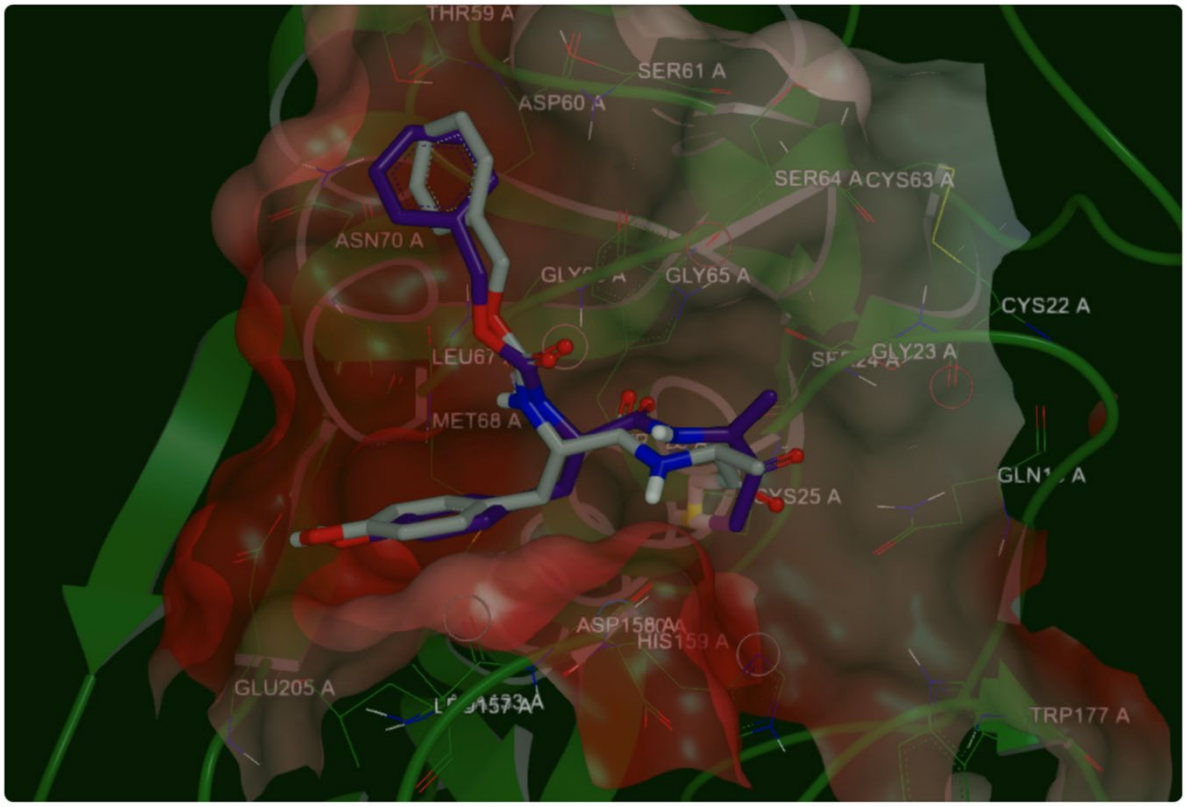
GOLD Orion® Integration User Guide



2025.1 GOLD release

www.ccdc.cam.ac.uk

GOLD Orion® Integration User Guide

For further support please contact support@ccdc.cam.ac.uk

Contents

Introduction

The GOLD Floes run a variety of docking calculations from within the Orion® platform using the CSD Ultra Large GOLD docking engine.

They accept input files generated in the Orion® platform (i.e. files prepared with SPRUCE) and input and configuration files generated locally with Hermes, and uploaded to the platform.

- GOLD Floe (Protein Input Dataset): input protein dataset prepared by SPRUCE
- GOLD Floe (Protein Input File): input protein file generated outside the Orion® platform in mol2 or pdb format

For each of these two floes there is an equivalent floe designed to run docking calculations with ultra-large libraries stored as "collections" in the Orion® platform:

- Ultra Large Docking GOLD Floe (Protein Input Dataset)
- Ultra Large Docking GOLD Floe (Protein Input File)

Please, make sure you use the most adequate floe for the type of calculation you need to run.

This implementation will not attempt to replicate all the options and functionalities available in the desktop app, but the most important ones currently available in GOLD to provide the user with the most impactful programmatic solutions to their drug discovery problems.

The list of features includes:

- Importing prepared receptors in the standard formats (pdb, mol2) and Orion® native formats (datasets)
- Selecting output names and destination folders in the Orion® platform
- Importing curated libraries of ligands from standard formats (sdf and mol2) and Orion® native formats (i.e datasets and OMEGA-generated libraries, like collections)
- Defining the binding site using a variety of methods available in GOLD and/or Orion®
- Setting the scoring function and genetic algorithm values
- Configuring and/or uploading a gold.conf file

Providing access to widely used functionalities:

- Standard docking with rigid receptor and rigid/flexible ligand
- Covalent docking with a single ligand or with multiple ligands via substructure match
- Connectivity for docking with Ultra Large Libraries

Further reading:

For more information about GOLD, please, consult the documentation:

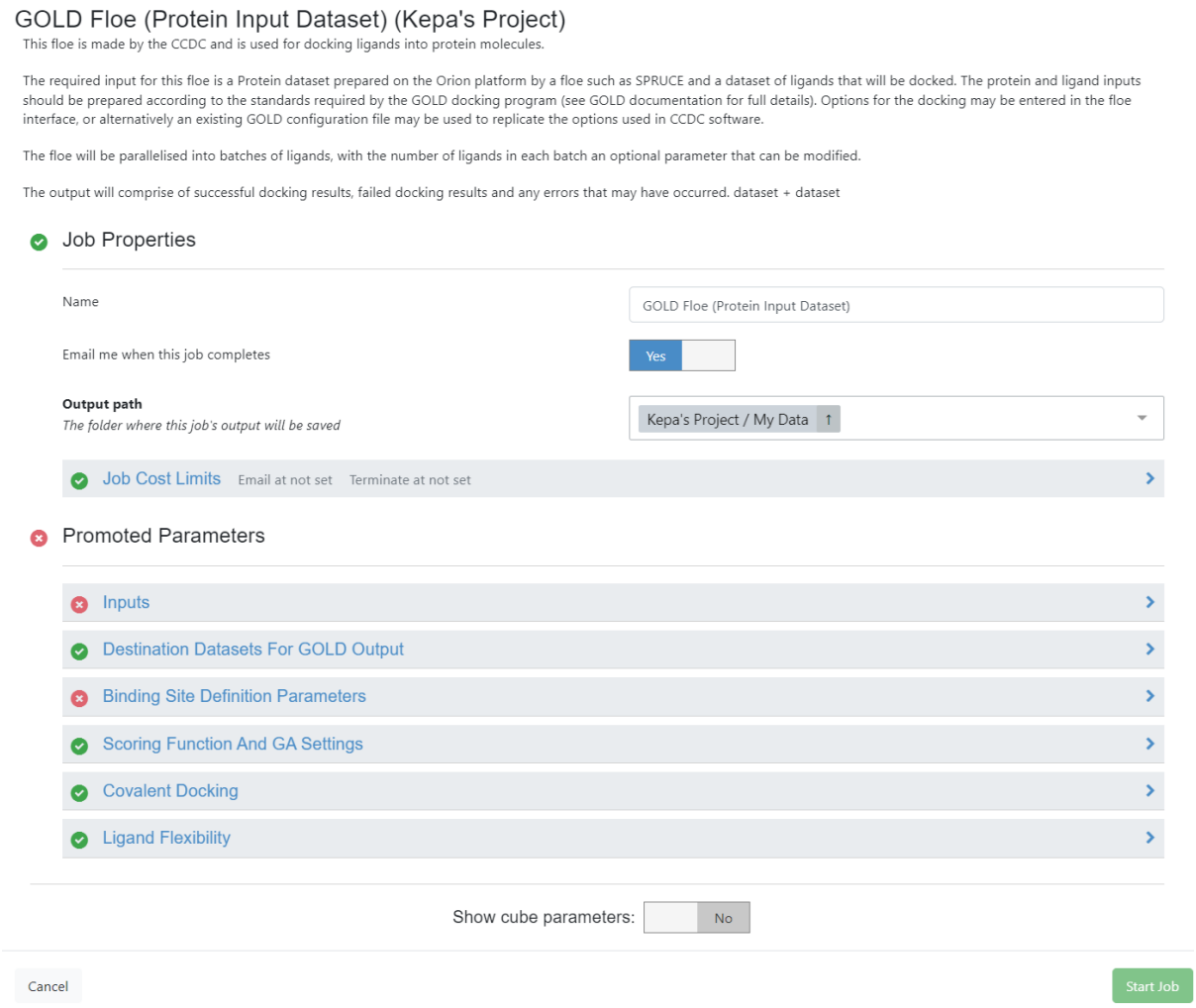
- GOLD User Guide: https://www.ccdc.cam.ac.uk/media/gold-1.pdf
- GOLD Configuration File User Guide: https://www.ccdc.cam.ac.uk/media/gold-configuration-1.pdf

The GOLD Floe Menu

The menus and submenus are concise, intuitive, and self-explanatory. They can be expanded by clicking the arrow on the right.

Parameters flagged with a cross in a red button are mandatory and users are required to introduce input values and/or files. Parameters displaying a check symbol in a green button are either optional or pre-assigned by default input values. In this case, users may not need to alter the settings to run a standard docking calculation unless they need a specific configuration.

The easiest way to set up the calculation is step by step clicking the different sections and providing the required values or accepting the by-default configuration.



Job Properties and Promoted Parameters

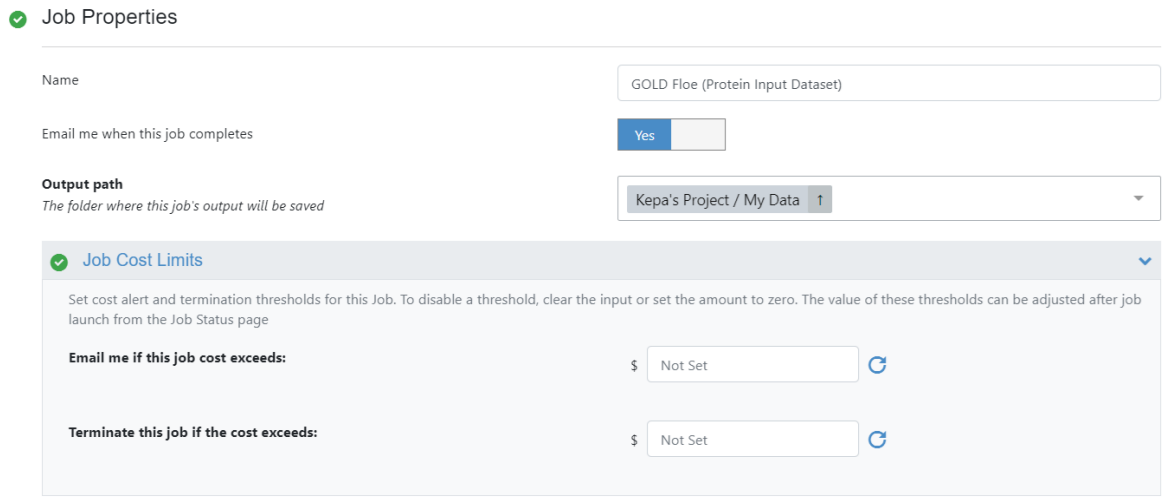
These are general options to identify the job, send alerts, and control the budget.

Name: Introduce a job name to identify the docking run.

Email: If set to yes, the Orion® platform will send the user an email when the calculation completes.

Output path: Path to a folder where all the calculation output files will be written.

Job cost limits: A practical functionality to avoid incurring in unexpected high costs during docking campaigns. If the job cost reaches the lower threshold, the Orion® platform will email the user with a warning. If the cost reaches the upper value, the platform will terminate the job.



Inputs and Outputs

All the fields in this section are mandatory.

To run a docking calculation with GOLD:

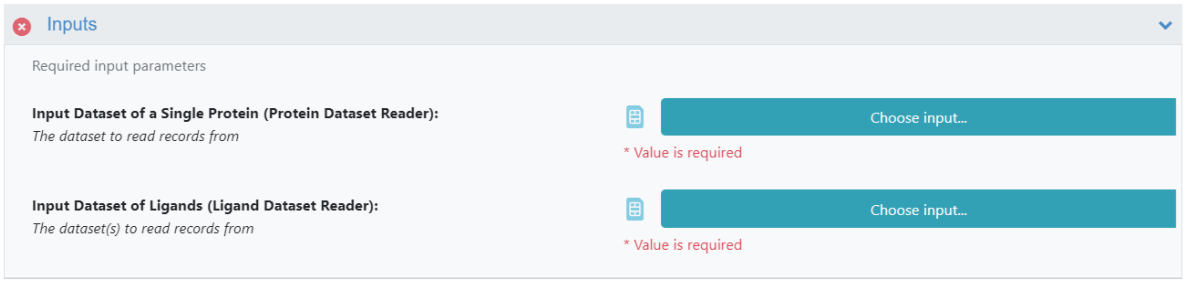
- The receptor/protein, must be prepared beforehand: add hydrogens, resolve occupancies, select tautomers and rotamers, remove unwanted waters and/or other solvents, ions, small ligands, organic ions, correct artifacts and chemical bonds.
- The libraries of ligands must be washed, sanitised, and contain 3D structures. If your input library contains SMILES strings, flat 2D structures, salts, hydrochlorides, or any type of artifact, you need to curate the library beforehand.

A good guide can be found in this recent paper: A practical guide to large-scale docking | Nature Protocols

This section allows the user to select the receptor, the library of ligands, and the output file names. The input text menus are slightly different depending on the GOLD Floe selected as they will accept different types of input files for receptors and ligands. The function, however, is the same. Please, use the recommended GOLD Floe for your combination of input files:

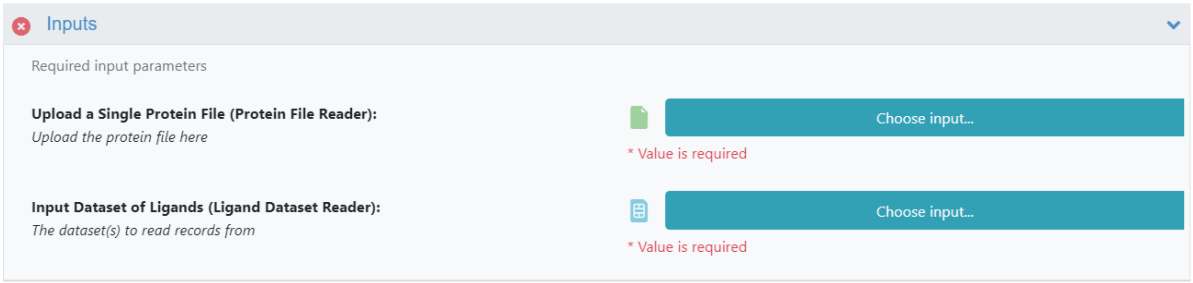
a) GOLD Floe (Protein Input Dataset)

Protein: Dataset from SPRUCE
Ligands: Dataset



b) GOLD Floe (Protein Input File)

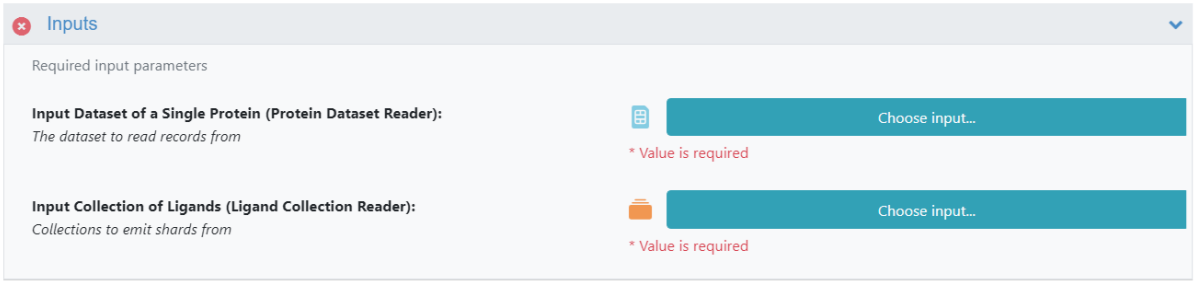
Protein: mol2 file prepared with Hermes
Ligands: Dataset



c) Ultra Large Docking GOLD Floe (Protein Input Dataset)

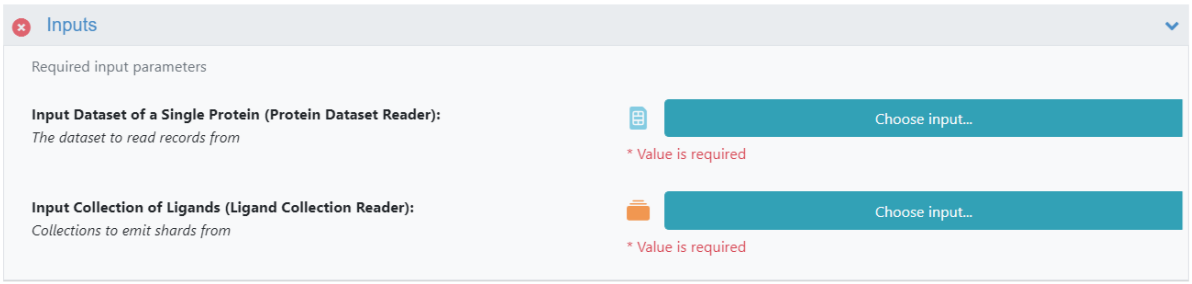
Protein: Dataset from SPRUCE

Ligands: A collection containing the Ultra Large library

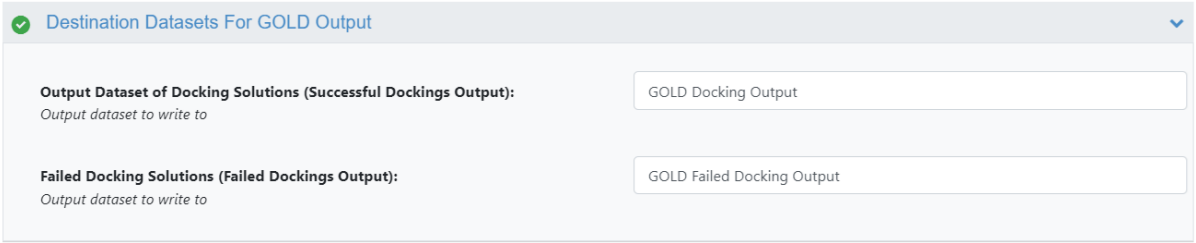


d) Ultra Large Docking GOLD Floe (Protein Input File)

Protein: mol2 file prepared with Hermes
Ligands: A collection containing the Ultra Large library



The successful docking results and the failed ligands will be saved in specific output files. These fields are mandatory.



Binding Site Definition Parameters

This section allows the user to define the binding site using the different options available.

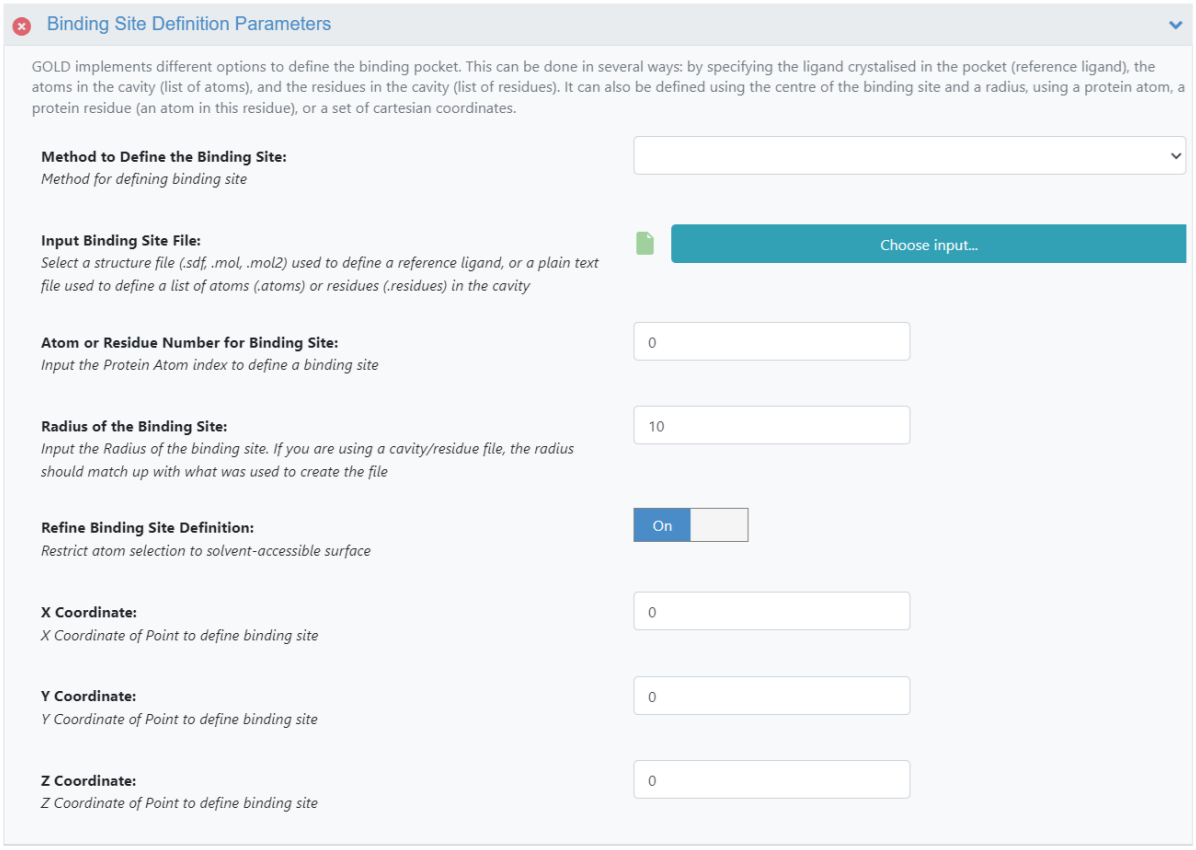
For more information about the cavity options implementation in GOLD, please, read the documentation in the GOLD User Manual, section 4.6, Defining the Binding Site: GOLD User Guide (cam.ac.uk) and the GOLD Configuration File User Manual, section 8 Flood Fill: GOLD Configuration File User Guide (cam.ac.uk)

Cavity selection is done in two steps:

- Select one of the cavity method definitions using the “Method to define the Binding Site” drop-down menu.
- Fill the boxes with the required information or path to files consistent with the cavity method selected.

ATTENTION:

- Avoid introducing conflicting or unnecessary information in the fields. For example, if the cavity method selected is “cavity from ligand”, load the reference ligand file and do not introduce any other value, for example, cartesian coordinates.
- It is strongly recommended to follow the guidance below to select the binding site. Specific types of GOLD floes have been created for different types of input files (datasets, files, and collections). Whenever possible, please, use the most appropriate cavity method.
- It is recommended to leave the “Refine Binding Site Definition” option ON (default) to allow the LIGSITE algorithm to refine the cavity.
- The radius parameter has a default value of 10 Angstrom. This is a reasonable value that provides good results in most of the cases in combination with the LIGSITE refinement.



a) Reference Ligand

If there is a ligand X-ray structure available, this is the recommended method. The reference ligand may be part of the original co-crystalized receptor-ligand Xray or Cryo-EM structure and it must be a fully sanitized 3D structure with the original cartesian coordinates. Acceptable file formats are sdf, mol2, and dataset.

- Method to define the Binding site à Reference Ligand

- Click on the “Choose input” green box next to “Input Binding Site File”, navigate to the desired location, and select the ligand file (mol2, sdf, dataset) that will be used to define the cavity

ATTENTION: This method can be used with the GOLD (Protein File) and GOLD (Protein Dataset) floes in the standard and ultra-large docking versions.

b) Cavity from Spruce protein design unit

This is the recommended method when the receptor input file is a dataset prepared with SPRUCE and the GOLD (Protein Dataset) floes are used. The object generated with SPRUCE contains cavity information and this information can be read if selected.

- Method to define de Binding Site à Cavity from Spruce protein design unit
- No further action required. The flow will read the input file generated by Spruce and extract the cavity automatically.

ATTENTION: Use the cavity from spruce option ONLY when the receptor input file is a dataset curated and generated with SPRUCE. Do not use this cavity option with other datasets or with the GOLD (Protein File) floes. Do not use “list of atoms”, “list of residues”, “protein atom”, and “protein residue”; or use with caution “cavity from ligand”, and “coordinates”, methods when the input receptor file has been generated with SPRUCE.

c) List of atoms

This file is normally created locally with Hermes when preparing a docking calculation and is saved by default with the name “cavity.atoms”. For further information, please consult the GOLD documentation: Gold User Manual, section 4.6.5 Defining a Binding Site from a List of Atoms or Residues.

- Method to define de Binding Site à List of atoms
- Click on the “Choose input” green box next to “Input Binding Site File”, navigate to the desired location, and select the “cavity.atoms” file containing the list of atoms that will be used to define the cavity.

ATTENTION: Use only the Protein Input File GOLD Floes (normal and Ultra Large Docking) and Hermes generated mol2 files for GOLD docking calculations. Be careful and check that the cavity file points to the right binding area. Do not use this cavity method with SPRUCE-generated receptor input files or datasets.

d) List of residues

This file is created manually. When specifying a list of residues, the residues can be extracted from any text file, including a standard GOLD solution file (GOLD writes the active site residues list to the solution files if the output of rotatable hydrogens is turned on). For further information, please consult the GOLD documentation: Gold User Manual, section 4.6.5 Defining a Binding Site from a List of Atoms or Residues.

- Method to define de Binding Site à List of residues
- Click on the “Choose input” green box next to “Input Binding Site File”, navigate to the desired location, and select the “cavity.residues” file containing the list of residues that will be used to define the cavity.

ATTENTION: Use only with the Protein Input File GOLD Floes (normal and Ultra Large Docking) and Hermes generated mol2 files for GOLD docking calculations. Be careful and check that the cavity file points to the right binding area. Do not use this cavity method with SPRUCE-generated receptor input files or datasets.

e) Protein atom

This method will generate a sphere of a given radius with centre at the cartesian coordinates of the protein atom selected. The algorithm LIGSITE (default = ON) will be used to refine the raw cavity from the atoms in the sphere.

- Method to define de Binding Site à Protein atom
- Introduce receptor atom number in field “Atom or Residue Number for Binding Site”
- OPTIONAL: Introduce a radius value in the field: “Radius of the Binding Site”. Default = 10 Angstrom.
- OPTIONAL: Refine Binding Site Definition: The algorithm LIGSITE is used for the automatic detection of potential small molecule binding sites in proteins. Default = ON.

ATTENTION: Use only the Protein Input File GOLD Floes (normal and Ultra Large Docking) and Hermes generated mol2 files for GOLD docking calculations. Be careful and check that the cavity file points to the right binding area. Do not use this cavity method with SPRUCE-generated receptor input files or datasets.

f) Protein residue

This method will generate a sphere of a given radius with centre at the cartesian coordinates of each of the atoms in the protein residue selected. The algorithm LIGSITE (default = ON) will be used to refine the raw cavity from the list of all atoms in the spheres.

- Method to define de Binding Site à Protein residue
- Introduce the atom number of one of the atoms in the reside in field “Atom or Residue Number for Binding Site”
- OPTIONAL: Introduce a radius value in the field: “Radius of the Binding Site”. Default = 10 Angstrom.
- OPTIONAL: Refine Binding Site Definition: The algorithm LIGSITE is used for the automatic detection of potential small molecule binding sites in proteins. Default = ON.

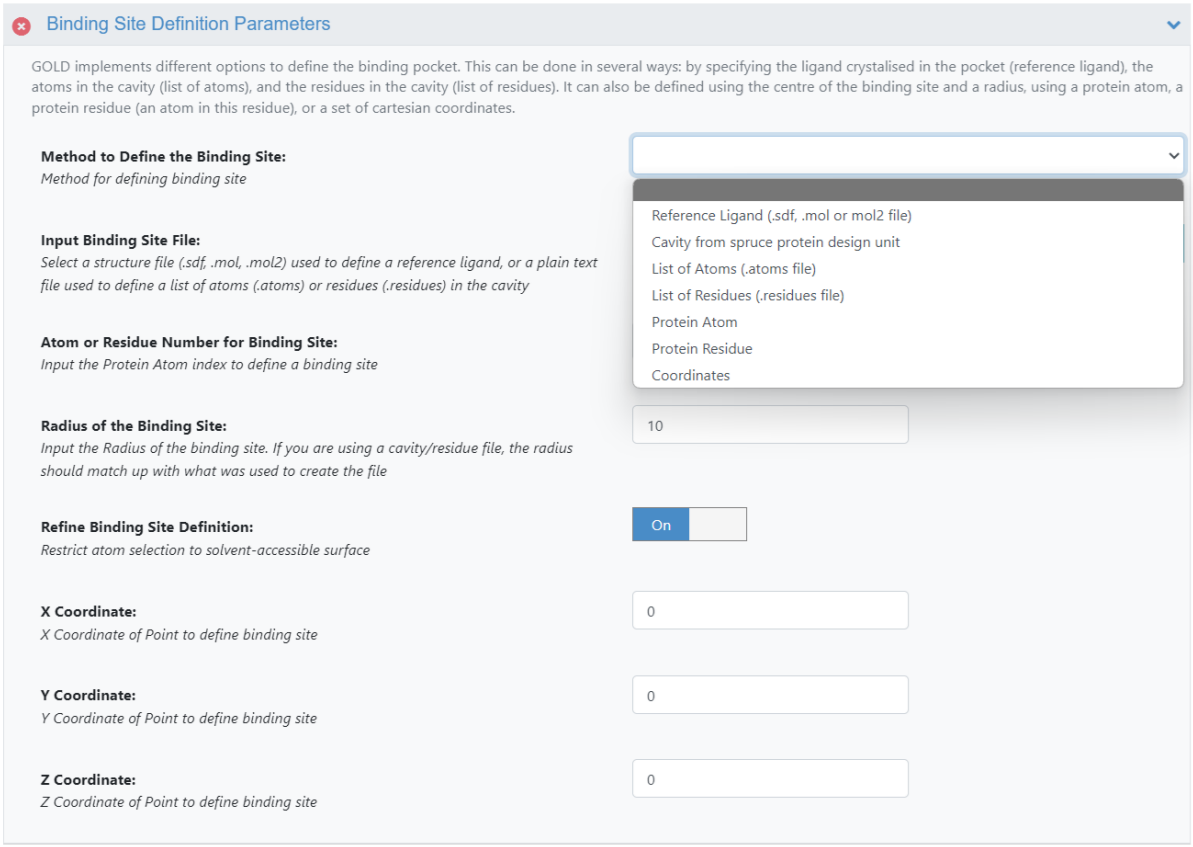
ATTENTION: Use only the Protein Input File GOLD Floes (normal and Ultra Large Docking) and Hermes generated mol2 files for GOLD docking calculations. Be careful and check that the cavity file points to the right binding area. Do not use this cavity method with SPRUCE-generated receptor input files or datasets.

g) Coordinates

This method will generate a sphere of a given radius with centre at the cartesian coordinates. The algorithm LIGSITE (default = ON) will be used to refine the raw cavity from the atoms in the sphere.

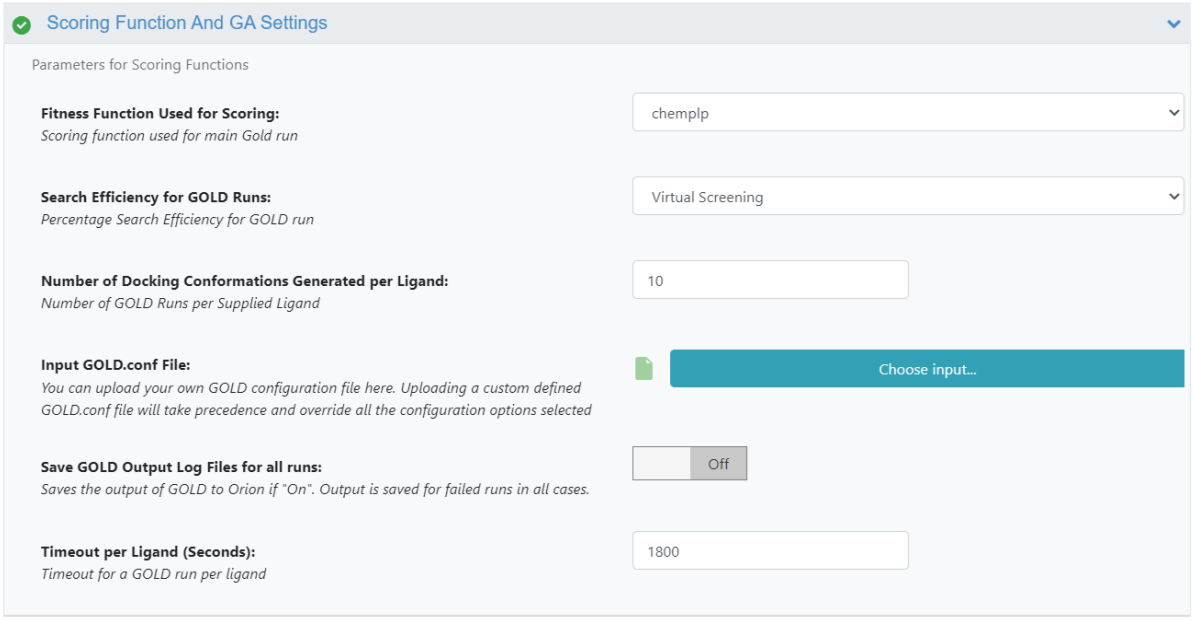
- Method to define the Binding Site à Coordinates
- Introduce the X, Y, and Z cartesian coordinates in the respective fields
- OPTIONAL: Introduce a radius value in the field: “Radius of the Binding Site”. Default = 10 Angstrom.
- OPTIONAL: Refine Binding Site Definition: The algorithm LIGSITE is used for the automatic detection of potential small molecule binding sites in proteins. Default = ON.

ATTENTION: This method can be used with the GOLD (Protein File) and GOLD (Protein Dataset) floes, both in the standard and ultra large docking versions.


Binding Site Definition Parameters
GOLD implements different options to define the binding pocket. This can be done in several ways: by specifying the ligand crystalised in the pocket (reference ligand), the atoms in the cavity (list of atoms), and the residues in the cavity (list of residues). It can also be defined using the centre of the binding site and a radius, using a protein atom, a protein residue (an atom in this residue), or a set of cartesian coordinates.
Method to Define the Binding Site:
Method for defining binding site
Reference Ligand (.sdf, .mol or mol2 file)
Cavity from spruce protein design unit
List of Atoms (.atoms file)
List of Residues (.residues file)
Protein Atom
Protein Residue
Coordinates
Input Binding Site File:
Select a structure file (.sdf, .mol, .mol2) used to define a reference ligand, or a plain text file used to define a list of atoms (.atoms) or residues (.residues) in the cavity
Atom or Residue Number for Binding Site:
Input the Protein Atom index to define a binding site
Radius of the Binding Site:
Input the Radius of the binding site. If you are using a cavity/residue file, the radius should match up with what was used to create the file
10
Refine Binding Site Definition:
Restrict atom selection to solvent-accessible surface
On
X Coordinate:
X Coordinate of Point to define binding site
0
Y Coordinate:
Y Coordinate of Point to define binding site
0
Z Coordinate:
Z Coordinate of Point to define binding site
0

Scoring Function and Genetic Algorithm (GA) Parameters

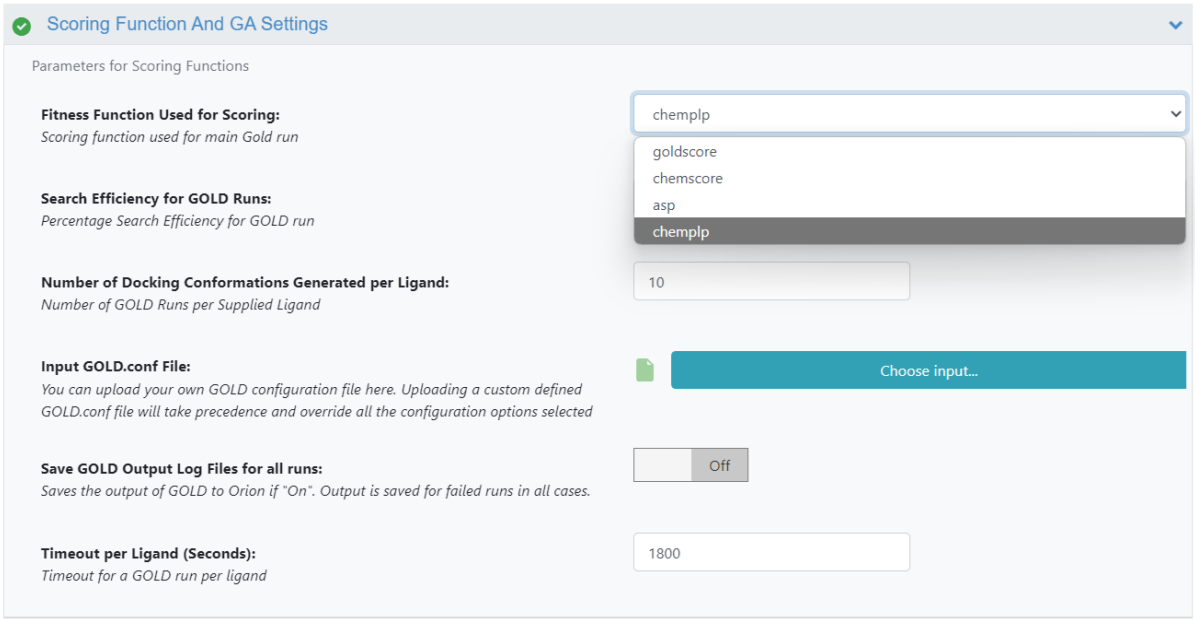
This section allows the user to configure the docking calculation by selecting the scoring function and search efficiency among the other options available.



- Fitness function used for Scoring:

GOLD implements four scoring functions: ChemPLP (default), GoldScore, ChemScore, and ASP. Detailed information about them can be found in the GOLD User Manual, section 8, Fitness Function: GOLD User Guide (cam.ac.uk)

Click the drop-down menu and select the scoring function. ChemPLP is currently the by-default option as it shows in general better performance.



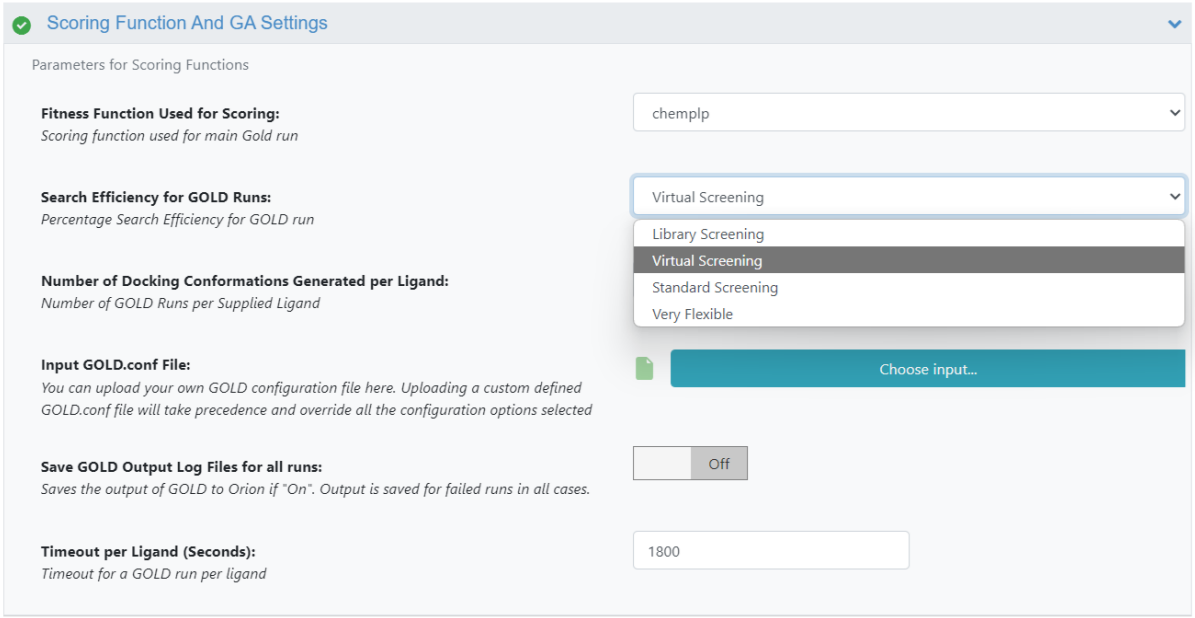
- Search Efficiency for GOLD Runs:

This parameter controls the level of exhaustiveness of the genetic algorithm used in the docking search. The complexity in the desktop app has been simplified to allow a better experience while maintaining good performance at the same time.

For more information about the search efficiency in GOLD, please read the documentation in the GOLD User Manual, section 12.3, Controlling Accuracy and Speed with Genetic Algorithm Parameter Settings: GOLD User Guide (cam.ac.uk)

There are four options:

- Library Screening: this sets the search efficiency at 10%. This is the fastest setting and therefore is the least reliable. Ideal for HTS with large and ultra-large libraries to filter out unlikely drug candidates.
- Virtual Screening (default): this sets the search efficiency at 30%. This setting is suitable for routine work and usually gives comparable predictive accuracy to the slower settings, unless the ligand has a large number of rotatable torsions.
- Standard Screening: this sets the search efficiency at 100%. GOLD will attempt to apply optimal settings for each ligand.
- Very Flexible: this sets the search efficiency at 200% and is recommended for large, highly flexible ligands. This setting delivers high predictive accuracy but is relatively slow. This option is not recommended for HTS with large and ultra-large libraries.



- Number of Docking Conformations Generated per ligand

This parameter controls the number of generated poses per ligand. For normal-size libraries, the by-default value is 10, however, for docking campaigns with large and ultra-large libraries, it may be reasonable to set it to something smaller to speed up the calculations.

- Input GOLD.conf File

The GOLD integration in Orion® includes a subset of all the functionalities in the desktop app. However, there is an option for the users to access some of them via the gold.conf file generated locally when preparing a docking calculation.

After some minimal edits, that configuration file can be uploaded to the Orion® platform and used for a docking calculation. This method will open the door to many parameters and user customizations including access to constraints and restraints, receptor Hydrogen bonds, and receptor flexibility.

ATTENTION: Some of the features in GOLD not available in the Orion® menu that require files for working will not work yet even if you upload a configuration file.

The gold.conf file configuration settings have precedence over the menu options, so, if you upload a configuration file, no parameter from the ORION® menu will be used.

Use preferably with the Protein Input File GOLD Floes (normal and Ultra Large Docking) and Hermes generated mol2 files for GOLD docking calculations.

As this is not the preferred mechanism to set docking calculations in Orion®, the gold.conf option has not been fully tested. Use this option with caution.

- Save GOLD Output Log Files

If ON, this switch will instruct the Floe to save all the configuration, results, log, and error files in a zip folder for analysis in case the calculation fails.

Default option: OFF

ATTENTION:

This zip file can be extremely large in the case of large and ultra-large docking calculations. Activate this option with caution.

- Timeout per Ligand (seconds)

This parameter sets a reasonable time limit per ligand for a docking calculation. If for any reason the calculation gets stuck with a problematic ligand, blocking or slowing down the job, the floe will stop trying to dock the ligand and will continue with the next one.

Covalent Docking

This section allows the user to configure the covalent docking calculations.

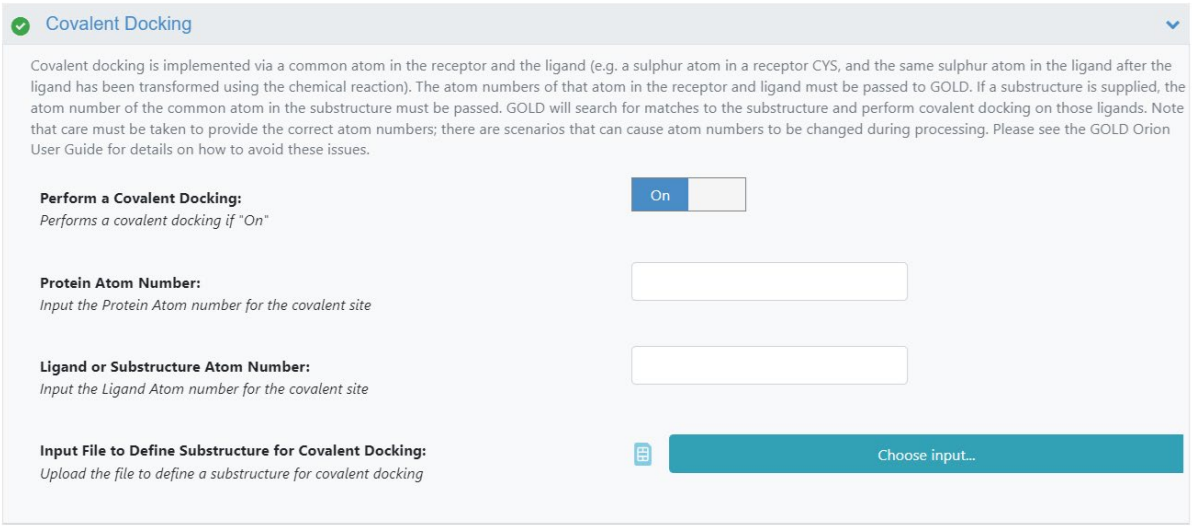
To activate this functionality, switch “Perform a Covalent Docking “ ON (default: OFF).

GOLD implements covalent docking in two ways: A covalent link for use with individual ligands (single ligand), and a substructure-based covalent link for use with multiple ligands which have a common functional group (multiple ligands). For more information about the covalent docking implementation in GOLD, please, read the documentation in the GOLD User Manual, section 6.6, Setting Up Covalently Bound Ligands: GOLD User Guide (cam.ac.uk)

This method requires a common atom between the receptor and the ligand. For example, if it is a reaction (image below) between a sulphur (CYS) and a Michael acceptor (alpha-beta unsaturated carbonyl)

the ligand should first be modified to look like the final product after the reaction has taken place. That means, a sulphur atom is added to the ligand (only the S atom is added, not the rest of the CYS side chain) to act as the common atom.

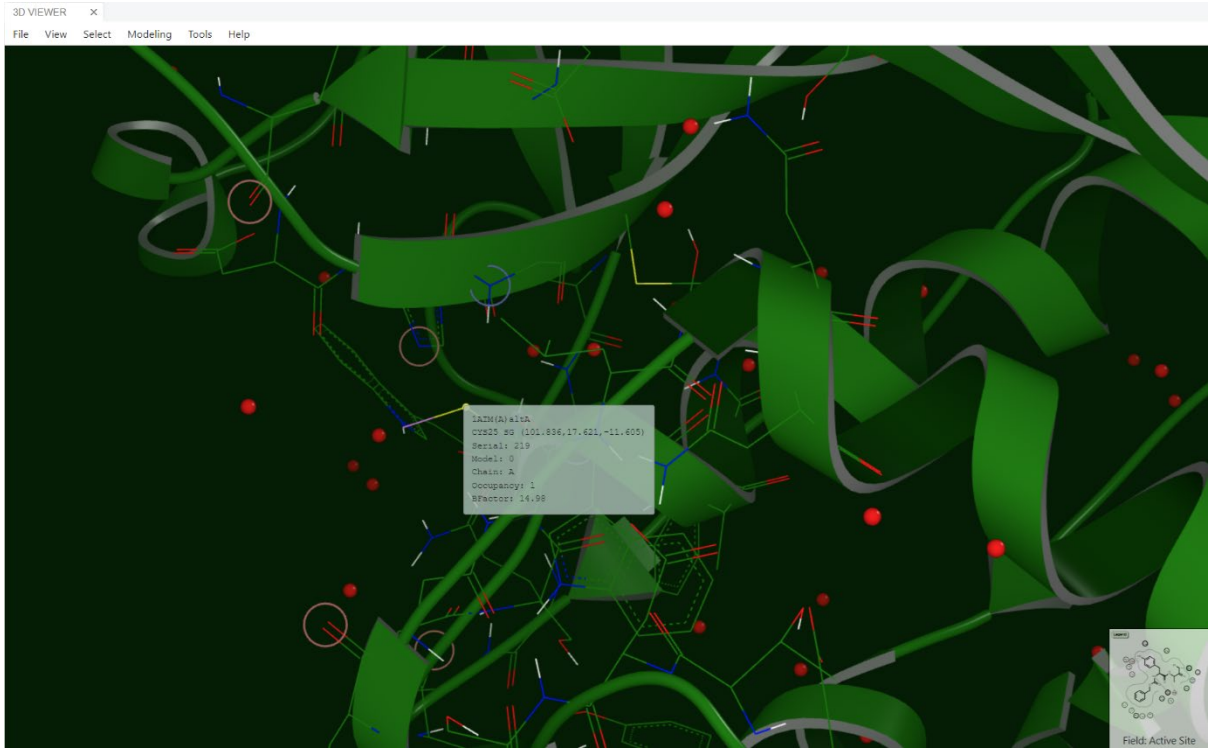
GOLD implements the covalent bond by adding a strong constraint to keep the common atom in both the receptor and ligand overlapping. For this, the user needs to introduce the ligand (or substructure) and protein atom indices in the specific fields; “Protein Atom Number” and “Ligand or Substructure Atom Number”:



The ligand atom index can be easily identified by opening the file with a text editor and reading the value in the file, or using a molecular viewer.

Similarly, the protein atom number can be identified by opening the file with a text editor and reading the value in the file or using a molecular viewer.

When preparing the receptor file in the Orion® platform, the receptor atom number can be identified using the "3D" modelling tab (left panel, towards the top): First, curate the raw pdb file with SPRUCE, view the output file generated as this file will be the input to the GOLD Floe (Protein Input Dataset). Click on the receptor atom selected to be the common atom (normally a Sulphur, or an Oxygen), and now, make ATL+click on the same atom. A small floating panel will be displayed showing the atom metadata.



The atom index required for the covalent docking constraint is the "serial" value, in this case 219.

Type of covalent docking implementations:

- A covalent link for use with individual ligands (single ligand)

This is the simplest covalent docking calculation. Follow the next steps to run it:

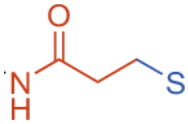
- Prepare the receptor and ligand input files as explained.
- Select the required job options in the flow menu.
- In the covalent docking section: activate the covalent docking calculation (switch to ON)
- Identify the common atom indices in receptor and residue,
- Introduce them in the "Protein Atom Number" and "Ligand or Substructure Atom Number" fields in the menu.
- Run.

- A substructure-based covalent link for use with multiple ligands which have a common functional group (multiple ligands).

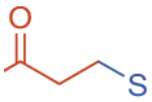
This is the implementation to run a covalent docking calculation with one receptor and an entire library of ligands.

If the user were to manually perform many single-ligand covalent docking calculations, they will need to run them one by one introducing each time a new atom index for each new ligand while everything else remains unchanged. This would be repetitive and prone to human errors. To make the calculation more efficient and user-friendly, GOLD uses a "substructure". This substructure is a mol2 file containing a few atoms defining the common substructure all ligands should match to react. For example, in the Michael addition to Cysteine:

The substructure would be something like this:



or even like this, if we needed a more generic pattern:



The substructure is not required to be a full molecule (there may be missing Hydrogen atoms and open valences) but all the atom types and connectivity must be correct.

In this type of docking, the "ligand atom index" is replaced by the "substructure atom index", which is the index of the common atom in the substructure, in this case, the Sulphur. When this type of calculation is selected, GOLD will perform a substructure search on each ligand in the library. If the substructure is present, GOLD will map the atoms and perform the covalent docking applying the constraint to the right common atom. If the substructure is not present, the ligand will be discarded, and GOLD will try with the next one in the library.

To run a covalent docking calculation with a substructure, follow the next steps:

- Prepare the receptor and ligand libraries as explained.
- Prepare a substructure mol2 file containing the key atoms in the substructure.
- Select the required job options in the flow menu.
- In the covalent docking section: activate the covalent docking calculation (switch to ON)
- Identify the common atom indices in receptor and substructure,
- Introduce them in the "Protein Atom Number" and "Ligand or Substructure Atom Number" fields in the menu.
- Run.

ATTENTION:

Using the right atom indices is crucial to perform a successful covalent docking calculation. During our tests we discovered that the atom indices in proteins may be altered in certain scenarios during the preparation process depending on the tools used (Hermes, Orion®, etc), the type of files (pdb, mol2) uploaded, and how they were generated.

To minimize potential problems related to unexpected atom renumbering, we strongly advise using one of these two options:

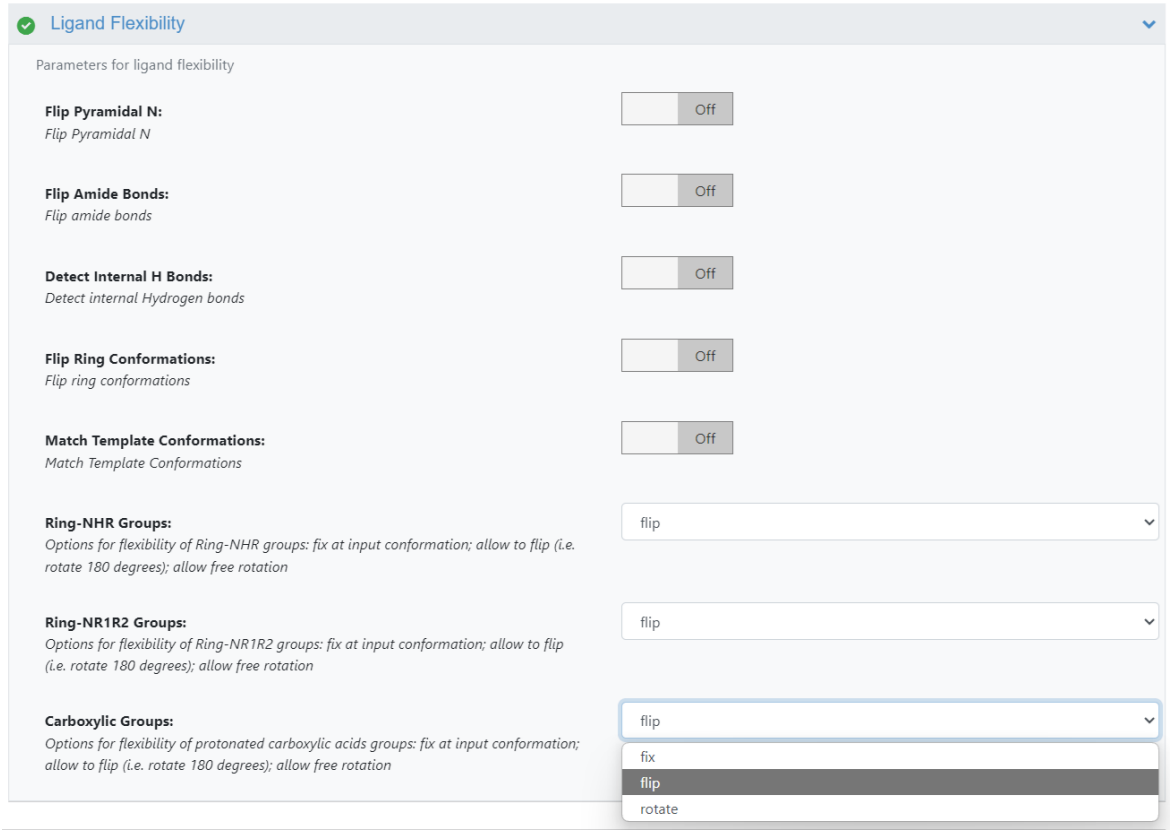
- If the protein input file is generated with SPRUCE in the Orion® platform from the raw pdb file, please use the "GOLD Floe (Protein Input Dataset)" and read the atom index as explained above. IMPORTANT: if the input file generated with Spruce contains multiple copies of the protein in the dataset, select ONLY ONE for the covalent calculation.
- If the protein input file is generated with Hermes and then uploaded to the ORION® platform, please, use the "GOLD Floe (Protein Input File)" and read the atom index in the mol2 file generated with Hermes. IMPORTANT: Be aware that opening this mol2 file and saving it again may cause an atom renumbering due to Hydrogen atoms being moved next to their residues. Please, upload the mol2 file generated by Hermes "as is".

Ligand Flexibility

This section allows the user to configure the ligand flexibility.

During a docking calculation, GOLD performs operations on the molecular torsions to generate new poses. GOLD implements as well other type of conformer generation mechanisms to explore the conformational space. This includes exploring ring conformations, inverting pyramidal nitrogen atoms, flipping amide bonds, and detecting internal hydrogen bonds. In the case of hydroxyl groups in carboxylic acids (-COOH), and secondary (Ring-NHR1) and tertiary (Ring-NR1R2) amino groups connected to aromatic rings, there are options to fix, flip and rotate.

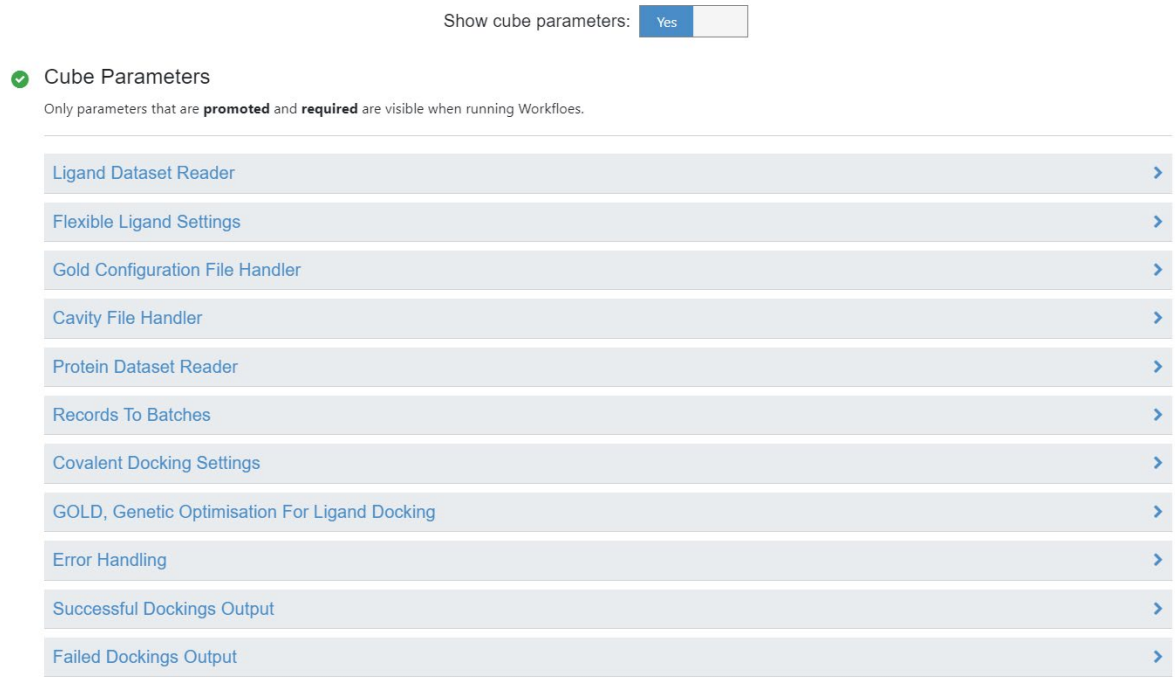
For more information about the ligand flexibility implementation in GOLD, please, read the documentation in the GOLD User Manual, section 9, Ligand Flexibility: GOLD User Guide (cam.ac.uk)



GOLD Cube Parameters

The GOLD Cube parameters can be displayed by enabling the option in the switch.

They are advanced technical options related to the cube configuration. They can be safely ignored by most users most of the time unless you have particular needs or you are a cube developer.



Release Notes

2025.1 Maintenance Release

- Updated version of GOLD to match latest desktop release: GOLD version 2025.1
- Updated to use Ubuntu 22 as base OS for Orion package
- Covalent docking enhancements in GOLD (only via Hermes)

2024.3 Release

- Initial release
- GOLD version 2024.1
- Functionalities available in Orion: Protein-ligand docking, pose prediction, multiple cavity selection methods, multiple scoring functions (ChemPLP, ChemScore, GoldScore and ASP), non-covalent & covalent docking, and ligand flexibility.
- Functionalities available in Orion via GOLD configuration file: receptor flexibility and multiple constrains (if they don't require extra files).

For further support on GOLD:

The Cambridge Crystallographic Data Centre

www.ccdc.cam.ac.uk

support@ccdc.cam.ac.uk

For further support on Orion:

OpenEye Scientific

www.eyesopen.com

support@eyesopen.com