MolPort Screening Compound® Phase Database

This document provides an overview of the MolPort Screening Compound Phase database distributed by Schrödinger. MolPort provides structure files for the screening deck in compressed SDF format. The SDF input is processed to generate structures suitable for property calculations and for the generation of the Phase database, as well as for the generation of the GPU shape screening bin file.

The structures are run through LigPrep in order to generate stereoisomers, as well as tautomers and charged states using Epik, including the calculation of metal-binding states. The LigPrep output is then imported into a Phase database to generate pharmacophore sites, conformers and fingerprints. Database subsets are provided for drug-like, lead-like, near-drug, and fragment structures. The LigPrep output is also used to generate the bin datafile for GPU shape screening. Updates to the distribution are provided on a quarterly basis. Included at the end of this document are the commands run to prepare the structures and to generate the Phase database and the shape screening bin file.

Table 1. Structure counts. Some structures are problematic and will not make it through the entire workflow.

Distribution	Count
Source SMI	4,973,086
2D Properties	4,972,412
3D variants	10,274,716

Classifications

Schrödinger classifies screening compounds based on properties calculated on a single, neutralized and desalted representation. Near drug-like compounds are those that fall close, but not quite into the drug-like property space. Drug-like compounds are expected to have properties similar to known marketed drugs. Lead-like compounds typically have a more restrictive set of properties that align with the goals of finding a hit and expanding that hit into a more drug-like compound. Fragments are compounds that are typically very small and are used to probe a target in order to determine the functionalities expressed by compounds that bind to a particular site. All other molecular structures are not given a classification.

Several rubrics are used in a hierarchical funnel to associate a molecular entity to a specific class. Note that even though a molecule may fit into several classifications, in practice, compounds are allowed to match and get assigned to categories further down the funnel in the order: near drug-like, drug-like, lead-like, to fragment. The count of structures in each class is summarized in table below.

Table 2. Classification criteria. These properties are discussed below. Note that NCC, NR, and HAC correspond to Num chiral centers, Num rings, and Num heavy atoms as calculated by ligfilter. Structures are assigned to a class by the last successful match on all criteria proceeding from right to left.

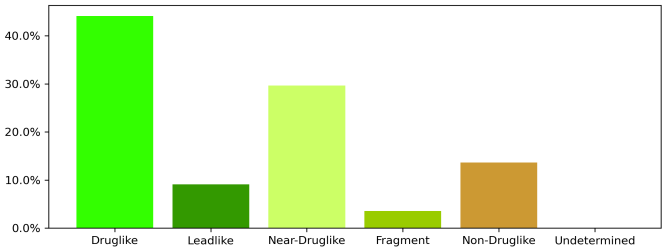
Near drug-like	Drug-like	Lead-like	Fragment
–1.5 ≤ AlogP ≤ 5.5	–1 ≤ AlogP ≤ 4	0 ≤ AlogP ≤ 3	AlogP ≤ 3
150 ≤ MW ≤ 575	250 ≤ MW ≤ 500	250 ≤ MW ≤ 375	MW > 110
30 < PSA < 150	50 < PSA < 130	PSA < 110	PSA ≤ 110
HBD ≤ 5	HBD ≤ 5	HBD ≤ 2	HBD ≤ 3
HBA ≤ 12	HBA ≤ 10	HBA ≤ 5	HBA ≤ 5
RB ≤ 10	RB ≤ 10	RB ≤ 10	RB ≤ 3
NCC ≤ 3	NCC ≤ 3	NCC ≤ 1	
			NR ≥ 1
			HAC ≤ 18

Table 3. Screening classification

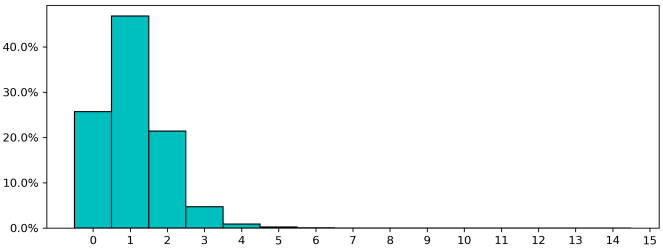
	Drug-like	Lead-like	Near drug-like	Fragment	Others
Count	2,193,461	451,568	1,474,451	175,840	677,092
Percentage	44.1%	9.1%	29.7%	3.5%	13.6%

Database Property Distribution

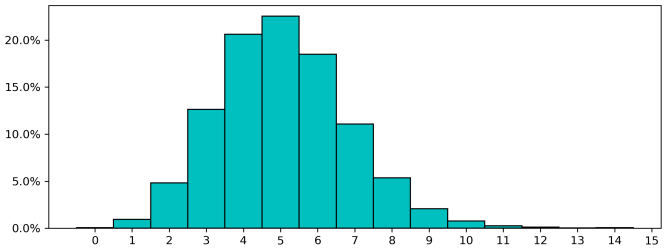
Class



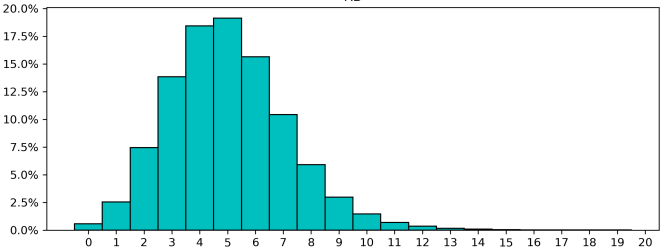
HBD



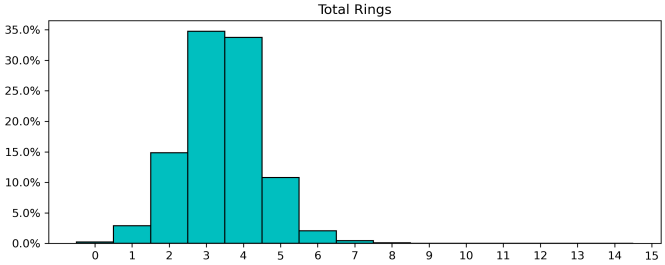
HBA



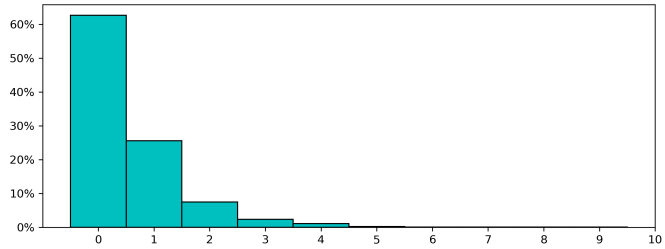
RB



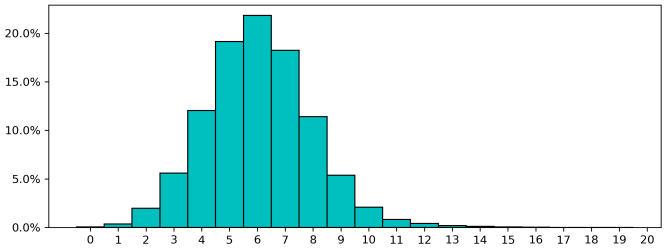
Total Rings



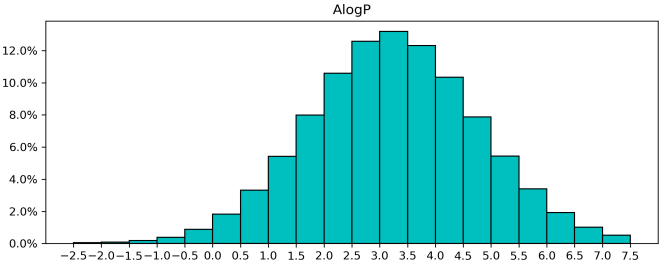
Stereo Centers



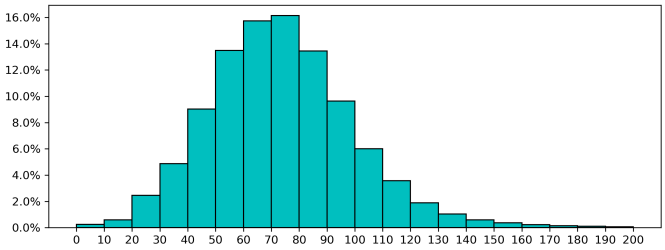
N+O



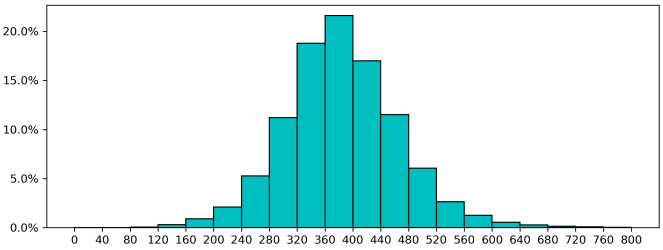
AlogP



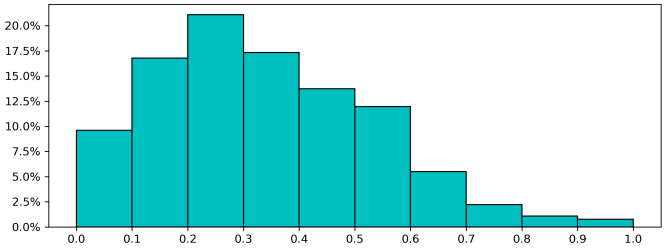
PSA



MW



Frac. CSP3



Property Description

A brief description of each property is provided below.

HBD:
Number of hydrogen bond donors

HBA:
Number of hydrogen bond acceptors

RB:
Number of rotatable bonds

Total Rings:
Number of rings

Stereo Centers:
Number of stereogenic centers

N+O:
Sum of nitrogen and oxygen atoms

AlogP:
Logarithm of the atomistic partition coefficient

PSA:
Fragment-based topological polar surface area

MW:
Molecular weight

Frac. CSP3:
Frequency of sp3-hybridized carbon atoms with respect to total carbon atom count

Schrödinger commands for creating the database

LIGPREP STRUCTURE PREPARATION
$SCHRODINGER/ligprep -epik -bff 16 -s 16 -pht 1.0 -emb -isd Molport_input.sdf.gz -omae ligprep.mae.gz

SHAPE BIN FILE CREATION
$SCHRODINGER/shape_screen_gpu generate -shape_data_treatment remote -shape_data_dir /path/2024q1_shape -source /path/ligprep.mae.gz -flex -shape_type pharm -sample thorough -limit 10 -conformer_format compact

PHASE DATABASE CREATION

Add records to the database
$SCHRODINGER/phase_database /path/2024q1_molport.phdb splice ligprep.mae.gz -new -fmt int -title s_lp_Variant -JOB 2024q1_molport_splice

Generate sites, conformers, fingerprints
$SCHRODINGER/phase_database /path/2024q1_molport.phdb revise -sites -confs all -fp -add dendritic,fdendritic,maccs,radial,ecfp4 -props -JOB 2024q1_molport_revise

Extract properties
$SCHRODINGER/phase_database /path/2024q1_molport.phdb extract -map

Generate subsets

-- Drug-like:
$SCHRODINGER/phase_database /path/2024q1_molport.phdb query subset_drug_like -where "r_canvas_AlogP >= -1.0 AND r_canvas_AlogP <= 4.0 AND r_canvas_MW >= 250.0 AND r_canvas_MW <= 500.0 AND r_canvas_PSA > 50.0 AND r_canvas_PSA < 130.0 AND i_canvas_HBD <= 5 AND i_canvas_HBA <= 10 AND i_canvas_RB <= 10 AND i_canvas_ChiralCenterCount <= 3"

-- Near-drug:
$SCHRODINGER/phase_database /path/2024q1_molport.phdb query subset_near_drug -where "r_canvas_AlogP >= -1.5 AND r_canvas_AlogP <= 5.5 AND r_canvas_MW >= 150.0 AND r_canvas_MW <= 575.0 AND r_canvas_PSA > 30.0 AND r_canvas_PSA < 150.0 AND i_canvas_HBD <= 5 AND i_canvas_HBA <= 12 AND i_canvas_RB <= 10 AND i_canvas_ChiralCenterCount <= 3"

-- Lead-like:

$SCHRODINGER/phase_database /path/2024q1_molport.phdb query subset_lead_like -where "r_canvas_AlogP >= 0.0 AND r_canvas_AlogP <= 3.0 AND r_canvas_MW >= 250.0 AND r_canvas_MW <= 375.0 AND r_canvas_PSA <= 110.0 AND i_canvas_HBD <= 2 AND i_canvas_HBA <= 5 AND i_canvas_RB <= 5 AND i_canvas_ChiralCenterCount <= 1"

-- Fragment-like:
$SCHRODINGER/phase_database /path/2024q1_molport.phdb query subset_fragment -where "r_canvas_AlogP <= 3.0 AND r_canvas_MW > 110.0 AND r_canvas_PSA <= 110.0 AND i_canvas_HBD <= 3 AND i_canvas_HBA <= 5 AND i_canvas_RB <= 3 AND i_canvas_RingCount >= 1 AND i_canvas_HeavyAtomCount <= 18"

-- Epik metal state
$SCHRODINGER/phase_database /path/2024q1_molport.phdb query subset_epik_metal_only -where "b_epik_Metal_Only == 0"