SEAware 1.6 User Manual

Updated July 2013

Questions? You can reach us at:

support@seachangepharma.com

a 409 Illinois Street
San Francisco, CA 94158
t +1 (415) 937 1732
f +1 (415) 534 1225
w seachangepharma.com
e info@seachangepharma.com

Table of Contents

1 Introduction 2
1.1 Overview 2
1.2 Installation 2
1.3 Hardware license activation 2

2 SEAview – a graphical SEA tool 3
2.1 Overview 3
2.2 Running SEAview 3
2.2.a Your first SEA prediction 3
2.2.b Running predictions using different SEA libraries 4
2.3 Running SEAview on many drugs at once (batch mode) 5
2.3.a Compound file formats used by SEAview 6
2.3.b Using SEAview with custom molecule fingerprints 6
2.4 Designing your own libraries for SEAview 9
2.4.a Viewing library information 9
2.4.b Exporting an existing library (library unpacking) 10
2.4.c Building SEAware libraries (library packing) 13
2.4.d Calculating background models (fit generation) 15
2.4.e Modifying an existing library 17

3 SEAshell – a command-line SEA tool 18
3.1 License management commands 18
3.2 SEAshell.exe tutorial 18

4 References 25

1 Introduction

1.1 Overview

Chemically similar drugs often bind to biologically diverse targets, making it difficult to predict what off-target effects a drug might have by protein structure or sequence alone. The Similarity Ensemble Approach (SEA) addresses this problem using a different strategy; it groups receptors according to the chemical similarity of their ligands, and can identify unknown relationships between ligands and receptors amenable to experimental testing. To do so, SEA uses a statistical model to correct for chemical similarity expected at random. For a full discussion, please see **4. References**.

1.2 Installation

After downloading, follow the instructions in the installation wizard to install SEAware on your local machine. The installer will automatically detect and install the 32-bit (x86) or 64-bit (x64) version, as is appropriate to your computer. SEAware currently runs on Microsoft Windows versions XP through 8.

1.3 Hardware license activation

Your SEAware license is specific to your computer; please be sure to generate the license request from the computer that you intend to use. Start SEAview, and you will be prompted to create a machine specific license request. This is because SEAware licenses are limited to one machine instance, whether physical or virtual. Use the generate button, fill out the license request (**Figure 1**), and send the saved request file to license@seachangepharma.com

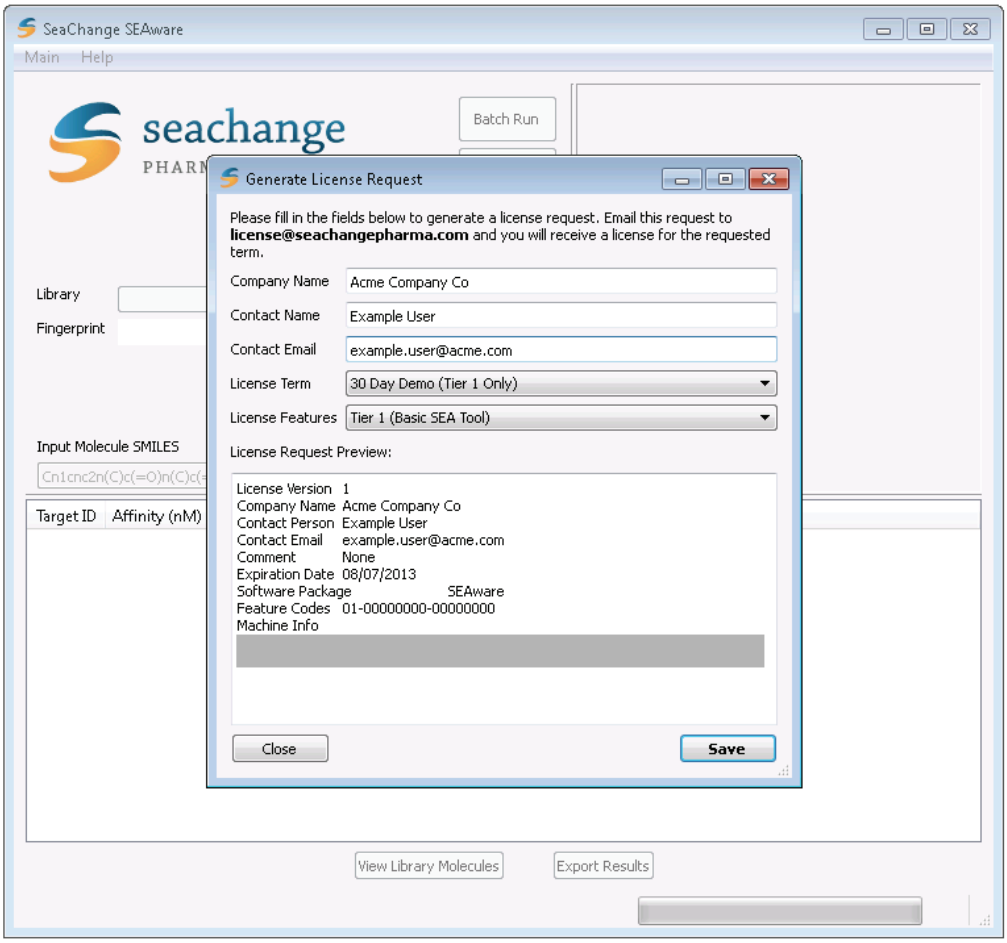


Figure 1 To generate your license, fill in and Save your request to a file, then send it to: license@seachangepharma.com

We will verify your request, and return your license key via email. When you receive it, move it to a safe location, and point the SEAview application to that location using Change License.

2 SEAview – a graphical SEA tool

2.1 Overview

This section is intended as a reference manual for the features of the SEAview tool. It is intended to give an overview of the ideas which form the basis of SEA and to detail the available user parameters. It is not intended to be a substitute for papers written on SEA (for this, please see **4. References**).

2.2 Running SEAview

The SEAview program lets you immediately start exploring with SEA. SEAview may take a minute to open, while it loads library data into the memory so that it can run predictions quickly. Once loading is complete, you can enter the structure of any small molecule into the search box and predict its targets using SEArch.

2.2.a Your first SEA prediction

SEAview operates on molecules in the SMILES format (Simplified Molecular Input Line Entry System; http://www.daylight.com/smiles). You can predict targets for a drug as follows:

1. Input the drug's SMILES; one quick source for these structures is PubChem (http://pubchem.ncbi.nlm.nih.gov). For caffeine, PubChem yields: `CN1C=NC2=C1C(=O)N(C(=O)N2C)C`
2. Press SEArch; a list of targets will appear in the table, sorted by strongest SEA p-value (**Figure 2** below).
3. Double-click on any target (or select it in the table and View Library Molecules) to view all the reference ligands for that target, sorted by their similarity to your drug.

The results table displays the following information about each target prediction:

Column name	Description
Target ID	An ID uniquely identifying this target within the current library. For the ChEMBL libraries distributed with SEAware, this is typically the accession number (http://www.uniprot.org).
Affinity (nM)	The affinity (to the closest log "bin") associated with the prediction. This is an experimental feature, which is described in greater depth in section 2.4.c below.
# Mols	The number of ligands known for this target within the current library.
P-Value	The SEA p-value of the prediction, where a p-value closer to zero (e.g., 1e-100) represents a stronger prediction. P-values approaching 1.0 (e.g., 1e-2) approach insignificance.
Max Tc	The maximum Tanimoto coefficient (Tc) present between the searched drug and its closest neighbor among the ligands already known to the target. The Tc is a pairwise score between molecules that ranges from 0.0 (no similarity) to 1.0 (identity). Any targets with Max Tc = 1.0 are shown at the top of table, because that means that the searched drug is in fact already *annotated* to that target in the current library.
Short Name	A short, human-readable name for the protein target. For the ChEMBL libraries distributed with SEAware (extracted from http://www.ebi.ac.uk/chembl), this is typically the protein's UNIPROT ID (http://www.uniprot.org).
Description	The longer, full-text name of the protein target.

Table 1 Description of data shown SEAview results table.

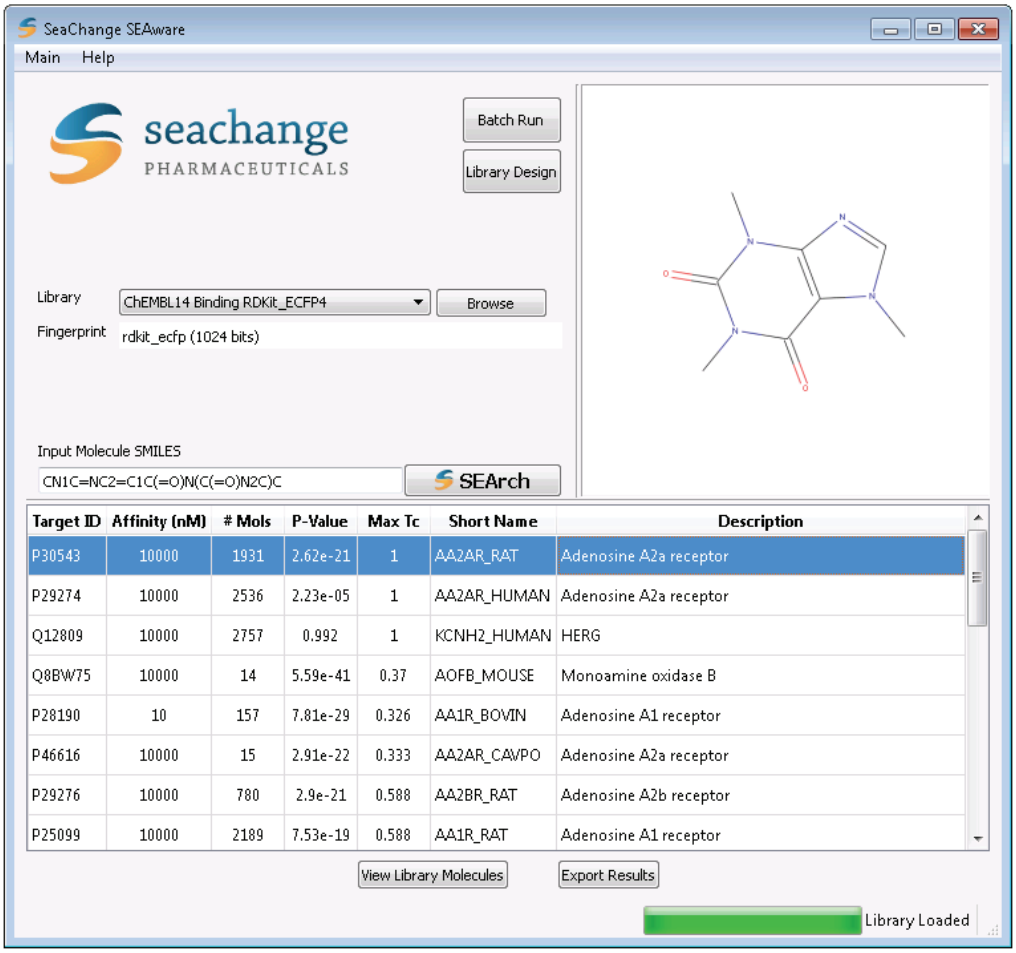


Figure 2 Targets predicted for caffeine.

You can save the full list of target predictions to an Excel file for later reference using Export Results. It is likewise possible to save the full list of ligands structures for any predicted target to a file by double-clicking on that target and then using Export Molecules.

2.2.b Running predictions using different SEA libraries

SEA can use any sufficiently large collection of ligand and target data as its reference panel from which to make predictions. In the SEAware package and across this manual, we refer to such a collection of targets and their ligands as a **SEA library**. Each SEA library is fully contained within a single file; which is identified by a `.sea` file extension.

SEAware is distributed with basic library files prepared from the publicly-accessible ChEMBL database (https://www.ebi.ac.uk/chembl). These libraries are enough to get started with SEA predictions across more than two thousand known therapeutic and molecular targets, leveraging over a quarter million ligands. These libraries are typically provided in two fingerprint formats, the impact of which is described in section 2.2.b.ii below.

2.2.b.i Changing and adding SEA libraries

To change your library, simply select a different one from among those provided in the drop-down box identified by "Library." To add a new library file (e.g., after downloading a ChEMBL SEA library update from the http://www.seachangepharma.com site), simply Browse to its `.sea` file. SEAview will keep the new library in its list of active libraries as long as the program is open.

Note: If you would like SEAview to "remember" a new SEA library across multiple openings and closings of the program, you should move or save the new `.sea` library file into the default library folder on your computer. This is the default folder that Browse opens on. In Windows 7 for instance, this location is typically:

`C:\Users\YOUR-WINDOWS-USERNAME\AppData\Roaming\SeaChange SEAware\data`

2.2.b.ii Molecular fingerprints (aka descriptors)

The molecular fingerprint (or descriptor) is the way that we computationally encode a small molecule for pairwise comparisons. The fingerprint method used affects how similar we consider any two small molecules to be to each other in chemical space, and by using different fingerprint methods, we can sometimes calculate entirely different SEA predictions, depending on the targets and drugs in question.

For this reason, SEAware is distributed with native support for two different fingerprint methods, both of which are implemented by RDKit (http://rdkit.org):

rdkit_ecfp. The default is the ECFP_4 (extended connectivity finger print) descriptor, which encodes 2D atom environments within a molecule via concentric expanding rings (4 = up to 4 bonds out). This descriptor has been shown to outperform other 2D chemoinformatic fingerprints on average (e.g., see Hert et al, *Org Biomol Chem* 2004, PMID: 15534703), as we have also observed.

rdkit_path. To complement the ECFP_4 descriptor, we also include a path-based option. Like ECFP_4, RDKit's path-based fingerprint is information-theoretic (the individual bit locations do not correspond to specific chemical moieties or single patterns). These path-based fingerprints are similar in their organization to those used by Daylight (http://www.daylight.com). In our experience, SEA typically yields fewer predicted targets when using rdkit_path fingerprints than it does when using rdkit_ecfp fingerprints, at least with the fingerprint parameters that we have explored so far.

For a more complete treatment of fingerprints and virtual screening, see section "6.1.2 Fingerprints" of the Daylight Manual at http://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

2.3 Running SEAview on many drugs at once (batch mode)

Note: This section describes a tier-2 feature, which requires a SEAware Pro or above license.

SEAview has been optimized for rapid, large-scale calculation. It is possible to run thousands, tens of thousands, or even hundreds of thousands of drugs and drug-like compounds through SEA at once. This is what the Batch Run feature enables.

In batch mode, you prepare all of your individual drugs or compounds into a single "Query Compounds" file, which is the input. Then:

1. You then select where to save the results file, which will be written in the Excel-compatible `.csv` format.
2. If your Query Compounds file is in `sd` format, you can optionally specify which `sd` field to use as the compound ID (if left blank, SEAview will use each entry's default title).
3. If your input file already contains precalculated fingerprints (supported in `.csv` input format only) and they are compatible with your currently selected library (see section 2.2.b.i above), you can select "Read from Query File" from the dropdown box to use them. By default, SEAview will instead calculate fingerprints for all the compounds in your query file to match those used in the currently selected SEA library.
    a. Note: If SEAview cannot calculate fingerprints to match (e.g., the current library is one you have built using custom fingerprints; see section 2.4 below), it will inform you that it cannot proceed.
4. Press **Run** to begin the batch calculation; progress will be shown by the bar at the bottom. When the run is complete, any relevant messages, warnings, or errors will be reported in the Log pane of the window (**Figure 3** below).

2.3.a Compound file formats used by SEAview

SEAview's batch mode can read compound query input files in several file formats (**Table 2** below). Optionally, SEAview can natively read compressed versions of any of these file formats, as long as they have been compressed using gzip (http://www.gzip.org) or bzip2 (http://www.bzip.org).

Format	Extensions	Description
SMILES	`.smi`; `.ism`; `.txt`	Standard SMILES format. Each line of the file should contain a compound's SMILES. It may optionally contain a unique ID for that compound, separated from the SMILES by a semicolon or whitespace (space, tab, etc). If no compound ID is provided, SEAview will fall back to identifying each compound by its line number in the file.
SD	`.sdf`; `.mol`	Standard SD format. You may optionally specify which field should be used as the compound ID's. All other fields will be ignored.
CSV	`.csv`	Comma-separated value format. It can be built by Excel (using File > Save As > Text CSV). At its simplest, this file can contain a header row, then the compound ID in the first column and the compound SMILES in the second column. This file format also allows you to import your own fingerprints, which is described in greater depth in section 2.3.b below.

Table 2 Input compound-file formats recognized by SEAview batch mode.

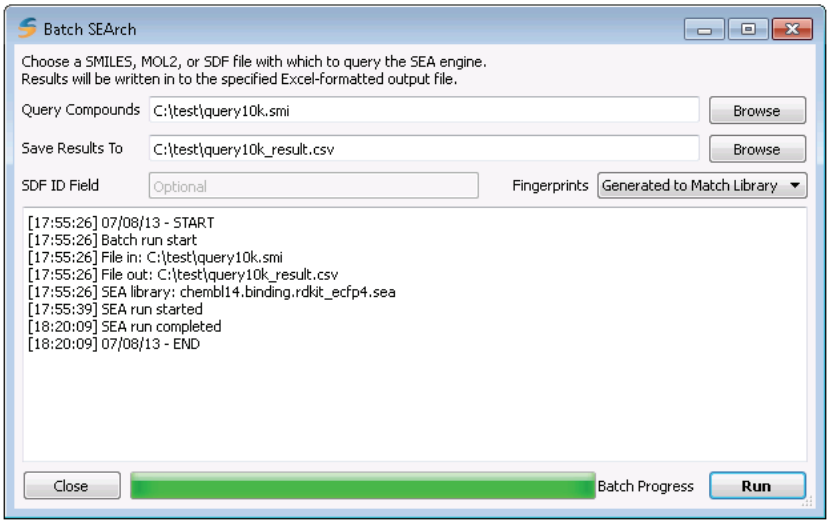


Figure 3 Completed SEAview batch run on 10,000 compounds at once, with final log data displayed.

2.3.b Using SEAview with custom molecule fingerprints

The SEAware package natively calculates two molecular fingerprint methods, described in section 2.2.b.ii above. However, you may wish instead to use fingerprints from your own preferred chemical information software packages, external to SEAware. To do so, you must prepare a compound query file in `csv` format as described in the following subsection 2.3.b.i.

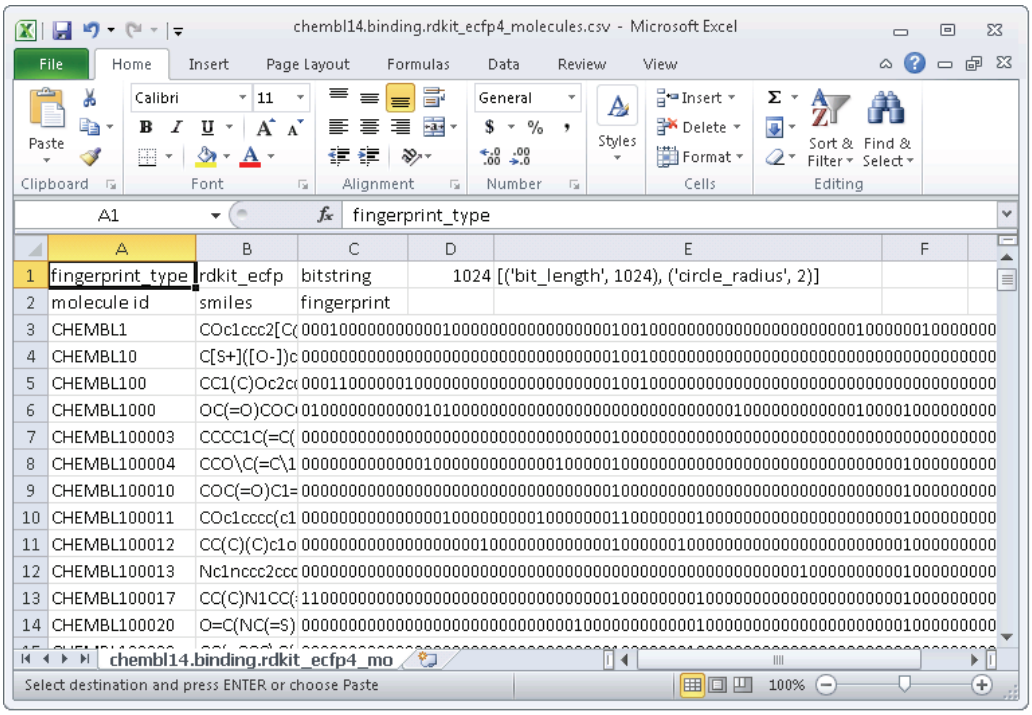
Note: Additionally, the SEA library you are using **must** have been prepared with the same fingerprint method; otherwise SEAview will detect the fingerprint mismatch and abort. See section 2.2.b.i above to import an externally-provided SEA library, or section 2.4 below to design and build your own.

2.3.b.i Preparing a compound batch file containing custom fingerprints

To use your own fingerprints (aka molecular descriptors) with SEAview, you must insert them into a `molecules-CSV` file that follows SEAview's special formatting. You can do this by writing your own custom conversion scripts to write a formatted `CSV` (**File 1** below) or by hand in Microsoft Excel (**Figure 4** below):

```
fingerprint_type,rdkit_ecfp,bitstring,1024,"[('bit_length', 1024), ('circle_radius', 2)]"
molecule id,smiles,fingerprint
CHEMBL1,COc1ccc2[C@@H]3[C@H](COc2c1)C(C)(C)OC4=C3C(=O)C(=O)C5=C4OC(C)(C)[C@@H]6COc7cc(OC
)ccc7[C@H]56,000100000000000010000000000000000010010000000000000000000000001000000100000000
000000000000000100000000000000000000000000000000000000000000000000000000000000000000000000
100000000000000000000000000000000000000000000000000000000000000000000000000001000000010000
000100000000000000000000000000000000000000000000000000000000100000001000000000000000000000
000000000000100000000001000000000000000000000000000000000100000000000000000000000000001000
000000001000000000010000000000000000000000000000000000000000000000000000000000001000000100
000000000000000000000000000000000000100000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000100000110000000000000000000000000000000000000010010
000000000000000000100000000100000000000000000000000000000000000000000000000000000010000000
000000000000000000000000000000000000000000000000000001000000010000000000000000000000000100
000100000010000000000100000000000000000000000000010000000000000000000100010000000000000000
0000000000000001000000001000000000000000000100000000010000
CHEMBL10,C[S+]([O-])c1ccc(cc1)c2nc(c3ccc(F)cc3)c([nH]2)c4ccncc4,000000000000000000000000000000000
100100000000000000000000000000000000000000000000000000000010000000000000000000000000000100
000000000000001000000000000000000010000000000000000000000000000001000000000010000000000000000
000000000000000000000000000000000000000000000010000000000000000000000000000000000000000000
000000000000000000000001000000000000000000000000000000000100000100000000000000010000100100
000001000000000000000000000000000000000001000000000000000000100000000000000000000001000000
100000000000000000000000000000000000000000000000000000000000000000000000000100000000000000
000000000000100000100000000000000000000000000000000000010000000000000000001000001000000000
000001000000000000000100000000000000000000000100000000000000000100001000000000000000000000
000000000000000000000000000000000000000000000000000000100000000000000000000000000000000100
000000001000000100000000000000000000000000000000000000000000100000001000000000000000000000
000000000000000000100000000000000000000000000000000000000000000000000000000000000000000000
000000000000
```

File 1 Example `molecules-CSV` file containing two header rows (in **bold**) followed by two data rows. Rows are shown with alternating background shading for clarity.



	A	B	C	D	E	F
1	fingerprint_type	rdkit_ecfp	bitstring	1024	[('bit_length', 1024), ('circle_radius', 2)]	
2	molecule id	smiles	fingerprint			
3	CHEMBL1	COc1ccc2[C(	0001000000000001000000000000000001001000000000000000000000010000010000000			
4	CHEMBL10	C[S+]([O-])c	0000000000000000000000000000000001001000000000000000000000000000000000000			
5	CHEMBL100	CC1(C)Oc2c(	0001100000010000000000000000000001001000000000000000000000000000000000000			
6	CHEMBL1000	OC(=O)COC(	0100000000001010000000000000000000000000000001000000000001000010000000000			
7	CHEMBL100003	CCCC1C(=C(	0000000000000000000000000000000001000000000000000000000000000000000000000			
8	CHEMBL100004	CCO\C(=C\1	0000000000001000000000000010000010000000000000000000000000000010000000000			
9	CHEMBL100010	COC(=O)C1=	0000000000000000000000000000000001000000000000000000000000000010000000000			
10	CHEMBL100011	COc1cccc(c1	0000000000000001000000001000000011000000010000000000000000000010000000000			
11	CHEMBL100012	CC(C)(C)c1o	0000000000000000001000000000000010000001000000000000000000000010000000000			
12	CHEMBL100013	Nc1nccc2ccc	0000000000000000000000000000000000000000000000000000000010000000001000000000			
13	CHEMBL100017	CC(C)N1CC(=	1100000000000000000000000000000001000000001000000000000000000001000000000			
14	CHEMBL100020	O=C(NC(=S)	0000000000000000000000000000001000000000000010000000000000000001000000000			

Figure 4 Example **`molecules-CSV`** file from **File 1** above with additional data rows, here shown in Excel.
Be sure to Save As **`CSV (Comma delimited) *.csv`**

As illustrated in **File 1** and **Figure 4** above, the **`molecules-CSV`** format contains two mandatory header rows, followed by a data row for each molecule. The formatting for these rows is further described in **Table 3**:

Row#	Col#	Value	Notes
1	1	**`fingerprint_type`**	Header formatting flag. Do not change this text.
1	2	`fingerprint-method`	Enter a brief, unique name to identify which fingerprinting method you have used. SEAview uses **`rdkit_ecfp`** and **`rdkit_path`** for its auto-generated fingerprints; these names are reserved.
1	3	**`bitstring`** or **`sea_native`**	This is the fingerprint format; it must be one of two values: **`bitstring`**: Fixed-length binary ("10001…") representation of fingerprint, as shown in **File 1** and **Figure 4**. **`sea_native`**: Fixed-length compressed Daylight base64 (hexadecimal) representation of fingerprint. See section 2.4.b below for an example of this format.
1	4	`fingerprint-length`	Enter the integer length of the fingerprint. For **`bitstring`** format, it is the length of the raw "100011…" binary character string. For **`sea_native`** format, it is the length of the *unpacked* fingerprint measured in double-bytes once it is loaded into memory.

1	5	additional-parameters or "[]"	Enter any additional parameters used to generate the fingerprint. Along with the **fingerprint_type** field, SEAview checks these parameters to confirm that the provided fingerprints are compatible with those in a SEA library. **Note:** If you do not wish to specify any additional parameters here, the field should contain the empty flag: "[]"
2	1	**molecule_id**	Default header.
2	2	**smiles**	Default header.
2	3	**fingerprint**	Default header.
data	1	molecule-ID	Data row. Enter a brief unique ID for the current molecule.
data	2	SMILES	Data row. Enter current molecule's structure in SMILES format.
data	3	raw-fingerprint	Data row. Enter current molecule's fingerprint, in **bitstring** or **sea_native** format matching that defined in the header (row 1, col 3), of fixed-length matching that defined in the header (row 1, col 4).

Table 3 Definition of rows and columns used in the **molecules-CSV** file format.

Once you have prepared your molecules into this **molecules-CSV** file format, you can input them to SEAview's batch mode, as described in section 2.3 above. Please note that the SEA library you use must contain fingerprints generated using methods and parameters exactly matching those that you have used in this **molecules-CSV** formatted compound input file, because it would not be meaningful to try to calculate SEA predictions between molecules and libraries that use different (and therefore incompatible) fingerprints. SEAview will attempt to detect this ahead of time if this occurs, and abort any calculations with an error reported in the log file.

2.4 Designing your own libraries for SEAview

Note: This section describes a tier-3 feature, which requires a SEAware Pro Designer license.

We distribute SEAview with several basic default SEA libraries that are derived from the public ChEMBL database; these are described in section 2.2.b above. However, you may wish to augment or replace these default libraries with custom libraries built instead from your own internal datasets. This is a way to leverage data from internal screening campaigns, proprietary databases, and novel assay results.

In our experience, building your own custom SEA libraries is especially powerful for improving SEA prediction accuracy when the compounds that you are investigating have been specifically designed *away* from those known in the public literature.

This section describes SEAview's ability to build, modify, and export information from SEA libraries when you are using proprietary target, assay, and/or ligand data. You can access all of these features via the Library Design button from the main SEAview window.

2.4.a Viewing library information

The main Library Design SEAview window shows a summary of all information stored in the currently-selected SEA library (**Figure 5**).

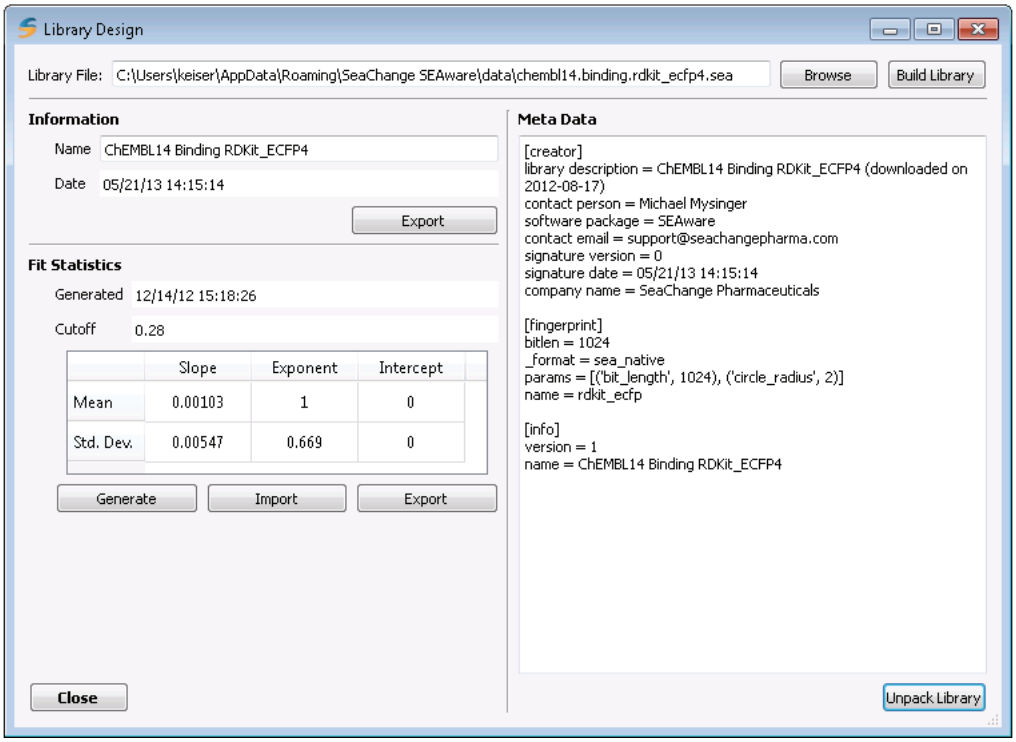


Figure 5 Main library design window in SEAview.

- **Meta Data**. This pane displays which fingerprint the library uses, as well as the contact information for the library's creator. All meta data can be exported into a text file if desired, using Export under **Information**. To export the actual targets and ligand data stored in the library, use Unpack Library (section 2.4.b below). To begin the process of creating an entirely new library, use Build Library (section 2.4.c below).
- **Fit Statistics**. This pane displays the current statistical SEA background model (aka the "fit") that has been calculated and stored within the current SEA library. Fits can be exported to and imported from simple text files in a standardized format using the appropriate buttons. To create a new fit, use Generate (section 2.4.d below).

2.4.b Exporting an existing library (library unpacking)

You can extract all of the ligand and target data contained within any SEA library into standardized Excel (or script) readable flat files; we call this process library "unpacking." To do so, Unpack Library from library design to display the Unpack Library window (**Figure 6**), specify the destination files and fingerprint format, then **Unpack to Files**.

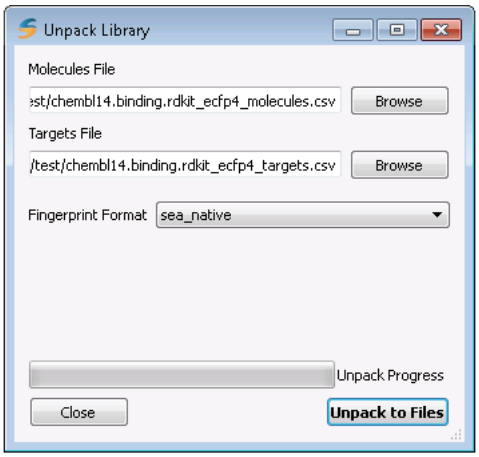


Figure 6 Library unpack window.

- **Molecules File**. This file also includes all ligand fingerprints. It is written in **`molecules-CSV`** format (as defined in section 2.3.b.i above). The fingerprint format corresponds to those listed in **Table 3**.
- **Targets File**. The targets file is written in SEAview's **`targets-CSV`** format, which can be read or written via Microsoft Excel, or via your own custom scripts. This is the primary format that SEAview uses to import/export target (or assay) data for use with its SEA libraries. An example of this format is shown in **File 2** and **Figure 7** below.

The process to import target and molecule data from these file format into a new custom SEA library is described later, in section 2.4.c.

target id,name,affinity,molecules,description
A1Z199,A1Z199_HUMAN,1000,CHEMBL1241688:CHEMBL1241689:CHEMBL1242391:CHEMBL1242482:CHEMBL1242484:CHEMBL1242581:CHEMBL483847:CHEMBL941,BCR/ABL p210 fusion protein
A1Z199,A1Z199_HUMAN,10000,CHEMBL1240546:CHEMBL1241498:CHEMBL1241499:CHEMBL1241591:CHEMBL1241592:CHEMBL1241688:CHEMBL1241689:CHEMBL1242298:CHEMBL1242389:CHEMBL1242390:CHEMBL1242391:CHEMBL1242482:CHEMBL1242483:CHEMBL1242484:CHEMBL1242579:CHEMBL1242580:CHEMBL1242581:CHEMBL1242669:CHEMBL1242670:CHEMBL1242760:CHEMBL1242857:CHEMBL1242858:CHEMBL483847:CHEMBL941,BCR/ABL p210 fusion protein
A1ZA98,A1ZA98_DROME,100,CHEMBL153015:CHEMBL153080:CHEMBL155513:CHEMBL155514:CHEMBL421992:CHEMBL440542,CG8425-PA [Drosophila melanogaster]
A1ZA98,A1ZA98_DROME,1000,CHEMBL153015:CHEMBL153080:CHEMBL153177:CHEMBL155513:CHEMBL155514:CHEMBL155621:CHEMBL356917:CHEMBL421992:CHEMBL440542,CG8425-PA [Drosophila melanogaster]
A1ZA98,A1ZA98_DROME,10000,CHEMBL152698:CHEMBL152843:CHEMBL153015:CHEMBL153080:CHEMBL153177:CHEMBL155396:CHEMBL155513:CHEMBL155514:CHEMBL155621:CHEMBL356917:CHEMBL421992:CHEMBL440542,CG8425-PA [Drosophila melanogaster]
A2AIR5,A2AIR5_MOUSE,100,CHEMBL22304:CHEMBL275528:CHEMBL284237:CHEMBL39664:CHEMBL43336:CHEMBL44073,Glutamate [NMDA] receptor subunit 3A
A2AIR5,A2AIR5_MOUSE,1000,CHEMBL22304:CHEMBL274440:CHEMBL275528:CHEMBL284237:CHEMBL289599:CHEMBL290048:CHEMBL37852:CHEMBL39664:CHEMBL40024:CHEMBL43336:CHEMBL44073,Glutamate [NMDA] receptor subunit 3A

File 2 Example **`targets-CSV`** file containing the header row (in **bold**) followed by seven data rows. Rows are shown with alternating background shading for clarity.

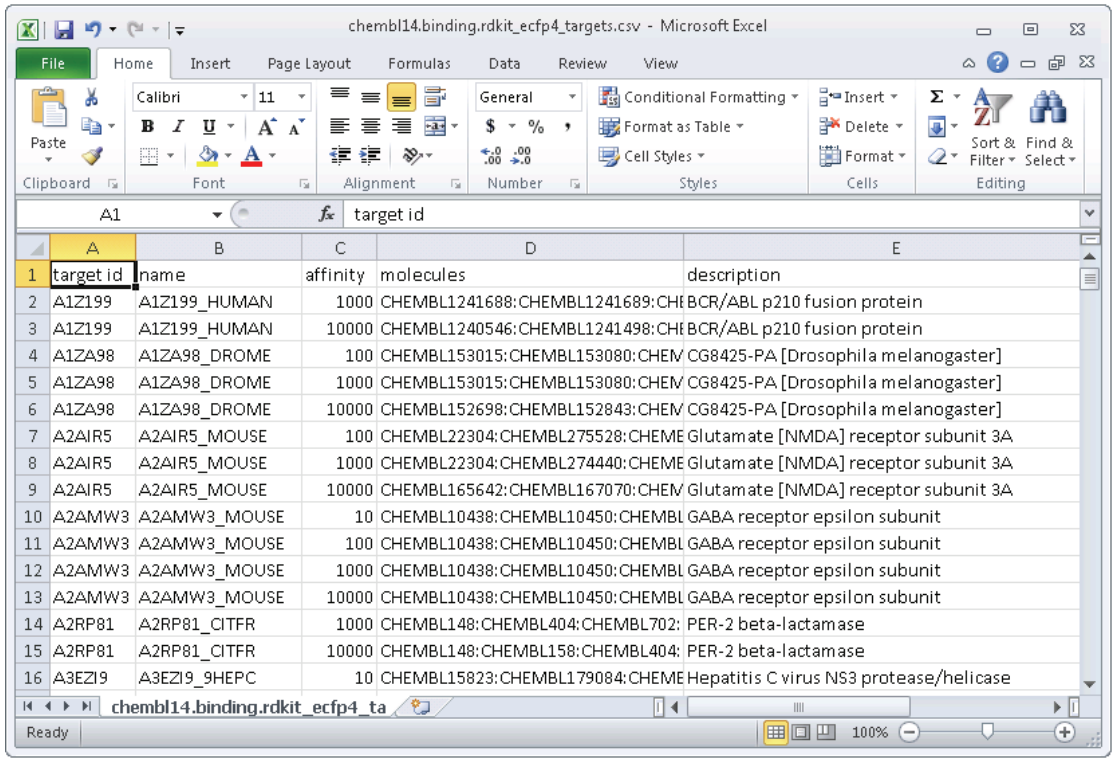


Figure 7 Example `targets-CSV` file from **File 2** above with additional data rows, here shown in Excel. Be sure to Save As `CSV (Comma delimited) *.csv`

As illustrated in **File 2** and **Figure 7** above, the `targets-CSV` format contains a mandatory header row, followed by a data row for each target (or assay). The formatting for these rows is further described in **Table 4**:

Row#	Col#	Value	Notes
1	1	`target_id`	Default header.
1	2	`name`	Default header.
1	3	`affinity`	Default header.
1	4	`molecules`	Default header.
1	5	`description`	Default header.
data	1	target-ID	Data row. Enter a brief ID for the target. Protein accession numbers work well for protein molecular targets, but this could also be an internal assay ID or other brief identifier.
data	2	target-name	Data row. Enter a brief human-readable ID for the target. This can be identical to the target-ID column, or for instance contain a UNIPROT (if target ID was a protein accession number, etc.).

data	3	`target-affinity` or **`None`**	Data row. Enter an optional subgrouping of the target. In the most common usage, this is an affinity "bin," where each bin comprises the set of ligands reported with that log-order affinity or better to the protein target. When SEAview runs predictions for a query drug against all of the targets in a given SEA library, it will report which target-subgroup (defined in this column) has the **strongest** SEA p-value for the target in question. If you do not wish to subgroup your targets, please instead fill this column with the empty flag: **`None`**. **Note:** The combination of (`target-ID` & `target-affinity`) must be **unique**, as the two together jointly serve as the unique ID for the ligand set representing this target and its subgrouping.
data	4	`target-molecules`	Data row. Enter the list of molecule IDs identifying the target's ligands. Each molecule ID should refer to a molecule entry in the corresponding **`molecules-CSV`** file (see section 2.3.b.i above). Molecule IDs in this column are separated by colons (**:**).
data	5	`target-description`	Data row. Enter a human-readable text description of the target. This will be displayed by SEAview and written to any batch output files.

Table 4 Definition of rows and columns used in the **`molecules-CSV`** file format.

2.4.c Building SEAware libraries (library packing)

This section describes the process to create new custom SEA libraries from your own target and ligand reference data; we call this process library "building" or packing.

To create a new SEA library, you need to:

1. **Prepare** the external target and ligand reference data in **`targets-CSV`** and **`molecules-CSV`** files, respectively.
2. **Build** the SEA library file from these two input files.
3. **Generate** or import a statistical background model and save it into the SEA library file.

These steps are described in the following subsections.

2.4.c.i Prepare input files

Prepare your reference target and ligand information in the required **`targets-CSV`** (**File 2** above) and **`molecules-CSV`** (**File 1** above). A good way to generate examples of these files is to try exporting one of the default ChEMBL-based libraries that we distribute with SEAview (see section 2.4.b above).

2.4.c.ii Build the SEA library file

From the main library design window, Build Library will bring up the library building window (**Figure 8** below):

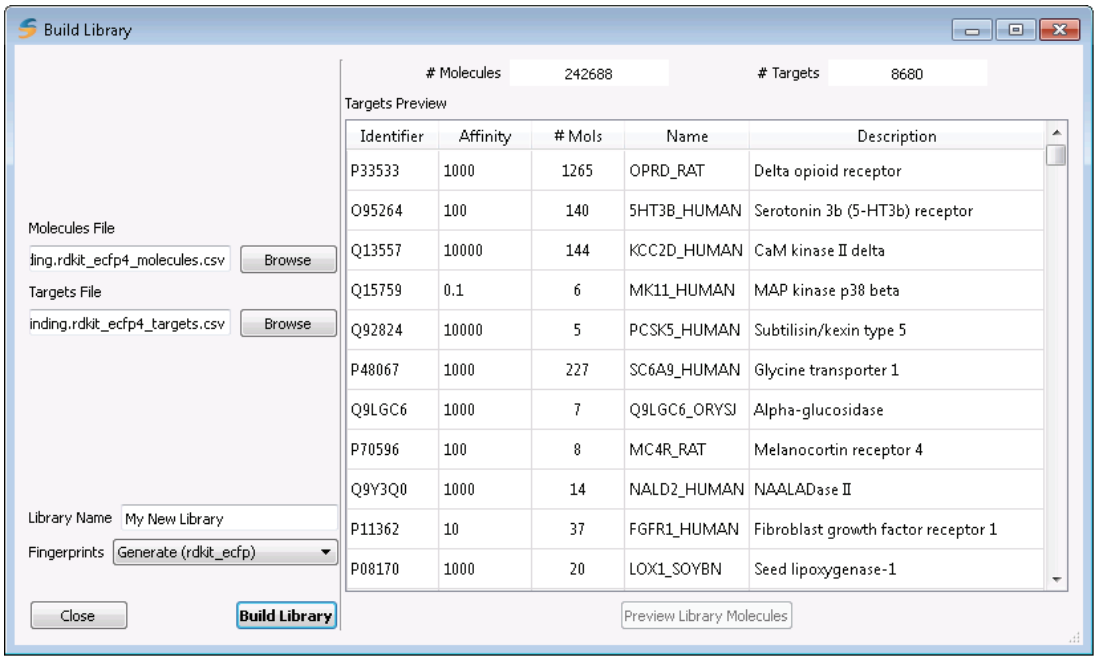


Figure 8 Build library window. A preview is automatically generated from the specified molecules and targets file. When you have confirmed that the data are being correctly read in, give the library a name, select the fingerprint method to use, and **Build Library**.

The build library window will automatically attempt to read in and provide a preview of the **`molecules-CSV`** and **`targets-CSV`** files that you provide. This gives you a chance to check for errors in the formatting of these input files, and to confirm from the molecule and target counts at the top that SEAview can successfully parse the files you have prepared. Additionally, you can double-click on any target row in the Targets Preview pane (or select the row and Preview Library Molecules) to confirm that the molecular structures for the target's ligands are correct.

After you have checked that the preview is correct, provide the following:

- Library Name. This is a human-readable name for the library as a whole.
- Fingerprints. If your **`molecules-CSV`** file already contains fingerprints, you are given the option to read these existing fingerprints from the file. Otherwise, SEAview can generate RDKit fingerprints for you (see section 2.2.b.ii above) during the library build.

Finally, Build Library and specify where the new **`.sea`** library file should be saved.

Note: As mentioned in section 2.2.b.i above, please be sure to save your new **`.sea`** library file to the default SEA libraries folder location if you would like SEAview to always automatically "remember" your new library in its dropdown list of libraries in the main SEA window. To make SEAview "forget" this library in the future, simply delete or remove its **`.sea`** library file from the default SEA library folder location.

On completion, the library build dialog will display a log (**Figure 9** below) of basic build process information along with any warnings or errors generated during the build attempt. A warning means that the process completed successfully, although it may have had to correct for a missing molecule ID or other situation for which it was able to recover and continue. An error means that the process failed and that the resulting **`.sea`** library file, if created, should be deleted and re-created after you have corrected the error.

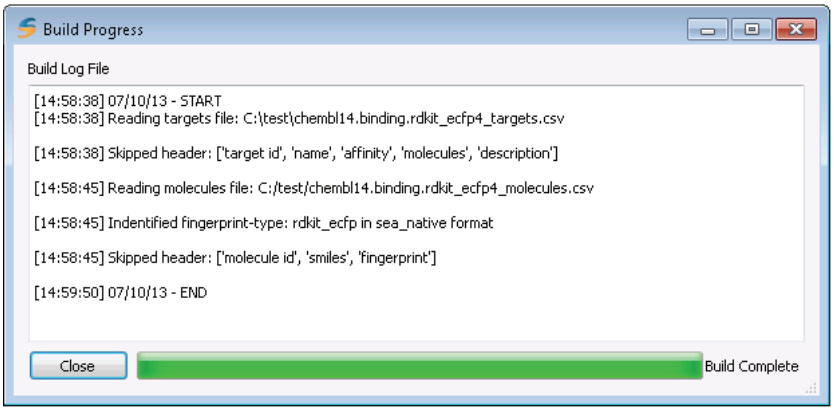


Figure 9 Example log showing successful completion of a library build, with no errors or warnings. This library was built using fingerprints already provided in the **`molecules-CSV`** input file; the log notes that it successfully identified the fingerprint-type from the file.

2.4.c.iii Generate or import a background model

SEAview cannot run any SEA calculations on your newly-created library until you have provided a statistical SEA background model (aka a "fit") for the library and saved it into the **`.sea`** library file. To import an existing SEA fit from another SEA library or a flat file, see **Fit Statistics** in section 2.4.a above. To generate a new statistical background from the ligand data in your new SEA library file, proceed to section 2.4.d below.

2.4.d Calculating background models (fit generation)

All SEA predictions are scored with respect to a statistical model for random chemical background (aka a "fit"). If your library is large enough (e.g., contains >100,000 ligands), we recommend that you build a SEA background model specific to each of your SEA libraries. This is the purpose of this section.

Note: If you are preparing a small SEA library (e.g., as a limited or focused target panel) that does not itself contain enough ligands to build a reliable statistical background, you can instead import an existing background fit from a larger SEA library (see section 2.4.c.iii above).

To calculate a new statistical background model for the current SEA library, press Generate under **Fit Statistics** in the main SEAview library design window. This brings up the library fit generation window. Start the calculation process with **Generate Background**. This is a one-time, calculation-intensive process and may take several minutes.

Figure 10 on the next page displays the fit generation window after initial background model calculations have completed.

Note: If you close this "Generate Library Fit" window before storing a fit in your SEA library via Save Fit, you will have to redo the **Generate Background** process above!

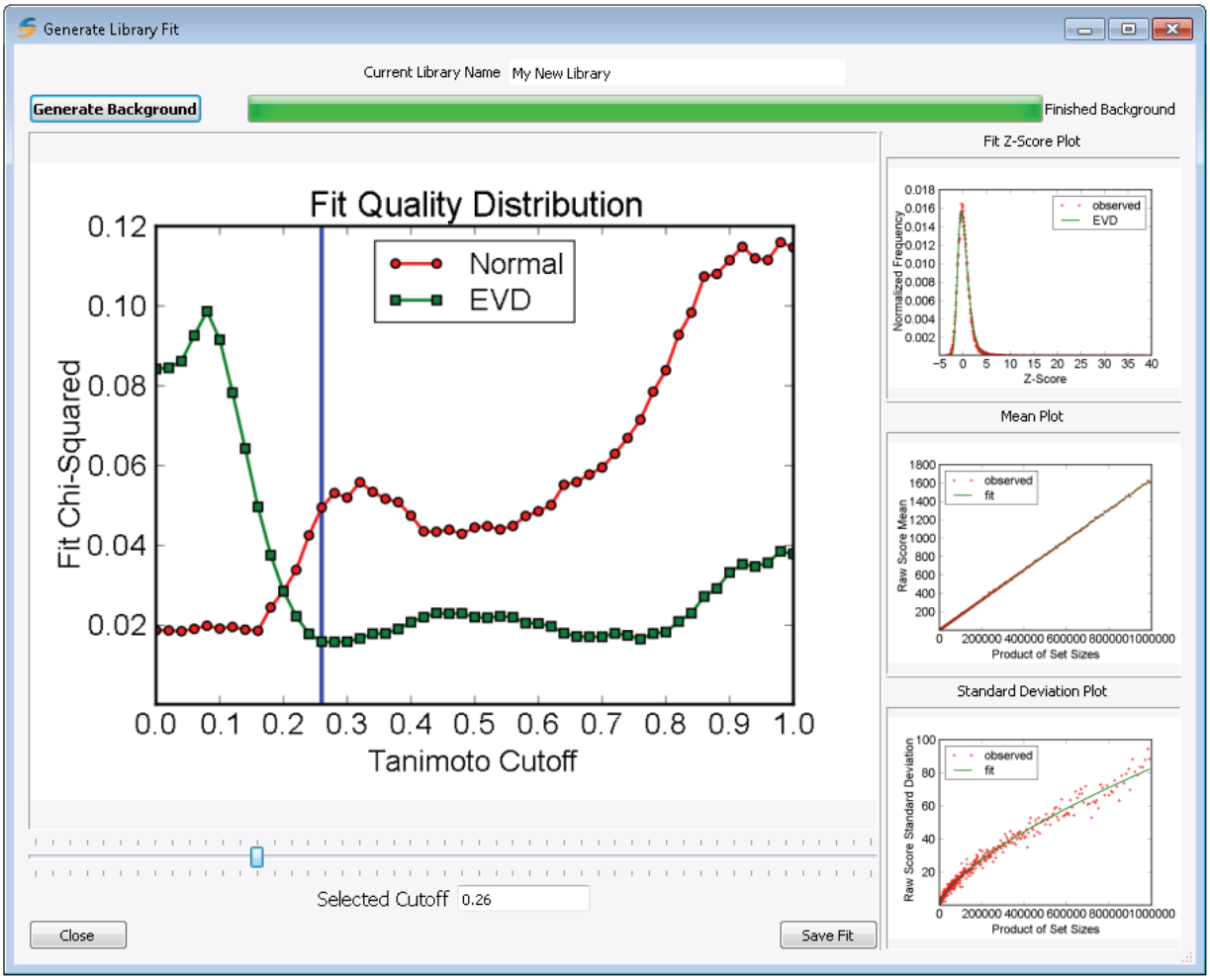


Figure 10 Example SEA background model plots at a good "Selected Cutoff" (=0.26 Tc, using the **`rdkit_ecfp`** fingerprint). The "Fit Z-Score Plot" at upper right shows a good fit to an Extreme Value Distribution (EVD), and the mean and standard deviation fits underlying the Z-score plot are also well-sampled and stable.

Selecting the Tanimoto Cutoff. After SEAview finishes the initial background model calculations, it will display a "Fit Quality Distribution" plot, which shows how well (y-axis) the model conforms to an Extreme Value Distribution (green; EVD), as compared to a Normal Distribution (red), across a range of possible Tanimoto Cutoffs (the x-axis).

A full discussion of the SEA background fitting is beyond the scope of this manual (see instead Keiser et al, *Nat Biotechnol*, 2007). For our purposes, however, a good fit is one where the EVD chi-square value (green dotted line) is small compared to the Normal chi-square value (red dotted line). SEAview attempts to suggest a good starting place for your "Selected Cutoff" based on an internal heuristic.

Note: The plots in this window are dynamic. As you use the "Selected Cutoff" slider, or directly type in a "Selected Cutoff" (which must be a number between 0.00 and 0.98, with step 0.02), the blue vertical line on the main plot will shift and all three of the smaller plots at right will be redrawn. You make any of the plots bigger by dragging its edge outward to take up more of the window.

Experimenting with different Selected Cutoffs will show you in real-time how the cutoff you choose affects underlying statistical distribution. Good cutoffs will result in a "Fit Z-Score Plot" that conforms to an EVD (as in **Figure 10**), while poor ones will devolve to a random Normal Distribution instead (as in.

Saving the fit. After you have selected a good Tanimoto Cutoff, you must save its statistics into your SEA library. To do so, press Save Fit, which will save it to the SEA library and close this "Generate Library Fit" window.

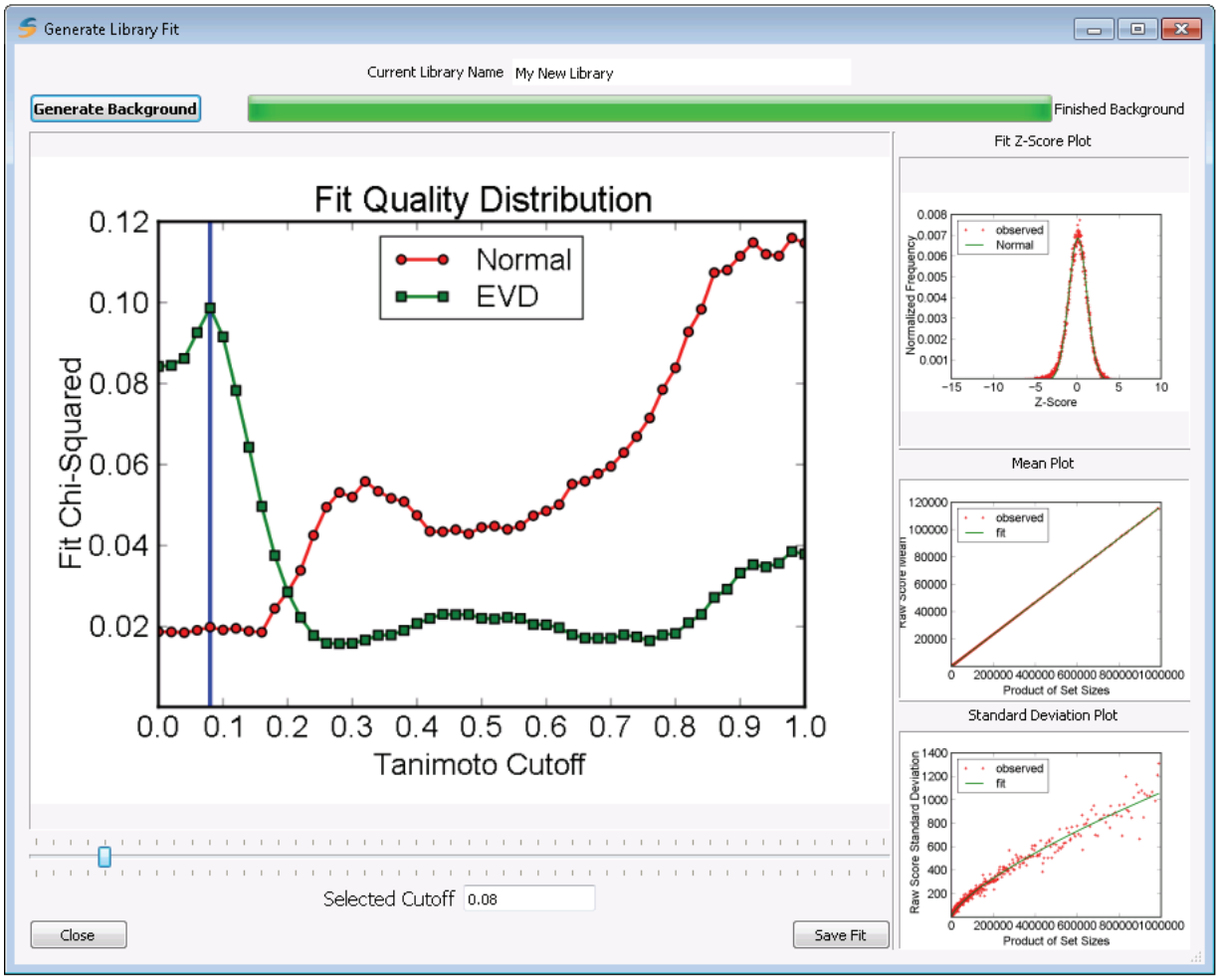


Figure 11 Plots illustrating a poor "Selected Cutoff" (=0.08 Tc; the vertical blue line). The "Fit Z-Score Plot" here favors a Normal Distribution, lacking the characteristic long "tail" of an EVD, which is achieved in **Figure 10**.

2.4.e Modifying an existing library

Sometimes you may wish to modify an existing SEA library, perhaps to add or remove a target's ligands, or even to add/remove targets entirely. The recommended way to do so is:

1. Unpack the library (section 2.4.b) and export its fit (section 2.4.a);
2. Modify the resulting **`molecules-CSV`** and **`targets-CSV`** files in Excel or by your own scripts;
3. Build a new library from the two modified **`CSV`** data files (section 2.4.c);
4. Import original or generate a new fit (section 2.4.c.iii).

Unless you are substantially changing the ligands and targets in the library (e.g., adding or removing >5% of the molecules), you do not need to generate a new fit.

3 SEAshell – a command-line SEA tool

In addition to the main SEAview application described in the previous section, SEAware is distributed with a SEAshell command-line application. SEAshell provides a "headless" command-line interface (CLI) to all of the major batch SEA prediction, molecular fingerprint import/export, and SEA-library management and design functionality present in SEAview. This makes it possible for you to integrate SEAshell with your internal discovery pipeline via batch scripting.

The quickest way to explore SEAshell commands, sub-commands, and available features is by making use of the `--help` (or `-h`) help flag:

```
C:\Program Files\SeaChange\SEAware>SEAshell.exe --help

usage: SEAshell.exe [-h] {license,fingerprint,batch,library} ...

seashell command-line interface.

optional arguments:
  -h, --help            show this help message and exit

subcommand help:
  {license,fingerprint,batch,library}
                        available subcommands.
    license             SEA license management.
    fingerprint         fingerprint conversion tools.
    batch               batch processing mode.
    library             SEA library management.
```

3.1 License management commands

Licensing: SEAshell follows the same license-based functionality as does SEAview. We describe it here.

SEAshell will intelligently provide access to commands and subcommands based on your software license level. If an entire command tree in `SEAshell.exe` is missing (such as `batch` or `library`), this is likely because your software is not licensed for these activities. At the "SEAware Primary" level (tier 1), `SEAshell.exe` provides tools for license management (section 3.1). Tier 2 ("SEAware Pro") and Tier 3 ("SEAware Pro Designer") provide more extensive `SEAshell.exe` functionality.

You can view your current license like this:

```
SEAshell.exe license display

Displaying current license information.
License File: C:/SOME/PATH/TO/YOUR/LICENSE/FILE.key
Expiration Date: YOUR-EXPIRATION-DATE
Company: YOUR-COMPANY
Contact Name: YOUR-NAME
Contact Email: YOUR-EMAIL
Comment: None
Tier: YOUR-LICENSE-TIER {1-3}
Features: 00000000-00000000
Completed.
```

The `SEAshell.exe license generate` and `set` subcommands let you create or change your current license.

3.2 SEAshell.exe tutorial

To introduce the interface, this section is a tutorial that walks through `SEAshell.exe`'s core functions.

Note: This section describes a tier-2 feature, which requires a SEAware Pro or above license.

This tutorial introduces the `library` and `batch` subcommands now available via `SEAshell.exe`. It uses the default ChEMBL-derived SEA library provided in the default library location folder (see section 2.2.b.i above) to step you through new functionality.

Note: We recommend you first make a backup of your default ChEMBL SEA library, in case you accidentally overwrite or replace important components of it via the `SEAshell.exe` commands during your testing or experimentation!

Note: You can access help for any `SEAshell.exe` command or subcommand with the `-h` flag (e.g., `SEAshell.exe -h`, `seashell library -h`, `seashell library pack -h`, etc.).

1. Copy the default ChEMBL SEA library to a testing one for this tutorial:

```
Start Menu -> SeaChange SEAware -> SEAshell
(current directory is C:\Program Files\SeaChange\SEAware)

> mkdir C:\tutorial
> copy "C:\Users\YOUR-WINDOWS-USERNAME\AppData\Roaming\SeaChange
SEAware\data\chembl14.binding.rdkit_ecfp4.sea" C:\tutorial\tutorial.sea
```

   Unpack the SEA library (this is analogous to exporting a SEAview library in section 2.4.b):

> **Note:** This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> SEAshell.exe library unpack C:\tutorial\tutorial.sea C:\tutorial\molecules.csv
C:\tutorial\targets.csv C:\tutorial\model.fit

Unpacking data from library.
Unpacking molecules.
Wrote fingerprint-type header row to: C:\tutorial\molecules.csv
Unpacking targets
Loading fit.
Completed.
```

2. You've extracted molecule, targets, and background model (fit) files from the library. The first two are in special **`csv`** formats, which are the same that you should use when preparing your own data:
   a. **`molecules-csv`** (format described in section 2.3.b.i above).
   b. **`targets-csv`** (format described in section 2.4.b above).
   c. **`model.fit`**: This flat-text file contains the parameters of the statistical model used for this SEA library. You can generate these for any library via the `SEAshell.exe library fit` subcommand (more on this later). It is identical to the fit file that you can Export using SEAview (see section 2.4.a).
3. Let's create a new SEA library from our files that uses **`rdkit_path`** fingerprints instead. But we're in a hurry and don't want to wait to fingerprint a quarter million molecules (**`rdkit_path`** is slower to generate than **`rdkit_ecfp`**). So first open up **`molecules.csv`** in Microsoft Excel (be sure to save as **`csv`**) or Notepad and delete all but the first 10,000 or so lines. Save this new file as **`molecules_10k.csv`**. Then:

> **Note:** This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> SEAshell.exe fingerprint -g rdkit_path C:\tutorial\molecules_10k.csv
C:\tutorial\molecules_10k_path.csv

Ignoring fingerprints in input file, use -f to override.
Generating rdkit_path fingerprints for 10000 molecules:

------------------ % Progress -------------------1
    1    2    3    4    5    6    7    8    9    0
----0----0----0----0----0----0----0----0----0----0
```



```
************************************************

Completed.
```

(**PS.** Your progress bar probably sat at zero for a long time, then jumped straight to 100% completed all at once. This is because we convert fingerprints in batches of 10,000 at a time, and there were only 10,000 in this file total. Bigger conversion runs will show incremental progress as each batch of 10,000 completes.)

4. Now that the new fingerprints are generated, let's convert them to **bitstring** format (they are currently in **sea_native** format because we left the generate command above on its default). We can convert as follows:

```
> SEAshell.exe fingerprint -f bitstring -c C:\tutorial\molecules_10k_path.csv
C:\tutorial\molecules_10k_path_bitstring.csv

Converting 10000 fingerprints from 'sea_native' to 'bitstring' format
Completed.
```

If you open the new file, you can see that the **fingerprint_type** header has changed and the fingerprints are now represented with "10100..." style binary character strings.

Note about the **fingerprint-type** header: This header is always in a standard format, as follows: **FLAG,fingerprint_name,fingerprint_file_format,bit_length,parameters**. If you wish to import your own fingerprints, you must include a header like this. The **FLAG** is always **fingerprint_type**, you can use your own unique standard names for **fingerprint_name**. And **fingerprint_file_format** should be either **sea_native** or **bitstring**. **bit_length** must match the fingerprint's actual bit length. The **parameters** field can have any parameters you wish to record--but you must always record them here exactly the same! This file format described in depth in section 2.3.a.

Important note! All of the fields in the **fingerprint-header** must *perfectly* match in order to compare fingerprints in different files or libraries.

5. Now that we have our new fingerprints, we can pack a new library (this is analogous to SEAview library building in section 2.4.c):

Note: This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> SEAshell.exe library pack C:\tutorial\tutorial_10k_library.sea
C:\tutorial\molecules_10k_path_bitstring.csv C:\tutorial\targets.csv

Packing data into library.
Logging to: C:\tutorial\tutorial_10k_library.log
Building library.
Reading targets.
Syncing.
Reading molecules.
Ignoring fingerprints in input file, use -f to override.
Generating rdkit_ecfp fingerprints for 10000 molecules:

------------------ % Progress -------------------1
    1    2    3    4    5    6    7    8    9    0
----0----0----0----0----0----0----0----0----0----0
**************************************************

Syncing.
Checking data consistency.
Note: No fit file specified, skipped.
Syncing.
Completed.
```


But wait, you might ask, why did this even work? After all, we only had 10,000 molecules (compared to ~250,000 originally from ChEMBL), but we still used the same full **targets.csv** file from ChEMBL. Weren't a lot of molecules missing?

The answer is: Yes, a lot were missing. So the `pack` command automatically pruned out all the targets or molecules references that weren't being used anymore before inserting them into the new **tutorial_10k_library.sea**. You can see a full log of what `pack` did for this here:

```
> more C:\tutorial\tutorial_10k_library.log
```

Logs are always generated for any major subcommand or operation, following the convention of **{outfile}.log**.

6. You can get some information about your new library by displaying its meta information (this is analogous to viewing library information in SEAview in section 2.4.a):

Note: This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> SEAshell.exe library display meta C:\tutorial\tutorial_10k_library.sea

Pulling meta from library.
[creator]
library description = seashell-generated
contact person = YOUR-NAME
software package = SEAware
contact email = YOUR-EMAIL
signature version = 0
signature date = 07/11/13 16:11:47
company name = YOUR-COMPANY

[fingerprint]
bitlen = 1024
_format = sea_native
params = [('bit_length', 1024), ('circle_radius', 2)]
name = rdkit_ecfp

[info]
version = 1
name = seashell-generated

Completed.
```

7. Oops! The fingerprint is wrong (we just made **rdkit_path** in **bitstring** format, right?). This is because we forgot to use the `-f` (`--fingerprints-supplied`) flag. So it generated its own (you can verify this by looking at the log). Run the command again, but with the `-f` flag this time:

Note: This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> C:\Program Files\SeaChange\SEAware>SEAshell.exe library pack
C:\tutorial\tutorial_10k_library.sea C:\tutorial\molecules_10k_path_bitstring.csv
C:\tutorial\targets.csv -f

Packing data into library.
Logging to: C:\tutorial\tutorial_10k_library.log
Building library.
> This will overwrite an existing library. Proceed? [y/N]: y
Clearing old library.
Reading targets.
Syncing.
Reading molecules.
Syncing.
```



```
Checking data consistency.
Note: No fit file specified, skipped.
Syncing.
Completed.
```

This was the same as providing fingerprints during a SEAview library build in section 2.4.c.i. Now `display` shows that we successfully imported the right fingerprints (also none were generated this time during the `pack`, above). Much better.

```
> SEAshell.exe library display meta C:\tutorial\tutorial_10k_library.sea

Pulling meta from library.
[creator]
library description = seashell-generated
contact person = YOUR-NAME
software package = SEAware
contact email = YOUR-EMAIL
signature version = 0
signature date = 07/11/13 16:29:24
company name = YOUR-COMPANY

[fingerprint]
bitlen = 2048
_format = bitstring
params = [('bit_len', 2048), ('bits_per_hash', 2), ('max_path_len', 7), ('min_bi
t_len', 2048), ('min_path_len', 1), ('target_density', 0.0), ('use_hydrogens', F
alse)]
name = rdkit_path

[info]
version = 1
name = seashell-generated

Completed.
```

8. Oh, we also forgot to include the fit file. Add it using `inject`:

Note: This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> SEAshell.exe library inject fit C:\tutorial\tutorial_10k_library.sea
C:\tutorial\model.fit

Injecting fit into library.
Completed.
```

Note: You can always export a library's fit directly to a file using the `library` subcommand with `display fit -o outfile`. This is like importing a fit file with SEAview (section 2.4.a).

9. OK, let's generate our own statistical model background fit from scratch. To do this, we'll use our copy of the original ChEMBL library that we made in Step 1 (because it has lots of targets and molecules in it).

   Note: Generating the background distribution in this step can take several minutes of calculation.

Note: This step describes a tier-3 feature, which requires a SEAware Pro Designer license.

```
> SEAshell.exe library fit C:\tutorial\tutorial.sea

Generating statistical library models.
> This will overwrite the library's existing fit. Proceed? [y/N]: y
Generating background distribution data
        Will sample 100000 random set comparisons
        Sampling product set sizes from 1 to 1000000
```



```
------------------ % Progress --------------------1
    1    2    3    4    5    6    7    8    9    0
----0----0----0----0----0----0----0----0----0----0
**************************************************

Waiting on processing queues

Fitting distributions to 51 background cutoffs

------------------ % Progress --------------------1
    1    2    3    4    5    6    7    8    9    0
----0----0----0----0----0----0----0----0----0----0
**************************************************

Distribution quality plot shown in browser window.
        Suggested cutoff from simplistic ratio analysis: 0.28
        Pick a point that minimizes EVD chi2 relative to normal chi2.

Enter your desired cutoff: .28
Fit saved to library file!
Completed.
```

The wizard displayed an image of the full distribution-fit (like the "Fit Quality Distribution" plot in **Figure 10** of the SEAview section) and asked what cutoff it should use; in the example here, we entered `0.28` (for approaches to cutoff choices, see Keiser et al, *Nat Biotech*, 2007 & Keiser et al, *Nature*, 2009). This actually matched the heuristically-generated suggestion of `0.28`; several values near this region would likely have been a good choice.

This fit wizard is analogous to the SEAview background model fit generation process described in section 2.4.d.

As before, you can display or write your new fit to file with the library subcommand's `display` command:

```
> SEAshell.exe library display fit C:\tutorial\tutorial.sea

Pulling fit from library.
# Generated 07/101/13 16:37:13
# fit: MU chisq: 50.7883 r2: 0.99988
# fit: SIGMA chisq: 70.8851 r2: 0.986831
TANI    0.28
MU      0.00103089      1       0
SIGMA   0.00601295      0.661249        0
Completed.
```

10. Now that we've explored ways to modify a library file and even generate new background fits from scratch, let's use it to for a SEA calculation predicting the targets of a large batch of 10,000 molecules at once. All of the SEA batch-prediction commands are available under the `SEAshell.exe batch` command tree, like this:

```
SEAshell.exe batch C:\tutorial\molecules_10k.csv
C:\tutorial\molecules_10k_predictions.csv --library
C:\tutorial\tutorial.sea --generate-fingerprint rdkit_path

SEA batch run started.
Precaching library: C:\tutorial\tutorial.sea

Ignoring fingerprints in input file, use -f to override.
Error: Active fingerprinter not compatible with your library.
Have fingerprint parameters changed?
```

Yikes! We're reusing the **`molecules_10k.csv`** file that we made back in step 3, and that's fine, but the problem is that we told `SEAshell.exe` to make RDKit Path fingerprints for it. But the **`tutorial.sea`** library

contains RDKit ECFP fingerprints (the default), so this wouldn't make any sense. You can't compare molecules represented by different fingerprint formats against each other. `SEAshell.exe` detected this and aborted.

Since generating new RDKit ECFP (aka `rdkit_ecfp`) fingerprints is the default action for any `SEAshell.exe batch` run, we can try again and this time just omit the `--generate-fingerprint` flag entirely (this may take 10-20 minutes depending on your computer):

```
C:\Program Files\SeaChange\SEAware>SEAshell.exe batch
C:\tutorial\molecules_10k.csv C:\tutorial\molecules_10k_predictions.csv --library
C:\tutorial\tutorial.sea

SEA batch run started.
Precaching library: C:\tutorial\tutorial.sea

Ignoring fingerprints in input file, use -f to override.
Generating rdkit_ecfp fingerprints for 10000 molecules:

------------------- % Progress --------------------1
    1    2    3    4    5    6    7    8    9    0
----0----0----0----0----0----0----0----0----0----0----0
**************************************************

Running SEA:

------------------- % Progress --------------------1
    1    2    3    4    5    6    7    8    9    0
----0----0----0----0----0----0----0----0----0----0----0
**************************************************

Completed.
```

SEA calculation complete! You can open up the resulting **`molecules_10k_predictions.csv`** output file in Excel or another program (watch out—this was a large calculation, so the results file is pretty big) to see the target predictions for all 10,000 input molecules that you just calculated.

Note: We could have saved a little time by passing the `--fingerprints-supplied` flag in the above command since **`molecules_10k.csv`** contains its own **`rdkit_ecfp`** fingerprints already. But that only works for **`molecules-CSV`** input files; if you're running batch searches on **`.SDF`** or **`.SMI`** files instead, `SEAshell.exe` has to generate molecular fingerprints on the fly like we just did.

Note: Steps 1 - 9 above explored ways to view, build, and modify your own SEA library files, which we used in our `SEAshell.exe batch` calculations. But of course you don't need to do all this every time you want to make a prediction; for instance, we could have just used one of the ready-to-go libraries that comes with SEAware instead.

This concludes the tutorial. Good luck & have fun!

4 References

These references summarize several Similarity Ensemble Approach (SEA) method details and scientific use cases.

- Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. **Nat Biotechnol**. 25 (2), 197-206 (2007).
- Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL. Predicting new molecular targets for known drugs. **Nature**. 462 (7270), 175-181 (2009).
- Degraw AJ, Keiser MJ, Ochocki JD, Shoichet BK, Distefano MD. Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs. **J Med Chem**. 53 (6), 2464-71 (2010).
- Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins J, Lavan P, Weber E, Doak AK, Côté S, Shoichet BK, Urban L. Large scale prediction and testing of drug activity on side-effect targets. **Nature**. 486 (7403), 361-7 (2012).

Website: For further information and product updates, visit http://www.seachangepharma.com

