structures,

$$S_{\mathrm{vol}}(i) = V_{\mathrm{common}}(i) / V_{\mathrm{total}}(i) \tag{1}$$

$V_{\mathrm{common}}(i)$ is the common or overlapping volume between ligand i and the reference ligand, while $V_{\mathrm{total}}(i)$ is the total volume occupied by both ligands. The default algorithm used to calculate the volumes is an approximation, in which only pairwise overlaps between the ligand atoms are considered, and intra-ligand overlaps are ignored. Although this algorithm overestimates the volumes, the net effect on the volume score is small.[1]

In principle, a reference pharmacophore could score well, even though it contains one or two very poor individual alignments. For this reason, user-adjustable cutoffs are applied to the RMSD values and vector cosines of each individual alignment. Any reference pharmacophore that violates a cutoff in any individual alignment is eliminated. A *site score* for each alignment is then computed based on the alignment score $S_{\mathrm{align}}(i)$ and the cutoff C_{align} by

$$S_{\mathrm{site}}(i) = 1 - S_{\mathrm{align}}(i) / C_{\mathrm{align}} \ . \tag{2}$$

This score is always between 0 and 1 because alignments with $S_{\mathrm{align}}(i) > C_{\mathrm{align}}$ are eliminated.

The site score, the vector score, and the volume score are combined with separate weights to yield a combined alignment score for each non-reference pharmacophore that has been aligned to the reference. If a non-reference ligand contributes more than one pharmacophore to the box, the pharmacophore yielding the best alignment to the reference is selected. The overall multi-ligand alignment score for a given reference pharmacophore is the average score from the best individual alignments.

1. Setting the environment variable `SCHRODINGER_PHASE_USE_OLD_VOLUME` to any value requests the more accurate volume evaluation that was used prior to Suite 2011.

After all pharmacophores in a box have been treated as a reference, the one yielding the highest multi-ligand alignment score is selected as the hypothesis for that box. The ligand that contributes the reference pharmacophore is referred to as the *reference ligand* for that hypothesis. The non-reference information is carried along with each hypothesis so that additional scoring can be performed using the optimal multi-ligand alignment.

Once hypotheses have been identified across all boxes, the lower scoring hypotheses can be eliminated by applying a percentage cutoff to the overall alignment score. In case the percentage filter yields a very small number of hypotheses, a minimum number of hypotheses can be specified.

After this stage of scoring is completed, the ranking of the hypotheses can be refined using volume and selectivity scoring. The overall volume score for a hypothesis is the average obtained by applying the formula given above to all non-reference ligands *i*. The volume score (S_{vol}) can be added to the overall score with its own user-adjustable weight (W_{vol}).

Selectivity is an empirical estimate of the *rarity* of a hypothesis, i.e., what fraction of molecules are likely to match the hypothesis, regardless of their activity toward the receptor. Selectivity is defined on a logarithmic scale, so a value of 2 means that 1 in 10^2 molecules would be expected to match the hypothesis. Higher selectivity is desirable because it indicates that the hypothesis is more likely to be unique to the active-set ligands. Selectivity is only a rough estimate of the rarity, so you should be careful not to place too much emphasis on it in the overall ranking of hypotheses. As with the other types of scores, the selectivity score (S_{sel}) can be added to the overall score with its own user-adjustable weight (W_{sel}).

If you choose to match less than the total number of chosen actives, you may wish to assign higher scores to hypotheses that match a greater number of the chosen actives. The reward comes in the form of W_{rew}^{m} , where W_{rew} is user-adjustable (1.0 by default) and m is the number of actives that match the hypothesis minus one. If W_{rew} is increased much above 1.0, care must be taken not to make it too large, or it may completely dominate the scoring function. For example, if you have 10 actives and W_{rew} is 1.4, this contribution to the score could have a value of 32. The other terms have a maximum value of 1.0.

Hypotheses for which the reference ligand has a high energy relative to the lowest-energy conformer for that ligand are less likely to be good models of binding, because of the energetic cost. You can include a penalty for high-energy structures by subtracting a multiple of the relative energy from the final score, $W_E \Delta E$.

Likewise, you can penalize hypotheses for which the reference ligand activity is lower than the highest activity, by adding a multiple of the reference ligand activity to the score, $W_{act} A$, where A is the activity.

The final scoring function—the *survival score*—has the following form:

$$S = W_{\text{site}} S_{\text{site}} + W_{\text{vec}} S_{\text{vec}} + W_{\text{vol}} S_{\text{vol}} + W_{\text{sel}} S_{\text{sel}} + W_{\text{rew}}^{m} - W_{\text{E}} \Delta E + W_{\text{act}} A \quad (3)$$

where the W's are the weights and the S's the scores.

If the hypothesis itself is a sufficient explanation of binding and activity, hypotheses for which inactives match well are unlikely to be good hypotheses. If the reason that inactives do not bind is steric hindrance rather than the lack of a particular pharmacophore feature, you can use excluded volumes to filter matches—see Section 6.7 on page 61.

You can penalize hypotheses that match inactives by calculating the survival score for the inactives, and subtracting a multiple of this score from the survival score for the actives. To ensure that inactives that do not match all sites in the hypothesis are penalized, their alignment score is adjusted. If a given inactive matches only k out of n sites in a hypothesis, the effective n-point alignment score is computed from the k-point alignment score as follows:

$$S_{\text{align},n} = \sqrt{W_k S_{\text{align},k}^2 + (1 - W_k) C_{\text{align}}^2} \quad (4)$$

where $W_k = k/n$. This score is then used in Equation (2) to calculate the site score. If an inactive fails to match at least 3 sites in the hypothesis, an unambiguous alignment cannot be obtained, and its contribution to the site score is 0. (This follows from setting k=0 in Equation (4).)

The inactive scoring function is the same as for actives. In addition to computing the inactive score, an adjusted survival score is calculated in which a multiple of the survival score of the inactives is subtracted from the actives survival score:

$$S_{\text{adj}} = S_{\text{actives}} - W_{\text{inactives}} S_{\text{inactives}} \quad (5)$$

6.2 Scoring the Hypotheses

The first task in the Score Hypotheses step is to align the actives to the hypotheses and calculate the score for the actives. To do this, click Score Actives. When you do so, the Score Actives dialog box (Figure 6.1) is displayed, in which you can examine and adjust the weights of the terms in the survival score, the alignment thresholds, and the filters on the number of hypotheses to keep. Details of these parameters are given below. When you have made any changes, click Start, set job parameters in the Start dialog box, and click Start again. When the job finishes, the surviving hypotheses and their scores are displayed in the Hypotheses table.

6.2.1 Scoring Method and Filtering

In this section, you can set the thresholds for filtering out hypotheses with low alignment scores and with poor feature matching, and set a limit on the number of hypotheses to keep.

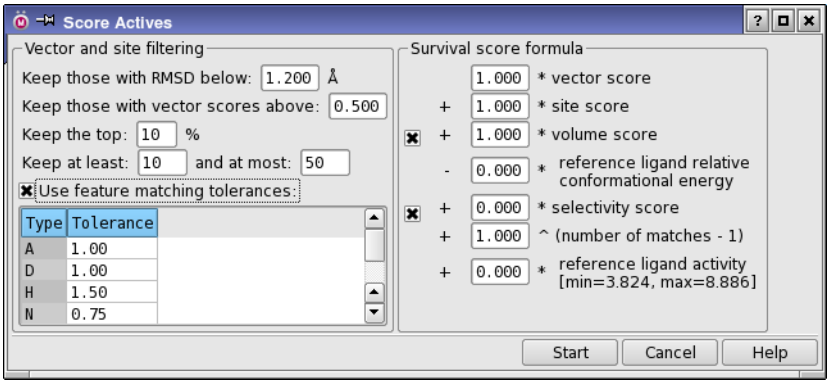

Figure 6.1. The Score Actives dialog box.

6.2.1.1 Alignment Scores

Vector and site alignment scores are computed first, and used to filter the hypotheses. You can set the following parameters, all of which are applied to filter the hypotheses:

Keep those with RMSD below *threshold* Å

Threshold for RMS deviation of the intersite distances of any contributing ligand from those of the reference ligand. The default is 1.2 Å.

Keep those with vector scores above *threshold*

Threshold for the variation in the alignment of vectors between any contributing ligand and the reference ligand. The maximum is 1.0, which corresponds to perfect alignment. The minimum is –1.0, which would keep all hypotheses, regardless of vector alignment.

6.2.1.2 Numerical Cutoffs

To limit the number of hypotheses, you can set the following cutoffs on the fraction or number of hypotheses to keep.

Keep the top *n* %

Limit on the percentage of hypotheses to keep, in order of combined alignment score.

Keep at least *n* and at most *m*

Lower and upper limits for the number of hypotheses to keep. If the percentage of hypotheses kept is lower or higher than these limits, these limits override the percentage limit.

6.2.1.3 Feature-Matching Tolerances

In addition to using the RMSD to filter out hypotheses, you can set matching tolerances on individual features. Features are considered to match if the site points are within the specified tolerance. This feature is useful if the RMSD matching is satisfied, but one or more features do not match well enough.

To apply feature-matching tolerances, select Use feature matching tolerances. The tolerances for each feature type are listed in the table below, and can be edited. All tolerances are applied: if you want to disable matching tolerances for a particular feature type, set the tolerance to a large value.

6.2.2 Survival Score Weighting Factors

The Weighting factors section of the Score Actives dialog box defines the survival score of the hypotheses, which is reported in the Hypotheses table of the Score Hypotheses step along with the individual scores that make up the survival score. The possible ranges for each score and weight are given in Table 6.1.

The lower end of the actual range for the vector score is limited by the cutoff specified in the Vector and site filtering section. Similarly, the maximum relative energy is limited by any cutoff you specified when generating conformers, such as in the Generate Conformers dialog box (see Section 3.3 on page 21).

The selectivity score weight is zero by default because it might eliminate useful hypotheses. Likewise, the energy and activity weights are zero by default.

Table 6.1. Maximum score ranges and allowed weight ranges in the survival score

Score	Score Range	Weight Range	Default Weight
vector score	–1.0 to 1.0	0.0 to 1.0	1.0
site score	0.0 to 1.0	0.0 to 1.0	1.0
volume score	0.0 to 1.0	0.0 to 1.0	1.0
selectivity score	0.0 to ∞	0.0 to 1.0	0.0
number of matches	1.0 to ∞	1.0 to ∞	1.0
reference ligand relative conformational energy	0.0 to ∞	0.0 to ∞	0.0
reference ligand activity	Determined by input	0.0 to ∞	0.0

The weight for the number of matches is raised to the power of the number of matches minus one. A value of 1.0 does not discriminate on the basis of the number of matches. This score can be useful when the required minimum number of actives is smaller than the total number of actives. Adjust this weight with caution: even a value of 2.0 can give large variations in survival scores, and dominate the survival score.

6.3 Scoring Inactives and Rescoring

If the hypothesis is a sufficient explanation of the activity of the ligands, then the inactive ligands can be expected to lack one or more of the features in the hypothesis. However, if it is not a sufficient explanation, the inactives might match all the features in the hypothesis. In that case the hypothesis could for example be missing a feature, or an account of steric clashes. In inactive scoring, survival scores are adjusted to penalize hypotheses that match inactives, assuming that the inactives fail to bind because they do not contain the true pharmacophore. While this condition is rarely satisfied by every inactive in a given set, at least some significant fraction of the inactives must lack the pharmacophore for this technique to be valid.

Once the hypotheses have been scored on the basis of the alignment of the chosen actives, you can calculate an adjusted score based on the alignment of the chosen inactives. The score is adjusted by subtracting a multiple of the survival score of the inactives from the survival score of the actives. To calculate it, click Score Inactives, specify a weight for the inactive score, and click Start, make settings in the Start dialog box and click Start. When the job finishes, the adjusted scores are displayed in the Survival-inactive column of the Hypotheses table.

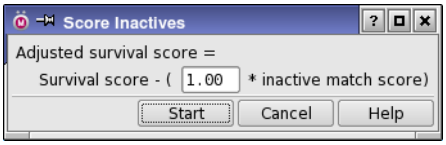

Figure 6.2. The Score Inactives dialog box.

If you want to apply a different scoring function to the surviving hypotheses, you can do so by clicking Rescore, and setting values for the coefficients (weights) of the scoring function in the Rescore Hypotheses dialog box (Figure 6.3). This dialog box contains the same controls as in the Score Actives dialog box. At the same time, you can adjust the weight of the inactives in the Survival-inactive score. The results of the rescoring are listed in the Post-hoc column of the Hypotheses table. These results correspond to the Survival score, and do not include any penalty for matching inactives.

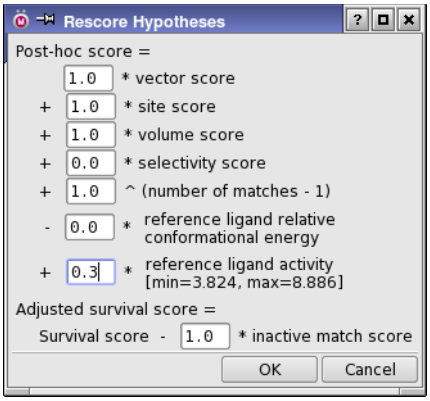

Figure 6.3. The Rescore Hypotheses dialog box.

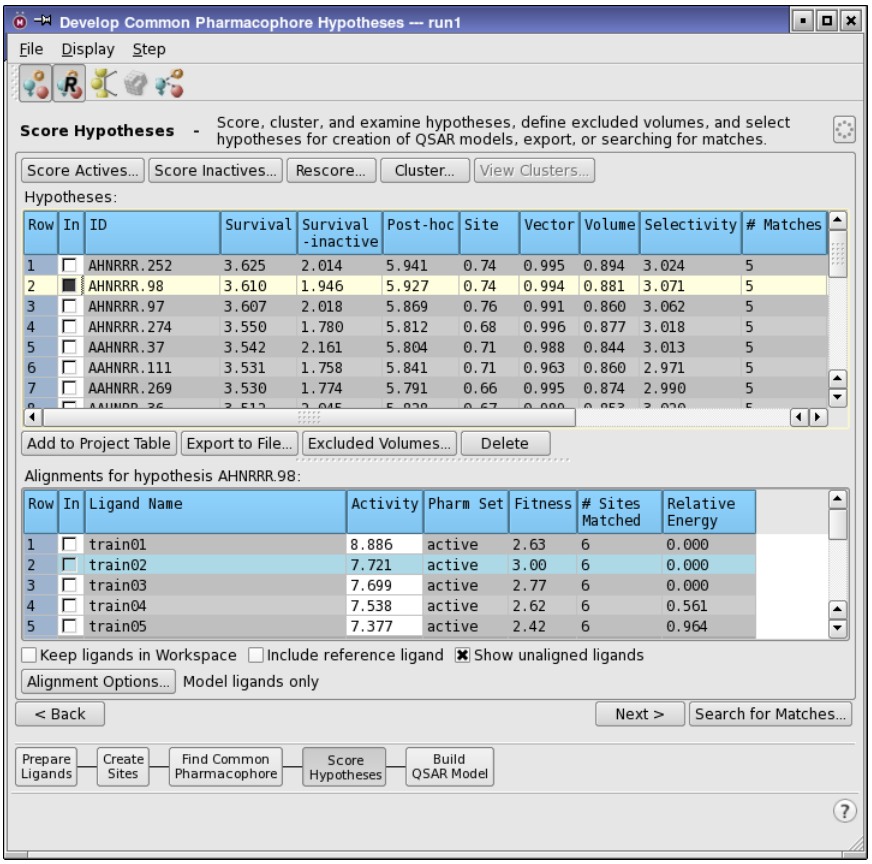

Figure 6.4. The Score Hypotheses step, after scoring.

6.4 Results of Scoring

The Hypotheses table displays the scores for each hypothesis. The In column can be used to display the hypothesis in the Workspace. The hypothesis ID is given in the second column, and consists of the variant name and an index. The remaining columns contain the various scores, whose definitions are given in Table 6.2. Extra columns are added when you cluster the hypotheses. You can sort the table by the values in a column by clicking the column heading. The data in the table is noneditable.

Table 6.2. Description of score columns in the Hypothesis table.

Column	Description
Survival	Weighted combination of the vector, site, volume, and survival scores, and a term for the number of matches. The weights of the volume score and survival score are set to 1.0 and 0.0 by default. The weights can be varied in the Score Actives dialog box. The minimum value of this score is 1.0.
Survival - inactives	Survival score for actives with a multiple of the survival score for inactives subtracted. The weight of the inactive survival score can be set in the Score Inactives dialog box.
Post-hoc	This score is the result of rescoring, and is a weighted combination of the vector, site, volume, and selectivity scores. You can set the weights in the Rescore Hypotheses dialog box, which you open by clicking Rescore.
Site	Site score. This score measures how closely the site points are superimposed in an alignment to the pharmacophore of the structures that contribute to this hypothesis, based on the RMS deviation of the site points of a ligand from those of the reference ligand.
Vector	Vector alignment score. This score measures how well the vectors for acceptors, donors, and aromatic rings are aligned in the structures that contribute to this hypothesis, when the structures themselves are aligned to the pharmacophore.
Volume	Measures how much the volumes of the contributing structures overlap when aligned on the pharmacophore. The volume score is the average of the individual volume scores. The individual volume score is the overlap of the volume of an aligned ligand with that of the reference ligand, divided by the total volume occupied by the two ligands.
Selectivity	Estimate of the rarity of the hypothesis, based on the World Drug Index. The selectivity is the negative logarithm of the fraction of molecules in the Index that match the hypothesis. A selectivity of 2 means that 1 in 100 molecules match. High selectivity means that the hypothesis is more likely to be unique to the actives.
# Matches	Number of actives that match the hypothesis.
Energy	Relative energy of the reference ligand in kcal/mol. This is the energy of the reference conformation relative to the lowest-energy conformation.
Activity	Activity of the reference ligand.
Inactive	Survival score of inactives. The scoring function is the same as for actives.

6.5 Examining Hypotheses and Ligand Alignments

Once you have generated scores for the hypotheses, you can examine the listed hypotheses, one at a time. To examine a hypothesis, select it in the Hypotheses table. The ligands are listed in the Alignments table. The first four columns contain the same information or controls as the Ligands table in the Prepare Ligands step. The columns of this table are described in Table 6.3.

The columns of this table are noneditable. The row for the reference ligand (the ligand that matches the hypothesis exactly) is colored light blue. Rows for aligned ligands are colored light gray; rows for unaligned ligands are colored dark gray. If you want to see only the aligned ligands in the table, deselect Show unaligned ligands. You can sort the table by the values in a column by clicking the column heading.

You can view the aligned ligands and information on their alignments in the Workspace by clicking the check box in the In column of the Alignments table. This column is multiple-select: you can add ligands to the display with shift-click and control-click. You can display and undisplay the hypothesis by clicking the View Hypothesis toolbar button. From the toolbar (or the Display menu) you can also display the distances between the site points and the angles between all sets of three site points.

Table 6.3. Description of Alignments table.

Column	Description
In	Inclusion status of the ligand. The square is filled if the ligand is included (black) and 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.
Activity	The ligand's activity. This value is editable.
Pharm Set	Indicates the status of a ligand in the set used to bind the pharmacophore model.
Fitness	Measures how well the conformer matches the hypothesis. The fitness score is a linear combination of the site and vector alignment scores and the volume score, and is related to the default survival score. The reference ligand, which matches exactly, has a perfect fitness score.
# Sites Matched	Number of sites on the ligand that matched the hypothesis.
Relative Energy	Energy of the best matching conformer relative to the lowest conformer. High relative energies indicate that the conformer is strained. A large proportion of high relative energies could indicate a poor hypothesis.

If you want to keep the same set of ligands in the Workspace when you change hypotheses, select Keep ligands in Workspace. You can include or exclude ligands, and the new list is used

when you change the hypothesis. The exception is that, if any included ligand is not aligned in the new hypothesis, it is not displayed.

If you want to include the reference ligand in the Workspace automatically when you display a hypothesis, select Include reference ligand.

It can also be useful to align and display the "non-model" molecules: the inactive molecules from the pharm set, the actives that do not match all site points, and the molecules that are not in the pharm set. To align these ligands, click Alignment Options, below the Alignments table.

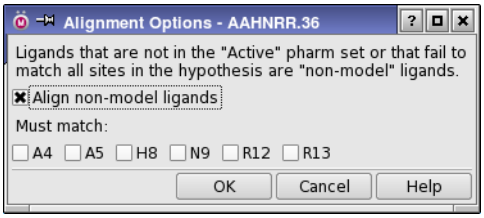

Figure 6.5. The Alignment Options dialog box.

In the Alignment Options dialog box (Figure 6.5) select Align non-model ligands. By default the molecules can match any three sites, but you can enforce matching at specific sites by selecting the sites under Must match. The tolerances for matching these sites are the tolerances specified in the Score Actives dialog box. When you have made your selections, click OK. The dialog box closes and the alignment is performed. The table is updated with information for all molecules that match three or more sites in the hypothesis. The rows for molecules that were not aligned are colored dark gray and are missing the information coming from the alignment.

If, in your examination of a hypothesis, you decide that the hypothesis is not a good one, you can delete it by clicking Delete.

If you have found one or more hypothesis that you want to use outside the Develop Common Pharmacophore Hypotheses workflow—for example, to search a database—you must make them available by selecting them in the Hypotheses table and performing one of the following actions:

- Add them to the Project Table, by clicking Add to Project Table. The selected hypotheses are automatically added to the Project Table when you click Search for Matches.
- Export them to a file by clicking Export to File. A file selector opens, in which you can navigate to the desired location and provide the file name. The extension is removed from the name you provide, so you do not need to add it.

Hypotheses are stored in the run as part of the project, but are not available for use in a search until they are either exported to an external file or added to the Project Table.

You can also add aligned ligands to the Project Table or export them to a file. To add aligned ligands to the Project Table, select the ligands in the Alignments table, then right-click in the table and choose Add Alignments to Project Table from the shortcut menu. The ligands are added to the Project Table as an entry group, with all the properties added by Phase. Likewise, to export aligned ligands to a Maestro file, select the ligands in the Alignments table, then right-click in the table and choose Export Alignments to File from the shortcut menu. A file chooser labeled Export Alignments opens, in which you can navigate to the desired location and enter the file name.

6.6 Clustering Hypotheses

Frequently, there are several hypotheses of a given variant that look very much alike and have very similar scores. In such situations, it is useful to cluster these hypotheses, using a suitable clustering algorithm, and showing only a single representative from each cluster. Details of the clustering method used by Phase are given in Section 12.6.6 on page 147.

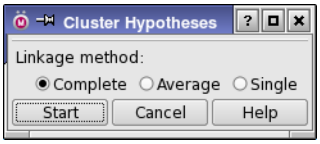

Figure 6.6. The Cluster Hypotheses dialog box.

To start the clustering job, click Cluster. The Cluster Hypotheses dialog box opens, allowing you to set the linkage method and start the job. The linkage method determines what kind of clusters are produced, as follows:

- Complete—The distance between clusters is the largest distance between any pair of objects (one object from each cluster). This option produces compact, spherical clusters.
- Average—The distance between clusters is the average distance between all pairs of objects in the two clusters.
- Single—The distance between clusters is the smallest distance between any pair of objects (one object from each cluster). This option produces diffuse, elongated clusters.

When the job has finished, three columns are added to the Hypotheses table that provide information about the clusters. These columns are described in Table 6.4.

Table 6.4. Description of clustering columns in the Hypothesis table.

Column	Description
Cluster Number	Index of the cluster that the hypothesis belongs to.
Cluster Size	Size of the cluster that the hypothesis belongs to.
Average Similarity	Average similarity of the hypotheses in the cluster.

You can view the cluster representatives by making settings in the View Clusters dialog box, which you open by clicking View Clusters.

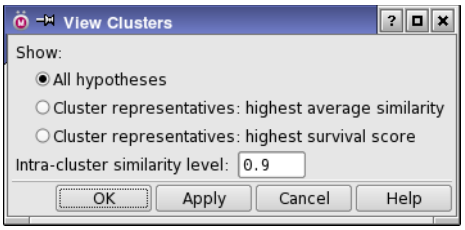

Figure 6.7. The View Clusters dialog box.

By default, all hypotheses are shown, but you can choose from the following options:

- All hypotheses—Show all hypotheses in the cluster.
- Cluster representatives: highest average similarity—Show the hypothesis that has the highest average similarity to the other hypotheses in the cluster. All other hypotheses in the cluster are hidden.
- Cluster representatives: highest survival score—Show the hypothesis that has the highest survival score in the cluster. All other hypotheses in the cluster are hidden.

The size of the clusters, and hence the representatives that are displayed, depends on the similarity threshold for the clusters. To change this threshold, enter a value in the Intra-cluster similarity level text box. The default is 0.9. When you have selected an option and set the similarity level, click Apply or OK to view the representatives.

6.7 Adding Excluded Volumes to Hypotheses

To increase the selectivity of a hypotheses when finding molecules that match it, you can add excluded volumes to the hypothesis. Excluded volumes represent regions that should not be occupied by any active ligand—for example, because the region contains receptor atoms that don't move on ligand binding or because it contains water whose displacement is unfavorable.

When used in a search for matches, ligands that have atoms in these excluded volumes are discarded.

Phase provides a few ways to add excluded volumes to a hypothesis. If the receptor is known, you can use the receptor to define the excluded volumes. If you have both actives and inactives, you can place excluded volumes in regions where inactives have atoms but actives don't. You can also place a full or partial shell of excluded volumes around one or more ligands.

You can add excluded volumes using the Excluded Volumes dialog box, which opens when you click Excluded Volumes. The following sections describe the different methods for adding excluded volumes.

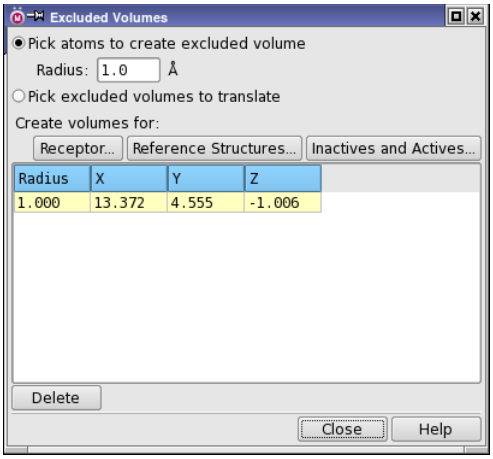

Figure 6.8. The Excluded Volumes dialog box.

Each method adds more volumes to the hypothesis, rather than replacing them. This can result in too many volumes, which will slow down any application of the excluded volumes. You can delete excluded volumes by selecting them in the table or the Workspace and clicking Delete.

The excluded volumes are represented by a set of spheres, whose coordinates and radius can be specified. Once the spheres are added (by any method), they are no longer connected in any way with the atoms or the method that you used to define them. You can subsequently change the radius of any sphere and its location by editing the table cells.

You can also move one or more spheres by selecting Pick excluded volumes to translate, then dragging the volumes to the new location. Dragging excluded volumes only moves them in the viewing plane. To switch to another plane you can rotate the view, for example with the Rotate around X axis by 90 degrees and Rotate around Y axis by 90 degrees toolbar buttons.

To display excluded volumes, select them in the table. You can select multiple excluded volumes in the table or in the Workspace, with shift-click and control-click. The spheres for the selected rows are highlighted in the Workspace. After you have dismissed the Excluded Volumes dialog box, you can view the excluded volumes in the Workspace using the toolbar button or from the Display menu.

6.7.1 Adding Excluded Volumes Manually

If you have included inactive molecules in the Phase run, you can use these molecules to add excluded volumes to locations where the actives do not have atoms, by picking atoms on the inactives in the Workspace.

To define one or more excluded volumes using the inactive molecules as a guide, first make sure that you have aligned the non-model ligands (as described on page 59), then display one or more of the active molecules along with one or more of the inactive molecules. Regions of space in which inactive molecules have atoms but active molecules do not are likely candidates for excluded volumes.

To define a set of volumes, select Pick atoms to create excluded volume, then pick a set of atoms in the Workspace that belong to the inactive molecule or molecules. After each pick, a new excluded volume sphere is displayed on the atom that you pick.

- If you want to place an excluded volume at the centroid of a set of atoms, use shift-click to add atoms to the first. After each pick, the centroid of the set of atoms is calculated and a sphere of the specified radius (default 1 Å) is displayed at the centroid. The atoms that you picked are marked with a small sphere.
- If you want to remove an atom from the set that defines the centroid, control-click the atom.
- If you want to set the default sphere radius before picking, enter the desired value in the Radius text box. You can always change the radius later by editing it in the table.

6.7.2 Adding Excluded Volumes from a Receptor Structure

If the reference ligand of your hypothesis is properly positioned with respect to a receptor structure, you can use the receptor structure to generate excluded volumes. The excluded volumes are placed on the receptor atoms that you select.

To add excluded volumes from a receptor structure, click Receptor under Create volumes for in the Excluded Volumes dialog box. The Excluded Volumes From Receptor dialog box opens.

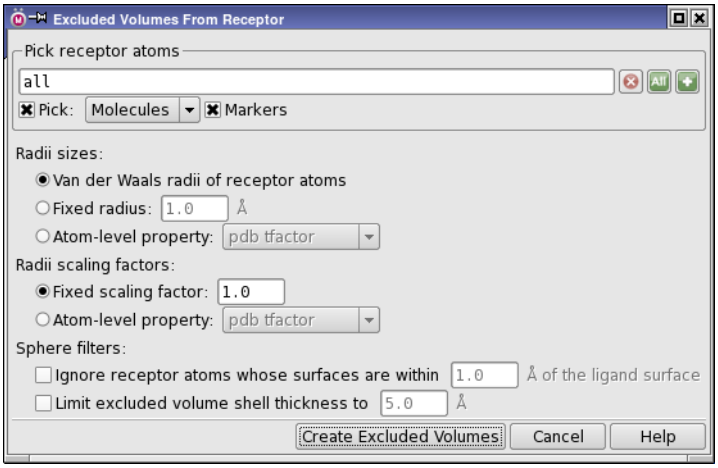

Figure 6.9. The Excluded Volumes From Receptor dialog box.

The steps to create an excluded volume from a receptor or part of a receptor are as follows:

1. Select atoms to define the part of the receptor that you want to create excluded volumes for, using the Pick receptor atoms picking tools. Excluded volume spheres are placed on the atoms that you select. When you select the atoms, you can restrict them to the binding site, or exclude a side chain or loop of the receptor that moves on binding, for example.
2. Set the radii of the excluded volume spheres.

   The options in the Radii sizes and Radii scaling factors sections allow you to set the sphere size at the level of individual atoms, if desired. There are three options for setting sphere size:

   - Van der Waals radii of receptor atoms—Use the van der Waals radii of the receptor atoms for the radii of the spheres.
   - Fixed radius—Set the radii of the excluded volume spheres to the value supplied in the text box, in angstroms.
   - Atom-level Maestro property—Set the radii of the spheres to the value of the atom-level Maestro property chosen from the option menu. Atoms with a zero or unspecified value of this property are skipped.

   In addition, you can specify a scaling factor, which is applied to the spheres. This is useful if, for example, you want to scale the van der Waals radii to soften the receptor surface. There are two options for scaling:

   - Fixed scaling factor—Set the scaling factor for the excluded volume spheres to the value supplied in the text box. The default is 1.0.

- Atom-level Maestro property—Set the scaling factor for the spheres to the value of the atom-level Maestro property chosen from the option menu. Atoms with a zero or unspecified value of this property do not have their excluded volume spheres scaled.

3. Filter out any unwanted excluded volume spheres.

   Spheres that are too close to the reference ligand should not be created, and spheres that are too far away from the reference ligand will not have any influence and will only slow the filtering of matches. The two options for filtering the spheres are:

   - Ignore receptor atoms whose surfaces are within *N* Å of ligand surface—Excluded volume spheres are not created for receptor atoms whose surfaces are within the specified distance of the reference ligand van der Waals surface.
   - Limit excluded volume shell thickness to *N* Å—Spheres located more than the specified distance from the reference ligand are not included.

4. Click Create Excluded Volumes.

A job is run to create the excluded volumes. When the job finishes, the excluded volumes are added to the table in the Excluded Volumes dialog box.

6.7.3 Adding Excluded Volumes Around Reference Structures

In the absence of a receptor, you can simulate the presence of a receptor by creating a shell of excluded volumes around the reference ligand of the hypothesis, or around selected atoms from one or more structures that are aligned to the hypothesis.

Using just the reference ligand, the excluded volume shell is a complement to the molecular volume, so a search with this kind of shell should produce similar results to a shape search. Likewise, placing a shell around several active ligands should produce results similar to a shape search with multiple query structures. Using multiple ligands is less restrictive, because it simulates the receptor flexibility.

If you have a set of active ligands in which parts of the structures overlap very well, but other parts vary in the region of space they occupy, you might conclude that the variable region is solvent-exposed, and then place a shell around only the highly aligned region. You can do this by picking the parts of the ligands that you want to use to create the shell.

To create excluded volumes around reference structures, click Reference Structures under Create volumes for in the Excluded Volumes dialog box. The Excluded Volumes From Reference Structures dialog box opens. Follow the steps below to create the volumes.

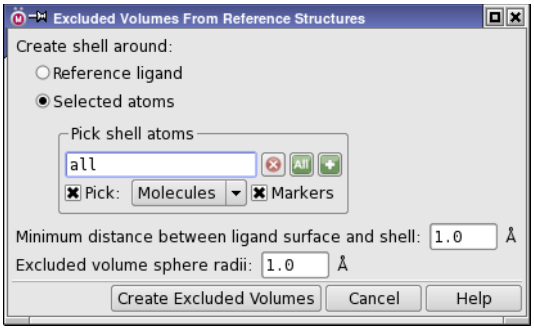

Figure 6.10. The Excluded Volumes From Reference Structures dialog box.

1. Include the structures that you want to use in the Workspace.
2. Select an option for the atoms around which to create the shell of excluded volumes:
   - Reference ligand—Create a full shell of excluded volumes around the reference ligand for the hypothesis. When running from the Develop Pharmacophore Model panel, the hypothesis is the one that is included in the Workspace.
   - Selected atoms—Create a partial shell of excluded volumes, around the atoms that are selected in the Workspace. These atoms should be part of one or more ligands that are aligned to the hypothesis. The atoms can be selected with the Pick shell atoms picking tools.
3. Specify the minimum distance between the van der Waals surface of any of the ligands used and the surface of an excluded volume sphere, in angstroms.

   This buffer distance can be considered to simulate receptor flexibility.
4. Specify the radius that is to be used for the excluded volume spheres in the Excluded volume sphere radii text box.

   Using a larger radius produces less spheres, but results in a less well-defined shape for the excluded region.
5. Click Create Excluded Volumes.

A short job is run to create the excluded volumes. When the job finishes, the excluded volumes are added to the table in the Excluded Volumes dialog box, and the Excluded Volumes From Reference Structures dialog box closes.

6.7.4 Adding Excluded Volumes on Inactive Ligands

Inactives that are inactive because they have atoms in locations that are detrimental to binding can be used to create excluded volumes. This can be done manually (see Section 6.7.1 on page 63) or it can be done automatically, as described here.

The basis of this procedure for generating excluded volumes is that any region of space in which inactives have atoms but actives do not corresponds to a region that is not favored for activity. These are considered to be “clashes”. The reason might be that the region has a receptor side chain that has to be moved out of the way, or that there are solvent molecules in this region that incur a penalty for replacement, or that there are unfavorable interactions with the receptor in this region. There are, of course, other reasons why molecules are inactive (or poor binders): for example, they might be missing a critical pharmacophore feature. When choosing the inactives, therefore, it is a good idea to choose those inactives that match all features in the hypothesis. The choice of actives should be as structurally diverse as possible, to avoid placing excluded volumes in regions that have no effect on activity (such as non-critical solvent regions).

The automatic procedure identifies atoms on the inactives that are considered to “clash”, and places excluded volumes on these atoms, with restrictions on the distance between atoms from the actives and atoms from the inactives.

Before you can run this procedure, you must have a file containing the inactives and a file containing the actives, and the structures must be aligned to the hypothesis. The steps are as follows:

1. In the Score Hypotheses step, click Alignment Options, to open the Alignment Options dialog box.
2. Select Align non-model ligands,
3. Select all the sites listed under Must match.

   This ensures that the inactives that are aligned are those that match all hypothesis sites.
4. Click OK.

The ligands are aligned and the dialog box closes. You can now select the actives and inactives.

5. Deselect Show unaligned ligands, so that only the aligned ligands are listed in the table.
6. Click the Activity column heading in the Alignments table to sort the ligands by activity.
7. Select the actives in the Alignments table.

   You should choose as many actives as you can, from both test and training sets.

8. Right-click and choose Export Alignments to File, to write a file containing the actives.
9. Select the inactives in the Alignments table, and export them to a file.
10. (Optional) Click the Activity column heading twice to restore the original order.

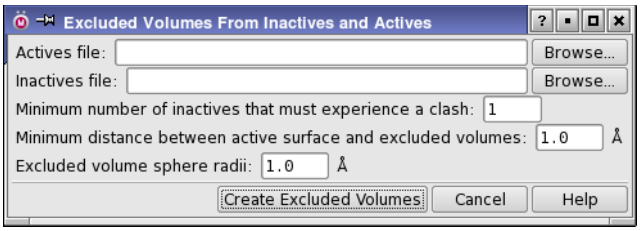

Figure 6.11. The Excluded Volumes From Inactives and Actives dialog box.

Now that you have the files, you can create the excluded volumes.

11. Click Excluded Volumes, to open the Excluded Volumes dialog box.
12. Click Inactives and Actives, to open the Excluded Volumes from Inactives and Actives dialog box.
13. Click the Browse button to select the actives file.
14. Click the Browse button to select the inactives file.
15. Specify the number of inactives that must experience a clash (have atoms in a region not occupied by active atoms).

    A single inactive might have atoms in a region not occupied by actives because that region is in the solvent, rather than in the protein, whereas multiple inactives with atoms in a particular region might indicate an unfavored region that is occupied by the receptor and there is a penalty for having atoms in that region.
16. Set the minimum distance between the van der Waals surface of any active ligand and the surface of an excluded volume sphere, in angstroms.

    This buffer distance can be considered to simulate receptor flexibility.
17. Specify the radius that is to be used for the excluded volume spheres.

    Using a larger radius produces less spheres, but results in a less well-defined shape for the excluded region.
18. Click Create Excluded Volumes.

A short job is run to create the excluded volumes. When the job finishes, the excluded volumes are added to the table in the Excluded Volumes dialog box, and the Excluded Volumes From Inactives and Actives dialog box closes.

6.7.5 Adding Excluded Volumes from the Command Line

You can also add excluded volumes to a hypothesis from the command line. Three utilities are provided that add excluded volumes based on the shape of the reference ligand, on steric clashes from inactive ligands, and on a receptor structure. For more information, see Section 12.8 on page 151.

6.8 Step Summary

To score hypotheses:

1. Click Score Actives.
2. Set scoring options in the Score Actives dialog box.
3. Score the hypotheses by clicking OK.

Optional tasks:

- Score inactives to generate an adjusted scoring function by clicking Score Inactives.
- Rescore the hypotheses with an adjusted scoring function by clicking Rescore.
- Export the selected hypothesis to a file, by clicking Export.
- Cluster the hypotheses, by clicking Cluster, and restrict the hypotheses shown to a representative of each cluster, by clicking View Clusters.
- Add excluded volumes to the selected hypothesis, by clicking Excluded Volumes.
- View hypotheses and alignments in the Workspace, using the toolbar buttons and the Alignments table.

To proceed to building QSAR models:

1. Select the desired hypotheses in the Hypotheses table.
2. Click Next.

To proceed to searching for matches:

1. Select the desired hypotheses in the Hypotheses table.
2. Click Search for Matches.

Building QSAR Models

Phase provides the means to build 3D QSAR models for a set of ligands that are aligned to a selection of hypotheses, and to visualize these models along with the ligand structures and the hypotheses. The QSAR models are developed from a series of ligands that have a range of activities. The usefulness of the QSAR model depends on how well the activity range is spanned, and how diverse the structures are.

In the Build QSAR Model step, you build QSAR models for the hypotheses selected in the Score Hypotheses step, using the activity data for all the available ligands. You can choose atom-based or pharmacophore-based models, select different training sets and test sets, vary the grid spacing, and visualize the model results. When you have built the models, you can use them to visualize parts of the ligands (atoms or pharmacophores) that contribute positively or negatively to activity, and to predict activities of matches to the hypotheses from a database.

When you have completed this step, you can export the hypotheses used to build the model to an external file for use with other projects, and you can continue directly to a search for matches to the hypotheses. Building QSAR models from aligned ligands without a hypothesis is described in Chapter 9.

7.1 Phase QSAR Models

Phase QSAR models are 3D QSAR models, in which chemical features of ligand structures are mapped to a cubic 3D grid. The ligands are first aligned to the set of pharmacophore features in the selected hypotheses using a standard least-squares procedure, as outlined in Section 6.1 on page 50. A rectangular grid is defined to encompass the space occupied by the aligned ligands. This grid divides the occupied space into *N* uniformly-sized cubes, typically 1 Å on each side.

The independent variables in the regression are the binary-valued occupancies (“bits”) of the cubes by structural components; the dependent variables are the activities. Because the number of bits is typically much larger than the number of training set molecules, the system is said to be highly *underdetermined*. For this reason, the regression is performed by a partial least squares (PLS) method, in which a series of models is constructed with an increasing number of PLS factors. The accuracy of the models increases with increasing number of PLS factors until over-fitting starts to occur. The independent variables can also be filtered using a t-value filter to eliminates independent variables whose regression coefficients are overly sensitive to small changes in the training set composition.

Phase offers two choices for the structural components that form the basis of the model: atoms and pharmacophore features.

In the atom-based QSAR models, the structural components of the ligands are represented by van der Waals models of the atoms in the ligands. Each atom is treated as a sphere whose radius is the van der Waals radius for the MacroModel atom type. To distinguish different atom types that occupy the same regions of space, atoms are divided into six classes:

- D—Hydrogen-bond donor (hydrogens bonded to N, O, P, S)
- H—Hydrophobic or nonpolar (C, H–C, Cl, Br, F, I)
- N—Negative ionic (formal negative charge)
- P—Positive ionic (formal positive charge)
- W—Electron-withdrawing (N, O; includes hydrogen-bond acceptors)
- X—Miscellaneous (all other types)

These classes have some correspondence to pharmacophore feature types, but atom classes are assigned using fixed internal rules, not the hypothesis feature definitions. The rules are generally consistent with the default pharmacophore feature definitions, but there are some important differences. For example, the pharmacophore feature definitions use complex rules to identify hydrophobic regions, whereas atom-based QSAR does not. Pharmacophore feature definitions can treat a given atom as part of two different pharmacophore sites, e.g., the nitrogen in a pyridine can be both an acceptor and part of an aromatic ring. Atom-based QSAR requires that each atom be assigned to only one category.

A given atom can occupy the space of one or more cubes in the grid. A cube is occupied by an atom of a particular class if the center of that cube falls within the radius of the atom. Each ligand can therefore be represented by a set of bit values (0 or 1) that indicate which cubes are occupied by atoms of each class. The independent variables used in the QSAR model are the $6N$ occupancies of the cubes and atom classes: each variable corresponds to a given cube and a given atom class, and can take the value 0 or 1.

In the pharmacophore-based QSAR models, the structural components of the ligands are represented by pharmacophore features with a specified radius. Only the pharmacophore features that are present in the hypothesis are used in the QSAR model. As for the atom-based models, the independent variables used in the QSAR model are the mN occupancies of the cubes by the m pharmacophore feature types: each variable corresponds to a given cube and a given feature type, and can take the value 0 or 1.

Once the occupancies are determined, a partial least-squares (PLS) regression analysis is applied to these binary-valued variables to obtain the QSAR model. The variables can be filtered Technical details on the regression analysis and the statistical measures used in the QSAR model are given in Appendix A.

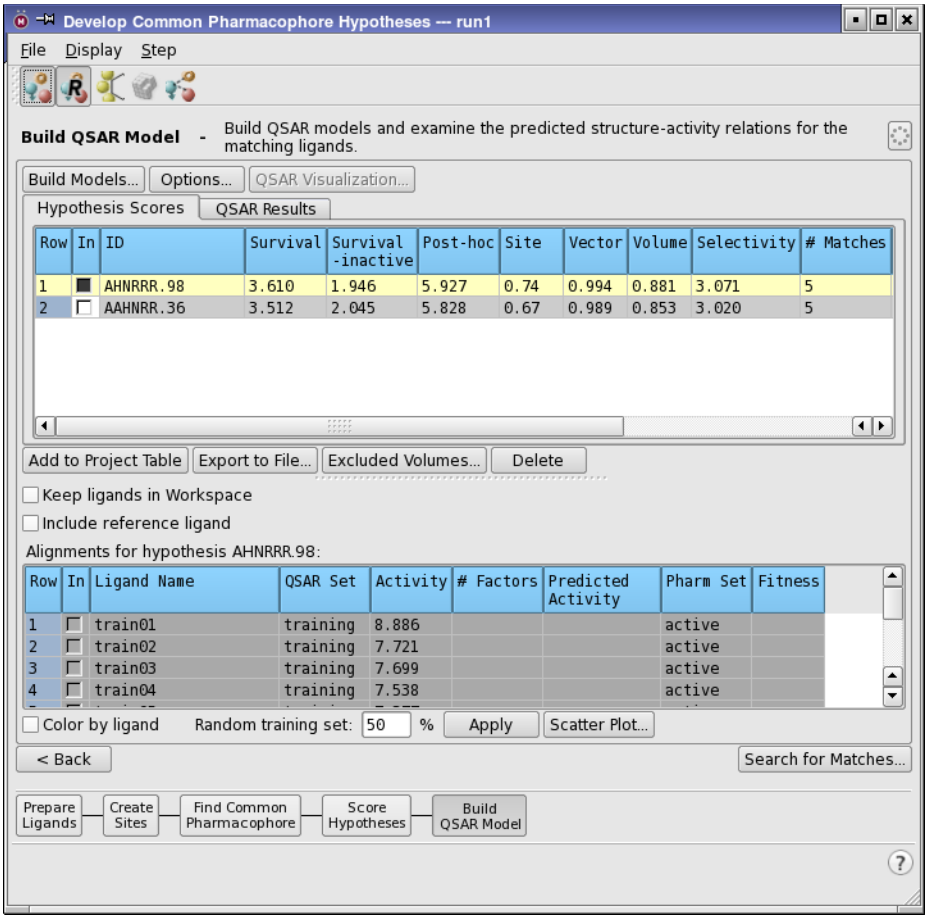

Figure 7.1. The initial view of the Build QSAR Model step.

Atom-based models are useful when features other than the pharmacophores are important to activity, such as steric clashes. However, their performance generally decreases as the diversity of the training set increases. If the structures in the training set contain a relatively small number of rotatable bonds and some common structural framework, then an atom-based model may work quite well.

Pharmacophore-based models assume that the activity is explained entirely by the pharmacophore model itself, and therefore cannot predict activities where other features are important to activity, such as steric clashes. If the structures in the training set are highly flexible or exhibit significant chemical diversity, a pharmacophore-based model may be more appropriate.

If the choice of model is not clear, it is easy to create both types of models and examine the test set statistics to see which approach produces models with the most predictive power.

7.2 Choosing a Training Set and a Test Set

The first task in this step is to choose a training set and a test set, and exclude ligands that you do not want in either set. To display the ligands in the Alignments table, click the In column for any hypothesis in the Hypothesis Scores table. It does not matter which hypothesis you select, because all ligands are listed for all hypotheses. Initially, all ligands are included in the training set, and all rows are colored dark gray, which indicates that there is no corresponding QSAR model. The data columns are empty, and are filled in after the QSAR models are built.

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 remainder are assigned to the test set. Ligands that are in neither set are not used in the selection. The seed for the random selection can be set as an option—see the next section.

7.3 Specifying Options for the QSAR Model

If you want to select the model type, set the grid spacing, choose the maximum number of PLS factors, or specify a seed for random selection of the training set, you can do so in the Build QSAR Model - Options dialog box, which you open by clicking Options.

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 in the Build QSAR Model step. However, if you change the value in the Random seed text box to any positive integer, you can ensure that the same random training set will be created each time you click Apply. The default value of zero ensures that the assignment is always random.

If you want to include the ligands that you designated as actives and inactives for the pharmacophore model development in the training set, and take a random sample of the rest of the ligands, you can select Keep actives and inactives in training set. The ligands whose pharm set property is Active or Inactive are included in the training set, and the remainder are sampled.

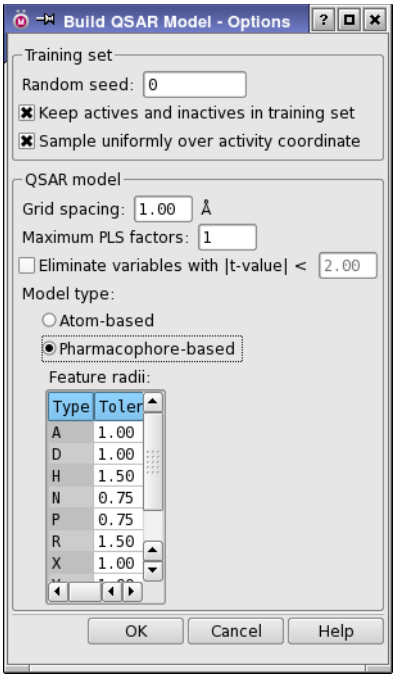

Figure 7.2. The Build QSAR Model - Options dialog box.

You can also ensure that the activities are sampled uniformly by selecting Sample uniformly over activity coordinate. The ligands (or ligand groups) are then sorted into bins by activity, and one ligand (or group) is taken from each bin. The number of bins is the number of training set ligands that are required (after subtracting the number actives and inactives, if you chose to include these in the training set).

The QSAR model partitions the space occupied by the ligands into a cubic grid. Any structural component can occupy part of one or more cubes. A cube is occupied by an atom or a feature if its centroid is within the radius of the atom or feature. You can set the size of the cubes by changing the value in the Grid spacing text box. The allowed range is 0.5 Å to 2.0 Å.

The regression is done by constructing a series of models with an increasing number of PLS factors. The accuracy of the models increases with increasing number of PLS factors until over-fitting starts to occur. You can adjust this value in the Maximum PLS factors text box. 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. This point usually occurs at 2 or 3 factors in Phase.

The independent variables can be filtered using a t-value filter. In this scheme, the standard deviation $\sigma(\beta_i)$ of the regression coefficients β_i is estimated from leave-*n*-out PLS models, and

variables whose coefficients have a t-value $\beta_i/\sigma(\beta_i)$ less than a threshold value are eliminated. To apply the filter, select Eliminate variables whose |value| < *x*, and enter the threshold in the text box. The advantage of the t-value filter is that it eliminates uninformative variables, reduces model complexity and increases model generality. It does, however, decrease the amount of information. One suggested approach for the use of this filter is to identify the most predictive models without the filter, choose the lowest maximum number of factors that is predictive, then apply a statistically significant but conservative filter. The default value of 2.0 is usually adequate.

To select the type of model, choose Atom-based or Pharmacophore-based. In the pharmacophore-based model, the features are represented by spheres whose radii is given in the Tolerance column of the Feature radii table. The features are those defined in the Create Sites step. You can change the feature radii by editing the values in the Tolerance column.

7.4 QSAR Model Results

After you have selected the test set and the training set and set any options, click Build Models. A Start dialog box is displayed, in which you can adjust job settings. When you click Start, the job is run. The predicted activities are displayed in the Alignments table (see Table 7.1), and parameters for the quality of the fit are displayed in the QSAR results table (see Table 7.2). These parameters are defined in Section A.3 on page 179. Each row presents a regression model with a given number of PLS factors.

Table 7.1. Description of the Alignments table columns.

Column	Description
In	Inclusion status of the ligand. The diamond has a cross in it 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 directly editing the table cells.
# Factors	Number of PLS factors used for the QSAR model.

Table 7.1. Description of the Alignments table columns. (Continued)

Column	Description
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.
Pharm Set	Status of a ligand in the set used to build the pharmacophore model.
Fitness	Fitness score from the scoring step.

Table 7.2. Description of the QSAR results table columns.

Column	Description
In	Inclusion status of the hypothesis. The diamond has a cross in it if the hypothesis is included in the Workspace, and is empty if the hypothesis is excluded. You can include and exclude hypotheses with click, shift-click, and control-click.
# 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-squared	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.
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.
Stability	Stability of the model predictions to changes in the training set composition. Maximum value is 1. This statistic can be used to compare models from different hypotheses.
RMSE	Root-mean-square error in the test set predictions.
Q-squared	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.

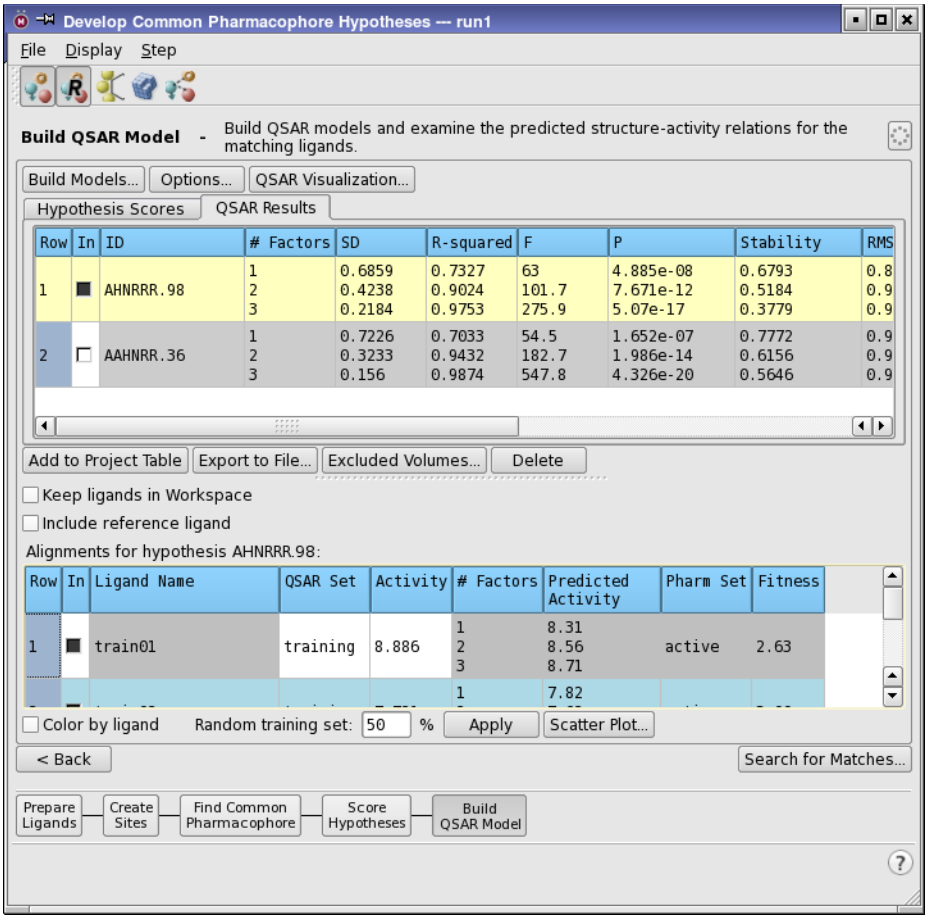

Figure 7.3. The Build QSAR Model step showing results of model-building.

The Alignments table functions the same as for the Score Hypotheses step, and you can select Keep ligands in Workspace to keep the same selection of ligands in the Workspace when you switch hypotheses. You can also select Color by ligand, to color the ligands in the Workspace and the rows in the Alignments table with a unique color. This helps to identify ligands when multiple ligands are displayed.

If you have more than one PLS factor in the model, you should examine the models produced to select the best model. For example, you can examine the predicted activities for the test set, and see at what point they begin to degrade, or you can compare the training set errors with the experimental uncertainty in the data.

It is important to recognize that there is no single statistic that unequivocally determines which model is best. The battery of statistics provided by Phase should be considered in totality, and some measure of common sense must be applied. For example, if the IC_{50} values are accurate to a multiplicative factor of 2, the corresponding $-\log[IC_{50}]$ are only accurate to $\pm\log(2)$. So if the SD statistic is smaller than this experimental uncertainty, then the data are clearly being over-fit, and the model is bound to yield spurious predictions on certain molecules outside the training set, even if the test set predictions appear satisfactory.

One way of assessing the results is to plot the experimental activities against the predicted activities. To do so, click Scatter Plot. The Phase QSAR - Scatter Plot dialog box is displayed, in which you can choose the number of PLS factors for the prediction, choose which ligands to plot results for, from All, Selected, Training set and Test set, and choose whether to draw the 45 degree line of perfect fit. When you click Plot in this dialog box, the scatter plot is displayed in a Scatter Plot panel, and the Manage Plots panel is also displayed.

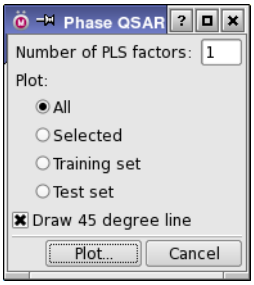

Figure 7.4. The Phase QSAR - Scatter Plot dialog box.

Each time you create a plot, it is added to the Manage Plots panel. You can show or hide plots with the tools in the panel—see Chapter 11 of the *Maestro User Manual* for more information.

The QSAR models are stored with the run, and can be used in the database search to predict activities for the hits. If you change the training set or the model parameters and build new QSAR models, these models overwrite the previous models. To save QSAR models for later use, you can export them with the hypothesis (click Export). The QSAR models are stored with the same file stem as the hypothesis data. If you intend to export more than one QSAR model for a given hypothesis, you must provide a different name for each copy of the hypothesis, or store each in a different location.

You can also add excluded volumes to the hypothesis before exporting it. The Excluded Volumes button opens the same dialog box as in the previous step—see Section 6.7 on page 61.

7.5 Viewing the QSAR Model

Once you have a QSAR model for a hypothesis, you can examine its 3D characteristics by displaying the QSAR model, the ligands, and the hypothesis in the Workspace. Each of these can be displayed independently. To display the hypothesis, the excluded volumes, or the QSAR model, click the appropriate toolbar button or choose the appropriate item from the Display menu. The buttons are described below:

View Hypothesis
Displays the selected hypothesis in the Workspace, as a spatial arrangement of feature symbols. For a description of these symbols, see Table 4.1 on page 34.

View Hypothesis Labels
Displays feature labels for the selected hypothesis in the Workspace.

View Excluded Volumes
Displays excluded volumes for the selected hypothesis in the Workspace.

View QSAR Model
Displays the QSAR model for the selected hypothesis.

View Site Measurements
Opens the View Site Measurements panel, in which you can select the intersite distances and angles of the hypothesis for display in the Workspace.

When you display the QSAR model, the cubes that represent the model are displayed in the Workspace, colored according to the sign of their coefficient values, which by default is blue for positive coefficients and red for negative coefficients. Positive coefficients indicate an increase in activity, negative coefficients a decrease. You can use the visualization of the coefficients to identify characteristics of ligand structures that tend to increase or to decrease activity.

In addition to viewing the model as a whole, you can examine the spatial distribution of contributions to the model by ligand, and by atom class or pharmacophore type either separately or in combination. These capabilities are available in the QSAR Visualization Settings panel, which you open by clicking QSAR Visualization Settings. The visualization tools provided in this panel help you to identify features of ligand structures that are likely to contribute to higher or lower activity.

For example, if you select Workspace ligands under View volume occupied by and choose an atom type from the View effects from list, you can include the ligands in the Workspace one by one (click the In column of the Alignments table) and see which parts of all ligands have a positive or a negative contribution to the activity due to the chosen atom type. This might give a clue to what functional groups are desirable or undesirable at certain positions in a molecule.

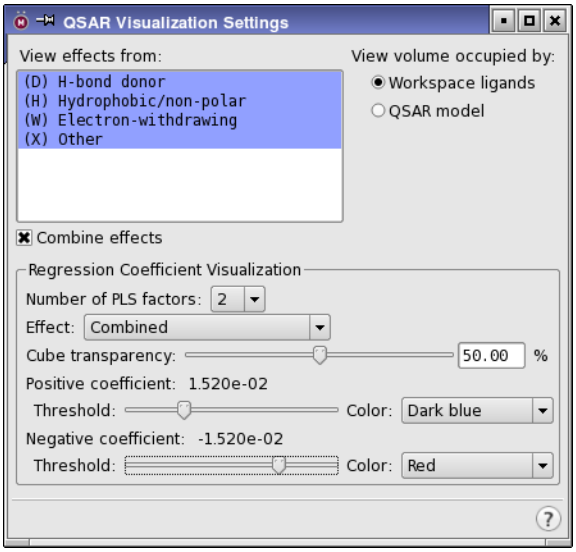

Figure 7.5. The QSAR Visualization Settings panel.

The QSAR Visualization Settings panel also has controls for the display of the model. For example, if you want to filter out cubes that have small coefficients, and therefore do not affect the activity much, you can use the sliders for the positive and negative coefficients in the Regression Coefficient Visualization section. You can also change the cube colors, and view effects from atom or pharmacophore type classes individually or in combination.

The three sections of the QSAR Visualization Settings panel are described below.

View Effects From

To view effects from one or more classes of atoms or pharmacophore types, choose the classes from the list. You can select multiple classes with shift-click and control-click. If you select Combine effects, the effects from all the selected classes are visualized; if you deselect this option, the effects from each class are visualized separately.

View volume occupied by

The two options under this heading allow you to choose whether to view the volume occupied by the QSAR model or by the ligands that are included in the Workspace.

- Workspace ligands— Display the cubes of the QSAR model grid that are occupied by the ligands that are in the Workspace.
- QSAR Model— Display all the cubes that are occupied in the QSAR model.

Regression Coefficient Visualization

This section provides controls for the choice of QSAR model and the display of its coefficients. For the coefficient sliders, cubes that have coefficients that are smaller in magnitude than the threshold are not displayed. This means that the coefficients that have the maximum magnitude are always displayed.

- Number of PLS factors—Select the number of PLS factors from the list to determine which QSAR model is displayed.
- Effect—Choose the effect for which the transparency, thresholds, and colors are to be set. The Combined choice applies when you select the Combine effects option.
- Cube transparency—Adjust the transparency of the cubes from 0% (opaque) to 100% (transparent) for the selected effect.
- Positive coefficient—Adjust the threshold for the display of positive regression coefficients with the Threshold slider, and choose the cube color from the Color option menu, for the selected effect.
- Negative coefficient—Adjust the threshold for the display of negative regression coefficients with the Threshold slider, and choose the cube color from the Color option menu, for the selected effect.

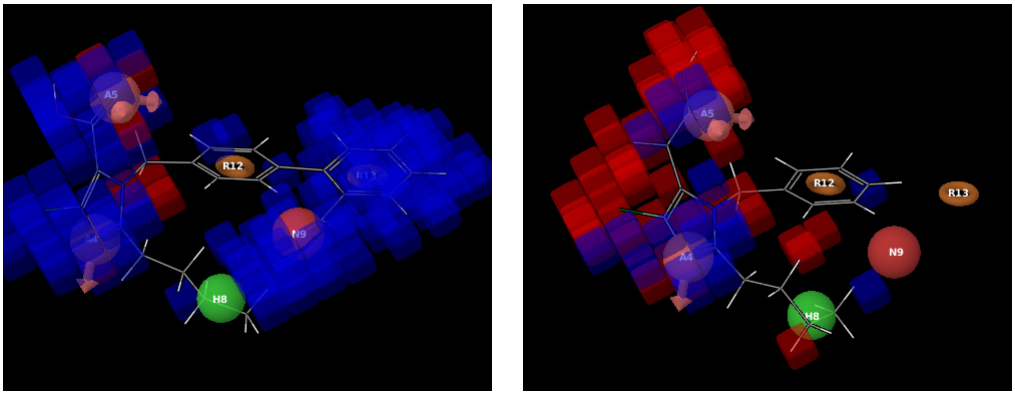

Figure 7.6. QSAR model for an active ligand (left) and an inactive ligand (right).

7.6 Continuing from the Build QSAR Model Step

When you have finished building QSAR models, you can close the Build Pharmacophore Model panel, export the hypothesis with its QSAR model, add aligned ligands and their properties to the Project Table or export them to a file, continue directly to searching for matches, or return to the previous step and select another set of hypotheses with which to build QSAR models.

To export the hypothesis, click Export. The Export Hypothesis dialog box is displayed, in which you can navigate to the desired location and provide the file name. The extension is removed from the file name, so it is not necessary to provide an extension. This means that you should not provide a name like `AAHNRR.26` because the `.26` will be removed. If you want to use a hypothesis name as it appears in the table, you should replace the period with another character, such as an underscore.

To add aligned ligands to the Project Table, select the ligands in the Alignments table, then right-click in the table and choose Add Alignments to Project Table from the shortcut menu. The ligands are added to the Project Table as an entry group, with all the properties added by Phase, including the activity predictions from the QSAR models and the QSAR set membership. The entries are selected and the first entry is included in the Workspace.

Likewise, to export aligned ligands to a Maestro file, select the ligands in the Alignments table, then right-click in the table and choose Export Alignments to File from the shortcut menu. A file selector labeled Export Alignments opens, in which you can navigate to the desired location and enter the file name.

To search for matches to a hypothesis, click Search for Matches. This button opens the Advanced Pharmacophore Screening panel, in which you can start a search for structures that match a hypothesis. All selected hypotheses from the Build Common Pharmacophore Hypotheses panel are loaded by default, and the first of these is selected in the Advanced Pharmacophore Screening panel.

To build a QSAR model for another set of hypotheses, click Back, or click Score Hypotheses in the Guide. You can then select hypotheses, and click Next or Build QSAR Model in the Guide to return to this step. When you do so, you are prompted to create a new run in which to store the hypotheses and the QSAR models. Each set of QSAR models is stored with its set of hypotheses in a separate run, so you can generate a QSAR model for as many hypotheses as you want. You can only store one set of QSAR models for a given hypothesis in the run, but you can always export a hypothesis (click Export Hypothesis) with the current model if you want to store more than one QSAR model for a given hypothesis, or create a new run.

7.7 Step Summary

To build QSAR models:

1. Display the ligands in the Alignments table.
2. Select the training set and the test set.
3. (Optional) Choose a model and set parameters in the Build QSAR Model - Options dialog box.
4. Click Build Models.

To export alignments:

1. Select the ligands in the Alignments for hypothesis table.
2. Right-click in the table and choose the destination from the shortcut menu.
3. If exporting to file, navigate to the location and name the file.

To proceed to searching for matches:

1. Select the desired hypotheses in the Hypotheses table.
2. Click Search for Matches.

Building and Editing Hypotheses

In the Develop Common Pharmacophore Hypotheses workflow, hypotheses are generated from a set of active molecules, automatically taking into account common features and excluding features that are not common. The process does not directly take into account any explicit knowledge about the binding of a particular molecule to the receptor. (Of course, you can display the possible hypotheses and select the ones that fit with your understanding of the binding.)

Phase provides the means to use knowledge of ligand binding directly in the construction of a hypothesis from a single molecule (the reference ligand). For this molecule, Phase generates all possible pharmacophore sites from a set of pharmacophore features. You then select the sites that are included in the hypothesis. The features are the same as those used in the Develop Common Pharmacophore Hypotheses workflow, and can be supplemented with custom features or custom patterns in the same way. You can also edit hypotheses that were exported from the Develop Common Pharmacophore Hypotheses workflow.

You can add excluded volumes to the hypothesis that you construct. QSAR models, however, require a set of ligands with activities, which are not available in this workflow.

8.1 The Manage Hypotheses Panel

The Manage Hypotheses panel provides controls for creating, editing, deleting, importing, and exporting hypotheses based on an individual structure. To open the Manage Hypotheses panel, choose Applications → Phase → Manage Hypotheses.

The panel consists of a toolbar, a table of hypotheses, and a set of action buttons. The toolbar contains three buttons, which are the same as in the Develop Common Pharmacophore Hypotheses panel, and allow you to view the excluded volumes and the intersite distances and angles:

View Hypothesis Labels
Displays feature labels for the selected hypothesis in the Workspace. The font size is the same as for atom labels, and is set in the Font tab of the Preferences panel.

View Excluded Volumes
View excluded volumes for the hypothesis in the Workspace as a set of spheres.

View Site Measurements
Opens the View Site Measurements panel, in which you can select the intersite distances and angles you want to display in the Workspace.

View Matching Tolerances
View feature-matching tolerances as semitransparent spheres whose radius is proportional to the tolerance.

The Hypotheses table lists the hypotheses that are available for editing. These hypotheses are stored as project entries, and can also be viewed in the Project Table. The Hypotheses table displays a filtered view of the Project Table data that includes only the entries from the Project Table that have hypotheses associated with them, and only the properties that are relevant to the hypothesis. (Entries that have hypothesis data are indicated by an H button in the Hyp column of the Project Table.) The structure that is stored in the project entry is the reference ligand for the hypothesis. For hypotheses without a reference ligand, dummy atoms are added to the entry at the site point locations.

You can select a single row of the table, for editing, deleting, adding excluded volumes, or export. The columns of the table are described in Table 8.1. The table is noneditable.

Table 8.1. Description of the Hypotheses table.

Column	Description
In	Inclusion status of the reference ligand and its hypothesis data. The diamond has a cross in it 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.
Title	The title of the project entry for the hypothesis. You can edit this column to change the title.
Hypothesis ID	Identifier of the hypothesis. For new hypotheses, the identifier is constructed automatically from the feature letters and the entry ID of the reference ligand. For hypotheses that were exported from a pharmacophore model development run, the identifier is the identifier from the run.
Entry ID	Project entry ID for the hypothesis
Entry Name	Project entry name for the hypothesis.
phase activity	The activity of the reference ligand, as stored in the Phase run from which the hypothesis originated.
QSAR	Indicates whether a hypothesis has an associated QSAR model.
Excluded Volumes	Indicates whether a hypothesis has associated excluded volumes.
Phase Run Name	The name of the Phase run from which the hypothesis originated.
Hypothesis Date	Date when the hypothesis was last modified.
#Sites	Number of sites in the hypothesis

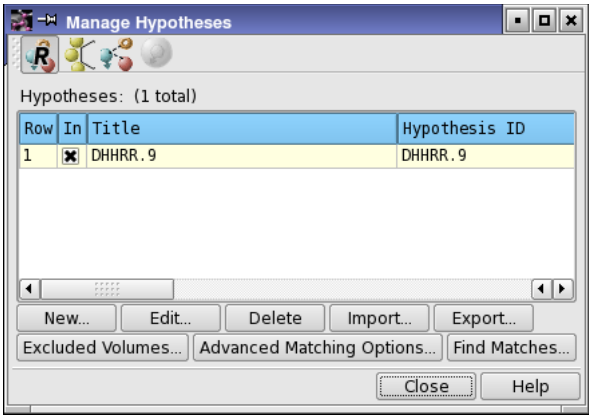

Figure 8.1. The Manage Hypotheses panel.

The Hypotheses table has a shortcut menu, which opens when you right-click in the table. The menu items and their actions are described in Table 8.2.

Table 8.2. Hypothesis table shortcut menu items.

Item	Description
Display	Control the display of the selected hypotheses and their reference ligands. Opens a submenu from which you can choose Hypothesis Only, Atoms Only, or Both.
Color By	Controls the color of the selected hypotheses and reference ligands. Opens a sub-menu from which you can choose Atom and Site Type, to color the atoms by atom type and the hypotheses by the feature type, as described inTable 4.1 on page 34; or Entry, to color the reference ligands and hypotheses with a different uniform color for each entry.
View	Controls the view of excluded volumes and intersite distances and angles. Opens a submenu from which you can choose the objects to display. Same as clicking the toolbar buttons.
Align	Align the selected hypotheses. The stationary ligand in the alignment is the ligand whose hypothesis you right-clicked.
Export	Export the selected hypotheses to external files. Opens a directory chooser in which you can navigate to the desired directory. The files are named with the hypothesis ID as the stem.
Find Matches	Opens the Find Matches to Hypothesis panel, to find matches for the hypothesis you right-clicked.

The action buttons, with the exception of the Delete button, open dialog boxes in which you can make the appropriate choices to perform the action. The buttons are described in Table 8.3.

Table 8.3. Action buttons in the Edit Hypotheses panel

Button	Action
New	Create a new hypothesis. Opens the Choose Reference Ligand dialog box, then the New Hypothesis dialog box.
Edit	Edit the selected hypothesis. Opens the Edit Hypothesis dialog box.
Delete	Delete the selected hypothesis from the project. This action removes the project entry, but does not remove hypotheses that are stored externally or in Phase runs.
Import	Import a hypothesis from disk. Opens the Import Hypothesis dialog box, in which you can navigate to the desired hypothesis. The file you select can be any of the hypothesis-related files.
Export	Export the selected hypothesis to disk. Opens the Export Hypothesis dialog box, in which you can navigate to a location to save the hypothesis.
Excluded Volumes	Add excluded volumes to the selected hypothesis. Opens the Excluded Volumes dialog box—see Section 6.7 on page 61.
Advanced Matching Options	Set up site-specific tolerances and matching criteria. Opens the Advanced Matching Options dialog box—see Section 11.4.6 on page 126.
Find Matches	Opens the Find Matches to Hypothesis panel, to find matches for the selected hypothesis. Active if only one hypothesis is selected in the table.

8.2 Creating New Hypotheses

New hypotheses can be created from an existing structure in two ways:

- You can use a set of pharmacophore features that Maestro uses to identify all the possible pharmacophore sites in the reference ligand. You then choose which sites you want to include in the hypothesis. These hypotheses are termed *ligand-based* hypotheses.
- You can place pharmacophore features at will in the Workspace, in relation to a reference ligand. The reference ligand is only there as a guide, and is deleted once the hypothesis is created. These hypotheses are termed *freestyle* hypotheses.

The hypothesis is created from an entry in the Project Table, which is converted into a hypothesis, with the possible loss of information. To ensure that you preserve the original entry, it is advisable to duplicate the entry before creating a hypothesis. This task is included in the procedures given below.

You can also edit the feature definitions used for either kind of hypothesis in the Edit Features dialog box—see Section 4.2 on page 34 for more information.

8.2.1 Ligand-Based Hypotheses

To create a ligand-based hypothesis, follow the steps below.

1. Do one of the following:

- Choose Applications → Phase → Create Pharmacophore Hypothesis Manually.
- Choose Tasks → Pharmacophore Modeling → Create Hypothesis Manually.
- In the Manage Hypotheses panel, click New.

The Choose Reference Ligand entry chooser opens. The table in the center lists entries from the Project Table. To control which entries are displayed, choose an item from the Choose entry from option menu. To add properties to the table, click Show Property and choose properties in the dialog box that opens. To sort the table, click one of the column headings or click Sort by Project Table order.

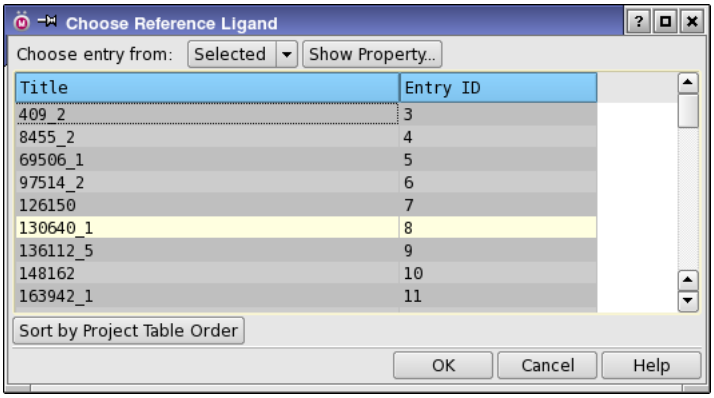

Figure 8.2. The Choose Reference Ligand dialog box.

2. Select the desired entry from the list.

The entry name appears in the Name text box. You should ensure that the structure in the entry is a 3D, all-atom structure. If it is not, the pharmacophore features are likely to be incorrectly assigned.

3. Click Choose.

The Choose Reference Ligand entry chooser closes and the New Hypothesis dialog box opens, with the Ligand-based tab displayed (Figure 8.3). The reference ligand is copied to a new entry in the Project Table for the hypothesis.

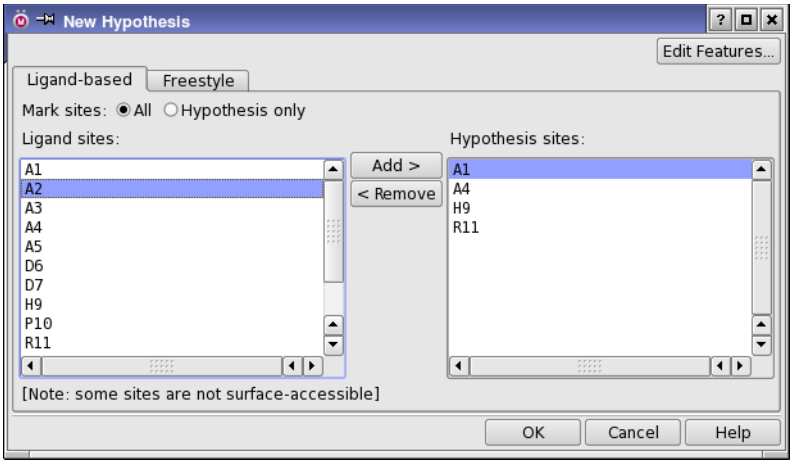

Figure 8.3. The New Hypothesis dialog box, Ligand-based tab.

This tab has two site lists: one of available sites in the reference ligand, and one of sites selected for the hypothesis, which is initially empty. You can choose which sites are displayed in the Workspace by selecting a Mark sites option. By default, all sites are marked. Note that if an atom has multiple sites associated with it, the markers are superimposed. When All is selected, the markers for sites that are not used in the hypothesis are displayed with a dimmer color than for the hypothesis sites.

4. Select the sites you want to include in your hypothesis from the Ligand sites list.

   You can select multiple sites with shift-click and control-click. Once you have selected sites, the Add button becomes available.

5. Click Add.

   The selected sites are added to the Hypothesis sites list. You can also select sites one by one and add them, and you can remove sites from the list. The Ligand sites list does not change when you add or remove sites.

6. When you have selected the desired sites, click OK.

   The New Hypothesis dialog box closes, and the hypothesis is displayed in the Workspace and added to the Hypotheses table of the Manage Hypotheses panel.

8.2.2 Freestyle Hypotheses

To create a new freestyle hypothesis, first follow Step 1 through Step 3 of the procedure above for ligand-based hypotheses. You can select sites in the Ligand-based tab before proceeding to freestyle site addition if you wish.

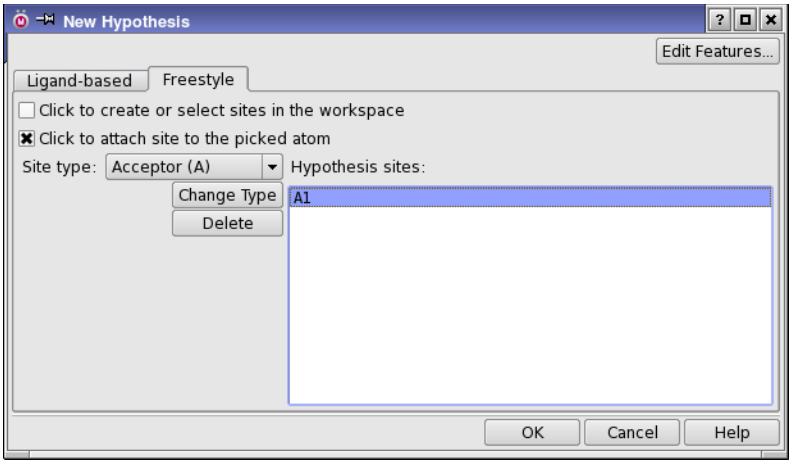

Figure 8.4. The New Hypothesis dialog box, Freestyle tab.

When you are ready to add free sites, click the Freestyle tab. This tab has a list of hypotheses sites, which is initially empty, and controls for choosing and changing the feature type, and placing sites in the Workspace. You can add sites at arbitrary locations, or on atoms.

To add sites in arbitrary locations:

1. Select Click to create or select sites in the Workspace.
2. Select the desired feature type from the Site type option menu.
3. Click in the Workspace where you want to place the feature.

   The site is placed in the *xy* plane (the plane of the screen) at z=0. You will probably need to rotate the ligand and move the site to position it precisely. To do so:

   a. Click on the site in the Workspace.

      The site turns red to indicate that it is selected.

   b. Drag with the middle mouse button to rotate the structure.

      You can also use the toolbar buttons to rotate around the *x* or *y* axis by 90°.

      

   c. Drag with the right mouse button to move the site.
4. Repeat Step 2 and Step 3 for each free site you want to add.

To add sites to atoms:

1. Select Click to attach site to the picked atom.
2. Select the desired feature type from the Site type option menu.
3. Click the atom in the Workspace on which you want to place the feature.
4. Repeat Step 2 and Step 3 for each site you want to add to an atom.

Once you have added sites, you can delete sites or change the site type.

- To delete sites, select them in the Hypothesis sites list and click Delete.
- To change the site type, select the sites in the Hypothesis sites list, choose the new site type from the Site type option menu, and click Change Type.

When you have selected all the sites, click OK. Because the freestyle hypothesis might have no relation to actual features on the initial reference ligand, the ligand will be removed when you save the hypothesis. A dialog box is displayed, giving you the choice to overwrite the existing entry in the project and remove the reference ligand (Overwrite), save the hypothesis as a new entry without a reference ligand (Save), or cancel and return to editing.

8.3 Editing Existing Hypotheses

As well as creating new hypotheses, you can edit hypotheses, including hypotheses that were exported from the Develop Common Pharmacophore Hypotheses panel. You cannot edit hypotheses from a pharmacophore model development run directly: you must export them first, then you can edit the exported version.

To edit an existing hypothesis, select the hypothesis in the Hypotheses table, and click Edit. The Edit Hypothesis dialog box opens. This dialog box is identical to the New Hypothesis dialog box. It has two tabs, Ligand-based and Freestyle. In addition to the two tabs, there is an Edit Features button, which opens the Edit Features dialog box. This dialog box allows you to edit the feature definitions, and is described in detail in Section 4.2 on page 34. If you change the surface area parameters for hydrophobic features to expose more hydrophobic sites (Section 4.2.7 on page 40), a change in the hypothesis must be made before the new sites are shown in the Workspace and the Ligand sites list.

The use of this dialog box to edit ligand-based and freestyle hypotheses is described in the next two subsections. You can edit ligand-based hypotheses in the Freestyle tab, but if you do so and save the changes, the reference ligand is discarded, and the hypothesis becomes a freestyle hypothesis.

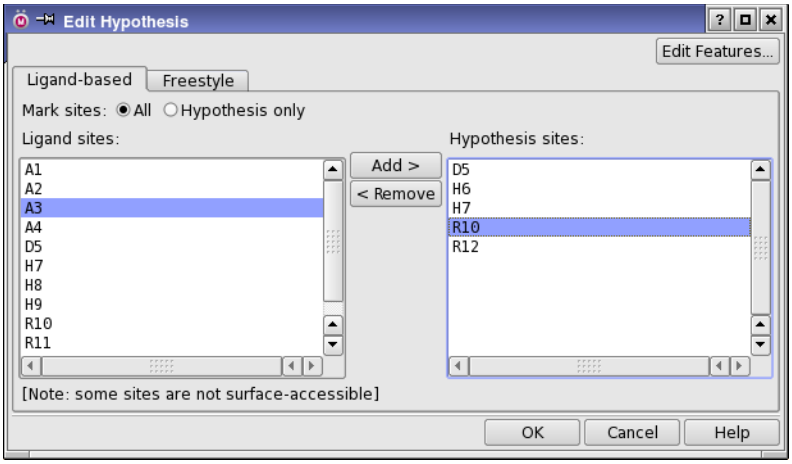

Figure 8.5. The Edit Hypothesis dialog box, Ligand-based tab.

8.3.1 Ligand-Based Hypotheses

In the Ligand-based hypotheses tab, you can add or remove sites from an existing ligand-based hypothesis. This tab is unavailable if you are editing a freestyle hypothesis.

You can choose which sites are displayed in the Workspace using the Mark sites options. By default, the hypothesis sites are marked. If you choose All, the markers for sites that are not used in the hypothesis are displayed with a dimmer color than those that are used in the hypothesis. Note that if an atom has multiple sites associated with it, the markers are superimposed.

- To add sites to the hypothesis, select the sites you want to add in the Ligand sites list, and click Add.

  The selected sites are added to the Hypothesis sites list. You can select multiple sites by shift-clicking and control-clicking. The Add button is only available when you have selected one or more sites in the Ligand sites list.

- To remove sites from the hypotheses, select the sites you want to remove from the Hypothesis sites list, and click Remove.

  The selected sites are removed from the Hypothesis sites list. You can select multiple sites by shift-clicking and control-clicking. The Remove button is only available when you have selected one or more sites in the Hypothesis sites list.

When you have made the desired changes, click OK. The Edit Hypothesis dialog box closes, and the changes to the hypothesis are applied.

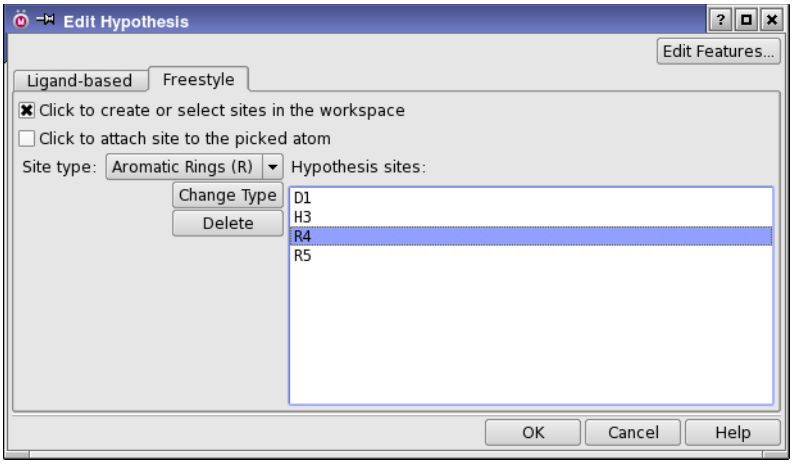

Figure 8.6. The Edit Hypothesis dialog box, Freestyle tab.

8.3.2 Freestyle Hypotheses

In the Freestyle tab, you can change the feature type for any site, add a site, move a site, and delete a site. You can select sites either by clicking on them in the Workspace or selecting them in the Hypotheses sites list. Selected sites are colored red in the Workspace.

To change the feature type for one or more sites:

1. Choose the desired type from the Site type option menu.
2. Select the sites in the Hypothesis sites list or the Workspace.
3. Click Change Type.

To add a site in an arbitrary location:

1. Choose the desired type from the Site type option menu.
2. Select Click to create or select sites in the Workspace.
3. Click in the Workspace where you want to place the feature.

To add a site to an atom:

1. Choose the desired type from the Site type option menu.
2. Select Click to attach site to the picked atom.
3. Click on the atom in the Workspace where you want to place the feature.

To reposition a site:

1. Select Click to create or select sites in the Workspace.
2. Click on the site in the Workspace.
3. Drag the site using the left mouse button.

   The site is moved in the *xy* plane (the plane of the screen).
4. As needed, rotate the Workspace contents with the middle mouse button, then drag the site again.

To delete sites:

1. Select the sites in the Hypothesis sites list or the Workspace.
2. Click Delete.

If the hypothesis had a reference ligand, a dialog box is displayed when you click OK, giving you the choice to overwrite the existing entry in the project and remove the reference ligand (Overwrite), save the hypothesis as a new entry without a reference ligand (Save), or cancel and return to editing.

Building QSAR Models from Ligands

If you have a set of aligned ligands, you can build Phase 3D QSAR models for these ligands without having a hypothesis. There are two kinds of 3D QSAR models available, atom-based and field-based. The field-based models are described in a separate document, *Field-Based QSAR*. This chapter focuses on atom-based models.

Atom-based models are based on atom types, and are the same as those that you build in the Develop Common Pharmacophore Hypotheses workflow, as described in Chapter 7. The methods used in atom-based Phase QSAR models are described in detail in Appendix A. You can use the results to predict activities for other molecules, display a scatter plot of predicted against experimental activities, and add the QSAR model to an existing hypothesis.

The models are built in the Atom-Based QSAR panel. To open this panels, choose Applications → Phase → Atom-Based QSAR or Tasks → QSAR → Atom-Based. The panel has many features in common with the Build QSAR Model step of the Develop Common Pharmacophore Hypotheses panel.

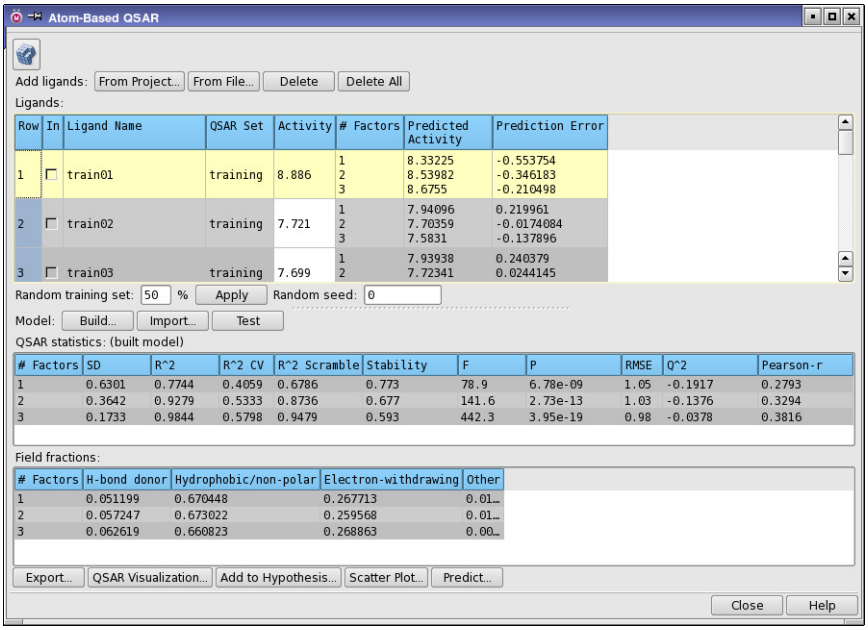

Figure 9.1. The Atom-Based QSAR panel.

9.1 Selecting Ligands

The first step is to select the ligands to use. The ligands you add must be fully prepared 3D structures that are properly aligned. 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*). Alignment can be done as follows:

- If the ligands were exported from the Develop Common Pharmacophore Hypotheses panel, they should already be prealigned to the pharmacophore model.
- If you want to align the ligands to a hypothesis, you can use one of the Pharmacophore Screening panels to run a screening job. The hit file from this job contains the aligned ligands.
- If you don't have a hypothesis, you can 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.

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.
- 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.

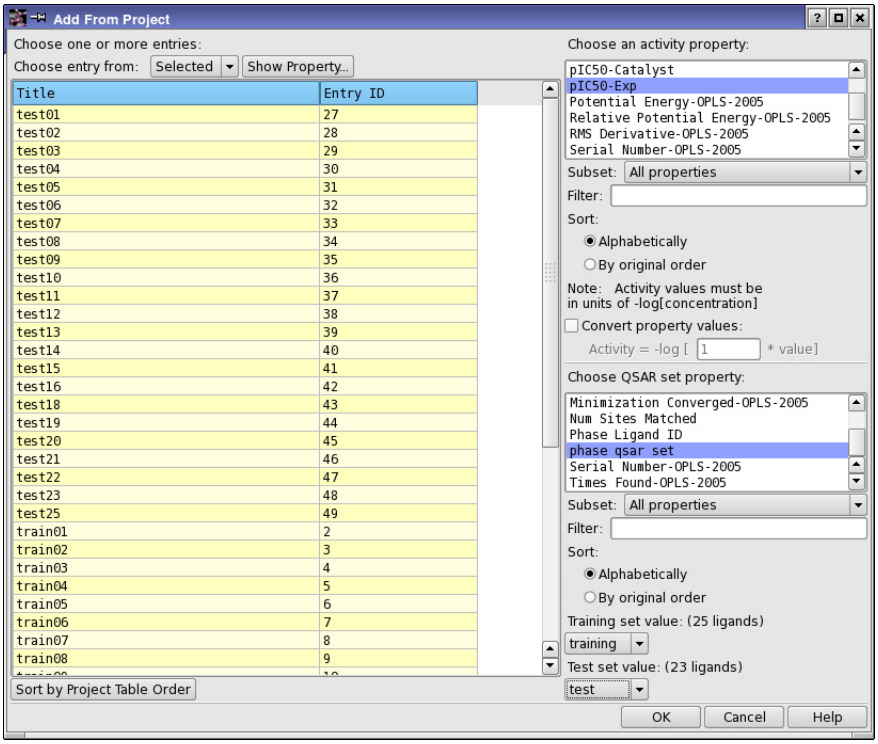

Figure 9.2. The Add From Project dialog box for QSAR models.

The ligands are displayed in the Ligands table when they are added. When the ligands are first read, the # Factors, Predicted Activity, and Prediction Error columns are empty. The values in these columns are added after the QSAR model is built. The table columns are the same as the Alignments table in the Build QSAR Model step of the Develop Common Pharmacophore Hypotheses panel—see Table 7.1 on page 76.

9.2 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 importing 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.

9.3 Building and Testing the Model

Once you have chosen the training and test sets, click Build to build the QSAR models. The Build Atom-Based Model dialog box opens, in which you can make settings for the type of QSAR model you are building, and then build the model.

- Grid spacing—Specify the grid spacing for the cubes that used to determine atom occupancy.
- Maximum PLS factors—Specify the maximum number of PLS factors. 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 rule, you should stop adding factors when the standard deviation of regression is approximately equal to the experimental error.
- 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.
- Number of ligands to leave out—Set the number of ligands to be used in the leave-N-out cross-validation statistics. The default is 1.

When you have made the desired 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 table is filled in.

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.

9.4 Examining the Model

There are several ways in which you can assess the accuracy of the model.

- Examine the QSAR statistics, which are described in Table 9.1. Definitions of the statistics can be found in Section A.3 on page 179. You can add more ligands to the test set, and update the statistics by clicking Test.

Table 9.1. 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 meaningless data, and should be low.
Stability	Stability of the model predictions to changes in the training set composition. Maximum value is 1. This statistic can be used to compare models.
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.

- Create a scatter plot of the experimental data against the predicted data. To do this, click Scatter Plot, which opens the Phase QSAR - Scatter Plot dialog box (see Figure 7.4 on page 79), in which you can select the number of PLS factors and the ligands to include in the plot, then opens the Manage Plots panel and the Scatter Plot panel to display the plot. See page 79 for more information.
- Visualize the QSAR model in the Workspace.

To visualize the QSAR model, click the View QSAR Model button at the top of the panel.

You can control the Workspace display from the QSAR Visualization Settings panel, which is described in Section 7.5 on page 80. To open this panel, click QSAR Visualization.

9.5 Using the Model

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

- 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.
- Add it to an existing hypothesis in the Project Table. To do so, click Add to Hypothesis and select the hypothesis in the entry chooser that is displayed. You can then export the hypothesis from the Project Table for external use.
- 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 the 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.

Creating and Managing a 3D Database

To search for matches to a hypothesis, it is often convenient to store the structures you want to search in a prepared 3D database. When you have prepared the database, the structures will be all-atom structures with reasonable 3D geometries. In addition, you can generate conformers and add site points for a given set of pharmacophore features to the structures. Once you have a database, you can perform management tasks such as adding and deleting structures, exporting structures, and creating subsets. These tasks can be carried out in two panels, the Generate Phase Database panel, and the Manage Phase Database panel.

A wider range of 3D database management tasks can be performed from the command line—see Chapter 13. For large databases, management should be done with the command-line tools, because the panel operations could become very slow.

Phase databases can also be used for Glide docking and with the VSW workflow.

10.1 Creating a Database

The creation of a database and addition of structures to the database can be performed in the Generate Phase Database panel. To open the Generate Phase Database panel, choose one of the following:

- Tasks → Pharmacophore Modeling → Generate Database
- Applications → Phase → Generate Phase Database

10.1.1 Input Structures

When you create a database, you can add structures to the database. You can also add structures to an existing database in this panel. At the top of the panel you specify the source of the structures and set various options for handling these structures.

You can add structures from a single file or from multiple files. The files must be in Maestro or SD format (compressed or uncompressed), or in SMILES format.

- To read from a single file, enter a file name in the Input structure file text box, or click Browse to navigate to the file.
- To specify multiple structure files with related names, you can use the wild card characters * and ? in the file name. These characters have their usual Unix file-matching meanings: ? matches a single character, and * matches zero or more characters.

- To specify multiple structure files with unrelated names, you can create a text file that contains a list of structure file names, and specify this text file in the Input structure file text box, or click Browse and navigate to it.

If you type in the text box, you must press ENTER to ensure that the name is read and the Using property option menu is populated.

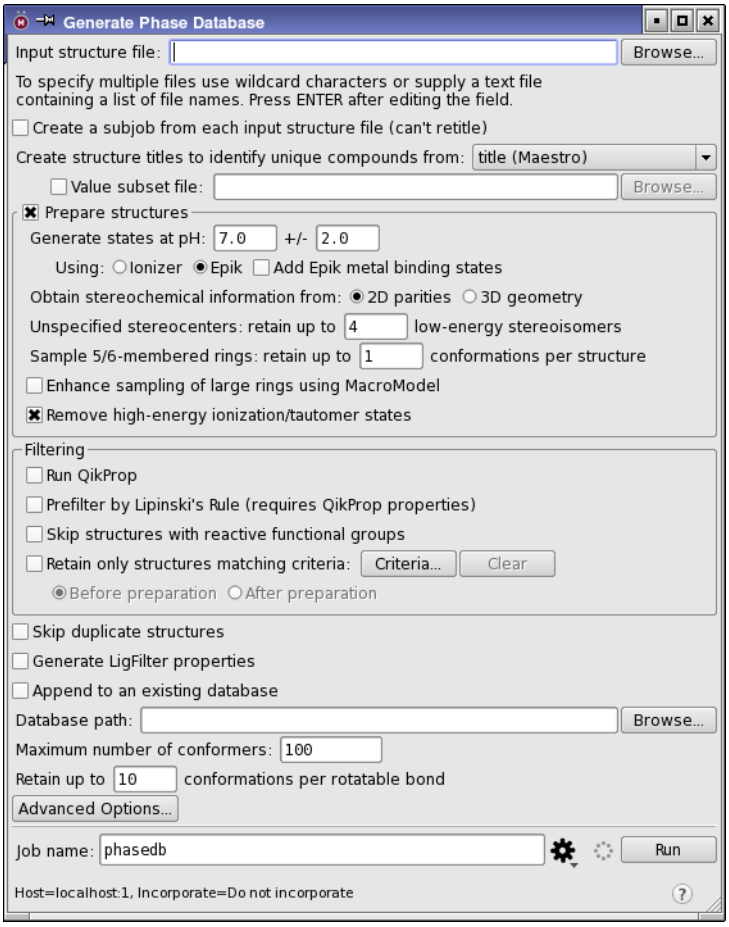

Figure 10.1. The Generate Phase Database panel.

The Create a subjob from each input structure file option allows you to run a separate subjob for each structure file. If you do, you cannot retitle the structures using the controls described below. In addition, no checking is done for duplicate structures.

The Create structure titles to identify unique compounds from option menu allows you to select the title for the structures. The title is used to associate structures that originate from the same compound, which is useful if the structures contain different ionization states or tautomers of the same compound. The property names are taken from the first structure in each file, and only those properties that exist in each file are presented. You should ensure that the property you choose exists for each structure in the file, not just the first. The option menu becomes available when a file with valid properties is specified.

When the title is set, a new property is created to store the original title.

The Value subset file option allows you to filter the input structures so that only those that contain certain values for the selected property are added. The values are specified in a file, one value per line. Enter the file name in the text box, or click Browse and navigate to the file. This option is only available if you use a property to create ligand titles.

Structures that are already in the database can be skipped, by selecting Skip duplicate structures, below the Filtering section. The check is done by generating a unique SMILES string for the structures and comparing the strings. The string is stored in the database for future tests.

10.1.2 Preparing the Structures

A 3D database should be built from all-atom structures with chemically reasonable 3D geometries. The input structures could be represented in 2D form, without explicit hydrogen atoms, or with counter ions and solvent molecules. In addition, the structures might not have chirality information or be in the appropriate ionization state for physiological conditions. If any of these is the case for your input structures, you must select Prepare Structures to obtain structures that are suitable for database searching. If the structures are already all-atom 3D structures, you can deselect this option.

If you need to prepare the structures, you can select from a range of options and settings for ionization, stereochemistry, and ring conformation:

- Generate states at pH controls—Generate ionization and tautomeric states that have significant probability in the given pH range. Enter the target pH and the range in the text boxes, and select the tool for generating ionization states. Epik is licensed separately, so you must have an Epik license to use it.
- Obtain stereochemical information from options—These options allow you to specify the source of stereochemical information. If you have 3D structures, select 3D geometry to determine the stereochemistry from the geometry. Otherwise, select 2D parities, to use information from the parity property in the input file. Any stereochemical center whose chirality is not determined from this information will have its chirality varied.

- Unspecified stereocenters: retain up to N low-energy stereoisomers—Enter the maximum number of stereoisomers to be retained by LigPrep in this text box. LigPrep generates up to 32 stereoisomers, which are then filtered to retain those with the lowest energies. Note that the LigPrep run preserves any existing chirality information in the input file, and selects starting chiralities based on the chemistry of naturally occurring steroids, fused rings and peptides.
- Sample 5/6-membered rings: retain up to N conformations per ligand text box—Enter the maximum number of ring conformations for 5- and 6-membered rings to be generated by LigPrep in this text box.
- Enhance sampling of large rings using MacroModel option—Sample conformations of 7-membered and larger rings with MacroModel after structure preparation. These ring conformations are not sampled by LigPrep.
- Remove high-energy ionization/tautomer states option—Select this option to remove ionization and tautomeric states that have high energies. These are states that are likely to have low populations at the prevailing conditions.

10.1.3 Filtering the Structures

If you want to add only those structures to the database that meet certain criteria, such as their suitability as drug candidates, you can apply a filter to the structures before they are added. Three options for filtering are available in the Filtering section of the panel, and a fourth to generate descriptors for filtering.

- Run QikProp—Select this option to run QikProp after the structure preparation is done. You must run QikProp if you want to prefilter the structures using the Lipinski Rule, or use QikProp properties for custom filtering. If your files already have QikProp properties, you do not need to run QikProp again.
- Prefilter by Lipinski's Rule—Prefilter the structures using Lipinski's Rule of 5 before addition to the database. This rule is described in Table 1.1 of the *QikProp User Manual*, under RuleofFive. Structures that do not satisfy this rule are not added. This option requires QikProp properties. If the input structure files do not have QikProp properties, select Run QikProp.
- Remove ligands with reactive functional groups—Prefilter the structures by removing structures that have reactive groups. The filtering is done by `ligfilter` with a predefined set of reactive groups, which are:

Acyl halides	Phosphinyl halides	Phosphines
Sulfonyl halides	Phosphonyl halides	Alkyl sulfonates
Sulfinyl halides	Alkali metals	Epoxides
Sulfenyl halides	Alkaline-earth metals	Azides
Alkyl halides without fluorine	Lanthanide series metals	Diazonium compounds
Anhydrides	Actinide series metals	Isonitriles
Perhalomethylketones	Transition metals	Halopyrimidines
Aldehydes	Other metals	1,2-Dicarbonyls
Formates	Toxic nonmetals	Michael acceptors
Peroxides	Noble gases	Beta-heterosubstituted carbonyls
R-S-O-R	Carbodiimides	Diazo compounds
Isothiocyanates	Silyl enol ethers	R-N-S-R
Isocyanates	Nitroalkanes	Disulfides

- Retain only ligands matching criteria—Prefilter the structures using `ligfilter`, retaining only those that match the criteria specified. To set up a custom filter or to read a filter file, click Criteria. This button opens the Filtering Options dialog box, which is described on page 13 of the *Virtual Screening Workflow* document. To clear the custom filter, click Clear. Select Before preparation or After preparation to determine the point at which the criteria are applied.

For more information on `ligfilter`, see Section 2.4 of the *General Utilities* manual.

10.1.4 Specifying the Database

A Phase 3D database must be stored on a file system that is accessible to all hosts that will read the database. Access to the database is not needed on the host from which you launch database jobs, only on the hosts that run the jobs.

In addition to host access, you should also consider whether other users will need to search or modify a database that you create. If so, you should choose a file system in which you can safely grant other users read and execute permissions to each directory along the path leading to the database. If you want to allow them to modify the database, they must be given write permissions as well as read and execute permissions.

- To specify the location of the database, enter the path to the database in the Database path text box, or click Browse and navigate to the database, whose extension is `.phdb`.
- If you are adding files to an existing database, select Append to an existing database. The database you specified must exist.

10.1.5 Generating Conformers, Sites, and Properties

Conformers are generated using the ConfGen facility, and site points are automatically added. This is the same procedure as is used when generating conformers during a search for matches. The Generate Phase Database panel offers only limited flexibility in the conformer generation process. You can restrict the number of conformers generated in the Maximum number of conformers text box. You can ensure a minimum coverage of conformational space by entering a value in the Retain up to N conformations per rotatable bond text box.

Further options for adding conformers are available in the Generate Phase DB - Advanced Options dialog box, which you open by clicking Advanced Options. You can choose whether to generate conformations for all structures (All molecules) or only those that do not have them (Auto-detect). The latter option is useful for adding pregenerated conformer sets. You can choose the conformational sampling method (Rapid or Thorough), and the treatment of rotations around amide bonds. You can specify an energy threshold for discarding conformers: those that are higher in energy than this amount above the lowest-energy conformer are discarded.

If you want to add structures without conformers, you can use the command-line tools, which are described in Chapter 13.

You can add properties to the structures in the database, by selecting Generate LigFilter properties. These properties are listed in the General Attributes tab of the Ligand Filtering panel.

10.2 Managing a Database

Database management tasks, such as adding, deleting, or exporting structures and creating and deleting subsets, can be handled in the Manage Phase Database panel. You can also use the command-line tools, which are described in Chapter 13, to manage databases. To open this panel, choose one of the following:

Maestro:

- Tasks → Pharmacophore Modeling → Manage Database
- Applications → Phase → Manage Phase Database

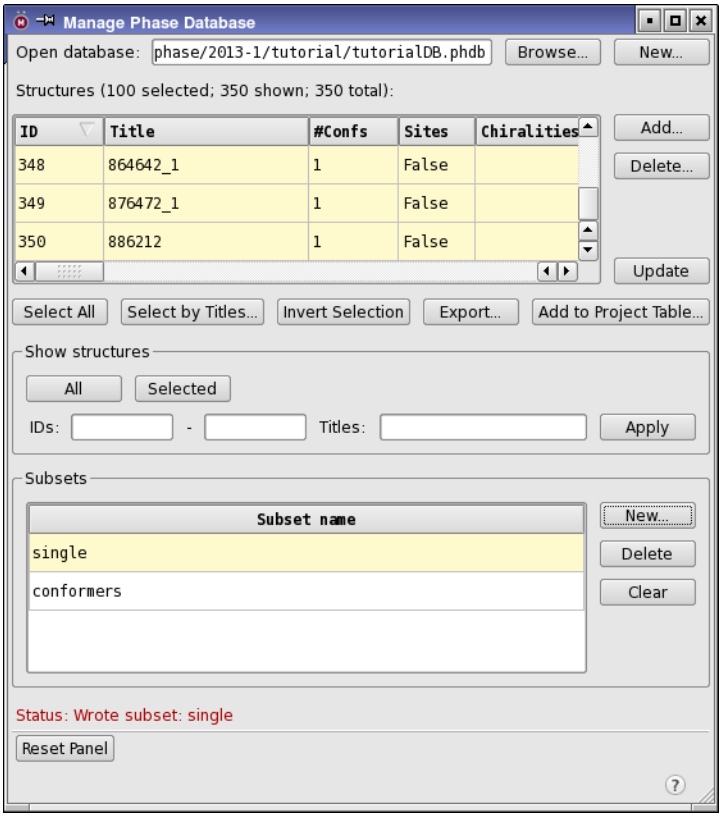

Figure 10.2. The Manage Phase Database panel.

10.2.1 Specifying the Database

You can specify an existing database, or create a new, empty database, on which to perform management tasks.

To open an existing database, enter the name of the database to use in the Open database text box, or click Browse and navigate to the database.

- If you open a database in the current format (the database name ends in `.phdb`). the structures in the database are listed in the Structures table.
- If you open a database in an earlier format, the Convert Phase Database panel opens, and you can convert it to the new format. When the conversion finishes, you can open the new database. The panel is unavailable until the conversion finishes.

To create a new, empty database, click New. A file selector opens, in which you can navigate to the location for the database and name the database. When you click Save, the database infrastructure is set up.

To close a database, click Reset. All the data is cleared from the panel.

10.2.2 Displaying and Selecting Structures

The structures in the database are listed in the Structures table, subject to the conditions set in the Show structures section. Each row represents a molecule, which can be included as a single structure or as a set of conformers. The first column shows the database record number (molecule ID), and the second column shows the structure title. The number of conformers is indicated in the #Confs column. The Sites column indicates whether sites have been generated for the molecule. The properties can be updated by clicking Update, which runs a `phase_database extract` task to extract the properties.

There are several ways of selecting structures in the table:

- Use the normal row selection actions: click, shift-click, and control-click.
- Click Select All to select all the structures in the database.
- Click Select by Titles to select structures by matching the titles with a list of titles read from a file. Opens a file selector, in which you can navigate to and open the file that lists the titles, one per line.
- Select a subset in the Subsets table to select all structures in the subset. If any structures in the subset are hidden, they are shown.

The structures that are listed in the table can be restricted in several ways, in the Show Structures section:

- To show only the structures that are currently selected, click Selected.
- To show only the structures in a particular range of IDs, enter the starting and ending IDs in the IDs text boxes, and click Apply.
- To show structures that match a filter on the title, enter text in the Filter text box and click Apply. This text can include the SQL wildcard character `%` for matching any number of characters. If you have a subset selected, the filter is applied to the structures in the subset.
- To remove the restrictions, click All.

10.2.3 Adding and Deleting Structures

To add structures to the database, click Add. A file selector opens, in which you can select a structure file. When you click Open, the file is used as input to a `phase_database import` job, which adds the structures to the database and generates sites for them. The structures must already be suitably prepared. If you want to perform preparation tasks on the structures, you can add structures to a database using the Generate Phase Database panel.

To delete structures from the database, select the structures in the table, then click Delete. When you confirm the deletion, a `phase_database delete` job is run to delete the structures.

10.2.4 Exporting Structures

The structures in the database can be exported to a file or added to the current Maestro project.

To export the selected structures to a Maestro (`.maegz`) or SD (`.sdf`) file, click Export. A dialog box opens, in which you can choose to export all conformers or only the first conformer for each molecule. When you click OK, a file selector opens, in which you can choose the format, navigate to the location, and name the file.

To add the selected structures to the Maestro project, click Add to Project Table. Each structure becomes a single entry. Conformer sets are added as entry groups.

10.2.5 Creating and Managing Subsets

You can create and manage database subsets in the Subsets section. These subsets are stored as subset files, ending in `_phase.inp`. The subsets are listed in the table, by name. The name is the file name without the `_phase.inp` extension.

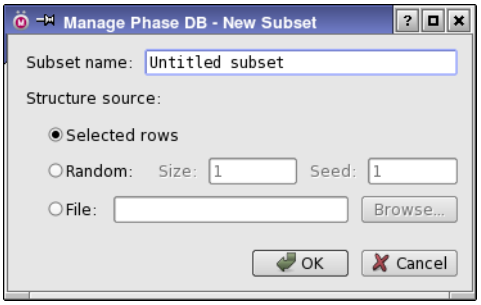

Figure 10.3. The Manage Phase DB - New Subset dialog box.

To create a new subset from the selected structures, from a random selection of the structures, or from a file, click New. This button opens the Manage Phase DB - New Subset dialog box, in which you can name the subset and select the structure source.

- If you select Selected rows, the subset is generated from the rows that are selected in the Structures table of the Manage Phase Database panel.
- If you select Random, you can specify the size of the subset (number of structures) and a seed for the random number generation.
- If you select File, enter the name of a subset file, or click Browse and navigate to the file. This option is useful for adding subsets that were generated or stored outside the database, with a new name.

When you click OK, the subset file is created, by appending the extension `_phase.inp` to the subset name.

To delete the selected subset, click Delete.

To clear the subset selection, click Clear.

10.3 Converting a Database

If you have databases from previous releases that you want to convert to the current release, you can do so in the Convert Phase Database panel, which you can open from Maestro by choosing Applications → Phase → Convert Phase Database.

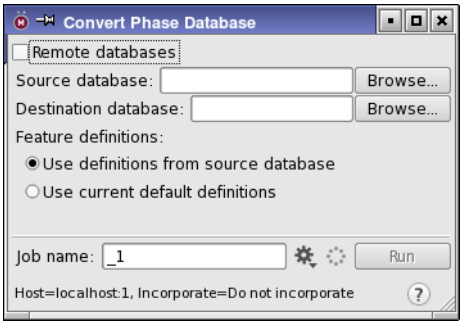

Figure 10.4. The Convert Phase Database panel.

The Convert Phase Database panel allows you to convert a database from the earlier format, in which the database was marked by a file whose name ends in `_phasedb`, to the current format, in which the database is a directory with the extension `.phdb`. The conversion is done by the `phase_database` application (see Section 13.1 on page 156), which is run as a job under Job Control. A new database is created without removing the old database.

To convert a database:

1. If the source database is on a remote host, select Remote database.

   The database conversion is done on the remote host.

1. Enter the full name of the old database in the Source database text box, or click Browse and navigate to the database.

   You must enter a name that ends in `_phasedb`. When the name is entered, the default destination location is filled in for in situ conversion: for an old database at *path*`/`*name*`_phasedb`, the default new name is *path*`/`*name*`.phdb`.

2. If you do not want to convert the database in situ, enter the full name of the new database in the Destination database text box, or click Browse and navigate to the database.

   You must enter a name that ends in `.phdb`. The file system on which the destination database will be written must exist on the host. The job name is set automatically from the base name of the new database.

3. Select an option for the feature definitions:

   - Use definitions from source database—Use the definitions from the old database. This option allows you to continue to use the old feature definitions with the new database format. This may be important for reproducing the results of previous runs with newer software.
   - Use current default definitions—Use the latest default feature definitions. The database conversion may result in a change in the sites identified on each molecule. Newer definitions are in general to be preferred, as they usually represent improvements in the science.

4. (Optional) Change the name in the Job name text box.

5. Click Run to run the job locally, or click the Settings button to open the Job Settings dialog box, where you can specify the host and number of processors and start the job.

   

   If you selected Remote database, you must specify a host on which the source database is available.

6. Choose whether to open the database or not.

   When the conversion finishes, you are prompted to open the database, if it was converted locally. If you click Open, the Manage Phase Database panel opens with the database loaded. If you click Don't Open, no action is taken on the database (it remains where it was, and is not opened).

After dismissing this dialog box, the status is reported as "Status: Conversion is complete" in the panel.

If the database conversion fails, you will need to determine why it failed, which you can do by examining the log file in the Monitor panel. If the source database was incomplete for some reason, the database cannot be converted, and you should consider removing the newly created database. If the conversion failed because of resource issues, you can restart the database conversion from the command line (see Section 13.1.5 on page 158). Conversion subjobs are automatically restarted up to 3 times if they fail.

To start another database conversion, choose Reset Panel from the Settings button menu to clear the panel, then specify the next database and convert it.

10.4 The eMolecules Database

Schrödinger provides the eMolecules database for purchase, prepared for use with Phase. Updates to this database will be available regularly, on a monthly basis. These updates can be requested from Schrödinger and installed with the `emol_update` script. To run the script, open a terminal window on Linux or Mac and set `$SCHRODINGER` to the installation path, or open a Schrodinger Command Prompt window on Windows.

To request an update:

1. Run the following command:

   `emol_update` *path*`/eMolecules.phdb`

   Prepend the command with `$SCHRODINGER/utilities/` on Linux or Mac.

2. Upload the resulting `eMolecules_`*date_time*`.zip` file to the support page, http://www.schrodinger/com/supportcenter.

   This file contains a record of the current state of your database. You will be sent an email message with the location of the database update on the FTP site for you to download.

To install the update:

1. Download the update from the FTP site.
2. Run the following command:

   `emol_update` *path*`/eMolecules.phdb` *update*`.zip`

   Prepend the command with `$SCHRODINGER/utilities/` on Linux or Mac.

Finding Matches to Hypotheses

Pharmacophore hypotheses can be used to search a set of structures for matches to the hypothesis. You can search (or "screen") a prepared 3D database or a file of 3D structures. Phase provides two approaches to screening for hits: a simplified approach in which you can either import or create a hypothesis, then screen your structure set with a minimal set of option; and an approach in which the full range of options for screening is available.

11.1 The Search Process

The search process proceeds as follows. First, the structure set is searched for geometric arrangements of pharmacophore sites that match the site types and intersite distances of the chosen hypothesis. For example, the hypothesis DHRR contains one donor (D), one hydrophobe (H) and two aromatic rings (R1, R2). These four pharmacophore features give rise to six unique intersite distances: dDH, dDR1, dDR2, dHR1, dHR2 and dR1R2. The structure set is scanned for occurrences of the four feature types for which the six intersite distances are sufficiently close to those of the hypothesis. When such an occurrence is found, the match is recorded.

Next, the relevant conformers are retrieved and aligned to the hypothesis. We refer to these conformers as *hits*. When hits are fetched, they are ordered and filtered, so that only a fraction of the total number of matches is presented. The hits are ordered first by their fitness score, then filtered by number, and by occupation of excluded volumes if these are defined. Finally, the activity is predicted if there is a QSAR model available. The resulting hits are added to the Project Table as an entry group.

The search offers considerable flexibility in the matching process. A hypothesis might have more features than are actually needed for binding, for example, but there might be some uncertainty about exactly which features are responsible. You can then require that only a certain minimum number of features must match. This is known as *partial matching*. As another example, you might know that a ligand cannot bind to a particular receptor unless it contains a positive site and an aromatic ring. You could then require these features to match. You might also know that a ligand cannot bind if it contains a hydrophobic site in some particular location, and you can require that this feature does not match. Specifying which features must match or must not match is done with a *site mask*. Also, because certain types of ligand-receptor interactions are stronger and more specific, it often makes sense to define different tolerances on matching different types of features. It might also be necessary to

distinguish between different instances of the same feature type. Both types of tolerances can be adjusted when a search is set up.

You can also search for matches in which there are fewer than 3 site points that match, either because the hypothesis only has 1 or 2 site points, or because only 2 matches are required in partial matching. In this case, of course, a full alignment cannot be done, only a translation, or a translation and a rotation about a single axis. This type of search makes most sense when the structures are prealigned, and only scoring is required.

11.2 The Fitness Score

Hits are first fetched in order of decreasing *fitness*. Fitness is a score that measures how well the matching pharmacophore site points align to those of the hypothesis, how well the matching vector features (acceptors, donors, aromatic rings) overlay those of the hypothesis, and how well the matching conformation superimposes, in an overall sense, with the reference ligand conformation. The fitness score is defined by

$$S = W_{\text{site}}\,(1 - S_{\text{align}}/C_{\text{align}}) + W_{\text{vec}}\,S_{\text{vec}} + W_{\text{vol}}\,S_{\text{vol}} + W_{\text{ivol}}\,S_{\text{ivol}}\,. \tag{1}$$

The terms in the score are described in Table 11.1. This score is a truncated version of the survival score for hypotheses. See Section 6.1 on page 50 for more information on the various terms in the scoring function and how they are defined.

By adjusting the parameters in the fitness function, you can control the order in which hits are returned. For example, if you want to emphasize the alignment of vector features, you could increase the vector weight. The volume term provides a means of forcing the shape of the hit to resemble that of the reference conformation. If the overall molecular superposition is most important, you could increase the volume weight.

If you choose to match fewer than the number of sites in the hypothesis ("partial matching"), the survival score is modified to penalize the hits that do not match all sites. If there are n sites in the hypothesis, and m sites are matched, the alignment score is modified as follows:

$$S_{\text{new}} = \sqrt{W_{\text{old}}\,S_{\text{old}}^2 + W_{\text{new}}\,P_{\text{align}}^2} \tag{2}$$

where $W_{\text{old}} = m/n$, $W_{\text{new}} = (n\text{-}m)/n$, and the alignment penalty P_{align} is adjustable, and has the default value 1.2.

If the hypothesis has fewer than 3 sites or a molecule matches on fewer than 3 sites, the alignment score is set to zero.

Table 11.1. Description of parameters in the fitness scoring function.

Parameter	Description
S_{align}	Alignment score: RMS deviation between the site point positions in the matching conformation and the site point positions in the hypothesis.
C_{align}	Alignment cutoff. User-adjustable parameter; default is 1.2.
P_{align}	Alignment penalty for partial matches. User-adjustable parameter; default is 1.2.
W_{site}	Weight of site score. User-adjustable parameter; default is 1.0.
S_{vec}	Vector score: average cosine between vector features in the matching conformation and the vector features in the reference conformation.
W_{vec}	Weight of vector score. User-adjustable parameter; default is 1.0
S_{vol}	Volume score: Ratio of the common volume occupied by the matching conformer and the reference conformer, to the total volume (the volume occupied by both). Volumes are computed using van der Waals models of all non-hydrogen atoms.
W_{vol}	Weight of volume score. User-adjustable parameter; default is 1.0
S_{ivol}	Included volume score: Ratio of the volume overlap between the matching conformer and the included volumes (if present) to the total included volume. Volumes are computed using van der Waals models of all atoms except nonpolar hydrogens.
W_{ivol}	Weight of volume score. User-adjustable parameter; default is 0.0

11.3 Setting Up a Simplified Search

The Simplified Pharmacophore Modeling and Screening panel provides a simple interface for the definition of a pharmacophore hypothesis and its use to screen a set of compounds. The range of features is limited, but it provides the main options. The full range of features is available in the Manage Hypotheses panel and the Advanced Pharmacophore Screening panel.

You can open the Simplified Pharmacophore Modeling and Screening panel as follows:

Maestro:

- Tasks → Pharmacophore Modeling → Simplified Modeling and Screening
- Applications → Phase → Simplified Pharmacophore Modeling and Screening

Elements:

- Tasks → Pharmacophore Search

BioLuminate:

- Tasks → Ligand Tasks → Pharmacophore Modeling → Simplified Modeling and Screening

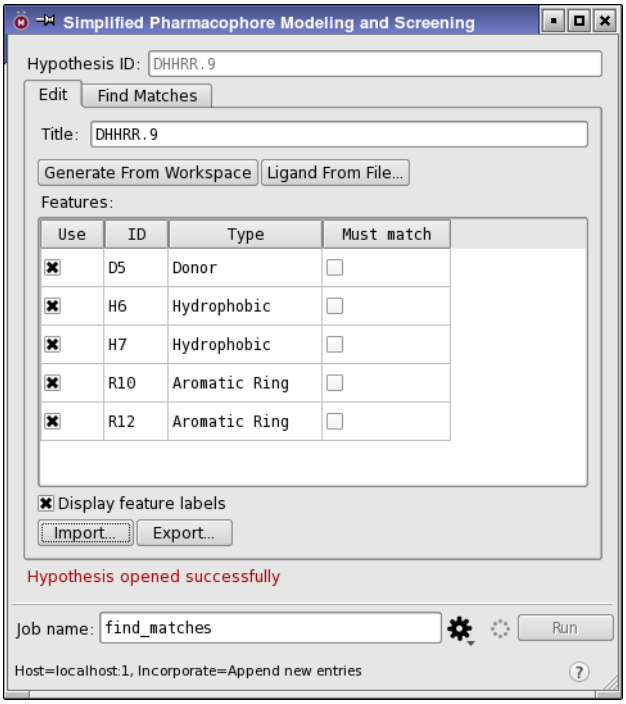

Figure 11.1. Simplified Pharmacophore Modeling and Screening panel - Edit tab.

11.3.1 Defining the Hypothesis

The hypothesis you use for screening can be imported or it can be created from a ligand.

- **To import a hypothesis:** Click Import and navigate to the hypothesis file in the file selector.

  The Features table is filled in, and you can make selections in the table before doing the search.

- **To create a hypothesis from a ligand structure:** Load and analyze a reference structure with one of the following methods:

  - Display the ligand in the Workspace and click Generate from Workspace.
  - Click Ligand from File and navigate to the file that contains the ligand. The file must contain a single structure.

  You can use a ligand that is part of a ligand-receptor complex, and the ligand is extracted for the hypothesis reference structure. The pharmacophore features are detected and listed in the Features table.

- **To change the title of the hypothesis:** edit it in the Title text box.

  This title is used as the entry title in the Maestro file that is written for the reference ligand. A default title is supplied when you create or import a hypothesis.

- **To select hypothesis features:** Check the boxes in the Use column of the Features table.

  Initially all boxes are checked, and the features are displayed in the Workspace on the ligand. When you clear a check box, the feature is undisplayed, and it is not used in the hypothesis for screening.

- **To require matches to features:** Check the boxes in the Must match column.

  By default, features are not required to match. These requirements are only needed when the number of sites that must match is less than the number of sites in the hypothesis (partial matching), which is decided in the Find Matches tab.

- **To export a hypothesis:** click Export, and navigate to a location in the file selector.

Note: If you edit the ligand structure in the Workspace, the panel is reset, because the feature list is no longer valid. You must then click Generate from Workspace to reload the Feature table.

11.3.2 Finding Matches

The steps for setting up a screening job are:

1. Choose the source of structures to screen against the hypothesis.
2. You can choose from a Phase 3D database, a file, or selected entries in the Project Table. If you choose a database or a file, you must supply the path in the File path text box, or click Browse to navigate to the file or database. For a file, the file type must be either Maestro or SD. If the database is not accessible to the host on which you are setting up the job, select the Remote option.
3. Select the source of the conformations to use:
   - Use existing conformers—Use the existing conformers in the database or the file. You must generate conformers before you run the search.
   - Generate during search—Generate the conformers as they are needed during the search.
4. Specify the number of site points in the hypothesis that must be matched in any structure for it to be considered a hit, in the Must match on at least N site points text box.

   If you specify less than the maximum number, the Must match settings in the Edit tab determine whether matching a particular site is required or optional.

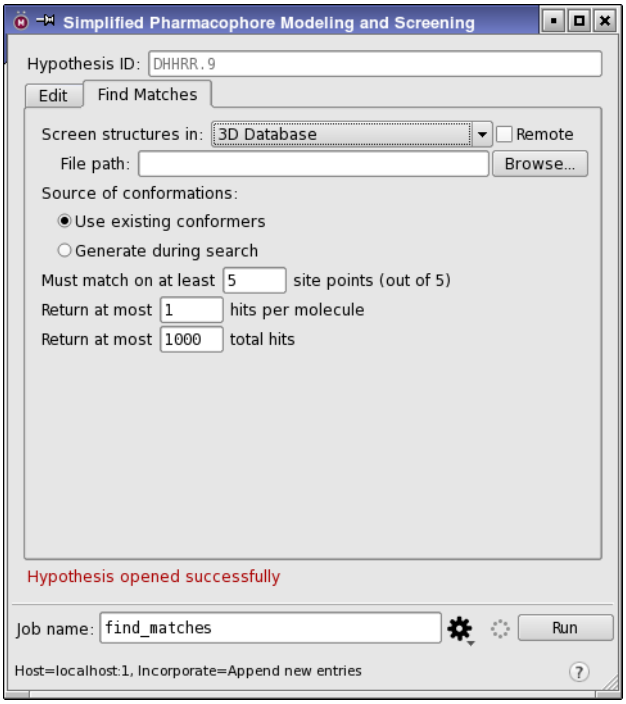

Figure 11.2. Simplified Pharmacophore Modeling and Screening panel - Find Matches tab.

5. Specify the number of hits to return per molecule in the Return at most *N* hits per molecule text box.

   The best scoring *N* hits for each molecule are returned. Some molecules can return more than one hit because different alignments or different conformers might match the hypothesis.

6. Limit the maximum number of hits returned, in the Return at most *N* total hits text box.

   The hits that are returned are the top-scoring hits, ordered from highest to lowest score of the best hit for each molecule.

When you have finished making the search settings, click Start. The Start dialog box opens, in which you can name the job, choose a host and the number of processors, and start the job.

Note: If you are screening a local database, Phase checks whether the feature definitions for the hypothesis are compatible with those in the database, and warns you. If the database is not accessible from the host where you start the job, no checking can be done, and the job will fail if the feature sets do not match.

11.4 Setting Up an Advanced Search

The Advanced Pharmacophore Screening panel offers the full flexibility for tuning the matching of structures to a hypothesis. To screen a structure set, you must select the source of the structures and a hypothesis, set any options for matching and for filtering and treatment of the hits, then click Start. The options are described in the following sections.

You can open the Advanced Pharmacophore Screening panel as follows:

Maestro:

- Tasks → Pharmacophore Modeling → Advanced Screening
- Applications → Phase → Advanced Pharmacophore Screening
- Click Search for Matches in the Develop Common Pharmacophore Hypotheses panel

BioLuminate:

- Tasks → Ligand Tasks → Pharmacophore Modeling → Advanced Screening

The search process is outlined in the next section, followed by a section on scoring of the hits. The subsequent sections describes how to set up a search, and the final section discusses the output.

If you want to run the job from the command line, or simply to generate the input files, click Write. A dialog box opens in which you can specify a name that is used for the file stem of the input files. When you click Write in this dialog box, the files are written to the current directory.

11.4.1 Selecting a Structure Source

The collection of structures that you use to search for matches can come from one of three sources: a prepared 3D database, a file, or the Project Table. The structures must be all-atom 3D structures.

- To search a 3D database, choose 3D database from the Search in option menu, then either enter a name in the File name text box, or click Browse and navigate to the desired database (`.phdb`). If you want to search a subset of structures from the database, enter the name of the subset in the Subset text box, or click Browse, and select a subset in the Select Subset dialog box. Subsets can be generated in the 3D database management process. By default, the database selected is the last database you used.

  If the database is located on a remote host on a file system that is not accessible to the local host, select Remote database. You must then enter the full path to the database on the remote file system in the File name text box. The usual checking done by Maestro is bypassed, and is done instead at the beginning of the search job.

You can specify a subset that is not part of the database, by entering the file name in the Subset text box. The name must end in `_phase.inp`. The file need not exist on the local file system: it can be on a remote host, such as the host where the database is stored.

- To search the structures in a file, choose File from the Search in option menu, then either enter a file name in the File name text box, or click Browse and navigate to the desired file. The file must be in Maestro or SD format, and can be compressed with gzip (`.mae.gz`, `.maegz`, `.sdf.gz`, or `.sdfgz`) or uncompressed.
- To search structures from the Project Table, select the entries to search in the Project Table, then choose Project Table (selected entries) from the Search in option menu.

11.4.2 Selecting a Hypothesis

Hypotheses are stored as entries in the Project Table. To choose a hypothesis for the search, click Choose. An entry chooser opens, with a list of entries that have hypotheses. Select the hypothesis, and click OK. The entry ID and the hypothesis ID are displayed in the Hypothesis text box.

If you open the panel from the Score Hypotheses step or the Build QSAR Model step of the Develop Common Pharmacophore Hypotheses panel, the selected hypotheses are added to the Project Table, and the first added hypotheses is chosen as the default hypothesis.

You can add hypotheses to the Project Table by importing them: click Import, and navigate to the desired hypothesis in the file selector that is displayed. However, you can only import hypotheses that were previously exported. You cannot import hypotheses from another project.

You can display the hypothesis, its excluded volumes, its intersite distances and angles, and its feature-matching tolerances in the Workspace by clicking the toolbar buttons. The first of these buttons displays the hypothesis; the rest of these buttons are the same as in the Manage Hypotheses panel, and are described in Section 8.1 on page 85.

The feature definitions for the hypothesis should match those in the database. If they do not match, you can compute pharmacophore sites as needed using the hypothesis feature definitions, without replacing the existing sites in a database.

11.4.3 Selecting the Source of Conformations

The search for matches to a hypothesis requires conformations of each molecule searched. If the source of structures includes sets of conformers, you can select Use existing conformers in the Searching tab. If the source does not include conformers, select Generate during search in the Searching tab, to generate them during the search. The conformer generation uses the ConfGen method (see Section 3.3 on page 21 for details), with a restricted set of options. You

can set options in the Conformer generation and refinement section of this tab. In either case, structures with different stereochemistry are considered to belong to different conformer sets.

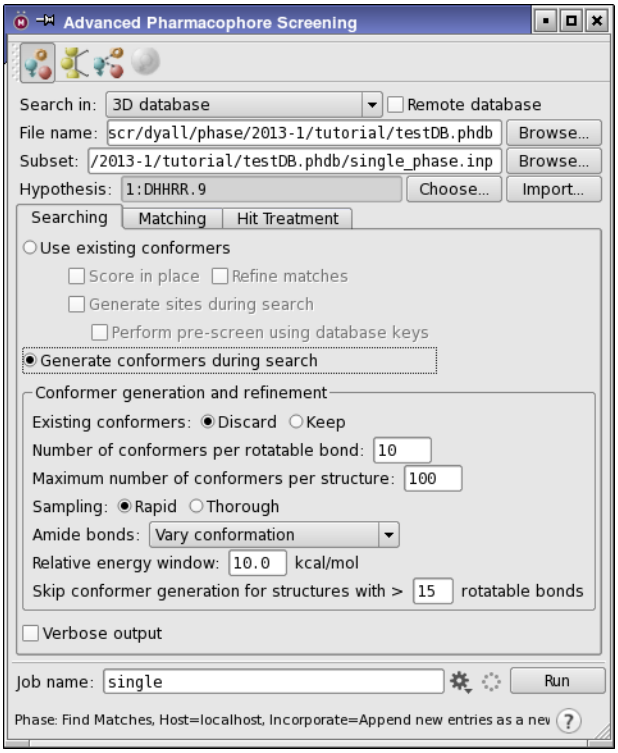

Figure 11.3. The Advanced Pharmacophore Screening panel, Searching tab.

Searching a set of molecules that does not include conformers is called "flexible searching" because the conformational flexibility is explored during the search. The storage requirements for flexible searching are much smaller than for standard searching, but the search can take up to 10 times longer.

Two other options are available when using existing conformers:

- Score in place—Score structures without doing any alignment to the hypothesis. This option is only available when you are searching in a file.
- Refine matches—Generate extra conformers for the top-ranked match from each molecule matched, using ConfGen with the options in the Conformer generation and refinement section of this tab. These conformers are subjected to the hit filters and are returned in the hit file. This procedure improves the fitness score, and can return matches when excluded volumes eliminate every match for some molecules.

Two more options are available when you are searching a database with existing conformers:

- Generate sites during search—Generate sites during the search using the hypothesis feature definitions, rather than those in the database. These sites are used only during the search, and are not written out anywhere. This option allows you to use different sites from those in the database without affecting the database.
- Perform pre-screen using database keys—To speed up the search, you can perform a pre-screening of the database using database keys. Use of these 3D keys rapidly filters out the majority of molecules that cannot possibly match the hypothesis. This option is recommended except when searching a small subset of the database or when the search is split across a large number of CPUs. It is not available if you generate sites during the search, because the keys are generated using the feature definitions.

11.4.4 Setting Options for a Conformational Search

If you are generating conformers during the search, or refining matches by performing a conformational search, you can set the parameters for the conformational search in the Conformer generation and refinement section of the Searching tab. Conformer generation is performed with ConfGen—see Section 3.3 on page 21 or the *ConfGen User Manual*.

- Existing conformers—When you generate conformers, you can discard the existing conformer set or you can keep it. If you keep the existing set, the new conformations are appended to the set. The set might therefore contain redundant conformers. If you discard the existing set, the new set will not contain the original conformer.

There are two options that control how many conformers are generated:

- Number of conformers per rotatable bond—This value specifies the maximum number of conformers to generate in the conformational search for each rotatable bond in the structure. The number given in the text box is multiplied by the number of rotatable bonds in the molecule to arrive at the maximum number of conformers.
- Maximum number of conformers—This value limits the total number of conformers returned. If the number of conformers generated is higher than this value, a sample of all the conformers generated is returned. If the maximum specified here is lower than that derived from the number of conformers per rotatable bond, the lower value is used.

The sampling of conformational space is controlled by two options.

- Rapid—All the core conformations are generated, and the conformations of the peripheral (rotamer) groups are sampled one by one.
- Thorough—A complete set of conformations is generated for both the core and the peripheral groups.

In addition, you can specify how amide bonds are treated by selecting from the Amide bond option menu. There are three options for treatment of amide bonds in the search. You can retain the original amide bond conformation in the input structure, you can set the conformation to trans, or you can vary the conformation. Varying the conformation allows the amide dihedral angle to take any value, not just cis or trans.

You can filter out conformers that are too high in energy, by specifying a maximum energy relative to the lowest conformation in the Relative energy window text box. Higher-energy conformers are discarded.

If the structure is too flexible, conformation generation can take a long time. You can set a maximum number of rotatable bonds that a structure can have in the Skip conformer generation for structures with greater than *N* rotatable bonds text box. Structures with more rotatable bonds are skipped, not only in the conformational search but in the search for matches.

11.4.5 Setting Options for Matching

In the Matching tab you can set various options to control the matching process.

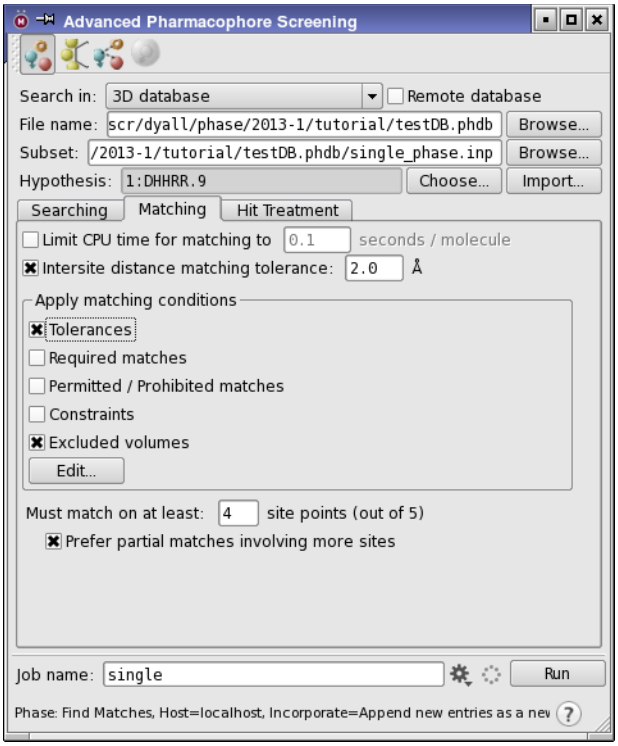

Figure 11.4. The Advanced Pharmacophore Screening panel, Matching tab.

- You can limit the CPU time spent per molecule by selecting Limit CPU time for matching to *N* seconds/molecule and entering a value in the text box. Limiting the CPU time is useful when doing partial matching for hypotheses with many sites, as the time spent finding matches to a given molecule may become very large because of combinatorial considerations. If the time limit expires, the matches found for the conformers that have been processed are returned.
- To change the tolerance for intersite distances, enter a value in the Intersite distance matching tolerance text box. Any intersite distance less than the specified value is considered to match. This tolerance may be adjusted when site-matching tolerances are in use.
- To set site-matching tolerances, require certain sites to match, allow or forbid sites to match other features, set constraints on intersite distances, angles, or dihedrals, select the relevant options in the Apply matching conditions section. If you want to change any of these conditions, click Edit, and make the desired settings for each site in the Matching Conditions dialog box. These settings are described in the next section. You can also choose whether to apply excluded volumes in this section, if the hypothesis has them.
- If you want to match fewer sites than there are in the hypothesis (partial matching), enter the minimum number of site points in the Must match on at least *N* site points text box. Matches are made on any number of site points from the minimum up to the total number of site points in the hypothesis, but matches that are made on less than the total number of site points are penalized in the survival score—see Section 11.2 on page 116.

  The minimum number is normally 3 site points, so that the hits can be aligned to the hypothesis. If you match only 2 site points, no alignment is done, and you cannot use excluded volumes, which require alignment, unless you are using prealigned conformers with Use existing conformers and Score in place selected in the Searching tab.
- If you are matching fewer than the maximum number of sites and want only the matches that have the greatest number of sites that match, select Prefer partial matches involving more sites. The number of sites that must match is reduced systematically from the maximum to the minimum number until matches are found.

11.4.6 Setting Site-Specific Matching Criteria

If you want to apply different matching criteria to each site or change the default criteria for each site, you can make settings in the Matching Conditions dialog box, which you open by clicking Edit in the Matching tab. You can select options to use site-matching tolerances, required site matches (“site mask”), and permitted and prohibited matches (matching rules); and make selections or set values for each of these features. If the hypothesis already has any of these features defined, they are read when the dialog box opens and used to set options.

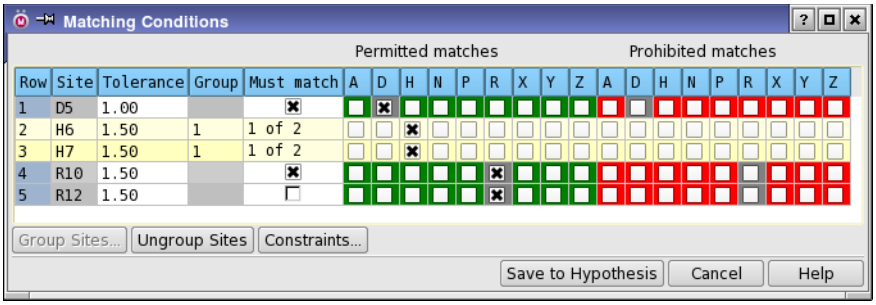

Figure 11.5. The Matching Conditions dialog box.

The dialog box displays a table containing the matching tolerance, site mask, and permitted and prohibited matches. Once you have made settings in the table, you must also select the appropriate Apply option to apply the settings during the search.

- To set matching tolerances for individual sites, enter the value in the Tolerance column of the table. The site on the matched structure must be within the specified distance of the site on the hypotheses, once the structure is aligned to the hypothesis.
- To require matches to certain sites, click in the Must match column for that site.
- To require a minimum number of matches to several sites, you can create a site group, and then select the number of these sites that must match. First, select the sites in the table, then click Group Sites. In the Group Sites dialog box, you can specify the number that must match. The number of sites that match and the number in the group are displayed in the Must match column, and the group index is displayed in the Group column. You can change the number that must match by clicking in the Must match column. This feature is for partial matching: you cannot require all members or no members to match (which you can do with ungrouped sites).
- To allow sites to match feature types other than that of the site (the “native” feature type), click the relevant feature column under Permitted matches in the table. The native feature type is always selected and cannot be deselected. For example, you might want to allow an aromatic ring to also match a hydrophobic site.
- To prohibit sites from inadvertently matching a particular type of site, click the relevant feature column under Prohibited matches in the table. The native feature type is always unselected and cannot be selected. This capability is useful if you are not matching all sites, to ensure that the sites that are not matched do not accidentally match some other site in the molecule that has an inappropriate type. For example, you might want to prohibit inadvertent matching of an ionic site to a hydrophobic site. The prohibited matches are eliminated after alignment on the basis of permitted matches, using the site-specific tolerances (half the intersite distance tolerance if positional tolerances are not defined).

If you want to set constraints on intersite distances, angles, or dihedrals that must be met when finding matches, click Constraints, and make the settings in the Constraints dialog box.

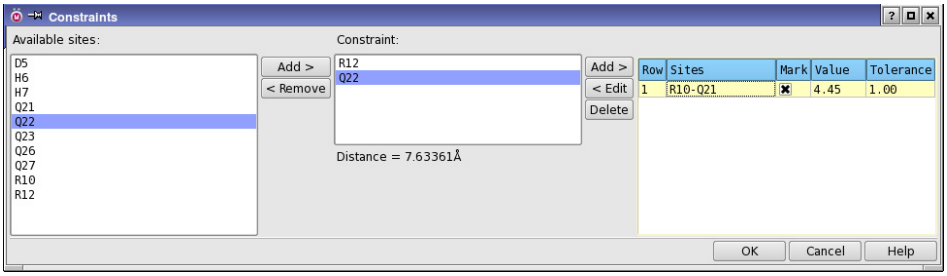

Figure 11.6. The Constraints dialog box.

- To define a constraint, select sites in the Available Sites list and add them to the Constraint list, one at a time, with the Add button. The sites must be added in the correct order for defining the desired angle or dihedral. If you make a mistake, you can remove a site from the Constraint list by selecting it and clicking Remove. When you have added the correct number of sites (2 for a distance, 3 for an angle, 4 for a dihedral), click the Add button to the right of the Constraint list to add it to the constraints table. The value of the distance, angle, or dihedral is listed in the table. A default tolerance is assigned, which you can change by editing the table cell. You can display the constraint in the Workspace by checking the box in the Mark column.
- To change the sites used in a constraint, select the constraint in the table and click Edit. The constraint is removed from the table and placed in the Constraint list, where you can add or remove sites.
- To delete a constraint, select it in the constraints table and click Delete.

When you have finished defining the constraints click OK. To apply the constraints, you must select the Apply option Constraints in the Advanced Matching Options dialog box,

If you require matches to fewer than the total number of sites and want hits for the minimum number of sites, you must also enter the minimum number in the Must match on at least *N* site points text box in the Matching tab.

You can save the matching criteria with the hypothesis, by making selections in the table and clicking Save to Hypothesis. The relevant files (`.dxyz`, `.cnst`, `.mask`, `.rules`) are then saved in the same location as the hypothesis: in the project, if the hypothesis came from the project, or to external files if the hypothesis was imported. For more information on these files and their use, see Section B.10, Section B.11, Section B.12, and Section B.13.

11.4.7 Setting Filtering and Scoring Options

At the end of the search, the matches are filtered to generate a reduced list of hits, and properties are calculated for these hits. Filters and scoring options are set in the Hit Treatment tab. The options that require files in the hypothesis (such as QSAR models, included volumes) are only active if the files are actually present in the hypothesis.

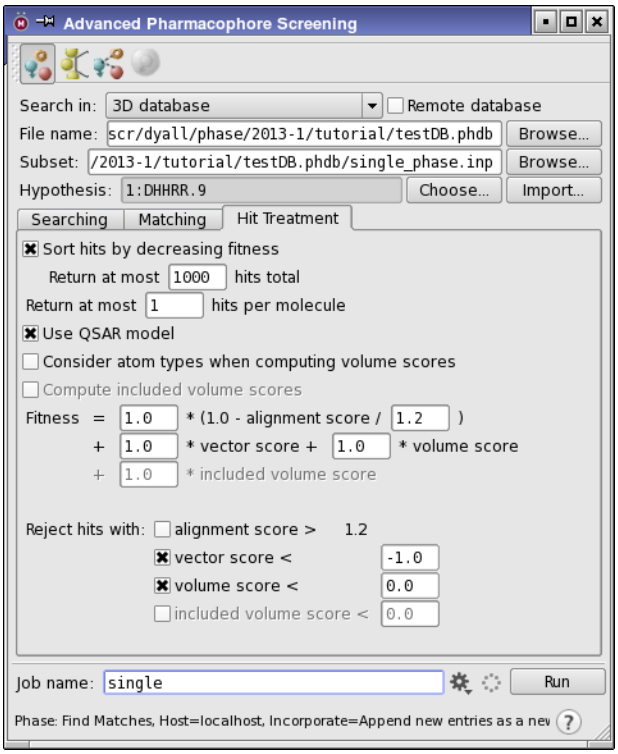

Figure 11.7. The Advanced Pharmacophore Screening panel, Hit Treatment tab.

- If you want the matches returned in order of fitness, select Sort hits by decreasing fitness.
- To limit the number of hits returned, enter values in the Return at most text boxes. Some molecules can return more than one hit because different alignments or different conformers might match the hypothesis.
- To calculate activities for the hits based on the QSAR model, if one is available, select Apply QSAR model. The hits are not ordered by the activities before they are returned, but you can sort by activity in the Project Table. Activities are calculated for each model, defined by the number of PLS factors in the model. This option is selected by default if the hypothesis has QSAR models, otherwise it is unselected by default.

- If you want the volume score to reflect the chemistry of the molecules rather than just the shape, select Consider atom types when computing volume scores. Atoms are then considered to overlap only if they have the same MacroModel atom type. This option favors alignments that superimpose chemically similar atoms.
- If the hypothesis has included volumes, you can compute a score based on the overlap of the van der Waals spheres of the hit with those of the included volume (nonpolar hydrogens in the hit are not included in the calculation). To do so, select Compute included volume scores. The score can be added to the fitness function or used to filter out hits. This option is only active if the hypothesis has an included volumes (`.ivol`) file.
- To change the fitness scoring function, enter new values of the weights in the text boxes.
- To filter out matches that do not satisfy criteria set on the alignment, vector, volume, or included volume scores, select the appropriate Reject hits with option. For the vector and volume scores, you can change the threshold for rejecting hits by editing the values in the text boxes.

11.4.8 Setting the Amount of Output

Below the tabs is an option for control of the amount of output. Selecting the Verbose output option provides details of each match in the output file. By default, only one line is written per match. You should only select this option if you are interested in the details of the matching, as it could generate a large amount of output if there are many matches.

11.5 Search Results

Each time you search the database, the hits are added as an entry group to the Project Table, where you can use the full range of applications and facilities available from Maestro. The fitness score and the activity predicted by the QSAR model (if any) are added as properties, along with a property that indicates which hypothesis was matched. A list of properties is given in Table 11.2.

You can view the hits superimposed on the hypothesis by including them in the Workspace, and displaying the hypothesis from the Find Matches to Hypothesis panel, using the toolbar buttons. If you want to cycle through the hits, you can use the ePlayer: select the hits in the Project Table, display the ePlayer toolbar, then click the Play forward button:

You can change the speed at which the hits are displayed in the ePlayer Options dialog box, which you open from the ePlayer menu.

Table 11.2. Maestro properties generated in a database search

Property	Description
phasedb index	Database record index.
Hit Source	Full path to the database used.
Ligand Name	Name of the ligand.
Conf Index	Index of the matching conformation.
Num Sites Matched	Number of sites matched.
Matched Ligand Sites	String that indicates which sites in the hit matched the hypothesis. The sites are listed in the order of occurrence in the hypothesis. Each site is indicated by the letter for the type of site that matched with the index of the ligand site matched in parentheses. For example `D(6)` means that site 6 on the ligand matched a donor in the hypothesis. A dash in parenthesis means that the hypothesis site did not match. An example of this property is as follows: `D(7) H(-) H(11) R(15) R(16)` When using feature-matching rules, it is the ligand feature that matches that is listed rather than the hypothesis feature. For example, allowing aromatic rings to match hydrophobes, you could have the following list of sites for the same hypothesis as in the previous example: `D(5) R(15) H(-) R(11) R(12)`
Align Score	Alignment score—see Table 11.1 on page 117
Vector Score	Vector score—see Table 11.1 on page 117
Volume Score	Volume score—see Table 11.1 on page 117
Included Volume Score	Included volume score—see Table 11.1 on page 117
Fitness	Fitness score as defined in Equation (1).
Pred Activity(*n*)	Predicted activity from the 3D QSAR model with *n* PLS factors.

Pharmacophore Model Development from the Command Line

To develop a pharmacophore model from a set of ligands, you must prepare the ligands, generate pharmacophore sites for the ligands, find common pharmacophores, then score the resultant hypotheses. You can also build QSAR models for any of the hypotheses. These five steps are referred to by the names used in the Phase panel in Maestro: Prepare Ligands, Create Sites, Find Common Pharmacophores, Score Hypotheses, and Build QSAR Model.

The first step, Prepare Ligands, is the only step that does not have Phase utilities to perform the task. The ligands you provide must be properly prepared and stored in one or more Maestro files prior to starting the workflow. Each molecule should be represented by multiple low-energy 3D structures that provide good coverage of that molecule's conformational space. See Chapter 8 of the *MacroModel User Manual* and the *ConfGen User Manual* for information on creating conformational models, and see the *LigPrep User Manual* for information on 2D-to-3D conversion and structure variation. If you want to develop QSAR models, then each molecule that will be used to train models should contain an activity property, expressed either in concentration units or as –log[concentration].

Within each Maestro file, conformers for a single molecule must be stored consecutively. If two consecutive structures differ only in their stereochemistry, they are treated as conformers of a single molecule unless the titles for those two structures are different.

12.1 Workflow Summary

The complete command-line pharmacophore model development workflow is outlined below, in terms of the scripts to run. These scripts are described in detail in the following sections; links to the relevant script are provided in the summary below. The starting point is one or more Maestro files containing multiconformer models for the ligands of interest. A Phase project is created from these ligands, after which a series of steps is followed, directly analogous to the Develop Common Pharmacophore Hypotheses workflow in Maestro.

Each step requires a Phase main input file, and other files that are stored in the current working directory, a subdirectory of this directory, or the Phase distribution. Output files are created in the current working directory, or specified subdirectories of this directory. Most of the required job setup, including creation of the input files, is handled with the `-setup` options of the utilities listed, and cleanup of temporary and intermediate files is done with the `-cleanup` option. For details on the Phase main input file, see Section B.2 on page 187.

Ligand structures and generated ligand-related information is stored in a subdirectory whose default name is `ligands`; this subdirectory is referred to as the *ligands directory*. Output of the scoring step is stored in a subdirectory whose default name is `result`; this subdirectory is referred to as the *results directory*.

The top-level programs that perform the work accept the standard Job Control options listed in Table 2.1 and Table 2.2 of the *Job Control Guide*, except for those related to Maestro projects. These options are represented by *job-options* in the syntax statements.

Create/Add to a Project:

```
$SCHRODINGER/utilities/pharm_project {-new|-add} [options]
```

Modify Master Data:

```
$SCHRODINGER/utilities/pharm_data [options]
```

Create Pharmacophore Sites:

```
$SCHRODINGER/utilities/pharm_create_sites -setup [setup-options]
$SCHRODINGER/phase_feature create_sites [job-options]
$SCHRODINGER/utilities/pharm_create_sites -cleanup
```

Find Common Pharmacophores:

```
$SCHRODINGER/utilities/pharm_find_common -setup [setup-options]
$SCHRODINGER/phase_partition find_common [job-options] or
$SCHRODINGER/phase_multiPartition find_common [options] [job-options]
$SCHRODINGER/utilities/pharm_find_common -cleanup
```

Score Hypotheses with Respect to Actives:

```
$SCHRODINGER/utilities/pharm_score_actives -setup [setup-options]
$SCHRODINGER/phase_scoring score_actives [job-options]
$SCHRODINGER/utilities/pharm_score_actives -cleanup
```

Score Hypotheses with Respect to Inactives:

```
$SCHRODINGER/utilities/pharm_score_inactives -setup [setup-options]
$SCHRODINGER/phase_inactive score_inactives [job-options]
$SCHRODINGER/utilities/pharm_score_inactives -cleanup
```

Cluster Hypotheses by Geometric Similarity:

```
$SCHRODINGER/utilities/pharm_cluster_hypotheses -setup [setup-options]
$SCHRODINGER/phase_hypoCluster cluster_hypotheses [job-options]
$SCHRODINGER/utilities/pharm_cluster_hypotheses
    -cleanup [cleanup-options]
```

Build QSAR Models:

```
$SCHRODINGER/utilities/pharm_build_qsar -setup [setup-options]
$SCHRODINGER/phase_multiQsar build_qsar [job-options]
$SCHRODINGER/utilities/pharm_build_qsar -cleanup
```

Preserve Project Data in a Tar Archive:

```
$SCHRODINGER/utilities/pharm_archive [options]
```

Once pharmacophore hypotheses and QSAR models have been developed, a number of other command line utilities may be run:

Align Project Ligands or New Molecules to a Pharmacophore Hypothesis:

```
$SCHRODINGER/phase_find_matches [options]
```

Align or Merge a Pair of Hypotheses:

```
$SCHRODINGER/utilities/align_hypoPair [options]
```

Create Excluded Volumes Automatically:

```
$SCHRODINGER/utilities/create_xvolShell [options]
$SCHRODINGER/utilities/create_xvolClash [options]
$SCHRODINGER/utilities/create_xvolReceptor [options]
```

Analyze QSAR Predictions within Hit Files:

```
$SCHRODINGER/utilities/phase_qsar_stats [options]
```

Visualize QSAR Models:

```
$SCHRODINGER/utilities/qsarVis [options]
```

12.2 Pharmacophore Model Development Utilities

The pharmacophore model development utilities are stored in `$SCHRODINGER/utilities`. Except where noted, all changes to pharmacophore project files should be done only through the use of the utilities listed below. Brief descriptions of the use of the utilities is given below; fuller descriptions are given in the following sections.

`pharm_help`	Prints a help message summarizing the command line pharmacophore model workflow, including all the utilities that follow.
`pharm_project`	Creates a new command line pharmacophore model project and adds molecules to an existing project.
`pharm_data`	Performs various operations on the project data.
`pharm_create_sites`	Does setup/cleanup for the job that creates pharmacophore sites.
`pharm_find_common`	Does setup/cleanup for the job that identifies common pharmacophores.
`pharm_score_actives`	Does setup/cleanup for the job that scores hypotheses with respect to actives.
`pharm_score_inactives`	Does setup/cleanup for the job that scores hypotheses with respect to inactives.
`pharm_cluster_hypotheses`	Does setup/cleanup for the job that clusters hypotheses by geometric similarity.
`pharm_cluster_modes`	Cluster ligands by possible binding mode.
`pharm_build_qsar`	Does setup/cleanup for the job that builds QSAR models.
`pharm_archive`	Preserves project data in a tar archive.
`pharm_align_mol`	Obsolete. Does setup/cleanup for the job that aligns project ligands or new molecules to a hypothesis.
`align_hypoPair`	Aligns/merges a pair of hypotheses.
`create_xvolShell`	Creates a shell of excluded volume spheres around one or more ligands. Provides a means of defining shape-based queries for database searching.
`create_xvolClash`	Creates excluded volumes using actives and inactives that have been aligned to a hypothesis. Excluded volumes are placed in locations that would cause steric clashes only for the inactives.
`create_xvolReceptor`	Creates excluded volumes using a receptor structure or a portion thereof.
`phase_qsar_stats`	Extracts statistics from a hit file that contains QSAR predictions.
`qsarVis`	Standalone graphical interface for visualizing QSAR models. Requires X terminal emulation.

In addition to the above utilities, the following programs are in the `$SCHRODINGER` directory. These programs accept the standard job options that are described in Section 12.1.

`phase_feature`	Creates pharmacophore sites.
`phase_partition`	Identifies common pharmacophores.
`phase_multiPartition`	Runs phase_partition one or more times to identify common pharmacophores containing the largest possible number of sites.
`phase_scoring`	Scores hypotheses with respect to actives.
`phase_inactive`	Scores hypotheses with respect to inactives.
`phase_hypoCluster`	Clusters hypotheses by geometric similarity.
`phase_multiQsar`	Builds 3D QSAR models for a collection of hypotheses, and generates a statistical summary for each model.
`phase_qsar`	Builds a single 3D QSAR model, and generates detailed output.

12.3 Setting Up a Phase Pharmacophore Model Project

Phase pharmacophore model projects are a collection of files, managed by a utility called `pharm_project`. These projects are *not* the same as the corresponding Maestro projects, but the results of pharmacophore model development—the hypotheses—can be imported into Maestro. In addition to managing the structures in the project with the utility `pharm_project`, you can add or change certain data associated with the structures with the utility `pharm_data`. These two utilities are described in the next two sections.

12.3.1 pharm_project

This utility creates a new command line pharmacophore model project or adds ligands to an existing project. Conformations must be generated ahead of time and stored in a Maestro file. Consecutive structures with identical titles and connectivities are treated as conformations of a single molecule. If you want to treat stereoisomers as a different molecule, you must use a different title for each stereoisomer, or use the `-stereo` option to check stereochemical properties in the structures. The syntax and option descriptions can be listed by running the command with the `-h` option.

Output Files

`ligands/`	Subdirectory that holds all structural data for the project ligands.
`ligands/*.mae`	Individual ligand files split out from the input files.

MasterData.tab	A specially formatted text file that holds project data required in various steps of the workflow. Certain modifications are permitted (by hand or through the use of pharm_data).
MasterData.backup	A backup copy of MasterData.tab. Used to revert changes you make to MasterData.tab. Do not modify.
ProjectLigands.inp	Ligand records file. Provides a compact summary of project data, and serves as a template for creating subsets of ligands to align to a hypothesis. While the file can be modified without affecting the integrity of project, it is recommended that you leave it as is, and make a copy of the file if you need to define a subset.
FeatureFreq.tab	Feature frequency file. Sets minimum and maximum allowed feature frequencies for common pharmacophore perception.
FeatureTol.tab	Feature matching tolerances that can be applied when hypotheses are scored with respect to actives.
pharma_feature.ini	Default pharmacophore feature definitions. You may replace this file with customized definitions, but it is strongly recommended that you do the customization with the Phase interface in Maestro.

12.3.2 pharm_data

This utility performs various operations on MasterData.tab and propagates any changes in this file to the rest of the project. This includes changes that may have been made by hand. The syntax and options can be listed by running the command with the -h option.

If you make changes to MasterData.tab, whether by hand or through operations supported by pharm_data, you must use the -commit option to update the Maestro files in the ligands subdirectory. If you do not update the Maestro files, the Phase utilities and programs will not use the modified values, because they read the property data directly from the Maestro files. If you plan to run pharm_data multiple times to make a series of changes, then you need only use-commit the final time you run pharm_data. You can even run pharm_data with nothing but -commit—this is how you can commit changes made by hand.

If you have completed any forward steps in the project workflow, the results generated in those steps may be invalidated by changes you make to MasterData.tab. When you attempt to commit the changes, you will be supplied with a list of files from forward steps that will be invalidated, and you will be given a chance to abort the commit operation. If you choose to abort, you can rerun pharm_data with the -restore flag to revert to the previous version of MasterData.tab (i.e., the data stored in MasterData.backup).

If your activities are expressed in concentration units (e.g. K_i or IC_{50} values) and you intend to create QSAR models, then you must use the `-log` and `-commit` options. You must also perform the `-log` conversion on concentrations if you plan to assign `PHARM_SET` categories using the `-active` or `-inactive` options, because these assignments are based on the assumption that the `ACTIVITY` property increases as potency increases.

12.4 Creating Sites

Pharmacophore sites are created by the `phase_feature` program. A set of pharmacophore feature definitions is applied to each ligand conformation, to identify the positions of all pharmacophores sites in that ligand.

12.4.1 pharm_create_sites

Performs pharmacophore site creation setup and cleanup. Requires completion of project setup. The syntax and options can be listed by running the command with the `-h` option.

Output Files

The files generated by the `-setup` option are:

`create_sites_feature.ini`	A copy of the default feature definition file, `pharma_feature.ini`.
`create_sites_phase.inp`	Main input file for `phase_feature`.

The file generated by the `-cleanup` option is:

`CreateSitesData.tab`	Summary of the pharmacophore feature counts for each ligand.

12.4.2 phase_feature

Generates pharmacophore sites for one or more ligands from a set of defined features. If you used `pharm_create_sites` to set up the job, *jobname* is `create_sites`, and the relevant input files are set up automatically.

Input Files

jobname`_phase.inp`	Phase main input file, which contains options that govern Phase behavior. See Section B.2 on page 187 for details of this file. The list of ligands should be restricted to the ligands to be used in the model development.

*jobname*_feature.ini	Feature definitions file. This file can be created from the template feature definitions file (pharma_feature.ini) by incorporating any changes to the standard features. The template file is located in $SCHRODINGER/phase-v*version*/data. It is strongly recommeded to edit this file using Maestro.
mmphob.ini	File that contains definitions for hydrophobic groups. This file is optional. Unless a local copy is supplied this file will be read from the default location in the mmshare installation.
ligand-name.mae.	Files containing ligand structures. These files should be stored in the ligands subdirectory, as specified in the Phase main input file. The default directory name is ligands. Each ligand file is a multi-conformer Maestro file. Ligand names should be listed in the Phase main input file as LIGAND_NAME = *ligand-name*. You should only list the active ligands to be used in the model.

Output Files

The following output files are created upon successful job completion:

*jobname*_phase.log	Log information, including the lists of mapped features for each ligand.
*ligand-name*_sites.phs	Pharmacophore site coordinates for the ligand specified by *ligand-name*. These files are created in the ligands directory.
*ligand-name*_xyz.phc	Atom coordinates of each conformer for the ligand specified by *ligand-name*. These files are created in the ligands directory, and are needed for running Phase scoring jobs. These files have a stripped-down format that allows rapid access to conformer structural data in subsequent steps of the workflow.

12.5 Finding Common Pharmacophores

Common pharmacophores are identified by the `phase_partition` program. All *n*-point pharmacophores from the `PHARM_SET` ligands are enumerated and filtered into a set of high-dimensional boxes. Pharmacophores in the same box are similar enough to be considered equivalent. Boxes with at least one pharmacophore from a sufficient number of actives are said to "survive" the partitioning process. See Chapter 5 for details on the process. The `phase_multiPartition` program runs `phase_partition` to identify common pharmacophore models with the highest number of *n*-point pharmacophores.

When setting up a `phase_partition` job, you must decide on the number of sites, how many ligands must match, and whether to restrict the number of features that can occur. The number of sites can range from three to seven, but the most meaningful and useful pharmacophore models typically contain between four and six sites. The recommended approach is to start with five or six sites, and decrease that number only if no common pharmacophores are found. You can do this automatically with `phase_multiPartition`. Likewise it is recommended to first require that all actives match the pharmacophore, and reduce the number only if no common pharmacophores are found. You can require certain actives to match and create ligand groups from structurally-related ligands (such as tautomers or ionization states) by adding data to the ligand blocks in `MasterData.tab`—see Table B.2 on page 184.

12.5.1 pharm_find_common

Performs setup and cleanup for common pharmacophore perception. Requires completion of the Create Sites step. The syntax and options can be listed by running the command with the `-h` option.

Output Files

The file generated by the `-setup` option is:

`find_common_phase.inp` Phase main input file for `phase_partition`.

The file generated by the `-cleanup` option is:

`FindCommonPharmData.tab` Summary of the number of boxes for each variant.

12.5.2 phase_partition and phase_multiPartition

The `phase_partition` program is used to find common pharmacophores for a given set of ligands. This step is also known as a partitioning job (from the name of underlying algorithm). This program can be run on multiple processors, which are specified with the `-HOST` option.

The phase_multiPartition program runs phase_partition multiple times, starting with the highest number of site points, and decreasing the number of points successively until common pharmacophores are found. Both programs use the same input file.

Syntax

```
$SCHRODINGER/phase_partition jobname [job-options]
$SCHRODINGER/phase_multiPartition jobname [-minSites n] [job-options]
```

The -minSites option specifies the minimum number of site points to consider.

Input Files

*jobname*_phase.inp	Phase main input file with options for this type of Phase job.
*ligand-name*_sites.phs	Site files for each of the ligands in the set, created by a phase_feature run. These files are located in the ligands directory, which is specified in the Phase main input file.
FeatureFreq.tab	Feature frequency file. Used to set mimimum and maximum allowed feature frequencies for common pharmacophore perception. See Section B.8 on page 203 for an example.

Output Files

The following intermediate and output files are created by the job in the working directory:

*jobname*_partition.inp	Partitioning input file, which is generated automatically from the Phase main input file. Used by the computational program, and is useful mainly for troubleshooting purposes.
*jobname*_partition.out	Output file. Contains some information about the job, but is mainly useful for debugging.
*jobname*_partition.log	Log file. Contains information on job progress, including boxes generated for each variant and eliminated variants.
*jobname*_partition_variants.tab	File containing list of variants and number of boxes for each variant.
*jobname*_boxes.tar	Archive of box file archives generated by the partitioning code. Box files are archived for each variant. This archive is used by the subsequent scoring job.

12.6 Scoring Hypotheses

The Score Hypotheses stage of the workflow involves calculating scores for each possible hypothesis based on ligand alignment, volume overlap, and various properties. Only the highest-scoring hypotheses are kept. For a detailed description of how scoring is done, see Chapter 6. Scoring does not eliminate redundant hypotheses that arise from site permutations, which are treated as distinct by the partitioning algorithm. Redundancies can be identified by applying a clustering technique based on geometric similarity.

Hypotheses are scored with respect to the active `PHARM_SET` ligands by the program `phase_scoring`. This process assigns numerical rankings to the pharmacophores within each surviving box from the Find Common Pharmacophores step. The highest scoring pharmacophore in a given box is designated as a hypothesis, and the ligand giving rise to that pharmacophore is known as its reference ligand. The scoring function considers the quality of the alignments afforded by each pharmacophore, along with a number of other user-configurable factors. See Section 6.1 on page 50 for more information on the scoring process.

In addition to scoring hypotheses, it is useful to eliminate hypotheses that are geometrically very similar or identical. It is not uncommon for two or more hypotheses to have very similar or even identical scores and physical characteristics. This is a consequence of the way in which common pharmacophores are perceived. Since the partitioning algorithm operates on an ordered set of intersite distances, it is necessary to consider all permutations among sites of the same type when enumerating pharmacophores. So, for example, the permutations $A_1H_2H_3R_4R_5R_6$ and $A_1H_3H_2R_4R_5R_6$ represent the same pharmacophore, but their 15-dimensional intersite distance vectors would generally not be identical, and they may in fact be dissimilar enough to end up in different boxes. As a result, a given box may have a mirror box that contains many (though not necessarily all) of the same pharmacophores, giving rise to a mirror hypothesis that is indistinguishable from the original, or nearly so. These sorts of redundancies are readily identified, by applying a technique that clusters hypotheses based on geometric similarity.

12.6.1 pharm_score_actives

Performs setup and cleanup for scoring of actives. Requires the completion of the Find Common Pharmacophores step. The syntax and options can be listed by running the command with the `-h` option.

If you modify the scoring function with any of these options, you should examine the range of values of the property to choose an appropriate weight. In general, it is advisable to ensure that contributions to the scoring function are in the range 0.0 to 1.0 in magnitude, to prevent the contribution from completely dominating the scoring function.

Output Files

The following files are created with the `-setup` option:

`score_actives_phase.inp`	Phase main input file.
`score_actives_feature.ini`	Feature definitions file, as provided to the prior `phase_feature` job.
`score_actives_boxes.tar`	Copy of the archive of box files generated by the `phase_partition` job.

The following files are generated with the `-cleanup` option.

`ScoreActivesData.tab`	Plain text summary of results in tabular form.
`ScoreActivesData.csv`	Summary of results in comma-separated value form.
hypoID`.def`	Feature definitions for the given hypothesis. Stored in the `hypotheses` subdirectory.
hypoID`.mae`	Maestro format file containing aligned actives for the given hypothesis. Stored in the `hypotheses` subdirectory.
hypoID`.tab`	Primary hypothesis data for the given hypothesis. Stored in the `hypotheses` subdirectory.
hypoID`.xyz`	Site coordinates for the given hypothesis. Stored in the `hypotheses` subdirectory.

12.6.2 phase_scoring

Scores and ranks pharmacophore hypotheses for actives. This program can be run on multiple processors, which are specified with the `-HOST` option.

Volume scoring is approximated by a pairwise sum of overlaps, which produces results that are very similar to rigorous volume scoring. To use the rigorous method (the default prior to Suite 2011) set the environment variable `SCHRODINGER_PHASE_USE_OLD_VOLUME` to any value.

Input Files

jobname`_phase.inp`	Phase main input file.
jobname`_feature.ini`	Feature definitions file, as provided to the `phase_feature` job.
`mmphob.ini`	Hydrophobic groups definitions file, as provided to the `phase_feature` job.

*jobname*_boxes.tar	Archive of box files generated by the phase_partition job. Box files are archived for each variant. This archive is expanded internally during execution.
ligand-name.mae	Files containing ligand structures, in the ligands directory, as provided to the phase_feature job.
*ligand-name*_sites.phs	Pharmacophore site coordinate files for each ligand, in the ligands directory, as generated by phase_feature.
*ligand-name*_xyz.phc	Ligand conformation files, in the ligands directory, as generated by phase_feature.
FeatureTol.tab	Feature-matching tolerances file. Optional.

Output Files

*jobname*_scoring.log	Log file. Contains information on job progress. Stored in the current directory.
*jobname*_scoring.tar	Archive file that contains all the results of the scoring job. Stored in the current directory.
*jobname_variant*_hypothesis.tab	File containing a list of hypotheses for the given variant, ordered according to hypothesis rank. Stored in the archive file.
*jobname_variant*_scores.out	File containing information about hypotheses for each variant. *variant* is encoded from the variant name as a string of integers; for example AADDH is encoded as 00112. Stored in the archive file.
*variant*_str_*N*.mae	Maestro format file containing ligand structures aligned onto the reference ligand for a given hypothesis. *N* is the unique identifier of the box from which this hypothesis came. Stored in the archive file.
*variant*_hyp_*N*.xyz	Site coordinate information file for the given hypothesis. Stored in the archive file.

12.6.3 pharm_score_inactives

Performs setup and cleanup for scoring of inactives. Requires the completion of the pharm_score_actives step. The syntax and options can be listed by running the command with the -h option.

Hypotheses are scored with respect to inactive molecules by the program phase_inactive. Survival scores are adjusted to penalize hypotheses that match inactives, on the assumption

that the inactives fail to bind because they do not contain the true pharmacophore. See Section 6.3 on page 55 for more information.

Output Files

The following files are created with the `-setup` option:

`score_inactives_phase.inp`	Phase main input file.
`score_inactives_inactive.inp`	Input file for `phase_inactive` (see Section B.4 on page 194).
`score_inactives_feature.ini`	Feature definitions file, as provided to the prior `phase_feature` job.
`score_inactives_hypoFiles.tar`	Archive of hypothesis files.
`score_inactives_ligandFiles.tar`	Archive of inactive ligand structure files.

12.6.4 phase_inactive

Scores pharmacophore hypotheses for inactives. Uses the same volume scoring method as for phase_scoring. This program can be run on multiple processors using the `-HOST` option (see Table 2.1 of the *Job Control Guide*).

Input Files

Uses the same input files as `phase_scoring`, and the following file in addition.

jobname`_inactive.inp` Phase inactives input file.

The main input file must list only the inactives in the `LIGAND_NAME` records. This is done automatically by `pharm_score_inactives`.

Output Files

The following output files are generated by this job:

jobname`_inactive.log`	Log file. Contains information on job progress. Stored in the current directory.
`ScoreInactivesData.tab`	Plain text summary of results in tabular form.
`ScoreInactivesData.csv`	Summary of results in comma-separated value form.

12.6.5 pharm_cluster_hypotheses

Performs setup and cleanup for the clustering of hypotheses. Clustering is performed by the program `phase_hypoCluster`. Requires completion of the `pharm_score_actives` step. The syntax and options descriptions can be listed by running the command with the `-h` option.

Output Files

The following files are created with the `-setup` option:

`cluster_hypotheses_phase.inp`	Phase main input file.
`cluster_hypotheses_hypoCluster.inp`	Input file for `phase_hypoCluster` (see Section B.5 on page 196).
`cluster_hypotheses_feature.ini`	Feature definitions file used to create hypotheses.
`cluster_hypotheses_hypoFiles.tar`	Archive of hypothesis files.

The following files are generated with the `-cleanup` option.

`ClusterHypothesesData.tab`	Plain text summary of results in tabular form.
`ClusterHypothesesData.csv`	Summary of results in comma-separated value form.

12.6.6 phase_hypoCluster

Hierarchical agglomerative clustering of hypotheses is performed by `phase_hypoCluster`, using a geometric similarity computed from the least-squares alignment of each pair of hypotheses i, j:

$$\mathrm{Sim}(i,j) = \frac{\langle\, i \mid j \,\rangle}{\sqrt{\langle\, i \mid i \,\rangle \langle\, j \mid j \,\rangle}}$$

where

$$\langle\, i \mid j \,\rangle = S_{\mathrm{site}}(i,j)\, W_{\mathrm{align}} + S_{\mathrm{vec}}(i,j)\, W_{\mathrm{vec}}$$

The site and vector scores are computed just as in the Score Actives step (see Section 6.1 on page 50). When more than one mapping is possible, the alignment yielding the highest similarity is used. Hypotheses that do not contain the same pharmacophore features (i.e., different variants) are assigned a similarity of zero, because the purpose of the clustering procedure is to distinguish hypotheses that are geometrically equivalent from those that are not.

Input Files

*jobname*_hypoCluster.inp Hypothesis clustering input file.

The remainder of the input files are specified in the hypothesis clustering input file—see Section B.5 on page 196.

Output Files

*jobname*_hypoCluster.log Log file. Contains information on job progress and results in readable form. Stored in the current directory.

The name of the output file containing the results of the cluster analysis is specified in the hypothesis clustering input file—see Section B.5 on page 196.

12.7 Building QSAR Models

The Build QSAR Model step develops a QSAR model based on partial least-squares (PLS) analysis for one or more hypotheses. The ligands are aligned to each hypothesis as part of the process. For more information on the QSAR model, see Section 7.1 on page 71.

12.7.1 pharm_build_qsar

Performs setup and cleanup for building a 3D QSAR model. Requires the completion of the pharm_score_actives step. A QSAR model is created for each hypothesis by the program phase_multiQsar. The syntax and options descriptions can be listed by running the command with the -h option.

Output Files

The following files are created with the -setup option:

build_qsar_multiQsar.inp	Input file for phase_multiQsar (see Section B.6 on page 197).
build_qsar_hypoFiles.tar	Archive of hypothesis files.
build_qsar_ligandFiles.tar	Archive of inactive ligand structure files.

The -cleanup option extracts the results archive from the phase_multiQsar job into the directory BuildQsarResults. These files are listed in the next section.

12.7.2 phase_multiQsar

This is a driver program that builds QSAR models for multiple hypothesis, by running `phase_qsar` on individual hypotheses and collecting the results. The syntax and options can be listed by running the command with the `-h` option.

Input Files

jobname`_multiQsar.inp`	Multiple QSAR model input file.

The remainder of the input files are specified in the multiple QSAR model input file—see Section B.7 on page 200.

Output Files

jobname`_multiQsar.log`	Log file. Contains information on job progress.
jobname`_multiQsar.tar`	Archive file (`.tar`). Contains files relating to QSAR models.
`BuildQsarData.tab`	Plain text summary of results in tabular form.
`BuildQsarData.csv`	Summary of results in comma-separated value form.

Files for each hypothesis are stored in the subdirectory specified by the `resultDir` keyword in the input file, which is set to `BuildQsarResults` by `pharm_build_qsar`, inside the archive file. These files are listed below.

hypoID`_align.mae`	Aligned structures for training and test set molecules that matched at least 3 sites in the hypothesis. Training set molecules appear first. Stored in archive file.
hypoID`.def`	Feature definitions (copied from input). Stored in archive file.
hypoID`.mae`	Reference ligand structure (copied from input). Stored in archive file.
hypoID`_order.dat`	File that defines the overall order of molecules in *hypoID*`_align.mae`.
hypoID`_pharm.dat`	The `pharmFile` required by `phase_qsar`, if a pharmacophore-based model was chosen (see Section B.7 on page 200).
hypoID`.qsar`	QSAR model file.
hypoID`_qsar.inp`	Main input file for `phase_qsar` job.
hypoID`.rad`	Copy of the feature radius file, if specified in the input file.
hypoID`.tab`	Primary hypothesis data, with QSAR model flag activated.

hypoID.tol	Copy of the feature cutoff file, if specified in the input file.
hypoID.xyz	Hypothesis site coordinates (copied from input).

12.7.3 phase_qsar

The phase_qsar program creates and applies grid-based 3D QSAR models. It makes activity predictions and generates detailed output for individual QSAR models for a single hypothesis. If pharmFile is specified in the input file, a feature-based QSAR model is developed or tested rather than an atom-based model. To generate a pharmFile, run phase_fileSearch on the Maestro file that contains the molecules of interest, with pharmFile specified in the input file to phase_fileSearch. The syntax and options can be listed by running the command with the -h option.

Input Files

*jobname*_qsar.inp	QSAR model input file, described in Section B.7 on page 200.

The other input files are specified in the QSAR model input file.

Output Files

*jobname*_qsar.log	Log file. Contains information on job progress.
*jobname*_qsar.out	Output file. Contains complete model statistics, Cartesian coordinates and regression coefficient for each bit in the model, training and test set predictions, and actual bit values for each molecule.

The other output files are specified in the QSAR model input file.

12.7.4 phase_qsar_stats

Extract statistics from Phase QSAR models and from hit files that contain QSAR predictions. The syntax and options can be listed by running the command with the -h option.

12.7.5 qsarVis

This utility allows you to visualize QSAR models that you create in command line projects. It launches a graphical interface entitled Visualization Tookit – OpenGL, with an interactive 3D image of the ligand, hypothesis, and QSAR model. The QSAR model is displayed in a similar way to the Maestro interface—see Section 7.5 on page 80 for more information. You can rotate, translate, and zoom in using the mouse, but the controls are different from those in Maestro. To rotate the image, drag with the left mouse button; to translate, drag with the middle mouse button; to zoom, drag with the right mouse button.

To change the visualization settings, you must start a new instance of `qsarVis`. However, if you place each instance in the background, you can display and compare QSAR models with various settings.

Note: This utility is only available on Linux platforms.

The syntax and options can be listed by running the command with the `-h` option.

12.8 Adding Excluded Volumes to a Hypothesis

A molecule may satisfy a pharmacophore model, but fail to bind to the associated receptor due to steric clashes. These clashes can be included by defineng excluded volumes, which are used to filter out matches that have any atoms inside these volumes. The Phase distribution contains three utilities, described in the following sections, that allow you to create excluded volumes in an automated fashion, using varying amounts of ligand and receptor information.

12.8.1 create_xvolShell

This utility creates a shell of excluded volume spheres to surround the supplied molecules. This shell defines the outer boundary of a shape-based constraint that can be applied when searching for matches to the hypothesis. The assumption is that the supplied molecules define the binding pocket, and molecules that do not fit in this shell will not fit into the receptor.

The syntax and options descriptions can be listed by running the command with the `-h` option.

12.8.2 create_xvolClash

Creates excluded volumes using actives and inactives that have been prealigned to a pharmacophore hypothesis. Excluded volumes are placed in locations that would cause steric clashes only for the inactives. The syntax and options can be listed by running the command with the `-h` option.

12.8.3 create_xvolReceptor

Creates excluded volumes from a receptor structure or a portion of a receptor structure. An excluded volume sphere is created for each atom in the receptor structure that satisfies the minimum and maximum distance criteria. The syntax and options can be listed by running the command with the `-h` option.

12.9 Other Utilities

12.9.1 pharm_archive

Archives forward steps in a Phase pharmacophore model project using `tar` and `gzip`, allowing data to be preserved before it is overwritten when a step is rerun. The syntax and options can be listed by running the command with the `-h` option.

12.9.2 align_hypoPair

Aligns one hypothesis onto another. Alignment is done using least-squares fitting of the matching site points in the two hypotheses, considering all possible mappings. Alignments are summarized to standard output in order of increasing RMSD. By default, only the best alignment is saved as a new hypothesis, but this can be overridden. The syntax and options can be listed by running the command with the `-h` option.

12.9.3 create_hypoConsensus

This utility creates a consensus hypothesis from a set of prealigned ligands or a set of hypotheses. The approach involves first performing complete linkage hierarchical clustering on each type of site: cluster all the acceptors, cluster all the donors, and so on. The clusters that are retained are those in which all the sites in a given cluster are within a user-specified distance. Then a representative site is chosen from each cluster. By default, the representative is the one closest to the centroid of the cluster. You can also control the number of ligands that must be represented in a cluster (default is all), and constrain the consensus hypothesis to contain only sites that are matched by an existing hypothesis. The consensus hypothesis that is created has no reference ligand, so the output consists only of the `.xyz` and `.def` files.

The syntax and options can be listed by running the command with the `-h` option.

12.9.4 phase_complex

Create a pharmacophore hypothesis from an all-atom fully prepared ligand-receptor complex. Sites are added to the hypothesis when any of the following complementary interactions are feasible:

Ligand...Receptor

A.....D
D.....A
H.....H
N.....P

P.....N
R.....R

The syntax and options can be listed by running the command with the `-h` option.

`phase_complex` *ligand receptor hypoID* [*options*]

Here, *ligand* is the Maestro or SD file containing the ligand, *receptor* is the Maestro file containing the receptor with no ligand, and *hypoID* is the prefix for all hypothesis files to be created. The ligand and receptor must be all-atom, 3D, fully prepared structures (for example, with LigPrep and the Protein Preparation Wizard).

`phase_complex` identifies and reports all complementary interactions between the ligand and receptor, and it adds each applicable ligand site to the hypothesis. You may see more interactions reported than there are sites in the hypothesis because a given site in the ligand may have more than one complementary site in the receptor, and vice versa. The options allow you to change the parameters for detection of complementarity, allow complementarity between hydrophobic and aromatic sites, to create a hypothesis for the receptor sites, and to create hypotheses based on projected points for acceptors and donors. The ligand and receptor hypotheses have the ligand and the receptor as reference structures, respectively.

Managing 3D Databases and Searching for Matches from the Command Line

Phase provides two ways of searching a set of structures for matches to a hypothesis: searching a plain Maestro or SD file, or creating a database for searching. In both cases, the process involves generating conformers of each structure and creating the site points. For a search on a file, the conformers can be detected in the file or generated during the search, and the site points are generated during the search. When a database is created, the conformers and the sites can be stored in the database. In this case, the search step only involves matching the sites to the hypothesis, which is much quicker than the conformational search. If you intend to search a set of structures more than once with the same set of features, you should consider creating a database and storing conformers in the database. Another benefit of creating a database is that you can define database subsets, which can be searched.

The structures that are used for searching must be all-atom, 3D structures in the correct ionization state. If the structures are not in this form—for example, if they are 2D structures—they must be prepared in the correct form first, which you can do with LigPrep. See the *LigPrep User Manual* for details. If you already have a database from some other source that matches these requirements, you can use it for Phase searches by exporting it to an SD file.

The input files can be in Maestro or SD format. If you need to convert between SD and Maestro format, you can do so with the utility `structconvert` or the utility `sdconvert` (see Section 1.4 of the *General Utilities* manual). You can also use `structconvert` to convert another format to Maestro or SD format.

The database must be stored on a file system that is accessible to all hosts that will be used to create, modify, or search the database.

The current database format was introduced in Suite 2011. It is 30% of the size of the previous format, and is correspondingly faster to search. If you have databases in the earlier format, you should convert them to use the new format—see Section 13.1.5 on page 158.

13.1 Managing Databases with phase_database

Databases can be created and managed with the program, `phase_database`. This program performs all the necessary database management tasks. The syntax is as follows:

`SCHRODINGER/phase_database` *database task* [*jobName*] [*options*]

where *database* is the full path to the database, which must have the extension `.phdb`. The database is a directory, *dbName*`.phdb`, which contains the database files (much like a Maestro project or a Canvas project). Information on the database structure is given in Section 13.5 on page 161. The default for the optional job name is `database_`*task*. Error messages are written to *jobName*`_errors.out`.

The allowed tasks are summarized in Table 13.1. Write permissions to the database are required for `import`, `revise`, `extract`, `delete`, and `convert`. For information about these tasks, you can use `-help_`*task* to get options for a particular task, rather than `-help`, which gives the full help message.

Table 13.1. Available tasks for phase_database.

Task	Action
`import`	Import structures into a new or existing database.
`revise`	Add sites, conformers or Canvas properties.
`extract`	Extract all properties into a single SQLite table.
`query`	Perform a property or substructure query.
`delete`	Delete records.
`convert`	Convert or merge one database into another.
`export`	Export database records to structure files.
`subset`	Create or operate on database subsets.

The options for each task are described in the usage message, which you can display by running the `phase_database` command with the -h option. Notes on the tasks are given in the following subsections. The standard Job Control options, which allow you to specify the host, set the priority, and so on, are accepted. These are described in Table 2.1 of the *Job Control Guide*. The `-LOCAL` and `-NOJOBID` options, described in Table 2.2 of the *Job Control Guide*, are also accepted.

Restarting of failed `import`, `revise`, and `convert` tasks is supported with the `-RESTART` option. Complete instructions on restarting a job are stored in the database, at *database*`/database_restart/README`. The `revise` and `convert` task subjobs are automatically

restarted if they fail, up to 3 times, after which a failure is recorded. You can set the number of retries with the SCHRODINGER_PHASE_MAX_RESTART environment variable. Restarting is also tried on each slave process run by a subjob. If the slave process itself fails, it is retried up to 5 times (settable with the SCHRODINGER_PHASE_MAX_RETRY environment variable). At a finer level, failed structures are retried once, and then skipped.

13.1.1 Import Task

New record numbers are written to the file *jobName*_new_phase.inp.

If using the -unique option, a summary of each duplicate structure that is rejected is written to the file *jobName*_reject.out, containing the SMILES string and the title of the input structure and the molID and title of the database structure, as shown in this example:

```
Rejected: [O-]C(=O)[C@H](C)[C@H](C([O-])=O)Cc1cc(C)cc(C)c1 "385089_3"
Duplicate of: block_1/mol_21 "385089_3"
```

All duplicates encountered in the database are listed. Duplicates in the input file are also listed.

13.1.2 Revise Task

The revise job can be distributed across multiple CPUs. Any combination of -sites, -confs, and -props is allowed, but at least one must be used.

Conformer generation is not exhaustive, and therefore depends to some extent on the input structures. However, the conformers generated should represent a reasonable sample, and the results of a search should not depend much on the input structures. If you want to be sure that you have a complete set of conformers, you should run a conformational search beforehand (with MacroModel, for example) and import the conformer sets.

The feature definitions used for the sites are taken from the definition file copied into the database at the point it was created. If you want to use custom feature definitions, you can copy the feature definition file into the database by running an inport with the -fd option. You should do this when you create the database or before you run a revise task for the first time. If you change the feature definitions after creating sites, you can get inconsistencies in the database, and you should run a revise task with -sites to ensure that there are no inconsistencies.

13.1.3 Extract Task

All properties that have been imported or computed are extracted and written to a single table in the SQLite database *database*/database.sqlite. A copy of the same data is written to *database*/database_props.csv.

13.1.4 Query Task

Matching record numbers are written to *jobName*_matches_phase.inp, and all properties for those records are written to *jobName*_matches.csv.

13.1.5 Convert Task

The destination database is the first argument in the phase_database command (i.e., *database*), and it must have been created (or be created) using phase_database. The job can be distributed across multiple processors.

New sites are created in the destination database if there is an upgrade in the storage format, the feature definitions differ, or if there are no sites for a given record. If none of these conditions is met, the default is to copy the source database sites to the destination database.

The -nosites option is intended for merging a source database with no conformers or sites into a destination database with no conformers or sites, without adding sites in the destination database. This option ensures that sites are not created in the destination database for source records that are missing sites. If there is no upgrade in the database format, existing sites in the source database are copied (regardless of changes in feature definitions).

13.1.6 Subset Task

The options -hits, -has, -titles, and -logic are mutually exclusive. An existing database is not required with -hits and -logic, so the *database* argument is parsed but not used.

13.2 Database Import with phase_multi_database

Importing a large number of structures can be distributed over multiple processors with the program phase_multi_database. This program actually creates several mini-databases inside the target database then renames them to become part of the main database. There are several limitations to this process: conformer sets cannot be imported, and no checking is done for redundant structures. The syntax of the command is:

```
phase_multi_database database inFile -HOST host:n [-new [-fd fdFile]]
    [-title propName] [-blimit maxRec] [-nosplit [-nocopy]] [-verbose]
    [-JOB jobName] [-TMPDIR dir] [-WAIT] [-NICE] [-RESTART]
```

The destination database, *database*, must be specified with an absolute path and an extension of .phdb. This database can be an existing database (the default), or can be created by using the -new option.

inFile specifies the source of cleaned 3D structures to import. It must be one of the following:

- Maestro file (`.mae`, `.mae.gz`, `.maegz`).
- SD file (`.sdf`, `.sd`, `.sdfgz`, `.sdf.gz`, `.sd.gz`).
- List file (`.list`). This is a text file that contains the names of one or more Maestro or SD files, with one name per line. Files must all be of the same type (i.e., all Maestro or all SD) and have the same compression state. There is no attempt to perceive conformers, so each structure will be stored in a separate database record.

The `-HOST` option specifies the number of processors to use on the specified host. This is a standard job control option; however for this job it must be in the format *host*`:`*n*. The number of processors, *n*, is also the number of mini-databases created, and is reduced if it exceeds the number of database blocks created. By default, a database block consists of 5000 records (structures in this context) but this value can be reduced with the `-blimit` option. If `-nosplit` is used, *n* must be equal to the number of input structure files.

The `-TMPDIR`, `-NICE`, and `-WAIT` options are Job Control options, described in Table 2.1 and Table 2.2 of the *Job Control Guide*. The remaining options can be listed by running the command with the `-h` option.

13.3 Running on Multiple Processors

Apart from the parallel import of structures performed with `phase_multi_database`, site creation jobs and database search jobs can be split across multiple processors on an appropriately configured cluster. The basic requirements for database creation and searching are as follows:

- The database must be located in a directory that is uniformly accessible to all nodes of the cluster on which jobs will be run.
- In the `$SCHRODINGER/schrodinger.hosts` file, each parallel queue that is used for database jobs should have a `tmpdir` entry with a path that is accessible to all nodes. See the *Installation Guide* for details of setting a `tmpdir` entry.

When you start a `phase_database` or `phase_find_matches` job, the `-HOST` option should specify the processors to use. If the processors are on a single host, you can add the number of processors after the host name and a colon—for example, `-HOST cluster:4`. One subjob is started on each processor.

The molecules in the database are distributed to the subjobs by dividing the records in the database into contiguous record ranges, so that each subjob has approximately the same number of records. This is done without regard to the record blocks. However, if the number of record

blocks in the database is smaller than the number of processors, the number of processors is reduced, and each subjob processes one block.

For example, if your database has 100,000 records and you run a job across 8 processors, then each subjob will have 12,500 records: the first will process records 1–12,500, the second will process records 12,501–25,000, and so on. But if the database has only 10,000 records and the block size is 5000 (the default) and you request 8 processors, the number of subjobs will only be 2, and the number of processors will also be reduced to 2.

13.4 Granting Access to a Database

When you create a database, you are the owner of all the files and directories associated with that database. As such, you will normally have read and write permissions to the files, and read, write, and execute permissions to the directories. You should therefore be able to modify or search a database that you create yourself. However, you may or may not want other users to be granted those same privileges.

If you want to be certain that you are the only person who can ever modify or search a particular database, then you should remove write permissions for all other users throughout the database tree, and remove read permission from the database. By default, the files and directories you create do normally not carry write permissions for other users, unless you change the default `umask`.

If you want to allow other users to search the database, but not to modify it, you must grant those users read and execute permissions to all directories throughout the database tree. The same read and execute permissions must also exist upward from the database directory to the file system mount point.

To remove all permissions for other users:

```
chmod -R g-wx,o-wx dbName.phdb
```

To grant permission to search but not to modify the database:

```
chmod g+rx dbName.phdb dbName.phdb/database_ligands
    dbName/database_ligands/block*
```

This command gives permissions only to members of your group, and assumes that they have execute permission to all directories above the database, but do not have write permission to the database. To grant permission to everyone, you could add the `o+rx` permission code.

13.5 Database Structure

The database is stored in a directory, *dbName*.phdb, which contains all the files associated with the database. The files and directories that are standard parts of the database are described in Table 13.2. You should not generally need to modify these files, but you may find it useful to examine them. In addition to these files, the database can contain subset files, named *subset*_phase.inp.

You should not create or store other databases inside the database directory, as this is likely to lead to job failure (The phase_multi_database program is an exception, as the databases it creates are only temporary).

Table 13.2. Database files and directories

File or Directory	Description
database.sqlite	An SQLite database with tables that hold global information about the database: dprop_name_table—Holds the name of the property used to detect and eliminate duplicates, s_phase_Unique_SMILES. dprop_values_table—Holds mol_id and s_phase_Unique_SMILES for each database record. props_table—Holds mol_id, title, num_confs, has_sites, plus all imported and computed properties for each database record. This table does not exist unless an "extract" job has been run, and it must be regenerated when changes to the database are made. props_map_table—Mappings of the column names in props_table to the CT property names in the block database files stored in database_ligands/. This table always exists, but is empty if properties have not been extracted. summary_table —Holds mol_id, title, num_confs, and has_sites for each database record. has_sites is 1 or 0, depending on whether or not sites are stored for a given record.
database_dbversion	Version file. Holds the Phase version number, the method of creation (always "CL"), and the storage format ("SQLite").
database_feature.ini	Feature definitions.
database_history.log	History of changes to the database, containing the date and the command issued.
database_info.log	Detailed information about changes to the database.
database_master_phase.inp	Master subset file.
database_props.csv	Copy of the data in props_table in CSV format. This file is only present if an extract job has been run.

Table 13.2. Database files and directories (Continued)

File or Directory	Description
database_props_map.csv	Copy of the mappings in props_map_table in CSV format. This file is only present if an extract job has been run with -map.
database_props_stats.csv	Statistics for the properties in props_table: count, min, max avg. count is the number of records for which the property exists. For string properties, avg is an empty value. This file is only present if an extract job has been run with -stats.
database_summary.csv	Copy of the data in summary_table in CSV format.
database_ligands	All block data is held under this directory. There is one SQLite file for each block of 5000 records, e.g., database_ligands/block_1/block_struct_1.sqlite. Structures, sites, CT-level properties, and 2D/3D indexes are stored in this file. When an extract job is run, the properties in these block files are extracted and written to the top-level file database.sqlite.
database_restart	This directory contains files that are required to restart a phase_database job. The files in this directory are updated as a job progresses. When the job and all of its subjobs finish normally, this directory is removed automatically.

13.6 Searching for Matches with phase_find_matches

The phase_find_matches program can be used to search one or more structure files, a Phase database, or a Phase command-line project, for matches to a pharmacophore hypothesis. The syntax is:

$SCHRODINGER/phase_find_matches *source hypoID jobname* [*options*]

source is the source of structures that will be searched, and must be one of the following:

- Maestro file (.mae, .mae.gz, .maegz).
- SD file (.sdf, .sd, .sdfgz, .sdf.gz, .sd.gz).
- List file (.list). This is a text file that contains the names of one or more Maestro or SD files, or one or more Phase databases, with one name per line. Mixing of Phase databases and structure files is not supported.
- Database created using phase_database (.phdb). Must include the absolute path. The database need not exist on the local machine, but it must be accessible to the host where the job is run.

- Ligand records file (`.inp`) that resides in a Phase command-line project. Must include the absolute path. The project need not exist on the local machine, but it must be accessible to the host where the job will run.

hypoID is the prefix for hypothesis files. At a minimum, the files *hypoID*`.xyz` and *hypoID*`.def` must be present. To use a reference ligand, the files *hypoID*`.mae` and *hypoID*`.tab` must also be present. If any other hypothesis files are present, they are used by default. You can disable their use with one of the `-no*` options.

The output is written to a Maestro file, *jobname*`-hits.maegz` by default; the `-osd` option writes an SD file instead, *jobname*`-hits.sdfgz`.

The options can be listed by running the command with the `-h` option. The syntax of the options that have dependencies or are mutually exclusive is as follows:

```
[[-distinct] | [-connect] [-stereo]]
[{-flex|-refine} [-sample rapid|thorough] [-max numConfs]
    [-bf numPerBond] [-ewin deltaE] [-amide vary|orig|trans]
    [-skip maxRot] [-append]]
[-nosort | -keep maxHits]
[-sites | -noindex]
-CHECKPOINT | -RESTART path | -NOCHECKPOINT
```

The standard Job Control options, which allow you to specify the host, set the priority, and so on, are accepted. These are described in Table 2.1 of the *Job Control Guide*. The `-LOCAL` and `-NOJOBID` options, described in Table 2.2 of the *Job Control Guide*, are also accepted.

Jobs can be distributed across multiple processors (with the `-HOST` option) for any input type. If searching files with pregenerated conformers, the number of processors must not exceed the number of files. If screening one or more databases, the number of processors cannot be larger than the number of records in the smallest database. In either of these cases, the number is reduced automatically if it exceeds the maximum allowed value.

Failed jobs can be restarted with the `-RESTART` option if the `-CHECKPOINT` option is used with the initial run. See the descriptions below for these options.

Searching for Molecules by Shape

There are occasions on which the shape of the molecule is its most important feature, and a search for molecules that are most similar in shape is needed. Phase provides this capability with the `shape_screen` program.

The `shape_screen` program can be used to screen one or more files or a Phase database against a shape query. Each conformer from a given molecule is aligned to the query in various ways, and a similarity is computed based on overlapping hard-sphere volumes. The conformer and alignment yielding the highest similarity for each molecule is written to a Maestro file, along with the similarity property `r_phase_Shape_Sim`, which appears in the Project Table as Shape Sim.

The shape query can be a single template molecule, or it can be a set of three or more spheres. The latter form provides a high degree of flexibility in designing the shape. For each molecule searched, `shape_screen` returns an aligned structure that provides the best overlap with the shape query. If you are searching a set of homologous structures and the shape query is a member of the same series, then `shape_screen` should return the most sensible alignment among all those available. If you are searching against structures that are not necessarily related to the shape query, `shape_screen` should return something that looks most like the query in an overall sense.

The shape search can treat all atoms as equivalent, or it can incorporate information on atom types as part of the search. Searching on atom types favors alignments that superimpose atoms of the same type. There are three possibilities for atom typing in a shape search: use of MacroModel types, typing by element, and use of Phase QSAR types. MacroModel atom types will impose the most stringent conditions on the matching, and Phase QSAR types will impose the most general conditions.

As an alternative, the shape search can treat each structure as a collection of pharmacophore sites, whose locations are assigned by applying Phase pharmacophore feature definitions. Overlapping volumes are computed only between sites of the same type.

Volume scoring as part of a search for matches to a hypothesis also uses molecular shape, but there are some important differences between volume scoring in a search for matches and shape-based queries. In the search for matches, the molecules are aligned to the hypothesis, and the volume overlap is then computed on the basis of that alignment. The molecular shape alignment might not be optimal in this case. Shape queries investigate a much greater variety of alignments than a typical search for matches to a hypothesis, and compute volume overlaps

that are closer to optimal. A second difference is in the algorithm used to evaluate the volume scores in the search for matches on the one hand, and the shape similarities on the other. However, the shape similarities are a form of volume scoring, and this term is used below for the shape similarities.

The search by shape can be set up and run from Maestro or run from the command line.

14.1 Running Shape Searches from Maestro

To run a shape search from Maestro, you use the Shape Screening panel, which you open by choosing one of the following:

- Applications → Shape Screening
- Tasks → Shape Screening

You can also open this panel from the Elements profile (Tasks → Ligand Tasks → Shape-Based Similarity Screening) and the BioLuminate profile (Tasks → Ligand Tasks → Shape Screening). The main panel contains the most commonly used options, but further options are available in a dialog box.

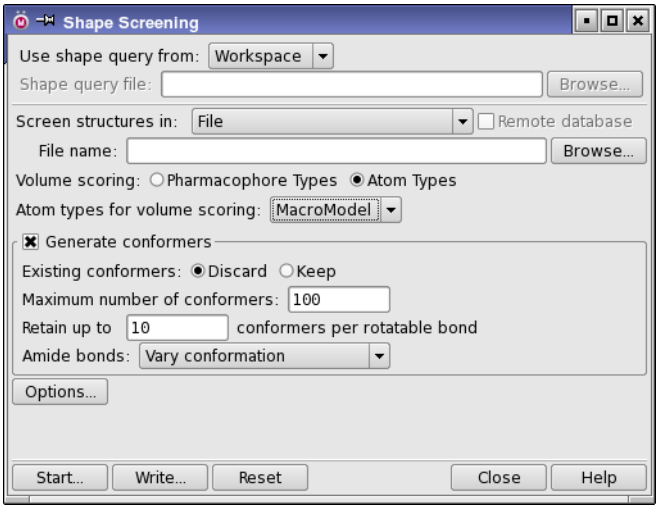

Figure 14.1. The Shape Screening panel.

To set up a default search by shape:

1. Choose the source of the shape query from the Use shape query from option menu.

   The options are Workspace and File.

   If you selected File, enter the file name in the Shape query file text box or click Browse to navigate to the file. You can read Maestro files, (`.mae`), SD files (`.sdf`), or Phase included volume files (`.ivol`).

   If you selected Workspace and there are multiple entries included in the Workspace, a query is constructed for each entry.

2. Choose the source of the molecules to screen from the Screen structures in option menu.

   The options are File, Phase database, and Project Table (selected entries). If you selected either of the first two, enter the file name or database name in the File name text box, or click Browse to browse to the file or database. If the database is not accessible on the local host, select Remote database.

   The file can be a Maestro file or an SD file, and can be compressed or uncompressed.

   If you want to limit the number of conformers generated per input structure, enter a value in the Maximum number of conformers text box, or enter a value in the Retain up to *N* conformers per rotatable bond text box, or both.

3. Choose a volume scoring option.

   - Pharmacophore types—Treat each structure as a collection of pharmacophore sites, whose locations are assigned by applying Phase pharmacophore feature definitions. Overlapping volumes are computed only between sites of the same feature type. Each feature is represented by a sphere of radius 2 Å.
   - Atom types—Treat each structure as a collection of atoms, whose volumes are defined by their van der Waals radii. Overlapping volumes are computed only between atoms of the same type.

   If you chose Atom types, you must also choose a category of atom types that you can use in volume scoring. The choices are:

   - None—Do not distinguish different types of atoms when calculating volume overlaps: all atoms are treated the same. This choice gives pure volume scoring.
   - MacroModel—Calculate volume overlaps only between atoms that have the same MacroModel atom type.
   - Elements—Calculate volume overlaps only between atoms of the same element.
   - QSAR—Calculate volume overlaps between atoms that have the same pharmacophore type (Acceptor, Donor, etc.) as defined for Phase QSAR models.

4. Choose whether or not to generate conformers, using the Generate conformers option.

   This option is off by default if you are screening a Phase database, which usually contains conformers. You can turn it on if the database does not contain conformers or if you want to generate conformers anyway.

   If the structures to be searched already include conformer sets or if you want to keep the input structure rather than only use it as a seed for the conformational search, you can choose to append the generated conformers by selecting Keep for Existing conformers. Otherwise, choose Discard.

   If you are generating conformers, choose an amide bond treatment from the Amide bonds option menu. The choices are:

   - Vary conformation—allow the conformation around the amide bond to vary freely.
   - Retain original conformation—do not vary the conformation around the amide bond, but keep the original conformation.
   - Set to trans—Set the conformation around the amide bond to trans (180°).

5. Click Start to set job options and start the job.

   If you are searching a Phase database, or if you are generating conformers when searching a file, you can distribute the job over multiple CPUs.

When the job finishes, the results are appended to the Project Table if you chose to incorporate them, with the property Phase Sim added.

If you want more control over the screening process and parameters, click Options, and make settings in the Shape Screening - Options dialog box.

- To generate more than one alignment of a given molecule to a given query, enter the number of alignments in the Write out at most *N* alignments per ligand text box. By default, only one alignment is returned.
- To apply excluded volumes in a search, select Use excluded volumes, and enter the name of an excluded volumes (`.xvol` or `.ev`) file in the text box, or click Browse to browse to the file.
- If you want polar hydrogens (which are almost always donors) to be considered in the shape search, select Include polar hydrogens. This option is not available if you chose Pharmacophore types for volume scoring in the main panel.
- To filter out conformers whose similarity to the query is less that a certain value, select Filter out conformers with similarity below and enter a value in the text box. The value must be in the range 0–1.

  If you are searching a Phase database and want to create a subset file based on the filter, select Create Phase subset file from hits.

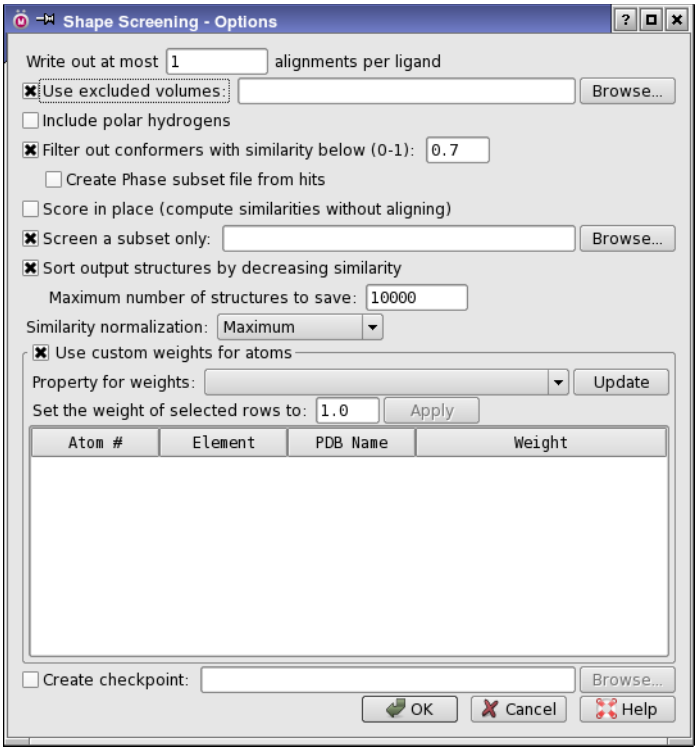

Figure 14.2. The Shape Screening - Options dialog box.

- To compute similarities without aligning the molecules, select Score in place.
- To sort the results by similarity, select Sort output structures by decreasing similarity, and enter a limit on the number of structures in the Maximum number of structures to save text box.
- To choose the normalization scheme used for calculating similarity, choose from the items on the Similarity normalization option menu. The options are Maximum, Minimum, Query structure, and Screen structure. These correspond to the values of the `-norm` argument to `shape_screen`. The default is Maximum.
- To weight particular atoms more than others in the queries, select Use custom weights for atoms, choose an atom-level property from the Property for weights option menu, and click Update. This option and its controls are not available if you chose Pharmacophore types for volume scoring in the main panel.

  The table is filled in with the list of atoms in the query with the defined weight. The weights are used to scale the atomic volumes used to compute the overlaps. The values

must be between 0 and 1. You should choose or create a property that scales down the volume of the less-important atoms.

You can change the weights by editing the table cells. To assign a weight to multiple cells, select the table rows, enter the value in the Set the weight of selected rows to text box, and click Apply.

- To create a checkpoint file for the job (so that it can be restarted if it fails), select Create Checkpoint, and enter the directory for the checkpoint file in the text box, or click Browse and navigate to the desired directory.

If you want to run the job from the command line, you can write out the structure input files by clicking Write. The command that corresponds to the settings in the panel is also displayed, so that you can copy and paste it.

14.2 Running Shape Searches from the Command Line

The command syntax for a shape search is as follows:

```
$SCHRODINGER/shape_screen -JOB jobname
    {CHECKPOINT|-RESTART path|-NO_CHECKPOINT}
    [program-options] [job-options]
```

The main job options are the same as for other Phase programs, and are described in Table 2.1 of the *Job Control Guide*. The program options, other job options, and the required arguments can be listed by running the command with the `-h` option. Shape screens can be run on both the current and the new database format. When you run the job, you must specify whether you want to create checkpoint files or not. If you do not create checkpoint files, the job cannot be restarted. To restart a job, you can use the syntax:

```
$SCHRODINGER/shape_screen -JOB jobname -RESTART path [-osub dbSubOut]
    [job-options]
```

The aligned structures are written to the file *jobname*`_align.maegz`. For each molecule, the alignments for a given query are stored consecutively, ordered by similarity. If sorting is requested, the blocks of alignments for each query are ordered by similarity, and the molecules are ordered by the maximum similarity for any query or alignment for that molecule. If sorting is not requested, the blocks of alignments for each query are in the order that the queries appear in the query input, and the molecules are in the same order as the molecules in the input.

To set up the file required for a SMARTS prealignment, you can display the query in Maestro, label the atoms with the atom numbers, then use the Find toolbar to find the desired SMARTS patterns. You can then copy the string and the atom numbers to the file. The atom numbers must be in the same order as the atoms in the SMARTS string.

14.3 Creating Included Volumes for Shape Queries

The simplest way to create an included volume file that can be used as a shape query is to rename an `.xvol` excluded volume file to have the extension `.ivol`. To create an excluded volume, use the Hypothesis Table panel in Maestro to create a hypothesis, add the desired included volumes as excluded volumes to this hypothesis, then export the hypothesis. You can then take the exported `.xvol` file and rename it for use with `shape_screen`.

You can also use following simplified format (which is the same format as the `.ev` excluded volume file):

NumSpheres
x1 y1 z1 r1
x2 y2 z2 r2
. . .

where *NumSpheres* is the number of spheres, and the *x*, *y*, *z*, and *r* values on any line are the coordinates of a sphere center and its radius.

A utility is available for creating included volumes, `create_ivolShape`, which is described in the next subsection. For converting included volumes files to Maestro files, you can use the utility `convert_ivolToMae`, which is described in the following subsection.

14.3.1 create_ivolShape

This utility creates an included volumes file to represent the "positive image" of a ligand or the "negative image" of a receptor, in terms of a set of spheres. For a receptor, you must define the binding pocket using a structure file that contains one or more ligands, or a box file that contains the limits of the box.

The syntax is as follows:

```
$SCHRODINGER/utilities/create_ivolShape -in maeFile
    {-pos positive-options | -neg negative-options}
    -out ivolFile [-append [-avoid dmin]]
```

The options are described in the usage message, which you can view by running the command with the `-h` option. You can only create a positive image or a negative image in a single run, but you can combine a positive image and a negative image with the `-append` option.

14.3.2 convert_ivolToMae

This utility converts an included volumes file to a structure in a Maestro file. This allows included volumes to be imported and visualized in Maestro without having to associate them with a Phase hypothesis. It also allows multiple sets of included volumes to be stored in a single Maestro file and supplied to `shape_screen` as multiple shape queries.

The utility works by creating a carbon atom for each included volume sphere. The positions of the spheres will be correct, but the radii will always be 1.7 angstroms (the van der Waals radius of carbon). If you use the Maestro file as a shape query, you should run `shape_screen` with the option `-atomWeights r_m_shape_weight` to scale the carbon van der Waals radii to the values that were present in the original included volumes file.

The command syntax is as follows:

```
$SCHRODINGER/utilities/convert_ivolToMae -in ivolFile -out maeFile
    [-append] [-title title]
```

The options can be viewed by running the command with the `-h` option.

14.4 Creating Consensus Shape Queries

Shape searches on multiple aligned structures do not work well because of the number of small interatomic distances. A more robust approach is to create a *consensus* shape, where closely superimposed atoms are replaced by a single representative sphere.

Consensus shapes can be created with the utility `create_shapeConsensus`. This utility performs hierarchical clustering with complete linkage on the heavy atoms of the input structures (or including polar hydrogens). By default, the hierarchical clustering merges atoms into clusters until it reaches the point where further merging would place two atoms from the same structure in the same cluster. Each cluster is then replaced by a single sphere whose location and radius can be set. This procedure automatically places atoms that are closely superimposed into the same cluster, and creates singleton clusters for atoms that have no counterpart in any of the other structures. The resulting shape is then essentially a super-structure of everything in the original file, with normal bond-length-like spacing between the closest pairs of spheres.

The syntax of the `create_shapeConsensus` command is:

```
create_shapeConsensus [options] inFile outFile
```

The input file can be in Maestro (`.mae`, `.maegz`, `.mae.gz`) or SD (`.sdf`, `.sdf.gz`, `.sdfgz`) format, and must contain two or more structures. The output file can be a included volumes (`.ivol`) or a Maestro (`.mae`) file for the consensus shape. If the output file is a Maestro file,

each sphere is represented by an unconnected carbon with two atom-level properties that relate the van der Waals radius of carbon to the sphere radius:

`r_m_sphere_scale`	The factor by which the carbon atom radius should be multiplied to obtain the radius of the consensus sphere.
`r_m_shape_weight`	The atom weight property to supply when using *outFile* as the shape query. The option to use when running `shape_screen` is `-atomWeights r_m_shape_weight`.

The options can be viewed by running the command with the `-h` option.

Phase QSAR Models

A.1 The Phase QSAR Methods

Phase QSAR models are developed from a series of molecules, of varying activity, that have all been aligned to a common pharmacophore hypothesis that is associated with a single reference ligand. QSAR models may be atom-based or pharmacophore-based: the former takes all atoms into account; the latter uses the pharmacophore sites that can be matched to the hypothesis.

A rectangular grid is defined to encompass the space occupied by a training set of aligned molecules. This grid divides space into uniformly-sized cubes, typically one angstrom on each side. In atom-based models, the grid is populated by van der Waals spheres, with radii that depend on the atom type. In pharmacophore-based models, the grid is populated by the pharmacophore sites that match the hypothesis, with each site represented by a sphere with a user-definable radius.

A given atom or pharmacophore site will occupy the space of one or more cubes in the grid. Occupation of a cube is deemed to occur if the center of that cube falls within the radius of the atom or site. A given cube may be occupied by more than one atom or site, and that occupation may come from the same molecule or from different molecules.

Each occupied cube gives rise to one or more volume bits. A volume bit is allocated for each different class of atom or site that occupies a cube.

In pharmacophore-based models, sites are assigned to classes that are determined by the feature definitions used to create the hypothesis (e.g., A, D, H, N, P, R). In atom-based models, there are 6 distinct atom classes that have some correspondence or similarity with pharmacophore feature types, but atom classes are assigned using fixed internal rules, not the hypothesis feature definitions:

- D – Hydrogen bond donor (hydrogens bonded to N, O, P, S)
- H – Hydrophobic/non-polar (C, H–C, Cl, Br, F, I)
- N – Negative ionic (formal negative charge)
- P – Positive ionic (formal positive charge)
- W – Electron-withdrawing (N, O)
- X – Miscellaneous (all other types of atoms)

So, if a particular cube is occupied by a “D” from molecule 1, an “H” from molecule 5, and a “P” from molecule 9, that cube would be allocated three volume bits. If a cube is never occu-

pied by any molecule in the training set, no volume bits would be allocated. Hence, each volume bit must be set by at least one molecule in the training set.

The pool of volume bits provides a means of characterizing the molecules. In atom-based models, the pattern of volume bits that are set by a molecule encodes the size, shape, and chemical characteristics of that molecule. In pharmacophore-based models, the pattern of set bits determines which subset of critical pharmacophore features that molecule contains, and the positions of those features in relation to other molecules.

If a binary scheme (0/1) is used to denote which bits are set by each molecule, a table of bit values may be assembled:

$molecule_1$	0	1	1	0	0	1	1	0	0	1	0	1	0	.	.	.
$molecule_2$	0	1	1	0	0	0	1	0	0	1	0	1	0	.	.	.
$molecule_3$	0	0	0	1	0	1	0	0	1	0	1	0	1	.	.	.
	.															
	.															
	.															

For atom-based models, there are usually several hundred or more volume bits for each series of aligned molecules. For pharmacophore-based models, that number is much smaller, usually only a few dozen. The number of bits increases as the grid spacing becomes finer, and, in the case of atom-based models, as the molecules become larger.

To generate a QSAR model, the 0/1 bit values are treated as independent variables in partial least-squares (PLS) regression analysis. This involves finding a linear least-squares relationship between the activity data and a special set of orthogonal factors that are linear combinations of the bit value variables.

More precisely, if there are n molecules in the training set and v volume bits, let the $n \times v$ matrix $\mathbf{X}$ represent the table of volume bits, and let the $n \times 1$ vector $\mathbf{y}$ represent the activity values for the training set molecules. Creation of the PLS regression model proceeds as follows:

Center each column of $\mathbf{X}$:

for $i = 1,\ldots,v$

$$\mu_i^x = \frac{1}{n} \sum_{k=1}^{n} X_{k,i}$$

for $k=1,\ldots,n$

$$X_{k,i} \rightarrow X_{k,i} - \mu_i^x$$

next k

next i

Center $\mathbf{y}$:

$$\mu^y = \frac{1}{n}\sum_{k=1}^{n} y_k$$

for k=1,…,n

$y_k \rightarrow y_k - \mu^y$

next k

Determine PLS factors and regression coefficients for up to m PLS factors ($m \le v$):

$\mathbf{X}_1 = \mathbf{X}$

for i = 1,…,m

Compute the vector of weights that define PLS factor i:

$$\mathbf{w}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{y} \Big/ \left|\mathbf{X}_i^{\mathrm{T}}\mathbf{y}\right|, \qquad \mathbf{w}_i \in \mathbf{R}^{v\times 1}$$

Project the rows of $\mathbf{X}_i$ onto factor i:

$$\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i, \qquad \mathbf{t}_i \in \mathbf{R}^{n\times 1}$$

Project $\mathbf{t}_i$ onto each column of $\mathbf{X}_i$:

$$\mathbf{p}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{t}_i \Big/ \left|\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i\right|, \qquad \mathbf{p}_i \in \mathbf{R}^{v\times 1}$$

Compute the ith PLS regression coefficient by projecting $\mathbf{t}_i$ onto $\mathbf{y}$:

$$\mathbf{b}_i = \mathbf{t}_i^{\mathrm{T}}\mathbf{y} \Big/ \left|\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i\right|, \qquad \mathbf{b}_i \in \mathbf{R}^{m\times 1}$$

Orthogonalize $\mathbf{X}_i$ with respect to PLS factor i:

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i\,\mathbf{p}_i^{\mathrm{T}}$$

next i

For a regression with m PLS factors, the fitted activities are then given by:

$$\hat{\mathbf{y}} = \mu^y + \sum_{i=1}^{m} b_i \mathbf{t}_i$$

To apply the m-factor PLS model to a new set of n_T ligands with bit value matrix $\tilde{\mathbf{X}}$, the regression coefficients $\mathbf{b}$ must first be translated back to the space of the original $\mathbf{X}$ variables:

Define

$$\mathbf{W} \equiv [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m], \qquad \mathbf{W} \in \mathbf{R}^{v \times m}$$

$$\mathbf{P} \equiv [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m], \qquad \mathbf{P} \in \mathbf{R}^{v \times m}$$

Then

$$\mathbf{b}^x \equiv \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{b} \qquad \mathbf{b}^x \in \mathbf{R}^{v \times 1}$$

The coefficients $\mathbf{b}^x$ may then be used to predict activities for the new ligands:

$$\hat{y}_k = \mu^y + \sum_{i=1}^{m} (X_{k,i} - \mu_i^x)\, b_i^x \qquad k = 1, \ldots, n_T$$

A.2 Phase Model Validation

Phase QSAR models do not use internal cross-validation techniques, but rather use distinct training and test sets. The use of a true test set is far superior to internal cross-validation techniques such as leave-n-out, where small subsets of the training set are temporarily held out and predicted using models built from the remainder of the training set. Leave-n-out, as it is usually applied, is not an unbiased validation technique because the activity data being predicted typically have some role in selecting or constructing the variables used in the series of models being built. This is especially true of partial least-squares (PLS) regression, the multivariate method that is used to develop Phase QSAR models. Statistics from PLS leave-n-out predictions will almost always be overly optimistic, because the latent variables included in each model are constructed from the full training set, so they correlate with all activities, even those of the molecules being predicted. Further, leave-n-out predictions are frequently used to arrive at an *optimal* number of PLS factors, but in fact, internal cross-validated statistics cannot provide a meaningful measure of how the model will actually perform when applied to new

molecules. As a result, the optimal number of PLS factors arrived at using this technique may very well correspond to a situation wherein the activity data have been seriously over-fit. For these reasons, Phase supports only the use of true, external test sets.

However, the use of leave-*n*-out techniques are useful for assessing the stability of the model to changes in the training set. In Phase QSAR models, leave-*n*-out models are built, and the R^2 value is computed between the leave-*n*-out predictions and the predictions from the model built on the full training set. This value is reported as the stability value, and has a maximum value of 1. If the stability value is high, the model built from the full training set is fairly insensitive to changes in that training set, i.e., the predicted values don't change much. Models with high stability are preferred because they are not overly dependent on the idiosyncracies of any particular training set. Stability should not be used to decide on an optimum number of PLS factors, but it can be helpful in choosing between models from different hypotheses whose other statistics are essentially the same.

A.3 Phase QSAR Statistics

This section defines the various statistical measures that are used in Phase QSAR models.

A.3.1 Training Set and Model

Statistical quantities describing the training set and the QSAR model are defined below.

m	number of PLS factors in the model
n	number of molecules in the training set
$df_1 = m + 1$	degrees of freedom in model
$df_2 = n - m - 2$	degrees of freedom in data
y_i	observed activity for training set molecule i
$\hat{y}_i$	predicted activity for training set molecule i
$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	mean observed activity
$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$	variance in observed activities

$sse = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ sum of squared errors

$\sigma_{\text{err}}^2 = \frac{sse}{n}$ variance in errors

$ssy = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ variance in model

$SD = \sqrt{sse / df_2}$ standard deviation of regression

$R^2 = 1 - \frac{\sigma_{\text{err}}^2}{\sigma_y^2}$ R-squared; coefficient of determination

$F = \frac{ssy / df_1}{sse / df_2}$ F statistic; overall significance of model

$P = B(df_1, df_2, \frac{df_2}{df_2 + F df_1})$ statistical significance; probability that correlation could occur by chance. The beta function $B(a, b, x)$ is defined by $B(a, b, x) = \int_0^x t^{a-1}(1-t)^{b-1} dt$

Note that R^2 can never be negative, because the regression coefficients are optimized to minimize *sse*. The worst-case scenario is when the independent variables have absolutely no statistical relationship with activity. Under those circumstances, the regression coefficients will all be zero, and the model will contain only an intercept parameter, the value of which will be $\bar{y}$. Thus every predicted activity will be $\bar{y}$, and σ_{err}^2 will be equal to σ_y, yielding $R^2 = 0$.

A.3.2 Test Set Predictions

Statistical quantities describing the test set predictions are described below.

T The test set of molecules

n_T number of molecules in T

y_j observed activity for molecule $j \in T$

$\hat{y}_j$ predicted activity for molecule $j \in T$

$RMSE = \sqrt{\frac{1}{n_T} \sum_{i \in T} (\hat{y}_j - y_j)^2}$ root-mean-squared error

$$Q^2 = R^2(T)$$ Q-squared

$$r = \frac{\sum_{j \in T} (y_j - \bar{y}_T)(y_j - \hat{\bar{y}}_T)}{\sqrt{\sum_{j \in T} (y_j - \bar{y}_T)^2 (y_j - \hat{\bar{y}}_T)^2}}$$ Pearson *r* value, Pearson correlation coefficient

The formulas for R^2 and Q^2 are equivalent, with the only difference being that Q^2 is computed using the observed and predicted activities for the test set. However, Q^2 can take on negative values. This happens whenever the variance in the test set errors is larger than the variance in the observed test set activities. Often, the test set does not have as large a range of activity values as the training set (so the variance in *y* is smaller), and the errors for the test set tend to be larger than those for the training set (so the variance in the errors is larger), It is therefore not uncommon to see negative Q^2 values from time to time.

Because all values are shifted by the sample means, the Pearson correlation coefficient is insensitive to systematic errors in the predictions, whereas Q^2 is not. So if the rank order of the activity predictions is basically correct, but there's a significant constant shift in the values compared to the observed activities, the Pearson correlation coefficient may still be quite high, even if Q^2 is small or negative.

Phase Input Files

In addition to structure files, Phase uses a variety of data input files. These files are described in the following sections. Normally you would not need to edit most of these files, as they are set up using the command-line utilities or from Maestro. However, some of the files used when searching for matches must be created by hand.

In this appendix, references to *utilities* are to programs or scripts in the `$SCHRODINGER/utilities` directory.

B.1 Master Data File

This file stores various pieces of ligand data that are used throughout the pharmacophore model development workflow. It is named `MasterData.tab`, and can be updated by the utility `pharm_data` or by hand. At the top of the file is a description of the data it contains, and a number of rules regarding how the data may be modified. The body of the file contains first a ligand property name block, followed by a set of ligand blocks, one per ligand in the project.

If you make any changes to `MasterData.tab`, whether by hand or through operations supported by the `pharm_data` utility, you must run `pharm_data` with the `-commit` option to update the Maestro files in the `ligands` subdirectory. If you do not update the Maestro files, various Phase modules will not be using the modified values because they read the property data directly from the Maestro files.

If you have completed any forward steps in the project workflow, the results generated in those steps may be invalidated by changes you make to `MasterData.tab`. When you attempt to commit the changes, you will be supplied with a list of forward files that will be invalidated, and you will be given a chance to abort the commit operation. If you choose to abort, you can rerun `pharm_data` with the `-restore` flag to revert to the previous version of the file, which is stored in `MasterData.backup`.

The ligand property name block keywords are described in Table B.1. These properties are set by `pharm_project` and must not be altered by hand. This block defines the names of certain properties that are relevant to the pharmacophore model development. The Maestro files in the `ligands` subdirectory contain these named properties.

Table B.1. Ligand property block keywords in the Master Data file.

Keyword	Description
LIGAND_NAME_PROPERTY	Name of Maestro property that defines the ligand name. Set automatically by pharm_project to s_phase_Ligand_Name.
PHARM_SET_PROPERTY	Name of Maestro property that defines the pharm set membership. Set automatically by pharm_project to s_phase_Pharm_Set.
LIGAND_GROUP_PROPERTY	Name of Maestro property that defines the ligand group membership. Set automatically by pharm_project to i_phase_Ligand_Group.
MUST_MATCH_PROPERTY	Name of Maestro property that records whether the ligand must match the pharmacophore. Set automatically by pharm_project to b_phase_Must_Match.
QSAR_SET_PROPERTY	Name of Maestro property that defines the QSAR set membership. Set automatically by pharm_project to s_phase_QSAR_Set.
ACT_PROPERTY	Name of Maestro property that stores the ligand activity values. Set by the -act option of pharm_project.
1D_PROPERTY	Name of Maestro property that defines the 1D_VALUE used to control which ligands can give rise to hypotheses. Set automatically by pharm_project to r_phase_Ligand_1D_Property.
CONF_PROPERTY	Name of Maestro property that defines the conformation-dependent quantity used in scoring. This is normally the relative conformational energy. Set by the -conf option of pharm_project.

The keywords that define the ligand data values in the ligand blocks are described in Table B.2. These data values are stored in the ligand Maestro files in addition to the master data file. Only the conformation-independent properties are kept in these blocks: the conformation-dependent property defined by CONF_PROPERTY is not stored here, but only in the ligand Maestro file.

Table B.2. Ligand block keywords in the Master Data file.

Keyword	Description
LIGAND_NAME	Ligand name, in the project. This is not the same as the title or the Maestro entry name, but is a unique identifier used in the pharmacophore model project. Do not modify.
TITLE	Title of the ligand. Taken from the Title property in the Maestro file. Do not modify.

Table B.2. Ligand block keywords in the Master Data file. (Continued)

Keyword	Description
PHARM_SET	Pharm set membership of the ligand. Can be modified by hand or using the -active and -inactive options of the pharm_data utility. Allowed values are active, inactive, and none. active Ligand is used to identify common pharmacophores and to score hypotheses. There must be at least two ligands with PHARM_SET = active inactive Ligand is used to measure the degree to which hypotheses discriminate actives from inactives by inactive scoring. none Ligand is not used in pharmacophore model development, but may be used in QSAR model development.
QSAR_SET	QSAR set membership of the ligand. Affected by the -train, -rand, and -pharm_set options of the pharm_data utility. Can be modified by hand. Allowed values are train, test, and none. train Ligand is used to develop QSAR models. You should use at least five training set ligands for each PLS factor. test Ligand is used to test QSAR models. none QSAR models are not applied to this ligand.
ACTIVITY	Ligand activity. Affected by the -log, -exp, and -multiply options of the pharm_data utility. Can be modified by hand. Values should increase as potency increases, as for example in –log K_i or –log IC50. If activity is unknown, the value should be set to missing.
1D_VALUE	A conformationally independent numerical property that may be used during hypothesis scoring to influence or control the selection of reference ligands. This property is added to the actives score when the -prop option is used with pharm_score_actives. If you assign a nonzero 1D_VALUE for certain actives, and a zero value for the remaining actives, you can force hypotheses to come from only those actives with a nonzero 1D_VALUE. Must be set by hand.
LIGAND_GROUP	Numerical identifier (group number) of ligand group. Ligands that have the same ligand group number belong to the same group. Ligands in the same group are treated as equivalent when finding common pharmacophores: to match the pharmacophore, only one member of the group has to match. Defining groups is useful for making tautomers, stereoisomers, and ions equivalent. By default, all ligands are in a separate group.
MUST_MATCH	Require this ligand to match when finding common pharmacophores. Allowed values are true and false. This property can be set to true only when PHARM_SET is active. All ligands in the same group must have the same value of this property. Default: false.

An example of the top of a Phase Master Data file is given below. This excerpt includes the header, the ligand property name block, and a few ligand blocks.

```
###############################################################################
#                                                                             #
# Phase Master Data File                                                      #
#                                                                             #
# You may change PHARM_SET, LIGAND_GROUP, MUST_MATCH, QSAR_SET, ACTIVITY and  #
# 1D_VALUE. To propagate changes to the project, use 'pharm_data -commit'.    #
# To revert to the most recently committed version of MasterData.tab, use     #
# 'pharm_data -restore'.                                                      #
#                                                                             #
# PHARM_SET:    Allowed values are "active", "inactive" and "none             #
#               active   - Used to identify common pharmacophores and to score #
#                          hypotheses.  There must be at least two ligands and #
#                          two ligand groups with PHARM_SET = active.         #
#               inactive - Used to measure the degree to which hypotheses     #
#                          discriminate actives from inactives.               #
#               none     - Not used in pharmacophore model development.       #
#                                                                             #
# LIGAND_GROUP: Allowed values are 0, 1, 2, etc. Ligands assigned to the same  #
#               group will be treated as interchangeable for purposes of      #
#               finding common pharmacophores.                                #
#                                                                             #
# MUST_MATCH:   Allowed values are "true" and "false". If "true", all common  #
#               pharmacophores will be required to match the ligand, or at    #
#               least one ligand from the group to which it is assigned.      #
#                                                                             #
# QSAR_SET:     Allowed values are "train", "test" and "none".                #
#               train - Used to develop QSAR models. Recommend at least five  #
#                       training set ligands for each PLS factor.             #
#               test  - Used to test QSAR models.                             #
#               none  - QSAR models not applied to these ligands.             #
#                                                                             #
# ACTIVITY:     Ligand activity.  Values should increase with potency, for    #
#               example, -logKi or -logIC50. If activity is unknown, the      #
#               value should be "missing".                                    #
#                                                                             #
# 1D_VALUE:     A conformationally-independent numerical property that may be #
#               used during hypothesis scoring to influence or control the    #
#               selection of reference ligands.                               #
#                                                                             #
###############################################################################
LIGAND_NAME_PROPERTY = s_phase_Ligand_Name
PHARM_SET_PROPERTY = s_phase_Pharm_Set
LIGAND_GROUP_PROPERTY = i_phase_Ligand_Group
MUST_MATCH_PROPERTY = b_phase_Must_Match
QSAR_SET_PROPERTY = s_phase_QSAR_Set
ACT_PROPERTY = r_phase_Ligand_Activity
```

```
1D_PROPERTY = r_phase_Ligand_1D_Property
CONF_PROPERTY = r_mmod_Relative_Potential_Energy-MMFF94s
###############################################################################
LIGAND_NAME = mol_1
TITLE = "endo-1"
PHARM_SET = active
LIGAND_GROUP = 1
MUST_MATCH = false
QSAR_SET = train
ACTIVITY = 5.509
1D_VALUE = 0.0
###############################################################################
LIGAND_NAME = mol_2
TITLE = "endo-2"
PHARM_SET = active
LIGAND_GROUP = 2
MUST_MATCH = false
QSAR_SET = train
ACTIVITY = 5.456
1D_VALUE = 0.0
```

B.2 Phase Main Input File

The Phase main input file (*jobname*_phase.inp) contains information that is used across the entire Develop Common Pharmacophore Hypotheses workflow. It contains sections relevant to all programs in the workflow and sections that are used only by specific programs. This file is created by the model development utilities. If you edit this file, the order of ligands must remain unchanged for an entire run. When you run Phase from Maestro, Maestro creates and updates a main input file for each run.

Each line of the Phase main input file contains a keyword-value pair separated by an equals sign (“=”), as follows:

keyword=value

Extra spaces are ignored, but blank lines are not permitted. Keywords can have string, integer, or real types. These types are enforced.

Keywords for the job name, file names, and directories are given in Table B.3. In the tables, the same text is used for the value as for the keyword, but it is set in a different font: for example, *ligand-name* represents the value of the keyword LIGAND_NAME.

Optional keywords for the Find Common Pharmacophores step are described in Table B.4. Optional keywords for the Score Hypotheses step are given in Table B.5. Each Score Hypotheses keyword has a default value, so it is not necessary to include any of them in the input file.

Changes to this file should be made with the utility programs described in Chapter 12.

Table B.3. Name and directory keywords.

Keyword	Description
BOXES_DIR	Name of directory where box files are stored. Box files are generated during the Find Common Pharmacophore step. These files contain data that is used as input for the scoring step. This keyword is used for both the Find Common Pharmacophore and Score Hypothesis steps. Optional keyword; the default is boxes.
JOB_NAME	Name given to each Phase job run with this input file. This name is used as a base for input and output files for the run. The name must match the name of the input file, *jobname*_phase.inp.
LIGAND_DIR	Directory where all ligand-related files are stored (the *ligands directory*). Should be a relative path. The ligand input files are multi-conformer Maestro files named *ligand-name*.mae. The output files include the conformer coordinate files, named *ligand-name*_xyz.phc, and the ligand sites files, named *ligand-name*_sites.phs. Optional keyword; the default is ligands.
LIGAND_NAME	Name of a ligand. This name is used to construct file names for ligand-related files. The ligand structure is contained in the file *ligand-name*.mae. For example, if the input file contains the line LIGAND_NAME=aspirin, there should be a Maestro file named aspirin.mae in the ligands directory. The input file should contain multiple lines of this kind, one for each ligand in the set. The order of these lines should *not* be changed during a run. For pharmacophore model development, the set should include only the active ligands on which the model is to be based. Do not change by hand.
RESULT_DIR	Name of the directory where results of the Score Hypothesis step are stored. Optional keyword; the default is result.

Table B.4. Optional Keywords for the Find Common Pharmacophores step.

Keyword	Description
NUM_SITES	Total number of sites in pharmacophore hypothesis (integer). Default is 5.
MIN_INTERSITE_DIST	Minimum distance between pharmacophore sites in angstroms (real). May be used to reject pharmacophores that contain, for example, an acceptor site and a donor site from the same oxygen. Default is 2.0 Å.
MAXIMUM_DEPTH	Number of times each side of the "box" is divided (integer). Default is 5.
INITIAL_BOX_SIZE	Initial box size in angstroms (real). This option should not appear in the Phase main input file. Set automatically using values for FINAL_BOX_SIZE and MAXIMUM_DEPTH as described below.
FINAL_BOX_SIZE	Final box size in angstroms (real). Default is 1.0 Å.
MIN_NUM_LIGANDS_PER_BOX	Minimum number of ligands or ligand groups that must be matched (integer).
MIN_MAX_SITES	Minimum and maximum number of sites for a feature type. Value is a string that contains 3 integers separated by commas with no spaces: *n1*,*n2*,*n3*. The first integer (*n1*) is the numerical code for the site type (see VARIANT_NAMES, below). The second integer (*n2*) is the minimum feature frequency and third integer (*n3*) is the maximum feature frequency. The maximum value of *n3* is 4. By default, the values of *n2* and *n3* are set to 0 and 4 for the standard features (A, D, H, N, P, and R), and to 0 and 0 for the custom features. If you change the defaults, the input file should contain multiple lines of this kind, one for each feature type.
VARIANT_NAMES	Comma-separated list of variants for which common pharmacophores are to be identified. These names are used to construct file names for variant-related files. Multiple lines of this kind can be used to specify the variants. By default, all variants are used. Each variant is a string of single-digit numbers in ascending order. The numbers encode the feature types, as follows: 0 Hydrogen-bond acceptor (A) 1 Hydrogen-bond donor (D) 2 Hydrophobic group (H) 3 Negatively-charged atom (N) 4 Positively-charged atom (P) 5 Projected point (Q)—not used 6 Aromatic ring (R) 7 Custom (X) 8 Custom (Y) 9 Custom (Z)

Table B.5. Optional keywords for the Score Hypotheses step.

Keyword	Description
ALIGN_CUTOFF	Maximum RMS deviation in angstroms of aligned site points from two ligands, in angstroms (real). Default is 1.2 Å.
ALIGN_WEIGHT	Weighting factor of the site alignment term in the survival score (real). See Section 6.2.2 on page 54 for definitions. Default is 1.0.
BOXES_TO_KEEP	Percentage of top-scoring boxes to be retained for volume scoring after the first pass (integer). Default is 10.
CONFORMATION_PROPERTY	Name and weight of a conformation-dependent property to use in property scoring. Value contains property name and weight, separated by a comma. Multiple conformation property entries can be specified in the input file. For example, to use MMFF relative conformation energies and weight –0.1 the value of CONFORMATION_PROPERTY is r_mmod_Relative_Potential_Energy-MMFF94s,-0.1.
FEATURE_ALIGN_CUTOFF_FILE	Name of the file that contains feature-matching tolerances. See Section B.9 on page 204 for the format of this file.
MAX_BOXES	Maximum number of boxes to be scored. Default is 50. Overrides percentage specified by BOXES_TO_KEEP.
MIN_BOXES	Minimum number of boxes to be scored, equivalent to the minimum number of returned hypotheses per variant. Overrides percentage specified by BOXES_TO_KEEP. Default is 10.
PENALTY_CONST	Used to penalize hypotheses that do not have matches for all ligands. Default is 1.1. A value of 1.0 means that no penalty is applied.
PROPERTY_NAME	Name of the property to use in property scoring, e.g. r_m_phase_activity for the activity value. Only one property can be specified, and the property must be conformation-independent.
PROPERTY_WEIGHT	Weighting factor for the property (activity) term in the survival score. See Section 6.2.2 on page 54 for definitions. Default is 0.0.
SELECTIVITY_WEIGHT	Weighting factor for the selectivity term in the survival score. See Section 6.2.2 on page 54 for definitions. Default is 0.0.
USE_PROPERTY	Calculate property scores. Value can be true or false. Default is false.
USE_SELECTIVITY	Calculate selectivity score.Value can be true or false. Default is false.
USE_VOLUME	Calculate volume scores. Value can be true or false. Default is true.

Table B.5. Optional keywords for the Score Hypotheses step. (Continued)

Keyword	Description
VECTOR_CUTOFF	Minimum vector score value needed to keep the hypothesis. Default is 0.5.
VECTOR_WEIGHT	Weighting factor of the vector term in the survival score (real). See Section 6.2.2 on page 54 for definitions. Default is 1.0.
VOLUME_WEIGHT	Weighting factor of the volume term in the survival score (real). See Section 6.2.2 on page 54 for definitions. Default is 1.0.

An example of a Phase main input file is shown below. This example includes options for which defaults exist.

```
JOB_NAME=index_26
MIN_INTERSITE_DIST=2
NUM_SITES=5
FINAL_BOX_SIZE=2
MAXIMUM_DEPTH=5
MIN_NUM_LIGANDS_PER_BOX=5
MIN_MAX_SITES=0,0,5
MIN_MAX_SITES=1,0,5
MIN_MAX_SITES=2,0,5
MIN_MAX_SITES=3,0,5
MIN_MAX_SITES=4,0,5
MIN_MAX_SITES=6,0,5
ALIGN_CUTOFF=1.2
ALIGN_WEIGHT=1
VECTOR_CUTOFF=0.5
VECTOR_WEIGHT=1
VOLUME_WEIGHT=1
SELECTIVITY_WEIGHT=1
BOXES_TO_KEEP=100
PENALTY_CONST=1
MIN_BOXES=10
MAX_BOXES=50
LIGAND_DIR=ligands
BOXES_DIR=boxes
RESULT_DIR=results
LIGAND_NAME=120_ligand
LIGAND_NAME=121_ligand
LIGAND_NAME=130_ligand
LIGAND_NAME=132_ligand
LIGAND_NAME=BAM_ligand
LIGAND_NAME=BMZ_ligand
```

B.3 Feature Definition File

This file contains definitions used to specify pharmacophore features. The default feature definition file, `phase_feature.ini`, is provided in `$SCHRODINGER/mmshare-v`*version*`/data`. This file contains commonly used definitions for the six basic feature types. You can create your own feature definition file for a particular phase run. The file should be stored in the working directory for the run, and should be named *jobname*`_feature.ini`, where *jobname* is the name of the current job as specified in the Phase main input file.

Feature definition files contain blocks of data for each feature type. Feature types can be either the default types, such as acceptor, donor, or hydrophobic, or custom features. Each feature has a geometry, which can be one of `point`, `group`, or `vector`, and a projected point type, which depends on the geometry. Projected point types include donor and a range of acceptor types for vector geometries, and aromatic ring for group geometries.

Each block of data for a feature has the following format:

`#FEATURE`	Beginning of a new feature type block
`#IDENTIFIER` *char*	Single character feature identifier
`#COMMENT` *string*	Feature comments
`#INCLUDE`	Beginning of include block
pattern1	Pattern to include
pattern2	Pattern to include
...	
`#EXCLUDE`	Beginning of exclude block
pattern1	Pattern to exclude
pattern2	Pattern to exclude
...	

The `INCLUDE` block must contain at least one pattern; the `EXCLUDE` block can be empty. The identifier character must be one of the standard set, A, D, H, N, P, R, X, Y, or Z.

Individual patterns have the following format:

string1 string2 int1 int2 int3 int4 int5 [`#` *string3*]

The components of the patterns are described in Table B.6.

Table B.6. Feature definition pattern components.

Component	Description
string1	SMARTS pattern. For hydrophobic or aromatic features this string may be `default`, indicating that the default mechanism that calls underlying libraries should be used instead of pattern matching.
string2	Geometry definition. The allowed values are `point`, `vector`, and `group`. The `point` and `vector` strings may be followed by the index of an atom in the SMARTS pattern, in parentheses: for example, `point(2)`. This index defines the point or vector atom, and by default is the first atom in the SMARTS pattern. The `group` string may be followed by a comma-separated list of atom indices, in parentheses, which define the group atoms. The default is all atoms.
int1	Reserved for future use. Set it to 1.
int2	Reserved for future use. Set it to 1.
int3	Projected point type, which can be one of the following: 0 no projected points –1 donor –2 acceptor, sp^3, 1 lone pair (lp) –3 acceptor, sp^2, 1 lone pair –4 acceptor, sp, 1 lone pair –5 acceptor, sp^3, 2 lone pairs –6 acceptor, sp^2, 2 lone pairs –7 acceptor, sp, 3 lone pairs –8 aromatic ring –9 acceptor, planar, 3 lone pairs
int4	Indicates whether this pattern is used (0) or ignored (1).
int5	Indicates whether this pattern is a default pattern (1) or not (0).
string3	Optional comments.

An example of a feature definition file is shown below. This file contains definitions of 3 types: acceptor, donor and hydrophobic.

```
#FEATURE
#IDENTIFIER  D
#COMMENT  donor: hydrogen atom attached to oxygen, nitrogen, sulfur or carbon
#INCLUDE
[#1][O;X2]            vector(1)  0 1 -1   0  1 # OH
[#1]S[#6]             vector(1)  0 1 -1   0  1 # SH
[#1][#7]              vector(1)  0 1 -1   0  1 # any NH
#EXCLUDE
[#1]OC(=O)            point(1)   0 1  0   0  1 # exclude carboxyl group
[#1]O[S;X3]=O         point(1)   0 1  0   0  1 #
```

```
#FEATURE
#IDENTIFIER  A
#COMMENT  acceptor: oxygen, nitrogen or sulfur with at least one lone pair
#INCLUDE
n1c[nH]cc1            vector(1)  0  1 -3  0  1 # his
O=[C,c]               vector(1)  0  1 -6  0  1 # carbonyl oxygen
[O;X2](~[A,a])C       vector(1)  0  1 -5  0  1 # oxygen with two lone pairs
#EXCLUDE
O=C[O-,OH]            point      0  1  0  0  1 #
[#7;X3][*]=[O,S]      point      0  1  0  0  1 # general amide
[N;X3](C)(C)[C;X3]    point      0  1  0  0  1 #
[N;X3][a]             point      0  1  0  0  1 # planar N bonded to aring
#FEATURE
#IDENTIFIER  H
#COMMENT  hydrophobic feature
#INCLUDE
default               point      1  1  0  0  1 # default calls mmphob library
#EXCLUDE
```

In addition to features, projected point features can be included in the feature definition file. These features are defined by a point at a specified distance along the vector from a donor or acceptor atom. The format of a projected feature block is the same as for a feature, except that the initial keyword is `#PROJECTED_FEATURE`, and there is an additional `#EXTEND_DISTANCE` keyword that defines the distance of the projected point site from the donor or acceptor atom. An example of a projected point feature for a donor is given below.

```
#PROJECTED_FEATURE
#EXTEND_DISTANCE  1.8
#IDENTIFIER  D
#COMMENT  donor: hydrogen atom attached to oxygen, nitrogen, sulfur or carbon
#INCLUDE
[#1][O;X2]            vector(1)  0 1 -1   0  1 # OH
[#1]S[#6]             vector(1)  0 1 -1   0  1 # SH
[#1][#7]              vector(1)  0 1 -1   0  1 # any NH
#EXCLUDE
[#1]OC(=O)            point(1)   0 1  0   0  1 # exclude carboxyl group
[#1]O[S;X3]=O
```

B.4 Inactives Scoring Input File

The input file for inactives scoring (`phase_inactive`) contains *keyword*=*value* strings that provide instructions for scoring hypotheses with respect to inactives. An exclamation point "!" may be used to add comments to input file. The allowed keywords and their values are given in Table B.7. This file is automatically generated by the utility `pharm_score_inactives`. A sample input file is shown below.

Table B.7. Keywords for inactives scoring.

Keyword	Description
`inactiveWeight`	Required. Weight of the inactives score in the final score.
`phaseOptionsFile`	Required. Phase main input file for scoring inactives. The following keywords must be set in this file: `FINAL_BOX_SIZE` From `phase_partition` job. `USE_VOLUME` From `phase_scoring` job. `ALIGN_CUTOFF` From `phase_scoring` job. `ALIGN_WEIGHT` From `phase_scoring` job. `VECTOR_WEIGHT` From `phase_scoring` job. `VOLUME_WEIGHT` From `phase_scoring` job. `LIGAND_NAME` For each inactive molecule.
`ligandArchive`	Required. Name of the archive file (`.tar`) containing the ligand multi-conformer Maestro files for each ligand specified in the main input file. Must be stored in the current directory.
`ligandDir`	Required. Name of the directory used to store the ligands when they were archived, and therefore of the temporary directory that they will be extracted to.
`hypoArchive`	Required. Name of the archive file (`.tar`) containing the hypotheses. For each hypothesis, the files *hypoDir*`/`*hypoID*`.mae`, *hypoDir*`/`*hypoID*`.tab` and *hypoDir*`/`*hypoID*`.xyz` must be present in the archive. Must be stored in the current directory.
`hypoDir`	Required. Name of the directory used to store the hypotheses when they were archived, and therefore of the temporary directory that they will be extracted to.
`survivalScore(`*hypoID*`)`	Required. Survival score from actives scoring for the given hypothesis. There should be one record containing a survival score for each hypothesis for which inactive scores are wanted. Inactive scores are calculated only for hypotheses whose survival score is listed. Records may be deleted for hypotheses whose inactive score is not required.
`featureFile`	Required. Name of the feature definition file that was used to create the hypotheses.
`tableFile`	Required. Name of plain text file containing results.
`csvFile`	Name of CSV file containing results.

```
inactiveWeight=1
phaseOptionsFile=score_inactives_phase.inp
ligandArchive=score_inactives_ligandFiles.tar
ligandDir=.ligands.tmp
hypoArchive=score_inactives_hypoFiles.tar
hypoDir=.hypotheses.tmp
```

```
survivalScore(DHHRRR_37)=14.9862
survivalScore(DHHRRR_40)=14.8790
...
survivalScore(AAADHH_16)=13.4861
survivalScore(AAADHH_12)=13.4861
featureFile=score_inactives_feature.ini
tableFile=ScoreInactivesData.tab
csvFile=ScoreInactivesData.csv
```

B.5 Hypothesis Clustering Input File

The input file for clustering of hypotheses (`phase_hypoCluster`) contains *keyword=value* strings that provide instructions for clustering hypotheses according to their geometric similarity. An exclamation point "!" may be used to add comments to input file. The allowed keywords and their values are given in Table B.7. This file is automatically generated by the utility `pharm_cluster_hypotheses`. A sample input file is shown below.

Table B.8. Keywords for hypothesis clustering.

Keyword	Description
`phaseOptionsFile`	Required. Phase main input file with combined options from `phase_partition` and `phase_scoring`. The following options must be set in this file: `FINAL_BOX_SIZE` From `phase_partition` job. `ALIGN_CUTOFF` From `phase_scoring` job. `ALIGN_WEIGHT` From `phase_scoring` job. `VECTOR_WEIGHT` From `phase_scoring` job.
`hypoArchive`	Required. Name of the archive file (`.tar`) containing the hypotheses. For each hypothesis, the files *hypoDir*`/`*hypoID*`.mae`, *hypoDir*`/`*hypoID*`.tab` and *hypoDir*`/`*hypoID*`.xyz` must be present in the archive. Must be stored in the current directory.
`hypoDir`	Required. Name of the directory used to store the hypotheses when they were archived, and therefore of the temporary directory that they will be extracted to.
`featureDefFile`	Required. Name of the feature definition file used to create the hypotheses.
`featureTolFile`	Name of the feature-matching tolerances file. Required only if feature-matching tolerances were used to create the hypotheses
`clusterFile`	Required. Name of the output file containing results of the cluster analysis. Usually named *jobname*`_hypoCluster.clu`.

Table B.8. Keywords for hypothesis clustering. (Continued)

Keyword	Description
`linkageMethod`	Method to be used for linking clusters. Allowed values are `single`, `average`, and `complete`. `single` Use the highest similarity between any two objects from the two clusters. Produces diffuse, elongated clusters. `average` Use the average similarity between all pairs of objects from the two clusters. `complete` Use the lowest similarity between any two objects from the two clusters. Produces compact, spherical clusters.
`survivalScore(`*hypoID*`)`	Required. Survival score from scoring of actives for the given hypothesis. There should be one record containing a survival score for each hypothesis that is to be clustered. Records may be deleted for hypotheses that you do not wish to cluster.

```
phaseOptionsFile=cluster_hypotheses_phase.inp
hypoArchive=cluster_hypotheses_hypoFiles.tar
hypoDir=.hypotheses.tmp
featureDefFile=cluster_hypotheses_feature.ini
linkageMethod=complete
clusterFile=cluster_hypotheses_hypoCluster.clu
survivalScore(DHHRRR_37)=14.9862
survivalScore(DHHRRR_40)=14.8790
survivalScore(DHHRRR_43)=14.8454
...
```

B.6 Multiple QSAR Model Input File

The input file for `phase_multiQsar` contains *keyword=value* strings that provide instructions for creation and use of multiple QSAR models. See Section 12.7.2 on page 149 for an overview. The file should be named *jobname*`_multiQsar.inp`. An exclamation point "!" may be used to add comments to the input file. The allowed keywords and their values are given in Table B.9. A sample input file is shown below.

Table B.9. Keywords for building multiple QSAR models with phase_multiQsar.

Keyword	Description
`actProperty`	Required. The name of the activity property exactly as it appears in the ligand Maestro files. Missing activities are set to zero.
`csvFile`	Name of CSV file containing a summary of statistics for each QSAR model.

Table B.9. Keywords for building multiple QSAR models with phase_multiQsar. (Continued)

Keyword	Description
`featureCutoffFile`	Name of file that defines positional tolerances for matching different types of pharmacophore features during ligand alignment. See Section B.9 on page 204 for a description of the format. Positional tolerances are enforced only after applying a tolerance to the intersite distances. If omitted, matching is done purely on intersite distances.
`featureFile`	Required. Name of feature definition file used to create hypotheses.
`featureRadiusFile`	Required if `modelType=pharm`. Name of file that defines feature radii.The format of this file is identical to *featureCutoffFile*, but the distances need not be the same.
`gridSpacing`	Grid spacing, in angstroms. Must lie between 0.5 and 4.0. The recommended and default value is 1.0.
`gzipAlignments`	Compress the alignments file with `gzip`. Allowed values: `true`, `false`. Default: `true`.
`hypoArchive`	Required. Name of archive file (`.tar`) containing hypothesis files for each hypotheses for which `hypoID` is specified. For each hypothesis, the files *hypoDir*`/`*hypoID*`.mae`, *hypoDir*`/`*hypoID*`.tab` and *hypoDir*`/`*hypoID*`.xyz` must be present in the archive.
`hypoDir`	Required. Name of the directory used to store the hypotheses when they were archived, and therefore of the temporary directory that they will be extracted to.
`hypoID`	Required. Hypothesis for which QSAR model is to be built. There should be one record containing this keyword for each hypothesis for which you want a QSAR model. This list need not contain every hypothesis in the hypotheses archive.
`ligandArchive`	Required. Name of archive file (`.tar`) containing ligand multiconformer Maestro files for each ligand specified in *phaseOptionsFile*.
`ligandDir`	Required. Name of the directory used to store the ligands when they were archived, and therefore of the temporary directory that they will be extracted to.
`maxFactors`	Required. Maximum number of PLS factors to include in the QSAR model. Models are created for the full sequence of models with the number of factors set to 1,...,`maxFactors`.
`modelType`	Type of model to build. Allowed values are `atom` and `pharm`. Default: `atom`.
`numTrain`	Required. Number of molecules in the training set. The first *numTrain* ligands from *phaseOptionsFile* (see below) are assigned to the training set, and the rest are assigned to the test set.

Table B.9. Keywords for building multiple QSAR models with phase_multiQsar. (Continued)

Keyword	Description
phaseOptionsFile	Required. Phase main input file with combined options for phase_partition and phase_scoring. The following keywords must be set in this file: FINAL_BOX_SIZE From phase_partition job. USE_VOLUME From phase_scoring job. ALIGN_CUTOFF From phase_scoring job. ALIGN_WEIGHT From phase_scoring job. VECTOR_WEIGHT From phase_scoring job. VOLUME_WEIGHT From phase_scoring job. LIGAND_NAME For each ligand in the training and test sets. LIGAND_NAME records for the training set should come first.
resultArchive	Required. Name of archive file (.tar) containing QSAR model results.
resultDir	Required. Name of directory for storing QSAR model results, relative to the current working directory.
tableFile	Required. Name of plain text file containing a summary of each QSAR model.
tvalueFilter	The minimum absolute t-value a bit must exhibit in order to be retained in the QSAR model. The t-value is the ratio of the regression coefficient to the standard deviation (i.e., uncertainty) in the coefficient. If using this option, a reasonable value of tvalueFilter is 2.0 or greater. In order to estimate the t-values, a series of leave-n-out PLS models are built from the training set. The options tvalueExclude and tvalueSeed may be specified to control this process. tvalueFilter is valid only when runMode=train.
tvalueExclude	Number of training set observations to exclude when estimating t-values. Must be smaller than numTrain. If omitted, tvalueExclude is set to numTrain/10. Valid only when tvalueFilter is specified.
tvalueSeed	Non-negative integer random seed used to select subsets to exclude when estimating t-values. If 0 or omitted, the current time is used to pick a random seed. Valid only when tvalueFilter is specified.
useVolumeGroups	Consider only atoms of the same MacroModel type when computing volume score overlaps. This favors alignments that superimpose chemically similar atoms. Allowed values are true and false. Default: false.

```
modelType=atom
numTrain=27
gridSpacing=1
maxFactors=3
actProperty=r_phase_Ligand_Activity
useVolumeGroups=false
phaseOptionsFile=build_qsar_phase.inp
ligandArchive=build_qsar_ligandFiles.tar
ligandDir=.ligands.tmp
hypoArchive=build_qsar_hypoFiles.tar
hypoDir=.hypotheses.tmp
hypoID=DHHRRR_37
hypoID=DHHRRR_40
hypoID=DHHRRR_43
...
hypoID=AAADRR_22
hypoID=AAADHH_16
hypoID=AAADHH_12
featureFile=build_qsar_feature.ini
tableFile=BuildQsarData.tab
csvFile=BuildQsarData.csv
resultDir=BuildQsarResults
resultArchive=build_qsar_results.tar
```

B.7 QSAR Model Input File

The input file for the QSAR module, `phase_qsar`, contains *keyword=value* strings that provide instructions for creation and use of the QSAR model. See Section 12.7.3 on page 150 for an overview. The file should be named *jobname*`_qsar.inp`. An exclamation point "!" may be used to add comments to input file. The allowed keywords and their values are given in Table B.10. A sample input file is shown below.

Table B.10. Keywords for the QSAR input file.

Keyword	Description
`actFile`	Text file containing activity data for the structures in `maeFile`. There should be one activity value per line, with no extraneous data or characters. Either `actFile` or `actProperty` (but not both) must be specified if `runMode=train`.
`actProperty`	The name of the activity property exactly as it appears in `maeFile`. Missing activities are set to zero. Either `actFile` or `actProperty` (but not both) must be specified if `runMode=train`.

Table B.10. Keywords for the QSAR input file. (Continued)

Keyword	Description
csvBits	Name of CSV file containing the volume occupation bits for each ligand in the QSAR model. The columns written are Index, Title, QSAR_Set, Activity, Bit1, Bit2, etc., where QSAR_Set is train or test, and training set data are written before test set data.This file can be supplied to the utility canvasPLS with the option -autoScaleOff to get the exact same model and predictions produced by phase_qsar.
csvLNO	Name of CSV file containing the cross-validated activities. The columns written are Index, Title, Activity, Pred(1),...,Pred(maxFactors).Valid only when runMode=train.
csvPred	Name of CSV file containing the predicted activities for each ligand in the QSAR model. The columns written are Index, Title, QSAR_Set, Activity, Pred(1),...,Pred(maxFactors), where QSAR_Set is train or test, and training set data are written before test set data.
duplex	A non-negative integer value that controls how the training set molecules are selected. If duplex=0, the first numTrain molecules in maeFile are used. If duplex>0, the value is treated as a random seed to sample numTrain molecules from maeFile. The default is duplex=0. Valid only when runMode=train.
featureRadiusFile	File that defines the size of pharmacophore features. Each line in the file should contain a 1-character feature type, followed by a radius in angstroms. These radii are used only in the model creation process. Valid only when runMode=train and pharmFile has been specified.
gridSpacing	The distance in angstroms between neighboring points in the 3D grid. The default is 1.0. Values may range from 0.5 to 4.0. Valid only when runMode=train.
LNO	Number of molecules to hold out when doing leave-n-out cross-validation. This keyword also turns on cross-validation. If the number of molecules in the training set is not a multiple of n, then some molecules will be held out more than once, and the predicted activity will be an average.
maeFile	Maestro file containing the molecules of interest. Required if runMode=train.
maxFactors	The maximum number of PLS factors to include in the QSAR model. Statistics and predictions are ultimately accessible for the full sequence of models with the number of factors set to 1,...,maxFactors. Valid only when runMode=train, in which case maxFactors must be specified.
modelFile	QSAR model file. The model is exported if runMode=train; the model is imported if runMode=test. Required if runMode=test.

Table B.10. Keywords for the QSAR input file. (Continued)

Keyword	Description
numTrain	The number of molecules in maeFile that will be used to train the model. The default is all molecules. Use the duplex option to control which molecules are assigned to the training set. Valid only when runMode=train.
outputFile	File for ordinary output. The default is to write to standard output.
pharmFile	Used for creating or testing a pharmacophore-based model. This file contains coordinates of site points that have been aligned to a particular hypothesis, one set of points for each molecule in maeFile. The file may be obtained by running phase_fileSearch on a Maestro file containing the molecules of interest. Valid when runMode=train or runMode=test. If runMode=test and modelFile contains a pharmacophore-based model, then pharmFile must be specified.
printBits	Boolean indicating whether or not volume occupation bit strings should be written to *outputFile*. printModel=true produces the full list of volume elements and atom classes that define the bit set. If runMode=train, bit strings for the training set and test set molecules are written separately. printBits=true is allowed only when maeFile has been specified.
printModel	Boolean (true or false) indicating whether or not a summary of the model should be written to outputFile. The default is printModel=false.
printPred	Boolean indicating whether or not predicted activities should be written to outputFile. If runMode=train, the training set and test set predictions are written separately. The default is printPred=false. printPred=true is allowed only when maeFile has been specified.
runMode	Legal values are train and test. Indicates whether a new model will be created (train) or whether an existing model will be imported (test). Required.
tvalueExclude	Number of training set observations to exclude when estimating t-values. Must be smaller than numTrain. If omitted, tvalueExclude is set to numTrain/10. Valid only when tvalueFilter is specified.
tvalueFilter	The minimum absolute t-value a bit must exhibit in order to be retained in the QSAR model. The t-value is the ratio of the regression coefficient to the standard deviation (i.e., uncertainty) in the coefficient. If using this option, a reasonable value of tvalueFilter is 2.0 or greater. In order to estimate the t-values, a series of leave-n-out PLS models are built from the training set. The options tvalueExclude and tvalueSeed may be specified to control this process. tvalueFilter is valid only when runMode=train.
tvalueSeed	Non-negative integer random seed used to select subsets to exclude when estimating t-values. If 0 or omitted, the current time is used to pick a random seed. Valid only when tvalueFilter is specified.

```
runMode=train
maeFile=steroids.mae           ! 31 molecules
actFile=steroids_act.txt       ! 31 activity values
modelFile=steroids_model.dat
numTrain=21
duplex=1234567                 ! Random split: 21 train / 10 test
gridSpacing=1.0
maxFactors=4
printModel=true
printPred=true
printBits=true
```

B.8 Feature Frequencies File

This file is used to set the minimum and maximum allowed feature frequencies for common pharmacophore perception. It is named `FeatureFreq.tab`. Each line contains the letter code for the feature type, followed by the minimum and the maximum number of occurrences of that feature in any hypothesis. The example below shows the default frequencies.

```
###############################################################################
#                                                                             #
# Feature frequency file.  Used to set mimimum and maximum allowed feature    #
# frequencies for common pharmacophore perception.   You may change these     #
# limits, but do not make any other modifications to this file.               #
#                                                                             #
###############################################################################
A 0 4
D 0 4
H 0 4
N 0 4
P 0 4
Q 0 0
R 0 4
X 0 4
Y 0 4
Z 0 4
END_OF_FEATURE_DATA
```

For example, the text `A 0 4` indicates that each common pharmacophore will be restricted to contain between zero and four acceptors (inclusive). If you had some prior knowledge of the problem at hand, you could adjust these frequencies to narrow the focus in accordance with that knowledge. For example, suppose it has been established that all actives bind to a specific site on the receptor through a hydrogen bond, where the ligand acts as an acceptor. In that case, you have justification to require that each common pharmacophore contain at least one acceptor.

B.9 Feature-Matching Tolerances File

This file is used to set feature-matching tolerances. When searching for matches, this file should be named *hypoID*.tol, where *hypoID* is the hypothesis identifier used to define other hypothesis-related files. Although this file contains no hypothesis-specific information, the naming convention is required for the file to be used when searching for matches for a specific hypothesis. You must make a copy of it with the appropriate name for each hypothesis for which you want to use feature-matching tolerances.

The file contains one line for each feature type for which tolerances are to be used. Each line consists of a single character feature type and a tolerance value in angstroms, separated by a space. If a feature type is omitted, a default tolerance of 1.0 is used for that feature type. The feature type ? can be used to define a default cutoff for any feature type not listed in the file. The following is a sample feature-matching tolerances file:

```
################################################################################
#                                                                              #
# Feature matching tolerances applied when hypotheses are scored with respect  #
# to actives.  You may change the tolerances, but do not make any other        #
# modifications to this file.  To completely disable the use of tolerances,    #
# remove the FEATURE_ALIGN_CUTOFF_FILE option from score_actives_phase.inp.    #
#                                                                              #
################################################################################
A 1
D 1
H 1.5
N 0.75
P 0.75
R 1.5
X 1
Y 1
Z 1
END_OF_FEATURE_DATA
```

B.10 Hypothesis-Specific Tolerances File

This file is used to set tolerances for the specific features of a hypothesis in a search for matches. For a specific feature, the matching feature in the hit must be within the specified distance of the feature in the hypothesis after the hit is aligned to the hypothesis. The file must be named *hypoID*.dxyz, where *hypoID* is the hypothesis identifier used to name hypothesis-related files. It contains multiple lines, each consisting of a single character feature type and a tolerance value, separated by a space. The file must contain one line for each site in the hypothesis, and there is a one-to-one mapping of the tolerances to the sites in the hypothesis. To see how the hypothesis maps to the reference ligand, you can use the Edit Hypothesis panel in Maestro.

The following is a sample hypothesis-specific tolerances file:

```
D 1.50
H 2.00
H 1.50
R 2.00
R 1.50
```

B.11 Matching Constraints File

This file is used to set tolerances for distances, angles, and dihedral angles between features in a search for matches. The file must be named *hypoID*`.cnst`, where *hypoID* is the hypothesis identifier used to name hypothesis-related files. It contains multiple lines, each consisting of a two to four sites, followed by the value of the parameter and a tolerance value, separated by spaces. The sites are identified by the letter code and the index from the *hypoID*`.xyz` file, with no spaces, and can include projected points, which have a letter code of Q. Two sites define a distance, three sites define an angle, and four sites define a dihedral[1]. The following example includes a distance constraint, an angle constraint, and a dihedral constraint:

```
A0 R10 5.583 1
D4 Q20 H5 51.1 10
Q12 R10 Q24 H5 143.3 20
```

When a database or file is searched, each match to the hypothesis is examined to see if the matched sites satisfy every constraint. When there is more than one projected point on a matched ligand site, every possible mapping to the reference ligand projected point (or points) is considered to see if the constraint can be satisfied.

You can set up constraints for an existing hypothesis with the program `phase_constraints` and Maestro. For more information on this program and the procedure, enter the following command:

```
$SCHRODINGER/run -FROM phase phase_constraints -h
```

Projected points are temporarily represented as feature type Z, because Maestro does not display features of type Q. This procedure supplies default tolerance values of 1.0 Å for distance constraints, 10° for angle constraints, and 20° for dihedral constraints, but you can edit the values in the file once it is generated.

1. A dihedral 1-2-3-4 is positive if 4 is clockwise from 1 when viewed down the 2-3 axis.

B.12 Site Mask File

This file is used to determine how specific sites or groups of sites are matched in a partial match. The file must be named *hypoID*`.mask`, where *hypoID* is the hypothesis identifier used to name hypothesis-related files. The file must contain one line for each site in the hypothesis, with an integer value on each line, the *mask value*. These values determine how the site is matched:

>1: Site is grouped.
1: Site must be matched.
0: Site can be matched but need not be matched (matching is optional)
–1: Site must not be matched.

For example, consider a hypothesis that contains the site types D, H, H, R, and R. To require that every partial match contains the donor site but not the second hydrophobic site, the corresponding site mask file should contain the following 5 lines:

```
4 D  1
5 H  0
6 H -1
9 R  0
11 R 0
```

The first two columns are the same as in the *hypoID*`.xyz` file. The third column contains the numbers that indicate how the site is matched.

Sites can be grouped, and matching requirements applied to the group rather than to individual sites. This feature is used for partial matching, in which the number of matches required is less than the size of the group. If you want to require all sites in a group to match, you should simply use a 1 for the mask value for each site rather than create a group. Likewise, you can only exclude a site from matching by specifying it as an individual site. A site mask can contain a mixture of grouped and ungrouped sites.

The rules for grouped sites are:

- Every member of a group must have the same mask value.
- Sites in different groups must have different mask values.
- A site cannot be a member of two groups.

The mask value encodes both the site grouping and the matching behavior. It is interpreted as $2^{\text{groupNumber}}$+minMatch, where groupNumber is the index of the group, and minMatch is the minimum number of sites that must match from the group. The value of minMatch must be less than $2^{\text{groupNumber}}$, which is the maximum size of the group. The encoding of the mask value for the first few values is shown in Table B.11.

Table B.11. Encoding of mask values for groups 1 and 2.

maskValue	groupNumber	minMatch
2	1	0
3	1	1
4	2	0
5	2	1
6	2	2
7	2	3

From the values in this table, it should be clear that there are multiple mask values that can be used to encode a given grouping and matching requirement. For example, if only 1 site from a given group must be matched, valid mask values are 3, 5, 9, 17, 33, and so on. If 2 sites from a group must be matched, valid mask values are 6, 10, 18, 34, ... If 3 sites from a group must be matched, valid mask values are 7, 11, 19, 35, ... This redundancy is necessary in order to support different groups for which the same number of sites must be matched.

As an example, suppose you have a 5-point hypothesis DHHRR where sites 1 and 3 are in a group, sites 2, 4 and 5 are in a group, and the minimum number of sites to match in the two groups are 1 and 2, respectively. Then any of the following site mask files are valid:

File 1:

```
1 D 3
2 H 6
3 H 3
4 R 6
5 R 6
```

File 2:

```
1 D 9
2 H 6
3 H 9
4 R 6
5 R 6
```

File 3:

```
1 D 3
2 H 10
3 H 3
4 R 10
5 R 10
```

An alternative for excluding sites from matching is to use a hypothesis rules file—see Section B.13 on page 208.

Note: The Phase 2.0 format, in which only the column of numbers is given (the third column in the above example files) is still supported, but the new format is recommended.

B.13 Hypothesis Rules File

Feature rules allow generalized matching according to permitted features and prohibited features for each site in the hypothesis. The file must be named *hypoID*`.rules`, where *hypoID* is the hypothesis identifier used to define other hypothesis-related files. This file will be used when you search for matches if it is present with the other input files, and `useFeatureRules` is not set to `false` in the input file.

The feature rules file must contain one line for each site in the hypothesis. Each line must contain the feature number followed by a string of features that are permitted at this site; these can optionally be followed by a string of features that are prohibited at this site. At a minimum, the feature rules file must contain the first two entries from each line in the corresponding *hypoID*`.xyz` file. The prohibited feature string is only used when doing partial matching.

Suppose you have the following *hypoID*`.xyz` file:

```
4 A 5.8663 -0.0455 1.5777
5 D 8.5193 0.6900 2.3939
6 H 9.5900 3.6582 1.8851
9 R 7.2866 2.6408 1.0380
11 R 0.8596 0.3614 1.2689
```

When finding matches to this hypothesis you want the following rules to apply:

- The hydrophobic site should match both hydrophobic and aromatic features.
- The aromatic sites should match both aromatic and hydrophobic features.
- The hydrophobic feature should never be overlaid onto an ionic site.

To achieve this behavior, you would construct the following feature rules file:

```
4 A
5 D
6 HR NP
9 RH
11 RH
```

The rule for site 6, which is a hydrophobic site, means that it can be matched to a hydrophobic or aromatic feature in the database. If partial matching is in use, each partial match that fails to include site 6 (i.e., it's not matched to an H or R in the molecule), is checked to make sure that it does not inadvertently match a negative or positive feature in the database molecule.

Vector scoring is turned off completely if any permitted feature string contains more than just the original feature type (as in the above example). This is done to avoid possible errors and inconsistencies when vector and non-vector features are aligned.

B.14 Excluded and Included Volume Files

There are two separate file formats for excluded and included volumes. Excluded volume data can be stored in a Maestro-format file with the extension `.xvol`, or in a plain text file ("native format") with the extension `.ev`. The Maestro-format file is used when importing and exporting excluded volumes from Maestro, and is also read by the programs that use excluded volumes. You can modify this file by hand if you wish, but it includes metadata that is used to construct tables in Maestro. The plain text file has the following format:

spheres hydrogens
x_1 y_1 z_1 r_1
x_2 y_2 z_2 r_2
...

where *spheres* is the number of excluded volume spheres, and *hydrogens* is a flag for considering hydrogen atoms when determining excluded volume violations, with values of 0 (don't consider hydrogens) or 1 (consider hydrogens). The remaining lines in the file list the cartesian coordinates of each sphere center and its radius.

Included volume data has the same format as excluded volume data, and can be stored in files of either format. The extension is the same for both formats, `.ivol`, and the format is determined from the file when it is read. The flag for considering hydrogens is ignored if it is present.

Phase Utilities

In addition to the utilities described in Chapter 12 and Chapter 13, there are several utilities provided with Phase that may be useful in some circumstances. These utilities are described in this appendix. The syntax and options descriptions for these utilities can be listed by running the command with the `-h` option.

C.1 phase_cluster_hits

Cluster the structures in a Phase hit file according to the sites that were matched. By default, hits are clustered according to the matched ligand site types. In other words, the Matched Ligand Sites strings in a given cluster are identical, except for the site numbers. For example, two structures with the following Matched Ligand Sites would be assigned to the same cluster.

```
A(1) D(5) H(-) R(12) R(-)
A(2) D(7) H(-) R(15) R(-)
```

The overall ordering of clusters is done as follows. The first cluster corresponds to all sites being matched, and clusters that match fewer sites come after clusters that match more sites. The ordering within a given cluster preserves the order in which the structures were encountered within the hit file. Thus if the hits are sorted by decreasing fitness, the hits within a cluster are also sorted.

The syntax is as follows:

`phase_cluster_hits` *hitFile outFile* `[-pos]`

where *hitFile* is the Phase hit file from a database or file search, and *outFile* is a Maestro or SD file with structures ordered by cluster. The property `i_phase_Cluster_Number` is added to each structure.

When run with `-pos`, only the positions in the Matched Ligand Sites string matter, and differences in the site types are ignored. This option is relevant only if feature-matching rules were applied during the search and one or more sites were permitted to match multiple types. For example, if ADHRR is allowed to match ADHRR or ADHHR, hits with the following Matched Ligand Sites would be assigned to the same cluster if this option is used.

```
A(1) D(5) H(-) R(12) R(-)
A(2) D(7) H(-) H(10) R(-)
```

C.2 convert_hypoDistToXYZ

The utility `convert_hypoDistToXYZ` creates a hypothesis `.xyz` file from a file containing intersite distances. The syntax of the command is as follows:

`convert_hypoDistToXYZ` *hypoID*

where *hypoID* is the prefix used to identify input and output files. The input file should be named *hypoID*.dist and it should contain the alphabetized variant, followed by the strict lower triangle of intersite distances, as indicated by the following example:

```
ADHP
d(D,A)
d(H,A) d(H,D)
d(P,A) d(P,D) d(P,H)
```

Here the `d(x,y)` are the distances.

Since intersite distances are unaffected by reflection operations, this program creates two mirror image hypothesis files: *hypoID*`_1.xyz` and *hypoID*`_2.xyz`. A least-squares technique is applied to align the two sets of sites, and the RMSD is reported, so it will be obvious whether the mirror images are equivalent.

There are no reference ligands for these two hypotheses, so two SD files, *hypoID*`_1.sdf` and *hypoID*`_2.sdf` are created to help visualize the points. To achieve colors similar to those used in the Phase GUI, the pharmacophore features in the SD files are represented by different elements—see Section C.5 for a description.

C.3 convert_hypoXYZToDist

The utility `convert_hypoXYZToDist` uses a hypothesis `.xyz` file to create a file containing intersite distances (`.dist`). The syntax of the command is as follows:

`convert_hypoDistToXYZ` *hypoID*

The output is the file *hypoID*`.dist`.

C.4 convert_hypoFeatures

The utility `convert_hypoFeatures` attempts to convert a hypothesis to use a new set of feature definitions. The syntax of the command is as follows:

`convert_hypoFeatures` *hypoID featureFile newHypoID*

where *hypoID* is the prefix used to name the files in the existing hypothesis, e.g., *hypoID*`.def`, *hypoID*`.mae`, *hypoID*`.tab`, *hypoID*`.xyz`. If these files are not in the current directory, you must include the path in *hypoID*. *featureFile* is the file containing the new feature definitions. *newHypoID* is the prefix for the converted hypothesis files, and must be different from *hypoID*.

C.5 create_hypoSDFile

The utility program `create_hypoSDFile` creates an SD file to help visualize hypotheses that have no reference ligand. To visualize the hypothesis, simply import the SD file into Maestro. The syntax of the command is as follows:

`create_hypoSDFile` *hypoID* `[-ref] [-mol2]`

The input hypothesis file should be named *hypoID*`.xyz`. The pharmacophore sites in that file are used to create the structure file, which by default is *hypoID*`.sdf`. You can instead create a MOL2 file *hypoID*`.mol2` with the `-mol2` option. The reference ligand from *hypoID*`.mae` is appended to the output file if you use the `-ref` option. This option requires the file *hypoID*`.tab` as well.

Pharmacophore features in the SD file are represented by atoms whose colors are similar to those used to render features in the Phase GUI, as follows:

A	oxygen (red)
D	nitrogen (blue)
H	boron (green)
N	bromine (dark red)
P	sodium (dark blue)
R	silicon (orange)

C.6 create_hypoFiles

This utility creates the `.def`, `.mae`, `.xyz` and `.tab` hypothesis files from a single reference ligand structure and a feature definition file. The syntax of the command is as follows:

`create_hypoFiles` *inFile hypoID* [*fdFile*]

where *hypoID* is the prefix used to name the hypothesis files, *inFile* is the Maestro or SD file that contains the reference ligand structure, and *fdFile* is an optional feature definition file. If this last file is omitted, the default feature definitions in the installation are used. Note that all surface-accessible sites in the reference ligand are included in the `.xyz` file. You can edit the hypothesis in the Edit Hypothesis panel.

C.7 phase_volCalc

This utility calculates a matrix of overlapping volume values between structures in one or two Maestro files. Volumes are computed by treating each molecule as a set of atomic spheres. The syntax and options descriptions can be listed by running the command with the `-h` option.

C.8 rmsdcalc

This utility computes the RMSD between each structure in a given Maestro file and a corresponding reference structure from a second Maestro file. The correspondence between structures is done by title, so reference titles must be unique, and each structure for which the RMSD is sought must have a title that matches one of the reference structure titles. Note that RMSD cannot be computed between two structures unless they have the same connectivity. The syntax and options descriptions can be listed by running the command with the `-h` option. The file specifications are given in Table C.1.

Table C.1. File specifications for the rmsdcalc command.

File Specification	Description
`-screen` *screenFile*	Maestro file (`.mae`, `.maegz`, or `.mae.gz`) containing the structures for which RMSDs are sought.
`-ref` *refFile*	Maestro file (`.mae`, `.maegz`, or `.mae.gz`) containing the reference structures.
`-out` {*csvFile*|*maeFile*}	Output file. If the extension is `.csv`, a comma-separated file is created with the title and RMSD for each structure in *screenFile*. If the extension is `.mae`, `.maegz`, or `.mae.gz`, a Maestro file is created with each structure from *screenFile* and the property `r_rmsdcalc_RMSD`.

C.9 flex_align

This utility uses the shape screening technology to flexibly align a set of structures to a flexible template, then reports the alignments associated with the template conformer that yielded the highest average similarity to all structures. The conformers can be generated during the job or they can be pregenerated. The syntax and options descriptions can be listed by running the command with the `-h` option.

The file containing the structures to align is specified with `-screen`, and the file containing the template structure is specified with `-shape`. Both files can be in either Maestro or SD format, compressed or uncompressed. If *shapeFile* and *screenFile* are the same, all pairs of

structures in the file are rigidly aligned to see which one is most suitable as a template. *shapeFile* cannot be the same as *screenFile* when using pregenerated conformers.

The job name is specified with `-JOB`, and the aligned structures are written to a compressed Maestro file named *jobName*`_flex_align.maegz`.

C.10 phase_align_core

This utility combines constrained conformational sampling with shape similarity to superimpose a set of ligands onto a rigid template in a consistent manner.

`phase_align_core` *template ligands* {`-fuzzy` *level* | `-core` *smartsFile* | `-mcs` *pwFile* `[-titles]`} [*options*]

where *template* is a Maestro or SD file containing a single 3D template structure to which the ligands are to be aligned, and *ligands* is a Maestro or SD file containing the ligand structures.

The options descriptions can be listed by running the command with the `-h` option. Standard Job Control options are supported, as described in Table 2.1 of the *Job Control Guide*, along with the `-LOCAL`, `-WAIT`, `-NOJOBID` options described in Table 2.2 of the *Job Control Guide*.

By default, the largest Bemis-Murcko scaffold that is shared by a given ligand and the template is identified and treated as the common core. The internal 3D coordinates of the ligand core atoms are reset to match those of the template, and the ligand is aligned to the template on the core atoms. Constrained conformational sampling of the ligand is done with the common core held fixed. The conformer with the highest overall shape similarity to the template is selected.

Aligned ligands are written to *jobname*`_align.maegz`, where *jobname* is the base name of the ligands file.

C.11 randsub

This utility selects a random subset of lines from a file and writes them to another file. The syntax and options descriptions can be listed by running the command with the `-h` option.

C.12 create_molSites

This utility creates a CSV file that contains the Phase features and associated atom numbers for the structure in the input Maestro file. The syntax and options descriptions can be listed by running the command with the `-h` option.

The output file contains lines for each site in the format:

index`,`*type*`,`*x*`,`*y*`,`*z*`,`*atom0*`[,`*atom1*`...]`

where

index is the site index,
type is the 1-character site type,
x is the *x* coordinate of the site,
y is the y coordinate of the site,
z is the z coordinate of the site, and
atomk is the index of the *k*th atom contributing to site.

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
- Phase purchaser (company, research institution, or individual)
- Primary Phase 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.

References

1. Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A., PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening. 1. Methodology and Preliminary Results, *J. Comput. Aided Mol. Des.* **2006**, *20*, 647–671.

2. 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.

3. 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.

4. 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.

5. 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.

Glossary

Active compound—A compound that shows high affinity for the biological target. Synonymous with the term *ligand.*

Active set—The set of active compounds that is used to develop a pharmacophore model. This set does not necessarily include all active compounds.

Excluded volume—A region of space in a pharmacophore hypothesis that should not be occupied by any atom of an active compound.

Feature—see **Pharmacophore feature**

Hit—A structure in a 3D database that is found to contain an arrangement of site points that can be mapped to the pharmacophore hypothesis. A hit is not necessarily active, but it is presumed to have a greater than average probability of being active if it was retrieved using a valid hypothesis.

Hypothesis—see ***n*-Point pharmacophore hypothesis**

Inactive compound—A compound that shows little or no affinity for the biological target.

Intersite distance—The distance between any two site points in a pharmacophore.

Ligand—see **Active compound**

Negative compound—A compound that is inactive, yet highly similar in structure to one or more known actives. Some compounds are negative because they lack certain key pharmacophore features found in true actives. Other negatives may actually satisfy exactly the same pharmacophore hypotheses as the actives, but possess extraneous structural characteristics that prevent binding.

Pharmacophore feature—A characteristic of chemical structure that may facilitate a noncovalent interaction between a ligand and a biological target. Examples are hydrogen-bond acceptor (“A”), hydrogen-bond donor (“D”), hydrophobe (“H”), positive ionic center (“P”), negative ionic center (“N”).

Pharmacophore site—The labeling and location of a particular pharmacophore feature within a molecule. For example, a hydrogen bond acceptor site could simply be a nitrogen atom which carries an available lone pair. A hydrophobic site might be a methyl carbon or the

centroid of a phenyl ring. The term *site point* is often used interchangeably with pharmacophore site.

***n*-Point pharmacophore**—Any 3D arrangement of *n* pharmacophore features.

***n*-Point pharmacophore hypothesis**—A specific 3D arrangement of *n* pharmacophore features, with associated uncertainties in the feature positions. High affinity ligands in their active conformations are expected to contain pharmacophore sites that can be mapped (within the limits of uncertainty) to any valid hypothesis. A given hypothesis may contain features that are associated with a single mode of binding, or it may contain features that are common to two or more modes of binding.

Reference ligand—The ligand that provides the pharmacophore that defines a hypothesis. In pharmacophore model development, this pharmacophore yields the highest multi-ligand alignment score for the active-set ligands. The reference ligand matches the hypothesis exactly, and has a perfect fitness score.

Site point—see **Pharmacophore site**

3D Database—A set of molecules, each of which is represented by one or more 3D conformational models, augmented with a pharmacophore-based representation of the molecules. A 3D database includes feature types and site point coordinates for each conformation.

Variant—The set of feature types in a pharmacophore. For example, the variant AHH indicates a 3-point pharmacophore containing one hydrogen bond acceptor and two hydrophobic sites.

Vector feature—A pharmacophore feature that contains directionality, such as a hydrogen bond acceptor, hydrogen bond donor, or aromatic ring. A vector feature does not necessarily have vector geometry.

Vector geometry—the geometric characteristics of hydrogen bond acceptors and donors. Refers to the direction of lone pairs in a hydrogen-bond acceptor or the direction of the heavy-atom–hydrogen-atom bond in hydrogen-bond donors. Features with vector geometry must be vector features.

Index

D

E

F

G

H

I

L

M

N

S

T

V

W

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®