Module 4 Background - Ligand Selection and Pre-Screen Preparation

Library Overview

As covered in Module 1, there are a number of large commercial screening libraries such as the Enamine REAL Database - "The current release of the *REAL* database comprises over 1.95 billion molecules which comply with "rule of 5" and Veber criteria: MW≤500, SlogP≤5, HBA≤10, HBD≤5, rotatable bonds≤10, and TPSA≤140." Other popular libraries are the Aldrich Market Direct Screening Collection (more info here), Molport's Screening Compound Database (more info here), and Enamine's Stock Screening Compounds Collection (more info here).

Another commonly used screening library is ZINC, which was developed at UCSF (see the most recent ZINC paper here). At zinc.docking.org you can access "a free database of commercially available compounds for virtual screening…[containing] over 230 million purchasable compounds in ready-to-dock, 3D formats."

In addition to commercial screening libraries, you can also build your own libraries through targeted enumeration. By using Pathfinder, a pseudo-retrosynthetic tool which decomposes a structure by looking for reactions to synthesize it from a library of 140 reactions, Reaction-Based Enumeration can generate synthetically-feasible libraries. You can specify which parts of the chemistry are varied and which should remain fixed, and then enumerate using 98 pre-loaded reactant libraries (you can also add your own libraries following this best practices document).

Please see the Enumeration Tools for Library Design tutorial for an overview of the different enumeration tools available within Maestro, and this paper that introduced Schrodinger's implementation of Reaction-Based Enumeration for more information.

Other large ligand libraries include:

- PubChem - a database of small molecules from the chemical and biological literature, hosted by the National Center for Biotechnology Information
- ChEMBL - a database of information about medicinal chemistry and biological activities of small molecules.
- The ZINC 15 database - a curated collection of commercially available chemical compounds prepared for virtual screening. ZINC is updated regularly and may be downloaded and used free of charge.
- Aldrich Market Select - a commercial library of over 8 million structures and 14 million in-stock products.
- MolPort - a commercial library with more than 7 million compounds purchasable from stock and over 20 million made-to-order compounds.

- WuXi LabNetwork - a commercial library with over 1.8 billion compounds
- Mcule Ultimate - a commercial library with more than 126 million compounds filtered for druglike properties

Also, please review the module 1 videos for more information about library selection.

Libraries for validation studies

The DUD-E directory is an enhanced and rebuilt version of DUD, a directory of useful decoys. DUD-E is designed to help benchmark molecular docking programs by providing challenging decoys. It contains: 22,886 active compounds and their affinities against 102 targets, an average of 224 ligands per target. 50 decoys for each active having similar physico-chemical properties but dissimilar 2-D topology.

In module 4, you will be filtering and preparing the DUD-E PLK1 set of actives and decoys for a validation study.

Library Filtering

Library filtering is an essential part of the pre-screening process. While we have previously demonstrated the value of screening large libraries, that value only exists when the compounds that are screened fall within the desired property space. If there are compounds that you would throw out anyway if you found them to be a potential hit from a screen, it makes sense to filter them out beforehand. While there is a point of view that it is worth screening compounds with liabilities since if there are found to be hits you can try to optimize it to address any concerns there might be, that is generally a much riskier and more time consuming approach.

Below you can find the breakdowns of different classification criteria that could be used for library filtering.

Near drug-like	Drug-like	Lead-like	Fragment
-1.5 ≦ AlogP ≦ 5.5	-1 ≦ AlogP ≦ 4	0 ≦ AlogP ≦ 3	Alogp ≦ 3
150 ≦ MW ≦ 575	250 ≦ MW ≦ 500	250 ≦ MW ≦ 375	MW > 110
30 < PSA < 150	50 < PSA < 130	PSA < 110	PSA ≦ 110
HBD ≦ 5	HBD ≦ 5	HBD ≦ 2	HBD ≦ 3
HBA ≦ 12	HBA ≦ 10	HBA ≦ 5`	HBA ≦ 5
RB ≦ 10	RB ≦ 10	RB ≦ 10	RB ≦ 3
NCC ≦ 3	NCC ≦ 3	NCC ≦ 1	

	NR ≧ 1
	HAC ≦ 18

**NCC, NR, and HAC correspond to the number of chiral centers, number of rings, and number of heavy atoms respectively.*

Pan-assay interference compounds (**PAINS**) are known to frequently give false positive results in biochemical assays, and are therefore frequently filtered out from screening libraries. A 2010 paper from the Walter and Eliza Hall Institute of Medical Research introduces a series of substructure filters that could be used for purging PAINS from libraries. There is, however, some controversy with the use of PAINS filters, with a group from the UNC Eshelman School of Pharmacy writing in 2017 paper that they "caution against the blind use of PAINS filters to detect and triage compounds with possible PAINS liabilities and recommend that such conclusions should be drawn only by conducting orthogonal experiments."

The final filter that we will cover is the popular Rapid Elimination of Swill (**REOS**) filter. First introduced in a 2003 paper out of Vertex, REOS combines both property-based and functional group-based filtering.

Property	Minimum	Maximum
HBD	0	5
HBA	0	10
Formal Charge	-2	+2
MW	200	500
Heavy Atoms	20	50
LogP	-2	5
Rotatable Bonds	0	8

The functional group filter includes nitro groups, long aliphatic chains, primary alkyl halides, aldehydes, peroxides and more.

Ligand Preparation

As a general rule, all ligands should be prepared before use in any virtual screening application. See the Introduction to Structure Preparation and Visualization tutorial for some more information and suggestions on how to best use LigPrep. Ligprep can process .mae, .sdf, .mol, .smi, .csv files, so it's a great way to bring in and process a 2D structure or SMILES string into

Maestro. We will be using ligand libraries consisting of SMILES strings in the Module 4 tutorials. If you are not familiar with the SMILES (.smi) format, you can read more about it here.

An important thing to remember when preparing ligands with LigPrep is that under Ionization you should choose a pH that agrees to the experimental/physiological/crystallization conditions. Additionally, when looking at enumerating stereoisomers follow the table below:

Settings	Recommendation
Retain specified chiralities (vary other chiral centers)	Use when your input is a 2D structure or SMILES string where some chiral centers have specified chirality (that you would like to retain) and others are not specified
Determine chiralities from 3D structure	Use when your input is a 3D structure and you are confident about the chirality of any chiral centers
Generate all combinations	Use when you are uncertain of the chirality of any chiral center, even if it is specified (or simply just want to enumerate the combinations)

It is worth noting that LigPrep does not output all minimized conformations of the structures (it just outputs low energy conformations). Conformational searching using Confgen is a part of the Glide docking, Shape screening, and Pharmacophore screening workflows so it is not something that needs to be done manually beforehand.

Receptor Grid Generation

Grid generation must be performed prior to running a virtual screen with Glide. The shape and properties of the receptor are represented in a grid by fields that become progressively more discriminating during the docking process. In addition to the tutorials in this Module, please see the Structure-Based Virtual Screening Using Glide tutorial for how to generate a receptor grid from a protein-ligand complex, add a hydrogen bond constraint, and how to generate a receptor grid from a SiteMap site. Please see the Advanced Settings Dialog Box for more information on how you can customize your grid.

Constraints can be used to incorporate experimental information into your docking model. Since you are able to toggle on and off constraints that are part of the receptor grid in the Ligand Docking panel, we suggest adding as many constraints into your grid as you think you might be

interested in for validating your model. That will save you time for regenerating the grid each time you think of a new constraint to add.

There are 5 different constraints that you could define through the Receptor Grid Generation panel: 1. Positional constraint 2. NOE (nuclear Overhauser effect) constraint 3. H-bond constraint 4. Metal constraint 5. Metal coordination constraint. Click here for more information about how each of these constraints are defined and can be used. Please note that core constraints, shape constraints, and torsional constraints are all defined in the Ligand Docking panel and not the Receptor grid generation panel.

All receptor grids should be thoroughly validated before being used prospectively. For a given target, you may develop several different grids (perhaps with different constraints or combinations of constraints), and as a first step you should test to see how well the cognate ligand (and if available, other actives that are known to have the same binding mode) re-docks. Additionally, you should combine a list of known actives with ~50 decoys per active (which can be generated here), dock them all, and then use the enrichment calculator to evaluate how well the different docking grids were able to separate the actives from the decoys.

Enrichment Analysis

The best way to determine the success of a retrospective virtual screen is by looking at the enrichment - i.e. the ability to separate binders from non-binders. There are many popular enrichment metrics (described in the table below) that are commonly used:

Metric	Description
Receiver Operator Characteristic (**ROC**) AUC	The value is bounded between 1 and 0, with 1 being ideal screen performance and 0.5 reflecting random behavior. The area under the curve is the probability that a randomly chosen known active will rank higher than a randomly chosen decoy. See this 2007 paper from Truchon and Bayly (equation A.8) for more information.
Robust Initial Enhancement (**RIE**)	Active ranks are weighted with a continuously decreasing exponential term. Large positive RIE values indicate better screen performance. See this 2007 paper from Truchon and Bayly (equation 18) for more information.
Boltzmann-enhanced Discrimination Receiver Operator Characteristic (**BEDROC**)	The value is bounded between 1 and 0, with 1 being ideal screen performance. The default alpha=20 weights the first ~8% of screen results. When alpha*Ra << 1, where Ra is the ratio of total actives to total ligands, and alpha

	is the exponential prefactor, the BEDROC metric takes on a probabilistic meaning. See this 2007 paper from Truchon and Bayly (equation 36) for more information.

Please see the Enrichment Calculator panel documentation for more information on the enrichment metrics that are available through Maestro.

In addition to these popular metrics, we also recommend paying attention to the hit rate. This is an easily explainable metric that is used to evaluate other types of screens and is ideally for communicating results to colleagues.

Selecting Probes and Generating a Shape Data File for Shape Screening

Below are recommendations based on extensive validation testing:

Topic	Recommendation
Ideal number of probes	10
Minimize output conformers?	No
Ideal Shape type	Pharmacophore

The best way to identify your probes would be to import the actives into Maestro, open the Canvas Similarity and Clustering panel, select the Dendritic Fingerprint type, cluster (aim for 10 clusters), and then create a group containing the structures nearing the centroid in each cluster. Make sure to also align your probes using the Ligand Alignment panel before loading them into the Shape Screening panel.

If you are running a screen using GPU Shape, you will need to first generate a Shape Data File. You can prepare the ligands and generate conformers from the Create Shape Data File panel, but if you are working off of a Phase Database, those steps are not necessary. Outside of this course, the Shape Data File generation job can be parallelized across as many CPUs as you'd like without checking out any license tokens.